Department of Mathematics
Munster Technological University
DATA9003: Research Project

The Impact of Crime Hot-Spots on House Prices in New York City

Eoin Keohane

R00209304

Supervisor: Dr. David Goulding

Submission Date: 28 Aug, 2022

Declaration of Independent Work

I, Eoin Keohane (the author), hereby declare this work to be my own.

This report was written entirely by the author, except where otherwise stated. The source of any material not created by the author has been clearly referenced.

The author confirms that they have read and understood the universities policies/procedures concerning academic honesty, plagiarism and infringements. The author understands that where breaches of this declaration are detected, these will be reviewed under any and all relevant university regulations and that breaches may incur penalties.

The author acknowledges the right of the academic department to request them to present for oral examination as part of the assessment regime for this module. The author understands that assessment material may, at the discretion of the internal examiner, be submitted to the university's plagiarism detection solution.

_____ _____

Eoin Keohane Date

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, Dr. David Goulding, for their advice and guidance throughout this project.

I would also like to take this opportunity to thank all the staff of the department, and of the university as a whole, who, despite the unique challenges of the last academic year, have continued doing all they can to support students.

Last, but certainly not least, I would like to thank my family who are a constant source of support and inspiration.

Abstract

Previous research suggests that crime negatively impacts property prices overall, but that crime hot-spots have a more significant impact than specific crime incidents. Using data from the boroughs of New York City (NYC), this project aims to identify and map crime hot-spots and examine their impact on house prices within the boroughs of NYC. By examining the geographic concentration of arrests affected by the New York City Police Department (NYPD) between the years of 2006 and 2020 hot-spots were identified at different geographic scales. A linear regression model with a sliding time window was used to assess the impact of crime hot-spots on property prices in the boroughs of Brooklyn, Queens, The Bronx and Staten Island. Evidence was found to support previous findings: that crime hot-spots have a significant effect on house prices. However the magnitude of this effect is not constant and is affected by citywide crime trends as well as other external factors. This study was severely limited by data constraints that highlight the necessity for a diverse range of descriptive attributes in the development of hedonic regression models.

Chapter 1

Introduction

When buying a home, safety is an important consideration for buyers [1,2]. As a result properties in high crime areas are often significantly cheaper than those in other, safer areas [3]. Historically, when trying to quantify this effect researchers have considered crime as a local externality. Thus, assuming that only crimes occurring in the immediate vicinity of a property affect its value [1–3]. More recent research suggests that the proximity of a property to crime hot-spots may have a more significant impact than specific crime occurrences in the locality of the property [4]. The geographic concentration of crime at so called "hot-spots" has long been an active area of research in the field of spatial criminology [5]. However [4] was the first to examine the impact of these hot-spots on property prices; finding that they had significant impact, particularly on houses. This study aims to provide a more thorough understanding of how and why crime hot-spots impact house prices as well as how this impact changes over time. The

study was conducted using data from the city of New York, chosen for the wealth of data that is freely available through the city's Open Data portal[1].

1.1 New York City

Located on the east coast of the United States, at the mouth of the Hudson river New York City (NYC) is perhaps the most well known city in the modern world. Renowned as a centre for commerce, retail and tourism; today NYC is home to a diverse population of over eight million people. Covering just over 300 square miles the city is made up of five boroughs: Manhattan, Brooklyn, The Bronx, Queens and Staten Island (Figure **??**). The city is served by the New York City Subway, a rapid transit system that covers all boroughs except Staten Island. Staten Island has transport links to the the other boroughs via the Verrazzano-Narrows bridge and the free, Staten Island ferry. Staten Island is served by its own rapid transit system, the Staten Island Railway, which runs along the south side of the island.

The following description of the recent history of the housing market in NYC is based heavily on [6]. Throughout the late 90s and early 2000's the housing market in NYC, and the US as a whole, experienced an unprecedented boom. This housing bubble was predominantly driven by a

[1] `https://opendata.cityofnewyork.us/`

Figure 1.1: The five boroughs of New York City

growth in suburban home ownership. When the bubble burst in 2008 New York was quick to recover, particularly in the more urbanised boroughs of Manhattan and Brooklyn. Manhattan's resilience was due in no small part to the emergence of a market for high-end, luxury apartments that catered to the boroughs growing population of wealthy residents. Changes in US federal policy throughout the late 20-th century caused an upward transfer of wealth resulting in an upper-class with more spending power than ever before and an ever widening gap between the rich and poor. The buying up of properties in mid- and down-town Manhattan by the wealthy elite signalled the beginning of the decline in the growth of suburban home

ownership and a renewed interest in urban centres. The rapid rise of property prices in Manhattan resulted in middle class buyers being pushed out to other boroughs, particularly Brooklyn and Queens. This meant a huge influx of middle class residents into previously low-income neighbourhoods. This influx coupled with a lack of development of affordable housing led to increases in rents ad property prices. And so the working class residents of these neighbourhoods, many of whom had lost their jobs or were at risk of job loss due to the Great Recession, were forced to move to more affordable neighbourhoods. This led to a vicious cycle of gentrification where the working class residents being pushed out were replaced by middle class residents with higher incomes, which allowed for more price increases which in turn put more pressure on the remaining low income residents. This phenomenon started in the lower income neighbourhoods of Manhattan, such as Harlem, and quickly spread to neighbourhoods in Brooklyn, including Williamsburg and Greenpoint. Consequently the housing market in NYC quickly bounced back from the 2008 crash and property prices throughout the five boroughs have continued to climb in recent years.

Crime rates in NYC climbed rapidly through the late 80s, reaching an all time high in 1992 [6]. In the following decade crime rates began to fall. This summary of the decline and its apparent cause is based primarily on [7]. In the absence of any other obvious triggers the decline was largely attributed to the organisational and strategic changes implemented by the NYPD. Throughout the 1990s the size of the police force

grew significantly. This increase in manpower was accompanied by changes in policing strategy and departmental organisation. The introduction of *CompStat* meant a more data-led approach, using crime statistics to inform policing strategies and resource allocation. This led to the targetting of so called problem areas or hot-spots. This was accompanied by the adoption of aggressive street policing tactics such as stop and frisk. These tactics led to a dramatic increase in the number of arrests made for minor offences, particularly among people of colour[2]. It is difficult to say which of these changes, if any, truly account for the dramatic fall of crime rates in NYC throughout the 90s and early 2000s. However crime rates did fall and continued to do so throughout the 2010s while the NYPD made efforts to distance themselves from the aggressive tactics used in the 90s and early 2000s in favour of less controversial methods [9, 10].

1.2 Aims of the Study

This study has two principle objectives:

The first is an examination of crime concentration within the five boroughs of NYC. Crime concentration can be observed at different geographic scales [5, 11]. For the purposes of this study particular attention is given to the concentration of crimes at street segments and across census tracts. By

[2]In 2013 the New York District Court found the NYPD's stop and frisk policy to be unconstitutional [8].

comparing the patterns of crime concentration at street segments in NYC with those observed in other urban centres it is hoped to provide further evidence for the existence of Wiesburd's *law of crime concentration* [11].

The second is an investigation into the impact of those census tracts identified as crime hot-spots on property prices in NYC. It has been shown that crime hotspots have a significant impact on the value of single family homes [4]. By examining how the magnitude of this impact changes over time it is hoped that some important questions can be answered: Is the impact of crime hot-spots on property prices constant over time? If not then what are the factors that cause this impact to change?

The remaining chapters of this report are as follows: Chapter 2 provides on overview of existing literature related to the research project, including previous works on the concentration of crime at place and the impact of crime on property prices. Chapter 3 gives a description of the data used to conduct the study and of the techniques used in the analysis of this data. Chapter 4 presents and interprets the empirical results obtained from this analysis and Chapter 5 summarises the main findings of the study and the conclusions drawn from them. Chapter 5 also includes some suggestions of possible avenues for future research as well as an acknowledgement of the limitations of this study.

Chapter 2

Literature Review

2.1 Crime Hot-Spots

Criminological research has often favoured the person as the principal unit of analysis in an effort to identify the behavioural and psychological factors that cause individuals to commit crimes [11]. Despite this the field of *spatial criminology* has remained an active area of research for almost 200 years [5]. Pioneered in France, in the early 19-th century by Adolphe Quetelet and André-Michel Guerry, [5] spatial criminology is a sub-discipline of criminology that investigates "explicitly spatial processes and relationships" [12].

Early works of spatial criminology concentrated on the comparison of macro-geographic areas, comparing crime across different districts, counties or towns. Though differences between the various regions could be ob-

served these large units of analysis were not entirely appropriate as there was still significant variations in crime within each unit [5]. Throughout the 20-th century there was a shift away from the use of macro-geographic units towards smaller meso-geographic units (i.e. wards, neighbourhoods and census tracts) within specific urban centres [5]. However there still remained some variation within these smaller units and so, more recently, there has been an increase in the number of researchers investigating crime at micro-geographic units, such as street segments and intersections [11]. While crime patterns may appear broadly similar at different scales the use of micro-geographic units highlights the variation that may be masked by larger units of analysis and reveals a more stable picture of crime concentration [13].

In their 2015 paper, Weisburd put forward the idea of the *law of crime concentration* [11], which states that:

"For a defined measure of crime at a specific micro-geographic unit, the concentration of crime will fall within a narrow bandwidth of percentages for a defined cumulative proportion of crime"

To support this proposition [11] collates data from a number of previous studies, some of which are discussed here. This data is used to identify micro-geographic hot-spots in Cincinnati, Seattle, Tel-Aviv, New York, Sacramento, Brooklyn Park, Redlands and Ventura.

As part of their 1989 study Sherman et al. examined the sources of police calls in Minneapolis over a one year period [14]. They found that 50.4% of calls to the police were generated from 3.3% of street segments/intersections. In 2014, Weisburd and Amram examined the locations of reported crime incidents in Tel-Aviv in the year 2010 and were able to show that 4.5% of street segments produced approximately 50% of the reported crime incidents. 0.9% of street segments accounted for 25% of all reported incidents [15]. Boivin and de Melo examined crime data from two major Canadian cities, Toronto and Montreal, from 2016, using intersections as their principal unit of analysis. They specifically looked at occurrences of burglaries, robberies and car theft in both cities. They found that 4.5% of intersections in Montreal produced 50% of the crimes in question. 3.7% of the intersections in Toronto accounted for the same fraction of crimes [16]. Jaitman and Nicolas investigated crime concentration at street segments in 6 Latin American cities (Belo Horizonte, Bogota, Montevideo, Sucre and Zapopan) and found that 50% of crime in these cities was produced by between 3% and 7.5% of street segments [17].

