in collaboration with

Softwarica College of IT & E-commerce | Coventry University

# Town Recommendation System

# And

# Report on an Individual Data Science Project


**Submitted to:**

Siddhartha Neupane

**Submitted by:**

Nirajan Mahato (220406)

**Table of Contents**

## List of Figures

## Abbreviations List

- CSV - Comma Separated Value
- LSOA – Lower Layer Super Output Area
- EDA – Exploratory Data Analysis
- DLM - Data lifecycle management

## Introduction

The decision to invest in property is considered significant, especially when looking at regions with distinct characteristics like Bristol and Cornwall in the United Kingdom. These areas offer a mix of urban and rural environments, each with its own unique lifestyle and potential for investment. An analysis of various critical factors, including housing prices, internet connectivity, local crime rates, and overall quality of life, has been requested to provide recommendations based on a thorough examination of these key factors affecting property investment.

The aim of this analysis is to provide a data-driven recommendation that takes into account budget constraints, safety considerations, and lifestyle preferences of potential investors. This report aims to guide them in choosing a location that offers the best combination of affordability, safety, reliable internet, and quality of life. The data needed for this analysis has been carefully chosen from reliable UK government sources and will be processed using R, a programming language suitable for statistical analysis and data visualization. This ensures that the results obtained are accurate and can support the decision-making process for this significant investment.

## Dataset obtained:

- **House Pricing Dataset (2020 to 2023):**

  This data is from the UK government and provides detailed info on property transactions from 2020 to 2023. It helps analyze housing market trends and assess property values, essential for understanding affordability and investment potential in different areas.

- **Broadband Speed Dataset:**

  Ofcom's dataset provides broadband performance data, including download and upload speeds, across different UK regions. It is crucial for evaluating internet connectivity in potential investment locations, especially for those who rely on stable internet for work and leisure.

- **Crime Dataset:**

  The UK police data portal contains records of reported crimes by type and location across various UK regions. This data helps assess the safety and security of different areas, which is crucial for potential property buyers. Understanding crime rates is key to evaluating the overall security and desirability of neighborhoods.

- **Population Dataset (2011 as baseline):**

  This dataset provides population figures from the 2011 census and estimates population growth up to 2023. It is crucial for analyzing demographic trends and understanding the impact of population changes on housing demand, local services, and regional attractiveness.

- **School Dataset:**

  From the UK government's school performance service, this dataset provides school performance data across regions, including key performance indicators. This information is crucial for families considering property investments, as school quality can influence decisions about where to buy. The dataset helps assess educational opportunities in potential investment regions.

# Data Cleaning and preprocessing

## House Prices

The data checking, cleaning, and preprocessing steps were crucial to ensure reliable and accurate analysis of housing prices. Initially, four separate datasets for housing prices from 2020 to 2023 were loaded into R, and their relevant columns were merged into a consolidated dataset using the bind_rows() function. Irrelevant columns were removed to streamline the dataset, followed by filtering to retain only records pertaining to Bristol and Cornwall. Missing values were handled by removing rows with excessive missing data (over 50% of their columns) and excluding columns with inadequate data. Remaining missing values in numeric columns were imputed with the median to provide a robust estimation. Duplicate entries were eliminated, and the Date column was standardized to a date format. The final cleaned dataset, organized in a tabular format, has rows representing unique property transactions and columns for specific attributes like price, date, post code, town, district, and county, facilitating comprehensive analyses such as average price calculations, regional comparisons, and correlation analyses. This structured approach supports efficient querying and manipulation of the data, enabling insightful evaluations of the housing market in Bristol and Cornwall.

```r
# Load necessary libraries
library(tidyverse)
library(dplyr)
library(ggplot2)

# Read housing data from CSV files for each year
housing_2020 = read_csv("D:/Sem4/Data Science/Obtained Data/Housing/pp-2020.csv")
housing_2021 = read_csv("D:/Sem4/Data Science/Obtained Data/Housing/pp-2021.csv")
housing_2022 = read_csv("D:/Sem4/Data Science/Obtained Data/Housing/pp-2022.csv")
housing_2023 = read_csv("D:/Sem4/Data Science/Obtained Data/Housing/pp-2023.csv")

# Rename columns to have consistent column names across all data frames
colnames(housing_2020) = colnames(housing_2021) = colnames(housing_2022) = colnames(housing_2023) =  c(
  "ID", "Price", "SaleDate", "Postcode", "Property_Type", "Old_New", "Durations", "PAON", "SOAN", "Street_Name",
  "Locality", "Town/City", "District", "County", "Category_PPD", "ActiveStatus"
)

# Combine all the yearly data frames into one data frame using pipelines
house_selling_clean = bind_rows(housing_2020, housing_2021, housing_2022, housing_2023) %>%
  # Subset to keep only the desired columns
  select(ID, Postcode, Price, SaleDate, `Town/City`, County) %>%
  # Remove rows with any missing values
  na.omit() %>%
  # Remove duplicate rows
  distinct() %>%
  # Filter the data for specific counties
  filter(County == 'CITY OF BRISTOL' | County == 'CORNWALL') %>%
  # Convert SaleDate to Date format and extract the year
  mutate(SaleDate = as.Date(SaleDate, format = "%Y-%m-%d"),
         SaleYear = format(SaleDate, "%Y"),
         # Extract and format ShortPostcode
         ShortPostcode = sub("(.{3,4})\\s*(.).*", "\\1 \\2", Postcode)) %>%
  # Select only required columns
  select(Price, SaleYear, Postcode, ShortPostcode, `Town/City`, County)

# Convert the cleaned data frame to a tibble for better display and manipulation
house_selling_clean = as_tibble(house_selling_clean)
```

*Figure 1: Code of Housing*

## Broadband

The broadband speed datasets were imported into R using the read.csv() function. Two separate files contained different broadband performance and coverage metrics. These datasets were combined with an inner join on the common key postcode space, aligning with pcds in the second dataset, to retain only records with matching postcodes. This integration was crucial to joining the broadband performance data with the relevant geographical coverage information, resulting in a more comprehensive dataset for analysis.

After the integration, unnecessary columns such as pcd7, pcd8, doterm, usertype, and various geographical codes were removed from the dataset using the select() function, focusing on essential variables related to broadband speeds and data usage. Furthermore, key columns were renamed for clarity and consistency, such as AvgDownloadSpeed, MinDownloadSpeed, MaxDownloadSpeed, AvgUploadSpeed, MedUploadSpeed, MinUploadSpeed, MaxUploadSpeed, and AvgDataUsage. Geographical identifiers were also standardized by renaming the pcds column to Post_code and ladnm to County, further enhancing the dataset's clarity.

The final dataset consists of rows representing individual postcodes, with columns capturing various broadband metrics such as average, minimum, maximum download and upload speeds, as well as average data usage. This organization allows for detailed analysis of broadband performance across different regions, facilitating comparisons between areas and identifying trends or disparities in internet connectivity. This structured data is saved as a new CSV file, ensuring that the preprocessing steps are well-documented and reproducible.

