# Explore the Relationship between House Prices and Crime Rate in the UK Using Machine Learning Techniques

Soudabeh Motevali
Department of Computer Science and Creative Technologies
The University of the West of England, UK
motevali.soudabeh@gmail.com

Hamzeh Aljawawdeh
Department of Software Engineering
Zarqa University, Jordan
hamzeh.aljawawdeh@zu.edu.jo

Sami Abuezhayeh
Department of Computer Science and Creative Technologies
The University of the West of England, UK
sami.abuezhayeh@uwe.ac.uk

Eman Qaddoumi
Department of Computer Science and Creative Technologies
The University of the West of England, UK
eman3.qaddoumi@uwe.ac.uk

**Abstract:** *In the UK, house price estimation is an important topic. It has been discussed in numerous academic research publications and official and business reports. A house's price can change depending on size, age, and location. One of the main characteristics of a property's neighbourhood is the crime rate; it is crucial to consider how it may affect the home's price. If crime rates rise, a neighbourhood may lose appeal to potential purchasers in favour of comparable-priced neighbourhoods. This study employs data from January 2020 to April 2022 to examine the relationship between crime rates and house prices in Bristol, UK. Crime datasets were taken from the UK Crime Stats website, while data on house prices was taken from the Price Paid Data of the HM Land Registry. 34,000 transaction records and criminal statistics are combined in the study to provide a comprehensive dataset for analysis. Two tree-based machine learning algorithms-Decision Tree (DT) and Random Forest (RF)-and Exploratory Data Analysis (EDA) were used to model and evaluate the data. The results show that the Random Forest outperformed the Decision Tree regarding prediction accuracy. According to the results, there is a substantial correlation between crime rates and house prices. While vehicle crimes and bike theft had a positive correlation with property values, violent crimes had the opposite effect. These revelations highlight the intricate relationship between different kinds of crime and the characteristics of the housing market, with considerable consequences for investors, policymakers, and real estate developers.*

**Keywords:** *Machine learning techniques, decision trees, random forest, house prices, crime rates, house price pre-diction.*

*Received August 28, 2024; accepted January 21, 2025*
*https://doi.org/10.34028/iajit/22/3/1*

## 1. Introduction

The value of houses can vary depending on characteristics, including location, size, age, and various other elements, including economic considerations. One of the most crucial elements affecting a house price is its location, which includes accessibility to public amenities like parks, schools, the police, hospitals, etc. The degree of security in a place is one of these elements, and it is typically stated by calculating the local crime rate. There are numerous reasons why the relationship between property values and crime rates should be analysed. There are challenges associated with its estimation using data from the real world. Policymakers should consider the direction and strength of this relationship for at least two additional reasons. First, it may impact municipal budgeting choices for crime prevention because increased tax revenues from higher property values may partially offset the costs of lowering crime in a particular location. Second, a localized group of disadvantaged residents can have their conditions improved by a policymaker by increasing crime reduction efforts in their neighbourhood. This could result in asset appreciation and higher well-being for the group in question [21]. Based on these facts, this study will look at house price estimation and how the frequency of various sorts of crime affects the price of a house.

The inspiration for this study originates from the need to address the gap in existing literature about the specific influence of various types of crime on house values. Despite extensive research on general house price estimation, this subject has received little attention. Accurate house price valuation techniques are critical due to the enormous financial consequences of property transactions for numerous parties. Furthermore, the societal impact of crime on community well-being and property values emphasises the significance of this research, which seeks to provide insights that can aid in crime reduction and community development.

Furthermore, recognising the economic impact of crime on property prices can have far- reaching consequences. Communities with dropping property values due to high crime rates may experience economic downturns, hurting local businesses and the economy. This study highlights these relationships and enhances economic stability through educated decision-making.
The research questions include:

- What is the extent of the impact of different types of crime on house prices in Bristol?
- How effective are machine learning algorithms in predicting house prices when considering crime data? Addressing these questions aims to provide valuable insights into the dynamics between crime and property prices, which can inform urban planning efforts and policy interventions.

This study uses machine learning methods to address several gaps in the relationship between crime rates and housing values. Here are the research gaps addressed by this work:

1. Limited uses powerful machine learning algorithms to forecast housing prices based on crime rates. This study uses two machine learning algorithms, Decision Tree and Random Forest, to estimate property prices and compare their performance and accuracy [18, 23].
2. A lacks datasets that combine precise crime statistics and home values over a long period. The creation of a new dataset that combines crime data from 28 months with house price data for Bristol, resulting in a comprehensive dataset for analysis [36].
3. Inadequate exploratory data analysis relating specific crime to housing values. Exploratory data analysis, skewness analysis, and data normalization are being carried out to provide deep insights into how various crimes (violent crimes, vehicle crimes, bike theft) correlate with property values [29].
4. A lack of awareness about which crime-related features impact property price predictions most. Using the feature importance technique to discover and rank the features with the highest influence on house price projections improves model interpretability.

Overall, this study adds to the current body of knowledge by utilizing advanced machine learning techniques, developing and analyzing a new dataset, and providing practical insights into the relationship between crime rates and house values. By filling these gaps, this study improves our understanding of how crimes affect housing markets and provides valuable tools for stakeholders to make informed decisions.

Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

### 1.1. Overview of the Study Area

The focus of this study is on the Bristol region. Bristol, England, is widely regarded as one of the most expensive cities in the United Kingdom. Bristol has a higher cost of living than the national average, with housing costs being a key contributor. The city's rent may rise due to increased demand for rental housing. The anticipated house price study in Bristol, UK, could be significant for many reasons:

1. Bristol is a well-known British city with a booming economy and a growing population. Because of the city's high demand for real estate, it's a great place to look into house price estimates.
2. The city has recently experienced tremendous economic growth, which has resulted in increasing real estate values. Everyone, including investors, residents, and government authorities, can profit by making educated judgements.
3. A range of factors, including regional demographics, transit networks, and employment prospects, influence Bristol's distinct property market.

As a result, it's a great case study for figuring out how various factors interact with the housing market. Bristol is one of the most appealing places in the UK for property purchasers, thanks to its strong economic growth, appealing lifestyle offers, and bustling business scene. Over the previous decade, average asking prices in Bristol have grown by 60%, while homeowners in Bristol's Easton neighbourhood have seen the UK's most significant increase in property prices - up 120% [42].

Bristol's map based on the postcode district is illustrated in Figure 1.

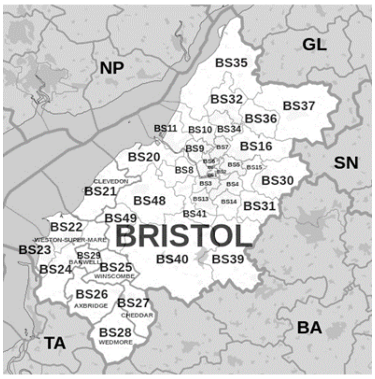


Figure 1. Postcode district in Bristol.

## 2. Materials and Methods

### 2.1. Housing Market Prices

Many countries rely on the housing industry as a growth engine and force generator to address unemployment and economic crises [33]. On the other hand, economic instability might impair the effectiveness of the housing market. In other words, the economy and housing are two interdependent parts. It is referred to as the housing market when people purchase or sell houses as a primary residence or investment. The most valuable possession many people will ever own is their house.

In the UK, two-thirds of families live in a home they own, and one-third rent [6].

Home price changes have been significant over time. In 1977, the average cost of a house was around £10,000. In 40 years, the average price has increased to £200,000 [6].

Some factors can affect a house's prices, value, and evaluation. The three categories that separate the elements influencing home values are location, structural, and neighbourhood circumstances [1].

Numerous studies have delved into the influential factors affecting housing prices, particularly emphasizing how important neighbours influence house prices. Although accessibility parameters, such as the adjusted Walk Score, are important, neighbour types and neighbourhood features emerge as critical considerations [41].

Osmadi *et al*. [24] and Zulkifly *et al*. [44] noted the importance of location factors in determining home prices in their studies. The property's location was classified as a fixed geographical attribute. These studies show a strong correlation between geographical factors, such as the distance to the nearest shopping mall or a location with views of hills or the ocean, and home prices, followed by community and economic factors.

The crime rate is one neighbourhood factor that has yet to be thoroughly researched about home values [44]. So, we will investigate the relationship between crime rate and property price.

## 2.2. Crime Rates

The number of crimes reported to law enforcement authorities per 100,000 people in a population is known as the "Crime Rate". The crime rate is determined by dividing the total number of reported offences by the population. Some evidence indicates that property crime and home values are correlated. According to Tita *et al*. [32], whereas total crime has no influence on home prices, violent crime significantly negatively impacts property values.

Recent research has shown interest in the relationship between crime rates and home prices. Bamiteko and Adebiyi [5] investigated how neighbourhood security impacted housing costs, which is directly tied to crime rates. They discovered that high crime rates had a detrimental effect on home values in some neighbourhoods. Furthermore, Zhang [43] created a decision tree-based system for predicting home prices, considering several variables, including the crime rate. They discovered that the number of dwellings, the local population's quality, and the crime rate were significant predictors of housing costs.

A study by Mcllhatton *et al*. [20] has shown how, depending on the type of crime, the type of property, and the geography, the effect of crime on pricing varies. For example, robbery and burglary are more prevalent in neighbourhoods with higher incomes. However, crimes like violence against persons, property dam- age, and drug offences are more prevalent in neighbourhoods with lower incomes. Braakmann [8] found that in England and Wales, each case of anti-social behaviour per ten residents in the same street leads to an approximately 0.6-0.8% drop in property prices. Crime outside of the respective street does not appear to matter.

## 2.3. Predication Models

Numerous machine learning forecasting models have been used to evaluate changes in housing costs in the UK and elsewhere while identifying the major driving factors. Figure 2 shows the different machine learning algorithms.

A systematic literature review conducted by Ja'afar *et al*. [15] on ma- chine learning techniques used in housing price predictions identified that the most popular method is supervised learning [4]. They also said that Random Forest, Decision Tree, Gradient Boosting, Neural Network, and Linear Regression are popular models used in prediction. The random forest, which can adapt well to different dataset situ datasets, is the best prediction model [15]. Compared to Gradient Boosting, Random Forest is superior for estimating rental costs [7].

Utilising 24,936 housing transaction records, [11] employs Extra Trees (ET), K-Nearest Neighbors (KNN), and Random Forest (RF) to predict property prices and then compares their results with those of a hedonic price model [31]. They suggest that these three algorithms outperform traditional statistical techniques regarding explanatory power and error minimisation.

Another model is Linear Regression (LR), widely used to forecast home prices [19]. LR is primarily used for forecasting, prediction, research discovery, and analysing the linear relationship between variables [7].

Advanced methodologies, such as those discussed by Aljawawdeh [3] in hybrid learning models, can enhance the analysis of this relationship by integrating traditional and modern techniques. Enriched datasets, similar approaches in Aljawawdeh and Nabot [2] coding education model, and compelling feature selection and data transformation methods, as highlighted by Takci and Nusrat [30], can improve the accuracy of house price estimations related to crime rates.

