Using Twitter data to understand the public response to a housing crisis

An analysis of 96 million tweets sent from London to understand what drives the public reaction to the London housing crisis

Master Thesis
Business Information Management

Rotterdam School of Management
Erasmus University Rotterdam

Ties Hagdorn
407131

Supervised by Tobias Brandt
Co-read by Anna Priante

June 30, 2020

Acknowledgements

I would like to thank...

> my thesis supervisor Tobias Brandt and co-reader Anna Priante for their invaluable guidance, advice and feedback.
>
> my friends and family for going out of their way to provide feedback on the draft version of this thesis.
>
> Marijne for her continuous encouragement and patience.
>
> my twin brother Ewout for keeping my spirits high since our simultaneous beginning.
>
> my dad for his sincere support and interest, and for allowing me to learn and study without worries for the past 20 years of my life.
>
> my mother, despite not being with us anymore, for being my inspiration in life and for the ideas that she provided through our connected thoughts.

Without you, I would not have been able to write this thesis.

Thank you.

Abstract

Providing housing for an increasing number of citizens is a challenging task for cities across the world in the face of global urbanisation. A metropolis that is currently dealing with a housing crisis, is London. Increased demand for housing outpaced supply, driving up the cost of houses 600% over the past 25 years. Understanding the drivers of public opinion on a housing crisis is valuable because it aids governmental bodies to undertake actions on the most critical components of a crisis and provides insights into how the public opinion may develop. This research analyses 96 million tweets sent from London from 2012 to 2018 as a novel technique to construct a representation of London's opinion on the housing crisis. Along with Twitter data, various housing-related variables served as input for two regression models to estimate the relationship between shifts in the London housing market and changes in tone and size of the public opinion on the housing market. The results show that the London housing market has become an increasingly discussed topic on Twitter and that the tone of the housing tweets has become more negative over the years. The regression models prove that housing cost, housing demand, annual income and the rate of homelessness have significant effects on both the tone and number of housing tweets sent from London.

Contents

List of Figures

List of Tables

Glossary

Abnormal sentiment The level of sentiment of a housing tweet compared to the average sentiment of all tweets.

Acceptance as homeless Household deemed in priority need through no fault of their own. Also known as *'statutory homelessness'*.

Borough A local authority districts in London. Acronyms for all 33 London boroughs are found in appendix A.

Dwelling A unit of accommodation which may comprise one or more household spaces.

GLA Greater London Authority, the governance body of London.

Household One person or a group of people who have the accommodation as their only or main residence.

Housing variable A variable that is considered to be a component of the housing market.

Housing tweet A Twitter message that is considered to be related to the housing market, because it matches the criteria specified in appendix E.

HPI Short for House Price Index, which measures the price changes of housing as a percentage change from January 2015 (which has HPI of 100).

Overcrowding A household is said to be overcrowded if they have fewer bedrooms available than the notional number needed.

Sentiment An attitude towards something. The sentiment of a text is a measure of tone, ranging from -1 (most negative) to +1 (most positive). Obtained through sentiment analysis.

Sentiment analysis The use of text analysis to identify, extract and quantify subjective information from text.

Sentiment driver A factor that affects a person's attitude or opinion towards something.

1 Introduction

The majority of the world population lived in rural areas in small communities for most of mankind's history. This trend has shifted dramatically in recent decades. In 1960, twice as many people lived in rural areas as in urban settings across the world (United Nations, 2007). Only 47 years later, in the year 2007, the United Nations estimated that for the first time in history, more than half of the world's population lived in cities. Nowadays, 55% of the world's population lives in urban areas, and this number is expected to increase to 68% by 2050 (United Nations, 2018).

In the face of rapid urbanisation, a growing concern for cities is providing housing to its increasing number of citizens. Demand for housing increases because of the growing population, but in many cities not enough houses are built to fulfil this increasing demand, due to limited physical space and building capacity. This shortage of housing drives up the house prices, causes overcrowding, and causes homelessness (Petkar, Macwan, & Takkekar, 2012). The problem of housing is a growing global concern and is already considered a crisis in various major cities like San Francisco, Amsterdam, Sidney, New York and London (Nijskens et al., 2019).

Various reports have shown the severity of the housing crisis in London. The Greater London Authority (GLA) stated that from 1997 to 2017, the population of London grew by 27%, and the number of jobs in London grew by 45%. Over the same period, the number of houses grew by only 18% (GLA, 2019). This mismatch of supply and demand has caused a surge in London house prices, which rose by 600% over the past 50 years, adjusted for inflation (GLA, 2019). The rising cost of housing has caused Londoners to spend an increasingly large amount of their disposable income on housing. An average Londoner had to spend 49% of their pre-tax monthly income on rent to live in a single bedroom house in London in 2019 (Trust for London, 2020). This number is 58% above the UK average, indicating the severity of the situation in London compared to the rest of the country. For some Londoners, these high costs have resulted in homelessness: one in 235 London households was accepted as homeless and in priority need in 2017 (Ministry of CLG, 2020). These statistics are reasons for 81% of Londoners to agree in 2018 that the London housing market is in a state of crisis (Page, 2018). A housing crisis is detrimental for a city, as proper housing is fundamental to the well-being of people and the functioning of society. Housing is even considered a basic human right by the Universal Declaration of Human Rights (1948):

> *"Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care and necessary social services..."*

Due to the nature of a crisis, the root cause of a crisis is complicated to tackle (Pearson & Clair, 1998). For the London housing crisis, this is not different: the solution is clear, but challenging to execute. The core to solving the London housing crisis is to build new houses at a rate that outpaces the demand for new houses (Marsden, 2015). However, this is easier said than done, because restricted space and building capacity limit the construction of houses, while demand keeps rising (GLA, 2019).

When tackling the root-problem of a crisis is complicated and time-intensive, various frameworks on crisis management suggest to focus on the symptoms of a crisis (Pearson & Clair, 1998; Hay, 2013). The evident symptoms of the London housing crisis are portrayed through the dissatisfaction of Londoners, as a consequence of the rising housing cost, overcrowding and homelessness (Page, 2018). Understanding changes in the public opinion as a symptom of a crisis can provide great value in managing the crisis (Fritsche, Jonas, & Kessler, 2011). City authorities can use the public opinion to prioritise what components of the crisis to focus on, to most effectively improve the public view. Systematically improving the public view of a crisis will naturally increase the overall satisfaction of the city population. Therefore, the importance of understanding the development of public opinion on such a relevant societal matter should not be underestimated.

Public opinion is often traditionally measured through large scale surveys. However, surveys are often time-consuming and costly to administer. Another critical shortcoming of surveys is the abundant presence of various types of response biases (Fowler, 2009). Over the past decades, new and innovative methods to obtain public opinion have emerged as an alternative to surveys These novel methods involve the collection and analysis of massive amounts of opinionated data, or *opinionated big data*. These novel techniques leverage opinionated big data to extract the public opinion in a more time- and cost-efficient manner than traditional survey methods. A source of opinionated big data is the social media platform Twitter.

Twitter has become an important way to communicate thoughts and activities. Over 300 million monthly active Twitter users send over half a billion Twitter messages (tweets) on an average day (Twitter, 2019). To put these numbers in perspective; if you are an average reader, over 500 000 tweets were sent since you started reading this introductory section. Big data analysis techniques are capable of collecting and analysing millions of tweets to obtain rich information. Though a tweet is limited to only 140 characters, tweets can be analysed in large quantities to construct a representation of the public opinion of *the real world*, which may reveal emerging patterns about how the masses vocalise their thoughts and opinions (O'Connor et al., 2010; Bollen, Pepe, & Mao, 2009; Cody et al., 2016). Twitter is as a valuable source of opinionated big data because tweets contain textual, temporal and spatial information. Using this information, tweets can be analysed to study *what* is said, *when* it is said, and *where* it is said. This study analyses 96 million tweets sent from London in the period of 2012 to 2018, to understand the public opinion on the London housing crisis. This research leverages the spatial aspects of the tweets by determining from which London borough the tweet was sent, which allows for comparing the sentiment and proportion of housing tweets between geographical areas.

The objective of this research is to investigate how the public opinion on the London housing market has changed over the years, and to understand the effect of shifts in the London housing market on the public opinion on the London housing market. This research defines *the public opinion* on the housing market in two parts. The first part is the tone of the discussion, measured through the level of positivity or negativity of housing-related tweets. The tone of a tweet is hereafter referred to as *sentiment*, which will be calculated by performing *sentiment analysis*. The second part of the public opinion is the size of the discussion, which is measured by the proportion of housing-related tweets. A housing-related tweet is a tweet that contains information about the housing market. Hereafter, these tweets are referred to as *housing tweets*.

The aforementioned research objective and motivation for this study are summarised in the following two research questions that will guide this research and will fulfil the research objective when answered:

RQ1 *How have the sentiment of housing tweets and the proportion of housing tweets sent from London changed over 2012-2018?*

RQ2 *What is the effect of shifts in the London housing market on the sentiment of housing tweets and the proportion of housing tweets sent from London over 2012-2018?*

In an attempt to guide the reader towards the answer to these research questions, this study is structured as follows. Chapter 2 reviews relevant literature, by firstly discussing how prior works have constructed the public opinion on the housing market, and identifying what housing-related variables drive housing sentiment (2.1). Section 2.2 reviews literature that studied how reflective Twitter is of the public opinion, confirming the use of Twitter as a source of opinionated data. Section 2.3 discusses the research gap this thesis fills.

Chapter 3 elaborates on the housing-related variables that were identified in the literature review. For each housing-related variable, chapter 3 develops a theory on how the variable is expected to affect the sentiment and proportion of housing tweets in London. The relationships are translated into hypotheses which will be investigated in this research to answer research question two.

Chapter 4 describes the data required to test the hypotheses, how this data is collected and prepared, and describes the methods used to analyse the data. In addition, this chapter explains the theory behind the statistical models that will make inferences about the effect of housing-related variables on the sentiment and proportion of housing tweets.

Chapter 5 describes the results of this research. Firstly, a descriptive analysis (5.1) shows how the sentiment and proportion of housing tweets have changed over the years, answering research question one. Secondly, two statistical models will estimate the effect of housing-related variables on the sentiment and proportion of housing tweets (5.2), which answers research question two.

Chapter 6 concludes this thesis by summarising the main findings, and subsequently discussing and interpreting the main findings. Afterwards, the implications of this research for various stakeholders are presented. Finally, the limitations and suggestions for future research are discussed.

2 Literature Review

This chapter reviews literature that composes the underlying theory of this research. Section 2.1 discusses different methods to measure housing sentiment and summarises the shortcomings of these methods. Simultaneously, this section aims at identifying which housing-related variables drive housing sentiment. Section 2.2 discusses literature that studied how reflective Twitter is of the public opinion. Section 2.3 concludes the literature review by summarising the most important findings, discussing the identified research gaps, and suggesting how this thesis will fill the identified research gaps and will overcome shortcomings of prior works.

2.1 Identifying drivers of housing sentiment

Housing sentiment is defined as *people's opinion about the housing market* (Case, Shiller, & Thompson, 2014). Various researchers have attempted to identify what exactly drives housing sentiment by studying how housing sentiment indices correlate to a range of housing market measures. This section particularly aims at providing an overview of what housing-related variables have been shown to drive housing sentiment. Understanding these drivers allows this thesis to assess which housing-related variables will be used to answer research question two. The housing-related variables that are identified as drivers of housing sentiment are hereafter referred to as *housing variables*.

Marcato and Nanda (2016) identify two primary methods to measure housing sentiment. The first method is to directly ask consumers about their opinions on the housing market through the use of surveys. The second method is indirectly measuring housing sentiment through underlying proxies.

2.1.1 Measuring housing sentiment directly through surveys

Surveys are a traditional way to obtain the public opinion a subject, and this is no different for the housing market. Housing surveys ask consumers various questions about their opinion on the housing market. The answers to these questions are quantified to indicate consumer sentiment towards the housing market. Various researchers have tried to identify the drivers of housing sentiment by studying how the results from housing surveys correlate with several housing-related measures.

Wilcox (2015) conducted a survey to construct a housing sentiment index. The resulting index was compared to a range of traditional housing market measures, which showed that house prices were negatively correlated to housing sentiment. Income, home buying and selling conditions and mortgage rate expectations were found to be positive indicators of housing sentiment.

Bork and Moller (2016) also administered a survey regarding home-buying conditions to construct a housing sentiment index. The index showed that increasing house and rent prices have a negative effect on housing sentiment. However, some respondents saw the rising house prices as beneficial, because the value of their property rose. The survey responses also showed that the economic status of a household is positively correlated with housing sentiment.

Instead of constructing their own housing survey, other researchers have used the results of existing surveys to construct housing sentiment indices to study how these indices correlate with housing-related variables. The *Reuters Survey of Consumers*, *Architecture Billings Index*, *Wells Fargo Housing Market Index* and *Fannie Mae National Housing Survey*, are examples of surveys that ask respondents about their housing-related views. Various studies used the results of these existing surveys to identify drivers of housing sentiment, which are elaborated on below.

Dua (2008) established a sentiment index by analysing the results of a housing survey conducted by the University of Michigan. The sentiment index was subsequently compared to various assumed determinants of housing sentiment to study correlations. The paper showed that house prices are negatively correlated with housing sentiment. Expected real disposable income and wealth were found to be positively correlated with housing sentiment.

Croce and Haurin (2009) researched the predictive power of housing sentiment constructed from the results of two housing surveys. The study showed that both surveys performed relatively well in predicting home permits, housing starts, and new home sales. However, Goodman (1994) found that the results of two different housing surveys were insignificantly correlated with the number of home sales and housing starts.

Using traditional surveys to measure housing sentiment have various shortcomings. Firstly, Wang and Hui (2017) mention the shortcoming of information asymmetry in the housing market between the demand (buyers/consumers) and supply-side (sellers/builders). The different quality and quantity of information between these groups may cause biased responses in surveys that measure housing sentiment. Secondly, a typical housing survey asks reflective questions about how the consumer thinks the present housing market compares to the past or future (Marcato & Nanda, 2016). A typical question is "*do you think today is a good time to buy a home, compared to last year?*". These type of reflective questions require the respondents to recall how they felt at a different point in time, which can result in a distorted representation of reality because respondents do not remember how they felt. Thirdly, surveys in general are prone to certain biases and have other disadvantages (not specific to surveys measuring housing sentiment). The non-response bias occurs when respondents disproportionately posses certain traits which can affect survey results, which is in line with the information asymmetry shortcoming presented by Wang and Hui (2017). Another survey shortcoming is the *demand characteristics bias*, which occurs when respondents alter their response merely because they are responding to a survey (Nichols & Maner, 2008). Fourthly, surveys are costly and timely to administer and require a population of willing participants (Fowler, 2009).

The aforementioned shortcomings of using surveys to measure housing sentiment were reason for academics to explore alternative methods. According to Marcato and Nanda (2016), these alternative methods can be grouped under the term *indirect methods*.

2.1.2 Measuring housing sentiment indirectly through proxies

This section discusses studies that have tried to measure housing sentiment without the use of surveys. The methods in these type of studies do not *directly* ask consumers about their sentiment on housing, and the methods are therefore named *indirect measures*. These indirect measures use various proxies in an attempt to construct housing sentiment. Using indirect methods to measure housing sentiment can be a difficult task, because typical sentiment proxies that exist for other markets (e.g. mutual fund flows as a proxy for the stock market (Baker & Wurgler, 2006)) are not as readily available for the housing market (Soo, 2013). Additionally, the indirect measures of the housing market are based on a relative small amount of transactions because houses are traded infrequently.

Notwithstanding these challenges, various researchers have attempted to measure housing sentiment through various indirect indicators. This study discusses these studies and highlights the drivers of housing sentiment that these studies identified.

Hui and Wang (2014) used over two million housing transactions records from Hong Kong between 1991 to 2011 to establish a sentiment index. Subsequent research by Wang and Hui (2017) used this index to find that housing sentiment is negatively correlated with house prices and rent, but positively correlated with household income and the size of the housing market.

Soo (2013) analysed the tone of housing-related news as an indirect measure for housing sentiment. The study showed a negative correlation between the tone of housing-related news and the house price index, and showed a positive correlation between housing sentiment and housing market size (transaction volume), housing supply (construction starts), and housing demand (population).

To conclude, the studies in this section 2.1 have identified that house prices negatively drive housing sentiment. Housing variables that positively drive housing sentiments are income, housing starts, housing demand, and housing sales. An indirect measure of housing sentiment that is yet unexplored is the use of Twitter. While Twitter has been playing an increasingly larger role in academic research (Zimmer & Proferes, 2014), the platform is yet to be used to obtain insights about the housing market. This study uses Twitter as a novel method to construct the public opinion on the housing market, and investigates the effect of the aforementioned housing-related variables on the public opinion of Londoners on the housing market. The following section (2.2) reviews literature that researched the potential of Twitter to measure public opinion.

2.2 How Twitter reflects the public opinion

Understanding the validity of Twitter as a measure of public opinion is core to this research, because this study uses tweets to construct the public opinion on the housing market. Section 2.2.1 firstly presents the advantages of Twitter as source of opinionated data. Section 2.2.2 discusses literature that studied the validity of tweets as a source for public opinion. Then, section 2.2.3 highlights studies that used this correlation to measure the public opinion on phenomena.

2.2.1 Tweets as a source of opinionated data

Twitter presents itself as a valuable source of opinionated data for three main reasons. Firstly, Bollen, Pepe, and Mao (2009) argue that the timeliness of Tweets can be advantageous. Tweets are so brief that they are generally associated with a specific thought at a specific moment, which is the timestamp of the tweet. Therefore, tweets reflect a much smaller temporal window than other types of user-generated texts.

Secondly, collecting large quantities of data on Twitter is relatively cost and time-efficient because the data is public and free, and can be collected in various ways (Kumar, Morstatter, & Liu, 2014). The method that this thesis uses to collect tweets will be discussed in chapter 4.

Thirdly, the content of tweets is very diverse. Twitter is used for daily chatter, open conversation, information sharing, public opinion, political discussion, and more (Dann, 2010). The content of a tweet conveys valuable information about the mood and opinion of the user that posted it (Bollen, Pepe, & Mao, 2009).

2.2.2 Is Twitter reflective of the public opinion?

Three prior works have shown that large quantities of tweets from different users can be aggregated to represent the public opinion (O'Connor et al., 2010; Bollen, Pepe, & Mao, 2009; Cody et al., 2016). These three studies are discussed in more depth in this section.

O'Connor et al. (2010) researched how Twitter sentiment corresponds with measures of public opinion obtained from polls. They performed a sentiment analysis on one billion tweets sent from 2008 and 2009. The aggregated sentiment of these tweets was found to correlate well (r = 0.73) with a renowned measure of consumer sentiment (*Gallup daily*). When 30-day smoothing was applied, the correlation increased further to 79%. These results were reason to conclude that aggregated Twitter sentiment captures large-scale public sentiment trends and that publicly available Twitter data can be a sound alternative to costly and time-intensive polling.

Bollen, Pepe, and Mao (2009) supported these findings by studying how the public sentiment as observed on Twitter compares to fluctuations in the stock market, crude oil price indices and major events in the media. A sentiment analysis was performed on 9 million tweets that were collected through a rule-based approach over a four-month period. The sentiment of all tweets was aggregated to a daily scale, and was found to correlate significantly with various measures of social, political, cultural and economic indicators of public mood. Based on these results, Bollen, Pepe, and Mao (2009) concluded that the sentiment level of aggregated tweets is able to reflect public moods and opinion.

In line with the two prior studies, Cody et al. (2016) investigated the extent to which tweets can be used to complement traditional public opinion surveys. The study also performed a sentiment analysis on tweets that were collected over a 7-year period. The resulting aggregated Twitter sentiment was compared to the Michigan Index of Consumer Sentiment data, showing a correlation of 67%. The study concluded that Twitter sentiment correlated reasonably well with the surveyed consumer sentiment.

2.2.3 How has Twitter been used to measure public opinion?

The three studies mentioned in the prior section have shown the correlation between Twitter sentiment and public sentiment, and all confirmed the validity of using Twitter sentiment to construct public opinion (O'Connor et al., 2010; Bollen, Pepe, & Mao, 2009; Cody et al., 2016). This confirmed relationship served as a motivation for a larger body of work to use Twitter to study the public opinion on various topics.

Firstly, Twitter has proven to be a source of valuable insights on *broad topics*. Twitter has been used to measure the public opinion on various refugee crises (Öztürk & Ayvaz, 2018; Pope & Griffith, 2016), climate change and natural disasters (Cody et al., 2015), terrorism (Kounadi et al., 2015) and nuclear-related issues (Kim & Kim, 2014). All these studies have collected tweets and performed sentiment analysis to observe how the public opinion on these topics has changed over time. This thesis studies the public opinion on the housing market over time and therefore falls within the category of studies that research the public opinion on *broad topics*.

Secondly, Twitter has also been studied to measure the public opinion on more *specific* events. For example, Magdy, Darwish, and Abokhodair (2015) analysed 900 000 tweets to observe the public response to the 2015 Paris attacks and how sentiment towards the Islam and Muslims changed on Twitter in the subsequent 50 hours after the attacks. Chew and Eysenbach (2010) researched the response of Twitter to the H1N1 virus by analysing H1N1-related tweets, which enabled health authorities to respond to concerns raised by the public on Twitter.