Studies have also indicated the presence of micro-geographic crime hotspots in suburban settings. Over a 15 year period, half of the reported crime incidents in Brooklyn Park were produced by approximately 2% of street segments, with 25% of reported incidents being produced by $< 1\%$ of street segments [18].

Studies have shown that these micro-geographic hotspots are stable over time. Weisburd examined the distribution of crime across street segments in Seattle over a 14 year period and found the crime concentrations to be generally stable [19]. 84% of the street segments examined were described as having "stable trajectories". The street segments that displayed decreasing trends were still among the most active crime hot-spots at the end of the 14 year period. While those with increasing crime trends already had relatively high crime levels at the beginning of the time period.

Braga et al. examined the distribution of gun violence across street segments in Boston over the 29 year period from 1980 to 2008 and found that offences were highly concentrated, with 4.8% of street segments producing 73.9% of serious gun attacks [20]. Segments were classified as "stable" and "volatile" based on the trends observed in their gun violence levels over time. Volatile segments accounted for only 3% of all street segments and produced 52.5% of serious firearm offences over the observed time period. These volatile hot-spots appear to have driven city wide gun violence trends over the course of the 29 year period under examination [20].

It is clear that the concentration of crime at micro-geographic hot-spots is a global phenomenon that can be observed in both urban and suburban settings. The stability of these micro-geographic concentrations over time goes to show that they are true hot-spots and that researchers are correct to describe crime hot-spots in terms of micro-geographic units, as the use

of larger macro- and meso-geographic units, while useful in certain contexts, can mask a significant amount of variation in crime concentrations. The existence of high crime areas does not preclude the possibility that a fraction, or even a majority, of the street segments/intersections within these areas may produce little to no crime.

Collectively, these works provide strong evidence for the existence of the so called *"law of crime concentration"* as defined by Weisburd [11].

2.2 Modelling House Prices

2.2.1 Crime and Property Prices

Hedonic regression models are often used to model property prices, predominantly in efforts to develop pricing indices or to quantify the impact of some characteristic on property prices.

Formalised by Rosen in 1974 , hedonic regression models assume that, for products of a given class, there exists a set of "hedonic prices" associated with the various characteristics/attributes of the product that determine the overall value of said product and estimates these prices using regression [21]. If one assumes the functional form of a hedonic model to be linear then it can be written as equation 2.1.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \epsilon \tag{2.1}$$

Where y is the price of the product, β_i are the regression coefficients that represent the hedonic prices of the corresponding attributes x_i and ϵ is some random error. Alternatively the hedonic regression model can be expressed using matrix notation (equation 2.2).

$$y = \beta X + \epsilon \tag{2.2}$$

Where y now represents a vector of observations of the sale price for various properties, β is the vector of regression coefficients/hedonic prices, X is the matrix containing observations of the various attributes for each of the observed properties and ϵ is a vector of random error. Note that in equation 2.2 the coefficient β_0, representing the y-axis intercept of the linear model (equation 2.1), has been absorbed into the vector β. This necessitates the addition of a column of 1s to the matrix X.

It is not feasible, and in most cases not possible, to identify/include every relevant attribute in a hedonic regression model. When applied to property prices there is no strong consensus as to the "best" attributes to include, however the most commonly used attributes fall into three broad categories:

- *Property Specific Attributes*
- *Neighbourhood Specific Attributes*
- *Spatial Attributes*

Property specific attributes are internal factors describing the structure of the property in question. Often used characteristics include the square-footage of the structure [3, 4, 22], the square-footage of any attached land[1] [3, 4], the number of rooms [1, 3, 22–24] and the age or year of construction of the property [4, 22, 25].

Neighbourhood specific attributes are external factors describing the local area in which the property is located, usually through the use of various socio-economic factors such as median household income [3, 23, 26, 27] and crime rate [3, 26].

It has been shown that higher crime rates are associated with lower property prices [1–3]. Lynch and Rasmussen examined the impact of crime on house prices using data from the sale of 2800 homes in Jacksonville, Florida and found that homes in high crime areas were significantly cheaper when compared with similar properties in areas with less crime [3]. Gibbons examined the impact of crime on property prices in London, England, with a specific focus on the impact of offences that result in damage to property[2] [2]. They hypothesised that, if incidents of criminal damage im-

[1] including gardens, driveways etc.
[2] i.e. burglary, vandalism

pacted property prices than it was not because of the costs incurred by the incidents themselves but rather because they resulted in an increased fear of victimisation [2].

Spatial attributes are explicitly spatial factors describing the location of the property in relation to various (dis)amenities. Some of the amenities considered in previous works include schools [27–29] and transport hubs[3] [27, 29]. In urban settings the distance to the central business district is sometimes considered as a relevant factor [4, 27].

Ceccato and Wilhelmsson investigated the impact of crime on house prices in Stockholm, Sweden over a single year (2013) [4]. Meso-geographic crime hot-spots were identified within the city and a new spatial attribute, measuring the distance from a given property to the centre of the nearest hot-spot, was introduced. When compared with the neighbourhood crime-rates it was found that the distance to the nearest crime hot-spot had a much greater impact on property prices than the local crime rate [4]. Ceccato and Wilhelmsson estimated that if one could move a property $1km$ further from a crime hot-spot the value of the property would increase by almost 3000 euros [4].

[4] does not discuss in detail the decision to use meso-geographic hot-spots over micro-geographic hot-spots[4]. However if the hypothesis proposed by

[3] i.e. train/bus stations

[4] This decision was likely made due to data constraints as it is mentioned that the *basområde*, the unit used, is the smallest area for which statistics are reported in Sweden [4]

Gibbons: that it is the fear of victimisation rather than the specific incidents that cause crime to impact property prices [2], is accepted than the use of meso-geographic hotspots is easily justifiable. As this fear can be assumed to pervade the wider meso-geographic area rather than be contained on the micro-geographic scale.

2.2.2 Time-Varying Regression Models

The standard hedonic model discussed in the previous section estimates constant parameters and so cannot account for changes in the value of attributes over time. Consequently it is best applied to relatively short time periods, no longer than twelve months. There exists a number of variations on the hedonic regression model that allow it to be applied to longer time periods. Three such methods are discussed here: dummy variables, temporal windows and state space modelling.

Dummy Variables

The dummy variable approach accounts for temporal changes in the value of a product over a number of discrete time periods with the introduction of a dummy variable for each period. Using notation from [30], which uses a semi-log linear functional form, this version of the hedonic regression model can be written as:

$$ln(p_i^t) = \beta_0 + \sum_{k=1}^{K} \beta_k z_{i,k}^t + \sum_{\tau=2}^{T} \delta_\tau d_{i,\tau}^t \tag{2.3}$$

The dummy variable $d_{i,\tau}^t$ takes on the value 1 when $t = \tau$ and 0 otherwise. Note that the attribute coefficients β_k remain time-invariant. So, while this approach may be suitable for the development of house price indices it is less suitable for estimating the impact of specific attributes.

Temporal Windows

The sliding-time-window approach subsets the data into a number of time periods or *windows* and fits a standard hedonic regression model to each subset. The subset may be fully discrete or overlapping depending on the choice of window width and step-size. The composite model can be written as:

$$y_i^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t x_{i,k}^t + \epsilon_i^t \tag{2.4}$$

Where the temporal subsets are indexed by $t \in \{0, 1, 2, ...T\}$. By comparing the estimates of the attribute coefficients β_k^t across different windows it is possible to see a picture of how the cost associated with a specific attribute changes over time.

State Space Models

State space models are a class of models that attempt to estimate the dynamic properties that govern the state of a system when these properties cannot be observed directly [31]. In the context of hedonic regression the unobserved state variables are the regression coefficients/attribute prices.

Rather than re-estimating the coefficients for each period the state space modelling approach allows for the coefficients to be redefined as stochastic, time-varying parameters. These stochastic parameters are often treated as random walks and estimated using Kalman Filter techniques. [32–34].

Chapter 3

Methodology

3.1 The Data

3.1.1 Crime

Crime hot-spots were identified based on the geographic concentration of arrests made by the NYPD. Arrests data from the NYPD for the years 2006 through 2021 is openly available through the NYC Open Data portal. Missing data in the property sales dataset (see section 3.1.2) meant that the year 2021 had to be excluded from further analysis consequently only arrests in the years 2006 through 2020 were considered.

The dataset includes details of the date and location at which the arrests were made, the offence committed and the demographic information of the arrestee. For the purposes of this research project only a subset of

the available variables was used, table 3.1 lists the variables used for the identification of crime hot-spots. The inclusion of variables detailing the time and place the arrests were made is self-explanatory. The law code is included as it allows arrests to be categorised by offence type through article numbers in New York State Penal Law (NYSPL). The dataset includes a negligible amount of missing data and very few duplicates. Duplicates and observations with missing data were removed. Note that while the dataset includes arrests for violations of a number of law types including NYSPL, vehicle traffic law and other local laws only arrests for violations of NYSPL were included for analysis.

