Impact of Crime Rate on the Housing Market in London

Suraj Divakala
Student number: 219430818

Supervisor: Christos Vasilakis

Word count: 12,005

Submitted on 5th September 2022
as a part of the requirement for completing the MSc in Business Analytics at the University of Bath

Abstract

This paper focuses on exploring the impact of different types of crimes on the housing market in London, England. A mathematical model is built that allows us to understand how each type of crime impacts the housing market. This model is based on historical data on crime and real estate sales in several boroughs in London. This paper merges these two datasets, i.e., historical crime data in London and house sales in London, to perform this analysis. This paper then proceeds to predict the future crime rates by using historical time series data. ARIMA models are built to predict the future crime counts monthly. These counts are then used as input in the previously mentioned mathematical model to predict the future sales of houses in London.

The different types of crimes this paper takes into account are violence against a person, theft and handling, robbery, drugs, criminal damage, burglary, fraud and forgery, sexual offences, and other notifiable offences. The following boroughs are included in this paper: Barking and Dagenham, Barnet, Bexley, Brent, Bromley, Camden, City Of London, Croydon, Ealing, Enfield, Greenwich, Hackney, Hammersmith, and Fulham, Haringey, Harrow, Havering, Hillingdon, Hounslow, Inner London, Islington, Kensington and Chelsea, Kingston upon Thames, Lambeth, Lewisham, London, Merton, Newham, North East, North West, Outer London, Redbridge, Richmond upon Thames, South East, South West, Southwark, Sutton, Tower Hamlets, Waltham Forest, Wandsworth, and Westminster.

Keywords: Housing market, Borough, ARIMA models, Criminal damage

Acknowledgement

I would like to acknowledge and give my thanks to my supervisor, Christos Vasilakis, who guided me through my dissertation program. I would also like to mention my professors, Melih Celik and Fotios Petropoulos, who helped me understand all the tools and techniques I have used for my dissertation.

I also want to thank all the faculty and student support team at the University of Bath School of Management for their support. And finally, I would like to express gratitude to my family and friends for their unwavering faith in me and their emotional support.

Table of Contents

1. Introduction

1.1 Problem Statement

Home, to most people, is a place where they feel safe both physically and emotionally. Living amidst a neighborhood full of people with malicious, evil, and malign intentions can never bring peace of mind to a person. Logically, one would not want to buy a house in a neighborhood where crime rates soar. However, several other factors come into play when it comes to a person's decision to buy a house.

Interest rates are one of the most influential factors when it comes to real estate sales. More often than not, people buy houses by taking home loans from banks, which finance their purchases in exchange for mortgage rates. Consistently low mortgage rates would slowly inflate the prices of houses; as the cost of borrowing decreases, more people are able to afford houses, thereby increasing the demand and concurrently the prices of houses (Nielsen, 2022). Secondly, economic factors also have an impact on the housing market. Economic indicators such as gross domestic product (GDP), disposable income, inflation, etc., all contribute to the health of the housing market (Nguyen, 2021).

The extensive research that countless entities across the globe have already put in to gauge the impact of crime on the housing market goes to show how difficult it is to develop accurate models using actual world data. Most papers focus on the average price at which properties are being sold to measure the impact of crime on the housing market. But, as mentioned before, prices of properties can be influenced by several factors other than crime. Moreover, the average costs of properties do not truly reflect the demand for real estate. The number of houses sold in a given time period, however, accurately represents the demand for real estate in a given region. Hence, this paper assumes the number of houses sold as a benchmark to gauge the impact of crime on the housing market.

The different types of crime considered in this paper are violence against a person, theft and handling, robbery, drugs, criminal damage, burglary, fraud and forgery, sexual offences, and other notifiable offences. Violence against a person includes incidents of homicide and violence with or without injury. Theft and handling include incidents of bicycle theft, shop lifting, theft from a person and any other kinds of stealing. Robbery includes all incidents of both business property robberies and private property robberies. Drug offences include incidents of drug trafficking and counts of illegal possession of drugs. Criminal damage includes incidents of vandalism, arson and all other types of damage to public or private property. Burglary count, just like that of robbery, include incidents of both residential burglaries and business or community burglaries. Fraud and forgery include all counts of deceit, contract by force, forging signatures, and malice. Sexual offences include rape, inappropriate touching, and non-consensual acts of intimacy. Other notifiable offences incidents such as violent disorder, offences against state, dangerous driving, perjury, soliciting for prostitution, possession of weapon intent to commit crime, etc. (Metropolitan Police Service, 2017)

1.2 Research Objectives

This paper uses time series data of the number of crimes committed in all the boroughs of London (Boysen, 2017). This dataset contains monthly data from January 2008 to December 2016 of the counts of all the different types of crime mentioned above.

The second dataset used for this paper is the number of houses sold in the same boroughs of London during the same period (Cirtautas, 2020).

The objective of this research paper is to mathematically define the impact of crime on the housing market. It also aims to predict the future values of house sales using this mathematical relationship and predicted values of crime in the future.

The first part of this paper aims to construct a mathematical model to depict the relationship between crime and number of houses sold. This model also illustrates which type of crime has the most impact and which type of crime has the least impact on house sales in London. Multiple linear regression is used to achieve this.

In the second part of the paper, forecasting model is built using the ARIMA method to predict the future crime rates. RStudio is used to build the ARIMA model and achieve this objective. These predicted values are then used as input in the regression model mentioned above to predict the number of houses sold.

1.3 Structure of the Dissertation

This paper is structured as follows.

Chapter 1 introduces the context of this research and its aims and objectives. Chapter 2 contains the literature review part of this research. Chapter 3 discusses the methodology employed to achieve the aforementioned objectives. Chapter 4 explores the datasets closely and the mathematical model defining the mathematical relationship between crime and house sales is discussed in detail along with the forecasting models built. Chapter 5 shows the findings of this paper i.e., the forecasted house sales. Chapter 6 discusses the limitations, possible future work and contribution to knowledge of this paper. And lastly, Chapter 7 summarizes the whole study.

2. Literature Review

2.1 Background

The housing sector is one of the most important economic sectors. This market represents the consumer spending of a particular geographical area. If the market is healthy and the real estate prices go up, homeowners are more confident and spend more. They also have the option of taking loans against the price of their house, which they would use to pay off debt or adjunct their pension fund. In the same way, in the case where the housing market is not healthy, homeowners still under mortgage have a fear that the price of the property will end up being lesser than the mortgage amount. Hence, they are more apprehensive and careful with their spending habits (Bank of England, 2020).

Zhu (2014) talks about how important the housing sector is for the financial health of the country. A country with a high number of homeowners is prepared to deal with adverse shocks. The mortgage market is a market derived from the housing market. These mortgage markets act as great platforms for the implementation of monetary policies. This is the easiest way for the government to control the cash in the market and prevent inflation. He also goes on to say that “housing makes up the largest component of wealth in many countries.” He insinuates that the housing market is a very sought after investment product.

Given the importance of this sector, a lot of studies have been done which attempt to understand the impact of different factors on it. In a study focused on the housing market in Malaysia, the various factors affecting the house prices are discussed – Foreign direct investments, gross domestic product, interest rates and unemployment. FDI, in this context, is where foreign nationals invest in the real estate market, thereby driving up the prices. The study also states that the GDP does not have a direct effect on the housing prices, but in the case of a collapse in the housing market, the GDP will definitely fall. Interest rates change the total cost of buying a house by large amounts. High interest rates will force people to rent and not own a property whereas lower rates will encourage more house purchases. Also, unemployment has an effect on both supply and demand. As the rate of unemployment increases, the demand takes a hit because people would not want to commit themselves to a big purchase without job security. In the same way, as the rate of unemployment increases, people who already own a home would not be willing to sell it and relocate without job security. So, both demand and supply take a hit if the rate of unemployment of a country goes up. Lastly, inflation is an important factor that affects the prices of houses in a country. Especially for brand new houses, rise in inflation increases the cost of construction. For renovated houses, inflation causes a hike in the cost of modifying as well. These will result in higher prices of new houses (Latif, et al., 2020).

Crime is another important factor that affects the housing market of a country, and that is the focus of this paper. Criminal activity is one of the biggest barriers to economic growth in many different countries. As mentioned above, one of the biggest boosters to the housing market is foreign direct investments. If the crime rate is high, these investments are discouraged due to lack of security and safety. The market becomes uncertain. Detotto and Otranto (2010) conducted research to study how the crime of a country affects its economic growth by taking Italy as an example. They proposed a state space model to closely analyze the implications of crime on the economy and applied this to the crime and economic activity data of Italy. With their study, they were able to prove that crime does in fact, disrupts the growth of an economy. They claim that a 1% rise in the crime rate reduces the actual economic growth by 0.00040%.

Now, let us look at studies that focused on the impact of crime solely on the housing market.

2.2 Previous study

A high rate of crime in a region where an individual is planning to buy a home is always a negative. The two main intents behind purchasing a house are the intent to purchase a place to live in and the intent to invest. With a high rate of crime, especially the types of crime mentioned in this paper, neither of these intents would be satisfied. No one

would want to live in a city with a high rate of crime unless they have no other options. Similarly, investors would not want to invest in an asset which is located in a high crime area. There would not be enough safety and security for these assets and the crime rate would decrease the demand for these properties, thereby decreasing the value and it would not be profitable for the investor.

Given how important the crime factor is for the housing market, a lot of research has been done to evaluate this relationship. Several researchers tried different methods to define this relationship between the housing market and the crime rate. Lynch and Rasmussen (2001) measured the impact of crime on house prices in Jacksonville, Florida. Rather than taking the number of crimes committed, they used the cost of crime as a measure to weigh the seriousness of the crime. They were able to conclude that even though the cost of crime had no impact whatsoever on the prices of the houses overall, they were comparatively much cheaper in areas with a high rate of crime.

In another research paper, the focus of which is quite similar to the current paper, Minghetti (2020) explores the correlation between crime and house prices in London. Contrary to the actual belief, a positive correlation is observed between them, which means that as crime increases, the price of the houses in that area increases. The author suggests that this is caused by endogeneity in the relationship between crime and property prices. He later introduces appropriate controls and distance from the nearest police station to the model and concludes that a 1% rise in crime decreases the value of a property in London by roughly 0.7%.

Another research carried out by Tita, Petras and Greenbaum (2006) first segregates the geographical area of the city of Columbus into neighborhoods based on the income of the people living there. Their results show that the overall impact of crime on the housing prices across the city is negligible. However, their analysis also provides an explanation as to why that finding is ambiguous. The crime data cannot be trusted completely as several crimes go unreported. This leads to unexpected results and false relationships. The authors also reference a few other research studies, for example the research study of Lynch and Rasmussen (2001) mentioned above, to further prove his point.

Segregating the geographical area in terms of income has revealed some important insights. Another study of the impact of crime on the housing market, conducted by Mclhatton, McGreal, Paz and Adair, uses the same technique. They were able to conclude that the impact of crime on the housing prices is very complex and it varies by the type of crime, type of property and the location of the crime and property. They also find that crimes such as burglary and theft are associated with high income neighborhoods and crimes such as criminal damage, violence against a person and drugs offences mostly occur in the low income neighborhoods.

Another study based in the LA count of the United States also gives some insights about the relationship between crime and house sales. This study is also based on spatial analysis. The underlying logic of the analysis of this study is that different types of crime have different impact distance. For example, drugs related crimes have shorter impact distance but crimes such as burglary and assaults, the impact distance is significantly higher. The study was able to conclude that correlation did exist between the value of properties and crime. The study was able to prove that the

housing prices is negatively associated with the levels of drugs related crimes and assaults. The spatial analysis also revealed that the farther a property is from an incident of a crime, such as vandalism and robberies, the higher the value of that property (De La Pax, et al., 2022).

The one thing that is common in all of these research studies is the approach for gauging the impact on the housing market. All of these papers use the value of the properties as a metric to measure the impact. This is the gap that this paper addresses.