Besides providing descriptive insights on the public opinion, as illustrated in the paragraphs above, the predictive power of Twitter on the real world has also been studied extensively. Various attempts were made to predict movements of financial markets using Twitter data (Ranco et al., 2015; Si et al., 2013). Bollen, Mao, and Zeng (2011) managed to predict the daily up and down directions of the Dow Jones using Twitter data and achieved an accuracy of 86.7%. Asur and Huberman (2010) used tweet sentiment and volume to predict box-office revenues of movies before their release. The results of the models that used Twitter data outperformed the Hollywood Stock Exchange predictions. Finally,various studies have used tweets to predict election results (Bermingham & Smeaton, 2011; Tumasjan et al., 2010).

2.3 Literature conclusion and research gap

Section 2.1 has shown how studies have measured housing sentiment. More importantly, the section provided an overview of which housing variables drive housing sentiment. The section concluded that house prices negatively drive housing sentiment, and that income, housing starts, housing demand, and housing sales are positive drivers of housing sentiment. Section 2.2 showed that Twitter is reflective of the public opinion when adequately analysed. Various studies were presented that confirmed the validity of using Twitter sentiment to construct public opinion (O'Connor et al., 2010; Bollen, Pepe, & Mao, 2009; Cody et al., 2016). This conclusion is crucial in understanding the value and relevance of the insights that will be provided by this research, because Twitter is a primary data source.

The research gap that this thesis will fill lies within the combination of section 2.1 and 2.2: Twitter has not yet been used as a way to observe the public opinion on the housing market. Therefore, this study is considered the first of its kind.

Furthermore, this research will be the first to relate drivers of housing sentiment to Twitter. For each housing variable that was identified in the literature review, chapter 3 will develop a new theory by hypothesising how each of the housing-related variables may affect the sentiment and number of housing tweets sent from London. Therefore, this thesis will expand existing literature on housing sentiment.

Moreover, this thesis will specifically use the spatial aspects of tweets to determine from which geographical area the tweet is sent. Therefore, this research is able to make inferences about the differences in sentiment and number of housing tweets sent from various geographical areas. None of the reviewed studies that use Twitter to measure public opinion has taken the geographical location of the tweets into account to measure sentiment.

Finally, through the use of Twitter data, this thesis is expected to overcome various shortcomings of traditional methods to measure housing sentiment as described in section 2.1.1. Firstly, information asymmetry is expected to be insignificant on Twitter, because of the wide variety of users that tweet about the housing market. Because Twitter is a public platform and has many different users, the combined sentiment on the housing market is assumed not to be affected by information asymmetry. Secondly, Twitter is expected to overcome the shortcoming of reflective survey questions, because Twitter does not require users to think about how they feel about the housing market compared to some point in time: tweets capture the raw and current feelings about a situation. Thirdly, in contrast to traditional surveys, Twitter is a cost and time-efficient way to extract the public opinion. Fourthly, using Twitter data overcomes various response biases of traditional surveys. When Twitter users tweet, they simply do not consider the fact that their messages may be used to measure. Fifthly, Twitter allows users to respond at their own convenience (Chisholm & O'Sullivan, 2017), and the users are publicly providing information on a plethora of topics. Therefore, Twitter overcomes the survey limitation of requiring a population of willing participants.

3 Theory Development and Hypotheses

The studies reviewed in section 2.1 of the literate review showed that house prices negatively drive housing sentiment. According to the literature, the following housing-related variables are positive drivers of housing sentiment: household income, housing starts, housing demand, and housing sales. These variables will hereafter be referred to as *housing variables*. This chapter elaborates on these housing variables, by developing a theory on how each variable is expected to affect the sentiment and proportion of housing tweets sent from London. Developing this theory is an essential part of this research, because the theory will be translated into hypotheses which will be tested by statistical models in chapter 5. The confirmation or rejection of the hypotheses developed in this chapter will answer research question two.

Table 2 summarises this chapter by showing how each housing variable is assumed to affect the sentiment of housing tweets (*sent. housing tweets)* and the proportion of housing tweets *(no. housing tweets)*. The theorised relationships are formulated as hypotheses (denoted by the letter *H*). The hypothesised effects of housing variables on the level of *sentiment* of housing tweets are denoted by the letter *a*. The hypothesised effects of housing variables on the *number* of housing tweets are denoted by the letter *b*. Finally, the units of analysis that these hypotheses apply to, are the 33 boroughs of London, which are the local authority districts of London. The following sections relate each housing variable to the London housing crisis, and discusses how it affects the sentiment and proportion of housing tweets.

Table 2: *Summary of hypotheses*

		Hypothesised effect on			
Section	**Variable**	**Sent. housing tweets**		**No. housing tweets**	
3.1	House prices	Negative	H1a	Positive	H1b
3.2	Income	Positive	H2a	Negative	H2b
3.3	Mortgage	Negative	H3a	Positive	H3b
3.4	Housing supply	Positive	H4a	Negative	H4b
3.5	Housing demand	Negative	H5a	Positive	H5b
3.6	Market size	Positive	H6a	Positive	H6b
3.7	Homelessness	Negative	H7a	Positive	H7b

3.1 Rising house prices as the main driver

The rising house prices are a fundamental indicator of the London housing crisis (Marsden, 2015). The cost of housing is strongly correlated with the price of housing: as house prices rise, cost of housing also rises (Gallin, 2008).

Since 1995, the house prices in London have increased by 600% (Marsden, 2015). London has always been an expensive city to live compared to the rest of the UK, but in 2016 the difference in average house prices in London and the rest of England had never been higher (GLA, 2019). According to Marsden (2015), from 2009 to 2014, London house prices have risen every year by 7.8% on average, which is three times the rate of the rest of the UK (2.6%).

As a consequence of the rapidly growing cost of housing, more than 25% of Londoners live in poverty when housing costs are taken into account (Department for Work and Pensions, 2018), and over half of London households have little to no savings (GLA, 2020b). Another consequence of the soaring house prices is delayed home-owning, or even being locked out of owning a home completely, which is a growing concern for more Londoners (Page, 2018).

For these reasons, the increasing house prices are expected to cause a more negative tone in housing tweets, as formulated in hypothesis H1a. The second hypothesis H1b states that the proportion of housing tweets will increase when house prices increase. The underlying assumption for hypothesis H1b is that Londoners tweet more about the London housing market when the crisis does not seem to improve. The logic behind this reasoning is as follows, and will be repeated for all the hypotheses to come that relate to the proportion of housing tweets (which are all denoted by the letter *b*): as the housing crisis grows as a problem, users will increasingly share thoughts, opinions, dissatisfaction and possible solutions on the housing crisis.

H 1a *Rising house prices in a borough are negatively correlated with sentiment of housing tweets sent from that borough.*

H 1b *Rising house prices in a borough are positively correlated with the proportion of housing tweets sent from that borough.*

3.2 Increased income as a reason for financial relief

London households spend an increasingly disproportionate amount of their disposable income on housing (Trust for London, 2018). In not a single London borough, the average monthly rent of a one-bedroom house costs *less* than 30% of the median pre-tax income in 2018 (GLA, 2019). On average, a Londoner would need to spend almost half of their pre-tax income on renting a one-bedroom house in 2019 (Trust for London, 2020). If the average increase in income outpaced the increase in cost of housing, the proportion of income spent on housing would be reduced. Various researchers have proven that an increase in disposable income is positively correlated with happiness (Diener, 2009).

For these reasons, housing tweets sent from boroughs with higher income are expected to contain a more positive sentiment, as formulated in hypothesis H2a. Additionally, the cost of housing is assumed to be less of a financial burden in boroughs where income is relatively high. For these affluent boroughs, the housing crisis is assumed to be less of an immediate threat and therefore less discussed on Twitter, resulting in fewer housing tweets. The aforementioned reasoning is formulated in hypothesis H2b.

H 2a *The average annual income in a borough is positively correlated with sentiment found in housing tweets sent from that borough.*

H 2b *The average annual income in a borough is negatively correlated with the proportion of housing tweets sent from that borough.*

3.3 The burden of high mortgage debts

The level of mortgage payments is expected to negatively drive housing sentiment because recurring mortgage payments are a burden and cause financial strain. The deposit for first-time home-buyers in London has risen to an average of £90 000, and first-time home-buyers are borrowing at an all time high of four times their annual salary (Council of Mortgage Lenders, 2017). For Londoners that already have a mortgage, one could expect that they would not wish the financial burden of disproportionately high mortgages onto others.

The immense financial strain of such high mortgage debts is reason to assume that an increase in mortgage debt is negatively correlated with the sentiment of housing tweets. Additionally, high mortgages are assumed to drive the discussion on the housing crisis. Both assumptions are captured in the following hypotheses:

H 3a *The average outstanding mortgage debt in a borough is negatively correlated with sentiment found in housing tweets sent from that borough.*

H 3b *The average outstanding mortgage debt in a borough is positively correlated with the proportion of housing tweets sent from that borough.*

3.4 Stale housing supply causing a shortage

London is on the rise: from 1997 to 2017, the population of London grew by 1.8 million (27% increase), and the number of jobs grew by 1.6 million (40% increase) (GLA, 2019). However, during these 20 years, only 470 000 additional homes were built (18% increase), indicating that not nearly enough houses were built to accommodate all Londoners properly. Calculations show that if the rate of housing growth would had kept pace with the population growth since 1997, approximately 700 000 homes would have been added to the 3.5 million homes in London today (GLA, 2017).

According to Marsden (2015), increasing housing supply reduces housing shortage. For this reason, housing tweets sent from London boroughs where many new houses are constructed are expected to contain a more positive tone, as formulated in hypothesis H4a. Though the stale housing supply is considered a primary cause of the housing crisis, it also plays a fundamental role in solving the housing crisis (Petkar, Macwan, & Takkekar, 2012). Therefore, boroughs constructing many new houses are actively working on solving the housing crisis by reducing housing shortage. Under the assumption that Londoners tweet less when the crisis situation does not worsen, hypothesis H4b is formulated.

H 4a *The level of housing supply in a borough is positively correlated with sentiment found in housing tweets sent from that borough.*

H 4b *The level of housing supply in a borough is negatively correlated with the proportion of housing tweets sent from that borough.*

3.5 Housing demand causing overcrowding

Housing demand is the opposite of housing supply and both variables are therefore expected to drive housing sentiment inversely. Housing demand has steadily increased over the years because of the growing population due to increased longevity and net additional migration (Department for Work and Pensions, 2018). The second driver of housing demand is increasing investment demand. London real estate is considered a safe haven for both domestic and foreign investors because of the long history of increasing house prices (Rossall, 2015).

As demand for housing is outstripping housing supply, competition for property, and overcrowding increase (Arestis & González, 2014; Quiggly, 1999). Figures show that the average household size has increased from 2.3 in 1991 to 2.79 in 2019, indicating more overcrowding (Office of National Statistics, 2019b). Additionally, according to a London-wide housing survey, around 250 000 homes were considered overcrowded in 2016, a number which is only expected to have climbed since (Office of National Statistics, 2017).

Because of the aforementioned reasons, an increase in housing demand is assumed to cause a more negative tone in housing tweets, as formulated in hypothesis H5a. Additionally, increasing housing demand is considered a fundamental issue of the housing crisis. Therefore housing demand is assumed to be positively correlated with the size of the discussion, as captured in hypothesis H5b.

H 5a *Housing demand in a borough is negatively correlated with sentiment found in housing tweets sent from that borough.*

H 5b *Housing demand in a borough is positively correlated with the proportion of housing tweets sent from that borough.*

3.6 Housing market size indicating movement

The size of the housing market is assumed to be a positive driver of tweet sentiment because it indicates movement in the market. In some cities dealing with a housing crisis, the number of housing sales decreases because the market is locked (Nijskens et al., 2019). The proportion of housing sales has not seen large deviations in London, indicating that the overall market has not slowed down yet (GLA, 2019). Under the assumption that a growing market volume indicates movement in the housing market and shows the housing market in London is not locked, the following hypotheses are formulated:

H 6a *The volume of the housing market in a borough is positively correlated with sentiment found in housing tweets sent from that borough.*

H 6b *The volume of the housing market in a borough is negatively correlated with the proportion of housing tweets sent from that borough.*

3.7 Homelessness means inequality

The rate of homelessness is expected to be a negative driver of housing sentiment through the idea of inequality and unfairness. Homelessness was not described as a driver of housing sentiment by the articles highlighted in section 2.1 of the literature review. However, the relationship between homelessness and housing sentiment is analysed in this study to further expand the literature on drivers of housing sentiment.

Though London is one of the richest cities in the world, homelessness is a prevalent problem. In 2017, 0.42% of London households were considered statutory homeless (Ministry of CLG, 2020). Statutory homelessness is defined as "*households that are owed a main homelessness duty to secure accommodation as a result of being unintentionally homeless and in priority need*" (GLA, 2020a). Besides the actual figures showing that homelessness is a present problem in London, a survey among 1000 Londoners in 2018 showed that 57% of Londoners see homelessness as a growing problem in the borough they live in (Page, 2018).

The feeling of inequality and unfairness as a consequence of homelessness is assumed to negatively affect the level of housing sentiment and positively drive the proportion of housing tweets, as formulated in the following hypotheses:

H 7a *The rate of homelessness in a borough is negatively correlated with sentiment found in housing tweets sent from that borough.*

H 7b *The rate of homelessness in a borough is positively correlated with the proportion of housing tweets sent from that borough.*

3.8 Conclusion theory development

This section has expanded literature on drivers of housing sentiment by developing a new theory on how various housing variables affect the sentiment and proportion of housing tweets from London. Additionally, this study adds a new assumed driver of housing sentiment, which is the rate of homelessness. The theorised relationships are captured in 14 hypotheses which will be researched throughout this research, and tested in chapter 5 to answer the research questions. The data and methods used in this thesis to test the hypotheses are described in the following chapter 4: *data and methods*.

4 Data and Methods

This chapter elaborates on the data selection, collection and preparation processes, after which the methods will be discussed that are used to analyse this data. The data collection process, the data sources and data preparation processes are explained in section 4.2 and 4.3 for housing data and Twitter data, respectively. Section 4.4 describes how missing values will be resolved to establish a balanced dataset. Section 4.5 explains the selection process of the statistical models. In this research, two statistical models will be used. The first model estimates the effect of housing variables on the *sentiment* of housing tweets. The second model will estimate the effect of housing variables on the *number* of housing tweets. The results of these models will answer the research question Section 4.6 evaluates the input variables for these models using multicollinearity as the evaluation criterion. The output and findings of the models will be discussed in chapter 5. Figure 1 illustrates the structure of this research.

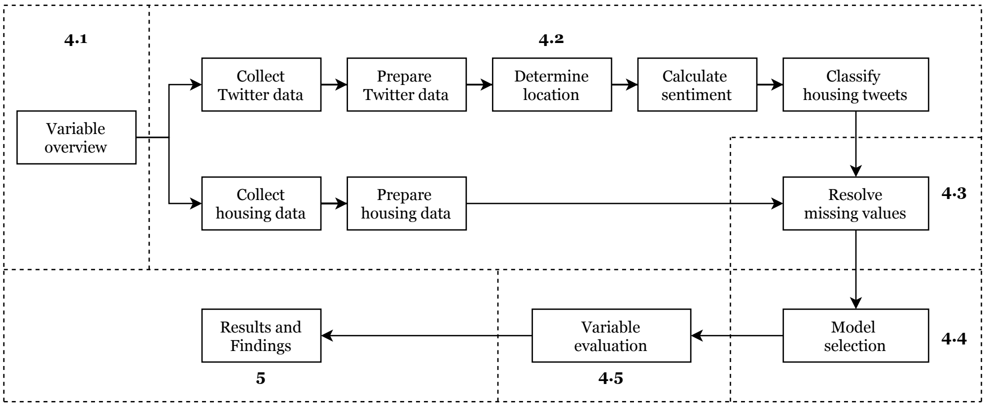


Figure 1: *Research design*

4.1 Variable overview

Table 3 summarises the variables that will be analysed in this research. These variables will serve as input for the models which will estimate the effect of housing variables on the sentiment and proportion of housing tweets. These variables are selected based on the literature review in chapter 2 and the theory developed in chapter 3. The variables can be split up into two categories, which are Twitter variables and housing variables. For each variable listed in table 3, the sections below explain how the variable is exactly measured, how the data is collected, and how the data is prepared for analysis. An in-depth table of all variables, measurement units, and how the variables should be interpreted can be found in appendix B.

Table 3: *Variables and data sources*

	Information	Source	Year
4.2	Tweet sentiment	Erasmus University Rotterdam	(2019)
4.2	Tweet count	Erasmus University Rotterdam	(2019)
4.3.1	House prices	HM Land Registry	(2019)
4.3.2	Housing supply	Ministry of CLG	(2019)
4.3.3	Housing demand	Office of National Statistics	(2019c)
4.3.4	Market size	HM Land Registry	(2019)
4.3.5	Mortgages	Council of Mortgage Lenders	(2019)
4.3.6	Income	Office of National Statistics	(2019a)
4.3.7	Homelessness	Ministry of CLG	(2020)

4.2 Twitter data collection and preparation

Twitter data can be obtained in various ways. Firstly, the official public Twitter API can be used to scan and store tweets that were sent in the last seven days to build up a large dataset if ran for a long time (McCormick et al., 2017). Secondly, some Twitter datasets are assembled by third parties and uploaded for public use. Thirdly, data can be directly obtained from Twitter as a company. The Twitter data used in this study was provided by dr. Tobias Brandt, assistant professor at the Rotterdam School of Management, Erasmus University Rotterdam. The dataset was obtained directly from Twitter and contains approximately 96 million tweets. All tweets in the dataset are sent from London with a time-stamp between 01/09/2011 to 31/04/2019. Every tweet in this dataset has a location tag attached to it, which will be explained in more detail below. The basic structure of the provided Twitter data set is described in appendix D. The dataset does contain more attributes, but those will be disregarded as they are irrelevant to this research.

After collection, the Twitter data is prepared in the following steps to extract the relevant information needed for the research. First, the location from which the tweet was sent is determined for every tweet based on the spatial data attached to tweets. Adding the location to the tweets is required to compare results between boroughs. The second step is to obtain the sentiment of every tweet. This step is required to observe trends in sentiment on Twitter over the years. Then, housing tweets need to be distinguished from other tweets, because this research is primarily interested in opinions on the housing market. Subsequently, the sentiment of housing tweets is compared to the average sentiment of all tweets, in the form of *abnormal sentiment*, to control for trends in overall sentiment over time. Finally, the number of housing tweets is compared to the number of overall tweets, to observe how many tweets are sent relatively to control for changes in overall tweet volume over time. The importance and added value of each preparatory step will be discussed in the sections below.

1. Assign location to all tweets (4.2.1)
2. Obtaining sentiment for all tweets (4.2.2)
3. Classifying housing tweets (4.2.3)
4. Calculating abnormal sentiment for housing tweets (4.2.4)
5. Obtaining proportion of housing tweets (4.2.5)

4.2.1 Assigning location to all tweets

For all 96 million tweets in the dataset, the location is assigned in the form of one of London's 33 boroughs. Assigning the borough to a tweet is of great value for two reasons. Firstly, it allows the observer to compare the results between geographical areas. Secondly, it allows the model to compare the tweets by borough, which improves model performance (Hsiao, 2007). The reason why model performance increases when the location is added to the data will be discussed in section 4.5.1.

Spatial data can be attached to a tweet in two ways. Firstly, users are given the option to manually add a location tag to their tweets. When a user wants to add the location to a tweet, Twitter provides the feature for the user to pick a location from the Google Places database, which stores all Google Maps locations. The locations stored in the Google Places database cover almost every location in the world, including businesses, landmarks, parks, and intersections (Google, 2020). As long as the location is available on Google Maps, it can be attached to a tweet. All locations in the Google Places database are uniquely identified by a variable called 'place_id'. All 96 million tweets in the dataset used for this research contain a 'place_id' variable, and were obtained in that format from Twitter by dr. Tobias Brandt.

The second way spatial data can be attached to a tweet is in the form of GPS coordinates. Users can choose to automatically add the coordinates of their position to tweets when they send a tweet (Twitter, 2019). Up until Twitter changed its privacy policy in April 2015 from an opt-out to an opt-in system, GPS coordinates were stored in the Twitter database by default. After April 2015, users had to give their permission in order for Twitter to store the location data of Twitter users in GPS form. After this policy change, the number of tweets containing GPS coordinates decreased to only about 1% of all tweets (Graham, Hale, & Gaffney, 2014). Therefore, not all tweets used in this research contain GPS coordinates, but every tweet does contain the ′place_id′ variable.

This study uses an algorithm that uses two methods to leverage both types of spatial data to determine from which borough the tweet was sent. The two methods used by the algorithm will be discussed in the sections below.

The first method uses the ′place_id′ data, and the second method uses GPS coordinates to determine from which borough a tweet was sent.

Method (1) Using ′place_id′ to determine borough

The ′place_id′ variable is a unique identifier in an incomprehensible format (e.g. `ChIJgUbEo8cfqokR5lP9_Wh_DaM`). To turn the incomprehensible ′place_id′ format into a readable location, the algorithm searches the Google Places database for the ′place_id′ and transforms it to the full name of the location as stored in the database (e.g. "Harrods Warehouse, Westminster"). Next, this readable full name string of the location is scanned for the name of any of the 33 London boroughs. If the name of a London borough is found in the full name string, that borough is assigned as a location to that tweet. If the name of a London borough is not found, method (2) will be used to estimate the location.