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
library(readr)


# Read in the CSV files
broadband1 <- read.csv("D:/Sem4/Data Science/Obtained Data/Broadband Speed/201805_fixed_pc_performance_r03.csv")
broadband2 <- read.csv("D:/Sem4/Data Science/Obtained Data/Broadband Speed/201809_fixed_pc_coverage_r01.csv")


# View the data (optional)
View(broadband1)
View(broadband2)


# Perform the inner join
combined_data_broadbrand <- broadband1 %>%
  inner_join(broadband2, by = c("postcode_space" = "pcds"))

# View the combined data (optional)
View(combined_data_broadbrand)

# Perform the inner join
combined_data_broadbrand <- filtered_LSOA %>%
  inner_join(broadband1, by = c("pcds" = "postcode_space"))

# View the combined data (optional)
View(combined_data_broadbrand)

# Example: Deleting cMaximum.upload.speed..Mbit.s.# Example: Deleting cMaximum.upload.speed..Mbit.s.# Example: Deleting colum
cleaned_data <- combined_data_broadbrand %>%
  select(-c(pcd7, pcd8,doterm, usertype, oa11cd, lsoa11cd,msoa11cd,ladcd, lsoa11nm, msoa11nm, ladnmw,dointr, postcode.area ,
            postcode))
View(cleaned_data)

print(colnames(cleaned_data))
```

*Figure 2:Code of BroadBand*

## Population

The population figures for the years 2020 through 2023 were projected using the available population data from 2011 as a base. A steady annual growth rate of approximately 0.561% was applied to estimate the population for each subsequent year. This growth rate was selected based on demographic trends and represents a realistic assumption for population growth over time. The method involved multiplying the 2011 population figures by a growth factor each year to generate new estimates for 2020, 2021, 2022, and 2023.

This approach enabled the simulation of population growth in a simple manner, assuming a consistent annual percentage increase in population for each area. Once the population estimates were calculated, they were stored as separate CSV files for each year, facilitating tracking of changes over time and comparison of different years as needed. The datasets are presented in a clear, tabular format. Each row in the table represents a specific geographical region, such as a district or county, with the columns displaying the estimated population for that region in the respective year. This uniform structure across all datasets allows for straightforward comparisons, whether examining how the population in a particular area has grown over the years or comparing different regions in the same year.

This organization of the data simplifies the analysis of population trends and comprehension of how these changes may impact other factors, such as housing demand, infrastructure needs, or public services. The consistent format across all the years ensures that the data can be easily merged, compared, or analyzed in various ways, supporting a wide range of analytical tasks. This approach offers a clear, manageable method to explore how population changes over time might influence broader trends and decisions, such as where to invest in

housing or improve local amenities.

```r
# Load necessary libraries
library(tidyverse)
library(dplyr)

# Read the 2011 population data from CSV
pop_2011 = read_csv("D:/Sem4/Data Science/Obtained Data/Population2011_1656567141570.csv", show_col_types = FALSE)

# View the original 2011 population data
View(pop_2011)

# Estimate the 2023 population by applying a growth factor to the 2011 population
pop_2023 = pop_2011 %>%
  mutate(Population = as.integer(Population * 1.00561255390388033))

# View the estimated 2023 population data
View(pop_2023)


# Assuming there's a dataset `PostcodeToLSOA` with 'ShortPostcode' and 'LSOA Code' columns
PostcodeToLSOA = read_csv("D:/Sem4/Data Science/Data-Science-Assignment/Cleaned Data/LSOA/Clean_Postcode_to_LSOA.csv")
View(PostcodeToLSOA)

population_with_lsoa <- pop_2023 %>%
  # Assume the population dataset also has a column named 'Postcode' to join with
  left_join(PostcodeToLSOA, by = c("Postcode" = "ShortPostcode")) %>%
  select(`LSOA Code`, Postcode, Population) %>%
  drop_na() %>%
  distinct()# Remove rows with NA values

view(population_with_lsoa)

colnames(population_with_lsoa) <- c("ID", "Short Postcode", "Population")

# View the merged dataset with LSOA codes
View(population_with_lsoa)

# Save Population2023 data to a new CSV file
write.csv(population_with_lsoa, "D:/Sem4/Data Science/Data-Science-Assignment/Cleaned Data/Population/Cleaned_Population_Data.c
```

*Figure 3: Code of Population*

**School:**

The datasets for the 2021-2022 academic year in Bristol and Cornwall are loaded for each region. The relevant columns for Bristol, such as the school's name, postcode, town, and key performance scores, were carefully chosen. The data was then filtered to remove any incomplete or incorrect entries, such as scores marked as "Not Entered" (NE) or "Suppressed" (SUPP), which could affect the analysis. Additional columns were added to label the data by academic year and county (Bristol). A similar process was applied to the Cornwall dataset to maintain consistency in data cleaning and structuring. After standardizing both datasets, they were combined into a single comprehensive dataset, including all schools from Bristol and Cornwall for easier comparison and analysis. Finally, the consolidated data was saved into a CSV file, making it ready for further analysis or reporting as needed.

```
bristol21=read_csv("D:/Sem4/Data Science/Obtained Data/school/bristol/2021-2022/801_ks4final.csv")
bristol22=read_csv("D:/Sem4/Data Science/Obtained Data/school/bristol/2022-2023/801_ks4final.csv")
cornwall21=read_csv("D:/Sem4/Data Science/Obtained Data/school/cornwall/2021-2022/908_ks4final.csv")
cornwall22=read_csv("D:/Sem4/Data Science/Obtained Data/school/cornwall/2022-2023/908_ks4final.csv")


# cleaning
clean_data = function(df) {
  df %>%
    select(SCHNAME, PCODE, ATT8SCR, TOWN) %>%
    mutate(ATT8SCR = as.numeric(ATT8SCR)) %>%
    filter(!is.na(ATT8SCR)) %>%
    filter(!is.na(ATT8SCR) & !is.na(PCODE) & !is.na(SCHNAME) & !is.na(TOWN))  %>%
    distinct()


}

# Cleaning all datasets
bristol21_clean = clean_data(bristol21)
bristol22_clean = clean_data(bristol22)
cornwall21_clean = clean_data(cornwall21)
cornwall22_clean = clean_data(cornwall22)

# Adding academic year and county identifiers
bristol21_clean = bristol21_clean %>% mutate(Academic_Year = "2021-2022", County = "Bristol")
bristol22_clean = bristol22_clean %>% mutate(Academic_Year = "2022-2023", County = "Bristol")
cornwall21_clean = cornwall21_clean %>% mutate(Academic_Year = "2021-2022", County = "Cornwall")
cornwall22_clean = cornwall22_clean %>% mutate(Academic_Year = "2022-2023", County = "Cornwall")

# Merging
combined_schooldata = bind_rows(bristol21_clean, bristol22_clean, cornwall21_clean, cornwall22_clean)

View( combined_schooldata)

write_csv(combined_schooldata, "D:/Sem4/Data Science/Data-Science-Assignment/Cleaned Data/School/cleanedschool.csv")
```

*Figure 4: Code of School*

## LSOA:

The process involved integrating two sets of data: one with information about house prices and the other matching postcodes to LSOA (Lower Super Output Area) codes. First, we cleaned the postcode to LSOA dataset by selecting only the necessary columns and renaming them for clarity. Then, this cleaned dataset was merged with the house price dataset using the postcode as a common link. The final step involved filtering out any rows with missing values or duplicates, resulting in a clean, consolidated dataset that connects house prices with their corresponding LSOA codes. This cleaned dataset was then saved for further analysis.