2.3 Research Gap

The real estate prices of a certain region have innumerable factors affecting them. As mentioned in Chapter 2.1, economic factors such as GDP, interest rates, inflation, FDI and unemployment all affect the real estate prices of a region (Latif, et al., 2020). Moreover, geographic factors, such as proximity to famous places, climate factors, such as temperature and urbanization levels all affect the prices of properties.

However, these factors do not affect the demand for the houses. The demand can be attributed to the number of houses being sold. When compared to the long term historical data, we can know determine whether the demand is high or low. And this true representation of demand is used as a metric to measure the impact of crime on the housing market in this paper.

The analysis done by Scott (1990) in his research study, which determines whether the prices of houses reflect the market fundamentals in the real estate market, further helps in proving my point. By applying regression tests, volatility tests and simple mean tests, he concludes that the prices in the real estate industry does not in fact reflect the market fundamentals. This goes to show that the prices of houses do not truly reflect the demand for said houses.

Given that the prices of houses are affected by so many factors, analyzing its variation just by relating it to crime will not give us accurate results. And by using the number of houses sold as a metric to represent the demand, this analysis can be helpful in predicting the future demand levels.

2.4 Datasets

There are two datasets being used for this paper. One is the crime dataset which was originally published by the Metropolitan Police Service. This was later summarized and published on Kaggle by Jacob Boysen. The data for the house sales has been published on Kaggle by Justinas Cirtautas who extracted it from the London Datastore.

Both these datasets are rich in facts and hence, have been used several times for other projects. The housing data, for example, was used to predict future sales based on historical data, to discover trends in the housing market over the long term etc. The crime data was used to make crime maps in London to highlight the geographical regions of the city where the frequency of crime is higher. This way, both datasets have a wide range of possible applications.

2.5 Papers using similar techniques

A lot of papers, although focusing on other topics, contributed to the working and theory behind this paper. Regression and forecasting are the two main concepts used in this paper. To understand the working and the limitations of these concepts, several other papers that used the same concepts were studied.

A paper titled 'Using regression analysis to predict the future energy consumption of a supermarket in the UK' has been used to understand the application of multiple linear regression on real world data. In this particular paper, the authors, Braun, Altan, and Beck (2014) use multiple regression to predict the impact of climate change on the energy consumption. A supermarket in the UK is the focus of their research and they use the data from this supermarket for their analysis.

Similar to the case of the current paper, this paper also builds the regression model on historical data and then uses future predicted values to determine the future costs. After doing so, they were able to conclude that the electricity use will increase by 2.1% whereas the gas consumption will drop by about 13%. However, these authors used SPSS to build their regression model, whereas this paper adopted the use of MS Excel to accomplish this task. Although the focus area of both the papers are very different, the methodology employed, and the flow of process is similar.

Another paper titled 'Analysis and Interpretation of Findings Using Multiple Regression Techniques' was of great help to this paper. The author, Hoyt (2006), illustrates how flexible the regression technique is and discusses its wide range of applications. The author also reiterates a point made in this paper earlier, that the metric of measurement, like the number of houses sold rather than the average price of a house, in the current paper should be chosen carefully.

Some other papers were also taken into study to understand the working of the forecasting model, that was employed in the current paper. Ho and Xie (1998) discuss the working of the ARIMA model. They compare it to the working of the traditional approach – Duane model. They analysis shows that ARIMA is an effective forecasting technique and gives satisfactory results. Using the results of their analysis as evidence, they conclude that the ARIMA modelling technique is a promising alternative for repairable system analysis.

Another paper published by Chen, Yuan and Shu (2008) also uses ARIMA modelling to predict the future values of crime in one of the cities in China. They use the historical dataset of 50 weeks to predict the values for the next one week. The concept of forecasting is thoroughly discussed and the accuracy of the ARIMA model is tested by comparing it to other forecasting models.

All these papers were found in either the Scopus database or in Google scholar and have helped in the understanding of the working knowledge of the different mechanisms employed in the current paper. All these papers together inspired the study reported in this paper.

In addition to these, dissertations presented in the library of the University of Bath, carried out by students of the previous classes proved to be of great help. Not only did it help me understand how to apply some of the techniques on real world data, but it

also served as a guide for structuring my own work. Two papers from the library have been used as reference for my dissertation which I would like to mention below.

Zeng (2021) submitted a dissertation titled 'Time Series Model and ARIMA-Genetic Algorithm for Forecasting Rice Production in Asia'. The methodology of forecasting employed in this paper was to build three forecasting models using simple moving average, exponential smoothing and ARIMA methods and then pick and develop the most promising model. The objective of this paper was to predict future food production which can then be compared to the demand to prepare for any shortages or abundance.

Xu (2018) submitted a dissertation titled 'Relationships Between Air Pollution and Weather Conditions/Living Habits, and Future Air Pollutant Emissions Forecasting, Evidence from: China, South Korea and India'. Similar to the current paper, this paper includes both analyzing the relationship between two variables and then using forecasting techniques to predict future values of a variable.

2.6 Motives behind crime

This paper also tries to understand the motives behind crime to get a complete understanding of the whole picture. It is important to know why crimes are being committed to think of ways to control it or to analyse its impact on an economic sector.

Ali and Ahmad (2015) conducted a research study to analyse the motivation for crimes in Pakistan. They conducted interviews with a few prisoners from a district jail in one of the cities in Pakistan. They collected data from 15 people from a total of 150 prisoners between the ages of twenty and forty.

They were able to find out that most of these criminals already had prior ties to the world of crime. Few of them were from families that were already heavily involved in crime. They know the criminal ways and techniques, which probably rubbed off on them from a very early age. Few of them were from law abiding families but got mixed up with the wrong people. Associating themselves with people who did not obey the law also rubbed off on them and they slowly slipped into that world.

Their study also discovered that most of these people feel that the law of the country is not fair and is unjust. They believe the laws are made to suit the interests of the rich and wealthy. Surprisingly, these prisoners agree that laws should exist and be enforced correctly to keep communities peaceful, but they believe that the policies should change to support all the people of the country equally. Some of these prisoners do not trust the justice system at all, because either they themselves or someone they know were wrongfully convicted of a crime. These people eventually started doing crime because they were anyway already labelled criminals.

On the whole, the prime motive behind engaging in criminal activities was to get rich easy and fast. They wanted a good socio-economic status in the easiest way possible, and crime was the way to go. The second most common motive was entertainment, especially in the younger side of the sample. The thrill of crime was what made them commit crimes. And lastly, satisfaction was claimed to be a motive behind crime for a couple of the prisoners.

The narratives from these prisoners also suggested that being a criminal involved extremely long training. It is not an easy way out. They train not only to commit a crime but also to get away safely and not get caught by the police post committing a crime. Most of these criminals were trained by their family or friends who were already associated with crime and unlawful activities. They were usually trained from when they were kids about the different techniques of theft and crime.

After thoroughly reading this study, self justification among the criminals' mindset is very apparent and obvious. They do not think of it as something bad. Although they regret being caught, they do not regret being criminals. If they were to be released from prison, they would go back to the unlawful ways immediately. One of the prisoner who was convicted of drugs related crime said that he considered what he did as service and not as something bad. He supplied a product to someone who needed it badly. That is how his illegal acts are justified in his head.

Now let us look at some more violent crimes and how the motives for those are different from those of non-violent crimes.

Morrall (2006) looks at the different motives and causes behind murder. To relate it to the current paper, the category of 'Violence against a person' involves murders. In his research, Morrall was able to state 4 main reasons behind murders and he calls them the 4 Ls – Lust, love, loathing and loot. Lust refers to the murders committed for sexual satisfaction, love refers to the mercy killings or the killing of someone to put them out of their misery. Loathing refers to the revenge killings or hate crimes. And lastly, loot is murders committed for financial gain i.e., for insurance payout, for inheritance or jobs as a contract killer.

Certain people commit murders because of schizophrenia. That is a serious mental disorder in which people interpret reality abnormally (Mayo Clinic Staff, 2020). They suffer from delusions and hallucinations. They imagine seeing something that is not actually there. They imagine that they are being harassed or troubled, when in fact nothing of the sort is happening. These schizophrenics act out in unexpected ways and constitute a major portion of murderers and serial killers. They are usually relentless and cruel towards their victim, punishing them and dealing damage way more than what is required to kill them (Amani et al., 2022).

To specifically talk about the crime rates in London, a research paper documenting the impact of crime on tourists in Jamaica is referenced (Alleyne and Boxill, 2003). This paper talks about how an increase in tourist influx from the European countries into Jamaica caused a hike in the crime rate. This paper was able to prove that the Jamaican economy had developed even with the crime rate that high. It goes on to explain that the revenues earned by the tourism sector was much higher than the losses incurred due to the crime rate.

However, the important point to take away from that research paper is how tourism brought about a hike in the crime rates. London is a hotspot for tourism. Home to some of the most famous buildings and architecture in the world, such as Big Ben, Tower bridge, Buckingham palace, etc. it attracts tourists in large numbers every year. There are also several football clubs based in London such as Chelsea, Arsenal, Tottenham etc. Fans from all over the world travel to London just to see their favorite team play. It is also home to some of the most famous stadiums in the world, where world famous

artists perform. With some of the best cuisines of the world available here, it attracts people from all over the country. When travelling, tourists usually carry a lot of cash with them, in the case that other methods of payment do not work. International tourists have important documents on their person, such as passports and citizenship cards. People attending concerts might be wearing expensive accessories or jewelry.

There is a lot at stake for these tourists if they find themselves in the victim role of a crime. So, they are easy targets for criminals to rob. And since London has an abundance of tourists, the rate of crime in London is consistently high even though the police forces are high and prepared for these outcomes.

3. Methodology

As mentioned before, this paper is based on two datasets, both of which have been taken from Kaggle. The dataset for crime was provided by Boysen (2017) which covers the number of criminal reports by month, borough and category from January 2008 to December 2016. The dataset for the house sales was provided by Cirtautas (2020), who extracted this data from several datasets released on London Datastore. Although this dataset contains data from January 1995 to January 2020, we only consider the 8 years between 2008 and 2016, the time period for which we already have the crime data.

3.1 Data Preparation

The housing in London dataset (Cirtautas, 2020) has been manipulated to be compatible with this research. As mentioned before, this dataset contains data from January 1995 to January 2020. However, the crime dataset (Boysen, 2017) only contains data from January 2008 to December 2016. Hence, only the data points from this time period are taken into account. The rest of the data is not included in this analysis.

Exploratory analysis revealed that the ‘Fraud and Forgery’ and ‘Sexual Offences’ has a count of 0 for every month except for the months of January and February of 2008. Given the presence of so many null values, these two categories of crime are omitted from this research. They would have had no impact on the mathematical model or can be forecasted. Hence, there is no point in considering them as factors anymore in this research.

Moreover, upon close examination, it has been identified that some of the boroughs for which housing data was provided, are not a part of the boroughs of greater London. Some of them are vague and some of them are in a part far away from London and hence are not included as boroughs in the crime dataset. Hence, these particular areas i.e., East midlands, West midlands, England, Ease of England, and Yorks and the Humber, have all been excluded from the housing dataset with the help of pivot tables in MS Excel.

After manipulating the datasets to be compatible with each other, both datasets are merged together to form a database of the counts of different types of crime committed and the number of houses sold every month from January 2008 to December 2016.

3.2 Regression

Regression is a mathematical way to explore the impact of one variable on another. It helps in understanding which factors have the greatest impact on an event and which factors have little to no impact on the outcome (Gallo, 2015).

In this instance, the different types of crime i.e., violence against a person, theft and handling, robbery, drugs, criminal damage and other notifiable offences, are all the independent variables. The underlying hypothesis of this paper is that these factors have an impact on the number of houses sold in the same region. Hence, the number of houses sold is the dependent variable.