Name	**Description**
arrest_key	Unique identifier for each arrest
arrest_date	The date on which the arrest was made
law_code	Law code for the offence committed
arrest_boro	The borough in which the arrest was made
latitude	latitude coordinate for the arrest
longitude	longitude coordinate for the arrest

Table 3.1: The variables taken from the arrests data set

The latitude and longitude coordinates match arrests to the nearest street section/intersection. Arrests that could not be accurately geo-located are listed as having been made at the NYPD station house of the relevant precinct. In order to remove these arrests a list of addresses for these station houses was scraped from the NYPD website[1]. This list was supplemented with a number of substations and correctional facilities. The list of addresses was geo-coded using *ArcGIS*[2] to obtain latitude and longitude coordinates that were then used to remove arrests listed as having

[1] `https://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page`
[2] See the code in appendix B

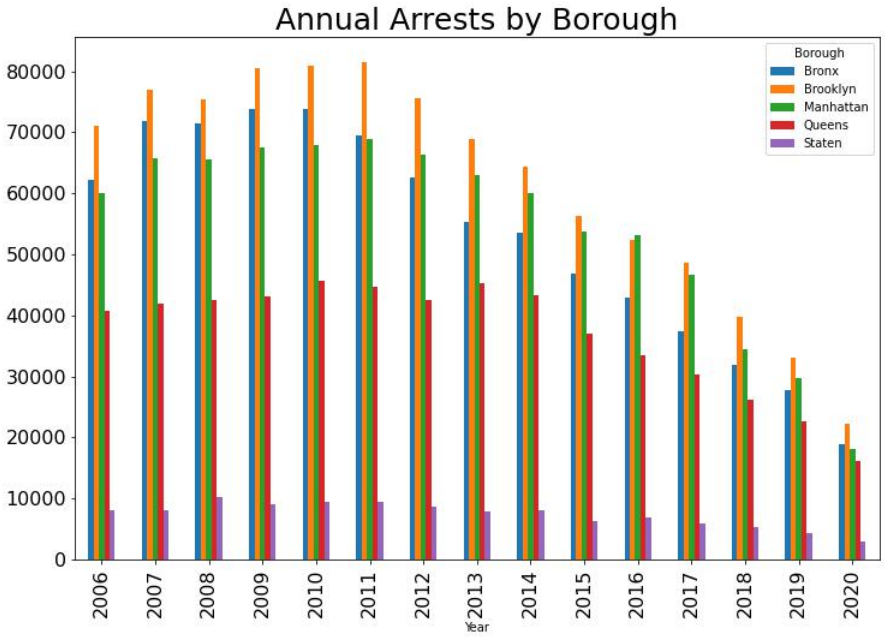


Figure 3.1: Annual arrests made by the NYPD by borough

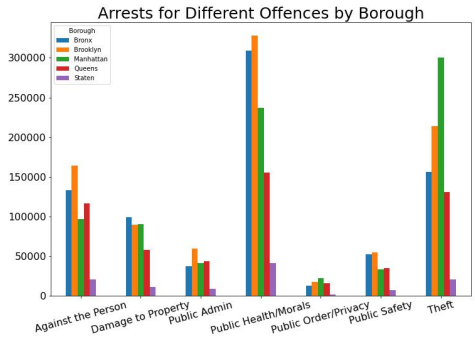
been made at these addresses from the dataset.

Arrests made for various white collar crimes, such as fraud and money laundering, were excluded from further analysis. Arrests for rarely occurring offences, such as terrorism, are likewise excluded.

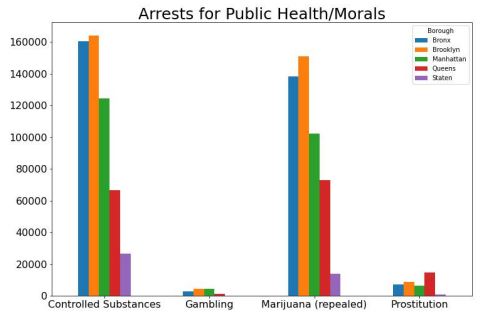
Looking at the number of arrests made annually by the NYPD in each borough (figure 3.1) shows a slight increase in the number of annual arrests between the years 2006 and 2010. Since 2010 the number of annual arrests has declined steadily

Violations of public health laws are the most common offences in every borough with the exception of Manhattan, where theft is most common (figure 3.2a). Almost all public health offences are drug related (figure 3.2b). Public health crimes, theft, violent crimes and property damage

are the four most common types of offence in every borough (figure 3.2a).



(a) Most arrests are for public health offences



(b) Almost all public health crimes are drug related

Figure 3.2: Drug related offences account for the majority of arrests

3.1.2 Property Sales

Annualised property sales data is published by the New York City Department of Finance (DoF) and is available on their website[3] for the years 2003 through 2021. Missing values within the dataset are heavily concentrated in the year 2021 and so the decision was made to exclude this year from further analysis. Consequently only sales data from the years 2006 through 2020 is considered.

The dataset contains information regarding the address, building class and square footage of each property as well as the date and price of the sale. The absence of coordinates meant that the addresses had to be geocoded. This was done through *ArcGIS* however due to the strict rate and usage limits the size of the dataset had to be reduced significantly to allow for

[3] `https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page`

geocoding to be completed within an appropriate timeframe. This was accomplished with the introduction of the following restrictions:

- Only consider sales that involved a single residential property
- Only consider properties of building class A1 and A5. The most common classes for one-family homes[4]
- Only consider properties whose sale price was between $150000 and $1000000

These restrictions meant that there was very little viable sales data left for the borough of Manhattan. Consequently the decision was made to exclude all property sales in the borough of Manhattan from further analysis.

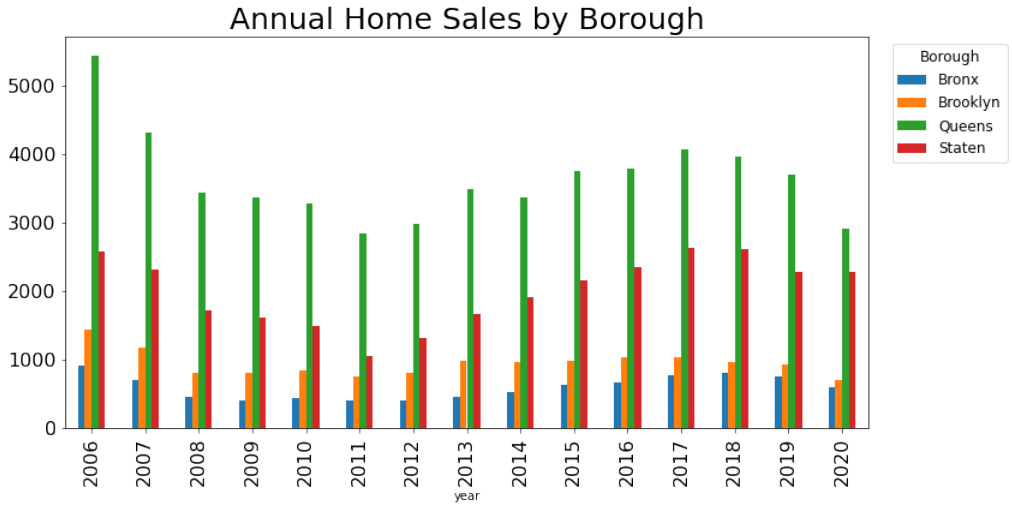


Figure 3.3: Annual house sales for each borough

Figure 3.3 shows that Queens was the borough with the most house sales for every year in the analysis, followed by Staten Island, Brooklyn and The

[4]A dictionary of NYC building classes is available at: `https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html`

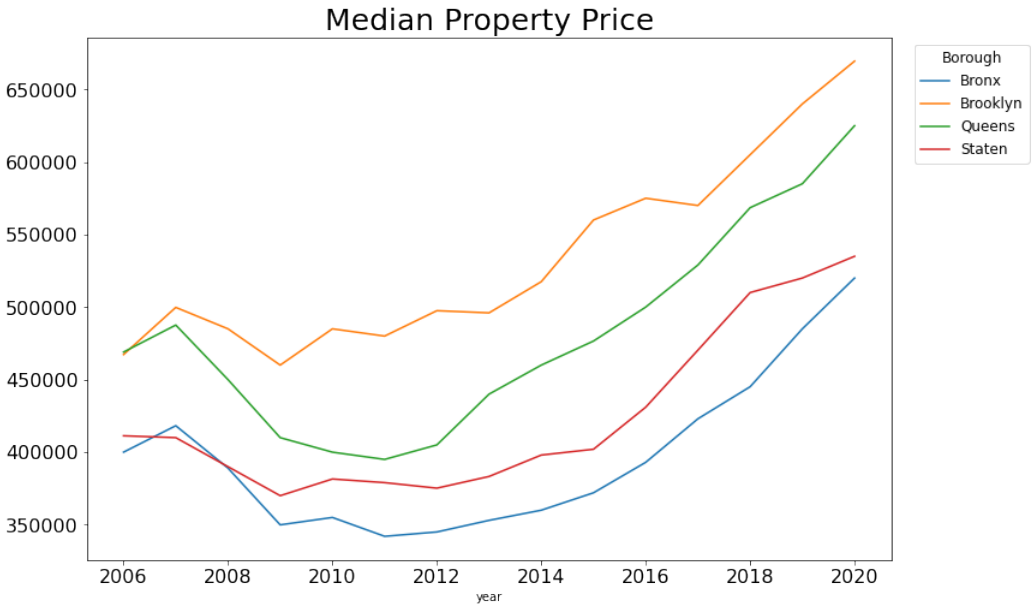


Figure 3.4: Median house prices over time for each borough

Bronx. The median house prices in all boroughs follows a similar trend. Figure 3.4 shows the drop in median price that occurred during the Great Recession of 2008/09 [35] followed by a steady increase in price over the course of the following decade.

The property sales data provides a limited number of property specific attributes, only including details for the age and square footage of each property. Efforts to find supplementary datasets that would allow for the introduction of neighbourhood specific attributes failed to produce any results. A dataset that provided suitable information for the entire 15 year period under inspection could not be found.

A number of spatial attributes were introduced in the hopes that they would prove relevant in determining the properties price. This required a number of supplementary datasets most of which were found via the NYC

Open Data portal. For each property the distance to the nearest school, park, university and transit station was calculated. As well as the distance to the centre of the nearest meso-geographic crime hot-spot and the distance to the shoreline. The variables that made up the supplemented sales dataset are described in table 3.2.

Name	Description	Unit	Mean	Std Dev
borough	The borough in which the property is located	-	-	-
zipcode	The zipcode in which the property is located	-	-	-
gross_sqft	The square footage any structures on the property, measured from the exterior walls	ft^2	1568.47	522.34
land_sqft	The square footage of any attached land including driveways, gardens etc.	ft^2	2843.99	1640.35
building_cls	The building class of the property	-	-	-
latitude	latitude coordinate of the property	deg	-	-
longitude	longitude coordinate of the property	deg	-	-
age	the age of the property at the time of sale	years	66.75	29.70
dist2school	Distance to the nearest school	km	0.47	0.27
dist2park	Distance to the nearest park	km	0.94	0.61
dist2sbwy	Distance to the nearest subway station	km	2.10	1.68
dist2uni	Distance to the nearest college/university	km	3.57	2.30
dist2shore	Distance from the shoreline	ft	7367.15	5482.36
dist2crime	Distance to centre of the nearest meso-geographic crime hot-spot	km	3.90	2.17
sale_date	The date of the sale	-	-	-
sale_price	The price of the sale	$ (USD)	490339.39	184019.49

Table 3.2: The supplemented property sales dataset

3.2 Identification & Mapping of Crime Hotspots

Arrests data is reported at the micro-geographic scale. Once the arrests dataset was appropriately filtered, as per the process described in section 3.1.1, it was a trivial task to determine the fraction of arrests produced

by each street segment both annually and over the course of the entire 15 year period of observation. These figures were then used to calculate the concentration of arrests for various cumulative proportions of crime. Micro-geographic hotspots were identified as those contributing to these cumulative proportions.