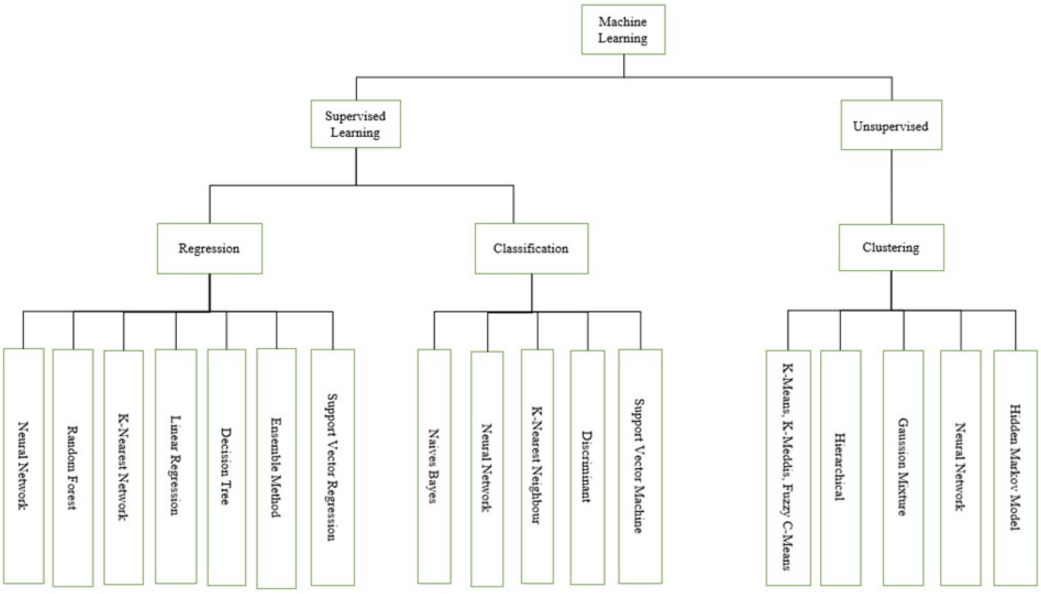


Figure 2. Type of machine learning algorithm.

## 2.4. Methods

Figure 3 shows the phases of the design science research methodology followed in this research.

A literature review has been conducted to identify the main problems related to house prices, outline the objectives, and reveal the methods used in the UK and other nations to forecast changes in home values.

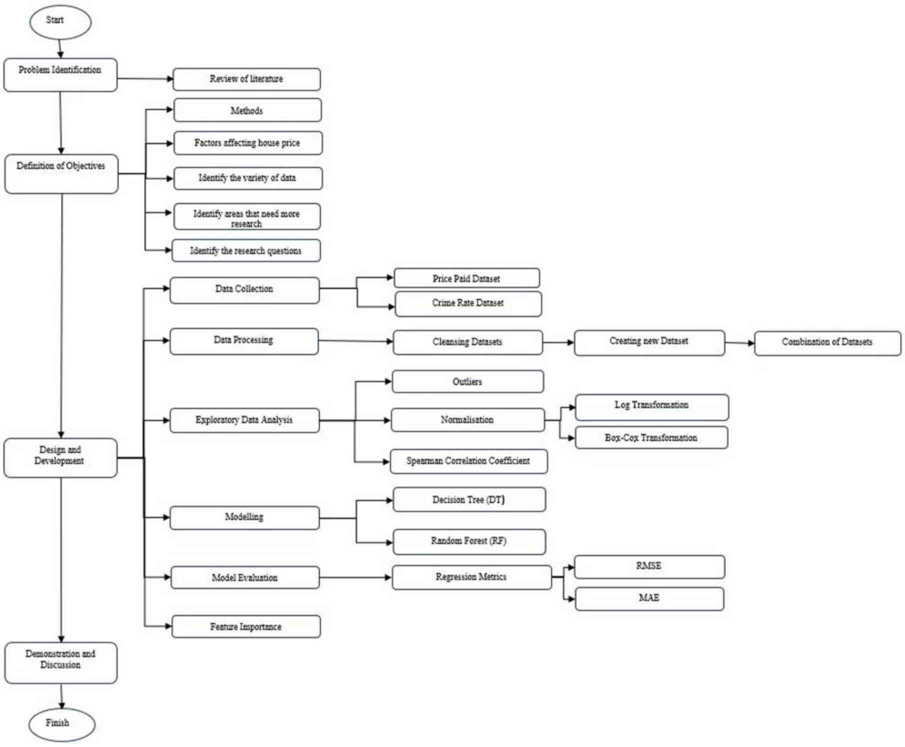


Figure 3. Methodology followed in this research.

The literature reveals that forecasting house prices has made significant strides, and predictive models have demonstrated some effectiveness in making precise projections of housing values. The use of machine learning in estimating housing prices with consideration for the potential impact of each factor is justified by the possibility that it will offer better solutions for improving the accuracy of predictions and assisting investors in identifying the best locations, prices, and factors that affect house prices.

According to a literature review, there is a gap in investigating the impact of the crime rate on house prices. This study uses machine learning to explore potential relationships between the crime rate and house price prediction. The neighbourhood is one of the most important factors influencing a home’s price, and the crime rate in each area is one of the indicators of a home's location.

This study chose the Decision Tree and Random Forest algorithms for their capacity to handle complex, non-linear relationships between variables and their robustness when dealing with big datasets. Decision trees provide a clear visualisation of decision-making processes, which can help stakeholders comprehend the model's predictions. Random Forest, an ensemble method, is well-known for its excellent ac- curacy and ability to avoid overfitting by combining numerous decision trees. This strategy was appropriate for the dataset, which had high variability and outliers.

## 3. Datasets

This project uses two datasets, and the following is a brief information regarding each of them:

### 3.1. Price Paid Data (from HM Land Registry)

HM Land Registry owns the "price paid data," which can be downloaded from the primary UK Government website [35]. More than 27 million records totalling 16 variables representing data on all property sales transactions in England and Wales from January 1, 1995, to the present day. The HM Land Registry Price Paid Data is described as "the official house price dataset in England." Around 34,000 entries from the Price Paid Data, which include all transactions in Bristol, UK, from January 1, 2020, to April 31, 2022, were used in this study.

### 3.2. Crime Rate Dataset (from UK Crime Stats)

Postcode-based monthly access to the dataset was used [34]. The dataset includes 12 different forms of crime: bur- glary, robbery, vehicle crime, violent crime, and anti-social behaviour. Other crimes include drug, shoplifting, criminal damage and arson, and theft. According to the hierarchy of possible punishment lengths, a collection of offences is thus listed as a single felony. The UK Crime Stats data description is presented in Table 1. Periodically, a report is posted on the police website, divided by areas and crime types. The Postcode districts with the most significant increases in "All Crime and ASB (AntiSocial Behaviour)" between January 2020 and April 2022 are shown in Table 1.

Table 1. Percentage change in crime rate over the period studied in Bristol.

| Postcode Districts with the largest increase in All crime & ASB (Between Jan 2020 and Apr 2022) | | | | | | |
|---|---|---|---|---|---|---|
| **Rank** | **Postcode Districts** | **Region** | **Total Jan 2020** | **Total Apr 2022** | **Difference** | **Percentage** |
| 233 | BS28 | Somerset | 10 | 14 | 4 | 40.00% |
| 431 | BS24 | North Somerset | 131 | 161 | 30 | 22.90% |
| 664 | BS27 | Somerset | 30 | 34 | 4 | 13.33% |
| 696 | BS48 | North Somerset | 90 | 101 | 11 | 12.22% |
| 714 | BS4 | Bristol | 424 | 473 | 49 | 11.56% |
| 785 | BS14 | Bristol | 221 | 242 | 21 | 9.50% |
| 835 | BS23 | North Somerset | 541 | 584 | 43 | 7.95% |
| 926 | BS34 | South Gloucestershire | 334 | 351 | 17 | 5.09% |
| 984 | BS11 | Bristol | 221 | 229 | 8 | 3.62% |
| 1083 | BS2 | Bristol | 318 | 321 | 3 | 0.94% |
| 1088 | BS30 | South Gloucestershire | 121 | 122 | 1 | 0.83% |
| 1196 | BS35 | South Gloucestershire | 91 | 90 | -1 | -1.10% |
| 1236 | BS36 | South Gloucestershire | 47 | 46 | -1 | -2.13% |
| 1267 | BS20 | North Somerset | 141 | 137 | -4 | -2.84% |
| 1308 | BS49 | North Somerset | 54 | 52 | -2 | -3.70% |
| 1349 | BS10 | Bristol | 358 | 341 | -17 | -4.75% |
| 1354 | BS5 | Bristol | 591 | 562 | -29 | -4.91% |
| 1400 | BS8 | Bristol | 217 | 204 | -13 | -5.99% |
| 1461 | BS7 | Bristol | 327 | 303 | -24 | -7.34% |
| 1507 | BS26 | Somerset | 23 | 21 | -2 | -8.70% |
| 1512 | BS13 | Bristol | 394 | 359 | -35 | -8.88% |
| 1528 | BS1 | Bristol | 983 | 893 | -90 | -9.16% |
| 1539 | BS9 | Bristol | 149 | 135 | -14 | -9.40% |
| 1569 | BS31 | Bath and North East Somerset | 166 | 149 | -17 | -10.24% |
| 1664 | BS40 | Bath and North East Somerset | 46 | 40 | -6 | -13.04% |
| 1672 | BS21 | North Somerset | 105 | 91 | -14 | -13.33% |
| 1729 | BS3 | Bristol | 401 | 341 | -60 | -14.96% |
| 1819 | BS37 | South Gloucestershire | 206 | 170 | -36 | -17.48% |
| 1879 | BS15 | Bristol | 405 | 326 | -79 | -19.51% |
| 1915 | BS22 | North Somerset | 263 | 207 | -56 | -21.29% |
| 1930 | BS6 | Bristol | 198 | 155 | -43 | -21.72% |
| 1992 | BS16 | Bristol | 667 | 508 | -159 | -23.84% |
| 2072 | BS41 | North Somerset | 36 | 26 | -10 | -27.78% |
| 2124 | BS39 | Bath and North East Somerset | 62 | 43 | -19 | -30.65% |
| 2198 | BS32 | South Gloucestershire | 108 | 67 | -41 | -37.96% |
| 2238 | BS25 | Somerset | 26 | 14 | -12 | -46.15% |
| 2277 | BS29 | Somerset | 14 | 5 | -9 | -64.29% |
| 2341 | BS80 | Bristol | 0 | 8 | 8 | N/A |
| 2342 | BS98 | Bristol | 0 | 0 | 0 | N/A |
| 2343 | BS99 | Bristol | 0 | 0 | 0 | N/A |

From the information in Table 1, it is clear that the percentage change in crime rates varies greatly across postcodes. Crime rates have increased significantly in some locations while decreasing in others. Postcodes such as BS28, BS24, and BS27 have shown significant positive percentage changes, indicating a substantial increase in crime rates. On the other hand, postcodes such as BS32, BS25, and BS29 show massive negative percentage changes, signifying a significant fall in crime rates in these areas. Some postcodes, like BS2 and BS30, show relatively minor variations in crime rates, either positive or negative, indicating relative stability. While the table provides insight into changes in crime rates, other contexts and considerations, such as demographic transitions, law enforcement techniques, economic changes, and more, are required to comprehend the causes of these changes entirely. The 28 monthly files containing the crime rate data were downloaded. It has been filtered by the postcode district of Bristol (BS), and the ”Date of Transfer” is between January 1, 2020, and April 30, 2022. This indicates that 33,715 records of residential property sale transactions in Bristol were the foundation for the model’s creation.