```
library(tidyverse)

# Load the data
postcode_to_lsoa <- read.csv("D:/Sem4/Data Science/Obtained Data/Postcode to LSOA.csv")
view(postcode_to_lsoa)


# Import the cleaned house price dataset
cleaned_house_prices = read_csv("D:/Sem4/Data Science/Data-Science-Assignment/Cleaned Data/House Pricing/house_selling_clean.csv"
View(cleaned_house_prices)

# Clean and join the data using pipes
postcode_lsoa_clean = postcode_to_lsoa %>%
  # Select only the necessary columns for the mapping
  select(pcds, lsoa11cd) %>%
  # Rename columns for consistency and clarity
  rename(Postcode = pcds, `LSOA Code` = lsoa11cd) %>%
  # Perform a right join with the cleaned house prices data on the Postcode
  right_join(cleaned_house_prices, by = "Postcode") %>%
  # Select only the relevant columns for the final dataset
  select(`LSOA Code`, Postcode, `ShortPostcode`, `Town/City`, County) %>%
  # Remove rows with missing values
  drop_na() %>%
  # Remove duplicate rows
  distinct()

# View the cleaned and joined dataset
View(postcode_lsoa_clean)

# Save the cleaned dataset to a new CSV file
write.csv(postcode_lsoa_clean, "D:/Sem4/Data Science/Data-Science-Assignment/Cleaned Data/LSOA/Clean_Postcode_to_LSOA.csv", row.n
```

*Figure 5: Code of LSOA*

## Crime:

The data files from different subfolders, each containing records of crime incidents, were filtered to remove any irrelevant or missing information. This involved eliminating empty crime types or absent geographical coordinates. Furthermore, a date column was added by converting the month information into a standard date format. Unnecessary columns were removed, and duplicates were eliminated to ensure the integrity of the data. Once the individual files were cleaned, they were combined into a single, comprehensive dataset. This dataset was then linked with geographical and population data using a file that maps postcodes to LSOA (Lower Layer Super Output Areas) codes. The LSOA information was filtered to focus on specific counties, such as Bristol and Cornwall. Finally, the cleaned crime data was merged with the LSOA data, along with updated population figures, allowing for an analysis of crime patterns in these regions, taking into account the population in 2023. The resulting dataset was saved for further analysis, providing a cleaned and comprehensive view of crime statistics linked with geographical and demographic data.

```
crime_data = "D:/Sem4/Data Science/Obtained Data/Crime Dataset"

# Listing all subfolders
folders = list.dirs(crime_data, full.names = TRUE, recursive = FALSE)

cleaned_data_list = list()

# Function to clean individual crime data files
clean_crime_data = function(file_path) {

  data = read_csv(file_path, col_types = cols(
    `Crime type` = col_character(),
    Month = col_character(),
    Longitude = col_double(),
    Latitude = col_double(),
    Context = col_character()
  ))

  # Cleaning the data
  data_clean = data %>%
    filter(!is.na(`Crime type`)) %>%

    mutate(
      Date = as.Date(paste0(Month, "-01"), format = "%Y-%m-%d")
    ) %>%
    drop_na(Longitude, Latitude)  %>%
    distinct()

  # Removing unnecessary columns
  data_clean = data_clean %>%
    arrange(Date) %>%
    select(-`Crime ID`, -`Reported by`, -Longitude, -Latitude, -Location, -`Last outcome category`, -Date, -Context)