The dependent variable and all the independent variables are continuous i.e., they do not have a limit on the number of values they can take. They can be any number between the lowest and highest points of measurement (McCue, 2007). Hence, we can use multiple linear regression to mathematically define the relationship between the dependent variable and the independent variables. Multiple linear regression analyses the relationship between the dependent and independent variables and enables us to know how strong the relationship is and the value of the dependent variable at a certain value of independent variables (Bevans, 2020)

Multiple linear regression is given by the following equation (Hayes, 2022)

$$y = b0 + x1(b1) + x2(b2) \dots xn(bn)$$

y = independent variable
x1, x2,...xn = independent variables
b0 = y-intercept (constant)
b1, b2,...bn = co-efficient for each independent variable

The regression model presented in this paper was built in excel using the Data Analysis ToolPak.

3.3 Forecasting

ARIMA models are built using RStudio to forecast the future values of the crime data.

ARIMA stands for 'Autoregressive Integrated Moving Average'. This is a statistical tool that can be fit on time series data to predict future values. Before we understand ARIMA models further, let us grasp a few concepts quickly.

A time series data is several data points indexed in time order. These data points are in chronological order and the time between successive data points is constant. This data can be yearly or monthly or weekly or daily or hourly. (Hayes, 2022)

Stationary data is data free of trends and seasonality. In the simplest way, data points are generated from events. The outcome of these events can change over time, but if

the way these events occur change, then a mathematical relationship cannot exist. Hence, it is always essential to convert a time series data into stationary data free of trends and seasonalities (Palachy, 2019). This can be done by differencing, which is a method widely adopted to convert non-stationary data into stationary data. It involves subtracting the previous observation from the current observation (Brownlee, 2017)

Now, let us get back to understanding more about ARIMA models. The best way to understand about ARIMA models is by diving it into parts –

1. AR
   The AR stands for auto regressive. This model predicts the future values based on historical data. It does not assume anything or take any other factors into consideration. It simply analyses historical data to determine trends and patterns and predicts the future values in the best way possible.

2. I
   The I is an abbreviation for integrated. It shows that the model has converted the time series data into stationary data. As mentioned above, trends and seasonalities are removed from the time series data. The number of times the data has been differenced is also noted.

3. MA
   The MA is an abbreviation for moving average. This means that this model assimilates the dependency between an observation and a residual error from a moving average model applied to lagged observations (Hayes, 2021).

These values are changed, and multiple forecasting models are built. Each forecasting model has 3 values associated with it called AIC, BIC, and AICc. The lower these values are, the better the model is. However, the auto ARIMA function in RStudio does this building and comparison for us and gives us the best possible model.

RStudio is an integrated development environment for the R programming language. One of its main purposes is to an open-source software for data analytics and statistical computing. In this paper, it has been used to perform exploratory analytics and to build forecasting models.

4. Data Analytics

This chapter closely examines the data to discover facts and underlying relationships. It also discusses how the regression model and the forecasting model are built. Its working is explained in detail and the results obtained from these models are discussed.

4.1 Removing collinearity

As mentioned in Chapter 3.1, two of the different types of crime – Sexual offences and fraud and forgery are mostly null values. Therefore, they have been omitted from this research. The other types of crimes are burglary, criminal damage, drugs, robbery, theft and handling, violence against the person and other notifiable offences.

It is important to note that there is a little degree of correlation between the different types of crime. Logically speaking, if the crime rate is high, the incidents are reasonably distributed among the different categories of crime. It is quite rare that a particular type of crime is high while the other types are low. Given that the motives behind committing these crimes are somewhat similar, there exists a certain degree of correlation between them.

Ideally, for regression, the independent variables must be truly independent. Independent from other events as well as each other. If these events are not truly independent, like in this case, it is a case of multicollinearity. In the case that the degree of multicollinearity is high, it can lead to incorrect results and interpretation of the regression model (Frost, 2017). Why is multicollinearity an issue?

When we interpret the results of the regression model, it tells us how the outcome of the dependent variable, the number of houses sold, will change with change in any one of the independent variables and the other independent variables being constant. However, with the presence of multicollinearity, change in one independent variable should cause a change in another independent variable as well. In this way, the whole accuracy and precision of the regression model takes a hit and the results cannot be trusted.

This paper adopted the use of Variance Inflation Factor to get rid of multicollinearity in the data. This factor estimates the inflation in the regression values caused due to multicollinearity. MS Excel has been used to calculate the VIF for the regression model. Each of the independent variables are regressed against the other independent variables to get the R^2 values. These values can then be plugged into the VIF formula

$$VIF = \frac{1}{1 - R^2}$$

Upon employing this to our dataset, it was observed that the crime type - Other notifiable offences, had the highest value of VIF. That means that the degree of correlation of this particular crime with the other crimes is the highest. Hence, given that this type of crime does more damage than good to our regression models, it is dropped from this research.

4.2 Relationship between crime rates and number of houses sold

To discover reliable relationships between crime rates and the number of houses sold, the paper also analyses the correlation between the crime rates and the number of houses sold.

Both the crime data and the number of houses sold data are monthly. Therefore, they are homogeneous and simple statistical correlation is employed. It is a measure of the degree of association between two variables (Glen, n.d). In this case, the degree of association between the different types of crime and the number of houses sold.

The table below (table 1) shows the correlation between the different types of crime and the number of houses sold in the boroughs of Westminster, Lewisham, Croydon, Greenwich and Hackney.

Borough	Burglary	Criminal Damage	Drugs	Robbery	Theft and Handling	Violence against a person
Westminster	0.28	-0.45	0.384	0.46	0.32	-0.41
Lewisham	-0.46	-0.89	-0.86	-0.68	-0.39	0.17
Croydon	-0.87	-0.65	-0.82	-0.80	-0.58	0.74
Geenwich	-0.75	-0.75	-0.41	-0.84	0.44	0.71
Hackney	0.67	-0.44	-0.88	0.58	0.78	0.56

Table 1: Correlation between different categories of crime and number of houses sold

Most of these values show that crime has a negative impact on the housing market. For example, both burglary and criminal damage have a significant negative impact on the number of houses sold on Greenwich. Criminal damage also heavily impacted the number of houses sold in Lewisham negatively.

A few anomalies in this matrix can be identified which are against the underlying hypothesis of this paper. For example, there is a high degree of positive correlation between violence against a person and the number of houses sold in the boroughs of Croydon and Greenwich. This can be attributed to the fact that the data under consideration is for a considerably short period of time.

However, it is important to remember that it is dangerous to draw conclusions from these observations given that this data is only for 8 years. Making decisions for the long-term using these insights is not wise. There might be several external factors such as increase in disposable income, change in government's policies, significant drop in mortgage rates, etc.

4.3 Trends in Crimes

4.3.1 Burglary

The sentencing council of UK recognizes burglary as an incident where the offender enters a property as a trespasser with the intention to steal, inflict harm or do unlawful harm.

Given that we have the time series data of counts of burglary for 8 years, the plot below (figure 1) shows the trend in this type of crime.

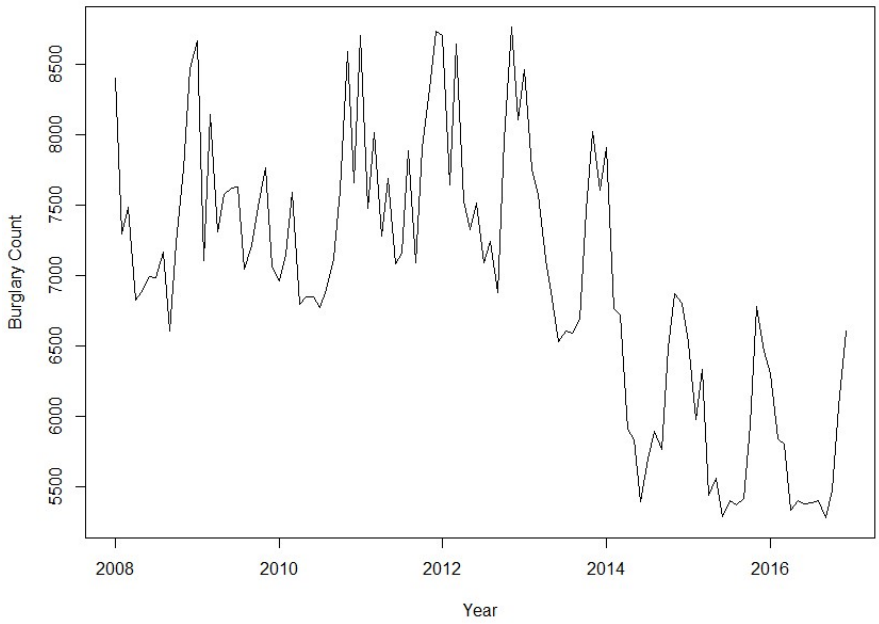

Figure 1: Time series plot of burglaries

This graph has been plotted in RStudio.

It can be seen that burglary has a negative trend. Although the counts of burglary have been somewhat steady from 2008 till 2013, we can observe a significant dip in 2014. Seasonality can also be observed in this data. From the beginning of every year the counts gradually decrease before shooting up at the end of the year. Let us see if this persists in the case of other crimes.

4.3.2 Theft

Theft is different from burglary. The sentencing council of UK explains theft in this way – "a person is guilty of theft if he dishonestly appropriates property belonging to another with the intention of permanently depriving the other of it."

The time series plot of this data looks like this (figure 2)

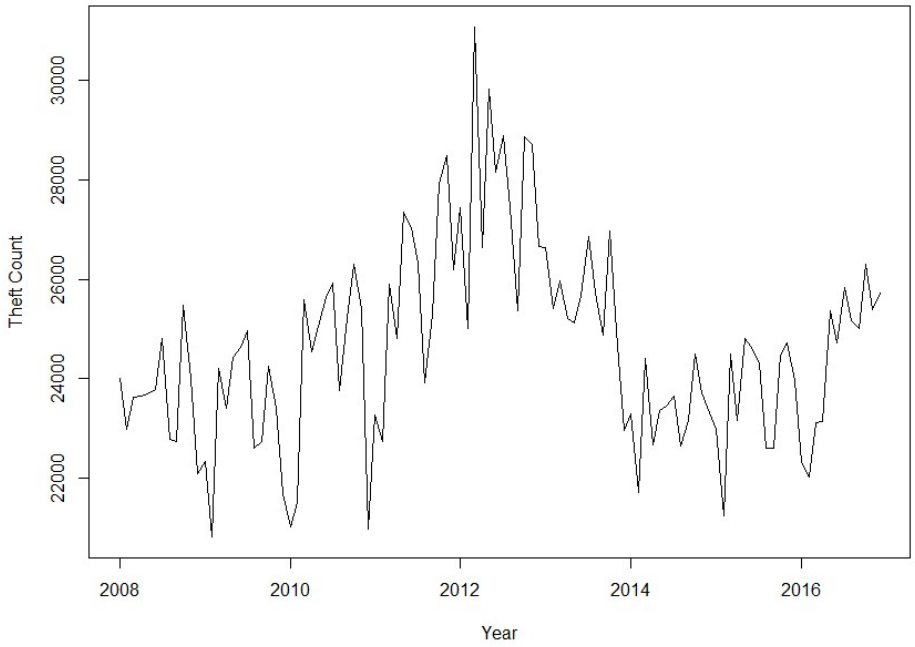

Figure 2: Time series plot of theft and handling

No trend exists in theft. The spread across the Y-axis is pretty small which indicates that the range is quite small. We can see a peak in the year 2012, when the count passed 30,000. Sligh seasonality exists in the data, we can see a steady rise in the count as soon as a new year begins.

4.3.3 Robbery

The sentencing council of UK defines robbery as an incident where a person is guilty of using force in order to steal from a person or from a business. Given the violent nature of this, robberies are dealt with more strictly than theft.