Method (2) Using GPS coordinates to determine borough

The second method utilises the longitude and latitude coordinates of a tweet to estimate from which borough the tweet was sent. The algorithm takes the coordinates attached to a tweet and checks within what borough the coordinates fall to determine the location of the tweet. The algorithm uses the GPS boundaries of all London boroughs to indicate which coordinates fall within which borough. The GPS boundaries of the boroughs are stored in the so-called shapefile format (*.shp*). A shapefile spatially describes geometries as polygons, which represent a geographical area when printed, as depicted in figure 2. The shapefiles of London boroughs are obtained from the Greater London Authority 2020c. For every tweet that contains GPS coordinates, the algorithm checks in which shapefile the coordinates of the tweet fall, and assigns the corresponding borough to the tweet. If the tweet contains GPS coordinates, this method never fails to assign a London borough to a tweet, because all tweets in the dataset are sent from London thus (if the tweet contains contain coordinates), the coordinate will always fall in one of the 33 London boroughs.

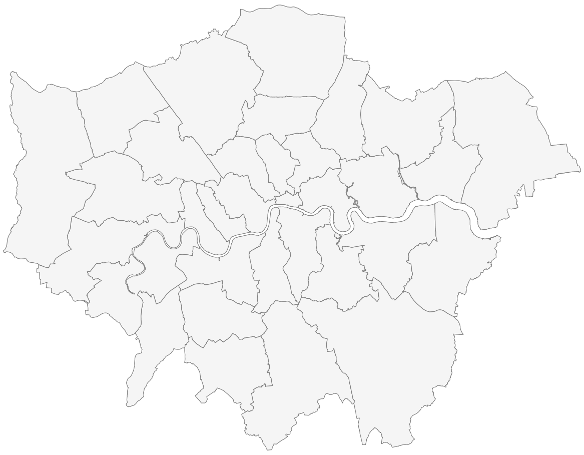


Figure 2: *Print of London borough shapefiles (GLA, 2020c)*

As mentioned before, all tweets in this dataset contain a ′place_id′ variable, but not all tweets contain GPS coordinates. Therefore, every tweet in the dataset can be placed into one of two categories. The first category consist of tweets that contain both ′place_id′ and GPS coordinates, which make up 45% of all tweets. The second category consist of tweets that contain only ′place_id′ and no GPS coordinates, which make up the remaining 55% of all tweets.

The algorithm handles both categories differently by using either a combination of method 1 and 2 or only method 2. Figure 3 shows a flowchart of how the algorithm handles both categories of tweets. The blocks in the bottom row of this diagram present the five possible outcomes of the algorithm. Both categories and the five possible outcomes are discussed below. An in-depth table with the distribution across all five possible outcomes per year can be found in appendix C.

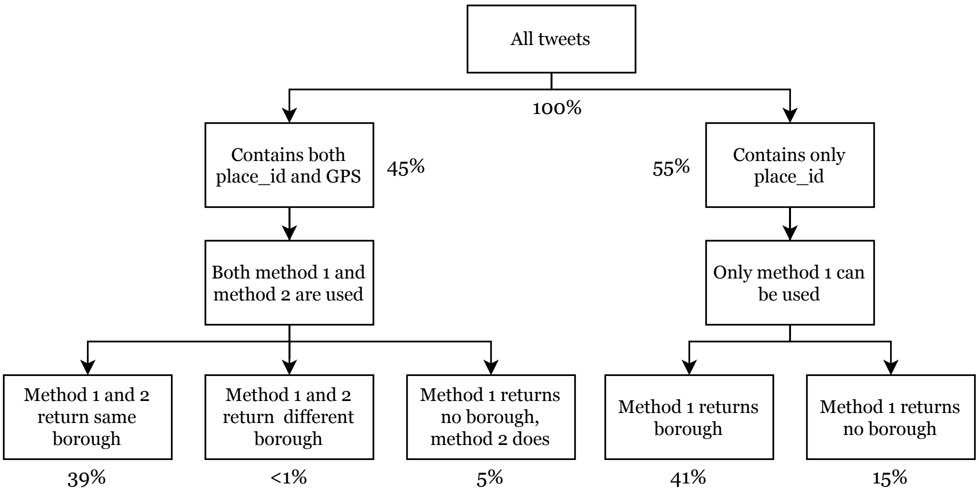

All tweets
100%
Contains both place_id and GPS
45%
55%
Contains only place_id
Both method 1 and method 2 are used
Only method 1 can be used
Method 1 and 2 return same borough
39%
Method 1 and 2 return different borough
<1%
Method 1 returns no borough, method 2 does
5%
Method 1 returns borough
41%
Method 1 returns no borough
15%

Figure 3: *How algorithm 1 determines location of tweets*

Category 1: tweets that contain both ’place_id’ and GPS coordinates

For the tweets that contain both ’place_id’ and coordinates, method (1) and method (2) are used in conjunction to estimate which borough the tweet was sent from. If both methods return the same borough, that borough is assigned to the tweet. The location of 39% of all tweets is determined this way.

However, for some tweets in this category, method (1) and method (2) return a different borough for the same tweet. This means that a user attached a different borough as ’place_id’ to the tweet, than the borough the user was located at when tweet was sent according to the GPS coordinates. This indicates that sometimes Twitter users do not tweet about activities as they are undertaking them, but rather at a different moment when they have moved to a different place, which is a shortcoming of Twitter data when used for spatial analysis (Abbasi et al., 2015). When the two methods return different boroughs for the same tweet, the borough that was returned by method (1) using ’place_id’ is considered to be the actual borough, because the user put in extra effort to attach the location to the tweet and actively wanted other users to see this location when reading the tweet. Fortunately, in only 0.6% of all tweets method (1) and (2) return a different borough for the same tweet.

For 15% of all tweets, method (1) fails to return a London borough because the full name string matched to the ’place_id’ did not contain the name of a London borough (e.g. "London King’s Cross railway station"). In this case, the algorithm resorts to method (2) to determine tweet location.

Category 2: tweets that contain only 'place_id' and no coordinates

When a tweet only contains 'place_id', and does not contain GPS coordinates, only method (1) can be used to determine the borough from which the tweet was sent. The location of 41% of all tweets was determined this way. Again, sometimes method (1) fails to assign a location to a tweet because the full name string matched to the 'place_id' did not contain the name of a London borough. Because the tweets in this category do not contain GPS coordinates, the algorithm is unable to resort to method (2) to assign a location to the tweet, and the algorithm fails is unable to assign a location to the tweet. This is the case for approximately 15% of all tweets. These tweets without a location will not be used as input for the statistical models but will be used in the descriptive analysis. These tweets without a location are further discussed in the section *resolving missing values* (4.4).

Figure 4 shows the pseudocode that represents how algorithm 1 handles the tweet categories and combines both methods to determine the location of tweets.

Algorithm 1: Pseudocode for location estimation algorithm

```
for tweet do
    check if contains 'place_id;
    if contains 'place_id then
        match 'place_id' with full place string;
        check full name string for any borough;
        if full name string contains any borough name then
            assign borough
        else
            "use coordinates to determine borough"
    else
        check if contains coordinates;
        if contains coordinates then
            for all borough shapefiles do
                check for coordinate;
                if coordinate is within shapefile then
                    assign borough based on coordinate;
                else
                    assign "no location"
        else
            assign "no location"
```

Figure 4: *Pseudocode for location estimation algorithm*

4.2.2 Obtaining sentiment of all tweets

The sentiment of a tweet is used to quantify the tone of the discussion about the housing market. The sentiment score of a tweet indicates how positive or negative the tweet is, ranging from -1 (most negative) to +1 (most positive). The sentiment is calculated for all 96 million tweets in the dataset. This is required in order to compare the relative (*abnormal*) sentiment of housing tweets to the average sentiment of all tweets, which will be described in section 4.2.4.

This study uses sentiment analysis to calculate the sentiment of a tweet. Sentiment analysis is defined as "*the computational study of people's opinions, sentiments, attitudes, and emotions toward entities and their aspects expressed in text*" (Liu, 2012). Various sentiment analysis techniques exist. The method used in this study to calculate sentiment is the **V**alence **A**wareness **D**ictionary and s**E**ntiment **R**easoner (VADER). VADER is a sentiment analysis method specifically developed to calculate the sentiment of short, informal microblog-like texts, such as tweets (Hutto & Gilbert, 2015). VADER uses a lexicon containing over 9000 words that were evaluated by independent human raters who scored all 9000 words on their level of valence, ranging from -4 (most negative) to +4 (most positive). VADER uses this rated lexicon to calculate a sentiment score for newly presented texts, like the tweets in this research. VADER calculates the sentiment score for a tweet by summing the valence scores of every word in a tweet and normalising it on a scale from -1 and +1 (Hutto & Gilbert, 2015). Additionally, VADER uses punctuation and emoticons to identify increased sentiment intensity. Examples of how VADER handles tweets are illustrated in table 4. However, VADER is only able to process English tweets, thus non-English tweets are excluded from this analysis.

Table 4: *Examples of VADER sentiment scores*

Example tweet	**Sentiment**	**Category**
`I’ll never be able to afford a house!! :(`	-0.5399	Negative
`UK house prices surpass 2007 peak, says Nationwide via @BBCNews #ukhousing #property`	0.0000	Neutral
`Pleased to see that affordable housing is on the up, with Westworth having built over 1,000 new units since 2010.`	0.4400	Positive

4.2.3 Classifying housing tweets

To research the public opinion on the housing market using tweets, a distinction needs to be made between tweets that are related to the housing market, and tweets that are not related to the housing market. A rule-based classifier classifies all 96 million tweets as "*housing tweet*" or "*not a housing tweet*". The input for this classifier is a set of *stemmed* housing-related queries. Stemming is a text normalisation technique in which words are reduced to its root form (Lovins, 1968). For example, the stem of the word "housing" and "house" is "hous". This technique is valuable as it expands the search query to match additional relevant tweets.

The search queries are formulated in the Structured Query Language (SQL). SQL is a programming syntax designed for managing and extracting data stored in databases (Groff & Weinberg, 1999). An algorithm scans the text field of every tweet and returns the tweets that match the conditions specified in the inclusion queries, which can be found in appendix E.

Although a sophisticated set of queries is used, the classifier can still make mistakes. The four different possible categorisation outcomes of the rule-based classifier are as follows. The classification can be either correct or incorrect. When correctly classified, the tweet is either correctly classified as 'housing tweet' (*true positive*), or correctly classified as 'no housing tweet' (*true negative*). When the classifier makes a mistake, the tweet is either incorrectly classified as 'housing tweet' (*false positive*), or incorrectly classified as 'no housing tweet' (*false negative*). Table 5 shows examples of tweets that fall in different four categories. The (stemmed parts of the) words that match the inclusion queries are underlined.

Table 5: *Types of classification*

Type	**Example Tweet**
True positive	`The rising prices of London property make me bawl my eyes out..`
False positive	`Just had a great house wine for a great price at @SJRestaurant`
False negative	`This tiny building next door was just sold for some stupid amount of money. so typical for London.`
True negative	`lol just bought this costume for a house party`

The first tweet in table 5 is a true positive, because it was correctly classified as "housing tweet" because it contains both the words "prices" and "property", which matched the inclusion queries. To improve the accuracy of the model, all of the inclusion queries consist of at least two housing-related words that both need to be present in the tweet before it is classified as 'housing tweet'. If only the word "house" is present in a tweet, there is not enough evidence to assume the tweet is related to the housing market and the tweet would therefore not be classified as "housing tweet".

The second tweet is incorrectly classified as "housing tweet" (false positive), because it contains the words "house" and "price", even though the tweet is not related to the London housing market. The primary reason for false positive classifications is that some inclusion queries are part of other words that are unrelated to housing (e.g. house music, house of parliament). To reduce the number of false positives to a minimum, a set of exclusion queries is established. The exclusion queries were constructed by manually reading 25 random samples of 100 housing tweets to observe what (combination of) words cause false positives. Tweets matching exclusion queries will be classified as "no housing tweet". A full list of exclusion queries and the explanation for each exclusion query can be found in appendix E.

The third tweet is a false negative because it is incorrectly classified as "no housing tweet". The text does not match any of the specified inclusion queries, even though the tweet is related to the housing market (i.e. "building" is not part of the inclusion queries, as it would return too many false positives).

The final category consists of true negatives, which are tweets that are not about the housing market and are marked as such because they do not match an inclusion query or they match an exclusion query. In this example, the tweet is considered a true negative because it matches the exclusion query "costume", which is considered an exclusion query because it contains the word "cost".

To assess the performance of the rule-based classifier, the distribution of all tweets across the four different classification outcomes is displayed in a *confusion matrix* shown in figure 5.

		Predicted outcome	
		Housing tweet	No housing tweet
Actual value	Housing tweet	True positive ≈ 93%	False negative
	No housing tweet	False positive ≈ 7%	True negative
		22 433	96 027 344

Figure 5: *Confusion matrix*

The matrix shows that out of a total of 96 049 777 tweets in the dataset, 22 433 tweets were classified as *housing tweet* and 96 027 344 tweets were classified as *no housing tweet*. The results classifier indicates that 1 in 4281 tweets sent from London matched an inclusion query and is considered a housing tweet. Considering the wide variety of discussed topics on Twitter (Zimmer & Proferes, 2014), this number seems appropriate.

To estimate the performance of the classifier, 25 random samples of 100 'housing tweets' were manually cross-checked. Out of the 2500 sampled 'housing tweets', 2313 (≈ 93%) tweets were actually related to the housing market (thus correctly classified), and 187 (≈ 7%) were not (thus incorrectly classified).

In contrast to the number of true and false *positives*, the number of true and false *negatives* is difficult to obtain, because the actual classification categories of all tweets are unknown. Of all tweet, 99.97% of the tweets were classified as *no housing tweet*. It is simply too much work to observe how many of the 96 027 344 'no housing tweets' were incorrectly classified as such.

4.2.4 Calculating abnormal sentiment for housing tweets

A reason for changes in sentiment of housing tweets may be that the overall sentiment on Twitter changes. In some months, Twitter users could simply be more positive or negative for various reasons. To control for these trends in overall sentiment, the sentiment score of housing tweets is compared to the average sentiment score of all tweets. This metric of relative sentiment is defined as *abnormal sentiment*. The abnormal sentiment of a housing tweet shows how far the sentiment deviates from the average sentiment of all tweets. The abnormal sentiment score is calculated for every tweet by subtracting the average of all tweets (sent from borough *i* in year *t*), from the average of the housing tweets (sent from borough *i* in year *t*). A housing tweet (sent borough *i* in year *t*) with an abnormal sentiment score of -0.05, means that the sentiment of that tweet is -0.05 below the average sentiment of all tweets sent from borough *i* in year *t*, on a scale from -1 to +1. The abnormal sentiment score of housing tweets in borough *i* in year *t* is calculated as expressed in equation (1):

$$AS_{it} = \bar{S}(h)_{it} - \overline{S}_{it} \tag{1}$$

where:

AS_{it} = abnormal sentiment of housing tweets in borough *i* in year *t*

$\bar{S}(h)_{it}$ = average sentiment of housing tweets in borough *i* in year *t*

$\bar{S}_{it}$ = average sentiment of all tweets in borough *i* in year *t* as defined in equation (2):

$$\overline{S}_{it} = \frac{\sum_{i=1}^{n} s_{it}}{\sum_{i=1}^{n} x_{it}} \tag{2}$$

where:

s_{it} = sentiment of tweets sent in borough *i* in year *t*

x_{it} = number of tweets in borough *i* in year *t*

4.2.5 Obtaining proportion of housing tweets

The number of housing tweets needs to be obtained to make any conclusions about how the proportion of housing tweets has changed over the years. An important factor that impacts the number of housing tweets, is the change in overall usage of Twitter. When Twitter becomes more popular, more tweets are sent in total, which also increases the total number of housing tweets. To control for changes in overall tweet volume, the number of housing tweets will be compared to the total number of tweets. The metric of relative tweet count is defined as the proportion of housing tweets. The proportion of housing tweets ($P(h)$) (in borough i in year t) is calculated by dividing the total number of housing tweets (in borough i in year t) by the total number of tweets (sent from borough i in year t). The resulting number is multiplied by 100 000 for readability and interpretation purposes. Therefore the proportion of housing tweets is defined as "*the number of housing tweets per* 100 000 *tweets*". Hereafter, the term *proportion of housing tweets* will be used to indicate the size of the discussion about the London housing market. The calculation for the proportion of housing tweets is expressed in equation (3).

$$P(h)_{it} = (\frac{\sum_{i=1}^{n} x(h)_{it}}{\sum_{i=1}^{n} x_{it}}) \times 100\,000 \tag{3}$$

where:

$P(h)_{it}$ = the proportion of housing tweets, defined as the number of housing tweets per 100 000 tweets

$x(h)_{it}$ = number of housing tweets in borough i in year t

x_{it} = number of tweets in borough i in year t

4.3 Housing data collection and preparation

This section describes the data source for each housing variable used in this research and explains how each housing variables is measured and prepared for analysis.

4.3.1 House prices

The rising house prices are measured through the house price index (HPI). The house price index tracks the changes in house prices relative to the house prices in January 2015 (which has an HPI of 100). Changes in the series represent increases and decreases in house prices. A house price index of 88.3 in borough i in year t should be interpreted as follows: *the average house prices in borough i in year t was 88.3% of the average house prices in borough i in January 2015.* The house price index is collected from the HM Land Registry (2019). The organisation measured the house prices and computed the HPI for every London borough since 1995. For some London boroughs, the Land Registry applied a three-month moving average to reduce volatility caused by a low number of housing sales transactions.

4.3.2 Housing supply

The level of housing supply is measured through the number of *net additional dwellings*. The number of net additional dwellings is measured by tracking newly built houses plus any gains or losses through housing conversions, change of use and demolitions. The level of housing supply is measured throughout the year for every London borough by the Ministry of Housing, Communities & Local Government (2019). The dataset contains annual numbers on housing supply for every borough.

4.3.3 Housing demand

Increasing longevity and immigration due to London being an attractive city to live in, have caused a rise in population and therefore a rise in people looking for a home to live in. Therefore, borough population is used as a measure of housing demand. The use of population size as an indicator of housing demand has been widely researched (Thompson, 1937; Mason, 1996; Mulder, 2008; Wang, Wang, & Zhang, 2015). The corresponding results of these studies validate the use of the borough population as a measure of housing demand. The annual data on borough-specific population is obtained from the Office of National Statistics (2019c).

4.3.4 Housing market size

The housing market size of London is measured through the number of property sales in London. Every London property sale is registered by the (HM Land Registry, 2019). The organisation has aggregated the number of property sales per borough per year, since 1995.

4.3.5 Mortgages

The mortgage variable is measured through the average cumulative value of outstanding residential mortgage lending per postcode sector in a borough. The data on mortgages is obtained from the Council of Mortgage Lenders, or CML (2019). The publicly available dataset contains the value of outstanding residential mortgage lending by postcode sector for 9273 Great Britain postcode sectors. The CML gathered the statistics per postcode sector from seven UK banks, which together account for 73% of mortgage lending in Great Britain (Council of Mortgage Lenders, 2019).

The mortgage data was prepared in the following way. The mortgages dataset presents the value of outstanding mortgages per *postcode sector*. London consists of 22 *postcode districts*, which are further subdivided in 487 *postcode sectors*, which in turn consist of all 642 752 unique London postcodes (Office of National Statistics, 2019d). To obtain the average cumulative value of outstanding mortgages per postcode sector *in a borough*, the postcode sectors need to be matched to a borough. However, the borders of the postcode sectors in London do not align perfectly with the boundaries of London boroughs (UK Post Office, 2020). For this reason, small parts of a postcode sector can be in multiple boroughs. To match each postcode sector to a borough, a list of all 642 752 London postcodes and their corresponding borough was obtained from the UK postcode directory (2019d). For each postcode sector, the most frequently occurring borough was calculated, which was considered the borough for that postcode sector.

4.3.6 Income

The income per London borough is measured through the gross annual pay before deductions (i.e. tax payments and insurance). The data is collected by the Office of National Statistics (2019a), through the Annual Survey of Hours and Earnings (ASHE). The ASHE is an annual survey that asks approximately 300 000 London workers about their income. The income figures in this dataset are based on the median rather than the mean because the median better controls for outliers, as it is less affected by citizens with very high income that would skew the distribution of the data (Office of National Statistics, 2019a). Furthermore, the data per borough is residence based, and not workplace based. This means that the average annual income for Westminster is the average income for the citizens that *live* in Westminster, and not for the citizens that *work* in Westminster.

4.3.7 Homelessness

The homelessness rate is defined as as "*the number of households per 1000 households that are owed a main homelessness duty to secure accommodation as a result of being unintentionally homeless and in priority need*" (Ministry of CLG, 2020). Since 2005, the rate of homelessness is monitored by the local London authorities and reported to the Ministry of Housing, Communities and Local Government (2020), which present the statutory homelessness rates per borough, per year.

4.3.8 Dummy variables

The statistical models that will estimate the relationship between housing variables and sentiment and proportion of housing tweets are regression models. Regression models requires numerical independent variables as input, and can not handle qualitative variables (Dunning, 2008). Therefore, qualitative variables should first be transformed into quantitative variables before used as input for regression models, which is done through *dummy variables*. Dummy variables are numeric stand-ins for qualitative independent variables, taking on the value of either 1 or 0 (Garavaglia & Sharma, 1998).