Before meso-geographic hot-spots could be identified a suitable meso-geographic unit had to be chosen. Zipcodes were briefly considered but ultimately it was decided to use census tracts as, prior to the recent spike in the use of micro-places such as street segments, these were the standard unit of analysis in spatial criminology [5]. GIS data describing the census tracts in NYC was gathered and used to determine the number of arrests made in each census tract. Choropleth maps were used to visualise the distribution of arrests across census tracts in each borough.

3.3 Hedonic Regression for Property Prices

It was decided that a linear regression model with a sliding-time-window would be used to estimate the impact of crime hot-spots on house prices. A rolling time window was applied rather than fully discrete windows in order to obtain a smoother picture of the coefficients evolution over time. Henceforth this model is referred to as the rolling regression model (RR model). The RR model uses a standard ordinary least squares regressor

to estimate model coefficients for each quarter based on sales data from the previous year, from the Q4-2006 through to Q4-2020.

In keeping with the method applied in [4] the distance to the crime hot-spot was measured as the distance to the centre of the nearest meso-geographic hot-spot rather than the nearest micro-geographic hot-spot. In doing so the model accepts that it is the fear of victimisation rather than specific incidents of crime that cause hot-spots to effect house prices. It was decided arbitrarily that only the five most concentrated census tracts within each borough would be considered hot-spots.

The data from each quarter was randomly divided into training (80%) and test (20%) sets. The RR model is fitted to the training set, while the remaining data is used as a hold out test set for model evaluation. At each time step the coefficient estimates are tested for statistical significance at the 5% significance level. The overall fit of the model is evaluated using the coefficient of determination (R^2) and mean absolute percentage error (MAPE). Various iterations of the model were developed using different combinations of the spatial variables with which the house sales data was supplemented. These variables were retained/omitted based on their effect on the adjusted coefficient of determination ($\tilde{R}^2$). In order to account for the changes in property prices across boroughs One-Hot encoding was used to introduce a set of dummy variables, one for each borough.

Chapter 4

Results

4.1 Crime Hot-Spots in New York City

Throughout the observation period arrests affected by the NYPD are highly concentrated geographically.

Table 4.1 shows the concentration of arrests at street segments for various thresholds. Half of the arrests made across all five boroughs over the 15 year period were produced by 4.07% of street segments with $< 1\%$ of street segments accounting for $> 25\%$ of arrests. These percentages are within the bandwidths observed by Weisburd et al when they proposed the law of crime concentration at place [11].

Significant concentration is likewise observed at the meso-geographic level. Figure 4.1 presents a choropleth map showing the distribution of arrests

Fraction of Arrests Produced (%)	Fraction of Street Segments (%)
10	0.11
15	0.25
25	0.75
50	4.07
90	32.88

Table 4.1: The concentration of arrests at street segments

across census tracts in the Bronx over the entire 15 year period of observation. The areas with the highest concentration of crime concentration are within the neighbourhoods of Mott Haven and Fordham. Similar patterns of crime concentration can be observed in the other boroughs, see the maps in appendix A. Table 4.2 lists the neighbourhoods in each borough that correspond roughly to the areas with the highest crime concentrations, as measured by the concentration of arrests effected by the NYPD.

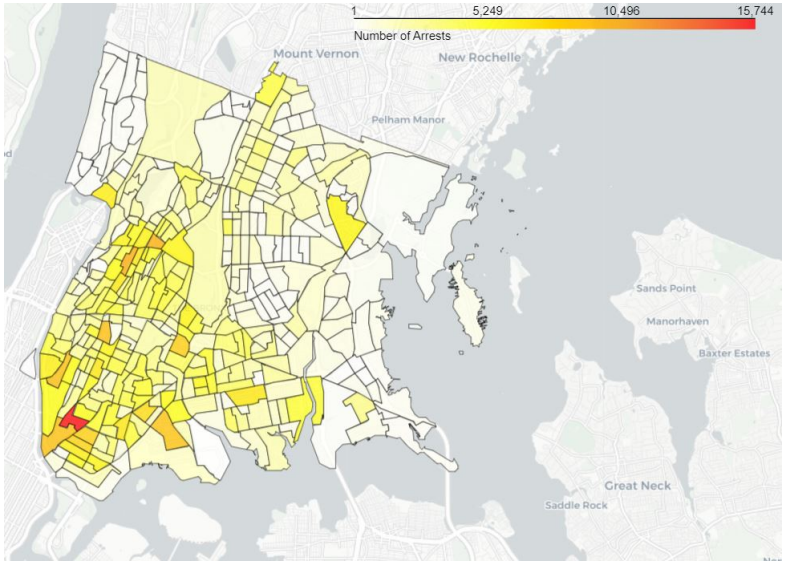


Figure 4.1: The distribution of arrests by census tract in the Bronx between 2006 and 2020

Manhattan	Brooklyn	Bronx	Queens	Staten
Garment District	Brownsville	Mott Haven	Jamaica	Clifton
East Harlem	Bedford-Stuyvesant	Fordham	Corona	Port Richmond

Table 4.2: High crime neighbourhoods

4.2 The Impact of Crime Hot-Spots on House Prices

Results for the investigation into the impact of crime hot-spots on house prices are presented two parts. Firstly, the overall goodness of fit of the model is discussed, alongside a brief description of the time evolution of the model coefficients. This discussion is followed by a more detailed look at the models estimate for the impact of crime hot-spots on house prices.

Experiments with the use of different combinations of supplementary spatial variables during model fitting produced little to no improvement in $\tilde{R}^2$. Consequently it was decided that all of the supplementary spatial variables would be omitted from the final model, with the exception of the distance to the nearest transit station which was retained.

Figure 4.2 shows the plots of the time evolution of all the coefficients in the final model. Points have been colour coded based on their statistical significance. Green points are coefficient estimates that are statistically significant while red points represents coefficient estimates that are not statistically significant. All tests for statistical significance were performed at a 5% significance level. The base price of properties in NYC has been rising steadily since the crash of 2008/09. Of the boroughs under consideration[1], Brooklyn is the most expensive borough in which to purchase property. Properties on Staten Island are generally cheaper than those in other boroughs. Note that a number of the estimates for the coefficients

[1] recall that Manhattan has been excluded from this analysis

Figure 4.2: Time evolution of model coefficients. Points in green are statistically significant at 5%, points in red are not.

relating to the boroughs of Queens and Staten fail to pass the hypothesis test for statistical significance. There appears to be a pattern in that it is the estimates that are closest to 0 that are not deemed to be statistically significant. It can be assumed that the borough in which a property is located will have a significant impact on its price, *ceteris paribus*. This would explain why estimates close to 0 for the Queens and Staten coefficients are not deemed statistically significant. There appears to be a slight increase over time in the value of a internal square footage, that is the square footage of the house itself (gross_sqft). While external square footage, that is the square footage of any attached lands such as gardens or driveways (land_sqft) maintains a relatively stable price of between $20 USD and $30 USD per square foot. Age has a negative effect on property price. Again coefficient estimates for the age attribute that are close to 0 are deemed not to be statistically significant while estimates of a greater magnitude are and so it is concluded that age truly does effect property prices. Properties further from subway stations are worth less, with the magnitude of this effect increasing over time. This could be indicative of a greater reliance on public transportation within the populace of NYC. With the value of the amenity increasing as its use becomes more important.

When applied to the training set the model achieves a coefficient of determination of $R^2 = 0.33$ and a MAPE of 27.5%. Performance on the test set is similar with $R^2 = 0.35$ and MAPE of 27%. So, the model only

accounts for approximately one third of the variation in property prices. Why does the model fail to account for for so much of the variation in house prices? It is thought that this failure is a result of the omission of some relevant variable(s). As discussed in chapter 2 most hedonic regression models applied to house prices use some combination of property specific, neighbourhood specific and spatial attributes. The property sales dataset does not include many property specific attributes. Attributes that have been shown to be relevant in previous works, such as the number of rooms in the house [24], are not available. No neighbourhood specific attributes are included in the model as efforts to source suitable data to develop such attributes for the entire 15 year observation period were fruitless. A number of spatial attributes were introduced but all failed to produce a significant improvement in model performance. It seems clear that some relevant variables have not been included in the model and that this has prevented the model from capturing a significant fraction of the variation in house prices.

Taking a closer look at the models estimate for the impact of crime hotspots on house prices (figure 4.3) one can see that these results agree with previous findings that crime hotspots have a significant impact on house prices [4], with houses costing more the further they are from crime hotspots. The magnitude of this impact increased sharply in the years 2006 through 2011. At the beginning of this period a property that was $1km$ further from a crime hotspot was estimated to cost between $2000 and

$4000 USD more than an identical property closer to the crime hot-spot. In 2011 this difference was estimated at close to $9000 USD. In subsequent years the model shows a slow but steady decline in the impact of crime hot-spots on house prices.

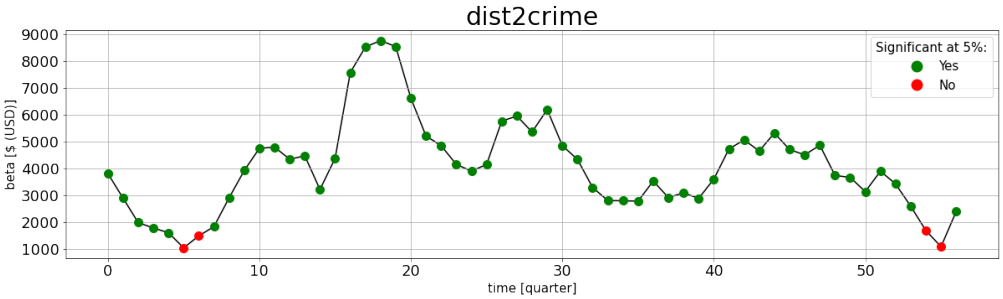


Figure 4.3: The impact of crime hot-spots on house prices

Recall from section 3.1.1 that the number of arrests effected annually by the NYPD increased slightly between the years 2006 and 2010 before steadily decreasing throughout the following decade. So, there are some qualitative similarities in the the number of annual arrests and the magnitude of the impact of crime hot-spots on house prices. If the number of annual arrests is assumed to be indicative of the citywide crime levels then this implies that the impact of crime hot-spots is lessened with citywide crime levels. However it has been shown that citywide crime trends are driven by the trends at hot-spots [20]. So, it is not only the concentration of crime but also the overall crime rate at a hotspot that determines its impact on house prices.