The time series data of robberies committed in the London area between 2008 and 2016 look like this on a graph (figure 3)

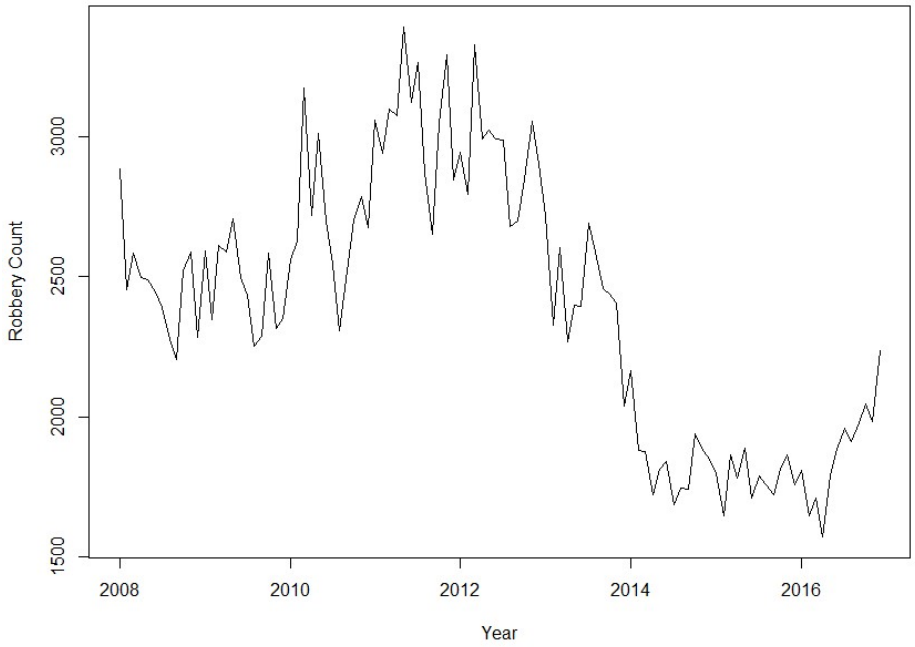

Figure 3: Time series plot of robberies

We can see a negative trend in the number of robberies. Although the counts have been somewhat constant till 2013, they dipped heavily right before 2014. However, a slight upward trajectory can be seen to the end of the plot, towards the end of 2016.

4.3.4 Criminal Damage

Criminal damage is an incident where an individual, either intentionally or recklessly, causes damage to property belonging to another individual or business. A few examples include vandalism, forced entry, graffiti, arson, etc. (Laver, n.d).

The time series data of criminal damage between 2008 and 2016 looks like this when plotted on a graph (figure 4)

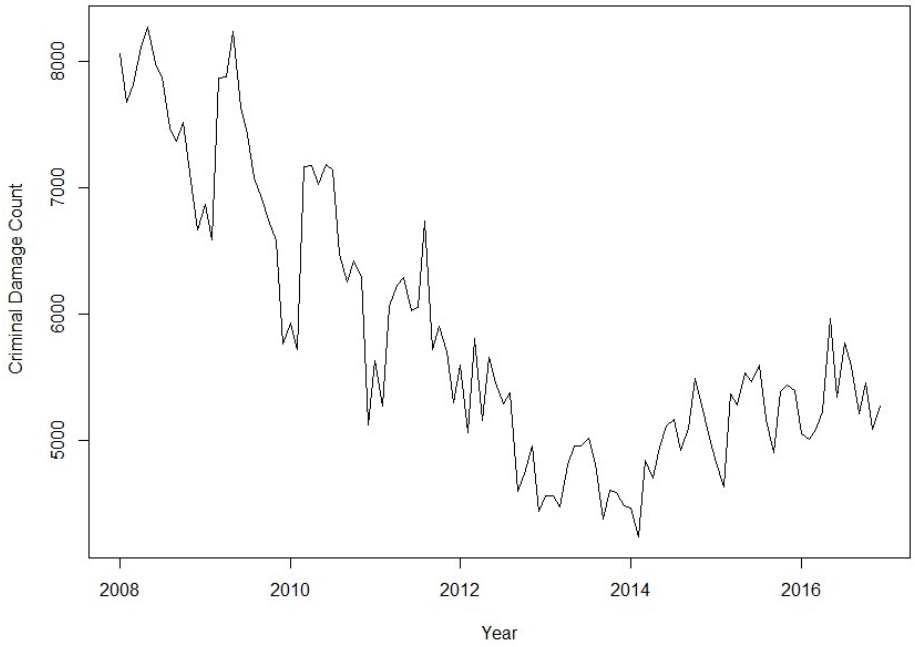

Figure 4: Time series plot of criminal damage

We can see a significant negative trend in criminal damage in the 8 years. However, similar to robberies, there seems to be an upward curve towards the end of 2016. There is not much seasonality that can be observed in this dataset.

4.3.5 Drugs

In the United Kingdom, drugs are classified into class A, B or C, depending on how dangerous or harmful they are to humans. The severity of punishment is the highest for crimes with class A drugs and comparatively lower for class B and lowest for class C. The different types of drug related crimes are possession of controlled drugs, supply of drugs, possession with intent to supply, importation of drugs and production of drugs (Crown Prosecution Service, n.d).

Let us look at the graph of the time series data of drugs from 2008 to 2016 depicted below (figure 5)

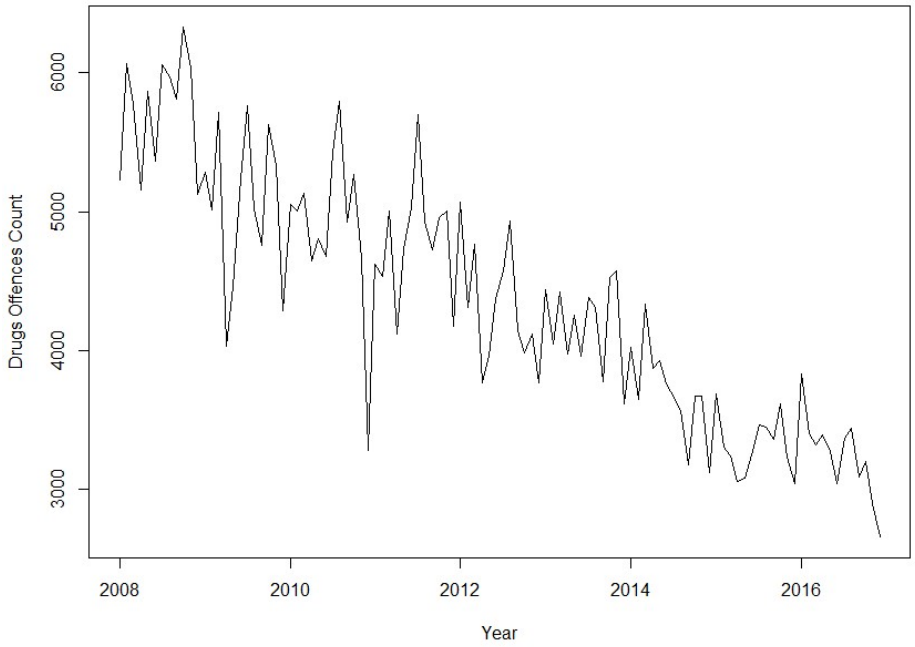

Figure 5: Time series plot of drug related crimes

We can see a very clear decreasing trend in the data. The change is also gradual and constant. We also see a decreasing curve towards the end of 2016, something that has not been observed in any other time series data. There is no significant seasonality that can be observed in this time series data.

4.3.6 Violence against a person

This type of crime includes a wide range of crimes from minor cases such as harassment and bullying to major cases such as assault and homicide. These kinds of crime are the most violent in nature. Every incident involves someone being the subject of harm, either physically or mentally.

The 8 years' time series data of violence against a person plotted on a graph looks like this (figure 6)

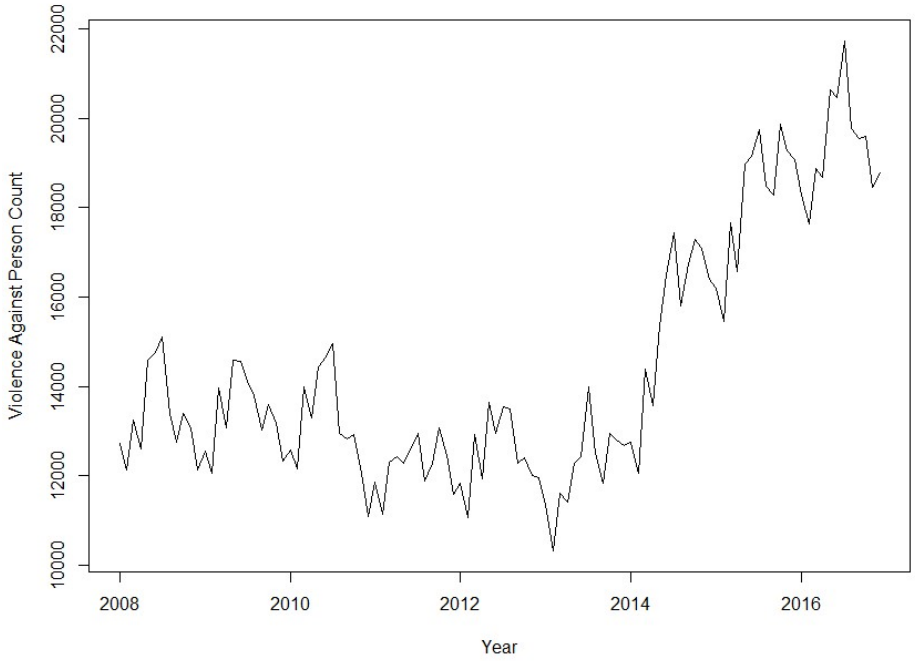

Figure 6: Time series plot of violence against a person

We can observe a significant increasing trend in this data. There is a spike in the year 2014. There exists a certain degree of seasonality in this time series dataset. The count is always highest during the middle of the year and decreases towards the end of the year.

4.4 Regression Model

This section deals with the explanation of the mathematical relationship between crime and the number of houses sold. As mentioned before, the data analytics toolpak in MS Excel has been used to build this regression model.

The independent variables are the counts of the different types of crime – theft and handling, robbery, drugs, criminal damage, burglary and violence against a person. The dependent variable is the number of houses sold.

The following mathematical model has been built to predict the number of houses sold from the crime rate based on the historical data of crime and houses sold from 2008 to 2016

The co-efficients are mentioned in the table below (table 2)

Independent Variables	Co-efficients
Y – Intercept	52585.51933
Violence Against the person	1.051279571
Theft and Handling	1.56698298

Robbery	-9.032268301
Drugs	1.26792589
Criminal Damage	-3.656277472
Burglary	-2.285725892

Table 2: Regression model co-efficients

The Y-intercept remains constant. The other coefficients are multiplied with the counts of the respective crimes. The sum of all these terms gives us the value of the number of houses sold. So, the regression equation is

$$y = 52585.5 + (x_1)\ 1.05 + (x_2)\ 1.56 + (x_3)\ (-9.03) + (x_4)\ 1.26 + (x_5)\ (-3.65) + (x_6)\ (-2.28)$$

Here,

y gives us the number of houses sold. The first term is the y-intercept, which remains constant. x_1 is the count of violence against a person. x_2 is the count of theft and handling incidents. x_3 is the count of robberies. x_4 is the count of drug related crimes. x_5 is the count of criminal damage. And lastly, x_6 is the count of burglaries.

The accuracy of this model can be increased greatly with more data. Given that the data we have is only of 8 years, this model may not be that accurate in the long term. This has to be kept in mind before using this model as a decision-making tool.

This model will be used later on in this paper to predict the future number of houses sold.

4.5 Forecasting Model

Time series data has three components to it. The trend, which is the long term direction of the data. The seasonality, which systematically changes in accordance with the calendar. And lastly, the irregular component, which are just short term fluctuations.

This paper employs the use of ARIMA models for forecasting. The ARIMA model is built by using the auto ARIMA function in RStudio. This function helps us build the most accurate and efficient forecasting model which helps us predict the future values of the time series. Let us understand the working of the auto ARIMA function.

ARIMA is an auto regressive model. That means that it uses historical data of the variable to predict the future values. However, for the model to do this, the data should be free of any trends and seasonalities. The auto ARIMA function in R does differencing, which is a widely adopted process to convert non-stationary data into stationary data. Differencing is a procedure where the observations are changed i.e., previous observations are subtracted from the current observations. This step can be repeated as well. Differencing once gets rid of all the linear trends. Doing it twice will remove quadratic trends. Differencing at a lag equal to the period i.e., monthly or yearly, will remove seasonality. ARIMA also incorporates the use of moving average method in its prediction.