  return(data_clean)
}
```

*Figure 6: Code of Crime*

# Exploratory Data Analysis

**Housing:**

The data is initially filtered to emphasize the relevant locations, followed by the creation of various visualizations to aid in understanding the trends in house prices. In the case of Bristol, a boxplot is generated to display the range of house prices in different towns within the city. This type of chart showcases the minimum, maximum, and average prices, as well as the variation in prices across different towns. By visualizing the distribution of house prices, the boxplot identifies towns with higher or lower average prices and highlights any outliers. Similarly, for Cornwall, another boxplot is produced to compare house prices in different towns within the county. This facilitates a visual comparison of house prices in various towns, enabling insights into which towns are more expensive or affordable. Subsequently, the average house prices for each town in both Bristol and Cornwall are calculated. These average prices are then presented using bar charts, which facilitate a clear and simple comparison of the average cost of purchasing a house in different towns. The bar charts utilize color-coding and a currency format for displaying prices, ensuring the information is comprehensible. Finally, a line chart is created to illustrate the change in average house prices from 2020 to 2023 for both Bristol and Cornwall. This line chart allows for the observation of trends in house prices over these years, indicating whether prices have been increasing or decreasing and illustrating how the two areas compare to each other over time. Overall, this comprehensive approach enables a detailed and visual analysis and comparison of house prices across different towns and over time, thereby enhancing understanding of the real estate market in Bristol and Cornwall.

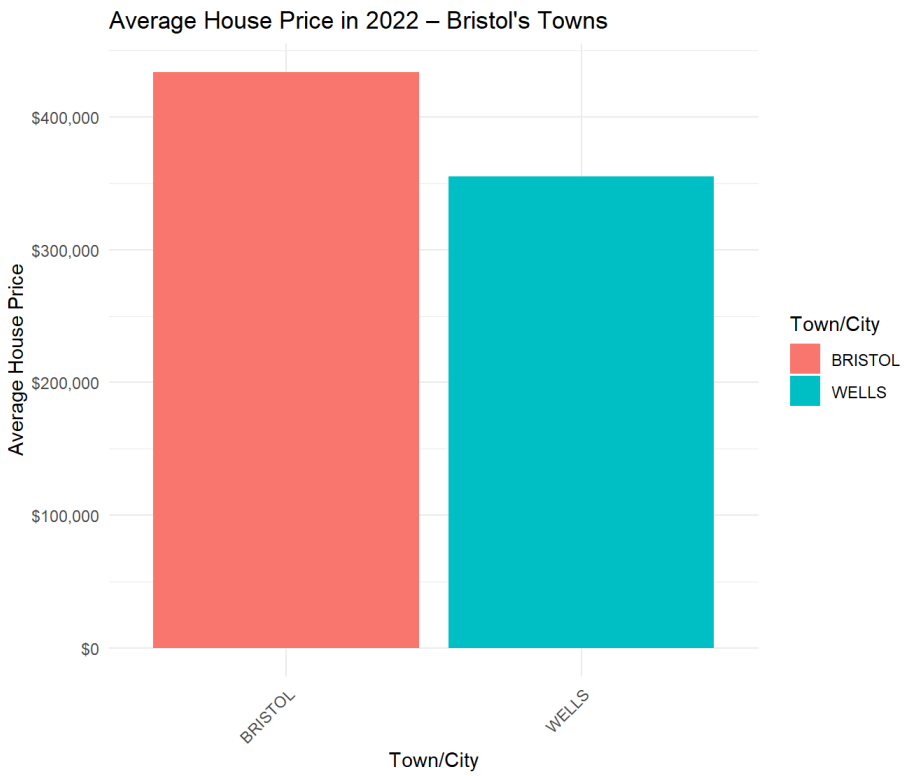



*Figure 7:Avg of Bristol*

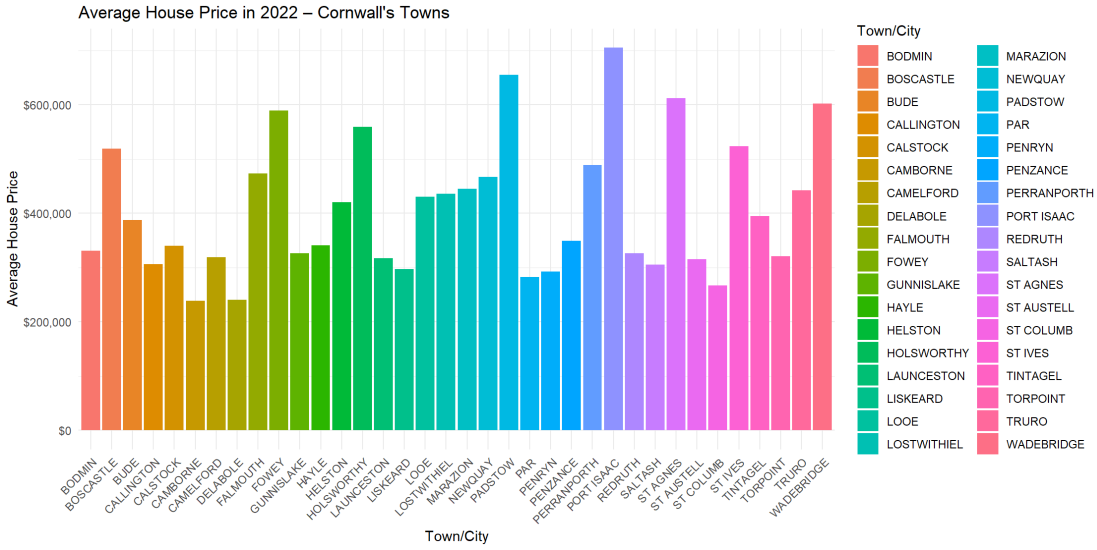



*Figure 8:Avg house price of cornwall*

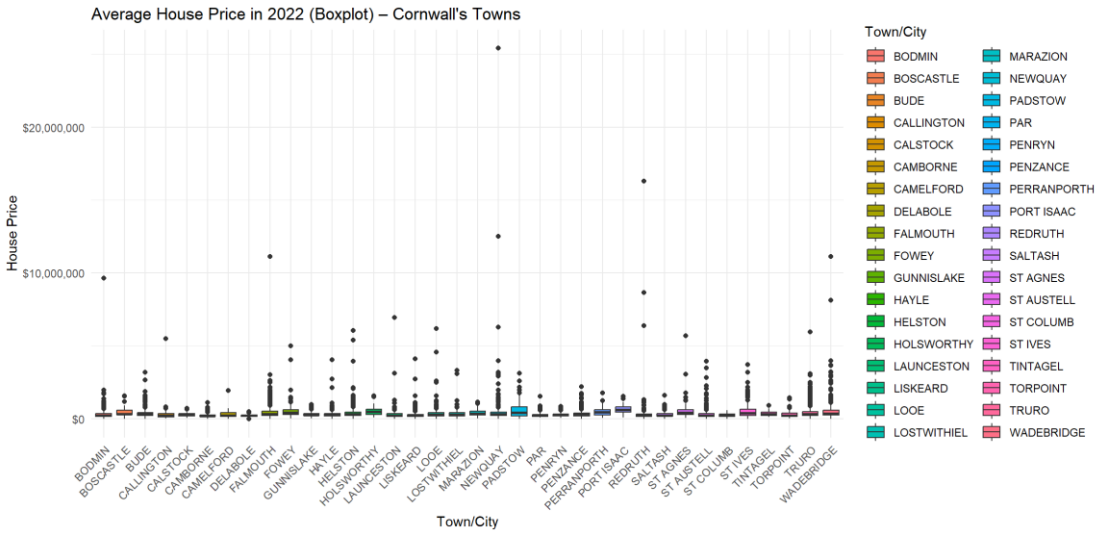



*Figure 9:Average house price in 2022(Bristol)*

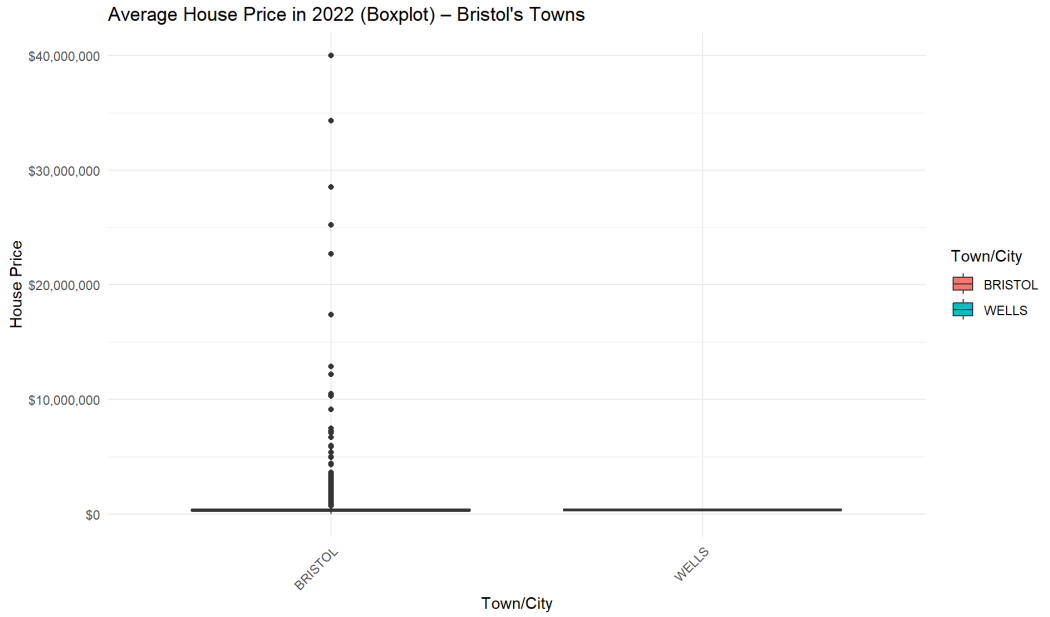



*Figure 10: Average house price from 2020 to 2023*

**Broadband:**

The data was initially cleaned to retain only the necessary information, including the speeds and data usage statistics for each area. This cleaned data was then saved for further analysis. To explore the broadband performance in these counties, a boxplot was created to visually compare the average download speeds between Bristol and Cornwall. The