A black background with blue text

Description automatically generated

**Town Recommendation System and Report on an Individual Data Science Project**

Sandesh Sapkota

Softwarica College of IT and E-Commerce, Coventry University

ST5014CEM Data Science for Developers

Siddhartha Neupane

August 2024

# Table of Contents

[Table of Contents 2](#_Toc175003413)

[Table of Figures 5](#_Toc175003414)

[Introduction 6](#_Toc175003415)

[Cleaning Data 7](#_Toc175003416)

[Cleaning Housing Data 7](#_Toc175003417)

[Cleaning Broadband Speed 7](#_Toc175003418)

[Cleaning Crime Data 8](#_Toc175003419)

[Cleaning School Data 8](#_Toc175003420)

[Exploratory Data Analysis 10](#_Toc175003421)

[Visualization of Housing Price 10](#_Toc175003422)

[Bar Chart: Average House Price by Town in 2023 10](#_Toc175003423)

[Bar Chart: Average House Price in 2023 11](#_Toc175003424)

[Line Chart: Average House Prices from 2020 to 2023 12](#_Toc175003425)

[Visualization of Broadband Speed Data 13](#_Toc175003426)

[Box plot: Average Download Speed by County 13](#_Toc175003427)

[Bar Chart: Average and Maximum Download Speeds in Cornwall 14](#_Toc175003428)

[Bar Chart: Average and Maximum Download Speeds Bristol 15](#_Toc175003429)

[Visualization of Crime Data 16](#_Toc175003430)

[Radar Chart: Vehicle Crime Rate per 10,000 people 16](#_Toc175003431)

[Pie Chart: Robbery in 2023 by Month 17](#_Toc175003432)

[Boxplot: Drug Offence Rate per 10,000 People in 2023 18](#_Toc175003433)

[Visualization of School Data 19](#_Toc175003434)

[Line Graph: Average Attainment 8 scores for 2021-2022 in Cornwall 19](#_Toc175003435)

[Line Graph: Average Attainment 8 scores for 2021-2022 in Bristol 20](#_Toc175003436)

[Linear Modeling 22](#_Toc175003437)

[Housing Price Vs. Average Download Speed 22](#_Toc175003438)

[Average Attainment 8 Score Vs Housing Price 22](#_Toc175003439)

[Housing Price Vs Drug Rates 23](#_Toc175003440)

[Average Download Speed vs Attainment 8 scores 24](#_Toc175003441)

[Town Recommendation System 25](#_Toc175003442)

[Overview 25](#_Toc175003443)

[Results 25](#_Toc175003444)

[Based on Housing Price Ranking 25](#_Toc175003445)

[Based on Crime Rates Ranking 26](#_Toc175003446)

[Based on School Performance 27](#_Toc175003447)

[Based on Broadband 27](#_Toc175003448)

[Overall Ranking 28](#_Toc175003449)

[Reflection 28](#_Toc175003450)

[Legal and Ethical Considerations 29](#_Toc175003451)

[Conclusion 30](#_Toc175003452)

[References 31](#_Toc175003453)

[Appendix 32](#_Toc175003454)

# Table of Figures

[Figure 1: Barchart: Average House Price by Town 11](#_Toc175003393)

[Figure 2: Bar Chart: Average House Price in 2023 12](#_Toc175003394)

[Figure 3: Line Chart of Average House Price 13](#_Toc175003395)

[Figure 4: Boxplot: Average Download Speed by County 14](#_Toc175003396)

[Figure 5: Barchart of Average and Max Download Speed in Cornwall 15](#_Toc175003397)

[Figure 6: Barchart of Average and Max Downlaod Speed in Bristol 16](#_Toc175003398)

[Figure 7: Radar Chart of Vehicle Crime Rate 17](#_Toc175003399)

[Figure 8: Pie Chart: Robberies by Month in 2023 18](#_Toc175003400)

[Figure 9: Boxplot : Drug Offence per 10000 19](#_Toc175003401)

[Figure 10: Linegraph of Attainment score in Cornwall 20](#_Toc175003402)

[Figure 11: Linegraph of Attainment score in Bristol 21](#_Toc175003403)

[Figure 12: House Price Vs Average Download Speed 22](#_Toc175003404)

[Figure 13: Average Attainment Vs House price 23](#_Toc175003405)

[Figure 14: House Price vs Drug Rate 24](#_Toc175003406)

[Figure 15: Average Downlaod Speed Vs Attainment 8 24](#_Toc175003407)

[Figure 16: Top 10 towns based on House Price 25](#_Toc175003408)

[Figure 17: Top 10 Towns based on Crime Rates\ 26](#_Toc175003409)

[Figure 18:Top 10 based on Schools 27](#_Toc175003410)

[Figure 19 Top 10 based on Broadband 27](#_Toc175003411)

[Figure 20 Overall Top 10 28](#_Toc175003412)

# Introduction

This assignment shows the process and outcomes of using data analysis to recommend the appropriate city in the United Kingdom to purchase a property in the United Kingdom and particularly in Cornwall or Bristol counties. Each city has their own advantages but buying a house requires detailed analysis of several factors influencing property valuation and quality of life. The primary factors that are usually considered while buying a property include the housing price, crime rate in the neighborhood, internet connectivity, education and so on. This project utilizes the data made available by the government of United Kingdom and public sources to facilitate buyers to make informed decisions.

The objective of this project is to develop a recommendation system that uses data to analyze and compare the cities of Bristol and Cornwall based on several factors. The system provides a score to each city of the two counties based on the given criteria and rank the towns accordingly. The final goal is to suggest the top three cities to purchase a house to the friend in the context. This report outlines the process of obtaining the data, cleaning and preprocessing, performing exploratory data analysis and linear modeling to develop the recommendation system.

# Cleaning Data

Data cleaning is an important step in the data science lifecycle which comes right after obtaining the datasets. This step ensures that the datasets are accurate and consistent and prepares them appropriately to be used for analysis. The datasets obtained from the UK government and other public institutions also had several inconsistencies and inaccuracies. Each dataset were carefully cleaned and prepared to be used for the further analysis and development of the recommendation system.

## Cleaning Housing Data

The housing price data from 2020 and 2023 were combined into one using the bind\_rows() function in R. Then the data was filtered to include the housing data of Cornwall and Bristol only. A new column was added named “Year” by extracting the year from the existing column “Transaction\_Date”. After that only the required columns were selected. Finally, the null values and redundant entries were removed using na.omit() and distinct() respectively. Now, the dataset was saved using the write\_csv() function.

## Cleaning Broadband Speed

The broadband speed dataset was loaded using read\_csv() function selecting the relevant columns. The the columns were renamed to simpler names for easier operation. For example, “Median download speed (Mbit/s)” was renamed to “MedianDownSpeed”. The final columns were ‘Postcode’, ‘MedianDownSpeed’, ‘MedianUpSpeed’, ‘AvgUpSpeed’, ‘MaxUpSpeed’, ‘AvgDownSpeed’ and ‘MaxDownSpeed’. Then the null values were removed using the na.omit() function. After that, the broadband data was merged with housing dataset using inner\_join() function in R. The two datasets were joined based on the common field “Postcode”. Finally, the redundant rows were removed using the distinct() function. In this way, the broadband speed dataset was cleaned and processed for further analysis. The cleaned dataset was then saved as a CSV using the write\_csv() function.