## 4. Data Processing and Cleansing

### 4.1. Price Paid Data Cleansing

Pre-processing is required as part of the data capture module to decrease the amount of data and execute calculations. Firstly, the house price data was cleaned. Then, the data of the crime files, which are 28 CSV files, was cleaned and combined with the house data, and finally, a final CSV file was created with the features of the house price and crime rate data. This dataset will be used in the statistical analysis stages if its outliers do not have much impact in the modelling stages. The following are included in this step:

1. Importing the original Price Paid Dataset, a 4.3 GB CSV file, and checking the basic information of the dataset.
2. Extracting the transactions during the study’s focal period.
3. Definition of the column header for the housing price dataset.
4. The study’s geographic scope uses price-paid data to extract transaction records for Bristol-based transactions. For that, the study required all of the locations whose postcodes begin with BS because the goal is to work with the data of the city of Bristol.
5. Unnecessary columns were removed from the study and will be used to merge two datasets, its null values may impact the outcome.
6. Unnecessary columns were removed from the study and will be used to merge two datasets, its null values may impact the outcome.
7. A new column combined two price-paid and crime datasets. Finally, the cleaned CSV file of house prices for Bristol was created.
8. A new column combined two price-paid and crime datasets. Finally, the cleaned CSV file of house prices for Bristol was created.
9. A new column combined two price-paid and crime datasets. Finally, the cleaned CSV file of house prices for Bristol was created.
10. A new column combined two price-paid and crime datasets. Finally, the cleaned CSV file of house prices for Bristol was created.

The study’s geographic scope uses price-paid data to extract transaction records for Bristol-based transactions. For that, the study required all of the locations whose postcodes begin with BS because the goal is to work with the data of the city of Bristol.

### 4.2. Crime Rate Data Cleansing

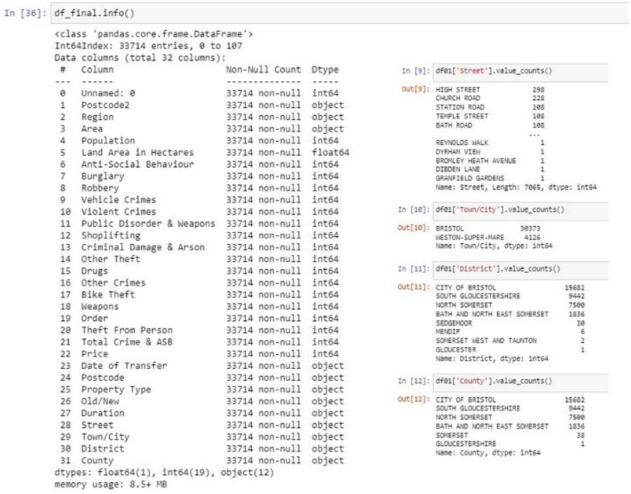
The crime data comes from the monthly reports on the UK crime statistics website. There are 28 files for 28 months (one file for each month from January 2020 to April 2022). These files must be cleaned as part of the data preparation process before being combined with the price-paid data and made into a CSV file, which is then available for analysis and modelling. The data-cleaning steps involve reading each crime CSV file separately and choosing rows whose postcode district starts with “BS.”

### 4.3. Combination of Datasets and the Preparation of the Final Dataset

This step will create a separate CSV file for each month. Then, the monthly crime data should be combined individually with the monthly house price data based on the postcode column. When matching values are in a field shared by both tables, records from the two tables are combined using an inner join operation. Finally, they will be combined based on the date.

The number of crimes committed increases as the population rises. Therefore, it can help determine if crimes have increased or decreased, considering population growth by looking at the crime rate. Consequently, it is necessary to prepare the combined file for analysis and modelling at this stage. One of the most critical measures that will be taken is to create a separate crime rate index for each crime. The steps of this preparation are as follows:

1. As we read the final single file and check the data information (Figure 4), some variables are unnecessary because they have no bearing on our research topic; therefore, eliminating them is preferable. Additionally, it will be simpler to work on the time variable in the following phases if we convert the date and time column to date format.



```
In [36]: df_final.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 33714 entries, 0 to 107
Data columns (total 32 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Unnamed: 0                 33714 non-null  int64
 1   Postcode2                  33714 non-null  object
 2   Region                     33714 non-null  object
 3   Area                       33714 non-null  object
 4   Population                 33714 non-null  int64
 5   Land Area in Hectares      33714 non-null  float64
 6   Anti-Social Behaviour      33714 non-null  int64
 7   Burglary                   33714 non-null  int64
 8   Robbery                    33714 non-null  int64
 9   Vehicle Crimes             33714 non-null  int64
 10  Violent Crimes             33714 non-null  int64
 11  Public Disorder & Weapons  33714 non-null  int64
 12  Shoplifting                33714 non-null  int64
 13  Criminal Damage & Arson    33714 non-null  int64
 14  Other Theft                33714 non-null  int64
 15  Drugs                      33714 non-null  int64
 16  Other Crimes               33714 non-null  int64
 17  Bike Theft                 33714 non-null  int64
 18  Weapons                    33714 non-null  int64
 19  Order                      33714 non-null  int64
 20  Theft From Person          33714 non-null  int64
 21  Total Crime & ASB          33714 non-null  int64
 22  Price                      33714 non-null  int64
 23  Date of Transfer           33714 non-null  object
 24  Postcode                   33714 non-null  object
 25  Property Type              33714 non-null  object
 26  Old/New                    33714 non-null  object
 27  Duration                   33714 non-null  object
 28  Street                     33714 non-null  object
 29  Town/City                  33714 non-null  object
 30  District                   33714 non-null  object
 31  County                     33714 non-null  object
dtypes: float64(1), int64(19), object(12)
memory usage: 8.5+ MB
```

```
In [9]: df01['Street'].value_counts()

Out[9]: HIGH STREET            298
        CHURCH ROAD            228
        STATION ROAD           108
        TEMPLE STREET          108
        BATH ROAD              108
                              ...
        REYNOLDS WALK            1
        DYRHAM VIEW              1
        BROMLEY HEATH AVENUE     1
        DIBDEN LANE              1
        GRANFIELD GARDENS        1
        Name: Street, Length: 7065, dtype: int64

In [10]: df01['Town/City'].value_counts()

Out[10]: BRISTOL              30373
         WESTON-SUPER-MARE     4126
         Name: Town/City, dtype: int64

In [11]: df01['District'].value_counts()

Out[11]: CITY OF BRISTOL                   15682
         SOUTH GLOUCESTERSHIRE              9442
         NORTH SOMERSET                     7500
         BATH AND NORTH EAST SOMERSET       1836
         SEDGEMOOR                            30
         MENDIP                                6
         SOMERSET WEST AND TAUNTON             2
         GLOUCESTER                            1
         Name: District, dtype: int64

In [12]: df01['County'].value_counts()

Out[12]: CITY OF BRISTOL                   15682
         SOUTH GLOUCESTERSHIRE              9442
         NORTH SOMERSET                     7500
         BATH AND NORTH EAST SOMERSET       1836
         SOMERSET                             38
         GLOUCESTERSHIRE                       1
         Name: County, dtype: int64
```

Figure 4. Checking variables' data types of the final dataset.

2. Create the rate column for each crime separately using the following formula:
3. CR is the rate of crime; TC refers to the number of each kind of crime recorded in each area, and P means population:

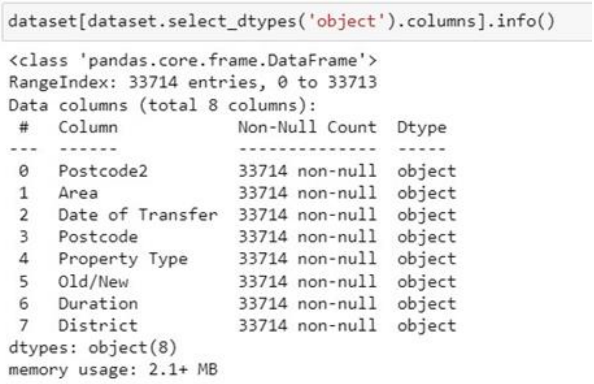
$$CR = TC/P * 100{,}000 \tag{1}$$

## 5. Exploratory Data Analysis

Data visualisation techniques are frequently employed in Exploratory Data Analysis (EDA), which examines and summarises large data sets. It makes finding patterns, identifying anomalies, testing hypotheses, or verifying assumptions simpler by determining how to alter data sources to achieve the desired answers. Exploratory data analysis is crucial for several reasons, such as:

1. Maximising the insight gained from a data set.
2. Identifying outliers.
3. Creating parsimonious models.
4. Identifying errors.
5. Determining the best factor settings.
6. Determining whether there are relationships among the exploratory variables.
7. Assessing the relationship between dependent and independent variables in more detail [25].

Following the techniques in Chatfeild [9], categorising them into graphical and non-graphical EDA and variate or non-variate EDA. The continuous and discrete variables will be divided in this stage and conduct limit analysis [3]. The discrete features will be reviewed in the form of a few graphs since this research focuses on the effect of the crime rate, and all crime rate features are continuous before moving on to modelling. The categorical variables are shown in Figure 5. The continuous variables were defined separately in a data frame, which will be used for the statistical analysis and modelling steps.



```
dataset[dataset.select_dtypes('object').columns].info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33714 entries, 0 to 33713
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Postcode2         33714 non-null  object
 1   Area              33714 non-null  object
 2   Date of Transfer  33714 non-null  object
 3   Postcode          33714 non-null  object
 4   Property Type     33714 non-null  object
 5   Old/New           33714 non-null  object
 6   Duration          33714 non-null  object
 7   District          33714 non-null  object
dtypes: object(8)
memory usage: 2.1+ MB
```

Figure 5. Categorical variables.

The features of data collection are summarised or described using descriptive statistics. Measures of central tendency, frequency distribution measures, and spread make up the three fundamental kinds of measures in descriptive statistics. To emphasise potential correlations between variables and to give basic information about the variables in a dataset, descriptive statistics can be helpful for both of these aims [2].

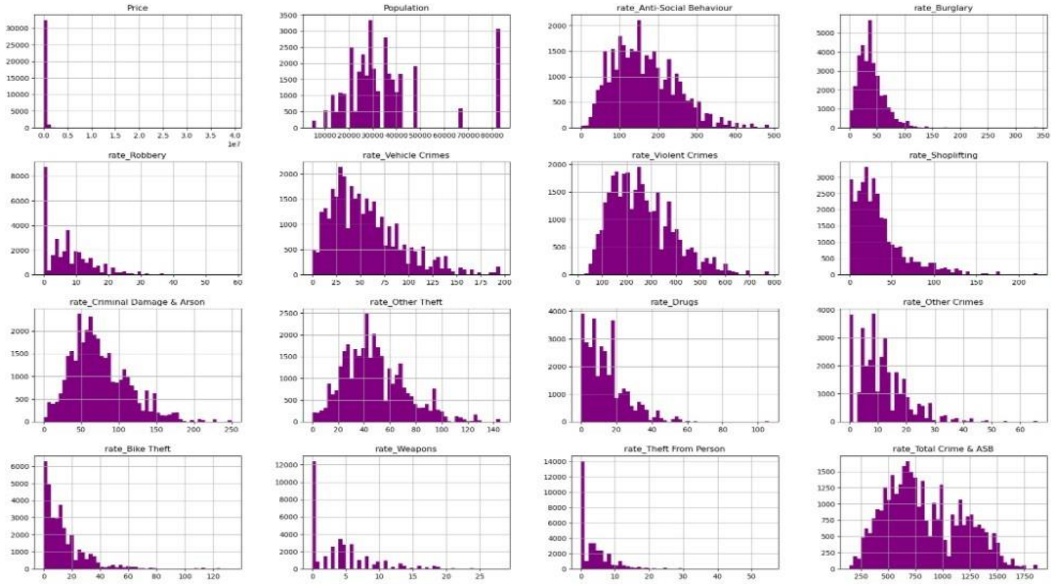
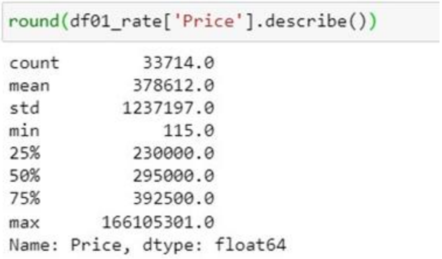


Figure 6. Plotting histogram for each variable.