The only non-numerical variable this research uses, is the variable of *location* in the form of London boroughs. For each borough, a dummy variable is created so they can be fed into the regression models. Table 6 illustrates how dummy variables work for all boroughs. All data observations for the borough of Barney will have a "1" in the column "d_barney", (which is the dummy variable created for the borough of Barney), and a "0" in the other columns of dummy variables for the other boroughs. Through the use of dummy variables, the regression model is able to control for the effect of location on the sentiment and proportion of housing tweets.

Table 6: *Example data to illustrate dummy variables*

month	**borough**	**sent**	**HPI**	..	d_barney	d_bexley	..	d_westmnr
2012-01	Barney	-0.09	98.1	..	1	0	..	0
2012-01	Bexley	-0.21	88.1	..	0	1	..	0
..	..	..	..	..	..	..	..	
2012-01	Westminster	-0.14	92.2	..	0	0	..	1

4.4 Resolving missing values and data irregularities

Not all data sources present the datasets in the same format. Some datasets provide information on every borough, other datasets on a different geographical level, like postcode sectors. Some datasets are collected monthly, other datasets on an annual basis. The availability of all data used for this research is summarised in table 7. When all data is combined into one large dataset, the discrepancy in data formats between individual datasets results in an unbalanced total dataset. An unbalanced dataset can have a negative effect on the validity of the statistical models (Kang, 2013). This section describes how a balanced dataset is created by resolving the missing values.

Table 7: *Data availability*

Variable	**Availability**	**Coverage**	**Periodicity**
Twitter data	01/09/2011 to 31/04/2019	Coordinate	Daily
House prices	01/01/1994 to 31/01/2020	Borough	Monthly
Housing market size	01/01/2004 to 31/12/2019	Borough	Annual
Housing supply	01/01/2004 to 31/12/2019	Borough	Annual
Mortgages	01/01/2012 to 31/06/2019	Zip-code	Quarterly
Income	01/01/2002 to 31/12/2019	Borough	Annual
Homelessness	01/01/2005 to 31/12/2018	Borough	Annual

First, the housing datasets are aggregated on a temporal and geographical level to transform the data into a balanced data format without missing values. The data is temporally aggregated by year, as this is the highest level of time granularity in any of datasets. Then, all housing datasets are aggregated per London borough, which is the highest level of geographical granularity in the datasets.

The Twitter data contained missing values on a temporal and geographical level, which were both resolved as follows. Firstly, the years 2011 and 2019 were dropped from the analysis, as the Twitter data from those years are incomplete: 2011 only contains the months from September and afterwards, and 2019 only contains the months up until April. If these month would have been included in the dataset, missing values would be created in other datasets that do not cover the years 2011 and 2019, and not all datasets are collected a monthly basis. This means that the datasets with a quarterly, biannual, or annual periodicity, would have to be interpolated to a monthly level, which would be too inaccurate for analysis.

The Twitter data was also missing values in the spatial aspect of the data. As described in section 4.2.1, 15% of all tweets were not assigned a location. These tweets will not be used as input for the models, but will be used in parts of the descriptive analysis, as they may enrich the descriptive results that will answer the research question one.

After aggregating all data on a temporal level and and geographical level, removing the years 2011 and 2019, and removing tweets without a location tag, a balanced dataset remained without any missing values.

Data irregularities

An exploratory data analysis was conducted to observe any curiosities in the datasets. The housing data did not contain any noteworthy deviations. However, the Twitter data did show some irregularities. During the exploratory data analysis, disproportional changes in tweet volume were observed in the borough Havering. The total number of tweets that were assigned the borough Havering dropped from 360 742 in 2014 to 9665 in 2019, which is a decline of 97%. This percentage is far above the average decline of 34% in the other boroughs over the same time period. After further data inspection it was found that since the month of May 2014, Havering was no longer listed as a ’place_id’ in the Twitter dataset, indicating an error in the dataset and explaining the sharp and sudden drop in tweet volume from that borough. Havering was excluded from the dataset because the sharp drop in number of tweets due to an error in the dataset a will skew and invalidate the model results.

4.5 Model selection strategy

This research uses statistical models to estimate the effect of housing variables on the sentiment and proportion of housing tweets. The selection of the correct model type is crucial to establish the most valid estimation. The type of model that should be selected depends, among other things, on the structure of the data and nature of the variables (Wooldridge, 2010). The theory that explains the reasoning behind selecting the right model is somewhat specific to the field of statistics and data analysis. However, this section attempts to explain the theory by directly applying it to the use case of estimating the relationships between the London housing market and Twitter.

4.5.1 Data format

After the data preparation process, the data covered 32 London boroughs over a period of seven years, resulting in 224 total observations (32 × 7). The 224 total observations can be considered as rows of a table. Each row, or observation, represents one borough, at one year. Each row of this table contains information about this borough in that year. To be more specific, each row contains information about the nine variables listed in 3. The information about these nine variables can be regarded as columns of a table, which has 224 rows. The total dataset therefore contains 2016 data points (224 × 9).

The data format of this final dataset is called *panel data*, which is a resulting data format when *time-series* and *cross-sectional* data are combined (Chamberlain, 1984).

The first data format that this dataset represents is called a *time-series*. Time-series observe how a single unit of analysis changes over various points in time (Hansen, 1995). The time points (years) are indicated by subscript t. To give an example, a time-series allows the observer to see how the abnormal sentiment of housing tweets and house price index in Westminster have changed over 2012 to 2018.

The second data format that this dataset represents is called *cross-sectional data*. Cross-sectional data shows how different *entities* (London boroughs in this case) at a single point in time (Levin, 2006). Hereafter, the term *entities* refers to the London boroughs, and vice versa. The boroughs are indicated by subscript i. To give an example, cross-sectional data would allow the observer to see how the abnormal sentiment in Westminster in 2012 is different from abnormal sentiment in Kensington in 2012.

When time-series and cross-sectional data are combined, the resulting data format is called *panel data*. Panel data observes the same entities over multiple time periods (Chamberlain, 1984). The main advantage of panel data is that it allows the observer to not only see processes of change over time, but simultaneously allows the observer to see differences between entities over time (Hsiao, 2007). As an example, this panel dataset allows the observer to see how abnormal sentiment in Westminster (i) in 2012 (t) is different from Kensington in 2017. The primary disadvantage of panel data models is that data collection is costly to collect, which is overcome through the use of Twitter data which is publicly available and easy to collect.

4.5.2 Panel data models

Associations and relationships in panel data can be estimated through the use of *panel data models*. Panel data models have various advantages over traditional models, two of which are key. The primary advantage of panel data models is that they are capable of making more accurate inferences than a single cross-section or time-series data model (Hsiao, 2007). The simplified reason is that panel data generally contains more data points, resulting in larger *sample variability* and *degrees of freedom*.

The second key advantage is that panel data allows to control for effects that are specific to entities in the form of a *fixed effects* or *random effects* panel data model (Wooldridge, 2010). This section describes the theory behind these models and provides arguments for which model type is the most appropriate. Two models will be used to answer research question two. The first model estimates the effect of housing variables on the sentiment of housing tweets. A second model will estimate the effect of housing variables on the proportion of housing tweets.

Homogeneous or heterogeneous?

Various types of panel data models exist, which can be split up into two main categories: homogeneous and heterogeneous models. This section describes both categories and provides arguments for which category will be selected. The foundation of both homogeneous and heterogeneous models is a *multiple linear regression model*, or *MLR* (Wooldridge, 2010). The formula for a MLR model is expressed in equation (4):

$$y = \alpha + \beta_1 x_1 + .. + \beta_k x_k + \epsilon \tag{4}$$

In this formula y is the dependent variable (i.e. sentiment or proportion of housing tweets). α represents the intercept of the equation. x represents the seven (k) independent housing variables (shown in table 3). β is the coefficient that represents the effect that housing variable (x) has on the dependent variable y. ϵ is the error term, which captures all other factors which influence the sentiment of housing tweets other than the independent (housing) variables that were included in the model (Wooldridge, 2010). To better understand the MLR formula, it displayed in a more descriptive manner in equation (5):

$$Sentiment_{housingtweet} = \alpha + \beta_{hpi} x_{hpi} + .. + \beta_{homeless} x_{homeless} + \epsilon \tag{5}$$

Two relationships will have to be estimated by two models to answer research question two. The first model estimates the effect (β) of housing variables (x_1, .., x_7) on sentiment of housing tweets (y_1). The second model will assess the effect of housing variables on the proportion (y_2) of housing tweets.

However, the MLR model is not the most appropriate model to estimate these relationships because it makes assumptions that do not hold for our data. The MLR model assumes that intercept α and the coefficients β do not vary across boroughs (i) over the years (t). In the MLR model, the differences across boroughs and time are only captured in the error term ϵ (Hurlin, 2018). Under this assumption that all parameters are common for all entities, we speak of a *homogeneous model*, which is the most restrictive type of panel data model (Katchova, 2013). The assumption that the coefficients are the same for all entities will likely not hold for our dataset: each housing-variable will probably have a different effect per borough.

The counterpart of the homogeneous model is the *heterogeneous model*, which assumes that the coefficients are not equal for all boroughs over the years (Hurlin, 2018). Heterogeneous models follow the general formula expressed in equation (6), in which α_i captures the effect of differences between boroughs.

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + .. + \beta_k x_{k,it} + \epsilon_{it} \tag{6}$$

As illustrated, the difference between homogeneous and heterogeneous panel data models is the presence of variation across entities, which is referred to as *heteroscedasticity*. A Breush-Pagan test is conducted to confirm the assumption that the housing variables have a different effects per borough. The Breush-Pagan test assesses the presence of heteroscedasticity by testing the following hypotheses (Breusch & Pagan, 1979). If the p-value is below 0.05, H_0 will be rejected and H_1 will be accepted (Breusch & Pagan, 1979).

H_0 $\alpha_i = 0$ *(variance across entities is zero, homoscedasticity exists, homogeneous model should be used).*

H_1 $\alpha_i \neq 0$ *(variance across entities is not zero, heteroscedasticity exists, heterogeneous model should be used).*

The Breush-Pagan test results in a test statistic (*Langrange multiplier*) with a p-value of 1.28×10^{-8}. Because this value is below 0.05 the null hypothesis is rejected. It can be concluded that variances across entities exist, and a heterogeneous model is most appropriate.

Fixed effects or random effects?

Within the area of heterogeneous panel models, two main types of models exist, being *fixed effects* models and *random effects* models (Katchova, 2013). The difference between a fixed effects and a random effects model lies in the type of variance between boroughs (α_i) (Chipperfield & Steel, 2012). A fixed effects model assumes that the effects of the housing variables (x_{it}) are identical for each borough, and a random effects model does not (Wooldridge, 2010). To estimate whether α_i is random or fixed, the Hausman test can be conducted, which tests the following hypotheses. H_0 states that the housing variables and borough-specific differences effects do not correlate significantly and a random effects model is preferred, and H_1 states the opposites and advocates the use of a fixed effects model. If the p-value is below 0.05, H_0 will be rejected and H_1 will be accepted (Hausman, 1978):

$$\begin{aligned} H_0 \qquad & cov(X_{it}, \alpha_i) = 0 \\ H_1 \qquad & cov(X_{it}, \alpha_i) \neq 0 \end{aligned}$$

The Hausman test returns a test statistic with a p-value of 0.2074. The p-value is above 0.05, thus the null hypothesis is accepted: the effects of housing variables are assumed to be uncorrelated with the boroughs. Therefore, a random effects model is considered a more efficient estimator than a fixed effects model and will be selected as the final model to assess the relationships between housing variables and the sentiment and proportion of housing tweets.

The random effects model assumes that the differences across boroughs (α_i) are not correlated with the housing variables (x_{it}). Therefore, differences between entities (α_i) are captured by the composite error term ϵ_{it}, instead of in the predictors (Park, 2011). The formula for the random effects models that will be used to to assess the relationship between housing variables and the sentiment and proportion of housing tweets is expressed in equation (7):

$$\begin{gathered} y_{it} = \beta_1 x_{1,it} + .. + \beta_k x_{k,it} + \epsilon_{it} \\ \epsilon_{it} = \alpha_i + e_{it} \end{gathered} \tag{7}$$

4.6 Variable evaluation

Choosing the right independent variables is important to obtain a valid estimation of the effect of housing variables on the sentiment and proportion of housing tweets. The selection of variables was thoroughly substantiated in the literature review (chapter 2) and the theory development (chapter 3). This section assesses the quality of the housing variables that were collected in section 4.3.

This research assesses the statistical quality of the variables on the level of multicollinearity. Multicollinearity refers to the concept of *highly correlated independent variables* (Wooldridge, 2010). Models that use highly correlated independent variables present less valid results (Daoud, 2017). The main problem caused by multicollinearity, is that it may influence the regression coefficient of one variable because the coefficient depends on the presence of other variables in the model through correlation (Daoud, 2017).

Two types of multicollinearity exist. The first type is *structural multicollinearity*, which occurs when predictors are created from other predictors (Daoud, 2017). To illustrate structural multicollinearity consider a model that uses three independent variables as input: a, b and c. The variables a and b are completely independent of each other, but the variable c is constructed by adding variable a and b together ($a + b = c$). In this model, structural multicollinearity occurs because the model would try to see what patterns are caused by variable c, though they are already explained by variables a and b.

The second type of multicollinearity is *databased multicollinearity*, which occurs through database related problems such as the inability to change the methods on which the data is collected in an observational study (Daoud, 2017).

Multicollinearity can be detected by measuring the variance inflation factor (VIF) of a variable. The VIF-score exists for all variables in a regression model and can be interpreted as *"the factor by which the variance of the β of regressor j is inflated because of correlation with other independent variables"* (Daoud, 2017). The formula for calculating the VIF-score of variable j is expressed in equation (8), where R_j^2 is the R^2 value obtained by regressing the variable j on the other variables.

$$VIF_j = \frac{1}{1 - R_j^2} \tag{8}$$

Variables with high VIF-scores could be problematic for the validity of the model (Daoud, 2017). By removing a variable with high VIF-score, the VIF-scores of other variables can be reduced, and therefore VIF-score is used as an argument to include or exclude variables from a model. However, there is no universally agreed-upon threshold at which a variable should be excluded from a model. Hair et al. (1995) state that variables with a VIF-score exceeding ten should not be included in the model, while Mela and Kopalle (2002) argue the VIF-score threshold depends on the circumstances and on what the research is trying to observe. Others say that a VIF-score of as low as five could be problematic for the validity of the model Daoud (2017).

Table 8 shows the variance inflation factors for the housing variables that this thesis uses as input for the random effects models to estimate the effect of these variables on the sentiment and proportion of housing tweets.

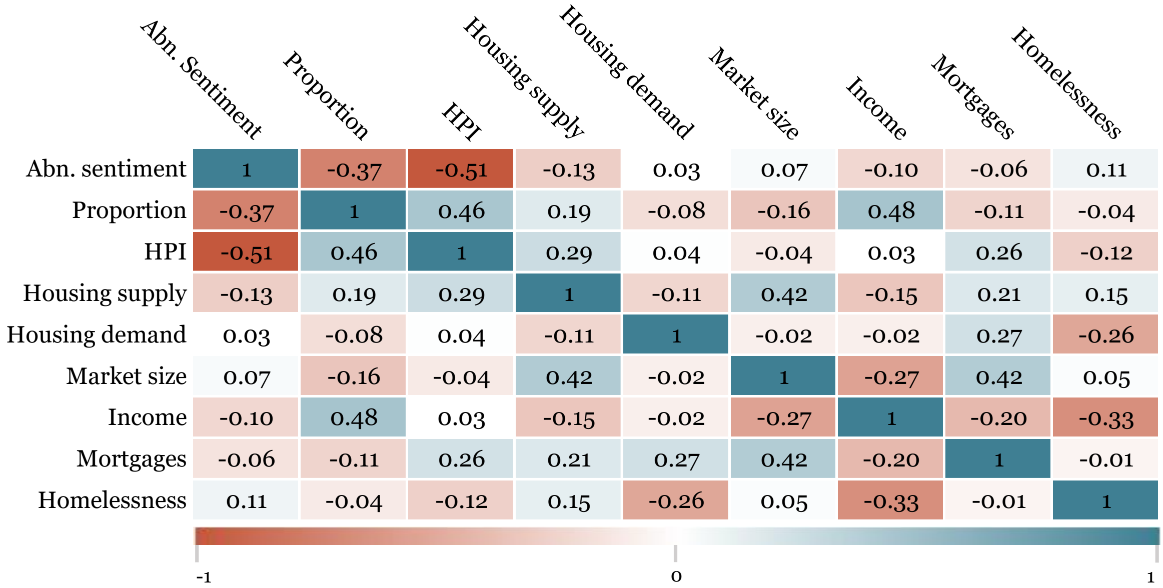
Table 8: *Variance Inflation Factors housing variables*

Variable	VIF-score
House price index	13.4
Housing supply	3.7
Housing demand	1.3
Market size	11.6
Income	8.8
Mortgage	12.9
Homelessness	4.8

The VIF-scores for the house price index, market size, mortgage and income variables indicate database-related multicollinearity for these four variables. To explain these FIV-scores, the correlation matrix in figure 6 can be consulted, which shows how each variable is correlated with every other variable.

The correlation matrix can be used in conjunction with the FIV-table to theorise various explanations for the VIF-scores. However, theorising all possible explanations for correlations between variables with high FIV-scores is beyond the scope of this research for three reasons.

Firstly, this study is of observational nature, and there is no way to change the



	Abn. Sentiment	Proportion	HPI	Housing supply	Housing demand	Market size	Income	Mortgages	Homelessness
Abn. sentiment	1	-0.37	-0.51	-0.13	0.03	0.07	-0.10	-0.06	0.11
Proportion	-0.37	1	0.46	0.19	-0.08	-0.16	0.48	-0.11	-0.04
HPI	-0.51	0.46	1	0.29	0.04	-0.04	0.03	0.26	-0.12
Housing supply	-0.13	0.19	0.29	1	-0.11	0.42	-0.15	0.21	0.15
Housing demand	0.03	-0.08	0.04	-0.11	1	-0.02	-0.02	0.27	-0.26
Market size	0.07	-0.16	-0.04	0.42	-0.02	1	-0.27	0.42	0.05
Income	-0.10	0.48	0.03	-0.15	-0.02	-0.27	1	-0.20	-0.33
Mortgages	-0.06	-0.11	0.26	0.21	0.27	0.42	-0.20	1	-0.01
Homelessness	0.11	-0.04	-0.12	0.15	-0.26	0.05	-0.33	-0.01	1

Figure 6: *Correlation matrix*

research design to reduce the multicollinearity, besides dropping variables. However, the goal of this research is to observe the effect that these housing variables have on the sentiment and proportion of housing tweets. Since the variables are core to this research, the variables will not be disregarded. Secondly, sentiment is an intangible and abstract concept, and is therefore expected to be affected by many different variables that may correlate with each other (Liu, 2012). Thirdly, correlation does not imply causation, and therefore it is difficult to isolate the exact reasons for the FIV-scores. To conclude, all variables mentioned in table 8 will be used as input for the random effects models, the result of which will be described in the following chapter 5: *results and findings*.

5 Results and Findings

This chapter presents the results of this research and is divided into two parts, that will each answer a research question. Firstly, section 5.1 presents a descriptive analysis of how the sentiment and proportion of housing tweets have changed over the years, answering research question one. The second section 5.2 presents the results of the two random effects models, which show the effect of housing variables on the sentiment and proportion of housing tweets. The results of the models will be used to confirm or reject all 14 hypotheses, answering the research question two. The results presented in this chapter will be interpreted and discussed in more detail in chapter 6: *discussion and conclusion*.

5.1 Descriptive analysis Twitter and housing data

To adequately interpret the spatial aspects of the Twitter data, one first needs to understand the distribution of housing prices in London. Figure 7 shows the average house prices and the percentage increase in house prices over 2012-2018 per London borough. The figure is an abstract geographical map of London, in which each block represents a borough. The colour of the blocks represents the value of the average house prices. The acronyms used in each block represent the full borough names. The explanation for each acronym can be found in appendix A.