Chapter 5

Conclusions

5.1 Main Findings

When Weisburd proposed the law of crime concentration they estimated the bandwidth for the concentration of 50% of crime in both small and big cities to be approximately 4%, between 2.1% and 6%. Likewise the concentration of 25% of crime was estimated to be between 0.4% and 1.6% [11]. The investigation into the concentration of crime in NYC found that between the years of 2006 and 2020 half of all arrests were produced by 4.07% of street segments, with 0.75% of street segments accounting for 25% of arrests. These concentrations fall cleanly within the bandwidths defined by Weisburd and provide further support for the existence of the law of crime concentration at place.

Evidence was found to support the findings of Ceccato and Wilhelmsson

[4]: that crime hot-spots have a significant effect on house prices. This effect is not constant but may change over time as hot-spots become more or less active. The link between the magnitude of the impact of crime hot-spots on house prices and citywide crime levels is discussed in chapter 4. While this interpretation does much to explain the gradual decline in the impact of crime hot-spots on house prices, observed between 2011 and 2020 it does less to explain the sharp increase in the magnitude of this impact observed between 2008 and 2011. In this time the number of annual arrests effected by the NYPD, used here as an indicator of citywide crime levels, rose only slightly, the spike in the magnitude of the impact of crime hot-spots appears disproportionate. What then triggered this sudden growth if not a increase in crime? Recall the conclusion of Gibbons: that it is the fear of victimisation rather than specific incidents that causes crime to impact urban property prices [2]. With this in mind consider the great recession of 2008/09: When the US housing market burst it triggered a global financial crisis that led to a surge in unemployment in many countries, including the United States [35]. In the following years millions of workers remained at risk of job loss [36]. Many assumed that the recession would result in a surge in crime rates but such a phenomenon was not observed [37]. Based on this it can be concluded that the sharp increase in the magnitude of the impact of crime hot-spots on house prices observed between 2008 and 2011 was triggered not by an actual increase in crime levels but rather by an increased fear of victimisation brought on by the fear and uncertainty introduced by the global financial crisis. The expected surge in crime rates

meant that safety was seen as more of a commodity and so buyers placed a greater value on it. When this surge failed to materialise and crime rates began to fall buyers placed less emphasis on safety and so the impact of crime hot-spots on house prices began to decline.

This research project has been affected by a number of limitations: Previous works have used racial makeup as a neighbourhood attribute [3], if there does exist a relationship between the price of a property and the racial makeup of the neighbourhood then the use of arrests data to identify crime hot-spots may introduce some bias. The demographic information of arrestees was not considered when determining crime hot-spots, a closer look reveals that the majority of arrestees were black or hispanic. Consequently the areas identified as hot-spots are more likely to be predominantly black or hispanic neighbourhoods. As a result the model may be attributing changes in price to the impact of crime hot-spots when in reality the differences are caused by racial discrimination in the housing market. Other possible sources of error include the functional form of the hedonic regression model. The assumption of a linear dependence between sale price and distance to a crime hot-spot assumes that a difference of $1km$ in distance from the nearest crime hot-spot will have the same impact on a property that is close to a hot-spot as it will on one that is already far away. It may be that the impact of crime hotspots is itself lessened with distance from a crime hotspot such that moving an already distant property further away will have a smaller effect than moving a property that

is closer to the hot-spot. The data constraints affecting the model have already been discussed in chapters 3 and 4. Suffice to say that the absence of a suitable number of property and neighbourhood specific attributes proved a serious limitation of this work.

5.2 Suggestions for Future Works

Future works may benefit from repeating this analysis with the use of a more complete dataset that includes the various property and neighbourhood specific attributes whose absence affected this work. A model built with such a dataset may more accurately reflect the impact of crime hotspots on house prices. The same might be true for models that utilise more advanced techniques, such as state space modelling. Once the impact of crime hot-spots on house prices is fully understood it may be worthwhile to examine their effect on other property types. [4] found that the impact of crime hot-spots was more pronounced for houses than for flats/-condominiums. The analysis could be expanded to investigate the impact of crime hotspots on commercial properties.

Existing works on crime concentration make mention of using the insights gained to inform policing policy [13]. This raises a number of important questions. Can so called 'hot-spot policing' lessen the impact of crime hot-spots on house prices? If so, how long do property prices take to

recover after the nullification of a crime hot-spot? These questions are all possible avenues for further research. The implication that the great recession increased the impact of crime hotspots on house prices could also be examined more closely. The interaction between macroeconomic factors such as unemployment rates and the effect of crime on property prices should also be investigated more thoroughly in future.

Bibliography

[1] V. Ceccato and M. Wilhelmsson, "Does Crime Impact Real Estate Prices? An Assessment of Accessibility and Location," in *Oxford Handbook of Environmental Criminology* (G. J. Bruinsma and S. D. Johnson, eds.), vol. 1, Oxford University Press, Feb. 2018.

[2] S. Gibbons, "The Costs of Urban Property Crime," *Economic Journal*, vol. 114, pp. F441–F463, Nov. 2004.

[3] A. K. Lynch and D. W. Rasmussen, "Measuring the impact of crime on house prices," *Applied Economics*, vol. 33, pp. 1981–1989, Dec. 2001.

[4] V. Ceccato and M. Wilhelmsson, "Do crime hot spots affect housing prices?," *Nordic Journal of Criminology*, vol. 21, pp. 84–102, Jan. 2020.

[5] M. A. Andresen, P. J. Brantingham, and J. B. Kinney, *Classics in Environmental Criminology*. Boca Raton, UNITED STATES: Taylor & Francis Group, 2010.

[6] R. Plunz, *A History of Housing in New York City*. Columbia University Press, Dec. 2018.

[7] F. E. Zimring, *The City That Became Safe: New York's Lessons for Urban Crime and Its Control*. Oxford University Press, Nov. 2011.

[8] "Floyd vs. City of New York," 2013.

[9] A. Watkins, "N.Y.P.D. Disbands Plainclothes Units Involved in Many Shootings," *The New York Times*, June 2020.

[10] A. Southall, "This Police Captain's Plan to Stop Gun Violence Uses More Than Handcuffs," *The New York Times*, Feb. 2022.

[11] D. Weisburd, "The Law of Crime Concentration and the Criminology of Place*," *Criminology*, vol. 53, no. 2, pp. 133–157, 2015.

[12] J. R. Hipp and S. A. Williams, "Advances in Spatial Criminology: The Spatial Scale of Crime," *Annual Review of Criminology*, vol. 3, no. 1, pp. 75–95, 2020.

[13] M. A. Andresen and N. Malleson, "Testing the Stability of Crime Patterns: Implications for Theory and Policy," *Journal of Research in Crime and Delinquency*, vol. 48, pp. 58–82, Feb. 2011.

[14] L. W. Sherman, P. R. Gartin, and M. E. Buerger, "Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place*," *Criminology*, vol. 27, no. 1, pp. 27–56, 1989.

[15] D. Weisburd and S. Amram, "The law of concentrations of crime at place: The case of Tel Aviv-Jaffa," *Police Practice & Research*, vol. 15, pp. 101–114, Apr. 2014.

[16] R. Boivin and S. N. de Melo, "The Concentration of Crime at Place in Montreal and Toronto," *Canadian Journal of Criminology and Criminal Justice*, vol. 61, pp. 46–65, Apr. 2019.

[17] L. Jaitman and N. Ajzenman, "Crime Concentration and Hot Spot Dynamics in Latin America," Working Paper IDB-WP-699, IDB Working Paper Series, 2016.

[18] C. Gill, A. Wooditch, and D. Weisburd, "Testing the "Law of Crime Concentration at Place" in a Suburban Setting: Implications for Research and Practice," *Journal of Quantitative Criminology*, vol. 33, pp. 519–545, Sept. 2017.

[19] D. Weisburd, S. Bushway, C. Lum, and S.-M. Yang, "Trajectories of Crime at Places: A Longitudinal Study of Street Segments in the City of Seattle*," *Criminology*, vol. 42, no. 2, pp. 283–322, 2004.

[20] A. A. Braga, A. V. Papachristos, and D. M. Hureau, "The Concentration and Stability of Gun Violence at Micro Places in Boston, 1980–2008," *J Quant Criminol*, vol. 26, pp. 33–53, Mar. 2010.

[21] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, vol. 82, pp. 34–55, Jan. 1974.

[22] K. C. Bishop and A. D. Murphy, "Valuing Time-Varying Attributes Using the Hedonic Model: When Is a Dynamic Approach Necessary?," *Review of Economics & Statistics*, vol. 101, pp. 134–145, Mar. 2019.

[23] B. Keskin, "Hedonic analysis of price in the istanbul housing market," *International Journal of Strategic Property Management*, vol. 12, pp. 125–138, Apr. 2008.

[24] S. Selim, "DETERMINANTS OF HOUSE PRICES IN TURKEY: A HEDONIC REGRESSION MODEL," *Doğuş Üniversitesi Dergisi*, vol. 9, pp. 65–76, Jan. 2008.

[25] C. Shimizu, H. Takatsuji, H. Ono, and K. G. Nishimura, "Structural and temporal changes in the housing market and hedonic housing price indices: A case of the previously owned condominium market in the Tokyo metropolitan area," *International Journal of Housing Markets and Analysis*, vol. 3, pp. 351–368, Jan. 2010.

[26] J. Zabel, "The hedonic model and the housing cycle," *Regional Science and Urban Economics*, vol. 54, pp. 74–86, Sept. 2015.

[27] N. C. Poudyal, D. G. Hodges, and C. D. Merrett, "A hedonic analysis of the demand for and benefits of urban recreation parks," *Land Use Policy*, vol. 26, pp. 975–983, Oct. 2009.

[28] H. Baba, H. Nishi, A. M. Seetharamapura, and C. Shimizu, "Dynamic Hedonic Analysis Using Time-Varying Coefficients: Application to

Dubai's Housing Market," *University of Tokyo Centre fod Spatial Information Science*, vol. 170, p. 18.

[29] S. N. Lieske, R. van den Nouwelant, J. H. Han, and C. Pettit, "A novel hedonic price modelling approach for estimating the impact of transportation infrastructure on property prices," *Urban Studies*, vol. 58, pp. 182–202, Jan. 2021.

[30] D. Melser, "The Hedonic Regression Time-Dummy Method and the Monotonicity Axioms," *Journal of Business & Economic Statistics*, vol. 23, no. 4, pp. 485–492, 2005.

[31] J. J. F. Commandeur and S. J. Koopman, "Introduction," in *An Introduction to State Space Time Series Analysis*, pp. 1–8, Oxford, UNITED KINGDOM: Oxford University Press, Incorporated, 2007.

