A black background with blue text

Description automatically generated

Prabin Tiwari

Softwarica College of IT and E-Commerce

ST5014CEM Data Science for Developers

Siddhartha Neupane

August 19, 2024

Table of Contents

[Introduction 5](#_Toc175004132)

[Data Cleaning 6](#_Toc175004133)

[Dataset Importing 6](#_Toc175004134)

[Cleaning Datasets 6](#_Toc175004135)

[Exploratory Data Analysis (EDA) 8](#_Toc175004136)

[Data Visualization 9](#_Toc175004137)

[EDA of Housing 9](#_Toc175004138)

[EDA of Broadband 13](#_Toc175004139)

[EDA of Crime 16](#_Toc175004140)

[EDA of School 19](#_Toc175004141)

[Linear Model 22](#_Toc175004142)

[House Price vs Average Download Speed 22](#_Toc175004143)

[Attainment 8 score VS House Price 24](#_Toc175004144)

[Average Download Speed VS Attainment 8 score 25](#_Toc175004145)

[House Price vs Drug Rates 26](#_Toc175004146)

[Recommendation System 27](#_Toc175004147)

[House Price Ranking 27](#_Toc175004148)

[Broadband Speed Ranking 28](#_Toc175004149)

[Crime Score Ranking 30](#_Toc175004150)

[School Score Ranking 31](#_Toc175004151)

[Overall Ranking 32](#_Toc175004152)

[Legal and Ethical Issues 33](#_Toc175004153)

[Reflection 33](#_Toc175004154)

[Conclusion 34](#_Toc175004155)

[References 35](#_Toc175004156)

[Appendix 36](#_Toc175004157)

[Data Cleaning 36](#_Toc175004158)

[Graphs 39](#_Toc175004159)

[Ranking Code 44](#_Toc175004160)

Table of Figures

[Figure 1: 10](#_Toc174997206)

[Figure 2: 11](#_Toc174997207)

[Figure 3: 12](#_Toc174997208)

[Figure 4: 13](#_Toc174997209)

[Figure 5: 14](#_Toc174997210)

[Figure 6: 15](#_Toc174997211)

[Figure 7: 16](#_Toc174997212)

[Figure 8: 17](#_Toc174997213)

[Figure 9: 18](#_Toc174997214)

[Figure 10: 19](#_Toc174997215)

[Figure 11: 20](#_Toc174997216)

[Figure 12: 21](#_Toc174997217)

[Figure 13: 23](#_Toc174997218)

[Figure 14: 24](#_Toc174997219)

[Figure 15: 25](#_Toc174997220)

[Figure 16: 26](#_Toc174997221)

[Figure 17: 28](#_Toc174997222)

[Figure 18: 29](#_Toc174997223)

[Figure 19: 30](#_Toc174997224)

[Figure 20: 31](#_Toc174997225)

[Figure 21: 32](#_Toc174997226)

[Figure 22: 36](#_Toc174997227)

[Figure 23: 37](#_Toc174997228)

[Figure 24: 37](#_Toc174997229)

[Figure 25: 38](#_Toc174997230)

[Figure 26: 39](#_Toc174997231)

[Figure 27: 40](#_Toc174997232)

[Figure 28: 41](#_Toc174997233)

[Figure 29: 43](#_Toc174997234)

[Figure 30: 44](#_Toc174997235)

[Figure 31: 45](#_Toc174997236)

[Figure 32: 46](#_Toc174997237)

[Figure 33: 47](#_Toc174997238)

[Figure 34: 48](#_Toc174997239)

# Introduction

In the context of helping international friends make an informed property investment decision in the United Kingdom, this report aims to analyze various factors including housing prices, internet connectivity, crime rates, and other significant aspects in the regions of Bristol and Cornwall. This objective is to utilize data science techniques to process and analyze datasets relevant to these factors, ultimately leading to a recommendation of the top towns based on their suitability for investment.