Plotting a histogram for each numerical attribute in Figure 6 can help you understand the data type we are working with. The distributions of columns population, Anti-Social behaviour, Burglary, Vehicle crimes, Violent crimes, Criminal damage and Arson, Other Theft, and Total crime and ABS seem to have a standard or bimodal data distribution.

According to the histogram, Figure 7 displays a summary statistic for the target variable, "Price". From January 1st, 2020, to April 30th, 2022, the average price of a house in Bristol was 378, 612 pounds. About 1, 237, 197 is the standard deviation, which depicts the range of values around the average price and denotes the significant departure of some values from the average. The price ranges from 115 (the lowest) to 166, 105, 301 (the highest). The price difference between a property's minimum and maximum appears quite large. While both are reasonable, a price of £115 can arise from a property auction with a starting bid of £1, and houses in the UK sell for millions of pounds. In the first quartile, 25% of the properties are listed for less than £230,000, while in the third quartile, 75% of the houses are priced below £392, 500. It should be emphasised that £295,000 is the median or second quartile.



```
round(df01_rate['Price'].describe())

count        33714.0
mean        378612.0
std        1237197.0
min            115.0
25%         230000.0
50%         295000.0
75%         392500.0
max      166105301.0
Name: Price, dtype: float64
```

Figure 7. Summary statistics of "price".

## 5.1. Outliers

The observations that appear to differ significantly from others in the sample are known as outliers. Outliers can occur for many reasons and represent inaccurate data. For instance, an experiment might not have been appropriately run, or the data might have been coded wrongly. In addition, the outputs of statistical modelling and data analysis might be affected by outliers. However, the most critical effects of outliers in the data set are their significant impact on the mean and the standard deviation. Understanding the shape of data is essential because it reveals where the majority of the information is located and, as a result, serves as the foundation for analysing outliers in the dataset.

It is evident from the boxplot diagram in Figure 8 that two outliers are significantly far from the rest of the data. After removing the two outliers, a statistical summary will need to be obtained by the describe command to get an overview of all numerical variables. The average decreased slightly (from 378, 612 to 371, 483), while the standard deviation decreased from 1, 237, 197 to 742, 150, meaning the data spread decreased significantly after removing these two points. The next indicator is the maximum data, which has reached 39, 300, 000 from 166, 105, 301.

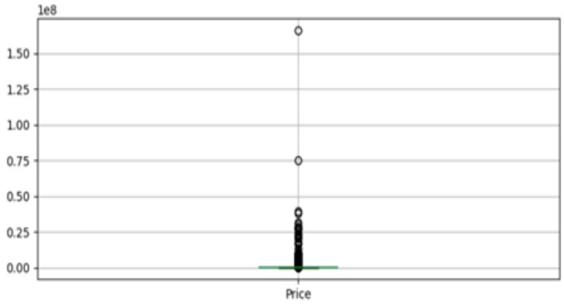


Figure 8. Two clear outliers in the main dataset.

Figure 9 shows the population in each postcode district in Bristol. The chart shows that BS16 has the highest population and BS41 has the lowest population. The rates of various types of crime in Figure 9 show that violent crimes have a significantly greater rate than others. Each region has an average of 263 monthly crimes, with the lowest number being 14. In other words, unlike all other forms of crime, this one has continued in all areas.

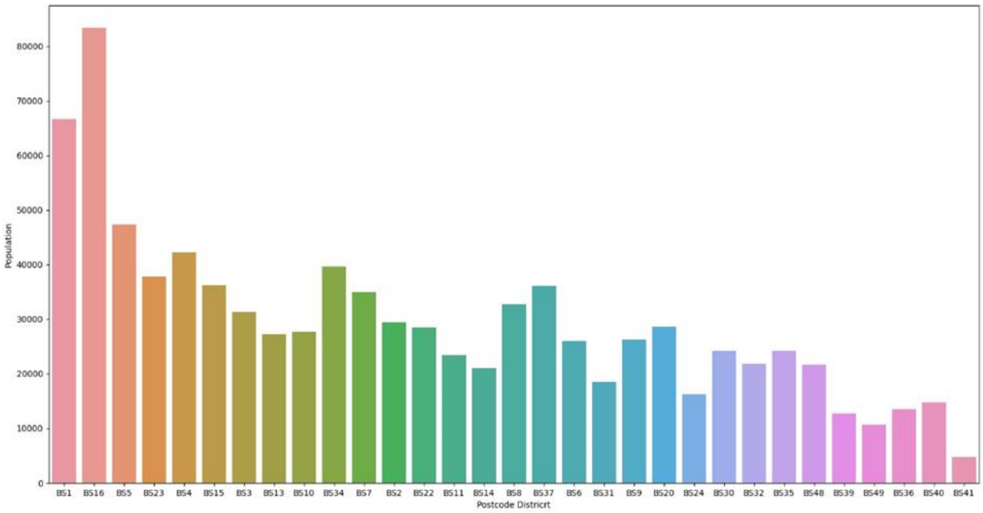


Figure 9. Bar chart of population in each postcode district.

Compared to other types, the maximum of Violent crimes was 778, which is significant. Figure 10 shows that the most violent crimes occurred in BS23 and the least in BS40.

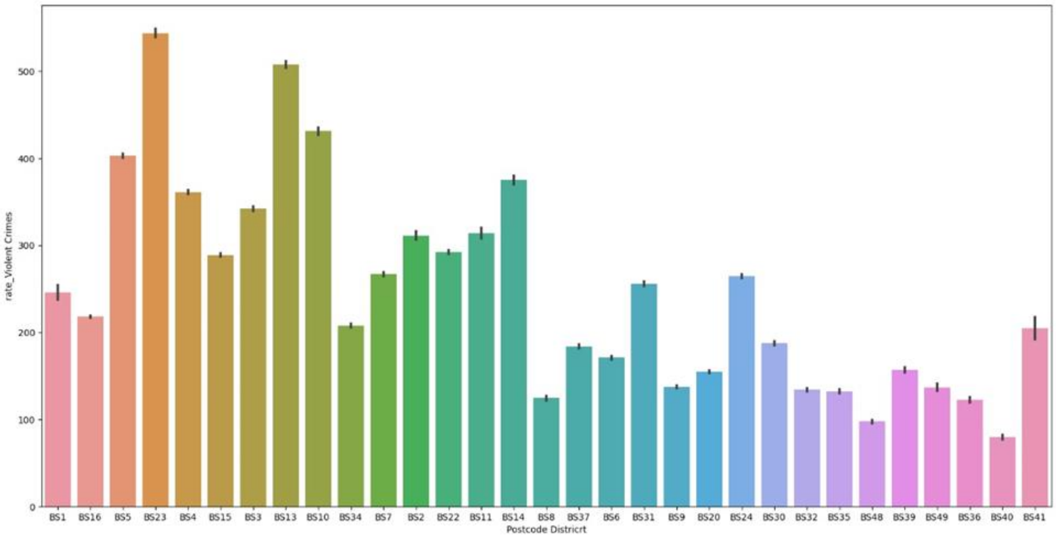
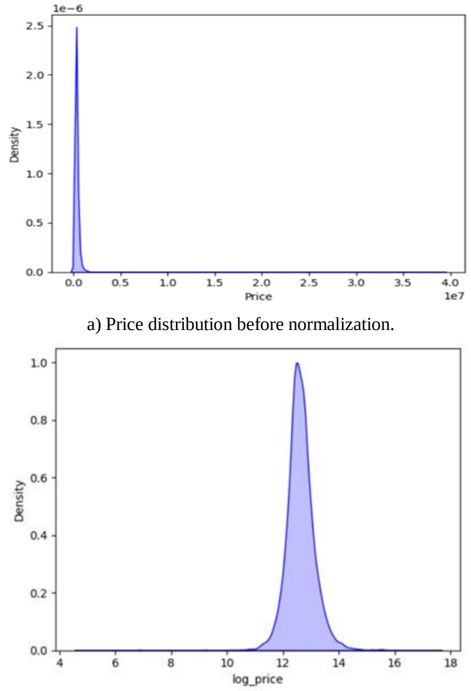


Figure 10. Bar chart of each postcode district’s violent crime rate.

## 5.2. Normalisation

It will be crucial to have the data normalised because the target variable in the dataset is left-skewed and over-dispersed. Normalisation is defined in statistical databases as the technique that ensures data is arranged logically and more sturdily. Instead of altering the values associated with entity attributes, normalisation creates structures based on the logical relationships and interconnections in the data. Considering this, normalisation can be defined as creating sets of relations with desirable features in light of the data needs [12]. Skewness is a metric for a distribution’s asymmetry. When a distribution’s left and right sides are not mirror reflections, it is asymmetrical. Right (positive), left (negative), or zero skewness can all apply to a distribution. Left-skewed distributions are longer on the left side of their peaks, while right-skewed distributions are longer on the right side. Due to its many advantages, a normal distribution is a probability and statistical concept frequently employed in scientific studies. Just one of these advantages is that normal distribution is transparent. It may be defined with the mean and variance, and its mean, median, and mode all have the same value. The price in this scenario does not follow a normal distribution. Although there are strategies to transform a distribution into a normal distribution with the least skewness, no distribution can be transformed into a perfectly normal distribution.

log transformation: The data transformation technique ”log transformation” [37] substitutes a log for each variable x. Typically, the analyst is in charge of selecting the logarithm base, and the goals of statistical modelling will influence their decision. When the bell curve does not fit the original continuous data, the data can be logged and converted to make it as ”normal” as possible, improving the validity of the statistical analysis results. In other words, the skewness of the initial data is reduced or eliminated via the log transformation. Figure 11 shows that the log transformation has been used in the “Price” variable using the NumPy package in Python. After performing a log transformation, it is evident that the Price distribution is somewhat close to normal, which increases the validity of statistical studies.



a) Price distribution before normalization.

b) Normalised price distribution.

Figure 11. Performing log transformation for price variable.

In this section, the data distribution is shown without taking outliers into account (prices less than £1,000,000 are considered). This step was implemented only to compare the modelling results. The data is generally symmetrical if the skewness is between -0.5 and 0.5. The data are moderately skewed if the skewness is between

-1 and -0.5 or 0.5 and 1. The data are significantly skewed if the skewness is more than -1 or equal to 1. By measuring the degree of skewness in Table 2, it appears that when the log transformation is applied, the distribution is still not entirely expected and has a left skew; however, after using Box-Cox, the Price variable has a roughly normal distribution. It should be mentioned that the price distribution is skewed when prices less than 1,000,000 are considered.