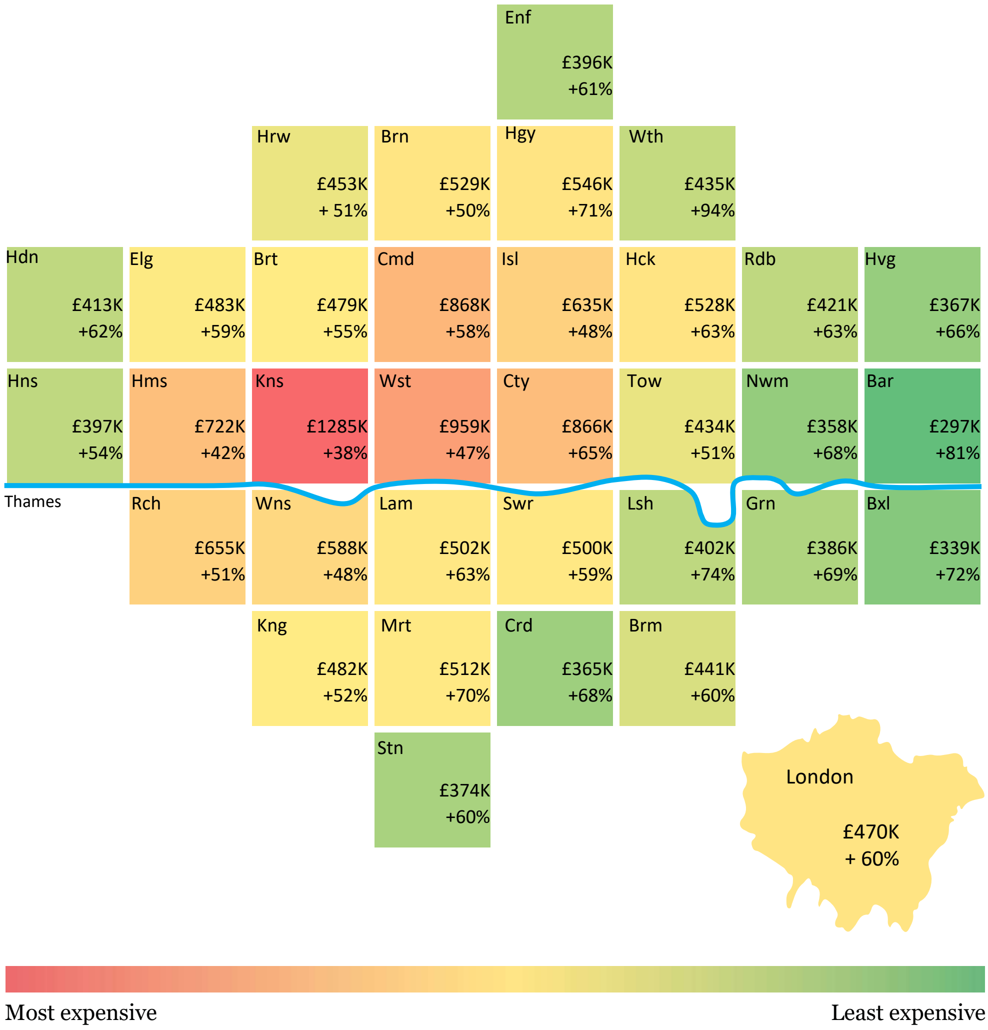


Figure 7: *Average house prices in 2018 and % increase over 2012-2018*

As can be observed from this figure, the average house price in London over 2012-2018 was £470 000. In this period, the average house prices in London increased by 60%. The most expensive boroughs are Kensington and Westminster, which are near the centre of London. The further out boroughs are from the city centre, the lower the average housing prices are. Another interesting observation is that the percentage increase in house prices is very different per borough, ranging from 38% in Kensington to +94% in Waltham Forest. In various boroughs where house prices are relatively low (Barnet, Bexley, Waltham Forest), the percentage increase since 2012 is high. In the most expensive boroughs, the percentage increase in house price since 2012 is relatively low.

5.1.1 Descriptive analysis of housing tweet sentiment

Figure 8 shows how the sentiment of housing tweets changed over the years 2012-2018 for all London boroughs. Each block represents a borough, and each column in a block represents the average abnormal sentiment of housing tweets in one year. For readability and interpretation purposes, the abnormal sentiment is standardised on a scale from -100 to +100. The y-axis is set equal for each block and ranges from 50 at the top, to -50 at the bottom. The middle of each block is 0, representing the average sentiment of all tweets sent from that borough. Each block contains seven columns, representing seven years. The number in the block represents the average annual abnormal housing tweets in that borough over 2012 to 2018.

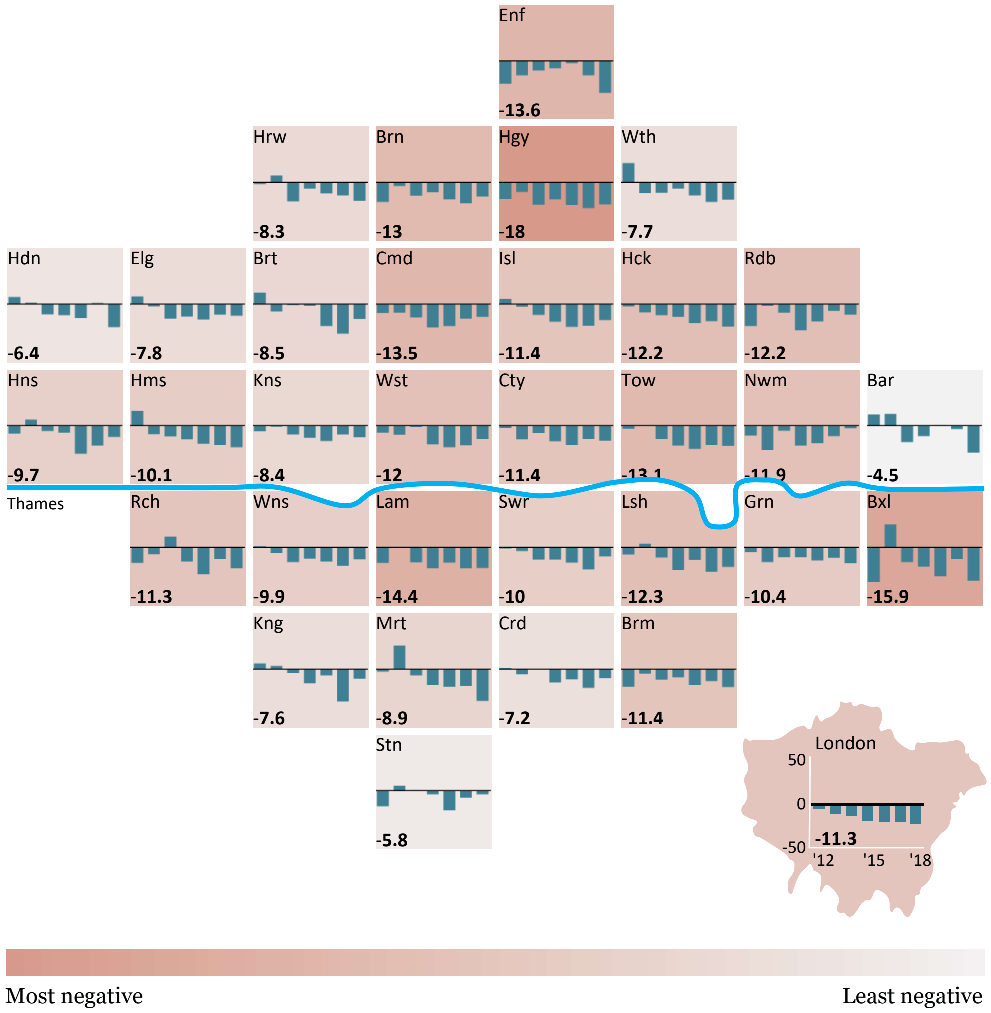


Figure 8: *Annual abnormal sentiment housing tweets per borough*

As can be observed in the bottom right of figure 8, the average abnormal housing sentiment has steadily decreased every year across London as a whole. Over the years 2012-2018, the average abnormal sentiment for housing tweets was -11.3, which means that the average sentiment of housing tweets was -11.3 more negative than the average sentiment of all tweets sent from London on a scale from -100 to +100 (which equates to 5.65%). Additionally, very few boroughs saw a year where the average abnormal sentiment was positive. Only the boroughs Waltham Forest, Hammersmith & Fulham, Barking & Daenham, Merton and Bexley saw a year where the average sentiment of housing tweets was significantly higher than the overall average sentiment of all tweets sent from those boroughs. Furthermore, this figure shows that housing tweets sent from Haringey were most negative about the housing market, with an average housing sentiment of -18 (9%) below the average sentiment of all tweets over 2012-2018 sent from that borough. A detailed table with all numbers for this figure can be found in appendix F.

Figure 9 shows the average sentiment of all tweets and the average sentiment of housing tweets across all boroughs for each month from 2012 to 2018. When the two lines of figure 9 are subtracted from each other, the *abnormal sentiment* is obtained.

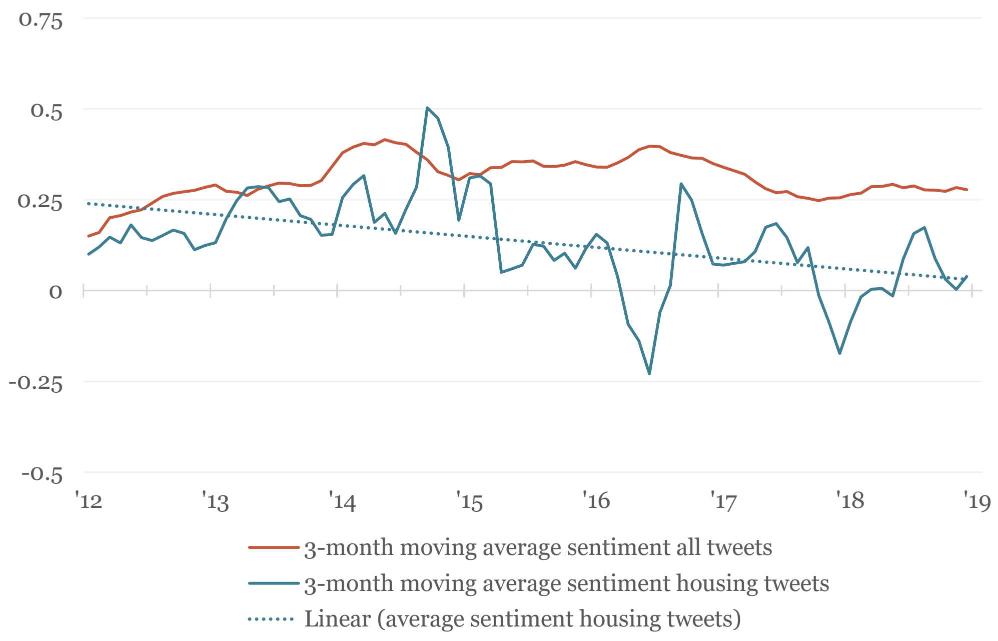


Figure 9: *Monthly average sentiment of tweets and housing tweets*

As the dotted trendline in figure 9 shows, the average sentiment of housing tweets declines over time. Another interesting observation is that the average sentiment for housing tweets is lower than the average sentiment of all tweets for almost every single month from 2012 to 2018, indicating that housing tweets contain a more negative sentiment than other tweets. The only period in which the average sentiment of housing tweets was more positive than the average sentiment of all other tweets, was during a brief period in the second half of 2014, as shown by the intersecting lines.

The rising house prices are hypothesised to be the primary driver of housing sentiment, as it most directly affects Londoners out of all housing variables. To illustrate the correlation between the house price index and abnormal sentiment of housing tweets, both variables are plotted in figure 10 in which each dot represents a borough at one year.

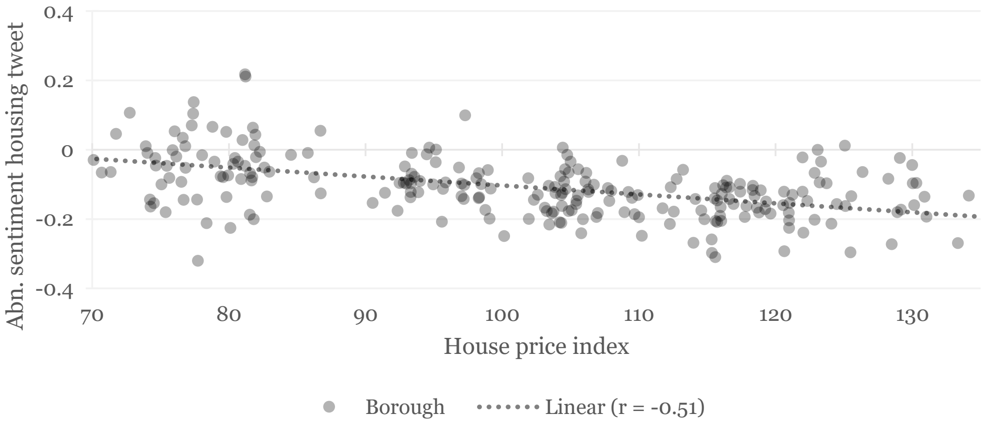


Figure 10: *Abnormal sentiment housing tweets vs. house price index*

Figure 10 shows a clear negative relationship between the house price index and abnormal sentiment found in housing tweets ($r = -0.51$). As hypothesised, this indicates that when house prices increase, the sentiment of tweets decreases. The figure shows that in the house price index range of 70 to 85, there are still some relatively positive housing tweets. However, when the house price index starts rising above 90, the sentiment of housing tweets drop well below the average of all tweets for almost every month in every single borough.

5.1.2 Descriptive analysis proportion of housing tweets

Figure 11 shows how the proportion of housing tweets has changed per over the years 2012-2018 for all London boroughs. Each block represents a borough, and each column in a block represents the average number of housing tweets per 100 000 tweets in that year. The y-axis is set equal for each borough: the top of the block measures 100 housing tweets per 100 000 tweets. The number in the cell represents the average number of housing tweets per 100 000 tweets over 2012 to 2018 in that borough. The London average is shown on the bottom right of the figure.

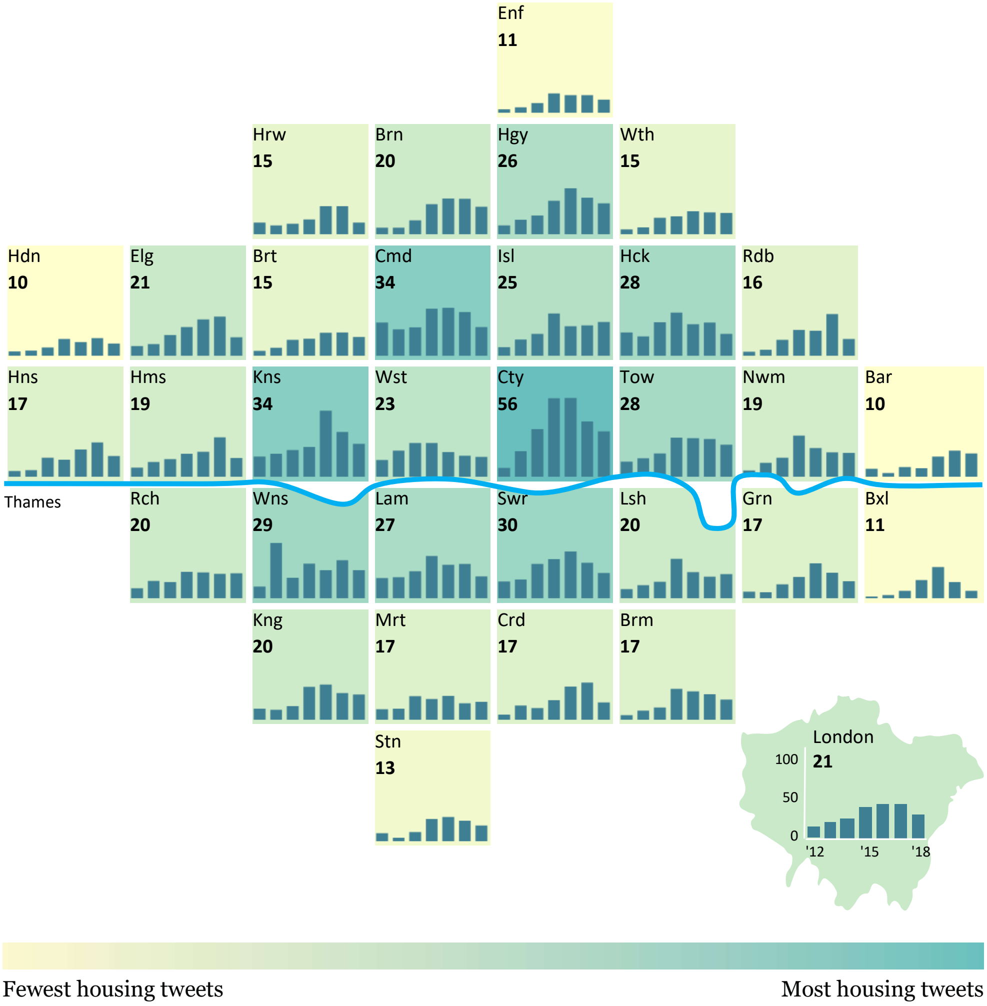


Figure 11: *Proportion of housing tweets per borough from 2012-2018*

As shown in the bottom right of figure 11, in the entire city an average of 21 housing tweets were sent for every 100 000 tweets in the 2012 to 2019 period. The boroughs with the highest proportion of housing tweets are City of London (56), Camden (34), and Kensington (34). The figure shows that the relative number of housing tweets is higher in the inner boroughs of London, where the house prices are high. Near the edges of London, where the house prices are much lower, there are fewer tweets sent about the housing market. This indicates a positive correlation between house prices and the proportion of housing tweets. A detailed table with all numbers used to create for this figure can be found in appendix G.

Figure 12 shows how the proportion of housing tweets and the house price index have moved over the period 2012-2018.

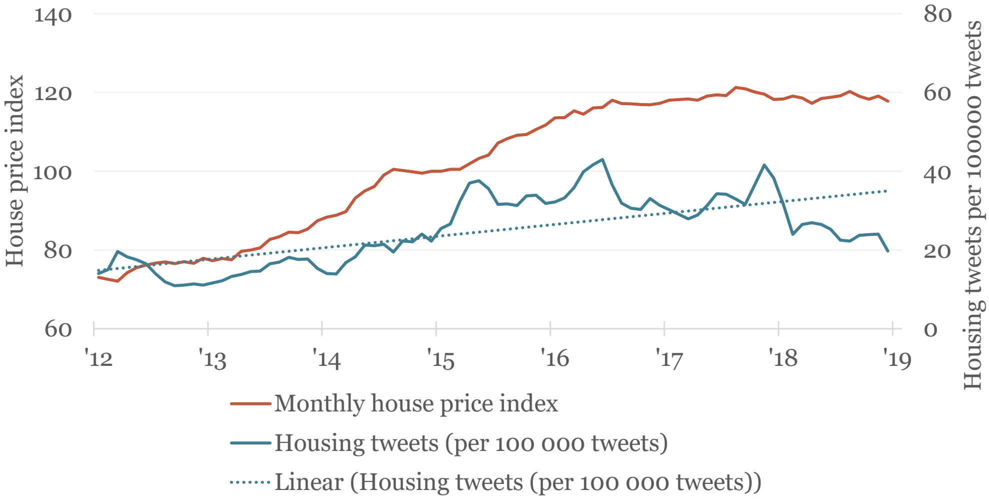


Figure 12: *House price index and proportion of housing tweets per month*

Figure 12 shows that the proportion of housing tweets steadily increased over the years 2013 to 2016, from approximately 15 to 40 housing tweets per 100 000 tweets. From 2016 to 2018, the number of housing tweets hovered around 35 per 100 000 tweets. In 2018, it dropped back to 20 housing tweets per 100 000 tweets. The figure also shows that the proportion of housing tweets and the house price index moved in similar fashion from 2012 to 2016, after which the two lines diverged. A possible explanation for this divergence is that the house price index stabilised in the years 2017 and 2018, and therefore there is less discussion about the rising house prices.

Both figure 11 and figure 12 show a relationship between the house prices and the proportion of housing tweets. Figure 11 shows that in the inner boroughs of London, where house prices are high, relatively more housing tweets are sent. Figure 12 shows that the house price index and proportion of housing tweets move in similar fashion. These observations indicate a positive correlation between rising house prices and the proportion of housing tweets. This assumed correlation between house price index and the proportion of housing tweets is displayed in figure 13. Each dot represents a borough at one year.

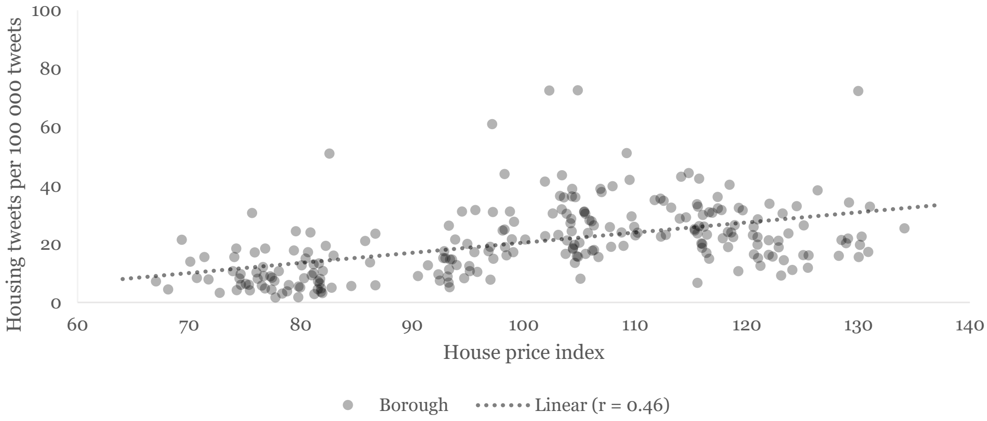


Figure 13: *House price index vs proportion of housing tweets tweets*

The figure shows a positive linear relationship ($r = 0.46$) between the house price index and the proportion of housing tweets. As hypothesised, this indicates that when house prices increase, the number of tweets increases. This figure shows more distinct outliers than the scatter plot for abnormal sentiment and house price index (figure 10). When cross-checking these outliers with figure 11, it shows the outliers all represent the borough *City of London*, which is the borough in which the relative number of housing-related tweets is highest, by far.