This project involved the complete data mining lifecycle, data collection and cleaning exploratory data analysis (EDA) and the development of a basic recommendation system. The datasets utilized for this project are sourced from the UK government’s open data repositories, ensuring their credibility and relevance. The analysis and recommendation system are developed using R language, leveraging libraries such as tidyverse, ggplot2, fmsb. Etc.

# Data Cleaning

Data cleaning is a fundamental process in data analysis. It involves dealing with missing numbers, eliminating or fixing inaccurate data, and getting the data ready for analysis. Data cleansing is crucial since it provides accurate analysis, boosts model performance, and assures the validity of recommendations. For this project, we cleaned and worked with the data using R's tidyverse library. An extensive description of the data cleaning procedure is provided below.

## Dataset Importing

In the initial phase of the analysis, the required datasets were imported into the environment using R. These datasets included housing prices, broadband speed, crime rates, and school data all of which are crucial for the analysis and recommendations. The “read.csv()” function in R was employed to load these datasets efficiently, enabling subsequent data manipulation and cleaning. Each dataset provided specific insights, housing prices offered average property costs across towns in Bristol and Cornwall, broadband speed data highlighted internet availability, crime rates detailed the frequency of various offenses, and school data provided metrics on educational performance.

## Cleaning Datasets

After importing the datasets, the data cleaning process was undertaken to ensure the datasets were accurate and ready for analysis. This process involved handling missing values by filling gaps with appropriate substitutes, such as the median or mean values. Data normalization was applied to conform the datasets to the third normal form(3NF), reducing redundancy and enhancing data integrity. Additionally, categorical data was transformed into numerical formats for ease of analysis, and outliers were addressed using statistical methods like the interquartile range (IQR) and Z-scores, either by capping or removing them. Finally, the cleaned datasets were integrated into a unified model, ensuring that each town was represented as a unique entity with associated attributes from all datasets.

Housing Dataset Cleaning

In this dataset cleaning, the primary objective was to clean and prepare the housing dataset for analysis. Essential columns such as Postcode and Town/City were extracted. The Postcode was standardized by trimming it to the first six characters to ensure consistency with other datasets. This cleaned data was subsequently used to join with cleaned crime dataset.

Broadband Dataset Cleaning

For the Broadband Dataset, the aim was to clean broadband availability and speed data. The dataset was filtered to focus on relevant geographic areas, such as Bristol and Cornwall. Column names and formats were standardized, duplicates and irrelevant rows were removed, and the data was merged with other datasets as needed, particularly by postcode.

Crime Dataset Cleaning

In the Crime Dataset cleaning process, the focus was on the data from Bristol and Cornwall for the year 2022 to 2024. Monthly crime datasets for each year and location were combined using “rbind” and the dataset was converted to a Tibble for better handling. Essential columns, including Month, LSOA code and Crime type, were selected. Duplicates based on LSOA code were removed, and the dataset was merged with LSOA data to include geographic information like streets and counties. Population data was also added to provide demographic context. Rows with missing values in key columns were filtered out to ensure data quality.

School Dataset Cleaning

The School Dataset was prepared by filtering data for schools within relevant geographic regions, cleaning and standardizing data formats, removing duplicates entries and merging other datasets using geographical identifiers like postcodes.

LSOA Dataset Cleaning

For the LSOA Dataset, the cleaning process involved selecting and renaming columns for consistency such as lsoa11cd, lsoa11nm, and pcds. To LSOA code, street, counties and postcode respectively. The data was filtered to retain only relevant areas like Bristol and Cornwall, and postcodes were standardized. Duplicates data were checked and removed.

Population Dataset Cleaning

The Population Dataset was cleaned by renaming columns for clarity e.g., renaming the postcode column to “postcode” and population count to “count” and postcodes were standardized to ensure proper joining with other datasets. This cleaned population data was used in conjunction with other datasets, such as crime data, to provide contextual information on population density.