## Cleaning Crime Data

The crime datasets from Bristol and Devon & Cornwall in the years 2020 to 2024 were loaded using the read\_csv(). Then all the datasets were combined together using the rbind() functions and converted into a tibble. Since the crime dataset did not include postcodes but includes the LSOA codes, another dataset postcodes to LSOA was also loaded and cleaned to join with the crime dataset. Only the required columns were selected from the LSOA and crime data. From the combined crime dataset, month, LSOA code, crime type, and falls within columns were selected. The columns were renamed for simplicity in further processing. Similarly, from the LSOA to postcode dataset, only ‘lsoa11cd’, ‘lsoa11nm’, ‘ladnm’, ‘pcds’ were selected. The data frame was then filtered to include the counties Bristol and Cornwall only. The duplicate values for LSOA code were checked and removed from both crime and lsoa datasets. Finally, the selected crime dataset was merged with the lsoa dataset using left join by the column LSOA code common in both datasets. Then two new columns ‘Year’ and ‘Month’ were created using the mutate function from the original Month column by trimming from 1 to 4 and 6 to 7 respectively. Now, the population data was also merged using left join. Finally, distinct() and na.omit() functions were used to remove the redundant rows and null values to clean the final data. Now, the dataset was saved using the write\_csv() function.

## Cleaning School Data

The performance school datasets from the academic years 2021 to 2022 and 2022 to 2023 were loaded using the read\_csv() function from the readR library. Then the datasets were filtered to only include the relevant columns ‘SCHNAME’, ‘PCODE’, ‘ATT8SCR’, ‘TOWN’. The columns ‘Year’ and ‘County’ were added to each dataset using the mutate() function. Then the data from both academic years were combined for each county using the rbind() function. The combined dataset was now cleaned by filtering out the non-numeric values ‘NE’ and ‘SUPP’ in the ATT8SCR column. The null values and redundant rows were removed using the na.omit() and distinct() functions respectively. The final dataset was then saved as comma separated values (CSV) using the write\_csv() function.

# Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach in data analysis which uses graphical representation and data visualization to summarize the features of data. After cleaning and preprocessing the datasets, the trends were analyzed by visualizing the data. Several visualization techniques such as box plots, line graph, radar chart, bar chart and pie chart were used for better understanding of distribution, relationships and main features of the data.

## Visualization of Housing Price

### Bar Chart: Average House Price by Town in 2023

The data was filtered to only include the year 2023 then the average housing price is calculated using the summarize() method. Then the data is visualized in a bar chart using the ggplot2 library. The cities were plotted in the x-axis and the average price was plotted in the y-axis in the bar chart. The bar chart shows the average housing price in various towns in Bristol and Cornwall for the year 2023. Boscastle town is seen to have a significantly higher average housing price among the cities. Boscastle is followed by Port Isaac, Padstow, and Waderbridge as the most expensive cities to own a house. Camborne, Callington, Redruth are some of the least expensive cities to consider.

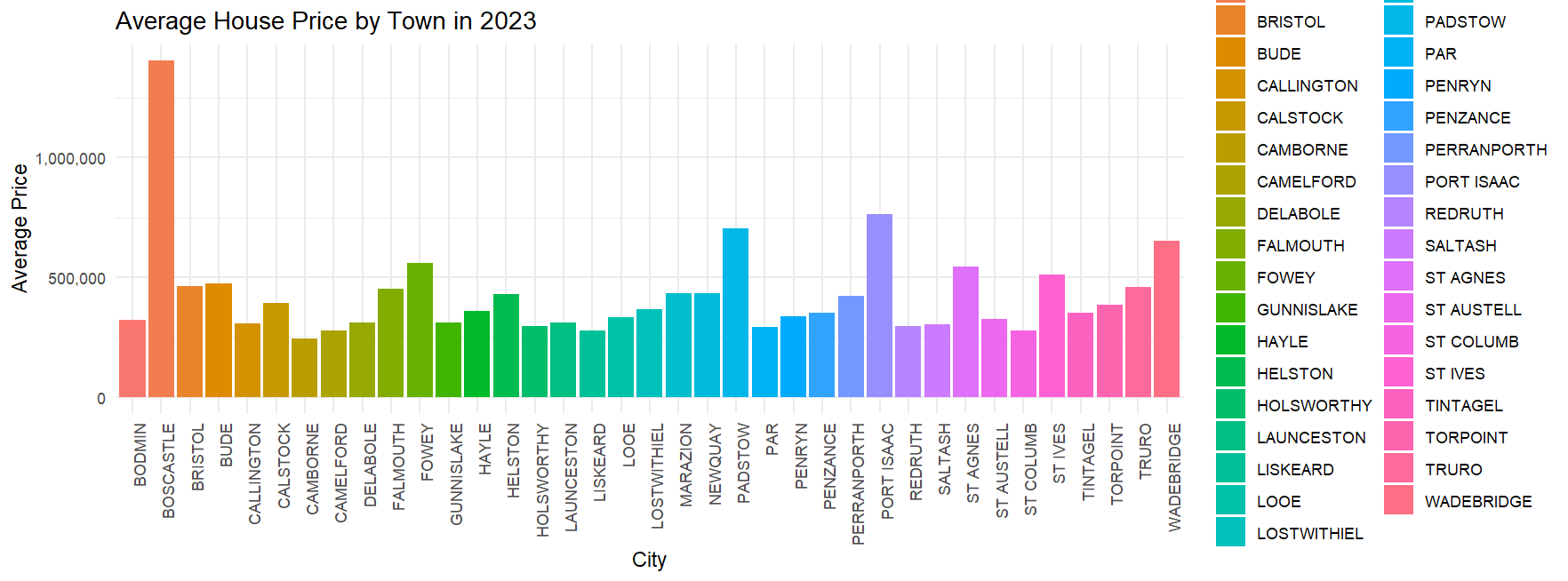


Figure 1: Barchart: Average House Price by Town

### Bar Chart: Average House Price in 2023

In this chart, the average house prices in 2023 were visualized by county wise in a bar chart. The data was first filtered to include only from the year 2023 and then grouped by County using group\_by() function. Then the average house price was summarized by calculating the mean of ‘Price column’. The data was then visualized in a bar chart using ggplot2 to visualize the County to the x-axis and average house price to the y-axis. The chart shows that average house prices are noticeably higher in Bristol compared to Cornwall.