5.1.3 Spatial analysis

This section emphasises the spatial aspect of the tweets. Although the spatial aspect of tweets does not contribute to answering the research questions, the spatial analysis does provide some interesting results that enrich this research. Figure 14 shows all housing tweets with GPS coordinates are plotted on the map of inner London. The colour of the dots represent the sentiment value: the reddest points represent the most negative tweets, and the more blue a point is, the higher the sentiment score of the housing tweet is. The black lines represent the borough borders. This figure excludes the outer boroughs, because the majority of tweets are sent from the inner boroughs, and it allows for an enlarged representation of the inner boroughs. A full map of London with all geotagged housing tweets can be found in appendix H.

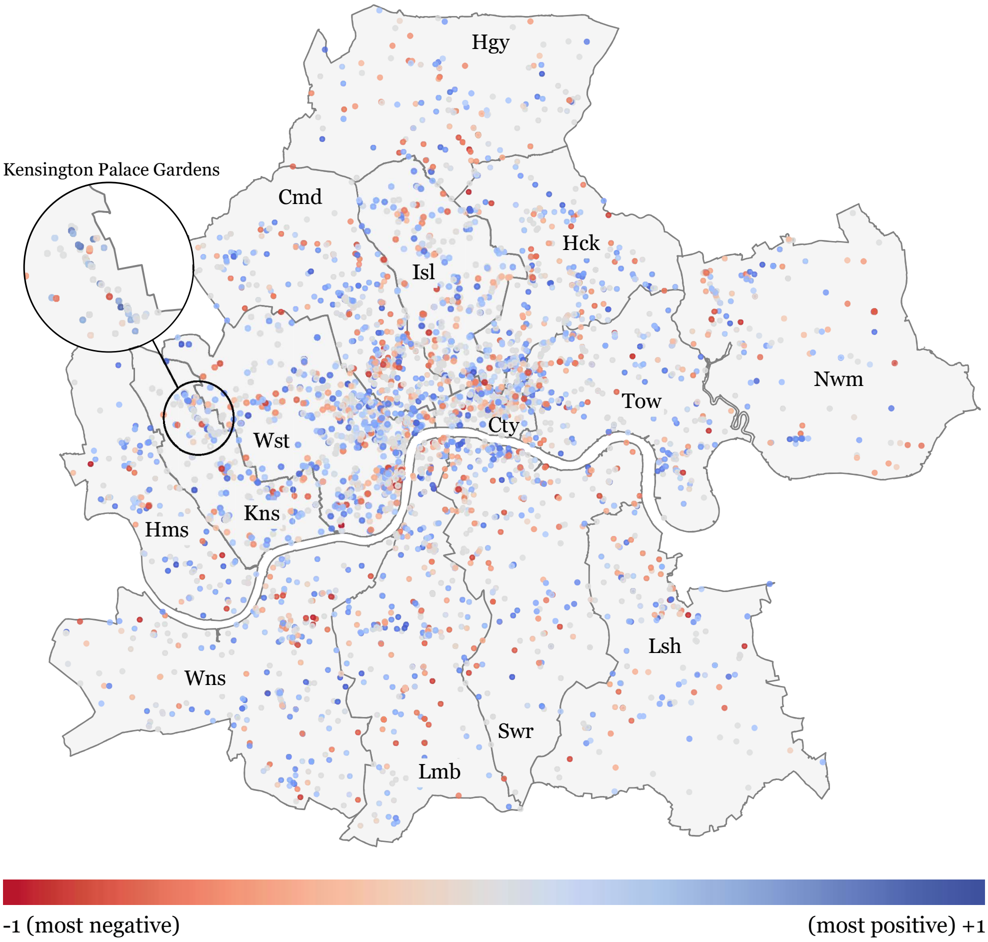


Figure 14: *Spatial distribution of geotagged housing tweets over inner London*

Some interesting insights can be derived from this figure. The map shows that within the inner boroughs, the density of tweets is highest for the boroughs Westminster, Camden and City of London, which are also the boroughs with the highest proportion of housing tweets, as shown in the previous figure 11.

The figure also shows other interesting patterns, such as the clear distinction of housing tweets sent from the street Kensington Palace Gardens. The average house price in this street is £33 million (Midolo, 2019), making it the most expensive street in England. The street shows only a few red dots, meaning that passers-by do not send many negative tweets. Perhaps the Twitter users are impressed by the imposing houses in this affluent area.

So far, the figures presented in the descriptive analyses have answered research question one by showing how the level of sentiment and proportion of housing tweets have changed over the past years. However, the descriptive statistics only showed *how* the sentiment and proportion of housing tweets changed over the years, but not *why*.

The results of the random effects models will explain *why* the sentiment and proportion of housing tweets changed over the years, and will answer research question two. The effects of the housing variables on sentiment and housing tweets will be statistically supported by the model results, which are presented in the following section 5.2.

5.2 Model results

Two models were constructed to answer research question two. The first model estimates the effect of changes in the housing market on the *sentiment* of housing tweets (5.2.1). The second model estimates the effect of changes in the housing market on the *proportion* of housing tweets (5.2.2). The results of both models will be discussed and explained in the following sections. Additionally, the hypotheses developed in chapter 3 will be rejected or accepted based on the significance level for each relationship. Appendix B serves as a guideline for variable interpretation and presents a summarised description of all variables. This chapter only presents the results of the models. The results will be interpreted in chapter 6.

5.2.1 Effect of housing variables on sentiment of housing tweets

Table 9 summarises the results of the model that estimates the effect housing variables on the sentiment of housing tweets. This table contains the estimated coefficients ($\hat{\beta}$) that represent the effect of the respective housing variable on the sentiment of housing tweets. The coefficient indicates how much the dependent variable changes on average, when the independent variable changes by one unit, while holding other independent variables constant (Dunning, 2008). This allows observing the isolated effect of an independent variable on the dependent variable.

To illustrate how the coefficients of this model should be interpreted, consider the coefficient of house price index, which is -0.2636. This coefficient should be interpreted as follows: for an increase in the house price index of one unit, the level of sentiment of housing tweets decreases by 0.2636. Because sentiment is standardised on a scale from -100 to +100 for interpretation and readability purposes, the coefficient can be divided by two to obtain the percentage change.

Table 9: *Summary of model results: sentiment of housing tweets*

Variable	**Coefficient**	**H**	**Conclusion**
House price index	-0.2636***	H1a	Accepted: for every point increase in HPI, abnormal sentiment of housing tweets decreases by 0.1318%
Income	+0.8531**	H2a	Accepted: for every additional £1000 in average salary, abnormal sentiment of housing tweets increases by 0.4265%
Mortgage	Insignificant	H3a	Rejected: the cumulative mortgage debt per postcode sector does not significantly impact abnormal sentiment of housing tweets
Housing supply	Insignificant	H4a	Rejected: the number of additional dwellings does not significantly impact abnormal sentiment of housing tweets
Housing demand	-0.0451*	H5a	Accepted: for every additional 1000 citizens, abnormal sentiment decreases by 0.0226%
Market size	Insignificant	H6a	Rejected: the number of housing sales does not significantly impact abnormal sentiment of housing tweets
Homelessness	+0.9024*	H7a	Rejected: for every additional household that is accepted as homeless per 1000 households, abnormal sentiment increases by 0.4512%

*p<0.1; **p<0.05; ***p<0.01

Table 9 is a summary of the more in-depth table 10. Table 10 displays all results of the model when different combinations of variables are used as input.

In total, seven combinations of variables are used as input for this first model, each containing an additional housing variable. The main observation from table 10 is that variable sets (1a) through (4a) yield significant results. The variables added in (5a), (6a) and (7a) do not significantly increase the explanatory power of the model. Therefore, (4a) will be considered the main model. This section describes the results of (4a) in more depth.

Table 10: *Model results: effect of variables on sentiment of housing tweets*

	Dependent variable: sentiment of housing tweets						
	(1a)	(2a)	(3a)	(4a)	(5a)	(6a)	(7a)
House price index	-0.1097*** (0.0051)	-0.2222*** (0.0306)	-0.2857*** (0.0003)	-0.2636*** (0.0427)	-0.2730*** (0.0483)	-0.2646*** (0.0491)	-0.1986*** (0.0661)
Homelessness		1.3770*** (0.3401)	0.7678** (0.3777)	0.9024* (0.3836)	0.8994* (0.3823)	0.4908 (0.4654)	0.5415 (0.2395)
Income			0.6104** (0.2975)	0.8531** (0.3304)	0.8620** (0.3343)	0.8143** (0.3320)	1.2014*** (0.4020)
Housing demand				-0.0451* (0.0264)	-0.0443* (0.0264)	-0.0441* (0.0255)	-0.0409* (0.0262)
Housing supply					0.0008 (0.0010)	0.0005 (0.0000)	0.0008 (0.0011)
Market size						0.0024 (0.0011)	0.0022 (0.0011)
Mortgage							-0.0844 (0.0628)
R^2	0.6496	0.7294	0.7379	0.7471	0.7476	0.7536	0.7559
F Statistic	43.17***	37.318***	15.730***	15.954***	15.469***	15.457***	15.161***

*p<0.1; **p<0.05; ***p<0.01

Firstly, the results show an increase in housing prices in a borough has a significant negative impact on the sentiment found in housing tweets sent from that borough ($\hat{\beta} = -0.2636, p < 0.01$). For every unit increase in the house price index, the average sentiment found in housing tweets decreases by 0.2636, or 0.1318%. Because the house price index has a significant effect on the proportion of housing tweets at the 5% significance level, hypothesis 1a is confirmed: rising house prices in a borough are negatively correlated with the sentiment of housing tweets sent from that borough.

Secondly, the rate of homelessness significantly affects the level of sentiment ($\hat{\beta} = 0.9024, p < 0.1$). For each additional household that is considered statutory homeless in a borough, the sentiment found in housing tweets sent from that borough increases by 0.9024 or 0.4512%. Though the finding is significant, it is opposite of what was hypothesised and above the 5% significance threshold. Thus, hypothesis 7a is rejected: the rate of homelessness in a borough is not negatively correlated with sentiment found in housing tweets sent from that borough.

Thirdly, the level of annual income in a borough significantly affects the sentiment of housing tweets ($\hat{\beta} = 0.8531, p < 0.05$). When the average annual income in a borough increases with £1000, the sentiment found in housing tweets sent from that borough increases by 0.8531, or 0.4265%. The relationship between the level of annual income and the level of sentiment in housing tweets is significant below a 5% level. This confirms hypothesis 2a: the average annual income in a borough is positively correlated with sentiment found in housing tweets sent from that borough.

Finally, the demand for housing has a significant negative effect on the tone of housing tweets ($\hat{\beta} = -0.0451, p < 0.01$). For every extra 1000 inhabitants in a borough, the sentiment found in housing tweets sent from that borough decrease by 0.0451, or 0.02255%. As this relationship is significant below the threshold of 5%, hypothesis 5a is confirmed: housing demand in a borough is negatively correlated with sentiment found in housing tweets sent from that borough.

As shown in model (5a), (6a), and (7a), the three remaining variables do not have a significant effect on the abnormal sentiment of housing tweets. Market size, measured through the number of property sales transactions, does not have a significant effect on the level of sentiment of housing tweets. A similar conclusion can be drawn for the housing supply variable, measured in the number of net additional dwellings. The average mortgage per postcode sector also does not significantly impact the sentiment of housing tweets sent from that borough. As these three variables have no significant effect on the level of sentiment of housing tweets, hypotheses 3a, 4a, and 6a are rejected.

The R^2 of the sentiment estimation model (4a) is 0.7471, indicating that the housing variables of this model explain 74.71% of the variation in abnormal sentiment of housing tweets.

5.2.2 Effect of housing variables on proportion of housing tweets

Table 11 summarises the results of the model estimating the effect of changes in the housing market on the proportion of housing tweets sent from London.

Table 11: *Summary of model results: proportion of housing tweets*

Variable	Coefficient	H	Conclusion
House price index	+0.4506***	H1b	Accepted: for every point increase in HPI, 0.4506 additional housing tweets are sent per 100 000 tweets
Income	-1.3076***	H2b	Accepted: for every additional £1000 in average salary, 1.3076 fewer housing tweets are sent per 100 000 tweets
Mortgage	Insignificant	H3b	Rejected: the cumulative mortgage debt per postcode sector does not significantly impact the proportion of housing tweets
Housing supply	Insignificant	H4b	Rejected: the number of additional dwellings does not significantly impact the proportion of housing tweets
Housing demand	-0.0245***	H5b	Rejected: for every additional 1000 citizens, 0.0245 fewer housing tweets are sent per 100 000 tweets
Market size	Insignificant	H6b	Rejected: the number of housing sales does not significantly impact the proportion of housing tweets
Homelessness	+0.8243**	H7b	Accepted: for every additional household that is accepted as homeless, 0.8243 additional housing tweets are sent per 100 000 tweets

*p<0.1; **p<0.05; ***p<0.01

Table 11 is again a summary of the more detailed table 12 which displays the results of the model when different combinations of variables are used as input. In total, seven sets of variables are used as input for this model, each containing an additional housing variable. Interestingly, the same four variables yield significant results for the proportion model, as shown by (4b). Again, adding more variables to the model does not yield additional significant results, as shown in by (5b), (6b) and (7b). Therefore, the results of model (4b) will be assessed.

Table 12: *Model results: effect housing variables on proportion of housing tweets*

	Dependent variable: proportion of housing tweets						
	(1b)	(2b)	(3b)	(4b)	(5b)	(6b)	(7b)
House price index	0.2350***	0.2890***	0.4386***	0.4506***	0.4518***	0.4549***	0.4905***
	(0.0126)	(0.0212)	(0.0290)	(0.0301)	(0.0332)	(0.0917)	(0.0824)
Homelessness		0.6850**	0.7513**	0.8243**	0.8247**	0.6764*	0.7029
		(0.3358)	(0.3639)	(0.3681)	(0.3698)	(0.4302)	(0.04383)
Income			-1.4392***	-1.3076***	-1.3088***	-1.3261***	-1.1169***
			(0.2126)	(0.2114)	(0.2143)	(0.2141)	(0.4214)
Housing demand				-0.0245***	-0.0246***	-0.0245***	-0.0228***
				(0.0080)	(0.0081)	(0.0082)	(0.0076)
Housing supply					-0.0001	-0.0002	-0.0001
					(0.0010)	(0.0011)	(0.0010)
Market size						0.0009	0.0008
						(0.0011)	(0.0010)
Mortgage							-0.0456
							(0.0853)
R^2	0.7915	0.9073	0.9231	0.9240	0.9240	0.9243	0.9245
F Statistic	347.48***	80.984***	67.094***	67.012***	63.515***	61.702***	59.953***

*p<0.1; **p<0.05; ***p<0.01

As observed in model (4b), the house price index has a significant positive effect on the proportion of housing tweets sent ($\hat{\beta} = 0.4506, p < 0.01$), indicating that for every unit increase in the housing price index, 0.4506 extra tweets are sent per 100 000 tweets. This relationship is significant below the 5% threshold, and therefore hypothesis 1b is accepted: rising house prices in a borough are positively correlated with the proportion of housing tweets sent from that borough.

Secondly, the rate of homelessness in London has a significantly effect on the proportion of housing tweets ($\hat{\beta} = 0.8243, p < 0.05$). The coefficient of 0.8432 can be interpreted as follows: for each additional household that is considered statutory homeless in a borough, 0.8243 more housing tweets per 100 000 tweets are sent from that borough. The significant relationships confirms hypothesis 7b: the rate of homelessness in a borough is positively correlated with the proportion of housing tweets sent from that borough.

The third significant variable is income, which has a negative, significant effect on the proportion of housing tweets ($\hat{\beta} = -1.3076, p < 0.01$). This result means that for every extra thousand pounds of average annual salary in a borough, 1.3067 fewer housing tweets are sent per 100 000 tweets. The significant effect of this relationship confirms hypothesis 2b: the average annual income in a borough is negatively correlated with the proportion of housing tweets sent from that borough.

The final significant variable is the housing demand, which shows that for every 1000 additional inhabitants in a borough, the number of housing tweets per 100 000 tweets decreases by 0.0245 ($\hat{\beta} = -0.0245, p < 0.01$). This effect is significant and negative, which contradicts hypothesis 5b, which expected that an increase in housing demand would increase the proportion of housing tweets. The significant relationship rejects hypothesis 5b: housing demand in a borough is negatively correlated with the proportion of housing tweets sent from that borough.

Similar to the sentiment estimation model, the variables of housing supply, housing market size, and the average cumulative mortgages per postcode sector are all insignificant predictors of the proportion of housing tweets as shown in columns (5b), (6b), and (7b). Therefore, hypotheses 3b, 4b, and 6b are rejected.

Model (4b) has a R^2 of 0.9240, indicating that 92.40% of the variance in the proportion of housing tweets can be explained by the independent variables in the model. A reason for the high R^2 may be the presence of dummy variables for each borough, which on their own explain variance in the proportion of housing tweets. It is interesting to note that the R^2 of the sentiment model is lower than the proportion model. A reason for this may be that sentiment is less quantifiable than the number of tweets, which is a more defined and observable variable.

6 Discussion and Conclusion

This chapter summarises the key findings of this thesis in section 6.1, interprets and discusses these findings in section 6.2. Section 6.3 discusses the implications and relevance of the findings on a societal and academical level. Finally, section 6.4 presents the limitations and suggestions for future research.

6.1 Main findings

Two research questions were formulated in chapter 1. The first research question is "*how have the sentiment of housing tweets and the proportion of housing tweets sent from London changed over 2012-2018?*". The second research question is "*what is the effect of shifts in the London housing market on the sentiment of housing tweets and the proportion of housing tweets sent from London?*"

Research question one was answered in the descriptive analysis in section 5.1. Firstly, the tone of the public opinion on the London housing market has become increasingly negative over the years. A clear negative correlation was shown between the rising house prices and the level of sentiment found in housing tweets: the tone of housing tweets gradually decreased from just below the average sentiment of all tweets in 2012, to well below the average sentiment of all tweets in 2018. In some boroughs, the average sentiment of housing tweets is up to 10% more negative than other tweets. The proportion of housing tweets, measured in the number of housing tweets per 100 000 tweets, gradually increased from from approximately 20 in 2013, to 40 in 2016, after which it stabilises at around 25 housing tweets per 100 000 tweets in 2018.

Research question two was answered by constructing two random effects models that estimated the relationship between housing variables and the sentiment and proportion of housing tweets. The results of the models as described in section 5.2, have shown that house prices, housing demand, annual income and the rate of homelessness all have a significant effect on both the sentiment and proportion of housing tweets, as summarised in table 13. These findings will be interpreted and discussed in section 6.2.

These answers to the research questions have fulfilled the research objective by showing how the public opinion on the London housing market has changed over the years, and by showing the effect of shifts in the London housing market on the public opinion on the housing market.

Table 13: *Summary main findings*

		Effect on	
Variable	**Measured in**	**Abn. sentiment**	**Proportion**
House prices	House price index	-0.1318%***	0.4506***
Housing supply	Additional houses	Insignificant	Insignificant
Housing demand	Population (x1000)	-0.0226%**	-0.0245***
Market size	Housing sales	Insignificant	Insignificant
Mortgages	Cumulative mortgage	Insignificant	Insignificant
Income	Annual salary (x£1000)	+0.4265%**	-1.3088***
Homelessness	Statutory homeless (per 1000 households)	+0.4512%*	0.8243***

*p<0.1; **p<0.05; ***p<0.01

6.2 Interpretation and discussion of findings

This research has proven that changes in four housing variables significantly affect both the level of sentiment and the proportion of housing tweets in London. These variables are house prices, housing demand, annual income and the rate of homelessness.

This study shows that for every additional percent increase of the house price relative to the house prices in January 2015, the sentiment of housing tweets decreases by -0.1318%, and 0.4506 more housing tweets are sent per 100 000 tweets. Though not particularly surprising, this relationship was not proven before. This result indicates that as housing prices rise, society becomes more vocal in an increasingly negative way on Twitter about the housing market. These findings are in line with prior works that found rising house prices to be a negative driver of housing sentiment (Wilcox, 2015; Bork & Moller, 2016; Dua, 2008; Wang & Hui, 2017; Soo, 2013). Various explanations for this relationship are that the rising house prices and cost of housing arguably affect Londoners most directly, out of all housing variables. The rising prices increases the concern of Londoners that they may never be able own a house. Also, the rising housing prices cause a decrease in disposable income of Londoners that rent, which is a phenomenon that the average Londoner will most likely not be too fond of.

The second conclusion of this research is that housing demand significantly impacts the sentiment and proportion of housing tweets, as for every additional 1000 inhabitants in a borough, the sentiment level of housing tweets from that borough decreases by 0.0226%, and 0.0245 fewer housing tweets are sent per 100 000 tweets. A possible explanation for why sentiment decreases when housing demand increases, is that the competition for housing rises when demand rises. This may give Londoners the idea that they have to try even harder to obtain a house for a fair price. Additionally, the sense of overcrowding may play a role in explaining the decreasing sentiment when demand for housing increases. The negative correlation between housing demand and the proportion of housing tweets is a surprising finding that contradicted hypothesis H5b. This indicates that when the population of a borough increases, fewer tweets are sent about the housing market. A possible explanation for this negative relationship may be that the increasing number of citizens in a borough may cause Twitter to be used in a more communicative way, in which the housing market is less discussed, proportionally. Another possible explanation may be that in more populated boroughs, more different topics are discussed on Twitter which causes the relative number of housing tweets to fall.