# Exploratory Data Analysis (EDA)

In the Exploratory Data Analysis (EDA) phase, the primary goal was to understand the underlying patterns, trends, and relationships within the datasets. This step is crucial in any data analysis project as it allows us to gain insights and prepare the data for further analysis, including modeling. We used various techniques such as summarizing data, visualizing distributions, and identifying outliers to explore the characteristics of the data.

## Data Visualization

Data visualization is the process representation of data through use of common graphics, such as charts, graphs or diagrams to make information more accessible and easier to understand. These techniques enable us to grasp the insights quickly that might be buried in a raw number by recognizing the patterns and trends in the visual data. Various visualization methods, such as box plots and line graphs are crucial tools in this process to make informed decisions and better understand the factors influencing outcomes like customer behavior and market trends.

In this project, data visualization techniques have been employed effectively to present the cleaned and processed data. By using visual tools like box plots to display data distribution and line graphs to track trends over time. These visualizations were specifically designed to address key questions in our analysis, such as examining the average house prices and identifying trends in crime rates. Making the data both actionable and insightful.

## EDA of Housing

House price data representation aims to visually analyze the real estate market by showing trends and variances in property values. For example, bar charts are used to show average property prices in different towns in 2023, and line graphs show price patterns over multiple years within several counties. Additionally, boxplots are used to compare housing prices across counties in 2023. These visualizations give analysts a full picture of the real estate sector, allowing them to detect patterns, analyze market movements, and make informed predictions.

Figure 1:

Bar chart of average house price in 2023

A screenshot of a graph

Description automatically generated

Figure 2:

Box plot of house prices

A screenshot of a graph

Description automatically generated

Figure 3:

Line chart of average house price

A graph showing the price of a house

Description automatically generated

## EDA of Broadband

The broadband speed dataset was analyzed to compare average and maximum download speeds in various cities and counties. A boxplot was created to illustrate the average download speed by county. This graph was useful for comparing internet speeds in different counties. We utilized bar charts to compare average and maximum download speeds in several towns/cities around Cornwall and Bristol. These graphics allowed for a clear comparison of broadband speeds within specified regions.

Figure 4:

Box Plot of Average and Max Download Speed

A screenshot of a computer

Description automatically generated

Figure 5:

Bar Chart of Avg and Max Download Speed

A screenshot of a graph

Description automatically generated

Figure 6:

Box plot of Average Download Speed

A screenshot of a graph

Description automatically generated

## EDA of Crime

Crime data representation use visual methods to investigate and analyze crime statistics, focusing on various crime categories and their distribution across counties. For example, radar charts show the pattern of vehicle offenses over time, indicating changes in crime rates. Pie charts depict the distribution of robberies by month in 2023, providing insight into temporal patterns. Additionally, boxplots compare drug offence rates across counties, allowing for a comparative analysis of crime severity.

Figure 7:

Radar Chart of Vehicle Crime Rate

A screenshot of a computer

Description automatically generated

Figure 8:

Pie Chart of Robberies

A colorful pie chart with numbers and a graph

Description automatically generated

Figure 9:

Box Plot of Drug Offense Rate

A screenshot of a computer

Description automatically generated

## EDA of School

School data representation comprises visualizing academic achievement metrics among educational institutions. In this context, boxplots are used to compare the average Attainment 8 scores for schools in various counties in 2022, and line graphs follow these scores over time for specific schools within counties such as Bristol and Cornwall.

Figure 10:

Box Plot of Average Attainment 8 Scores

A graph with red and blue squares

Description automatically generated

Figure 11:

Average Attainment score of Bristol

A graph with purple lines and dots

Description automatically generated

Figure 12:

Average Attainment Score of Cornwall

A graph with lines and dots

Description automatically generated with medium confidence

# Linear Model

Linear Model also known as “Linear Regression” is a statistical modeling technique used to estimate the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. With one explanatory variable, it is a simple linear regression, while the inclusion of multiple variables is termed multiple linear regression. The model assumes that the relationship between the variables is linear, and it aims to estimate the parameters that best describe this relationship by minimizing the differences between observed and predicted values. Linear regression models are widely used due to their simplicity and the ease with which they can be fitted and interpreted, making them fundamental tools in statistical analysis and predictive modeling.