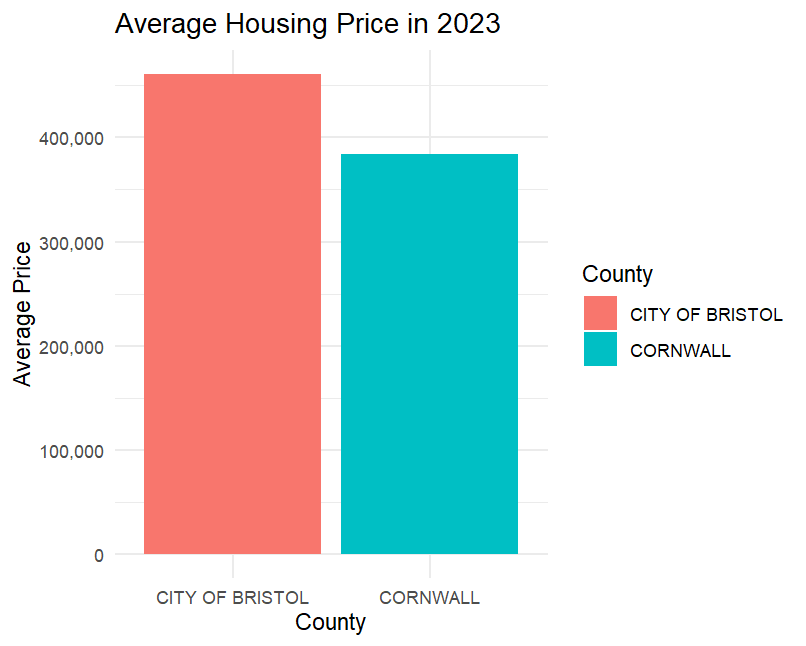


Figure 2: Bar Chart: Average House Price in 2023

### Line Chart: Average House Prices from 2020 to 2023

The data was filtered so that the year lies between 2020 to 2023 and grouped by Year and County. Then the average price was calculated using summarise() method. Then the line chart was plotted using the ggplot2 library with year on the x-axis and average price price on the y-axis. Looking at the line chart, the city of Bristol has consistently had higher prices compared to Cornwall throughout the year. After 2022, the housing market in Cornwall has seen more stability while the gap between the two counties has widened more in recent years.

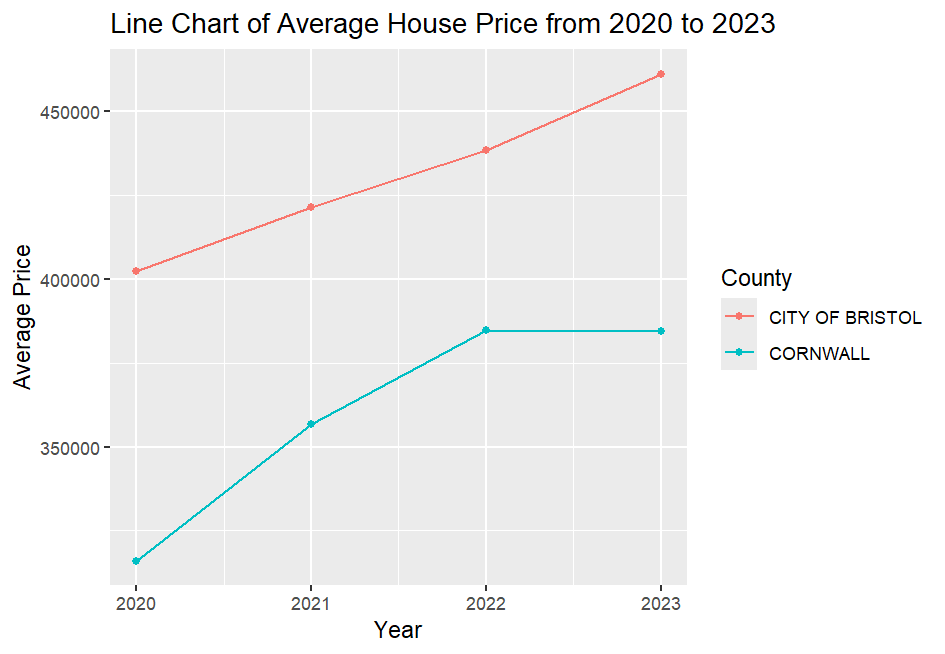


Figure 3: Line Chart of Average House Price

## Visualization of Broadband Speed Data

### Box plot: Average Download Speed by County

The cleaned broadband data is loaded from the csv file. Then the box plot is created using ggplot2 by grouping the average download speed and county. The box plot then visualizes the distribution of download speeds across the two counties. The box plot shows that Bristol has higher and more consistent download speeds with median around 80-90 Mbit/s while Cornwall has lower speed with median of 50 Mbits/s. The outliers denote the areas with significantly higher or lower speeds in both counties.

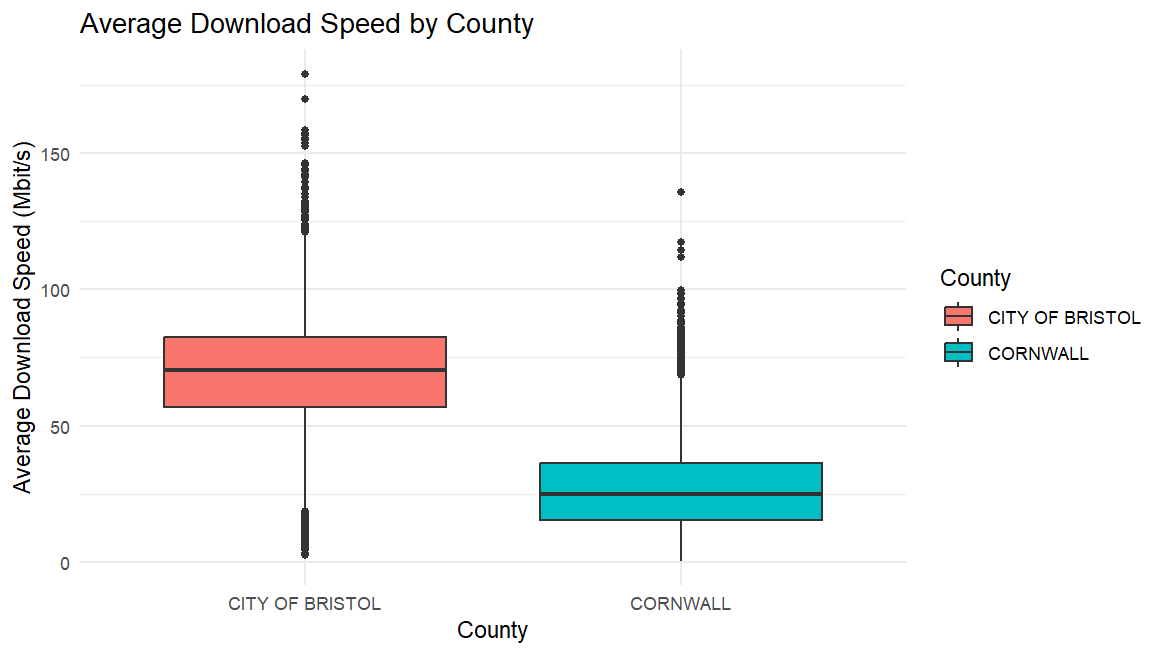


Figure 4: Boxplot: Average Download Speed by County

### Bar Chart: Average and Maximum Download Speeds in Cornwall

The data was filtered to only include the county Cornwall and grouped by Town/City. Then the average and maximum download speeds were calculated using the summarize() method. The bar chart was plotted using ggplot2 which reveals significant differences in internet quality across Cornwall. The towns Saltash and St Austell have very high maximum speeds while the average among most towns is moderate.