Table 2. Comparing skewness.

| Data | Skewness |
|---|---|
| Main dataset | 84.362 |
| Remove two outliers | 30.388 |
| Log transformation | -1.217 |
| Box-Cox transformation | 0.264 |
| Prices less than 1,000,000 | 1.397 |

The following section will use the original data to determine the correlation. The types of distributions described in this section will subsequently be used in the modelling stage, where the best one will be chosen by analysing the model for each one separately.

Spearman correlation coefficient: Three types of correlations are commonly found: Kendall [38] , Pearson [39], and Spearman [40]. One of the most applied statistical correlations is the Pearson correlation. It works as a measure of the degree and direction of a linear relationship between two variables. For this metric, two primary presumptions are:

1. Outliers in the data can significantly impact the Pearson correlation [16].
2. Normally distributed data is required for the analysis.

The Pearson $r$ correlation is calculated using Equation (2) below:

$$rxy = \frac{n\sum xiyi - \sum xi \sum yi}{\sqrt{n\sum xi^2 - (\sum xi)^2}\sqrt{n\sum yi^2 - (\sum yi)^2}} \tag{2}$$

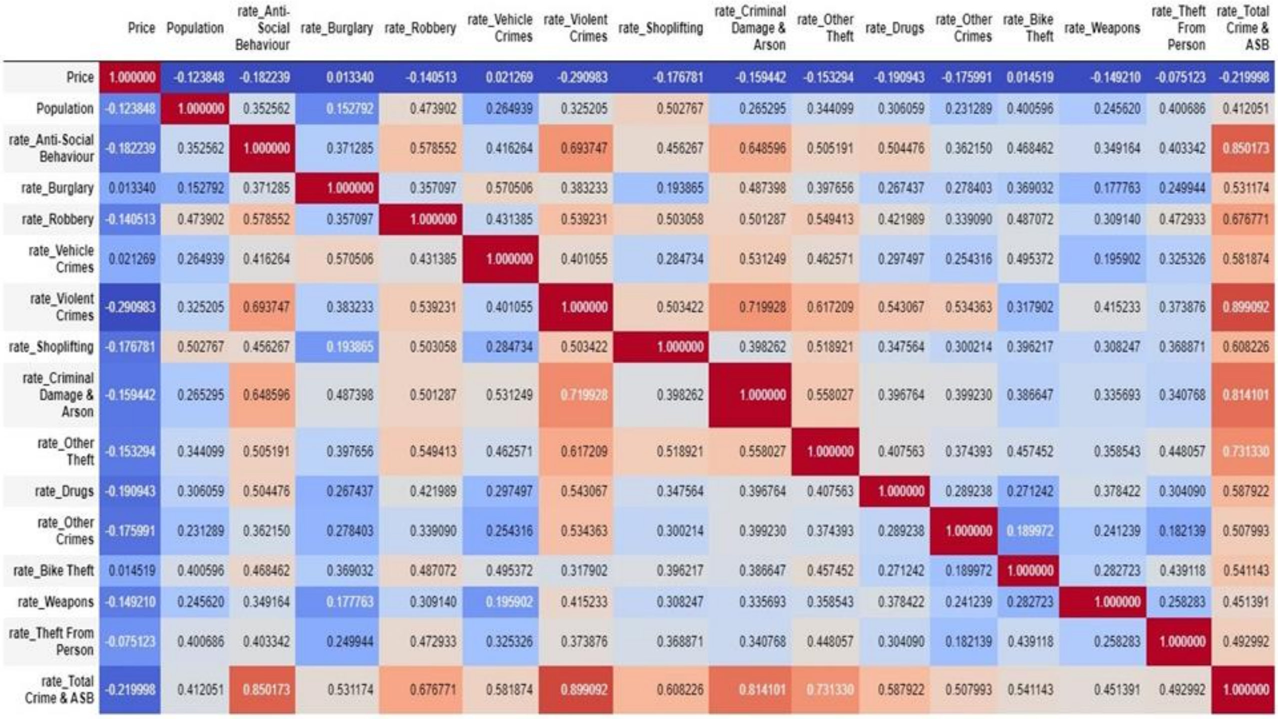
$rxy$=Pearson correlation coefficient between $x$ and $y$
$n$=number of data
$xi$=value of $x$
$yi$=value of $y$

A non-parametric test called Spearman rank correlation as- sesses how closely two variables are related. When the vari- ables are measured on at least an ordinal scale, the Spearman rank correlation test is the ideal correlation analysis because it carries no assumptions about the data distribution. The Spearman rank correlation is calculated using Equation (3):

$$\rho = 1 - \frac{6\sum di^2}{n(n^2-1)} \tag{3}$$

$P$=Spearman rank correlation.
$di$=the variation in the rankings of related variables.
$n$=number of data.

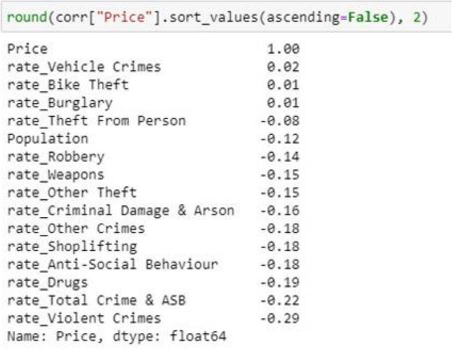
The Spearman correlation and the Kendall correlation are both non-parametric correlations. It can be applied to continuous or ordinal data. This statistic measures the relationship between two variables. Spearman's correlation coefficient was utilised because the primary data have a non-normal distribution and contain many outliers. A table displaying correlation coefficients between variables is called a correlation matrix. The correlation between the two variables is displayed in each table cell. They can provide information regarding the nature (shape), direction, and intensity (degree) of the relationship between the two variables. The correlation matrix for the House and Crime dataset is shown in Figure 12.



| | Price | Population | rate_Anti-Social Behaviour | rate_Burglary | rate_Robbery | rate_Vehicle Crimes | rate_Violent Crimes | rate_Shoplifting | rate_Criminal Damage & Arson | rate_Other Theft | rate_Drugs | rate_Other Crimes | rate_Bike Theft | rate_Weapons | rate_Theft From Person | rate_Total Crime & ASB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Price | 1.000000 | -0.123848 | -0.182239 | 0.013340 | -0.140513 | 0.021269 | -0.290983 | -0.176781 | -0.159442 | -0.153294 | -0.190943 | -0.175991 | 0.014519 | -0.149210 | -0.075123 | -0.219998 |
| Population | -0.123848 | 1.000000 | 0.352562 | 0.152792 | 0.473902 | 0.264939 | 0.325205 | 0.502767 | 0.265295 | 0.344099 | 0.306059 | 0.231289 | 0.400596 | 0.245620 | 0.400686 | 0.412051 |
| rate_Anti-Social Behaviour | -0.182239 | 0.352562 | 1.000000 | 0.371285 | 0.578552 | 0.416264 | 0.693747 | 0.456267 | 0.648596 | 0.505191 | 0.504476 | 0.362150 | 0.468462 | 0.349164 | 0.403342 | 0.850173 |
| rate_Burglary | 0.013340 | 0.152792 | 0.371285 | 1.000000 | 0.357097 | 0.570506 | 0.383233 | 0.193865 | 0.487398 | 0.397656 | 0.267437 | 0.278403 | 0.369032 | 0.177763 | 0.249944 | 0.531174 |
| rate_Robbery | -0.140513 | 0.473902 | 0.578552 | 0.357097 | 1.000000 | 0.431385 | 0.539231 | 0.503058 | 0.501287 | 0.549413 | 0.421989 | 0.339090 | 0.487072 | 0.309140 | 0.472933 | 0.676771 |
| rate_Vehicle Crimes | 0.021269 | 0.264939 | 0.416264 | 0.570506 | 0.431385 | 1.000000 | 0.401055 | 0.284734 | 0.531249 | 0.462571 | 0.297497 | 0.254316 | 0.495372 | 0.195902 | 0.325326 | 0.581874 |
| rate_Violent Crimes | -0.290983 | 0.325205 | 0.693747 | 0.383233 | 0.539231 | 0.401055 | 1.000000 | 0.503422 | 0.719928 | 0.617209 | 0.543067 | 0.534363 | 0.317902 | 0.415233 | 0.373876 | 0.899092 |
| rate_Shoplifting | -0.176781 | 0.502767 | 0.456267 | 0.193865 | 0.503058 | 0.284734 | 0.503422 | 1.000000 | 0.398262 | 0.518921 | 0.347564 | 0.300214 | 0.396217 | 0.308247 | 0.368871 | 0.608226 |
| rate_Criminal Damage & Arson | -0.159442 | 0.265295 | 0.648596 | 0.487398 | 0.501287 | 0.531249 | 0.719928 | 0.398262 | 1.000000 | 0.558027 | 0.396764 | 0.399230 | 0.386647 | 0.335693 | 0.340768 | 0.814101 |
| rate_Other Theft | -0.153294 | 0.344099 | 0.505191 | 0.397656 | 0.549413 | 0.462571 | 0.617209 | 0.518921 | 0.558027 | 1.000000 | 0.407563 | 0.374393 | 0.457452 | 0.358543 | 0.448057 | 0.731330 |
| rate_Drugs | -0.190943 | 0.306059 | 0.504476 | 0.267437 | 0.421989 | 0.297497 | 0.543067 | 0.347564 | 0.396764 | 0.407563 | 1.000000 | 0.289238 | 0.271242 | 0.378422 | 0.304090 | 0.587922 |
| rate_Other Crimes | -0.175991 | 0.231289 | 0.362150 | 0.278403 | 0.339090 | 0.254316 | 0.534363 | 0.300214 | 0.399230 | 0.374393 | 0.289238 | 1.000000 | 0.189972 | 0.241239 | 0.182139 | 0.507993 |
| rate_Bike Theft | 0.014519 | 0.400596 | 0.468462 | 0.369032 | 0.487072 | 0.495372 | 0.317902 | 0.396217 | 0.386647 | 0.457452 | 0.271242 | 0.189972 | 1.000000 | 0.282723 | 0.439118 | 0.541143 |
| rate_Weapons | -0.149210 | 0.245620 | 0.349164 | 0.177763 | 0.309140 | 0.195902 | 0.415233 | 0.308247 | 0.335693 | 0.358543 | 0.378422 | 0.241239 | 0.282723 | 1.000000 | 0.258283 | 0.451391 |
| rate_Theft From Person | -0.075123 | 0.400686 | 0.403342 | 0.249944 | 0.472933 | 0.325326 | 0.373876 | 0.368871 | 0.340768 | 0.448057 | 0.304090 | 0.182139 | 0.439118 | 0.258283 | 1.000000 | 0.492992 |
| rate_Total Crime & ASB | -0.219998 | 0.412051 | 0.850173 | 0.531174 | 0.676771 | 0.581874 | 0.899092 | 0.608226 | 0.814101 | 0.731330 | 0.587922 | 0.507993 | 0.541143 | 0.451391 | 0.492992 | 1.000000 |

Figure 12. Correlation matrix.