## House Price vs Average Download Speed

In this analysis, the code explores the relationship between house prices and average download speeds. It selects relevant columns from the housing and broadband datasets, merges them by postcode and creates a scatter plot. The plot includes a linear regression line to illustrate the potential influence of internet speed on house prices.

Figure 13:

House Price Vs Average Download Speed

A graph with a purple line

Description automatically generated

## Attainment 8 score VS House Price

This section examines the connection between educational attainment and house prices. It filters the school data for the year 2022 and merges it with housing data by postcode. The resulting scatterplot, enhanced with a linear regression line, which visualizes how the average Attainment 8 score, measure of student performance, correlates with house prices, potentially indicating the value of living in areas with better educational outcomes.

Figure 14:

Attainment 8 Score Vs House Price

A graph with a line and a purple line

Description automatically generated with medium confidence

## Average Download Speed VS Attainment 8 score

The analysis examines whether there was a relationship between broadband speed and educational performance. It merges the broadband data with school data for 2022 by postcode and generates scatterplot. The plot includes a linear regression line which helps to assess whether faster download speeds are associated with higher Attainment 8 scores, exploring the impact of internet access on education.

Figure 15:

Download Speed Vs Attainment 8 Score

A graph with a purple line

Description automatically generated

## House Price vs Drug Rates

In this analysis, the focus was on how drug crime rates affect house prices. The code filters and aggregates crime data for drug-related offenses in 2023, then merges it with housing data by postcode. A scatter plot with a linear regression line is created to examine whether higher drug crime rates in an area are linked to lower house prices, reflecting the impact on property values.

Figure 16:

House Price Vs Drug Rates

A screen shot of a graph

Description automatically generated

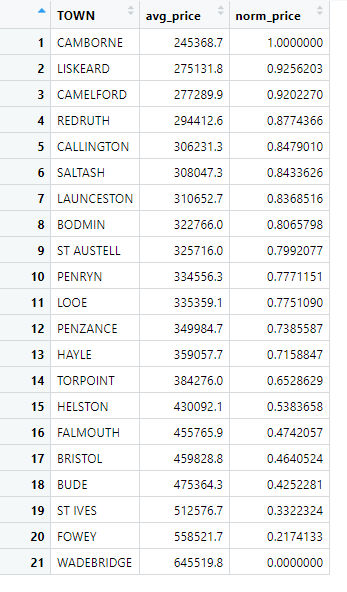
# Recommendation System

## House Price Ranking

The housing recommendation system identifies the town with the highest average housing price, which is Camborne. The system ranks towns based on the normalized housing prices, with the highest normalized score indicating the most expensive area. The recommendation is made by analyzing the housing data, filtering it to include only relevant fields like the average price for the year 2023, and then ranking the towns based on the normalized values. Camborne tops the list with an average price of £245368.7 and a normalized price score of 1.0000000, making it the most expensive town for housing in this analysis.

Figure 17:

House Price Ranking

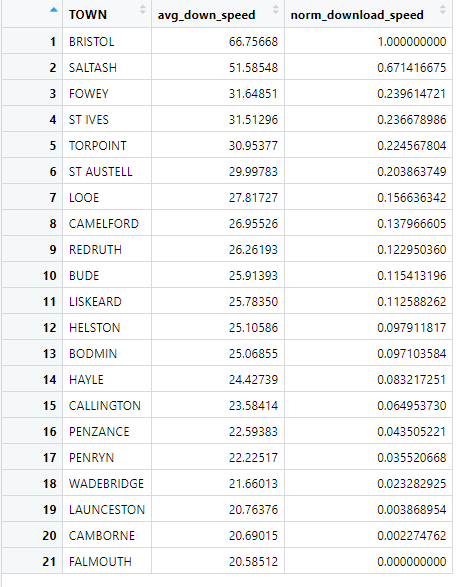


## Broadband Speed Ranking

The broadband recommendation system ranks towns based on average download speeds. The recommendation identifies Bristol as the top town, with an average download speed of 66.75 Mbit/s and a normalized download speed score of 1.0000000. This ranking is determined by grouping the broadband data by town, calculating the mean download speed, and then normalizing these values. Bristol's high-speed internet makes it the top recommendation for broadband.

Figure 18:

Broadband Ranking



## Crime Score Ranking

For crime rates, the system recommends Fowey as the town with the highest crime rate, with a crime rate of 98 and a normalized crime rate score of 1.0000000. This recommendation is derived from filtering crime data for the year 2023, calculating the total number of crimes by postcode, and aggregating these values by town. Towns are then ranked based on their normalized crime rates, highlighting Fowey as the area with the most concerning crime statistics.

Figure 19:

Crime Ranking

A table with numbers and numbers

Description automatically generated

## School Score Ranking

The school recommendation system evaluates towns based on their average attainment 8 scores, with St Ives ranking at the top. St Ives has an average attainment 8 score of 52.40 and a normalized score of 1.0000000. This ranking was obtained by grouping school data by town, calculating the mean attainment 8 scores, and normalizing these values to identify the top-performing educational areas.

Figure 20:

School Ranking

A table of numbers and numbers

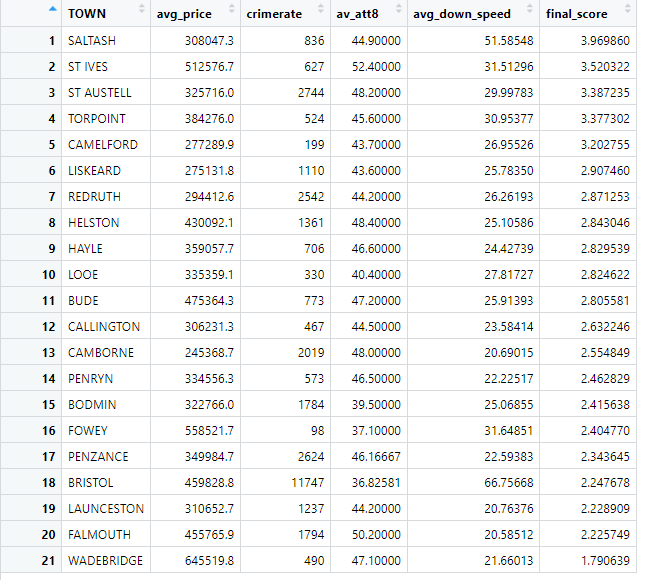
Description automatically generated

## Overall Ranking

The overall recommendation system combines the scores from housing, broadband, crime, and schools to provide a comprehensive ranking of towns. Saltash emerges as the top town with a balanced performance across all categories. It has an average housing price of £30,8074.20, a crime rate of 836, an average attainment 8 score of 44.90, and an average download speed of 51.58548 Mbit/s. The final score of 3.969860 makes Saltash the top overall recommendation for living, considering all the factors analyzed followed by St Ives and St Austell.

Figure 21:

Final Ranking



# Legal and Ethical Issues

The datasets for this analysis were obtained from the official UK government website, ensuring that all data is publicly available and utilized in accordance with legal standards. The materials were obtained with the explicit intention of being used for educational purposes, in accordance with the terms of use specified on the government’s website. This method ensures that the analysis is carried out within the legal and ethical parameters, while also protecting data privacy and intellectual property rights. The datasets, which include information on housing, broadband speeds, school performance, and crime rates, are intended solely for academic and research purposes, demonstrating the commitment to appropriate data handling and usage.