range and distribution of speeds within each county were highlighted by this plot, making it easier to see which areas have higher or lower speeds overall. Further analysis was done by creating bar charts that show both the average and maximum download speeds for each postcode within Bristol and Cornwall. These charts allow for a detailed comparison of broadband performance at a more granular level, illustrating which specific postcodes have better connectivity. This analysis provides a clear view of broadband speed distribution across Bristol and Cornwall, offering valuable insights into how different areas compare in terms of internet performance.

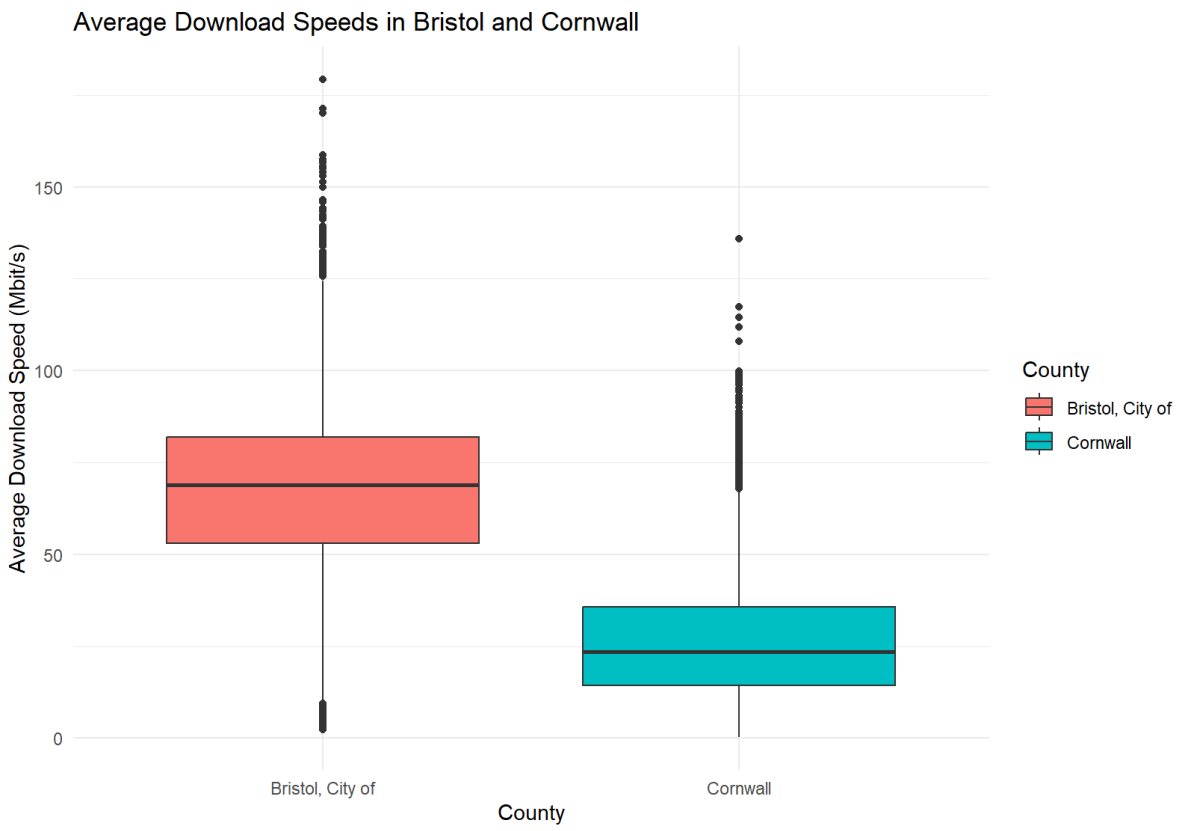
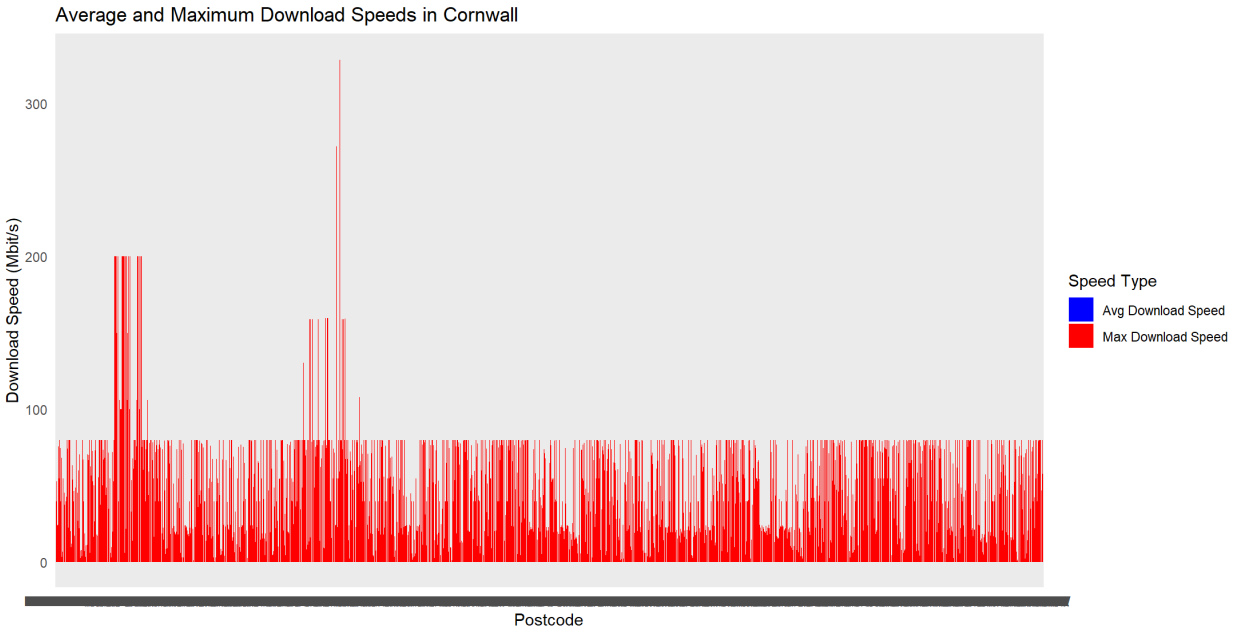



*Figure 11: Average speeds in Bristol and Cornwall*



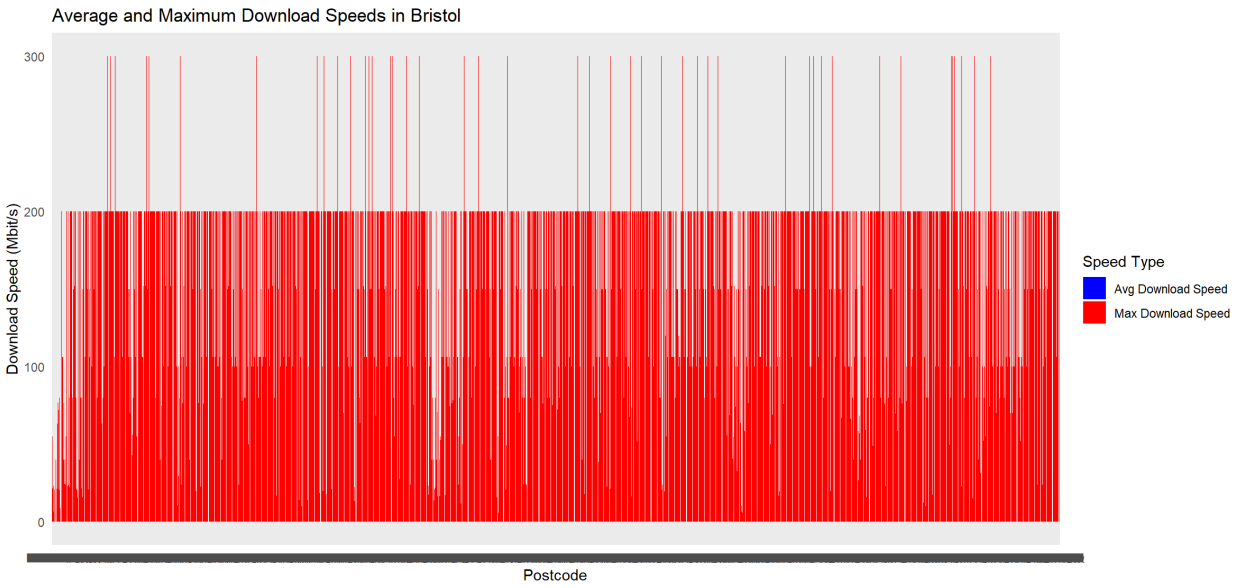



*Figure 12: Average and Maximum Speeds in Bristol*

## Crime:

A variety of visualizations were employed to analyze and compare crime rates in Bristol and Cornwall, with a focus on specific types of crime and time periods. For instance, drug offense rates in 2022 were compared using a boxplot, revealing differences in crime distribution between the two counties. Additionally, a radar chart was used to examine vehicle crime rates per 10,000 people