[32] H. S. Guirguis, C. I. Giannikos, and R. I. Anderson, "The US Housing Market: Asset Pricing Forecasts Using Time Varying Coefficients," *J Real Estate Finan Econ*, vol. 30, pp. 33–53, Feb. 2005.

[33] A. N. Rambaldi and C. S. Fletcher, "Hedonic Imputed Property Price Indexes: The Effects of Econometric Modeling Choices," *Review of Income and Wealth*, vol. 60, no. S2, pp. S423–S448, 2014.

[34] R. Schulz and A. Werwatz, "A simple state space model of house prices," in *Applied Quantitative Finance: Theory and Computational Tools* (W. Härdle, T. Kleinow, and G. Stahl, eds.), pp. 283–307, Berlin, Heidelberg: Springer, 2002.

[35] I. Islam and S. Verick, "The Great Recession of 2008–09: Causes, Consequences and Policy Responses," in *From the Great Recession to Labour Market Recovery: Issues, Evidence and Policy Options* (I. Islam and S. Verick, eds.), pp. 19–52, London: Palgrave Macmillan UK, 2011.

[36] R. Torres, "The global jobs crisis and beyond," tech. rep., Internat. Inst. for Labour Studies, Geneva, 2009.

[37] R. Rosenfeld, "Crime and the Great Recession: Introduction to the Special Issue," *Journal of Contemporary Criminal Justice*, vol. 30, pp. 4–6, Feb. 2014.

Appendix A

Maps

Presented are a collection of maps showing the distribution of arrests across census tracts in each of the five boroughs of NYC for the entire 15 year period of observation.

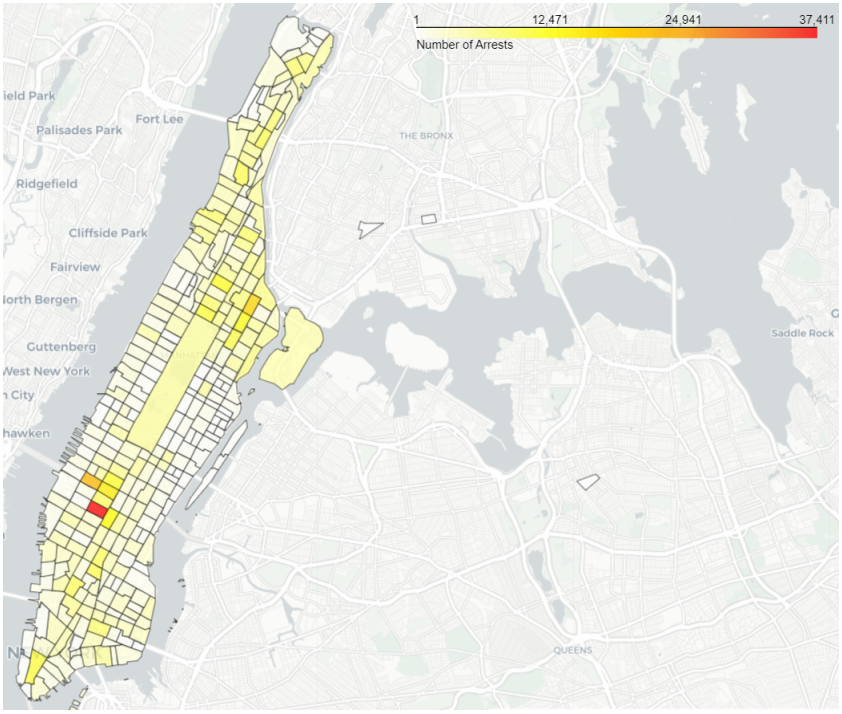


Figure A.1: The distribution of arrests by census tract in the Manhattan over 15 years (2006 - 2020)

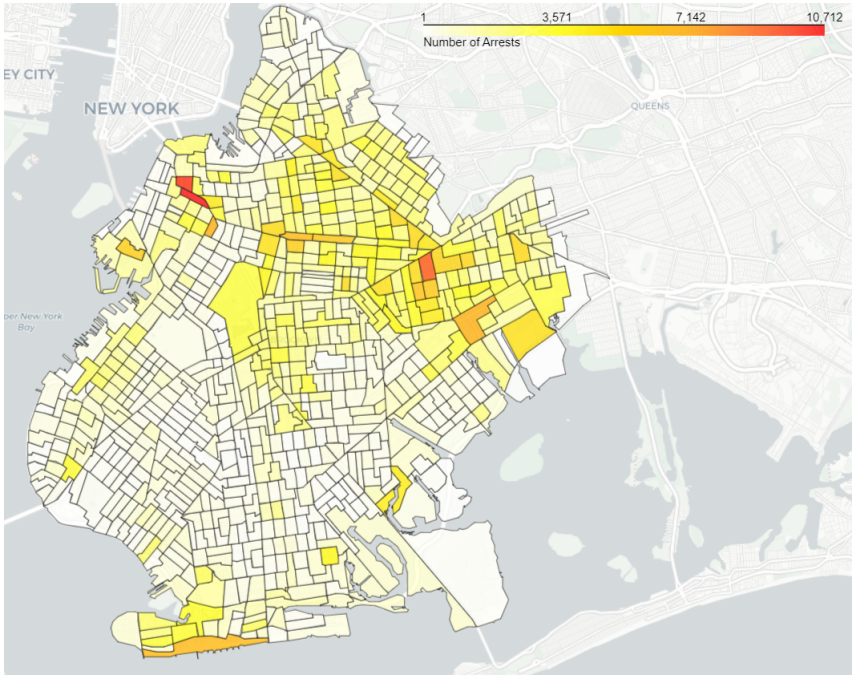


Figure A.2: The distribution of arrests by census tract in the Brooklyn over 15 years (2006 - 2020)

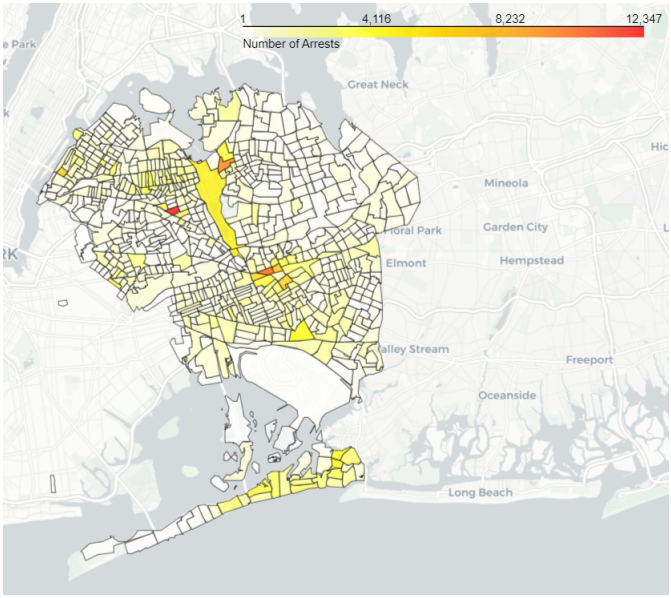


Figure A.3: The distribution of arrests by census tract in the Queens over 15 years (2006 - 2020)

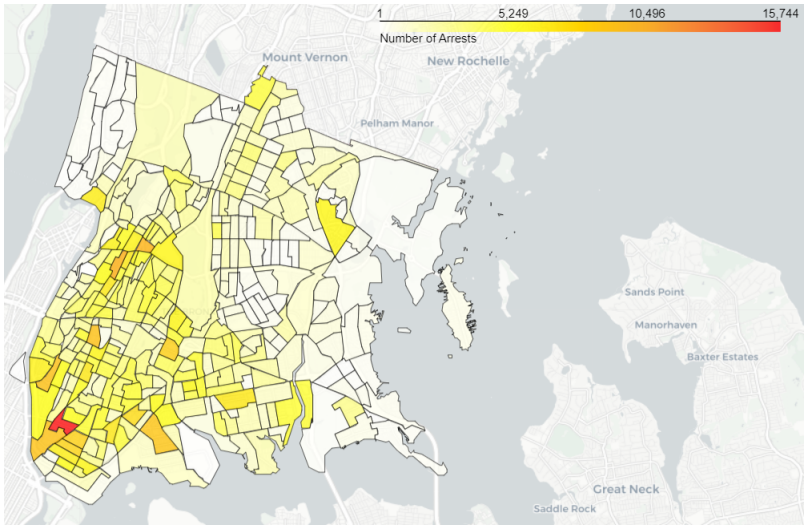


Figure A.4: The distribution of arrests by census tract in The Bronx over 15 years (2006 - 2020)

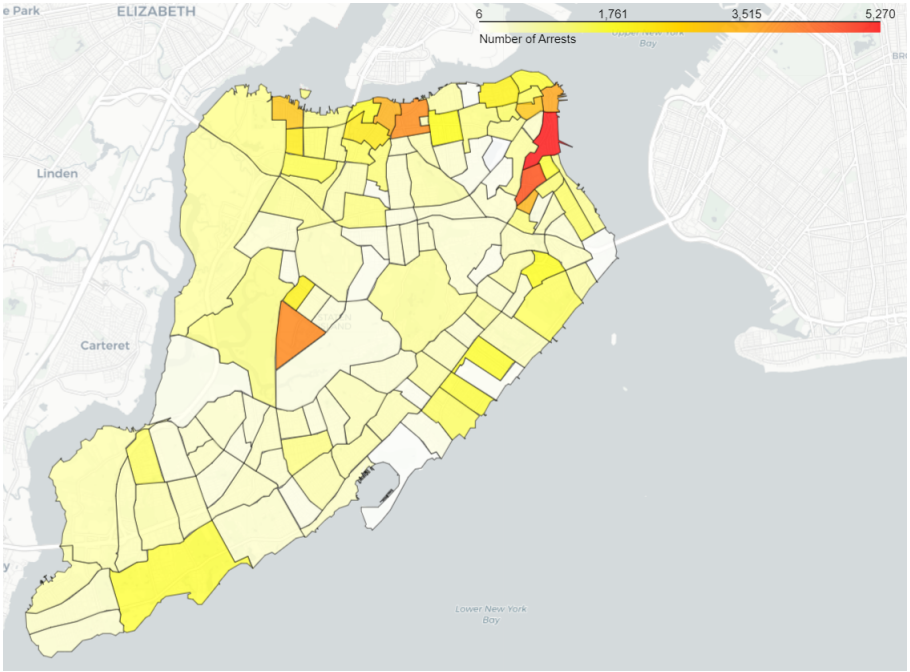


Figure A.5: The distribution of arrests by census tract on Staten Island over 15 years (2006 - 2020)

Appendix B

Code Snippets

The following is a collection of extracts from the python code used in the implementation of this research project. Not all of the code used in the implementation has been included, extracts have been chosen because they provide some key functionality or posed some novel technical challenge. A full collection of the code used to implement this project is available on *GitHub*[1]. Note that due to the omission of certain large data files this code is not executable, the collection is purely archival.