The outcomes of sorting the feature-target correlation values with Price (Figure 13), which is our target variable, are as follows: Given that positive correlation refers to a relationship between two variables that change together, there is a positive correlation between Price and three different sorts of crimes, including vehicle crimes, bike theft, and burglary. Even though these numbers are minimal, their positivity is significant. This interpretation implies that more vehicle-related crimes are likely in locations with higher house prices. Theft of bicycles and burglaries have also increased in regions with higher house prices. Given the Figure's values, the Price negatively correlates with the other crime categories, meaning these crimes will occur less frequently in places with higher house prices. For instance, property prices will likely be lower in locations with higher violent crime rates.



```
round(corr["Price"].sort_values(ascending=False), 2)
```

```
Price                           1.00
rate_Vehicle Crimes             0.02
rate_Bike Theft                 0.01
rate_Burglary                   0.01
rate_Theft From Person         -0.08
Population                     -0.12
rate_Robbery                   -0.14
rate_Weapons                   -0.15
rate_Other Theft               -0.15
rate_Criminal Damage & Arson   -0.16
rate_Other Crimes              -0.18
rate_Shoplifting               -0.18
rate_Anti-Social Behaviour     -0.18
rate_Drugs                     -0.19
rate_Total Crime & ASB         -0.22
rate_Violent Crimes            -0.29
Name: Price, dtype: float64
```

Figure 13. Correlation values vs price.

In order of the strongest correlation, these crimes have a negative link with house prices: Violent Crimes, Drugs, Anti-Social Behaviour, Shoplifting, Other Crimes, Criminal Damage and Arson, Other Theft, Weapons, Robbery, and Theft from Person.

## 6. Data Modelling

### 6.1. Using Python Programming Language

Due to its independent platform and widespread use in the programming community, Python is ideally suited for machine learning. Diagramming data flow is the process of data modelling. A diagram of the data flow into and out of the database is the designer's first step when developing a new or alternative database structure. To effectively meet the needs for data flow, this flow diagram is used to specify the characteristics of the data formats, structures, and database handling functions [30]. The data model continues to exist after the database has been created and made available as the documentation and rationale for why the database was created and how the data flows were planned. A thorough and optimal data model aids in developing a simplified, logical database that removes redundancy, facilitates effective retrieval and lowers storage needs. Additionally, it provides all systems with a ”single source of truth”, a necessity for efficient operations and demonstrable adherence to rules and laws [25].

### 6.2. Scikit-Learn Library

The Scikit-Learn class has been employed in the model framework creation in this study. Scikit-Learn is a free and open-source machine learning package based on the Python computer language. The NumPy, SciPy, and Matplotlib libraries from Python are the foundation for this library. These include K-Means, feature selection, pre-processing, spectral clustering, SVR, SVM, random forest, and numerous regression, classification, clustering, dimensionality reduction, model selection, and pre-processing techniques [28]. Hao *et al*. [13] found the characteristics that set the Scikit-Learn library apart from other machine-learning tools. A community review process that helps determine which approaches to include and which to discard or leave out guided our coverage. As a result, a balance between ML's selectivity and comprehensive coverage is achieved. The implementation of the algorithm is secondly optimised for computational effectiveness.

## 7. Model Building

In machine learning, generalising and learning from training data creates a mathematical representation. The created ma- chine learning model is then used to anticipate and produce outcomes with new data. Data that was ready for modelling was divided into training and test sets. Because they utilised each algorithm's default parameters, these models have been called "baseline". This research has used two algorithms, Decision Tree [27] and Random Forest [14]. The use of tree- based models has a variety of advantages. Random Forest is used for:

1) Powerful outcomes.
2) Allows for nonlinear and unsteadiness of variables, and Decision Tree is used for:
   a) Reading and understanding simply.
   b) It is the only scaling-invariant method used in supervised learning. The dataset is prepared for usage by algorithms to train and test our model after the data preparation processes. As a result, the entire dataset is divided into two portions, with 20% used to test the model and 80% used to train it. Variable *X* is used to specify all crime categories (as features or independent variables), and variable *Y* is used to set "price" (as a target or dependent variable). After reading the data, it has been divided into two sets: a training set for the regressor model to be trained on and a test set for the trained model to be used. By determining whether the trained model is accurate, the results from the test set will be compared to the actual outcomes. Therefore, the test size that has been picked is equal to 20% of the overall dataset that will be randomly divided as the test set, and

the remaining 80% will make up the training set used to train the model. The Decision tree and random forest algorithms will be implemented in this section over the following four steps:

1. Model application to the main data (only two outliers are removed).
2. Model application to the data after log transformation.
3. Model application to the data on which the Box-Cox is utilised.
4. The application of a model to data where outliers have virtually been eliminated. (Prices less than $1,000,000 are taken into account).

Each step of the process will involve evaluating the model's performance, and the best models will then be chosen by comparing the regression metrics (RMSE and MAE)

### 7.1. Decision Tree (DT)

This study uses decision trees to estimate how property values relate to district crime rates. Therefore, it is possible to describe it as a flowchart-like model for data analysis that offers a tree-like framework for decision-making and identifying the class and category of a specific data set. The DecisionTreeRegressor module is imported from the sklearn library once the model is divided and prepared for training. The regressor goes through numerous optimisation techniques, including Gradient Descent and Backpropagation, during this training process before creating the Decision Tree Regressor model. After preparing the model, its accuracy must be tested on the test set. This step tests the decision tree algorithm-built model on the previously divided test set.

After that, another simple data frame with two columns must be created, one for the test set's actual values and the other for the predictions. This stage enables us to contrast the outcomes of the developed model.

### 7.2. Random Forest (RF)

One of the practical machine learning algorithms, random forest, produces fantastic outcomes even without parameter tuning. This technique is one of the most popular machine learning algorithms for classification and regression applications because of its accessibility. Using the Random Forest Algorithm has several advantages. One of the most important ones is that it decreases the possibility of overfitting and the time needed for training. It also provides a high degree of accuracy. The Random Forest algorithm efficiently runs in extensive databases by approximating missing data and generating highly accurate predictions. Like the Decision tree algorithm steps, once the model is divided and prepared for training, the 'RandomForestRegressor' module is imported from the sklearn library. After preparing the model, the Random Forest algorithm-built model is tested. A second straightforward data frame with two columns, one for the test set's actual values and the other for the predictions, is now required. We can compare the results of the constructed model at this point.

The graphs in Figures 14 to 16 show that the predicted prices are sometimes higher than the actual prices, occasionally lower, and sometimes a very close or identical match.

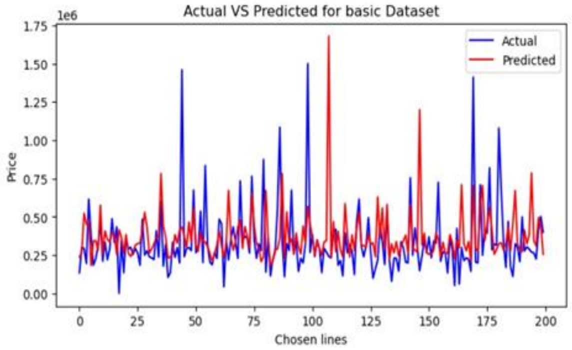


Figure 14. Predicted prices versus actual prices for basic dataset by RF model.

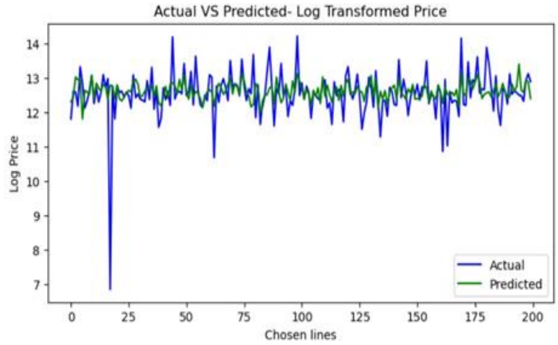


Figure 15. Predicted prices versus actual prices for log-transformed Price by RF model.

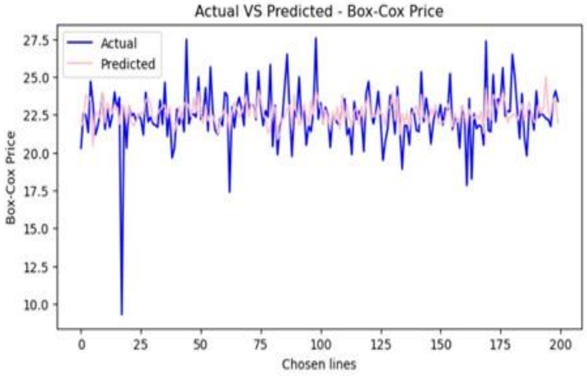


Figure 16. Predicted prices versus actual prices for Box-Cox price by RF model.

### 7.3. Random Forest Regression

By mixing the predictions from various models, ensemble learning is a supervised learning technique used in machine learning to enhance overall performance. Figure 17 shows an example of Ensemble Learning found in Chouinard's work [10].

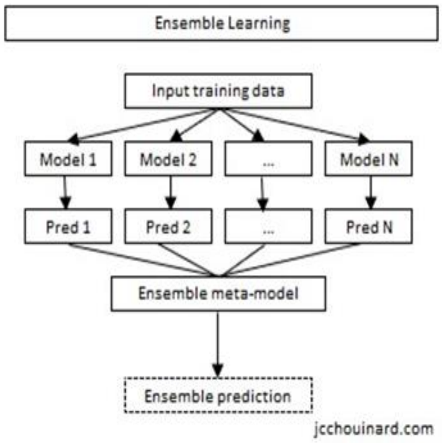


Figure 17. Ensemble learning algorithm.

Random Forest is an ensemble method that handles classification and regression tasks. It uses several decision trees and the bagging technique, often known as bootstrapping and aggregation. This method's fundamental principle is integrating several decision trees to get the final result rather than depending solely on one decision tree. Numerous decision trees are used in Random Forest's base learning models. The concept of Random Forest Models is explained in detail on the ML Science website [11]. (Figure 18)

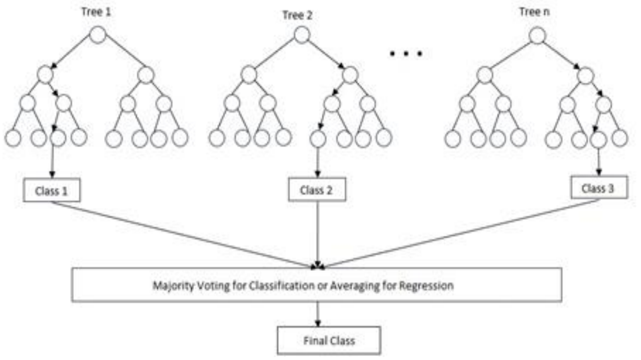

Tree 1
Tree 2
Tree n
Class 1
Class 2
Class 3
Majority Voting for Classification or Averaging for Regression
Final Class

Figure 18. Random forest algorithm.

The random forest algorithm's steps are as follows:

1. From a data set with k records, n random records are selected and used in the Random Forest algorithm. In this study, a random forest will have five decision trees (*n* estimators=5). Each decision tree will be trained on a separate subset of the data, accomplished using random record selection.
2. A unique decision tree is built for each sample. The fit method is used in this step to train each decision tree in the random forest using the supplied data *x* and goal values *y*.
3. Each decision tree will produce an output. In this step, predict the output values for the test data x-test using each decision tree in the ensemble using the trained random forest.
4. The final result is evaluated using a majority vote or an average for classification and regression. In the final phase, the random forest regression model's performance is evaluated using the Root Mean Squared Error (RMSE). While RMSE is not directly about majority voting or averaging (frequently used in classification), it is a measure used for evaluating the ensemble's overall prediction performance, which matches the concept of assessing the final result.