from 2020 to 2023, highlighting trends over these years. A pie chart was utilized to illustrate the robbery crime rate in April 2023, showing the proportional distribution of these crimes between Bristol and Cornwall. Lastly, monthly drug offense rates in 2022 were tracked using a line chart, providing insights into fluctuations in drug-related crimes throughout the year. These visualizations collectively offer a comprehensive view of crime patterns across both counties, emphasizing variations in crime rates by location and time.

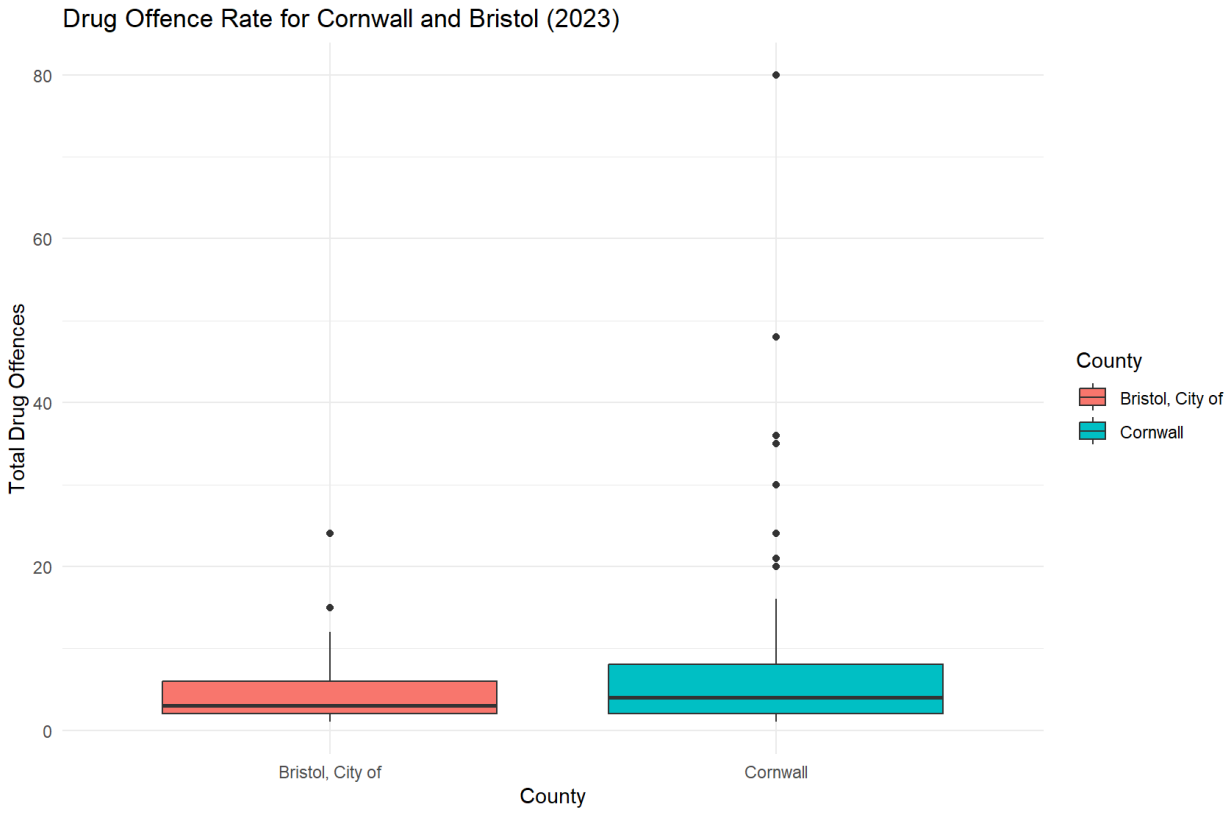



*Figure 13: Drug offence rate in both counties towns or districts in the year 2022*

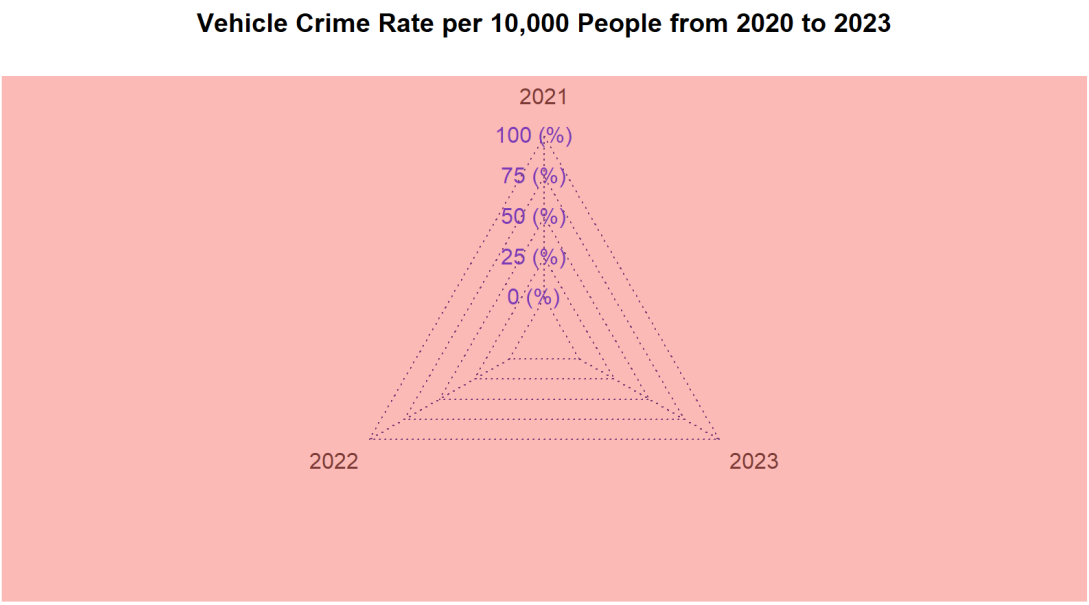



*Figure 14: Vehicle Crime Rate per 10000 people in the Specific month of your choice in year 2022*

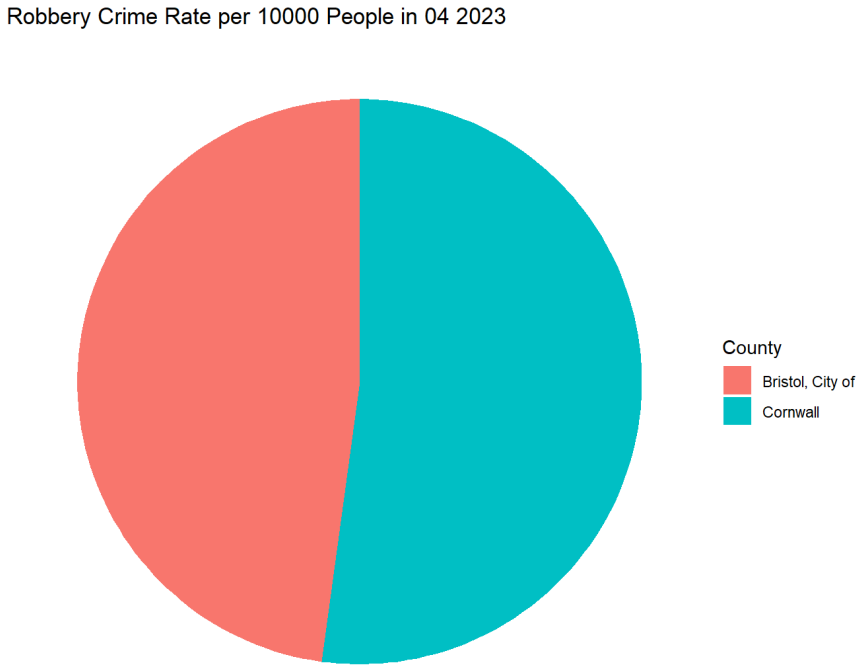



*Figure 15: Robbery crime rate per 10000 people in the specific month of your choice in year 2022*

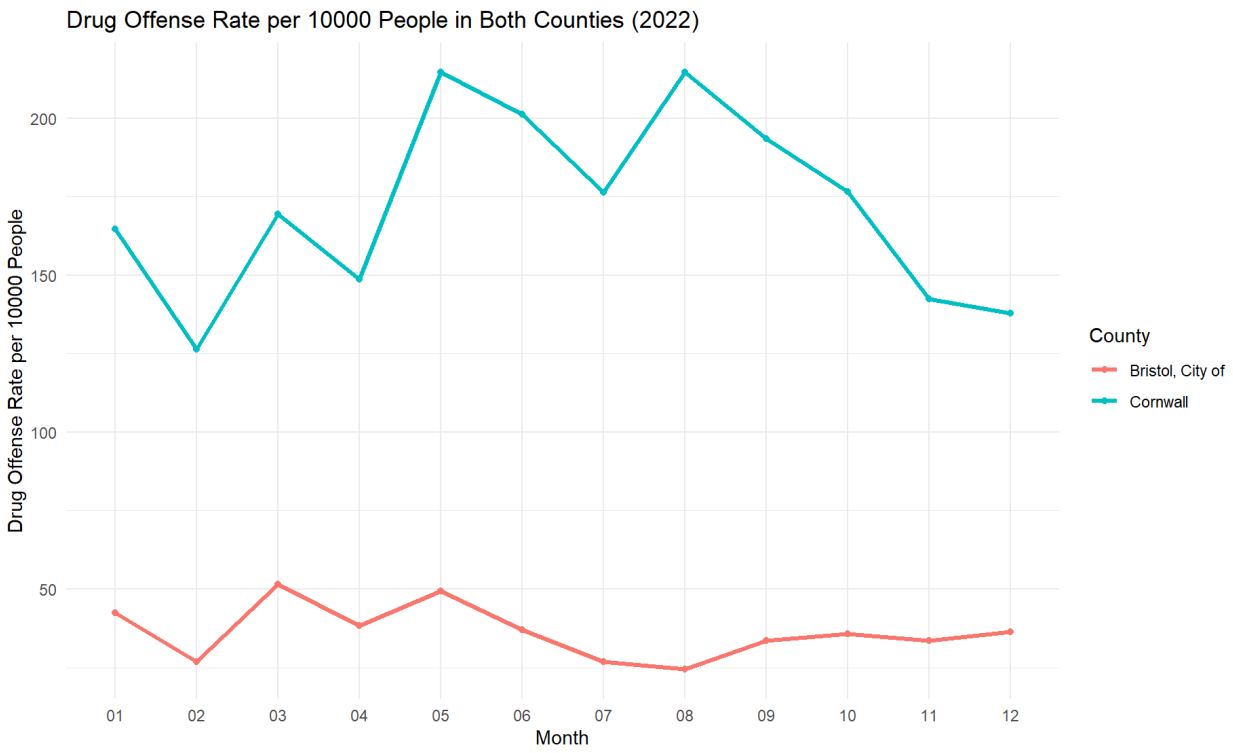



*Figure 16: Drug offense rate per 10000 people in both counties*

**School:**

The visualizations were created to analyze and compare how schools in Bristol and Cornwall performed in the 2021-2022 academic year based on their Attainment 8 scores. These scores represent the average scores achieved by