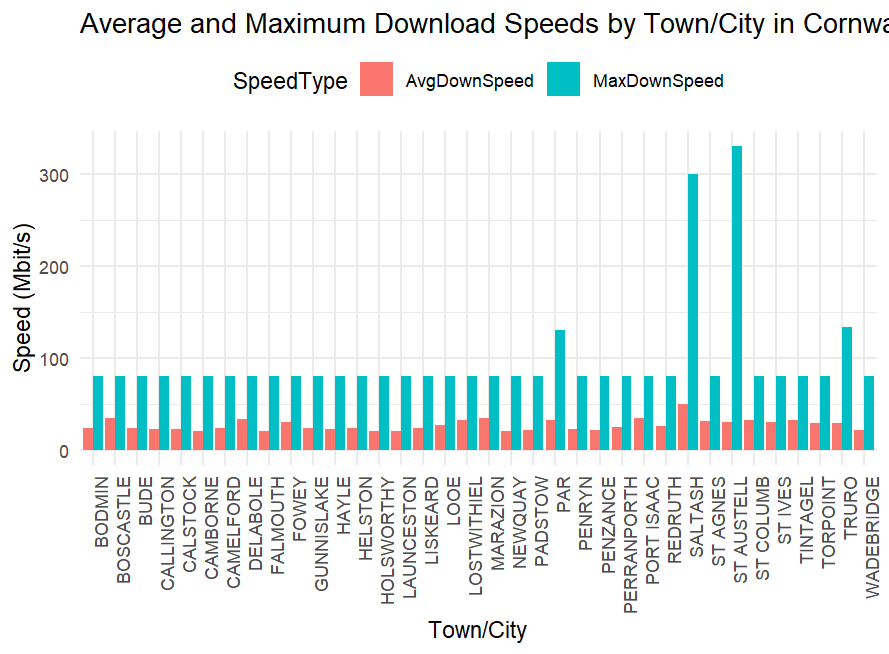


Figure 5: Barchart of Average and Max Download Speed in Cornwall

### Bar Chart: Average and Maximum Download Speeds Bristol

The data was filtered to include the data of Bristol only and grouped by Town/City. Then the average and maximum speeds were calculated using summarize() function before visualizing the bar chart using ggplot2. From the bar chart, the city’s internet service is not uniformly distributed as there is a big gap between the maximum possible speed and the average speed most users get in the City of Bristol.

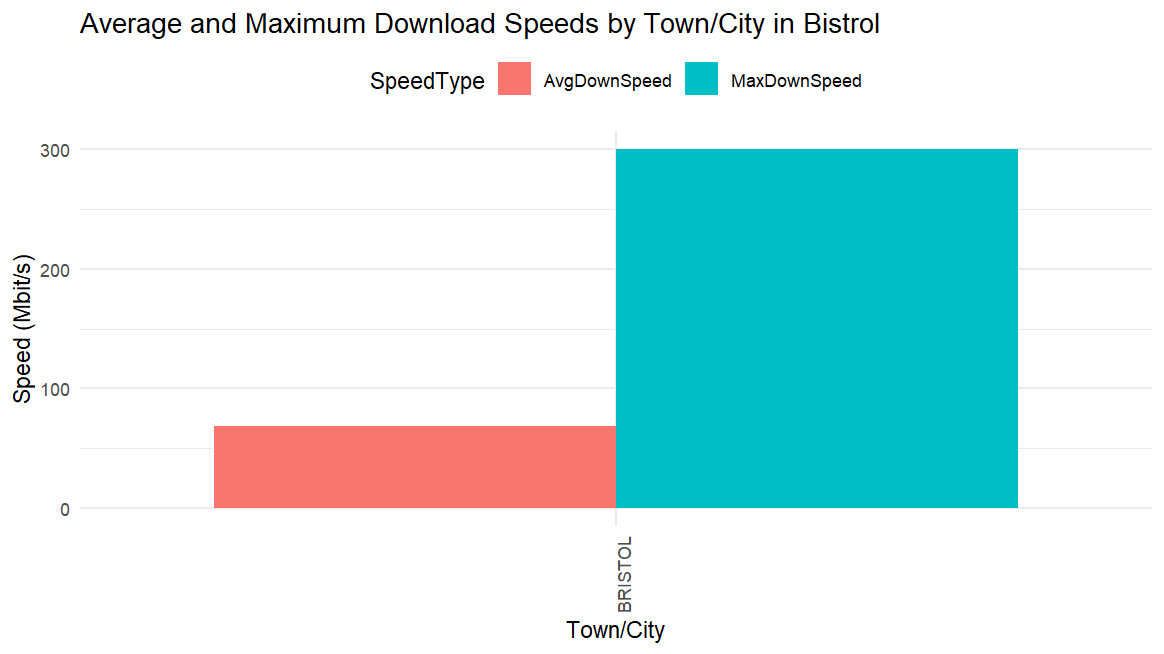


Figure 6: Barchart of Average and Max Downlaod Speed in Bristol

## Visualization of Crime Data

### Radar Chart: Vehicle Crime Rate per 10,000 people

The cleaned dataset is filtered to only include vehicle crimes and grouped by Year. Then the total crime is calculated by using the summarise() method. Then the data frame is prepared for the radar chart and plotted using the radarchart() function from ‘fmsb’ library in R. The chart shows the vehicle crimes were consistent from 2021 to 2024 with a few variations. The highest was in 2021 and the lowest was in 2023. Overall, the crime rates were stable across the years.