The three components mentioned above, constitute three different parameters of an ARIMA model. An ARIMA model has three parameters – p, d and q.

I. p
This is the number of lag observations in the model. That is, the number of auto regressive terms the model uses in order to make future predictions. For example, in a model with p value equal to 1, the model uses the preceding observation to predict the next observation. In a model with p value equal to 2, the model uses two consecutive observations to predict the next one.

II. d
This is the number of times the data has to be differenced to convert it into stationary data. Differencing once will remove all the linear trends and doing it twice will remove quadratic trends. Differencing it with a lag equal to the period of the time series data i.e., lag of 12 for monthly data, removes seasonality.

III. q
This is the order of the moving average. It is the number of lagged forecast errors in the prediction equation (Nau, n.d).

ARIMA also constructs models for time series data with seasonality. The model denotes the parameters for this data separately. These parameters are usually mentioned in capital letters. Hence, an ARIMA model is represented in this way

ARIMA (p, d, q) (P, D, Q)

Different models can be built using the same data with different values of p, d, and q. Each model gives different results. So let us understand how the auto ARIMA function picks the best model.

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are two of the most widely used selection criteria for forecasting models. These values represent how well the model fits onto the historical data. This method of selection is much easier when compared to cross validation (Tran and Arabnia, 2015). Using these values, the auto ARIMA function in R returns the best possible model to predict the future values.

This method has been employed to predict the future values of the different types of crime. Let us discuss the parameters and the AIC values of each of the models.

4.5.1 Forecast violence against a person

The first model built predicts the count of violence against a person for the 5 years after 2016.

Its parameters are ARIMA (0, 1, 1) (0, 1, 1). The value of p is 0, which means that this model does not use one observation to predict the next one. The time series data of violence against a person is differenced once to get rid of the trend component, which is represented by the d value equal to 1. And lastly, the model has a moving average order equal to 1, as shown by the value of q.

This model has an AIC value of 1484.52, an AICC value of 1484.78 and a BIC value of 1492.18.

4.5.2 Forecast theft and handling

The second model built predicts the count of theft and handling for the 5 years after 2016.

Its parameters are ARIMA (2, 0, 0) (2, 1, 0). The value of p is 2, which means that this model uses two consecutive observations to predict the next one. The time series data of theft and handling is not differenced, because no trend was evident as can be seen in Chapter 4.3.2, which is represented by the d value equal to 0. And lastly, the model has a moving average order equal to 0, as shown by the value of q.

This model has an AIC value of 1627.97, an AICC value of 1628.64 and a BIC value of 1640.79.

4.5.3 Forecast robbery

The third model built predicts the count of robberies for the 5 years after 2016.

Its parameters are ARIMA (0, 1, 1) (1, 0, 1). The value of p is 0, which means that this model does not use an observation to predict another one. The time series data of robberies is differenced once, to get rid of the trend, which is represented by the d value equal to 1. And lastly, the model has a moving average order equal to 1, as shown by the value of q.

This model has an AIC value of 1410.65, an AICC value of 1411.04 and a BIC value of 1421.34.

4.5.4 Forecast drug crimes

The fourth model built predicts the count of drug related crimes for the 5 years after 2016.

Its parameters are ARIMA (0, 1, 2) (2, 0, 0). The value of p is 0, which means that this model does not use an observation to predict another one. The time series data of robberies is differenced once, to get rid of the trend, which is represented by the d value equal to 1. And lastly, the model has a moving average order equal to 2, as shown by the value of q.

This model has an AIC value of 1568.03, an AICC value of 1568.62 and a BIC value of 1581.39.

4.5.5 Forecast criminal damage

The fifth model built predicts the count of criminal damage crimes for the 5 years after 2016.

Its parameters are ARIMA (0, 1, 1) (0, 1, 2). The value of p is 0, which means that this model does not use an observation to predict another one. The time series data of robberies is differenced once, to get rid of the trend, which is represented by the d value equal to 1. And lastly, the model has a moving average order equal to 1, as shown by the value of q.

This model has an AIC value of 1372.2, an AICC value of 1372.65 and a BIC value of 1382.42.

4.5.6 Forecast burglary

The sixth model built predicts the count of burglaries for the 5 years after 2016.

Its parameters are ARIMA (0, 1, 1) (2, 1, 0). The value of p is 0, which means that this model does not use an observation to predict another one. The time series data of robberies is differenced once, to get rid of the trend, which is represented by the d value equal to 1. And lastly, the model has a moving average order equal to 1, as shown by the value of q.

This model has an AIC value of 1401.2, an AICC value of 1401.64 and a BIC value of 1411.41.

5. Results

This section shows the results obtained from the aforementioned forecasting models and predicts the future values of the number of houses sold from these values and the regression model mentioned in Chapter 4.4.

5.1 Forecasting results

The forecasting models discussed in Chapter 4.5 are used to predict the values of the counts of different types of crimes. This forecasted data is also monthly and is calculated till the December of 2021.

5.1.1 Future incidents of violence against a person

These (table 3) are the results for the crime category of violence against a person from January 2017 to December 2021 derived from the forecasting model

	2017	2018	2019	2020	2021
January	18418.73	19460.87	20503	21545.14	22587.27
February	17718.64	18760.78	19802.91	20845.05	21887.18
March	19411.58	20453.72	21495.85	22537.99	23580.12

April	18830.91	19873.04	20915.18	21957.31	22999.45
May	20537.24	21579.37	22621.51	23663.64	24705.78
June	20686.22	21728.35	22770.49	23812.62	24854.75
July	21541.27	22583.4	23625.53	24667.67	25709.8
August	20099.43	21141.56	22183.7	23225.83	24267.97
September	19933.61	20975.74	22017.87	23060.01	24102.14
October	20579.16	21621.29	22663.43	23705.56	24747.7
November	19949.75	20991.89	22034.02	23076.16	24118.29
December	19720.36	20762.49	21804.63	22846.76	23888.89

Table 3: Forecasted values of violence against a person

These forecasted values can also be plotted on a graph along with the historical data on RStudio, which looks like this (figure 7)

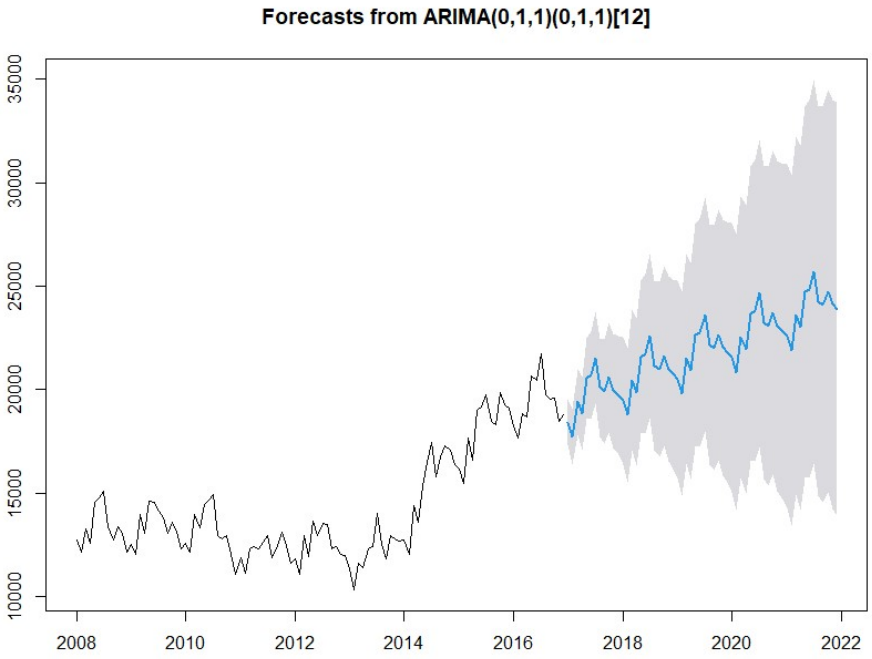

Figure 7: Time series plot of forecasted violence against a person

The grey shaded region indicates the range of the values. We can say with 95% confidence that the future value will definitely lie in the grey area. From this graph plot, it can be observed that seasonality persists. The changes seem to be consistent with different months of the calendar.

5.1.2 Future incidents of theft and handling

These (table 4) are the results for the crime category of theft and handling from January 2017 to December 2021

	2017	**2018**	**2019**	**2020**	**2021**
January	24102.75	24135.59	23668.9	23974.86	24049.55
February	23392.56	23013.06	22888.69	23185.4	23128.2

March	25042.86	25194.78	24582.04	24897.12	25018.91
April	24228.47	24216.52	23995.94	24182.26	24209.97
May	25687.29	25777.59	25792.55	25821.31	25829.86
June	25286.99	25403.73	25294.59	25366.75	25400.95
July	25846.62	25642.83	25890.15	25912.87	25829.69
August	24883.59	24384.24	24878.32	24894.39	24716.4
September	24924.28	24359.98	24789.45	24857.36	24678.25
October	26230.64	25862.08	26187.73	26224.81	26099.29
November	25407.6	25423.67	25519.93	25512.55	25496.17
December	25308.09	25082.17	25457.94	25408.66	25301.93

Table 4: Forecasted values of theft and handling

The graph plot of these forecasted values looks like this (figure 8)

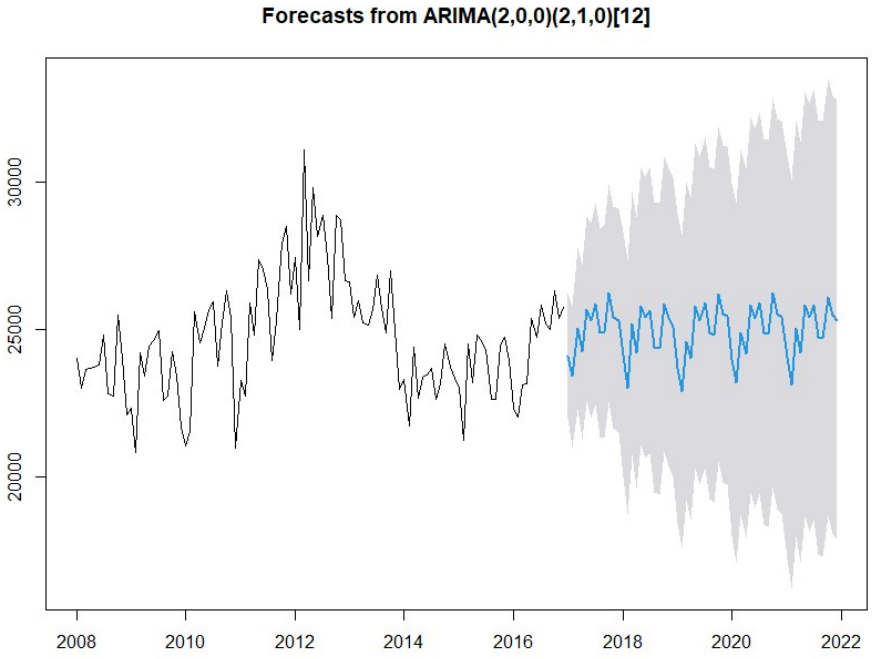

Figure 8: Time series plot of forecasted theft and handling

We can observe that there is not any presence of a trend in the forecasted data. Seasonality persists, however.