# 8. Model Evaluation

## 8.1. Regression Metrics

The commonly used metric, Root Mean Square Error (RMSE), assesses how accurately a model predicts quantitative data. It calculates an observed standard deviation from the model's predicted value. The observed value equals the sum of the expected value and reliably dispersed random noise with a mean zero, according to [26]. The model is judged effective at predicting the observed data if the noise is low, as RMSE indicates. This means that in the House Price Estimation Framework, the model is considered good at forecasting the observed data if RMSE is low for the model or decreases when comparing model results. The model, however, needs to consider critical aspects of the data if RMSE is high for models or rises when similar model results are compared (Table 3). Mean Absolute Error (MAE) refers to the average of the absolute errors. As a result, it represents the mean of the absolute value of the difference between the anticipated value and the actual value. MAE displays the degree of the expected error relative to the predicted outcomes. A low or declining MAE indicates that the projected and actual property prices are relatively close (Table 4).

Table 3. Comparison of results to evaluate models by RMSE.

| Data | RMSE for decision tree | RMSE for random forest |
|---|---|---|
| Remove two outliers | 675,739.3 | 650,049 |
| Log transformation | 0.538 | 0.52 |
| Box-Cox transformation | 1.541 | 1.489 |
| Prices less than 1,000,000 | 139,855.587 | 135,577.971 |

Table 4. Comparison of results to evaluate models by MAE

| Data | MAE for decision tree | MAE for random forest |
|---|---|---|
| Remove two outliers | 165,055.718 | 160,645.296 |
| Log transformation | 0.345 | 0.335 |
| Box-Cox transformation | 1.005 | 0.974 |
| Prices less than 1,000,000 | 101,279.474 | 98,343.512 |

RMSE and MAE were calculated and shown in Tables 3 and 4 after applying Decision Tree, and Random Forest algorithms were used to four datasets. Since RMSE and MAE are the lowest for Random Forest models when the values for the two techniques are compared, it can be said that Random Forest models were the best when RMSE and MAE scores were for all datasets considered. Another critical point is that RMSE and MAE had the lowest values when Log transformation was used for the target variable (Price). In other words, after employing Log transformation, the predicted values of prices were the most accurate.

## 8.2. Feature Importance

The feature's importance reflects how many variables influence the model's forecasting ability. Generally, it establishes how much a particular variable benefits a given model and prediction. For the best-performing model, feature importance is applied here. Before that, a decision tree visualisation is employed to demonstrate how underlying data predicts ”Price” and to emphasise important decision tree insights. For example, in Figure 19, for Log transformed price, which was the best performing model, a Decision Tree using Matplotlib has been visualised and then using feature importance most influenced features on house prices will be introduced.

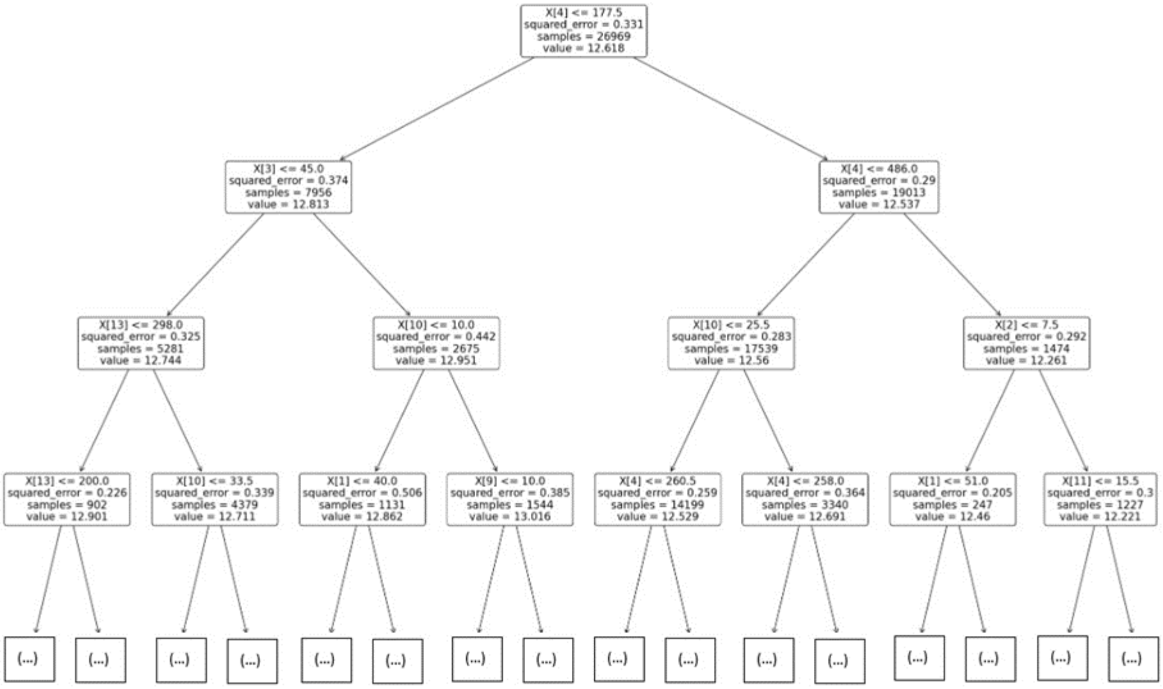
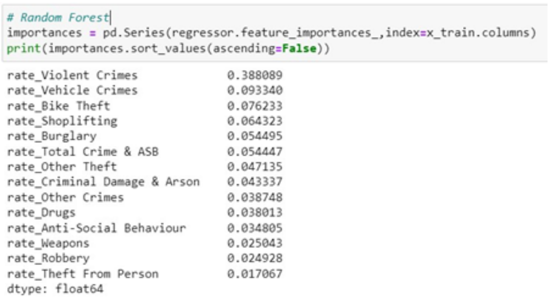


Figure 19. Visualising decision tree for log-transformed price.

Figure 20 displays the ranked importance of criminal categories for variable price prediction. Given that the scores indicate how important each feature is, a higher score means that the particular feature will significantly impact the predictive model. Again, it can be seen that "violent crimes" are the critical element. In other words, violent crimes, followed by "vehicle crimes" and ”bike theft", might be used to measure the price prediction.



```
# Random Forest
importances = pd.Series(regressor.feature_importances_,index=x_train.columns)
print(importances.sort_values(ascending=False))

rate_Violent Crimes              0.388089
rate_Vehicle Crimes              0.093340
rate_Bike Theft                  0.076233
rate_Shoplifting                 0.064323
rate_Burglary                    0.054495
rate_Total Crime & ASB           0.054447
rate_Other Theft                 0.047135
rate_Criminal Damage & Arson     0.043337
rate_Other Crimes                0.038748
rate_Drugs                       0.038013
rate_Anti-Social Behaviour       0.034805
rate_Weapons                     0.025043
rate_Robbery                     0.024928
rate_Theft From Person           0.017067
dtype: float64
```

Figure 20. Feature importance for RF algorithm for log-transformed price.

It should be noted that these characteristics also demonstrated the strongest correlation with house prices. As a result, it is critical to consider the frequency of these three crimes when estimating housing prices because they can affect actual prices.

## 9. Conclusions and Contribution to Knowledge

This section discusses the outcomes of exploratory data analysis, correlation analysis, skewness analysis, data normalisation, and the application of baseline models to estimate or predict house prices. The model accuracy evaluation used both regression measures (RMSE and MAE). The model with the highest performance was Random Forest. The first three features that influenced our algorithm were introduced using the feature importance technique in the final section. These features include, in that order, "violent crimes", "vehicle crimes", and "bike theft". It should be noted that these characteristics also demonstrated the strongest correlation with house prices. As a result, it is critical to consider the frequency of these three crimes when estimating housing prices because they can affect actual prices.

This study has developed a deeper understanding of machine learning by investigating how machine learning systems can be built to learn from and employ a variety of data sources to generate insights that support the decision-making and actions of users in the housing market. For example, the price of the home eventually serves as the standard for decisions like where to build or invest in a house for the highest return on investment. These stakeholders will have a baseline that supports their choices if the price of a home can be reliably anticipated based on a variety of known criteria and potential future neighbourhood, economic, or infrastructure developments.

The research has created a new dataset (crime rate categories and house prices in each district). Models have been developed following exploratory data analysis, and regression metrics have been used to assess them. The most effective performance model was chosen, and the most important features were then determined with feature importance. The main contribution of this study to knowledge is to reduce the research gap by using machine learning algorithms to generate insights that improve decision-making for various stakeholders. The list of the most significant contributions to this research is summarised below. First, a new dataset combining. housing prices and crime rates was developed. This step involved merging 28 crime datasets for 28 months (from January 2020 to April 2022) with price- paid data that had been selected by month depending on the postcode district in Bristol. It is a whole new dataset that may be examined to extract information. One of the primary design outputs of this research is based on the design science approach.

Second, significant outliers and substantially leftward skew- ness of prices' distribution caused the creation of 8 machine learning models, which use two separate machine learning techniques (DT and RF) as the baseline features.

Third, studies have already been done on how machine learning algorithms react when predicting price, algorithm comparisons, model accuracy, and predictors.

Fourth, various evaluation metrics react to features and machine learning algorithms differently. The type of machine learning problem typically determines the choice of assessment measures, as was covered in the Methodology chapter. MAE and RMSE were used in this study due to regression problems. Beyond them, however, a combination of business issues or project objectives, various variables, and machine learning algorithms may impact the evaluation selection.

Fifth, the Feature Importance Technique was used to identify variables affecting the model's forecast ability, as clarified in the EDA section.

After implementing statistical tests and machine learning algorithms to determine whether there was a correlation be- tween the crime rate and the price of houses in each location, the findings that support the probability of a relationship's existence have been partially supported by the existing literature. The Random Forest regressor models outperformed the decision tree models after datasets were used for training and testing. Other articles have supported this finding, and [17] recommend using this regression model to estimate future home prices.

It might be challenging to deal with crime data; thus, it's critical to understand that reported crime rates can vary from actual crime rates due to over or under-reporting. Another crucial point is that, generally, crime rates will be correlated with other area characteristics that are likely to impact crime. Understanding the complex relationship between regional crime rates, economic conditions, and other factors impacting corporate location decisions is critical for policymakers and businesses. By considering these aspects, stakeholders can make more informed decisions promoting economic growth and development while addressing social issues like crime rates.

## References

[1] Alfiyatin A., Febrita R., Taufiq H., and Mahmudy W., "Modeling House Price Prediction Using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp.323-326, 2017. DOI:10.14569/IJACSA.2017.081042

[2] Aljawawdeh H. and Nabot A., "CASL: Classical, Asynchronous, and Synchronous Learning Model towards a Universal Hybrid E-Learning Model in Jordan Universities," *in Proceedings of the 22$^{nd}$ International Arab Conference on Information Technology*, Muscat, pp. 1-9, 2021. DOI:10.1109/ACIT53391.2021.9677410

[3] Aljawawdeh H., "An Enriched E-Learning Model to Teach Kids in Arab Countries How to Write Code," *in Proceedings of the International Arab Conference on Information Technology*, Abu Dhabi, pp. 1-10, 2022. DOI:10.1109/ACIT57182.2022.9994110

[4] Aljawawdeh H., Simons C., and Odeh M., "Metaheuristic Design Pattern: Preference," *in Proceedings of the Companion Publication of the Annual Conference on Genetic and Evolutionary Computation*, Madrid, pp. 1257-1260, 2015. https://doi.org/10.1145/2739482.2768498

[5] Bamiteko O. and Adebiyi O., "Effect of Neighborhood Security on Housing Price in Lagos," *American Journal of Environmental and Resource Economics*, vol. 5, no. 4, pp. 80-85, 2020. DOI:10.11648/j.ajere.20200504.11

[6] Bank of England. How does the Housing Market Affect the Economy? https://www.bankofengland.co.uk/explainers/how-does-the-housing-market-affect-the-economy, 2023. Last Visited, 2024.