B.1 Geocoding Addresses

The following code block defines a function that takes as input a *pandas* DataFrame with the columns `address` and `zipcode` that provide the street addresses and zipcodes of a number of properties. The function uses

[1] `https://github.com/EK-DATA9003/DATA9003-Research-Project`

python's *geopy* package to query the *ArcGIS* API and obtain geographic coordinates for each property. The parameter `SalesDir` is optional and is used to provide a path to a directory in which the resulting collection of addresses and coordinates can be saved. This allows users to avoid having to repeat the geocoding process which can take some time (when applied to the filtered house sales dataset it took almost 72 hours to get results)

```
import os
import pandas as pd
from geopy.extra.rate_limiter import RateLimiter
from geopy.geocoders import ArcGIS

def GeocodeAddresses(sales_df, SalesDir=None):

    # only the street address and zipcode are required for geocoding
    mydf = sales_df.loc[:, ["address", "zipcode"]]

    # a single property may have been sold multiple times
    mydf.drop_duplicates(inplace=True)

    # combine the street address and zipcode into the full address
    mydf["FullAddress"] = mydf["address"] + ", NEW YORK, NEW YORK, " +
   mydf["zipcode"].astype(str)
    mydf.rename(columns={"address":"StreetAddress"},
                inplace=True)

    # initialise the connection to ArcGIS
    locator = ArcGIS(user_agent="NYCsales")

    # 1 - object to geocode addresses with a delay between calls (avoid
    usage limits)
    geocode = RateLimiter(locator.geocode, min_delay_seconds=0.1)

```

```
    # 2- - create location column
    mydf['location'] = mydf['FullAddress'].apply(geocode)

    # 3 - get longitude, latitude and altitude from location column (
    break tuple into multiple columns)
    mydf['point'] = mydf['location'].apply(lambda loc: tuple(loc.point)
    if loc else None)
    mydf[['latitude', 'longitude', 'altitude']] = pd.DataFrame(mydf['
    point'].tolist(),
                                                               index=
    mydf.index)

    # remove unneeded columns
    mydf.drop(["FullAddress", "altitude"],
              axis=1,
              inplace=True)

    # save the results to the SalesDir directory, if specified
    if SalesDir is not None:
        filepath = os.path.join(SalesDir, "PropertyCoords.csv")
        mydf.to_csv(filepath,
                    index=False)

    return mydf
```

B.2 Calculating Distances for Spatial Attributes

The following code block defines two functions. The first `NearestNeighbours` takes as input two dataframes: `FROM` and `TO` and returns for each point in `FROM` the distance to the nearest point in `TO`. This is used to calculate the distance of each property from the nearest school, subway station etc. The

dataframes must specify the locations of points with two distinct columns giving the latitude and longitude in degrees.

The second function, `DistanceToShore`, takes a dataframe of properties with latitude and longitude columns and calculates the distance of each property in ft to the NYC shoreline. Because this code was written as part of a python package rather than as an isolated file the reading in of static assets (i.e. the GIS data for the shoreline) must be managed with *importlib*. Note that `NearestNeighbours` function could also have been implemented using *geopandas*' built in functionality but *sklearn*'s *BallTree* method is much more efficient for large dataframes. The built in functionality is used here because it was considered easier than trying to transform the shoreline to a representation as a series of distinct latitude, longitude pairs.

```
import pandas as pd
import geopandas as gpd
import numpy as np
from sklearn.neighbors import BallTree
from importlib import resources

def NearestDistance(FROM, TO):
    F = pd.DataFrame()
    T = pd.DataFrame()

    # get coordinates in radians
    for col in FROM[["latitude", "longitude"]]:
        F[col] = np.deg2rad(FROM[col].values)
    for col in TO[["latitude", "longitude"]]:
        T[col] = np.deg2rad(TO[col].values)
```

```
16
17      # initialise BallTree object
18      Btree = BallTree(T.values,
19                       metric='haversine')
20
21      # query the Btree object for the nearest neighbour of each property
22      distance, index = Btree.query(F.values, k=1)
23
24      # distances returned are for unit sphere
25      # convert to kilometres by multiplying by earth's radius: 6371km
26      distance = distance * 6371
27
28      return distance.flatten()
29
30
31  def DistanceToShore(houses):
32      # create GeoDataFrame to use coords as GIS data
33      geodf = gpd.GeoDataFrame(houses,
34                               geometry=gpd.points_from_xy(houses.
    longitude,
35                                                           houses.
    latitude))
36      geodf.set_crs(epsg=4326,
37                    inplace=True)
38      # change crs to EPSG:2263, this means resulting distances will be
    in ft
39      geodf.to_crs(epsg=2263,
40                   inplace=True)
41
42      # read in GIS data for NYC shoreline
43      with resources.path("DATA9003/assets", "shoreline.geojson") as
    filepath:
44          geofile = open(filepath, "r")
45          shoreline = gpd.read_file(geofile)
46          shoreline.to_crs(epsg=2263, inplace=True)
```

```
        geofile.close()

    # geopandas has some functionality for calculating distances
    between different geometries
    def getdist(point):
        distances = shoreline.distance(point)
        return min(distances)

    # get the distance to shore for each property
    dist2shore = geodf.geometry.apply(getdist)

    return dist2shore
```

B.3 The Rolling Regression Model

The following code defines the function used to divide the data in training and test set, ensuring that the same proportions are taken from each time period, as well as the RollingRegression class. This class includes functions to fit the model, estimate house prices using the fitted model and generate plots showing how the model coefficients change over time. While *sklearn* does provide support for stratified splitting of train and test sets it insists that the data be shuffled first. This is not appropriate here because it is important that the data remains in chronological order. Hence a new function had to be defined.

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
import statsmodels.api as sm
```

```
import matplotlib.pyplot as plt
from matplotlib.lines import Line2D

def TrainTest(X, y, t):
    # for each quarter in t split the data into training (80%) and test (20%) sets

    for quarter in t.unique():
        X_t = X.loc[(t == quarter)]
        y_t = y.loc[(t == quarter)]
        t_t = t.loc[(t == quarter)]

        X_train_t, X_test_t, y_train_t, y_test_t, t_train_t, t_test_t = train_test_split(X_t, y_t, t_t, test_size=0.2, random_state=1)

        # concatenate the sets for each time set to get overall training and test set
        if "X_train" not in locals():
            X_train = X_train_t
            X_test = X_test_t
            y_train = y_train_t
            y_test = y_test_t
            t_train = t_train_t
            t_test = t_test_t
        else:
            X_train = pd.concat([X_train, X_train_t])
            X_test = pd.concat([X_test, X_test_t])
            y_train = pd.concat([y_train, y_train_t])
            y_test = pd.concat([y_test, y_test_t])
            t_train = pd.concat([t_train, t_train_t])
            t_test = pd.concat([t_test, t_test_t])

    return X_train, X_test, y_train, y_test, t_train, t_test

class RollingRegression:
```

```
37      # steps to initialise an instance of the Rolling Regression Model
38      def __init__(self, X, y, t):
39          # initialise window width and step size of sliding window
40          self.window = 4
41          self.stepsize = 1
42
43          # store all data as instance variables
44          self.data = X
45          self.response = y
46          self.time = t
47
48          # some basic error checking
49          n_obs = X.shape[0]
50          if (len(t) != n_obs):
51              raise ValueError("t should be of length {} but is of length
    {}".format(n_obs, len(t)))
52          if (len(y) != n_obs):
53              raise ValueError("y should be of length {} but is of length
    {}".format(n_obs, len(y)))
54
55          # calculate the number of windows needed to cover the entire
    data set
56          n_endog = X.shape[1]
57          n_windows = (len(t.unique()) // self.stepsize) - (self.window -
     self.stepsize)
58
59          # initialise arrays for output to 0
60          self.coeffs = np.zeros((n_windows, n_endog))
61          self.pvals = np.zeros((n_windows, n_endog))
62          self.Rsq = np.zeros(n_windows)
63          self.Rsq_adj = np.zeros(n_windows)
64          self.Fstat = np.zeros(n_windows)
65
66      # function to fit the RR model
67      def fit(self):
```

```
68          # indices
69          idx = 0
70          lwr = 0
71          upr = self.window
72
73          # iterate through the data set
74          while upr <= len(self.time.unique()):
75              window = list(self.time.unique())[lwr:upr]
76
77              # filter for data in the current windoe
78              x_window = self.data.loc[self.time.isin(window), :]
79              y_window = self.response.loc[self.time.isin(window)]
80
81              # fit a linear regression model to the window
82              model_window = sm.OLS(y_window, x_window, hasconst=True)
83              res_window = model_window.fit()
84
85              # add the coefficients and evaluation metrics to the
    relevant arrays
86              self.coeffs[idx] = np.array(res_window.params)
87              self.pvals[idx] = np.array(res_window.pvalues)
88              self.Fstat[idx] = np.array(res_window.fvalue)
89              self.Rsq[idx] = np.array(res_window.rsquared)
90              self.Rsq_adj[idx] = np.array(res_window.rsquared_adj)
91
92              # update indices
93              lwr = lwr + self.stepsize
94              upr = upr + self.stepsize
95              idx = idx + 1
96
97          # convert 2D arrays to dataframes
98          self.coeffs = pd.DataFrame(self.coeffs,
99                                     columns=self.data.columns)
100         self.pvals = pd.DataFrame(self.pvals,
101                                   columns=self.data.columns)
```

```
102
103     # function to predict the sale price of new observations using the
    fitted model
104     def predict(self, X, t):
105         # initialise array for output
106         ypred = np.zeros(len(X))
107
108         # iterate through the observations by index
109         for i in range(len(X)):
110             # get the attribute values
111             obs_i = np.array(X.iloc[i, :])
112
113             # get the relevant coefficients for the time
114             # any quarters before Q4-2006 use the coeffs for Q4-2006
115             if t.iloc[i] <= 3:
116                 beta = self.coeffs.iloc[0, :]
117             else:
118                 beta = self.coeffs.iloc[t.iloc[i] - 4, :]
119
120             # calculate the estimate for this observation
121             ypred[i] = np.dot(obs_i, beta)
122
123         return ypred
124
125     # function to plot the coefficients over time
126     def PlotCoefficients(self, coeff=None):
127         time = self.coeffs.index
128
129         # define legend elements: green dot = yes, red dot = no
130         legend_elements = [Line2D([0], [0],
131                                   marker="o",
132                                   color="w",
133                                   markerfacecolor="green",
134                                   markersize=10,
135                                   label="Yes"),
```