# Reflection

To ensure the accuracy of the results, data was carefully gathered from trustworthy sources, with a focus on acquiring information directly from the official UK government website. The project used R programming, a well-known data science tool, to carry out the analysis. Recognizing the possibility of inaccuracies in raw data, a strong emphasis was placed on the cleaning process. This involved removing irrelevant, incomplete, or inaccurate data to create polished datasets, which were then stored as .csv files.

Throughout this project, various visualization techniques, such as box plots, bar graphs, and linear graphs, were used to make the cleaned data easier to understand. Linear modeling was employed to explore relationships between different factors, including the connection between house prices and download speeds, as well as crime rates and educational attainment. This analysis helped in categorizing counties by their performance in these areas. In the end, the county with the most favorable outcomes was identified, and recommendations were made based on this comprehensive comparison.

# Conclusion

The integration of several datasets and the use of modern data science techniques were critical in this project. By using R programming for complete data cleaning and visualization, insightful patterns emerged, revealing the complex links between property values, broadband speeds, crime rates, and academic accomplishment. The linear modeling approach provides a quantitative foundation for comprehending these relationships, demonstrating how each element affects county outcomes. This study successfully illustrated how merging numerous data sources can provide a comprehensive understanding of regional dynamics. The findings underline the importance of data science in translating complex facts into actionable insights, directing decision-making, and identifying major predictors of living conditions. Finally, this research provides a strong framework for evaluating and comparing county performance, allowing for better informed, data-driven decisions.

# References

* *Price paid data*. (2024, July 26). GOV.UK. <https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads>
* Takyar, A., & Takyar, A. (2023, May 6). *Exploratory Data Analysis - a Comprehensive guide to EDA*. LeewayHertz - AI Development Company. <https://www.leewayhertz.com/what-is-exploratory-data-analysis/>
* Wikipedia contributors. (2024, August 9). *Linear regression*. Wikipedia. <https://en.wikipedia.org/wiki/Linear_regression>
* Zabrodski, A. (2024, June 14). *Introduction to data cleaning in R – dataquest*. Dataquest. <https://www.dataquest.io/course/r-data-viz-2/>
* Hashemi-Pour, C., Brush, K., & Burns, E. (2024, August 16). *What is data visualization and why is it important?* Business Analytics. <https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization>
* Hall, D. (2024, June 18). *Data Cleaning: Definition, Techniques & Best Practices for 2024*. TechnologyAdvice. <https://technologyadvice.com/blog/information-technology/data-cleaning/>

# Appendix

GitHub Link: <https://github.com/ItachiPrabin/Data-Science>

Google Drive Link: <https://drive.google.com/drive/folders/1te--i11iZGIE90UxuDT4-3a79klfw7kO?usp=drive_link>

## Data Cleaning

Figure 22:

House dataset cleaning

A computer code with text

Description automatically generated with medium confidence

Figure 23:

Broadband Dataset Cleaning

A screenshot of a computer program

Description automatically generated

Figure 24:

Crime Dataset Cleaning

A computer screen shot of a computer code

Description automatically generated

Figure 25:

School Dataset Cleaning

A screenshot of a computer

Description automatically generated

## Graphs

Figure 26:

Housing Dataset

A screenshot of a computer program

Description automatically generated

Figure 27:

Broadband dataset

A screenshot of a computer

Description automatically generated

A computer screen shot of text

Description automatically generated

Figure 28:

Crime Dataset

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

Figure 29:

School Dataset

A screenshot of a computer program

Description automatically generated

## Ranking Code

Figure 30:

House Ranking Code

A computer screen shot of a program

Description automatically generated

Figure 31:

Broadband ranking code

A computer code on a white background

Description automatically generated

Figure 32:

Crime ranking code

A screenshot of a computer

Description automatically generated

Figure 33:

School ranking code

A screenshot of a computer program

Description automatically generated

Figure 34:

Final Ranking code

A screen shot of a computer code

Description automatically generated