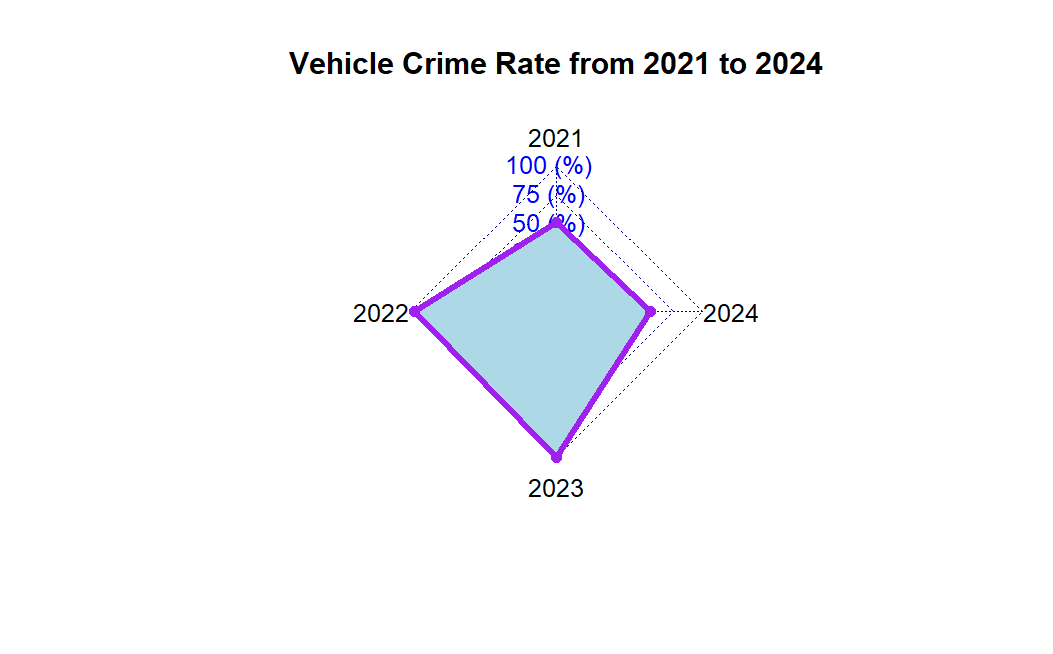


Figure 7: Radar Chart of Vehicle Crime Rate

### Pie Chart: Robbery in 2023 by Month

The data was filtered and summarized for robbery rates in each month in 2023. The data is then visualized in a pie chart using ggplot2. The chart shows that there was highest percentage of robberies at 13.9% in January, while April and September had the lowest at 5.1%. Other than that, the rate of robberies is even throughout the year.

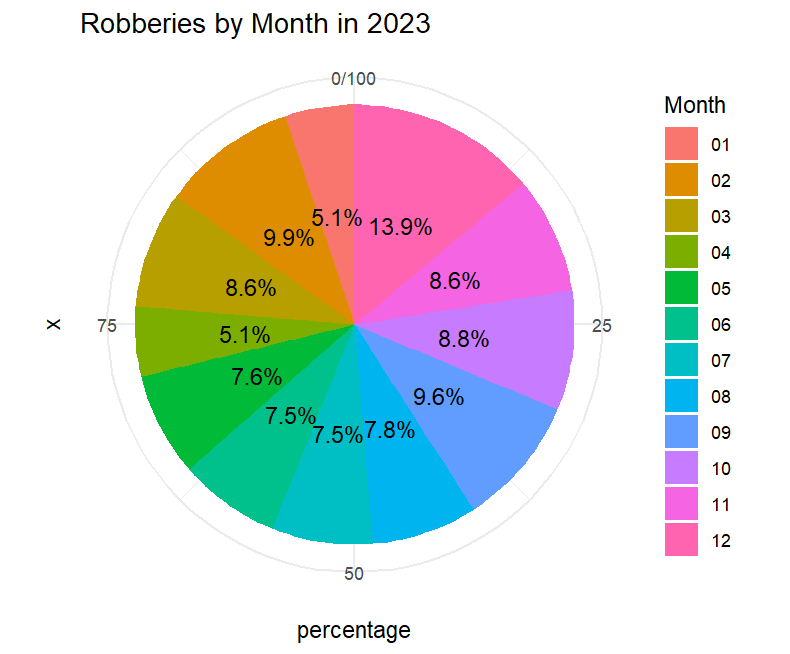


Figure 8: Pie Chart: Robberies by Month in 2023

### Boxplot: Drug Offence Rate per 10,000 People in 2023

The boxplot visualizes the distribution of drug related crimes in 2023 per 10,000 people in Bristol and Cornwall. The boxplot shows that the drug rates are higher in Cornwall compared to Bristol. Similarly, Cornwall has a wider spread indicating differences in offences in different regions while Bristol shows uniform distribution of data.

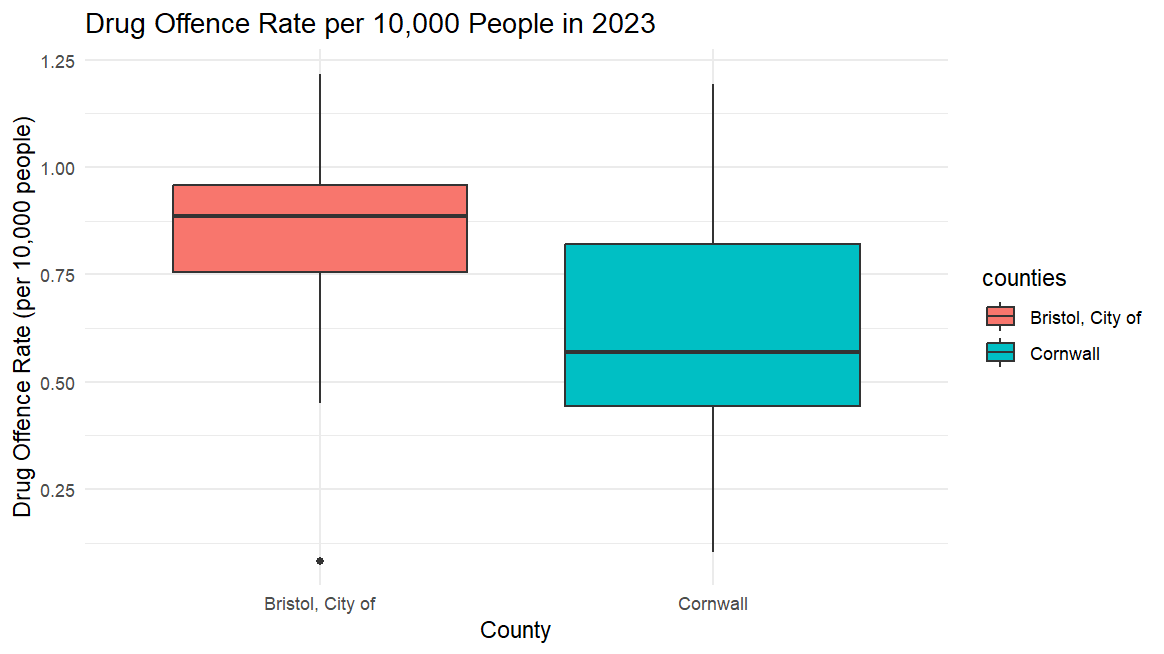


Figure 9: Boxplot : Drug Offence per 10000

## Visualization of School Data

### Line Graph: Average Attainment 8 scores for 2021-2022 in Cornwall

Attainment 8 scores is a way to measure the average performance of school students in the UK. The line line graph is created to show the average attainment 8 scores for different towns in Cornwall in academic year 2021-2022. Town is taken in x-axis and scores is taken in y-axis of the graph. It shows a significant variation across the towns. Towns like Cornwall, St Ives performed well while others like Truro have performed extremely poor. The others had an above average performance.