Thirdly, the average level of annual income in a borough significantly affects the sentiment housing tweets. For every additional £1000 average annual salary in a borough, the level of sentiment found in housing tweets increases by 0.4265%, and 1.3088 fewer tweets per 100 000 tweets are sent from that borough, confirming the findings of Wilcox (2015), Bork and Moller (2016), Dua (2008), Wang and Hui (2017), Soo (2013) that income is a positive driver of housing sentiment. A possible explanation for this positive relationship is that in affluent boroughs, citizens are not as worried about the London housing market since they are able to afford the increasing housing costs. Another reason may be that in these boroughs more people already own a house, and increasing house prices may even be considered beneficial, as it increases the value of their property. Another reason for the increased level of sentiment in more wealthy boroughs may be that the houses in these boroughs are nicer than in less wealthy boroughs, which may cause an overall more positive feeling towards housing.

The fourth significant finding is that the rate of homelessness positively affects the level of sentiment and proportion of housing tweets. For each additional household per 1000 households that is accepted by the local authorities as homeless and in priority need, the sentiment of housing tweets increases by 0.4512%, and 0.8243 additional housing tweets are sent. The positive relationship between homelessness and housing sentiment is surprising, as homelessness was hypothesised to negatively impact the level of sentiment (H7a). It is difficult to think of reasons why an increase in homelessness would cause an increase in housing sentiment. Also, this relationship was relatively low in significance ($p < 0.1$) and therefore this may be a spurious relationship.

This study also concludes that the three variables of housing supply, market size, and mortgage level do not significantly affect either the level of sentiment or the proportion of housing tweets. These insignificant relationships may be explained as follows. For the variables of housing supply and market size, the results of constructing new houses and a growing housing market may possible only be noted after a certain time lag. This study does not consider a time lag for these two variables, which, in hindsight, maybe an explanation for why these variables do not significantly affect the sentiment or proportion of housing tweets. Another possible explanation is that the level of housing supply and the size of the housing market do not impact the population as directly as for example the rising house prices. The relationship between mortgages and the sentiment and proportion of housing tweets is also found to be insignificant. An explanation for this insignificance may be that the level of mortgage as measured in this research is affected by other factors that were not considered in this study. A possible factor is that citizens move to a postcode sector and take out a new mortgage and therefore affect the cumulative outstanding mortgage debt.

6.3 Research relevance and implications

This study can be seen as a pioneering effort since it is the first of its kind to use Twitter data to observe changes in the public opinion on a housing market. Additionally, it has shown what components of the London housing market have the most effect on the public opinion on the housing crisis of Londoners. The methods and results presented in this study are relevant to various stakeholders on a practical and academic level.

Firstly, understanding the effect of various housing variables on the public opinion on a housing crisis is relevant for cities all over the world. City authorities can use these data-driven insights to prioritise what components of the housing crisis to focus on, to most effectively improve public opinion. Systematically improving the public view of the housing situation will increase overall satisfaction and happiness of the city population. As this study is tailored towards London, the results are most applicable to London specifically. However, the practical implications extend to cities well beyond just the UK capital. The results and frameworks of this study can assist any other city in the world by taking preemptive measures to make sure the city does not undergo the same fate as London.

Additionally, these insights assist policymakers in assessing how the public opinion on the housing market develops in the future. For example, one can estimate how the public opinion on the housing market will develop when observing trends in house prices, housing demand, levels of income and homelessness.

Furthermore, this study has shown that Twitter data can be a source of insights about the public opinion on a crisis. This research observes the how the public opinion on a housing crisis specifically is formed, but the methods used in this study also apply to measure the public response to other phenomena. Therefore, this research is also relevant for cities struggling with other types of crisis than a housing crisis.

Besides practical implications, this study also has valuable academic implications. This study shows how the public opinion is affected by changes in the housing market, which is a novel way of observing the public opinion on the housing market. Using the novel ways presented to measure the public opinion on the housing market builds upon a larger body of works about the use of Twitter as a measure of public opinion (O'Connor et al., 2010; Bollen, Pepe, & Mao, 2009; Cody et al., 2016). By leveraging the large quantity of opinionated data on Twitter to construct the public opinion on the housing market, this research addresses the following six gaps that were left open in prior work on measuring housing sentiment (Wilcox, 2015; Bork & Moller, 2016; Dua, 2008; Croce & Haurin, 2009; Wang & Hui, 2017; Baker & Wurgler, 2006; Soo, 2013; Hui & Wang, 2014).

Firstly, this study expanded the existing literature on housing sentiment by researching the rate of homelessness in a city as a driver of housing sentiment. Secondly, analysing multiple years of historical Twitter data allowed this research to observe how the public opinion on the housing market in London has changed on a very detailed temporal level. This is an advantage over to traditional methods of measuring public opinion, that analyse a snapshot of the public opinion at a single moment in time. Thirdly, this research used a more cost- and time-efficient method of data collection by using publicly available Twitter data instead of using surveys. Fourthly, prior studies on housing sentiment have constructed sentiment on the housing market by asking reflective questions that are difficult to answer (Fowler, 2009), which is overcome by using Twitter, because it does not require users to think about how they feel about the housing market compared to some point in time. Fifthly, through the use of publicly available Twitter data, this study avoids various types of response biases that commonly occur when administering surveys. Finally, most studies researching housing sentiment or studies measuring public response on Twitter did not incorporate spatial data. This research, however, used the spatial aspects of tweets which increased model performance and provided a better view of how the public responds to a crisis, and allowed for comparing results between geographical areas.

Finally, this study assists researchers in the field of social media research to understand how Twitter and social media in general can be used to measure the public response to phenomena. The methods and frameworks presented in this research could be used for measuring sentiment about any topic on Twitter.

6.4 Limitations and future research

This study has provided relevant insights on the topic of public opinion-forming and the London housing market. The research provides a solid basis for further research in various ways, as this research can be generalised in several directions. However, before encouraging other researchers to further elaborate on the topic and methods presented in this thesis, it is important to note the limitations of this study. The limitations of this study are divided into two categories that will both be discussed. First, the limitations are discussed that should be considered when interpreting the results. The second type of limitations are aspects that were beyond the scope of this research, but would have been interesting to study and are therefore presented as suggestions for future work.

The first limitation of this research lies within two possible shortcomings of Twitter data. Users can manually add any place in the world to a tweet using the 'place_id' option, without necessarily being physically present at that location (Abbasi et al., 2015). As an example, a person in the borough Bexley could have manually added the borough Westminster as a 'place_id' to the tweet. Twitter is unable to verify if this location tag corresponds with the exact location from which the tweet was sent, which introduces a possible location bias. The second bias of Twitter is a population bias, as Twitter is not fully representative of the entire population, because not everyone is active on Twitter (Ruths & Pfeffer, 2014).

Secondly, part of this research uses the concept of sentiment as a dependent variable, which presented various challenges. However, these challenges could be overcome in future research. Sentiment is an intangible concept and is an emotion or feeling someone has. Sentiment can be based on various constructs at once, and changes in sentiment are therefore challenging to isolate (Liu, 2012). Because of this, as described in section 4.6, housing sentiment might be a result of variables that are not completely uncorrelated, introducing multicollinearity. Examples of unobserved effects that could impact the tone of a tweet are the individual moods of users, political decisions regarding the housing market, or any housing-related events that may have happened. Though this research did manage to uncover a part of the drivers of housing-related sentiment, more unobserved drivers exist. Discovering the unobserved drivers of housing sentiment and using these variables as additional inputs for future models will further improve understanding the public opinion on housing sentiment.

A suggestion for further research is the use of additional aspects of Twitter that were beyond the scope of this research, but could have provided valuable insights. This research only studies the sentiment level of tweets, but disregards the actual text of the tweets. Therefore the only conclusions that could be drawn about *how* things are said on Twitter, and not *what* is said on twitter. Future research could study *what* is said about the housing market through the use of advanced language processing methods like topic modelling. Another aspect of the tweets that was beyond the scope of this research is the communicative component of Twitter. By using the "@<username> <message>" syntax, users can to mention other users in their tweets. Another option is to *retweet* the tweet of another user, which shares the tweet to the timeline of a user for their followers to see. The retweet feature was disregarded in this study because the dataset did not contain or specify retweets. These communicative aspects of Twitter can be studied to observe communication patterns and information flows between users. For this research specifically, these communicative patterns would have added value by showing how the housing market is discussed between users.

Another suggestion for future research is to use different types of classifiers to distinguish relevant tweets. The rule-based classifier used in this study shows few false positives (high precision), but was not able to correctly classify every single tweet describing the London housing market as "housing-related", resulting in inescapable false negatives that did not match the inclusion queries (appendix E). Therefore, not all housing tweets sent from London were analysed in this study. Expanding and improving the queries would have increased the number of housing tweets that could be analysed, which in turn would have enriched results. A second way to possibly improve the classification of housing tweets, is by developing a machine learning text classifier. A training set with housing tweets can be used to train a machine learning text classifier to improve the recall rate of the classification process. This will result in a larger number of retrieved housing tweets, but precision might decrease through the incorrect classification of tweets.

Finally, the methods and results of this study can be be generalised in several directions. The first way is that the study can be conducted for different geographical locations. As long as the language of the tweets can be analysed for its sentiment and the proper data is available, the theory and methods provided in this research can be reproduced for any other city.

Secondly, the topic of analysis can be generalised to any other societal problem. Examples of research topics that would be interesting to study using Twitter data are public health issues or social stratification.

Thirdly, this study uses Twitter as a source of opinionated data, but the methods presented in this study are applicable to any other types of opinionated data. Examples of different platforms which can be used to collect large quantities of publicly available opinionated data are YouTube, Reddit, Facebook or Instagram. However, these data sources do not contain spatial data. For all three types of generalisations, it would be interesting to see if the results confirm the methods of this study.

7 References

Abbasi, Alireza et al. (2015). “Utilising Location Based Social Media in Travel Survey Methods”. In: *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN'15*. New York, New York, USA: ACM Press, pp. 1–9. DOI: 10.1145/2830657.2830660.

Arestis, P. and A.R. González (2014). “Modelling the housing market in OECD countries”. In: *International Review of Applied Economics* 28.2, pp. 131–153. DOI: 10.1080/02692171.2013.828683.

Asur, Sitaram and Bernardo A. Huberman (2010). “Predicting the Future with Social Media”. In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, pp. 492–499. DOI: 10.1109/WI-IAT.2010.63.

Baker, Malcolm and Jeffrey Wurgler (2006). “Investor Sentiment and the Cross-Section of Stock Returns”. In: *The Journal of Finance* 61.4, pp. 1645–1680. DOI: 10.1111/j.1540-6261.2006.00885.x.

Bermingham, Adam and Alan F Smeaton (2011). “On Using Twitter to Monitor Political Sentiment and Predict Election Results”. In: *Psychology*, pp. 2–10.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011). “Twitter mood predicts the stock market”. In: *Journal of Computational Science* 2.1, pp. 1–8. DOI: 10.1016/j.jocs.2010.12.007.

Bollen, Johan, Alberto Pepe, and Huina Mao (2009). “Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena”. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*. URL: http://arxiv.org/abs/0911.1583.

Bork, Lasse and Stig Moller (2016). “A New Index of Housing Sentiment”. In: *SSRN Electronic Journal* March 2020. DOI: 10.2139/ssrn.2867855.

Breusch, T. S. and A. R. Pagan (1979). “A Simple Test for Heteroscedasticity and Random Coefficient Variation”. In: *Econometrica* 47.5, p. 1287. DOI: 10.2307/1911963.

Case, Karl E., Robert J. Shiller, and Anne K. Thompson (2014). “What Have They Been Thinking? Homebuyer Behavior in Hot and Cold Markets A 2014 Update”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2580196.

Chamberlain, Gary (1984). “Panel data”. In: pp. 1247–1318. DOI: 10.1016/S1573-4412(84)02014-6.

Chew, Cynthia and Gunther Eysenbach (2010). “Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak”. In: *PLoS ONE* 5.11. Ed. by Margaret Sampson, e14118. DOI: 10.1371/journal.pone.0014118.

Chipperfield, James O. and David G. Steel (2012). “Multivariate random effect models with complete and incomplete data”. In: *Journal of Multivariate Analysis* 109, pp. 146–155. DOI: 10.1016/j.jmva.2012.02.014.

Chisholm, Elinor and Kimberley O’Sullivan (2017). “Using Twitter to Explore (un)Healthy Housing: Learning from the #Characterbuildings Campaign in New Zealand”. In: *International Journal of Environmental Research and Public Health* 14.11, p. 1424. DOI: 10.3390/ijerph14111424. URL: http://www.mdpi.com/1660-4601/14/11/1424.

Cody, Emily M. et al. (2015). “Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll”. In: *PLOS ONE* 10.8. Ed. by Sune Lehmann, e0136092. DOI: 10.1371/journal.pone.0136092.

Cody, Emily M. et al. (2016). “Public Opinion Polling with Twitter”. In: *Social and Information Networks*. URL: http://arxiv.org/abs/1608.02024.

Council of Mortgage Lenders (2017). *Mortgage lending statistics*. URL: https://www.cml.org.uk/industry-data/industry-data-tables/.

Council of Mortgage Lenders (2019). *Lending by UK Postcode Sector*. URL: https://www.ukfinance.org.uk/data-and-research/data/mortgages/mortgage-lending-within-uk-postcodes.

Croce, Roberto M. and Donald R. Haurin (2009). “Predicting turning points in the housing market”. In: *Journal of Housing Economics* 18.4, pp. 281–293. DOI: 10.1016/j.jhe.2009.09.001.

Dann, Stephen (2010). “Twitter content classification”. In: *First Monday*. DOI: 10.5210/fm.v15i12.2745.

Daoud, Jamal I. (2017). “Multicollinearity and Regression Analysis”. In: *Journal of Physics: Conference Series* 949, p. 012009. DOI: 10.1088/1742-6596/949/1/012009.

Department for Work and Pensions (2018). *Households below average income: An anlysis of the income distribution 1994/95 to 2015/16*. URL: https://www.gov.uk/government/statistics/households-below-average-income-199495-to-201516.

Diener, Ed (2009). “Assessing Subjective Well-Being: Progress and Opportunities”. In: pp. 25–65. DOI: 10.1007/978-90-481-2354-4{_}3.

Dua, Pami (2008). “Analysis of Consumers’ Perceptions of Buying Conditions for Houses”. In: *The Journal of Real Estate Finance and Economics* 37.4, pp. 335–350. DOI: 10.1007/s11146-007-9084-0.

Dunning, Thad (2008). “Model specification in instrumental-variables regression”. In: *Political Analysis* 16.3, pp. 290–302. DOI: 10.1093/pan/mpm039.

Fowler, Floyd (2009). *Survey Research Methods (4th ed.)* 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. DOI: 10.4135/9781452230184.

Fritsche, Immo, Eva Jonas, and Thomas Kessler (2011). “Collective Reactions to Threat: Implications for Intergroup Conflict and for Solving Societal Crises”. In: *Social Issues and Policy Review* 5.1, pp. 101–136. DOI: 10.1111/j.1751-2409.2011.01027.x.

Gallin, Joshua (2008). “The Long-Run Relationship Between House Prices and Rents”. In: *Real Estate Economics* 36.4, pp. 635–658. DOI: 10.1111/j.1540-6229.2008.00225.x.

Garavaglia, Susan and Asha Sharma (1998). “A Smart Guide to Dummy Variables”. In:

GLA (2017). *Housing in London 2017*. Tech. rep. URL: https://data.london.gov.uk/download/housing-london/27e10d40-bb04-4028-95a6-606bd13d7777/Housing-in-London-2017-report.pdf.

GLA (2019). *Housing in London 2019: The evidence base for the Mayor’s Housing Strategy*. Tech. rep. London: Greater London Authority.

GLA (2020a). *Economic Fairness*. URL: https://data.london.gov.uk/economic-fairness/living-standards/homelessness/.

GLA (2020b). *Family Resources Survey*. URL: https://www.gov.uk/government/collections/family-resources-survey--2.

GLA (2020c). *Statistical GIS Boundary Files for London*. URL: https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london.

Goodman, John (1994). “Using Attitude Data to Forecast Housing Activity”. In: *Journal of Real Estate Research* 9.4, pp. 445–453. URL: https://www-jstor-org.eur.idm.oclc.org/stable/44095512.

Google (2020). *Google: Place_ID*. URL: https://developers.google.com/places/place-id.

Graham, Mark, Scott A. Hale, and Devin Gaffney (2014). “Where in the World Are You? Geolocation and Language Identification in Twitter”. In: *Professional Geographer* 66.4, pp. 568–578. DOI: 10.1080/00330124.2014.907699.

Groff, James and Paul Weinberg (1999). *SQL - The Complete Reference*. Berkeley: Osborne/McGraw-Hill. URL: https://docs.microsoft.com/en-us/sql/odbc/reference/structured-query-language-sql.

Hair, Joseph et al. (1995). *Multivariate Data Analysis*. 7th ed. New York: Macmillan Publishing Company.

Hansen, Bruce E. (1995). "Time Series Analysis". In: *Econometric Theory* 11.3, pp. 625–630. DOI: 10.1017/S0266466600009440.

Hausman, J. A. (1978). "Specification Tests in Econometrics". In: *Econometrica* 46.6, p. 1251. DOI: 10.2307/1913827.

Hay, Colin (2013). "Treating the Symptom Not the Condition: Crisis Definition, Deficit Reduction and the Search for a New British Growth Model". In: *The British Journal of Politics and International Relations* 15.1, pp. 23–37. DOI: 10.1111/j.1467-856X.2012.00515.x.

HM Land Registry (2019). *UK House Price Index*. URL: https://www.gov.uk/government/publications/about-the-uk-house-price-index/about-the-uk-house-price-index.

Hsiao, Cheng (2007). "Panel data analysis—advantages and challenges". In: *TEST* 16.1, pp. 1–22. DOI: 10.1007/s11749-007-0046-x.

Hui, Eddie Chi-man and Ziyou Wang (2014). "Market sentiment in private housing market". In: *Habitat International* 44, pp. 375–385. DOI: 10.1016/j.habitatint.2014.08.001.

Hurlin, Christophe (2018). "Linear Panel Models and Heterogeneity". PhD thesis. University of Geneva. URL: https://www.univ-orleans.fr/deg/masters/ESA/CH/Geneve_Chapitre1.pdf.

Hutto, C J and Eric Gilbert (2015). "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Kang, Hyun (2013). "The prevention and handling of the missing data". In: *Korean Journal of Anesthesiology* 64.5, p. 402. DOI: 10.4097/kjae.2013.64.5.402.

Katchova, Ani (2013). "Panel Data Models". URL: https://sites.google.com/site/econometricsacademy/econometrics-models/panel-data-models.

Kim, Dong Sung and Jong Woo Kim (2014). "Public Opinion Mining on Social Media: A Case Study of Twitter Opinion on Nuclear Power". In: *Advanced Science and Technology Letters* 51, pp. 224–228. DOI: 10.14257/astl.2014.51.51.

Kounadi, Ourania et al. (2015). “Exploring Twitter to Analyze the Public’s Reaction Patterns to Recently Reported Homicides in London”. In: *PLOS ONE* 10.3. Ed. by Lidia Adriana Braunstein, e0121848. DOI: 10.1371/journal.pone.0121848.

Kumar, Shamanth, Fred Morstatter, and Huan Liu (2014). *Twitter Data Analytics*. SpringerBriefs in Computer Science. New York, NY: Springer New York. DOI: 10.1007/978-1-4614-9372-3.

Levin, Kate Ann (2006). “Study design III: Cross-sectional studies”. In: *Evidence-Based Dentistry* 7.1, pp. 24–25. DOI: 10.1038/sj.ebd.6400375.

Liu, Bing (2012). “Sentiment Analysis and Opinion Mining”. In: *Synthesis Lectures on Human Language Technologies* 5.1, pp. 1–167. DOI: 10.2200/S00416ED1V01Y201204HLT016.

Lovins, Julie Beth (1968). “Development of a stemming algorithm”. In: *Mech. Transl. Comput. Linguistics* 11, pp. 22–31.

Magdy, Walid, Kareem Darwish, and Norah Abokhodair (2015). “Quantifying Public Response towards Islam on Twitter after Paris Attacks”. In: URL: http://arxiv.org/abs/1512.04570.

Marcato, Gianluca and Anupam Nanda (2016). “Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market”. In: *Journal of Real Estate Research* 38.2, pp. 165–204. URL: https://ssrn-com.eur.idm.oclc.org/abstract=2487104.

Marsden, Joel (2015). “House prices in London – an economic analysis of London’s housing market”. London. URL: https://www.london.gov.uk/sites/default/files/house-prices-in-london.pdf.

Mason, Andrew (1996). “Population and housing”. In: *Population Research and Policy Review* 15, pp. 419–435. URL: https://link-springer-com.eur.idm.oclc.org/content/pdf/10.1007/BF00125863.pdf.

McCormick, Tyler H. et al. (2017). “Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing”. In: *Sociological Methods & Research* 46.3, pp. 390–421. DOI: 10.1177/0049124115605339.