[7] Borde S., Rane A., Shende G., and Shetty S., "Real Estate Investment Advising Using Machine Learning," *International Research Journal of Engineering and Technology*, vol. 4, no. 3, pp. 1821-1825, 2017.

[8] Braakmann N., "The Link between Crime Risk and Property Prices in England and Wales: Evidence from Street-Level Data," *Urban Studies*, vol. 54, no. 8, pp. 1990-2007, 2017. DOI:10.1177/0042098016634611

[9] Chatfield C., "Exploratory Data Analysis," *European Journal of Operational Research*, vol. 23, no. 1, pp. 5-13, 1986. https://doi.org/10.1016/0377-2217(86)90209-2

[10] Chouinard J., Decision Trees in Machine Learning, https://www.jcchouinard.com/decision-trees-in-machine-learning/, Last Visited, 2024.

[11] Choy L. and Ho W., "The Use of Machine Learning in Real Estate Research," *Land*, vol. 12, no. 4, pp. 740, 2023. https://doi.org/10.3390/land12040740

[12] Eessaar E., "The Database Normalization Theory and the Theory of Normalized Systems: Finding a Common Ground," *Baltic Journal of Modern Computing*, vol. 4, no. 1, pp. 5-33, 2016. https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/4_1_3_Eessaar.pdf

[13] Hao J. and Ho T., "Machine Learning Made Easy: A Review of Scikit-Learn Package in Python Programming Language," *Journal of Educational and Behavioral Statistics*, vol. 44, no. 3, pp. 348-361, 2019. https://doi.org/10.3102/1076998619832248

[14] IBM, What is Random Forest? https://www.ibm.com/topics/ random-forest, 2023, Last Visited, 2024.

[15] Ja'afar N., Mohamad J., and Ismail S., "Machine Learning for Property Price Prediction and Price Valuation: A Systematic Literature Review," *Planning Malaysia*, vol. 19, no. 3, pp. 411-422, 2021. DOI:10.21837/pm.v19i17.1018

[16] Kim Y., Kim T., and Ergun T., "The Instability of the Pearson Correlation Coefficient in the Presence of Coincidental Outliers," *Finance Research Letters*, vol. 13, pp. 243-257, 2015. https://doi.org/10.1016/j.frl.2014.12.005

[17] Kudavkar A., Sadanad S., Nhavkar M., and Wagh S., "House Price Prediction Using Machine Learning Algorithm," *International Research Journal of Engineering and Management Studies*, vol. 3, no. 4, pp. 1-6, 2019. https://www.researchgate.net/publication/350430324_House_Price_Prediction_Using_Machine_Learning_Algorithm

[18] Madhukar B., Bharathi S., and Ashwin M., "Classification of Breast Cancer Using Ensemble Filter Feature Selection with Triplet Attention Based Efficient Net Classifier," *The International Arab Journal Information Technology*, vol. 21, no. 1, pp. 17-31, 2024. https://doi.org/10.34028/iajit/21/1/2

[19] Mayer M., Bourassa S., Hoesli M., and Scognamiglio D., "Estimation and Updating Methods for Hedonic Valuation," *Journal of European Real Estate Research*, vol. 12, no. 1, pp. 134-150, 2019. https://doi.org/10.1108/JERER-08-2018-0035

[20] McIlhatton D., McGreal W., De La Paz T., and Adair A., "Impact of Crime on Spatial Analysis of House Prices: Evidence from a UK City," *International Journal of Housing Markets and Analysis*, vol. 9, no. 4, pp. 627-647, 2016. http://dx.doi.org/10.1108/IJHMA-10-2015-0065

[21] Minghetti A., Exploring Why Crime and House Prices Correlate Positively in London, https://api.semanticscholar.org/CorpusID:212411930, Last Visited, 2024.

[22] Moody J., What Does RMSE Really Mean? https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e, Last Visited, 2024.

[23] Narayanan L., Krishnan S., and Robinson H., "A Hybrid Deep Learning Based Assist System for Detection and Classification of Breast Cancer from Mammogram Images," *The International Arab Journal Information Technology*, vol. 19, no. 6, pp. 965-974, 2022. https://doi.org/10.34028/iajit/19/6/15

[24] Osmadi A., Kamal E., Hassan H., and Fattah H., "Exploring the Elements of Housing Price in Malaysia," *Asian Social Science*, vol. 11, no. 24, pp. 26-38, 2015. http://dx.doi.org/10.5539/ass.v11n24p26

[25] Patil P., What is Exploratory Data Analysis? https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15, Last Visited, 2024.

[26] SAP Insights, What is Data Modeling? https://www.sap.com/products/technology-platform/datasphere/what-is-data-modeling.html, Last Visited, 2024.

[27] Scikit-Learn., Decision Trees, https://scikit-learn.org/stable/modules/tree.html, Last Visited, 2024.

[28] Scikit-Learn., Scikit-Learn Machine Learning in Python, https://scikit-learn.org/stable/, Last Visited, 2024.

[29] Susan J. and Subashini P., "Deep Learning Inpainting Model on Digital and Medical Images-A Review," *The International Arab Journal Information Technology*, vol. 20, no. 6, pp. 919-936, 2023. https://doi.org/10.34028/iajit/20/6/9

[30] Takci H. and Nusrat F., "Women. Highly Accurate Spam Detection with the Help of Feature Selection and Data Transformation," *The International Arab Journal of Information Technology*, vol. 20, no. 1, pp. 29-37, 2023. https://doi.org/10.34028/iajit/20/1/4

[31] The Science of Machine and AI Random Forest Models, https://www.ml-science.com/random-forest, Last Visited, 2024.

[32] Tita G., Petras T., and Greenbaum R., "Crime and

Residential Choice: A Neighborhood Level Analysis of the Impact of Crime on Housing Prices," *Journal of Quantitative Criminology*, vol. 22, pp. 299-317, 2006. https://doi.org/10.1007/s10940-006-9013-z

[33] Trinh H., Khan M., Squires G., and Mareic M., Housing Price, Financial Development, Energy Intensity, FDI Inflows: Global Evidence, https://www.nzae.org.nz/wp-content/uploads/2021/07/Trinh.pdf, Last Visited, 2024.

[34] UK Crime Stats. Crime by Postcode District, https://ukcrimestats.com/Postcode_Districts/, Last Visited, 2024.

[35] UK Government. Price Paid Data, https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads#single-file, Last Visited, 2024.

[36] Vengaloor R. and Muralidhar R., "Deep Learning Based Feature Discriminability Boosted Concurrent Metal Surface Defect Detection System Using YOLOV-5s-FRN," *The International Arab Journal Information Technology*, vol. 21, no. 1, pp. 94-106, 2024. https://doi.org/10.34028/iajit/21/1/9

[37] West R., "Best Practice in Statistics: The Use of Log Transformation," *Annals of Clinical Biochemistry*, vol. 59, no. 3, pp. 162-165, 2022. DOI:10.1177/00045632211050531

[38] Wikipedia, Kendall Rank Correlation Coefficient, https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient, Last Visited, 2024.

[39] Wikipedia, Pearson Correlation Coefficient, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient, Last Visited, 2024.

[40] Wikipedia, Spearman's Rank Correlation Coefficient, 2023, https://en.wikipedia.org/wiki/Spearman%27s_rank Correlation_Coefficient, Last Visited, 2024.

[41] Wittowsky D., Hoekveld J., Welsch J., and Steier M., " Residential Housing Prices: Impact of Housing Characteristics, AccessiBility and Neighbouring Apartments-A Case Study of Dortmund, Germany," *Urban, Planning and Transport Research*, vol. 8, no. 1, pp. 44-70, 2020. https://doi.org/10.1080/21650020.2019.1704429

[42] Woolf A., 8 Reasons to Invest in Property in Bristol and Where to Buy, https://aspenwoolf.co.uk/resources/property-news/category/tips/8-reasons-to-invest-in-property-in-bristol-and-where-to-buy/, Last Visited, 2024.

[43] Zhang Z., "Decision Trees for Objective House Price Prediction," *in Proceedings of the 3rd International Conference on Machine Learning, Big Data and Business Intelligence*, Taiyuan, pp. 280-283, 2021. DOI:10.1109/MLBDBI54094.2021.00059

[44] Zulkifley N., Rahman S., Ubaidullah N., and Ibrahim I., "House Price Prediction Using a Machine Learning Model: A Survey of Literature," *International Journal of Modern Education and Computer Science*, vol. 12, no. 6, pp. 46-54, 2020. DOI:10.5815/ijmecs.2020.06.04

**Soudabeh Motevali** holds a Master's in Data Science from the University of the West of England (UWE), Bristol, with distinction (2023), and a Bachelor's in Statistics from Delijan University, Iran. Her studies focused on data analysis, statistical modeling, and research methods. She has worked in various sectors, starting as an accountant in shipping and logistics, then moving into financial roles, including internal auditing and financial data analysis in investment services. Her current research, based on her master's thesis, looks at the link between house prices and crime rates using machine learning. Passionate about data-driven decision-making and statistical modeling, she is committed to advancing these fields and improving business strategies.

**Hamzeh Aljawawdeh** is a distinguished software engineering professional holding a PhD in the discipline. He has a solid background in academia and industry. He has significantly contributed to developing and advancing software engineering methodologies and practices. His research interests include Business Process Models, Software Architecture, Genetic Algorithms, and E-Learning, which have led him to produce numerous publications, conference presentations, and fruitful collaborations with other leading experts in the field. In addition, through their work, they have contributed to enhancing software engineering techniques and fostered a deeper understanding of the interplay between technology and educational practices. In addition to his academic pursuits, Hamzeh has extensive experience in the software engineering industry. As a software engineer, he has skillfully led teams in developing and successfully deploying large-scale projects across various sectors. These accomplishments have cemented Hamzeh's reputation as a skilled and reliable software engineering professional.

**Sami Abuezhayeh** PhD, is a lecturer in the computing department at the University of the West of England. Sami has extensive industry experience as a consultant in media content management and city development. His research interests lie in the intersection of the built environment and management information systems, with a particular focus on enhancing Decision-Making Processes (DMP) through the implementation of Knowledge Management (KM) and Business Process Management (BPM) activities within the construction sector. Sami has supervised numerous dissertations and research group projects.

**Eman Qaddoumi** PhD, is a Fellow of the Higher Education Academy. With a demonstrated history of working in the higher education sector within and outside the UK, Eman is a lecturer and researcher. Currently, Eman holds a position as a lecturer in the Data Science department and is a member of the Computer Science Research Centre within the College of Arts, Technology, and Environment at the University of the West of England (UWE), Bristol, UK. Eman's research interests include data science, software quality, and requirements modelling, with a particular focus on semantic modelling in the systems of systems context. Eman has taken the lead in developing several modules in higher education.