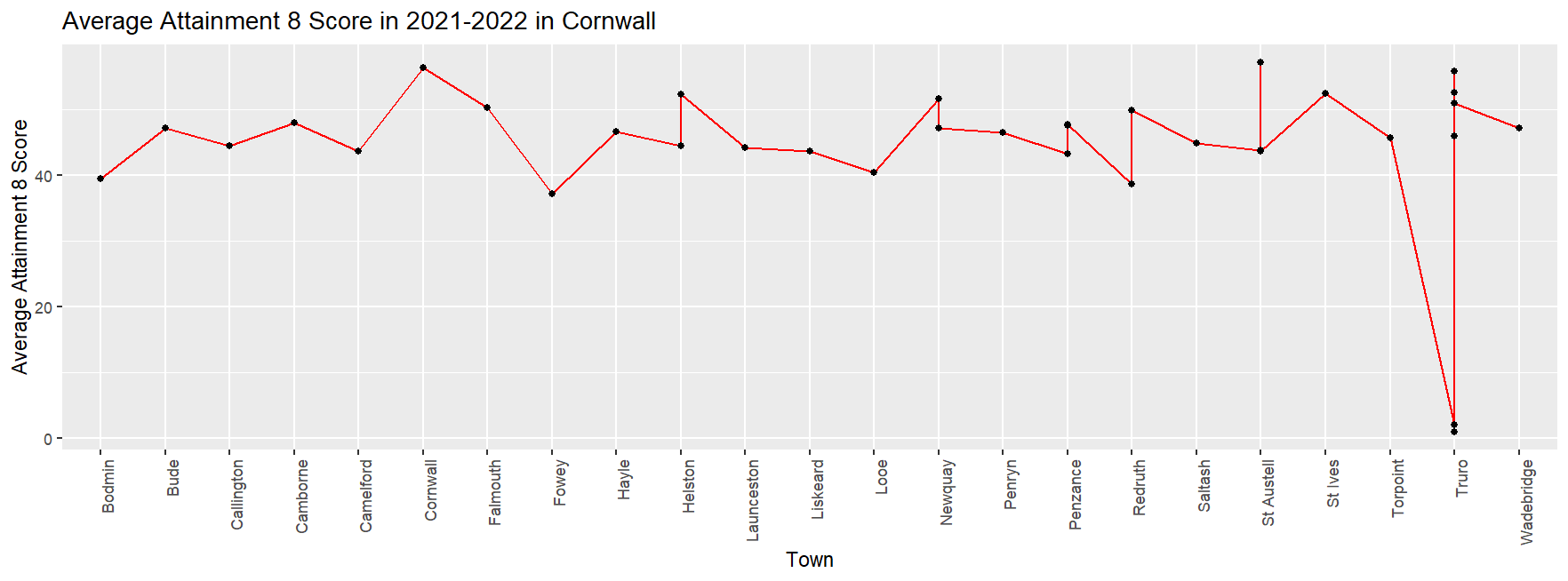


Figure 10: Linegraph of Attainment score in Cornwall

### Line Graph: Average Attainment 8 scores for 2021-2022 in Bristol

For Bristol, the visualization for attainment 8 scores was done using ggplot2 based on Schools as there weren’t significant data for the towns. In the line graph, x-axis represented Schools and y-axis contained their attainment scores. The graph shows schools like Montpelier High School peak the graph while schools like Venturers Academy show lower scores. This shows high academic disparities across the Bristol County.

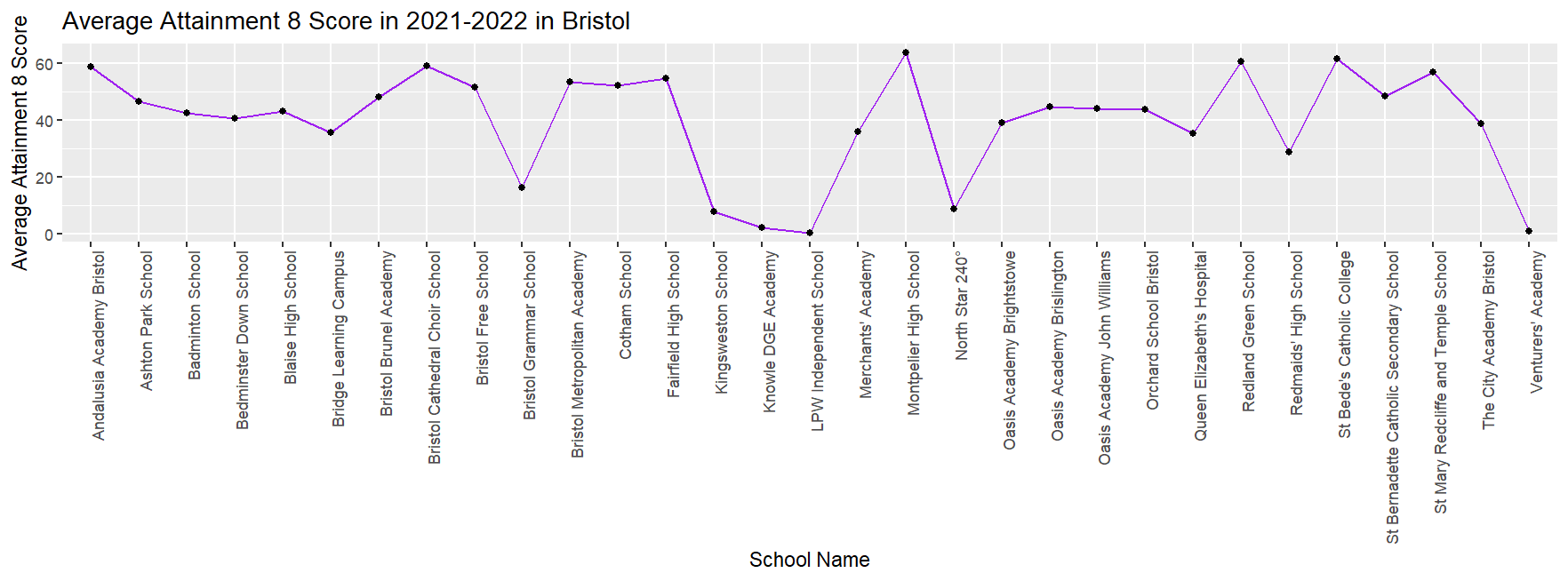


Figure 11: Linegraph of Attainment score in Bristol

# Linear Modeling

Linear Modeling is a statistical method used in data analysis to analyze the the relationship between two or more variables. It shows the linear relationship between a dependent variable and and one or more independent variables expressed in the form of a linear equation. This approach was used to identify the trends, patterns and make predictions based on the data of Cornwall and Bristol.

## Housing Price Vs. Average Download Speed

The shows a positive correlation between house price and average download speed. This suggests that higher the internet bandwidth, higher will be the housing price. This is likely because people with nicer house often want a good internet connection. However, there is a lot of outliers especially at the lower end of the house prices.