Mela, Carl F. and Praveen K. Kopalle (2002). “The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations”. In: *Applied Economics* 34.6, pp. 667–677. DOI: 10.1080/00036840110058482.

Midolo, Emanuele (2019). *Billionaires’ Boulevard Is Still The Most Expensive Street In England, Despite Brexit*. URL: https://www.forbes.com/sites/

emanuelemidolo/2019/09/27/billionaires-boulevard-is-still-the-most-expensive-street-in-england--despite-brexit/.

Ministry of CLG (2019). *Total Number of Net Additional Dwellings per London Borough*. URL: https://www.gov.uk/government/statistical-data-sets/live-tables-on-net-supply-of-housing.

Ministry of CLG (2020). *Live tables on homelessness*. URL: https://www.gov.uk/government/statistical-data-sets/live-tables-on-homelessness.

Mulder, Clara (2008). "Housing and population: a two-sided relationship". In: *Sixty-ninth session of the UNECE Committee on Housing and Land Management*. Amsterdam. URL: https://pdfs.semanticscholar.org/897f/7c0f24e14995e7155973c3ba27279d38ecf9.pdf.

Nichols, Austin Lee and Jon K. Maner (2008). "The Good-Subject Effect: Investigating Participant Demand Characteristics". In: *The Journal of General Psychology* 135.2, pp. 151–166. DOI: 10.3200/GENP.135.2.151-166.

Nijskens, Rob et al. (2019). *Hot Property: the Housing Market in Major Cities*. Ed. by Rob Nijskens et al. Amsterdam: Springer International Publishing. DOI: 10.1007/978-3-030-11674-3.

O'Connor, Brendan et al. (2010). "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series". In: *International AAAI Conference on Weblogs and Social Media*. Vol. 11. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842.

Office of National Statistics (2017). *English Housing Survey 2015 to 2016: headline report*. Tech. rep. URL: https://www.gov.uk/government/statistics/english-housing-survey-2015-to-2016-headline-report.

Office of National Statistics (2019a). "Employee earnings in the UK: 2019". In: URL: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2019.

Office of National Statistics (2019b). *Families and households in the UK: 2019*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2019.

Office of National Statistics (2019c). *London ward-level population estimates*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/.

Office of National Statistics (2019d). *ONS Postcode Directory*. URL: http://geoportal.statistics.gov.uk/datasets/ons-postcode-directory-may-2019.

Öztürk, Nazan and Serkan Ayvaz (2018). “Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis”. In: *Telematics and Informatics* 35.1, pp. 136–147. DOI: 10.1016/j.tele.2017.10.006.

Page, Ben (2018). *Londoners in 2018*. Tech. rep., p. 10. URL: https://www.ipsos.com/ipsos-mori/en-uk/londoners-say-brexit-crime-and-housing-are-top-issues-facing-capital.

Park, Hun Myoung (2011). “Practical Guides To Panel Data Modeling: A Step by Step Analysis Using Stata”. URL: https://www.iuj.ac.jp/faculty/kucc625/method/panel/panel_iuj.pdf.

Pearson, Christine M. and Judith A. Clair (1998). “Reframing Crisis Management”. In: *Academy of Management Review* 23.1, pp. 59–76. DOI: 10.5465/amr.1998.192960.

Petkar, Arati, Dr Macwan, and Dhiraj Takkekar (2012). “Urbanization and its Impact on Housing”. In: *International Journal of Multidisciplinary Research* 1, pp. 116–121.

Pope, David and Josephine Griffith (2016). “An Analysis of Online Twitter Sentiment Surrounding the European Refugee Crisis”. In: *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. SCITEPRESS - Science and Technology Publications, pp. 299–306. DOI: 10.5220/0006051902990306.

Quiggly, John M (1999). “Real Estate Prices and Economic Cycles”. In: *International Real Estate Review* 2, pp. 1–20.

Ranco, Gabriele et al. (2015). “The effects of twitter sentiment on stock price returns”. In: *PLoS ONE* 10.9, pp. 1–21. DOI: 10.1371/journal.pone.0138441.

Rossall, Valentino (2015). *Solving the UK Housing Crisis*. Tech. rep. London: The Bow Group.

Ruths, Derek and Jürgen Pfeffer (2014). “Social media for large studies of behavior”. In: *Science* 346.6213, pp. 1063–1064. DOI: 10.1126/science.346.6213.1063.

Si, Jianfeng et al. (2013). “Exploiting topic based twitter sentiment for stock prediction”. In: *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* 2.2011, pp. 24–29.

Soo, Cindy K. (2013). “Quantifying Animal Spirits: News Media and Sentiment in the Housing Market”. In: *SSRN Electronic Journal* 1200. DOI: 10.2139/ssrn.2330392.

Thompson, Warren S. (1937). “Population Growth and Housing Demand”. In: *The ANNALS of the American Academy of Political and Social Science* 190.1, pp. 131–137. DOI: 10.1177/000271623719000115.

Trust for London (2018). *Housing, Trust for London*. URL: https://www.trustforlondon.org.uk/issues/housing/.

Trust for London (2020). *London rent as a percentage of gross pay*. URL: https://www.trustforlondon.org.uk/data/rent-affordability-borough/.

Tumasjan, Andranik et al. (2010). “Predicting elections with Twitter: What 140 characters reveal about political sentiment”. In: *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media* 10, pp. 178–185.

Twitter (2019). *Twitter Q1 2019 Earnings Report*. URL: https://investor.twitterinc.com/home/default.aspx.

UK Post Office (2020). *UK Postcode Finder*. URL: https://www.postoffice.co.uk/postcode-finder.

United Nations (1948). *Universal Declaration of Human Rights*. Paris. URL: https://www.refworld.org/docid/3ae6b3712c.html.

United Nations (2007). *World Urbanization Prospects*. Vol. 12, pp. 197–236. DOI: 10.4054/demres.2005.12.9.

United Nations (2018). “2018 Revision of World Urbanization Prospects”. In: *Department of Economic and Social Affairs*.

Wang, Zhangming, Chengzhang Wang, and Qian Zhang (2015). “Population Ageing, Urbanization and Housing Demand”. In: *Journal of Service Science and Management* 08.04, pp. 516–525. DOI: 10.4236/jssm.2015.84052.

Wang, Ziyou and Eddie Chi-man Hui (2017). “Fundamentals and Market Sentiment in Housing Market”. In: *Housing, Theory and Society* 34.1, pp. 57–78. DOI: 10.1080/14036096.2016.1196240.

Wilcox, James A. (2015). “The Home Purchase Sentiment Index: A New Housing Indicator”. In: *Business Economics* 50.4, pp. 178–190. DOI: 10.1057/be.2015.27.

Wooldridge, Jeffrey (2010). *Econometric Analysis of Cross Section and Panel Data*. 7th ed. Cambridge: MIT Press.

Zimmer, Michael and Nicholas John Proferes (2014). “A topology of Twitter research: disciplines, methods, and ethics”. In: *Aslib Journal of Information Management* 66.3. Ed. by Axel Bruns and Dr Dr Katrin Weller, pp. 250–261. DOI: 10.1108/AJIM-09-2013-0083.

8 Appendices

A London borough acronyms

Acronym	Borough	Acronym	Borough
Bar	Barking and Dagenham	**Hns**	Hounslow
Brn	Barnet	**Isl**	Islington
Bxl	Bexley	**Kns**	Kensington and Chelsea
Brt	Brent	**Kng**	Kingston upon Thames
Brm	Bromley	**Lam**	Lambeth
Cmd	Camden	**Lnd**	London as a whole
Cty	City of London	**Lsh**	Lewisham
Crd	Croydon	**Mrt**	Merton
Elg	Ealing	**Nwm**	Newham
Enf	Enfield	**Rdb**	Redbridge
Grn	Greenwich	**Rch**	Richmond upon Thames
Hck	Hackney	**Swr**	Southwark
Hms	Hammersmith and Fulham	**Stn**	Sutton
Hgy	Haringey	**Tow**	Tower Hamlets
Hrw	Harrow	**Wth**	Waltham Forest
Hvg	Havering	**Wns**	Wandsworth
Hdn	Hillingdon	**Wst**	Westminster

B Variable description and interpretation

Variable	Measured in	Example	Interpretation
Abnormal sentiment	Sentiment score	-0.105	The sentiment score in this tweet is -0.105 lower than the average sentiment score found in tweets sent from this borough, this year, on a scale from -1 to +1.
Proportion	Tweets per 100 000 tweets	54.32	A total of 54.32 housing tweets were sent per 100 000 tweets in this year, from this borough.
House price	House Price Index (HPI)	88.3	The house price in this borough in this year was 88.3% of the average house price in January 2015 in London.
Housing market size	Property sales transactions	4055	4055 property transactions were recorded in this year, in this borough.
Housing supply	Net additional dwellings	1247	1247 dwellings were added this year, to this borough.
Mortgages	Average total outstanding mortgage per postcode sector (in million £)	166	In this borough, on average, the inhabitants of a postcode sector had a combined mortgage of £166 million this year.
Income	Annual income (in thousand £)	34.06	The median annual income of citizens living in this borough is £34 060.
Homelessness	Number of homeless and in priority need per 1000 households	2.1	A total of 2.1 persons per 1000 households is accepted as homeless and in priority need in this borough, this year.

C Location algorithm results per year

	Tweets	(1)	(2)	(3)	(4)	(5)	(6)
2012	11 638 282	7 475 692	124 387	2 640 430	34 493	1 363 280	12%
2013	15 046 497	10 974 782	144 655	1 744 685	62 036	2 120 339	14%
2014	17 053 109	11 078 060	148 537	1 570 847	1 828 439	2 427 226	14%
2015	15 456 857	4 133 306	72 886	7 297 442	921 677	3 031 546	20%
2016	15 182 733	1 565 400	24 008	10 821 003	624 977	2 147 345	14%
2017	10 299 503	1 158 220	13 376	7 048 193	569 307	1 510 407	15%
2018	11 372 796	864 577	11 858	8 231 209	533 399	1 731 753	15%
Total	96 049 777	37 250 037	539 707	39 353 809	4 574 328	14 331 896	
	100%	39%	1%	41%	5%	15%	

Column (1): both methods were used and assigned the same borough
Column (2): both methods were used and assigned a different borough
Column (3): only place_id was used to assign borough
Column (4): only GPS coordinates were used to assign borough
Column (5): no borough was assigned
Column (6): percentage of tweets that were not assigned a borough

D Twitter data structure

Variable	Explanation
text	Text of the tweet
lang	Acronym of the tweet language
created_date	Date on which the tweet was sent
created_timestamp	Timestamp on which the tweet was sent
geo_lat	Coordinates from which the tweet was sent
geo_lon	Coordinates from which te htweet was sent
place_id	Unique place identifier of location attached to tweet

E Queries for classifying housing tweets

Inclusion Query	Exclusion Query
"hous*" AND "*pric*"	"prick" AND "parliament" AND "household" AND "stock"
" rent " AND "London"	
" rent " AND "*pric*"	"prick" AND "stock"
"apartment*" AND "*pric*"	"prick" AND "stock"
" hous*" AND "market"	"household" AND "stock"
" hous*" AND "supply"	"household" AND "parliament"
" hous*" AND "crisis*"	"household"
" hous" AND "cost"	"costume" AND "costa"
" hous*" AND "afford*"	"household"
"property" and "pric*"	"prick" AND "stock"
"property" and "market*"	"stock" AND "marketing"
"London hous*"	"household" AND "parliament"
"#londonhousing*"	

** indicates the word is lemmatized*

Exclusion Query	Explanation
"prick"	consists of the query "pric*"
"parliament"	often associated with "House" (of parliament)
"household"	consists of the query "hous"
"stock"	often co-occurred with stock "price" and "market"
"marketing"	consists of the query "market"
"music"	associated with "house music"
"costa"	consists of query "cost"

F Abnormal sentiment of housing tweets per borough per year

Abr.	Borough	2012	2013	2014	2015	2016	2017	2018	Avg.
Cty	City of London	-2.0	-12.6	-6.7	-14.5	-17.7	-12.2	-13.9	-11.4
Bar	Barking and Dagenham	10.3	10.6	-15.3	-9.8	0.7	-3.1	-24.9	-4.5
Brn	Barnet	-18.0	-3.4	-12.4	-9.2	-15.7	-19.3	-13.2	-13.0
Bxl	Bexley	-32.0	21.1	-13.7	-17.9	-26.8	-10.8	-31.0	-15.9
Brt	Brent	10.5	-6.7	-0.9	-1.5	-20.2	-27.2	-13.5	-8.5
Brm	Bromley	-16.4	-4.1	-9.7	-7.6	-14.6	-11.0	-16.5	-11.4
Cmd	Camden	-8.1	-7.9	-12.2	-21.5	-20.1	-13.3	-11.7	-13.5
Crd	Croydon	1.1	-4.6	0.1	-12.2	-9.4	-17.2	-8.3	-7.2
Elg	Ealing	7.0	-2.1	-13.4	-11.5	-14.1	-9.8	-10.8	-7.8
Enf	Enfield	-21.1	-13.4	-8.8	-6.7	-2.2	-13.4	-29.6	-13.6
Grn	Greenwich	-4.5	-13.6	-9.0	-9.2	-12.1	-9.7	-14.8	-10.4
Hck	Hackney	-2.2	-7.7	-10.1	-11.9	-17.5	-15.7	-20.5	-12.2
Hms	Hammersmith and Fulham	13.7	-7.9	-9.9	-12.5	-16.7	-17.7	-19.9	-10.1
Hgy	Haringey	-15.4	-8.8	-20.7	-15.6	-20.8	-23.9	-20.4	-18.0
Hrw	Harrow	-1.6	6.4	-17.6	-5.6	-10.3	-12.1	-17.1	-8.3
Hvg	Havering	5.2	-18.8	-6.9	-17.4	20.5	-4.4	-26.9	-7.0
Hdn	Hillingdon	6.6	1.2	-9.5	-10.1	-12.9	1.2	-21.3	-6.4
Hns	Hounslow	-7.4	5.5	-5.1	-6.6	-25.9	-18.4	-10.4	-9.7
Isl	Islington	4.6	-2.4	-10.0	-16.2	-20.9	-20.0	-14.7	-11.4
Kns	Kensington and Chelsea	-5.2	-0.9	-8.2	-11.2	-14.2	-8.2	-10.7	-8.4
Kng	Kingston upon Thames	5.4	2.8	-3.6	-13.0	-5.7	-29.8	-8.9	-7.6
Lam	Lambeth	-14.4	-0.6	-13.9	-19.4	-14.2	-19.3	-19.1	-14.4
Lsh	Lewisham	-6.5	3.4	-9.5	-21.1	-11.6	-22.6	-18.2	-12.3
Mrt	Merton	-2.4	21.7	-5.9	-14.8	-16.2	-15.3	-29.2	-8.9
Nwm	Newham	-9.3	-22.6	-4.8	-18.3	-16.2	-9.6	-2.4	-11.9
Rdb	Redbridge	-20.0	-1.5	-7.8	-24.1	-16.0	-6.4	-9.7	-12.2
Rch	Richmond upon Thames	-14.4	-6.4	9.9	-13.0	-24.8	-10.8	-19.5	-11.3
Swr	Southwark	-0.9	-3.4	-11.3	-11.4	-14.2	-20.5	-8.4	-10.0
Stn	Sutton	-14.4	4.4	0.6	-3.4	-18.1	-6.6	-3.4	-5.8
Tow	Tower Hamlets	-2.9	-0.1	-12.2	-18.3	-21.4	-17.8	-18.5	-13.1
Wth	Waltham Forest	17.8	-10.0	-9.8	-5.6	-12.1	-18.0	-15.9	-7.7
Wns	Wandsworth	1.0	-5.1	-13.7	-10.4	-13.0	-16.9	-11.0	-9.9
Wst	Westminster	-6.4	-8.4	-1.3	-17.3	-19.9	-18.0	-12.4	-12.0
Lnd	London	-4.6	-7.9	-9.4	-13.1	-12.7	-14.9	-16.7	-11.3

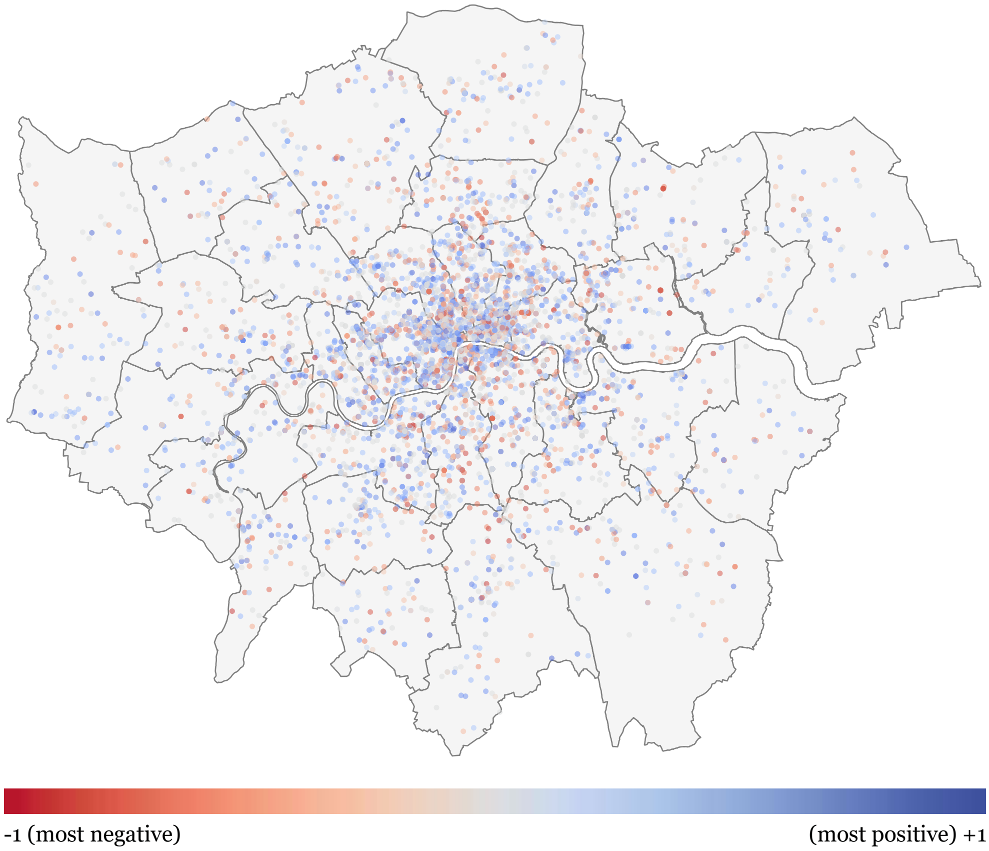
The numbers indicate how much the level of sentiment of housing tweets deviated from the average sentiment of all tweets from that borough in that year on a scale from -100 to +100.

G Proportion of housing tweets per borough per year

Abr.	Borough	2012	2013	2014	2015	2016	2017	2018	Avg.
Cty	City of London	8	24	44	73	73	51	42	56
Bar	Barking and Dagenham	7	3	9	8	19	24	22	10
Brn	Barnet	6	6	13	28	33	33	25	20
Bxl	Bexley	2	3	7	17	29	15	7	11
Brt	Brent	4	7	15	16	21	21	17	15
Brm	Bromley	4	8	11	29	26	24	19	17
Cmd	Camden	31	24	26	44	44	40	26	34
Crd	Croydon	5	13	11	18	31	34	16	17
Elg	Ealing	9	11	19	26	34	36	17	21
Enf	Enfield	3	5	9	18	16	16	12	11
Grn	Greenwich	6	6	13	20	32	24	16	17
Hck	Hackney	22	18	31	40	29	31	20	28
Hms	Hammersmith and Fulham	8	14	16	21	23	37	17	19
Hgy	Haringey	8	13	17	31	42	34	28	26
Hrw	Harrow	11	8	10	14	26	26	11	15
Hvg	Havering	2	5	5	8	9	72	0	5
Hdn	Hillingdon	4	5	7	16	13	16	11	10
Hns	Hounslow	5	6	18	16	25	32	19	17
Isl	Islington	8	15	20	39	27	28	31	25
Kns	Kensington and Chelsea	18	21	25	28	61	41	30	34
Kng	Kingston upon Thames	10	9	12	30	32	25	23	20
Lam	Lambeth	19	19	25	39	31	32	20	27
Lsh	Lewisham	8	12	15	36	24	20	22	20
Mrt	Merton	10	10	22	19	22	15	16	17
Nwm	Newham	6	13	18	38	26	23	22	19
Rdb	Redbridge	4	6	15	24	23	38	16	16
Rch	Richmond upon Thames	9	16	15	24	24	23	23	20
Swr	Southwark	16	17	32	36	43	33	23	30
Stn	Sutton	7	3	8	21	22	19	14	13
Tow	Tower Hamlets	14	17	22	36	36	35	29	28
Wth	Waltham Forest	5	6	15	17	21	20	20	15
Wns	Wandsworth	11	51	19	32	26	35	26	29
Wst	Westminster	16	24	31	31	23	19	18	23
	Untagged	4	7	11	30	30	32	24	20
Lnd	London	9	13	17	27	29	30	20	21

Measured in number of housing tweets per 100 000 tweets

H Spatial distribution of geotagged housing tweets in Greater London


-1 (most negative)
(most positive) +1