5.1.3 Future incidents of robbery

These (table 5) are the results for the crime category of robbery from January 2017 to December 2021

	2017	**2018**	**2019**	**2020**	**2021**
January	2172.206	2243.992	2302.189	2349.369	2387.618
February	2053.19	2147.505	2223.966	2285.954	2336.208

March	2133.072	2212.266	2276.468	2328.517	2370.714
April	2041.453	2137.99	2216.252	2279.7	2331.138
May	2146.968	2223.532	2285.601	2335.922	2376.717
June	2148.774	2224.995	2286.788	2336.884	2377.497
July	2185.605	2254.854	2310.995	2356.508	2393.406
August	2154.461	2229.606	2290.526	2339.914	2379.953
September	2166.561	2239.415	2298.479	2346.361	2385.18
October	2225.503	2287.2	2337.218	2377.767	2410.641
November	2212.417	2276.591	2328.617	2370.794	2404.988
December	2256.768	2312.547	2357.766	2394.426	2424.146

Table 5: Forecasted values of robberies

The graph plot of these forecasted values looks like this (figure 9)

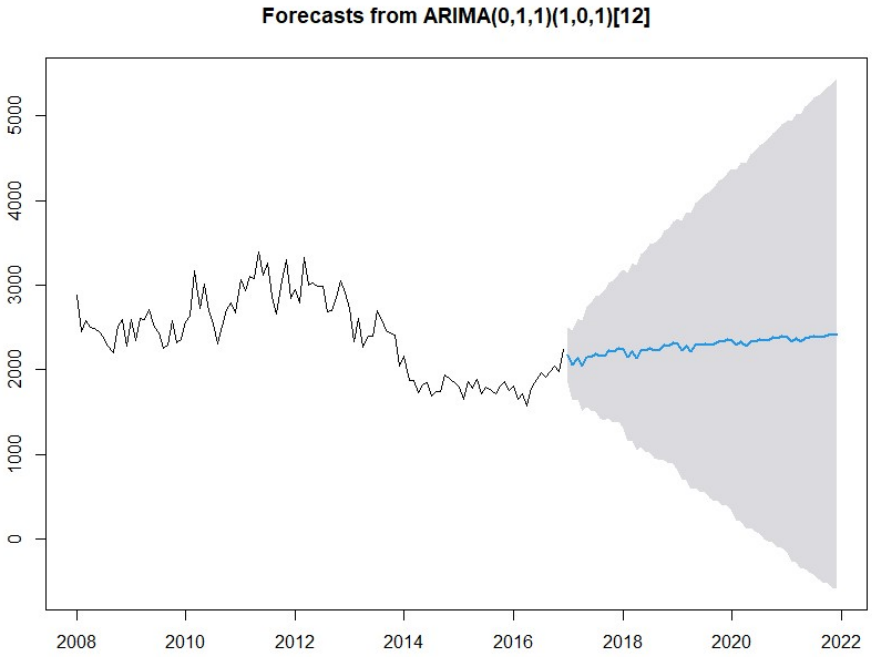

Figure 9: Time series plot of forecasted robbery

The grey area of this forecasting very wide. The low variation in the historical data values can be one of the reasons why this model is not effective in predicting the future values with confidence.

5.1.4 Future incidents of Drugs related crimes

These (table 6) are the results for the crime category of drugs from January 2017 to December 2021

	2017	2018	2019	2020	2021
January	3280.317	3201.261	2924.292	2802.647	2637.467

February	2994.489	2914.533	2704.258	2602.709	2474.476
March	2942.054	2864.732	2664.754	2567.586	2445.45
April	2879.005	2875.329	2638.927	2564.532	2432.597
May	2855.788	2816.06	2609.99	2528.291	2408.103
June	2860.1	2708.975	2579.048	2469.375	2375.711
July	3062.919	2924.044	2738.751	2617.703	2494.999
August	3077.064	2960.686	2756.545	2640.079	2510.089
September	2928.528	2753.084	2624.18	2503.603	2407.057
October	3080.914	2852.076	2724.921	2580.25	2477.101
November	2807.051	2624.847	2528.704	2415.084	2335.79
December	2645.287	2465.307	2405.017	2303.448	2244.4

Table 6: Forecasted values of drug related crimes

The graph plot of these forecasted values looks like this (figure 10)

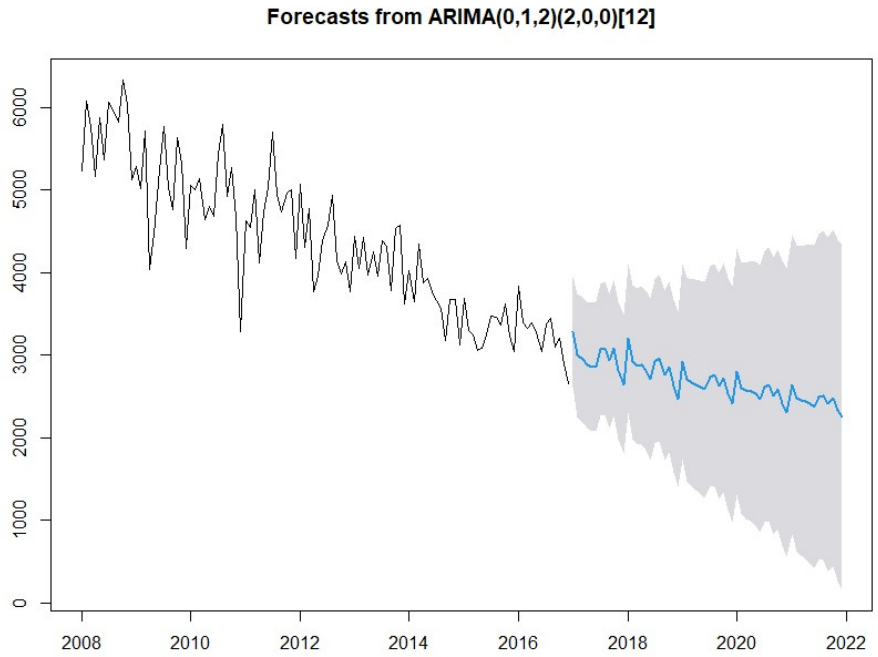

Figure 10: Time series plot of forecasted drug related crimes

A clear downward trend can be easily observed in the forecasted data. There exists a slight degree of seasonality as well.

5.1.5 Future incidents of Criminal Damage

These (table 7) are the results for the crime category of criminal damage from January 2017 to December 2021

	2017	2018	2019	2020	2021
January	5052.202	4947.691	4826.788	4705.886	4584.984

February	4907.385	4763.502	4642.6	4521.698	4400.795
March	5247.834	5226.89	5105.988	4985.085	4864.183
April	5299.16	5245.025	5124.123	5003.221	4882.318
May	5791.252	5625.278	5504.376	5383.474	5262.572
June	5426.302	5373.415	5252.513	5131.611	5010.709
July	5649.843	5515.238	5394.336	5273.434	5152.532
August	5472.089	5316.628	5195.726	5074.824	4953.922
September	5135.633	4986.629	4865.726	4744.824	4623.922
October	5377.56	5247.241	5126.339	5005.437	4884.535
November	5132.206	5075.815	4954.913	4834.011	4713.109
December	5040.57	4894.934	4774.032	4653.13	4532.228

Table 7: Forecasted values of criminal damage

The graph plot of these forecasted values looks like this (figure 11)

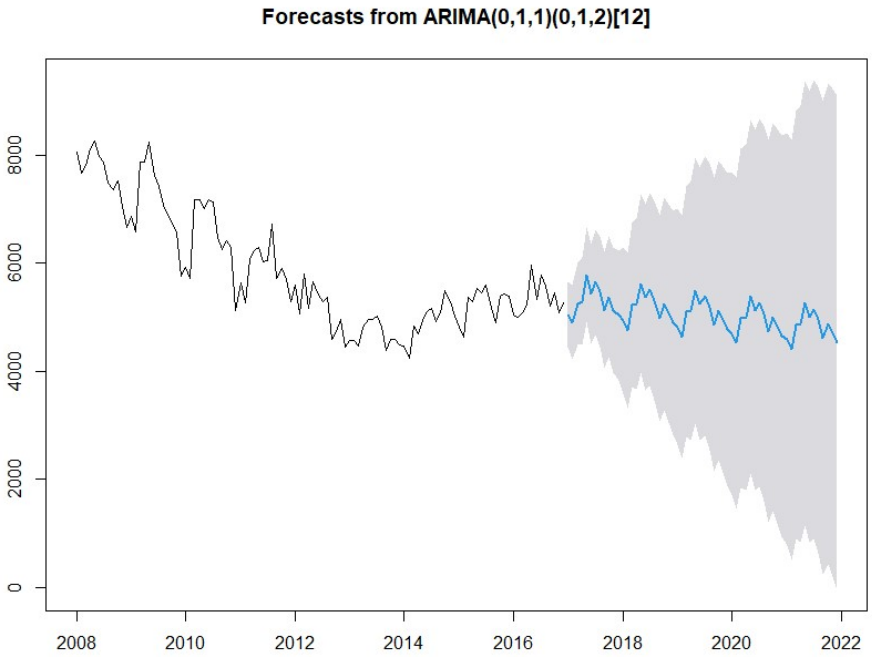

Figure 11: Time series plot of forecasted criminal damage

No trend can be seen as the forecasted values plot lie in-between 4000 and 6000 after December 2017. Seasonality does persist for the forecasting of this particular crime.

5.1.6 Future values of burglary

These (table 8) are the results for the crime category of burglary from January 2017 to December 2021

	2017	2018	2019	2020	2021
January	6356.018	5560.339	5261.255	5050.631	4500.221

February	5502.558	4958.252	4683.286	4316.382	3871.628
March	5544.671	5159.118	4719.971	4390.856	4007.443
April	4797.934	4387.72	4119.733	3673.941	3285.371
May	4796.962	4469.061	4176.849	3704.319	3349.383
June	4492.226	4243.122	4045.559	3457.788	3138.585
July	4687.3	4351.974	4116.256	3604.823	3248.482
August	4791.262	4365.512	4151.963	3673.282	3279.861
September	4701.943	4343.216	4058.552	3597.195	3229.617
October	5252.159	4802.633	4392.131	4076.328	3667.07
November	5786.086	5517.706	5000.715	4666.525	4329.687
December	5808.582	5458.827	5286.608	4734.786	4374.328

Table 8: Forecasted values of burglaries

The graph plot of these forecasted values looks like this (figure 12)

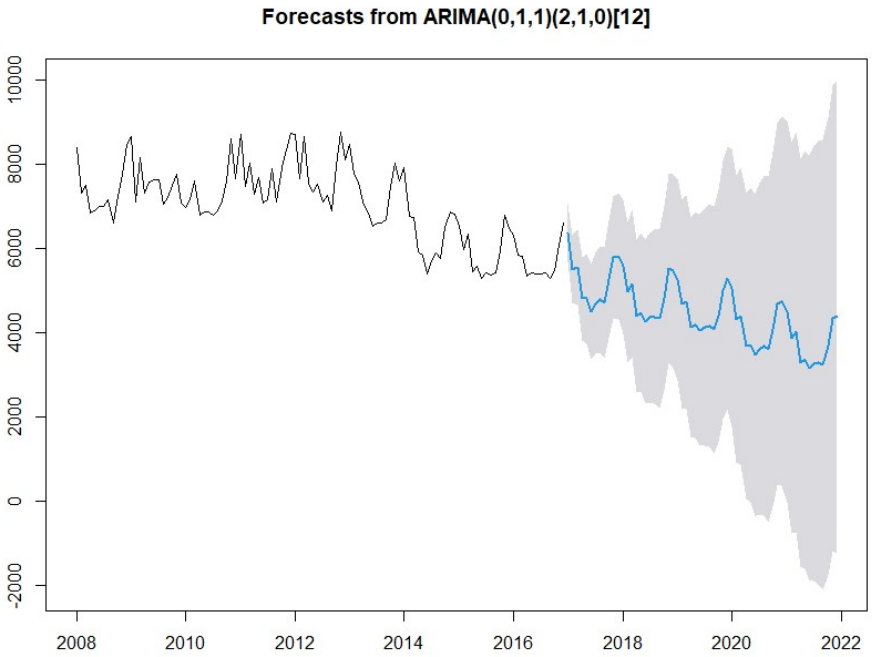

Figure 12: Time series plot of forecasted burglary

A constant and significant downward trend can be observed in the forecast plot of this type of crime. Seasonality also persists in the forecasted values.

5.2 Predict the number of houses sold

All the above mentioned forecasted values are used as input in the regression model discussed in Chapter 4.3, which will help us predict the number of houses sold by using the mathematical relationship between crime and the number of houses sold.