```
                           Line2D([0], [0],
                                  marker="o",
                                  color="w",
                                  markerfacecolor="red",
                                  markersize=10,
                                  label="No")]

    # if a coefficient isn't specified then plot all coeffs in a grid
    if coeff is None:
        fig, ax = plt.subplots(5, 2, figsize=(40, 50))

        # grid indices
        row = 0
        col = 0

        # add a plot of each coeff to the grid
        # colour code points based on their significance at 5%: Green = Yes, Red = No
        for beta in self.coeffs.columns:
            pass_ = self.pvals[beta] <= 0.05
            fail_ = self.pvals[beta] > 0.05

            ax[row, col].plot(time, self.coeffs[beta], c="#baafaf", linestyle="--", zorder=1)
            ax[row, col].scatter(time[pass_], self.coeffs.loc[pass_, beta], c="green", zorder=2)
            ax[row, col].scatter(time[fail_], self.coeffs.loc[fail_, beta], c="red", zorder=3)
            ax[row, col].set_title(beta, fontsize=30)
            ax[row, col].tick_params(axis='x', labelsize=20)
            ax[row, col].tick_params(axis='y', labelsize=20)

            ax[row, col].legend(handles=legend_elements, title="Significant at 5%:")
```

```
165
166                 #update indices
167                 row = row + 1
168                 if (row > 4) and (col == 0):
169                     col = 1
170                     row = 0
171
172         # if a coefficient is specified then plot only that coefficient
173         else:
174             pass_ = self.pvals[coeff] <= 0.05
175             fail_ = self.pvals[coeff] > 0.05
176
177             plt.figure(figsize=(20, 5))
178             plt.plot(time, self.coeffs[coeff], c="black")
179             plt.scatter(time[pass_], self.coeffs.loc[pass_, coeff], c="
    green")
180             plt.scatter(time[fail_], self.coeffs.loc[fail_, coeff], c="
    red")
181             plt.legend(handles=legend_elements, title="Significant at
    5%:")
182             plt.grid()
183             plt.title(coeff, fontsize=30)
184             plt.xticks(fontsize=18)
185             plt.yticks(fontsize=18)
```

Appendix C

Dashboard

A dashboard presenting some additional visualisations and a summary of the main findings of the project was developed but due to time constraints the final dashboard still suffered from technical bugs and has not been published. The code used to create the dashboard is included in the gitHub archive but again is not executable due to the omission of large data files. Figures C.1 & C.2 shows some screenshots of the dashboard.

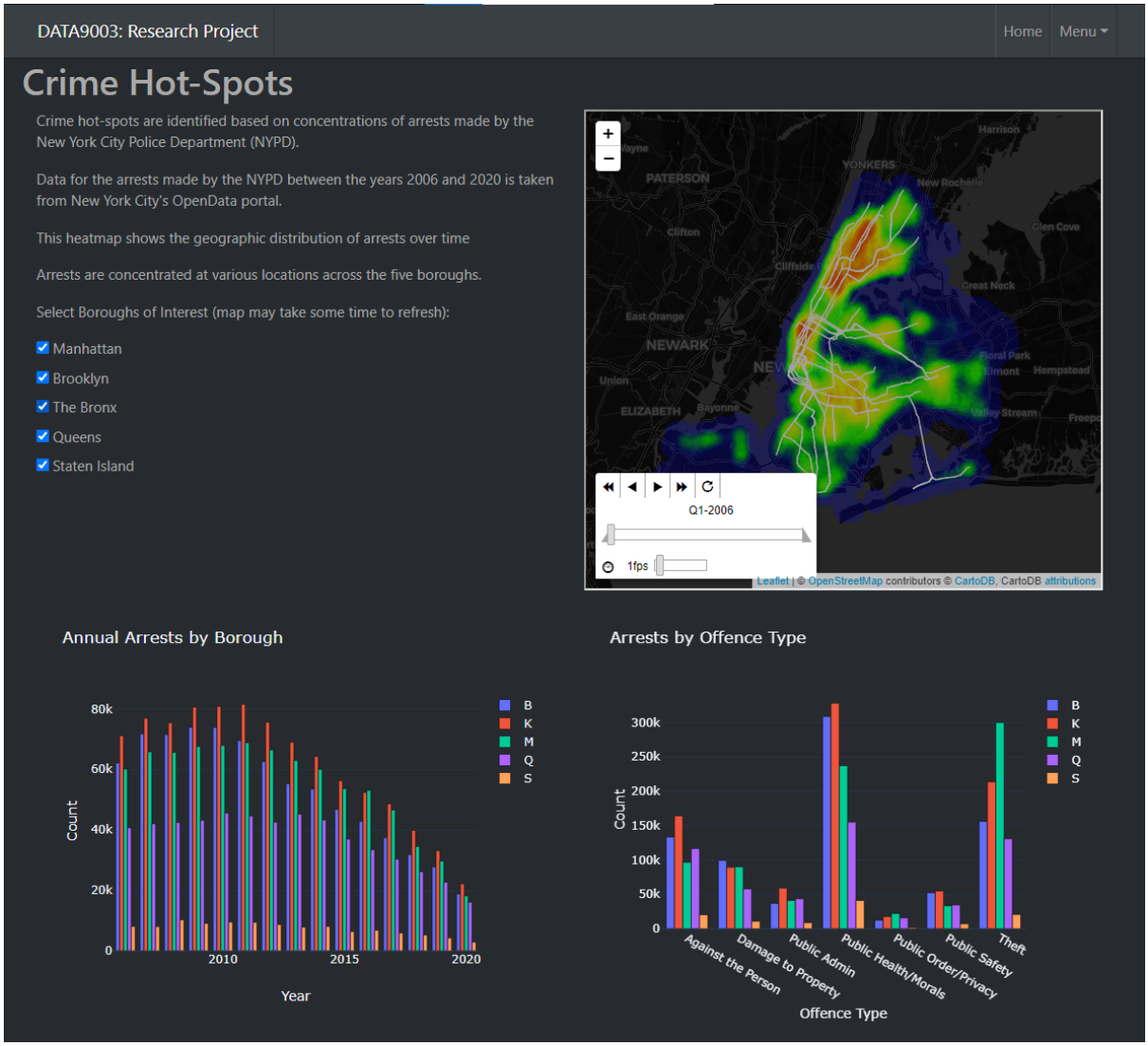


Figure C.1: Dashboard page exploring the arrests data

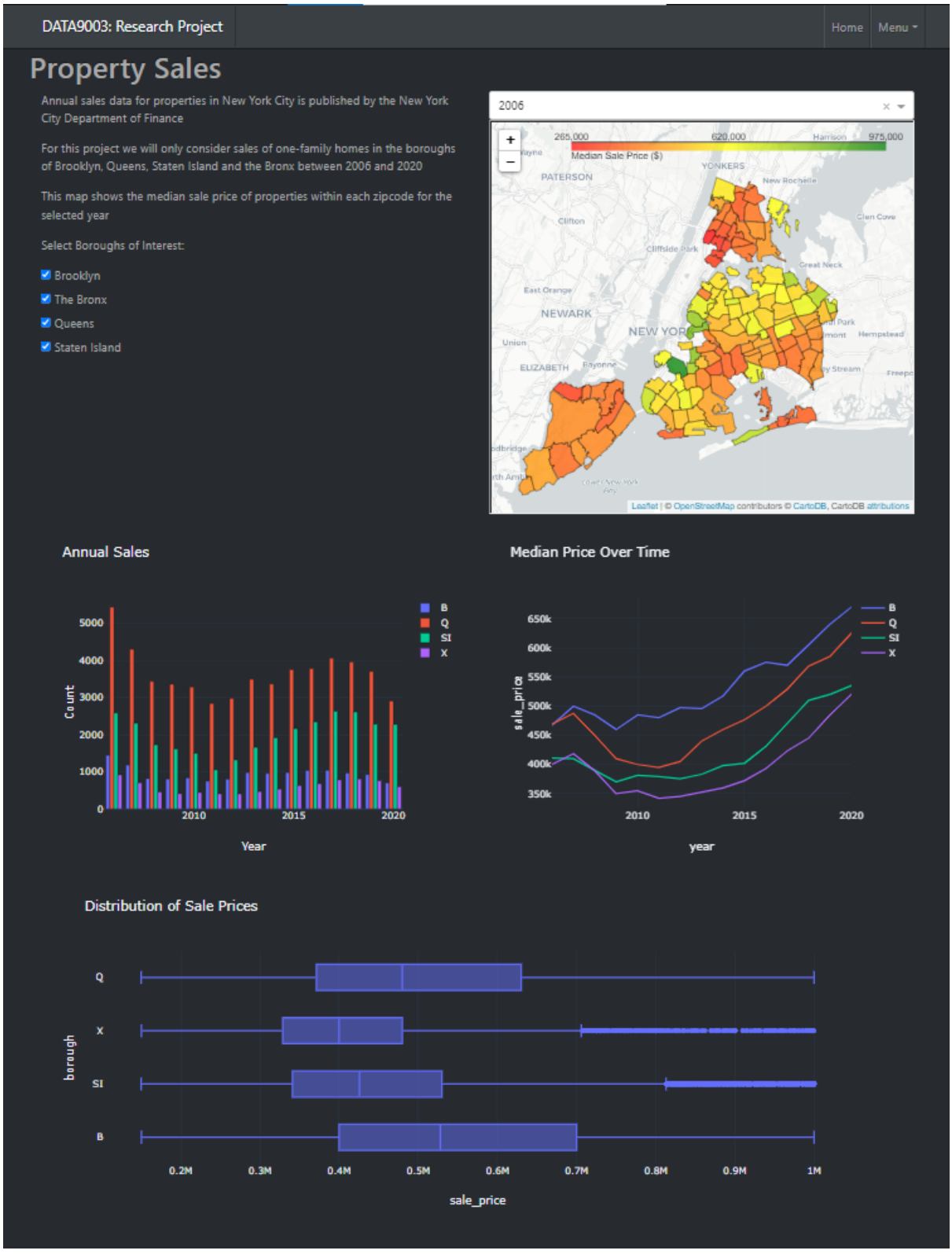


Figure C.2: Dashboard page exploring the house sales data