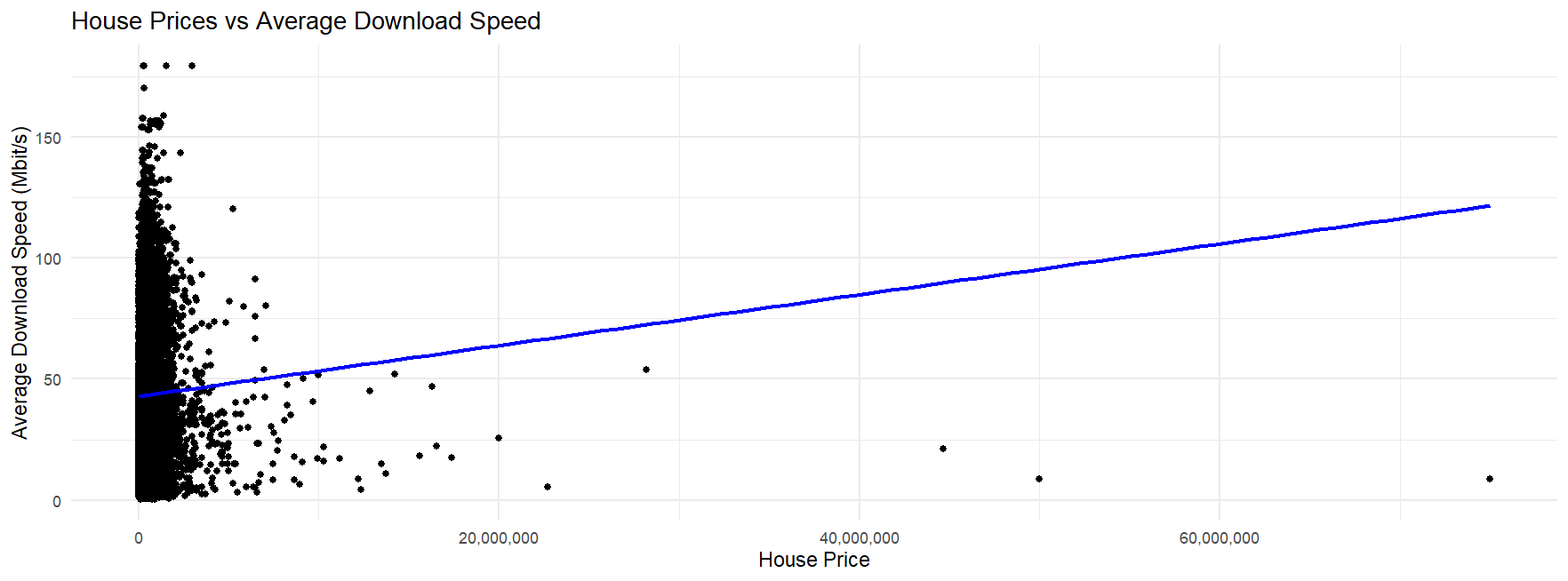


Figure 12: House Price Vs Average Download Speed

## Average Attainment 8 Score Vs Housing Price

The scatterplot shows a weak positive relationship between the attainment 8 score and housing prices in the two counties. This shows that the areas with good educational performance have higher housing price. The relationship is weak because the data points are widely scattered indicating that attainment 8 score might not have significant impact on the price.

A graph with a line

Description automatically generated

Figure 13: Average Attainment Vs House price

## Housing Price Vs Drug Rates

The plot shows a negative correlation between the housing prices and drug crime rates in the area. This suggests that the areas with expensive houses have lower drug offence rate while cheaper neighborhoods have higher drug crime rates. This might be because social factors like this affect the real estate market.

A graph with a line going up

Description automatically generated

Figure 14: House Price vs Drug Rate

## Average Download Speed vs Attainment 8 scores

A weak positive correlation is seen between the average download speed and attainment 8 scores of a region. Even though the scattered data points suggest the relationship was not too strong, it could be said that the better internet has a slight impact in the academic performance of schools.

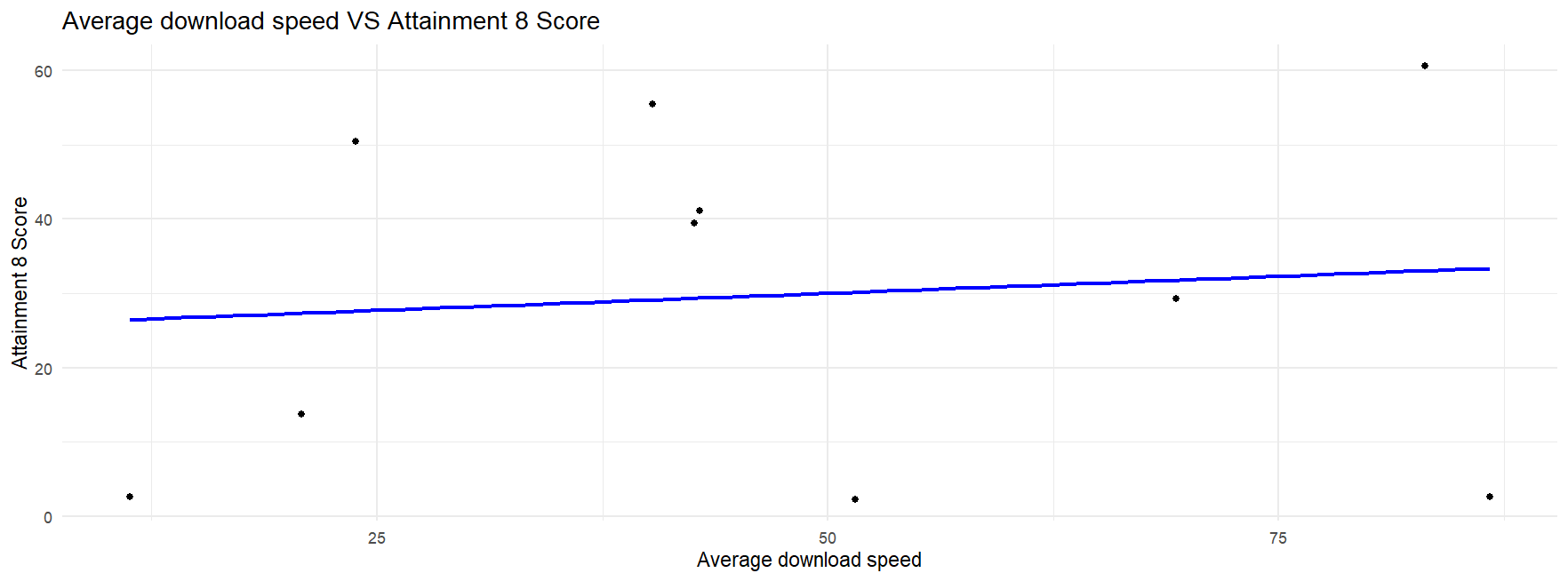


Figure 15: Average Downlaod Speed Vs Attainment 8

# Town Recommendation System

## Overview

After cleaning the obtained datasets, the data were prepared and visualized for exploratory data analysis, prepared linear models to compare the relationships between the different parameters in deciding the house purchase. Now, the task was to identify the most suitable town to purchase the house considering affordable housing, low crime rates, good internet connectivity, and good school performance measured by the attainment 8 scores. The preprocessed datasets were loaded and filtered and processed to calculate average for each factors by town. The data was then merged into a single data frame and normalized to ensure they can be compared later on. Finally, the towns are ranked based on the scores. Higher scores indicate more favorable to purchase a house or property in that location.