The sum product function in MS Excel is used to calculate the future values of the number of houses sold. These are the results (table 9)

	2017	2018	2019	2020	2021
January	61256.24	63855.48	64468.6	66386.72	68744.55
February	62600.23	63918.1	64932.1	67084.67	68932.75
March	64836.94	66315.09	67062.88	69253.18	71322.03
April	65217.12	66552.93	67350.84	69531.97	71369.3
May	66517.42	68368.91	69775.99	71880.65	73722.11
June	68066.83	69228.02	70323.34	72725.98	74561.21
July	68503.94	69737.23	71459.17	73636.83	75369.75
August	66190.63	67219	69209.88	71273.06	72904.57
September	67216.78	67912.37	70076.71	72189.92	73814.03
October	67461.12	68635.72	71008.94	72776.82	74625.51
November	64957.37	66087.1	68365.45	70130.36	72002.85
December	64238.22	65579.7	67614.93	69876.74	71727.71

Table 9: Forecasted values of the number of houses sold

All the historical data plotted along with this predicted data plotted on a graph looks like this (figure 13)

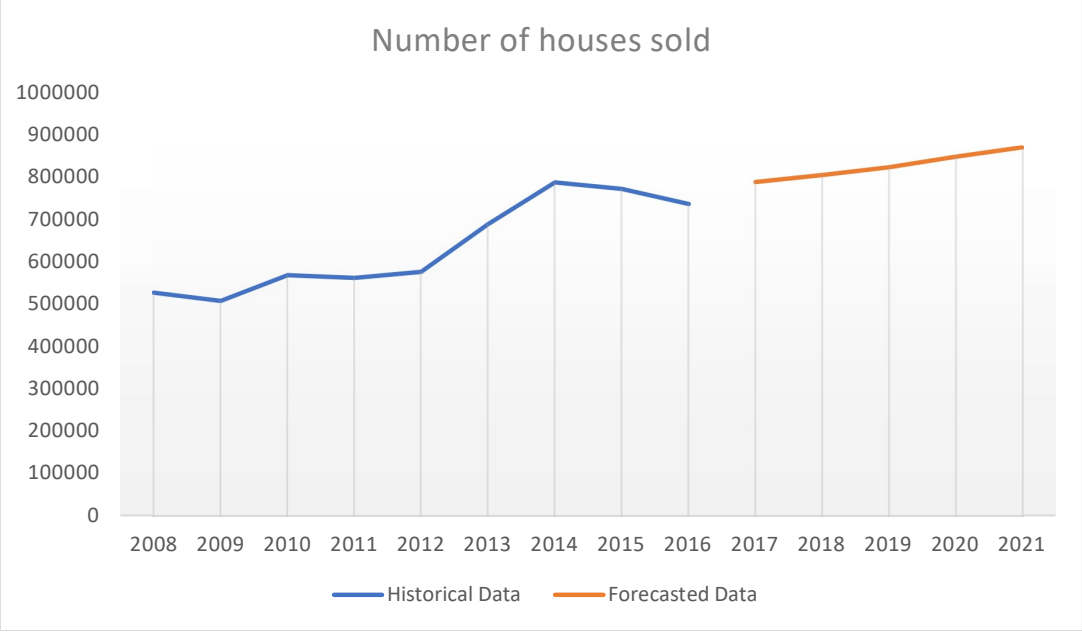

Figure 13: Time series plot of historical and forecasted number of houses sold

This graph shows the total number of houses sold in a year. The historical data is till 2016 and from 2017 till 2021 is the values we have calculated above using our regression model. The forecasted values show a positive linear trend over the 5 years.

It is not wise to use these results to make decisions. This paper takes a new approach to analyzing the impact of crime on the housing market, and at least until these models are backed up by a larger database or until the model is developed further by adding more controls, these results must be taken with a pinch of salt.

6. Limitations, future work and contribution to knowledge

6.1 Limitations

6.1.1 Underreporting of crimes

The whole paper is based on the fact that every single crime committed is reported, which is not usually the case. There are several crimes that go unreported. Data reporting is not always up to the mark and any faults in the database can cause the conclusions to be wrong.

Also, some neighborhoods are just notorious for street crime, irrespective of the count of crimes that have occurred there. This paper does not account for such instances.

6.1.2 Limited Data

Although the housing dataset was for a long period of time, the crime dataset was only for the 8 years between 2008 and 2016. In order to make both the datasets homogeneous, they were converted into monthly data from January 2008 to December 2016.

This is a significantly small database to construct regression models and forecasting models upon. As the historical data gets bigger, the accuracy of the model is increased and better will be the results derived.

6.1.3 Old data

The data we have ends in the December month of 2016. Even if the prediction model was built to forecast values for the next 5 years i.e., till 2021, it is still a year less than the year in which this paper is being submitted.

When the correct data after 2016 is gathered, it can be added to the current regression and prediction models to increase its efficacy and accuracy in analyzing the relationship between crime and house sales and in forecasting future values respectively.

6.1.4 Low accuracy

The accuracy of the regression model is measured by the value of R^2, which is pretty low for the regression model in this paper.

It can be attributed to the fact that the database upon which it was built was too small, a problem that can be solved in the future once more data is generated. Also, if cost

of crime is employed into the regression model, feature selection can be done to pick out the factors that are the most significant and remove the factors that do not have much impact.

6.1.5 Null values

Two of the types of crimes – sexual offences and fraud and forgery, had only null values and have hence been omitted from this research. Sexual offences are a serious crime and will definitely have some impact on the housing market. Upon acquiring this data, it would be beneficial to add it to the regression model to enhance its accuracy.

6.2 Future work

Given that this paper adopted a fairly new approach, to consider the number of houses sold rather than the average price of houses as a metric, there is plenty of scope for future research. In fact, both these approaches can be combined as well to open an entirely new avenue of research.

Based on the literature review, some of the techniques adopted by a few papers are inspired. Implying them with this approach would give us some unique insights and more reliable results. For example, cost of crime can be calculated for each type of crime and weights can be assigned based on this cost. The higher the cost of crime, the higher will be the weight assigned. In this way, all crime types are not treated equally, and the more severe crimes are given more importance. This will also help in getting rid of the independent variables that are not significant from the regression model.

Another concept observed through literature review was spatial analysis. I believe there is a lot of scope for spatial analysis in this topic. We can implement details like nearest police station, distance of crime from the residential areas, etc. Implementing spatial analysis can also help us determine which parts of the city have the highest demand from the historical data of house sales.

Methods other than ARIMA such as exponential smoothing or simple moving average can also be used to predict the future values. And then, the model with the highest level of accuracy can be picked and developed to achieve the required results. Although, ARIMA is proven to be the best option most times, there is no harm in checking other methods as well.

There is also another unique observation made by Tita, Petras and Greenbaum (2006) in their research study. They claim that a certain type of crime has a different impact on different neighborhoods of the city. They justify this finding by mentioning the higher options of the high income group to respond to a crime, such as private security and alarm systems. This can also be done by using the approach used in the current paper and by employing spatial analysis.

6.3 Contribution to knowledge

This study documents research that has been conducted on the impact of crime on the housing market. The demand of the housing market is taken as a metric to measure

this impact, which bridges the knowledge gap in this field as most research papers take the average prices of the houses sold as the metric.

This study would benefit governments by aiding their decision making process for policies. By predicting future house sales, they can ensure that that predicted demand can be met by increasing the supply. In the housing market, there is usually a time lag between the rise of demand and the actual supply. By predicting the future sales early, this problem can be omitted. In addition to that, the literature review explaining about the patterns in crime and what the motives behind crime are can also aid the policy making process.

This study would also benefit police departments. By predicting future crime rates, they can allocate their resources accordingly. If the results say that a certain violent crime is going to increase in the near future, more experienced officers can be assigned to deal with it and volunteers or part time workers can be employed for less priority work. By constantly predicting future crime rates, the police department can ensure that the challenges are not too overwhelming. If future work is done by using spatial analysis, the police department can also plan patrolling routes and position themselves accordingly in order to be better equipped to face any kinds of threats.

Lastly, investors would also be benefitted by this study. As mentioned before, the real estate market attracts a lot of investors. The only motive of investors is to maximize their returns. They can use this model to predict the future demand for real estate in a particular region and use the results to aid their decision making process about buying a house.

7. Summary

The objective of this paper was to determine the impact of crime on the housing market in London. In order to achieve this, crime data from January 2008 to December 2016, which has been released by the metropolitan police service and housing data released by the London Datastore has been used. The data was modified to be homogeneous i.e., converting it into monthly data for the same time period.

Unlike many other research studies in this field, this paper did not use the average prices of houses as a metric to gauge the impact of crime. Instead, it uses the number of houses sold as metric to measure it.

The crime dataset included different types of crime. Few of these different categories, such as sexual offences and fraud and forgery, only had null values and have hence been omitted from the research. Multicollinearity was discovered using the variance inflation factor in MS Excel. It was observed that the category ‘Other notifiable offences’ had a very high degree of collinearity with the other types of crimes and has hence been removed from further research.

A correlation matrix was made to check the correlation between different types of crime and the number of houses sold in some of the boroughs of London. Some of the values of this matrix, however, were opposite to what was expected. The crime rate had a positive correlation with the number of houses sold in the same region. This can be attributed to the fact that the data was limited.

Later, the different categories of crimes were converted into time series data and were plotted on a graph using RStudio. The different components of time series data i.e., trend and seasonality, were analyzed and discussed for each category of crime.

To achieve the first objective of this research paper, the regression model has been built using the historical data from January 2008 to December 2016. The regression equation is

$$y = 52585.5 + (x_1)\ 1.05 + (x_2)\ 1.56 + (x_3)\ (-9.03) + (x_4)\ 1.26 + (x_5)\ (-3.65) + (x_6)\ (-2.28)$$

Here, the values of x_1, x_2 and so on are values of different categories of the crime. The second objective of this research paper was to predict the future house sales, and this is achieved by substituting those x_1, x_2 and so on values with future predicted values of crime rates.

After the relationship between the different types of crimes and the number of houses sold is mathematically defined, the former is predicted using a forecasting model. This forecasting model was built in RStudio using the ARIMA method. Six different models were built, one for each category of crime. After eliminating sexual offences, fraud and forgery, and other notifiable offences for the reasons stated above, we are left with burglary, robbery, theft and handling, drugs, violence against a person and criminal damage.

All of these six categories of crime are predicted for the next 5 years from 2016 i.e., till December 2021. The six models are the best possible ARIMA models because auto ARIMA function has been used in RStudio. This function builds several models and then compares the AIC values, which is the accuracy values, of all these models and selects the best one. Hence, using these models, all the future values of the six different categories of crime are predicted. These values are monthly data from January 2017 till December 2021.

These values are then used as input in the regression equation mentioned above to predict the future values of the number of houses sold for all the months mentioned above.

This prediction not only helps governments and police departments to prepare for future trends and launch policies, but also aids the decision making process of an investor. Given the fairly new approach of using the number of houses sold as a measure of demand, the other techniques applied in the other approach can be replicated with this approach to derive new insights.

In conclusion, the housing market is a very important sector of the economy of a country. The government and the judiciary system should take appropriate measures to keep the crime rate under control so that the housing market can continue to thrive. This analysis will hopefully pave way for a new way of mathematically defining the relationship between crime and the housing market, which can be used to achieve the aforementioned objectives.