## Results

### Based on Housing Price Ranking

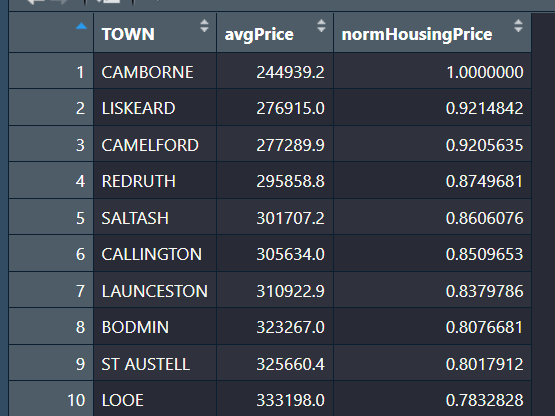


Figure 16: Top 10 towns based on House Price

### Based on Crime Rates Ranking

A screenshot of a computer

Description automatically generated

Figure 17: Top 10 Towns based on Crime Rates\

### Based on School Performance

A screenshot of a computer

Description automatically generated

Figure 18:Top 10 based on Schools

### Based on Broadband

A screenshot of a graph

Description automatically generated

Figure 19 Top 10 based on Broadband

### Overall Ranking

A screenshot of a computer

Description automatically generated

Figure 20 Overall Top 10

Ergo, after analyzing the towns on different parameters, the towns SALTASH, ST AUSTELL and ST IVES can be recommended for purchasing a house since these cities provide the most balanced parameters optimal housing price, less crime rates, good internet and good education.

## Reflection

The relevant datasets were obtained from reliable sources of the official websites UK government and other public organizations to maintain the credibility of the data. Then the analysis was done using the R programming language to clean, process, visualize, construct linear models and finally rank the towns based on different parameters. The data cleaning included filtering required columns, removing null values and redundant fields. Then the data were visualized using different visualization methods such as box plots, bar charts, line graphs and pie charts to find the patterns and trends. Similarly, linear modeling was done to analyse and compare the relationships between different parameters. This helped to classify the counties based on their performance in various factors. Finally, all the parameters were analyzed to identify the most favorable characteristics and recommendations were made on the basis of different comparative results.

# Legal and Ethical Considerations

The project has been done by considering all the ethical and legal compliances practiced in the industry. The datasets were obtained from the official websites of the UK government that are publicly available and approved by the government. The data weren’t manipulated to push a certain political ideology and are completely based on the UK government approved sources.

# Conclusion

This project demonstrated the use of data analysis to develop a recommendation system to figure out the optimal town in the Bristol and Cornwall counties in the United Kingdom to purchase a house. By cleaning and preprocessing the datasets obtained from credible sources, the parameters were compared to identify Saltash, St Austell and St Ives as the top three most favorable cities to buy a house considering affordable pricing, low crime rates, good internet and strong academic performance of local schools. It was a great learning experience to explore the capacities of data analysis in recommendation systems.

# References

Sisense. (2018, June 12). What is Data Cleaning? | Sisense. <https://www.sisense.com/glossary/data-cleaning/>

Stedman, C. (2022, January 28). data cleansing (data cleaning, data scrubbing). Data Management. <https://www.techtarget.com/searchdatamanagement/definition/data-scrubbing>

Data downloads | data.police.uk. (n.d.). <https://data.police.uk/data/>

Connected Nations 2018: Data downloads. (2019, January 2). www.ofcom.org.uk. <https://www.ofcom.org.uk/phones-and-broadband/coverage-and-speeds/data-downloads>

*RPUBS - Introduction to Linear Modeling in R*. (n.d.). <https://rpubs.com/AnthonyCorbisieri/1100671>

*Radar*. (n.d.). <https://plotly.com/r/radar-chart/>

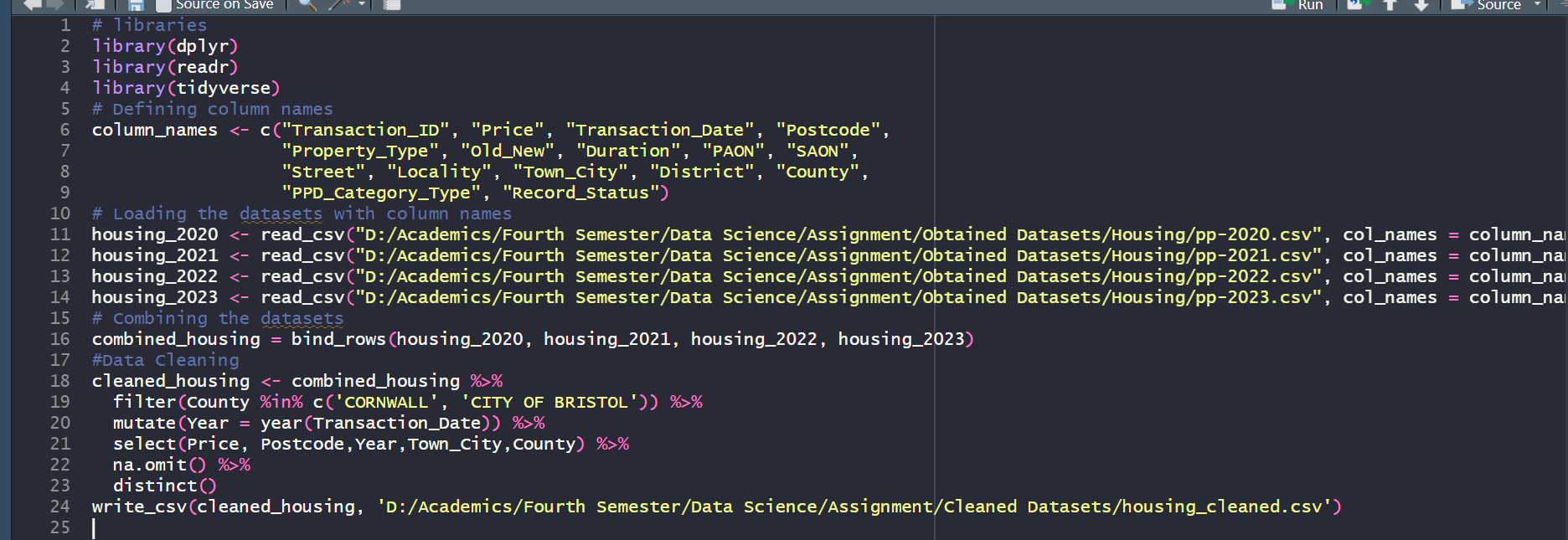
Chatterjee, S. (2024, February 26). *5 Ethical aspects for data science professionals to consider*. Emeritus Online Courses. <https://emeritus.org/blog/data-science-and-analytics-data-science-course-curriculum/>

# Appendix

**Github Link:** <https://github.com/sandeshs0/data-science-assignment>

**Google Drive Link:** <https://drive.google.com/drive/folders/11tJCS5Z5XRncfJ_cnh9a0_XBJJ31dKoX?usp=sharing>

**Cleaning.R**

****

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A black background with white text

Description automatically generated**

**Graphs.R**

**A screenshot of a computer program

Description automatically generated**

**A computer screen shot of text

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screen shot of a computer code

Description automatically generated**

**A computer screen shot of text

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

**LinearModel.R**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

**Ranking.R**

**A screen shot of a computer program

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A computer screen shot of a computer code

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**