Appendix 1

R Code for forecasting model

```
library(forecast)

## Warning: package 'forecast' was built under R version 4.0.5

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

data <- read.csv("Final dataset.csv", stringsAsFactors = FALSE)
data$date = as.Date(data$Month, format="%m%d%Y")

#Crate ts datasets for each variable
ts_violence <- ts(data$Violence.Against.the.Person, start=c(2008,1), end =
c(2016,12), frequency = 12)
ts_theft <- ts(data$Theft.and.Handling, start=c(2008,1), end = c(2016,12),
frequency = 12)
ts_robbery <- ts(data$Robbery, start=c(2008,1), end = c(2016,12), frequenc
y = 12)
ts_drugs <- ts(data$Drugs, start=c(2008,1), end = c(2016,12), frequency =
12)
ts_criminaldamage<- ts(data$Criminal.Damage, start=c(2008,1), end = c(2016
,12), frequency = 12)
ts_burglary <- ts(data$Burglary, start=c(2008,1), end = c(2016,12), freque
ncy = 12)

#Plot the ts
plot(ts_violence, xlab="Year", ylab="Violence Against Person Count")

plot(ts_theft, xlab="Year", ylab="Theft Count")

plot(ts_robbery, xlab="Year", ylab="Robbery Count")

plot(ts_drugs, xlab="Year", ylab="Drugs Offences Count")

plot(ts_criminaldamage, xlab="Year", ylab=" Criminal Damage Count")

plot(ts_burglary, xlab="Year", ylab="Burglary Count")
```

```
#Build arima model
model1 <- auto.arima(ts_violence)
model2 <- auto.arima(ts_theft)
model3 <- auto.arima(ts_robbery)
model4 <- auto.arima(ts_drugs)
model5 <- auto.arima(ts_criminaldamage)
model6 <- auto.arima(ts_burglary)

#View model characteristics
model1
```

```
## Series: ts_violence
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##           ma1     sma1
##       -0.2707  -0.7236
## s.e.   0.0926   0.1200
##
## sigma^2 = 312638:  log likelihood = -739.26
## AIC=1484.52   AICc=1484.78   BIC=1492.18
```

```
model2
```

```
## Series: ts_theft
## ARIMA(2,0,0)(2,1,0)[12]
##
## Coefficients:
##          ar1     ar2     sar1     sar2
##       0.4915  0.4016  -0.4754  -0.3898
## s.e.  0.0952  0.0941   0.1012   0.0944
##
## sigma^2 = 1191955:  log likelihood = -808.99
## AIC=1627.97   AICc=1628.64   BIC=1640.79
```

```
model3
```

```
## Series: ts_robbery
## ARIMA(0,1,1)(1,0,1)[12]
##
## Coefficients:
##           ma1    sar1     sma1
##       -0.2661  0.8107  -0.4772
## s.e.   0.1163  0.1401   0.2192
##
## sigma^2 = 28550:  log likelihood = -701.32
## AIC=1410.65   AICc=1411.04   BIC=1421.34
```

```
model4
```

```
## Series: ts_drugs
## ARIMA(0,1,2)(2,0,0)[12]
##
## Coefficients:
##           ma1      ma2    sar1    sar2
##       -0.4875  -0.3300  0.3075  0.4613
## s.e.   0.0951   0.1042  0.0853  0.0953
##
## sigma^2 = 115329:  log likelihood = -779.02
## AIC=1568.03   AICc=1568.62   BIC=1581.39
```

```
model5
```

```
## Series: ts_criminaldamage
## ARIMA(0,1,1)(0,1,2)[12]
##
```

```
## Coefficients:
##           ma1     sma1     sma2
##       -0.4923  -0.4745  -0.2156
## s.e.   0.1029   0.1278   0.1346
##
## sigma^2 = 97014:  log likelihood = -682.1
## AIC=1372.2   AICc=1372.65   BIC=1382.42
```

```
model6
```

```
## Series: ts_burglary
## ARIMA(0,1,1)(2,1,0)[12]
##
## Coefficients:
##           ma1     sar1     sar2
##       -0.4206  -0.7607  -0.5487
## s.e.   0.1034   0.0984   0.0948
##
## sigma^2 = 124501:  log likelihood = -696.6
## AIC=1401.2   AICc=1401.64   BIC=1411.41
```

```
#Forecast future values for the next 5 years
forecast_violence <- forecast(model1, level = c(95), h=5*12)
forecast_theft <- forecast(model2, level = c(95), h=5*12)
forecast_robbery <- forecast(model3, level = c(95), h=5*12)
forecast_drugs <- forecast(model4, level = c(95), h=5*12)
forecast_criminaldamage <- forecast(model5, level = c(95), h=5*12)
forecast_burglary <- forecast(model6, level = c(95), h=5*12)


#Plot forecast
plot(forecast_violence)

plot(forecast_theft)

plot(forecast_robbery

plot(forecast_drugs)

plot(forecast_criminaldamage)

plot(forecast_burglary)
```

Referencing

Ahmad, F. and Ali, R., 2015. The motivation for crimes: Experiences of criminals from district jail Karak, Khyber Pakhtunkhwa, Pakistan. 7. Pp. 16-28

Amani, M., Boumeslout, S., Bencharif, M. and Belkourissat, M., 2022. Caractéristiques et particularités des homicides commis par des schizophrènes. *Annales Médico-psychologiques, revue psychiatrique*, 180(6), pp.S66-S74.

Bank of England., 2020., *How does the housing market affect the economy?* [Online]. Available from: https://www.bankofengland.co.uk/knowledgebank/how-does-the-housing-market-affect-the-economy#:~:text=The%20housing%20market%20is%20closely,or%20pay%20off%20other%20debt.
[Accessed on 20/08/22]

Bevans, R., 2020. *Multiple Linear Regression – A Quick Guide (Examples)* [Online]. Scribbr. Available from: https://www.scribbr.com/statistics/multiple-linear-regression/
[Accessed on 29/07/22]

Boysen, J., 2017. *London Crime Data, 2008-2016* [Online]. Kaggle. Available from: https://www.kaggle.com/datasets/jboysen/london-crime

Braun, M., Altan, H. and Beck, S., 2014. Using regression analysis to predict the future energy consumption of a supermarket in the UK. *Applied Energy*, 130, pp.305-313.

Brownlee, J., 2017. *How to Remove Trends and Seasonality with a Difference Transform in Python* [Online]. Machine Learning Mastery. Available from: https://machinelearningmastery.com/remove-trends-seasonality-difference-transform-python/#:~:text=Differencing%20is%20a%20method%20of,structures%20like%20trends%20and%20seasonality.
[Accessed on 28/07/22]

Chen, P., Yuan, H. and Shu, X., 2008. Forecasting Crime Using the ARIMA Model. *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*.

Cirtautas, J., 2020. *Housing in London* [Online]. Kaggle. Available from: https://www.kaggle.com/datasets/justinas/housing-in-london

Crown Prosecution Service. *Drug Offences* [Online]. Available from: https://www.cps.gov.uk/crime-info/drug-offences
[Accessed on 28/09/22]

De La Paz, P.T., Berry, J., McIlhatton, D., Chapman, D. and Bergonzoli, K., 2022. The impact of crimes on house prices in LA County. *Journal of European Real Estate Research*, 15(1), pp. 88-111.

Detotto, C. and Otranto, E., 2010. Does Crime Affect Economic Growth?. *Kyklos*, 63(3), pp.330-345.

Frost, J., 2017. *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions* [Online]. Statistics By Jim. Available from: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ [Accessed on 30/07/22]

Gallo, A., 2015. *A Refresher on Regression Analysis* [Online]. Harvard Business Review. Available from: https://hbr.org/2015/11/a-refresher-on-regression-analysis [Accessed on 01/08/22]

Glen, S., n.d. *Correlation in Statistics: Correlation Analysis Explained* [Online]. Statistics How To. Available from: https://www.statisticshowto.com/probability-and-statistics/correlation-analysis/ [Accessed on 01/08/22]

Hayes, A., 2021. *Autoregressive Integrated Moving Average (ARIMA)* [Online]. Investopedia. Available from: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical,values%20based%20on%20past%20values. [Accessed on 10/08/22]

Hayes, A., 2022. *Time Series Definition* [Online]. Investopedia. Available from: https://www.investopedia.com/terms/t/timeseries.asp [Accessed on 10/08/22]

Hayes, A., 2022. *Multiple Linear Regression (MLR)* [Online]. Investopedia. Available from: https://www.investopedia.com/terms/m/mlr.asp#:~:text=Key%20Takeaways-,Multiple%20linear%20regression%20(MLR)%2C%20also%20known%20simply%20as%20multiple,uses%20just%20one%20explanatory%20variable. [Accessed on 27/08/22]

Hoyt, W., Leierer, S. and Millington, M., 2006. Analysis and Interpretation of Findings Using Multiple Regression Techniques. *Rehabilitation Counseling Bulletin*, 49(4), pp.223-233.

Latif, N. S. A., Rizwan, K. M., Rozzani, N. and Saleh, S. K., 2020. Factors Affecting Housing Prices in Malaysia: A Literature Review. *International Journal of Asian Social Science*, 10(1), pp. 63-67.

Laver, N., n.d. *Criminal Damage* [Online]. In Brief. Available from: https://www.inbrief.co.uk/offences/criminal-damage/ [Accessed on 26/08/22]

Lynch, A. and Rasmussen, D., 2001. Measuring the impact of crime on house prices. *Applied Economics*, 33(15), pp.1981-1989.

Mayo Clinic Staff, 2020., *Schizophrenia* [Online]. Mayo Clinic. Available from: https://www.mayoclinic.org/diseases-conditions/schizophrenia/symptoms-

causes/syc-20354443#:~:text=Schizophrenia%20is%20a%20serious%20mental,with%20schizophrenia%20require%20lifelong%20treatment.
[Accessed on 02/09/22]

McCue, C., 2007. *Data Mining and Predictive Analysis* [Online]. ScienceDirect. Available from: https://www.sciencedirect.com/topics/computer-science/continuous-variable
[Accessed on 27/08/22]

McIlhatton, D. et al., 2016. Impact of crime on spatial analysis of house prices: evidence from a UK city. *International journal of housing markets and analysis*., 9(4), pp.627-647.

Metropolitan Police Service, 2017. Recorded Crime: Geographic Breakdown. *London Datastore* [Online]. London: Metropolitan Police Service. Available from: https://data.london.gov.uk/dataset/recorded_crime_summary
[30/07/22]

Minghetti, A., 2020. *Exploring why crime and house prices correlate positively in London.* Explore Econ Undergraduate Research Conference (BSc Economics). University College London.

Morrall, P., 2006. Murder and society: why commit murder?. *Criminal Justice Matters*, 66(1), pp.36-37.

Nau, R., n.d. *Statistical forecasting: notes on regression and time series analysis* [Online]. Available from: https://people.duke.edu/~rnau/411arim.htm#:~:text=A%20nonseasonal%20ARIMA%20model%20is,errors%20in%20the%20prediction%20equation.
[01/08/22]

Nguyen, J., 2021. *4 Key Factors That Drive the Real Estate Market* [Online]. Investopedia. Available from: https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp
[Accessed on 28/08/22]

Nielsen, B., 2022. *How Interest Rates Affect the Housing Market* [Online]. Investopedia. Available from: https://www.investopedia.com/mortgage/mortgage-rates/housing-market/
[Accessed on 28/08/22]

Palachy, S., 2019. *Stationarity in time series analysis* [Online]. Towards Data Science. Available from: https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322
[Accessed 29/08/22]

Scott, L., 1990. Do prices reflect market fundamentals in real estate markets?. *The Journal of Real Estate Finance and Economics*, 3(1), pp.5-23.

Sentencing Council. *What is the difference between theft, robbery and burglary?* [Online]. Available from: https://www.sentencingcouncil.org.uk/blog/post/what-is-the-difference-between-theft-robbery-and-burglary/
[Accessed on 28/08/22]

Tran, Q. N., Arabnia, H., 2015., *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*. Elsevier.

Xu, K., 2018. *Relationships Between Air Pollution and Weather Conditions/Living Habits, and Future Air Pollutant Emissions Forecasting, Evidence from: China, South Korea and India*. Dissertation (MSc Business Analytics). University of Bath School of Management

Zeng, Z., 2021., *Time Series Model and ARIMA-Genetic Algorithm for Forecasting Rice Production in Asia.* Dissertation (MSc Business Analytics). University of Bath School of Management.

Zhu, M., 2014. *Housing Markets, Financial Stability, and the Economy* [Online]. International Monetary Fund. Available from: https://www.imf.org/en/News/Articles/2015/09/28/04/53/sp060514
[Accessed 28/08/22]

*****