students in eight subjects, including English and math, and are used to assess overall student achievement. Firstly, a boxplot was generated to show the distribution of Attainment 8 scores for schools in both Bristol and Cornwall. This plot helps to compare the range, median, and variation in scores between the two counties, providing insights into the overall performance and any potential outliers. The analysis then looked at individual schools within each county. For Bristol, a line chart was created to display the average Attainment 8 score for each school, illustrating how each school performed relative to others in the area. A similar line chart was also generated for schools in Cornwall. These charts help to identify trends in school performance within each county, making it easier to see which schools are excelling and which may need additional support. Overall, these visualizations provide a comprehensive view of student achievement in Bristol and Cornwall, highlighting differences and similarities in school performance across the two regions.

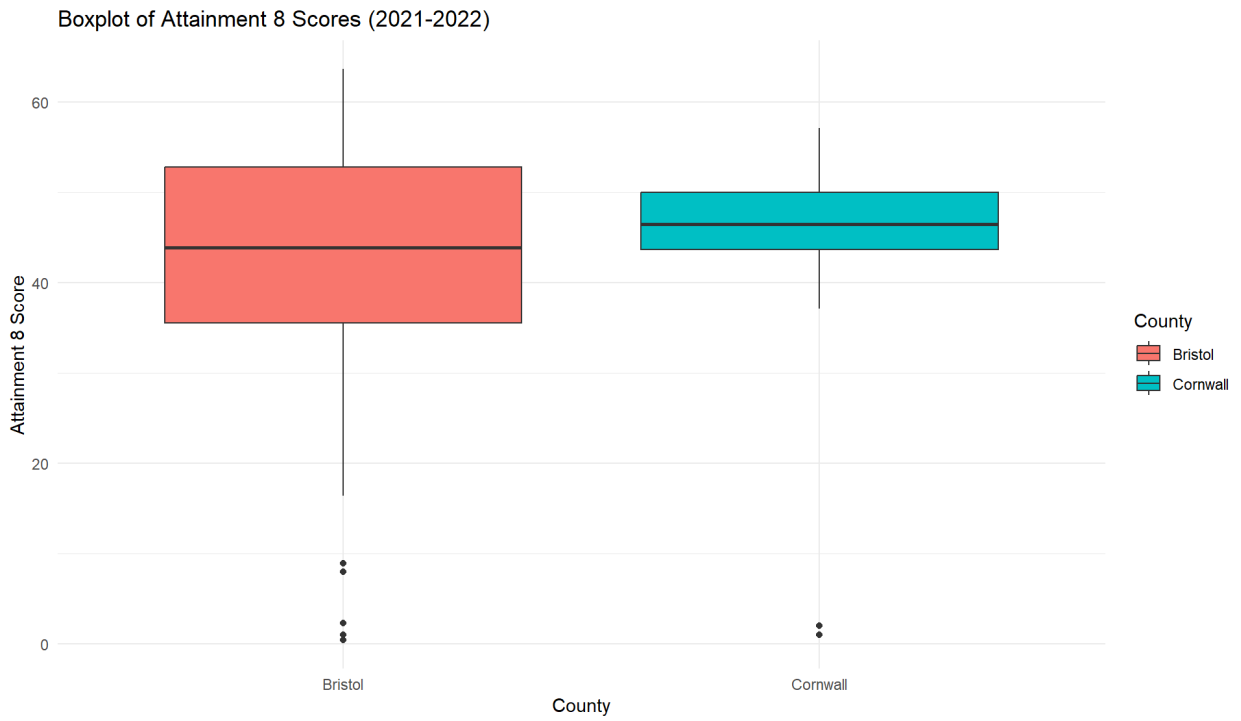



*Figure 17: Average Attainment 8 score in the year 2021-2022*

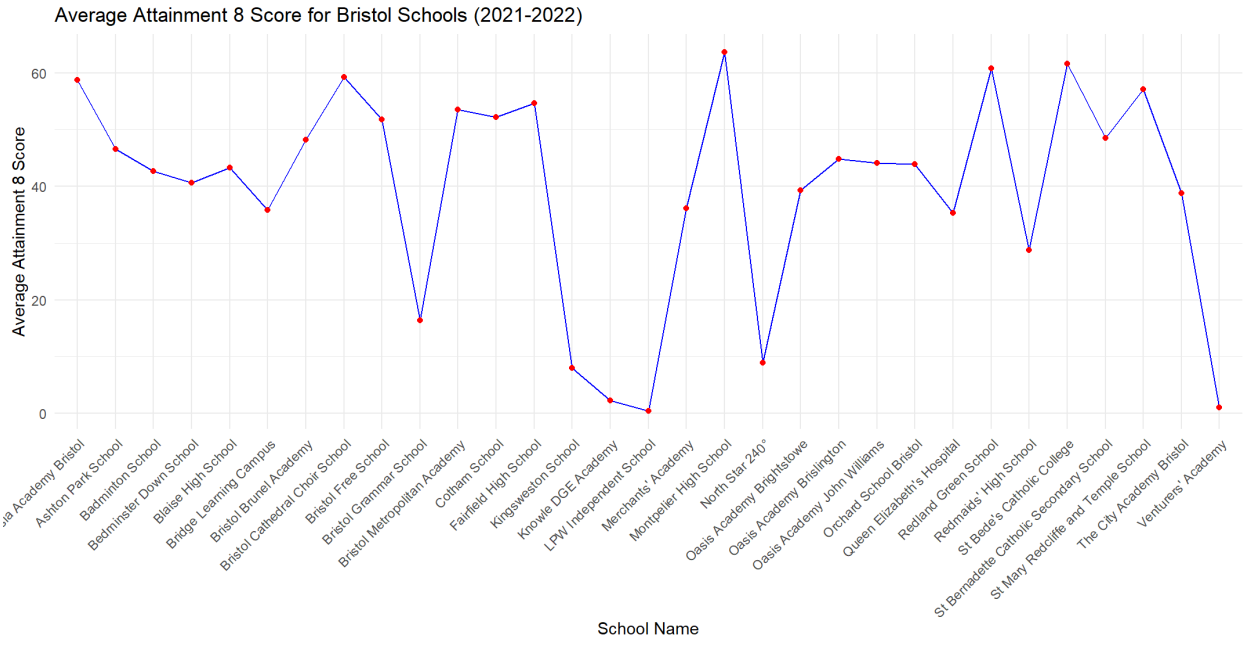



*Figure 18: Bristol average attainment 8 score in academic year 2021-2022*

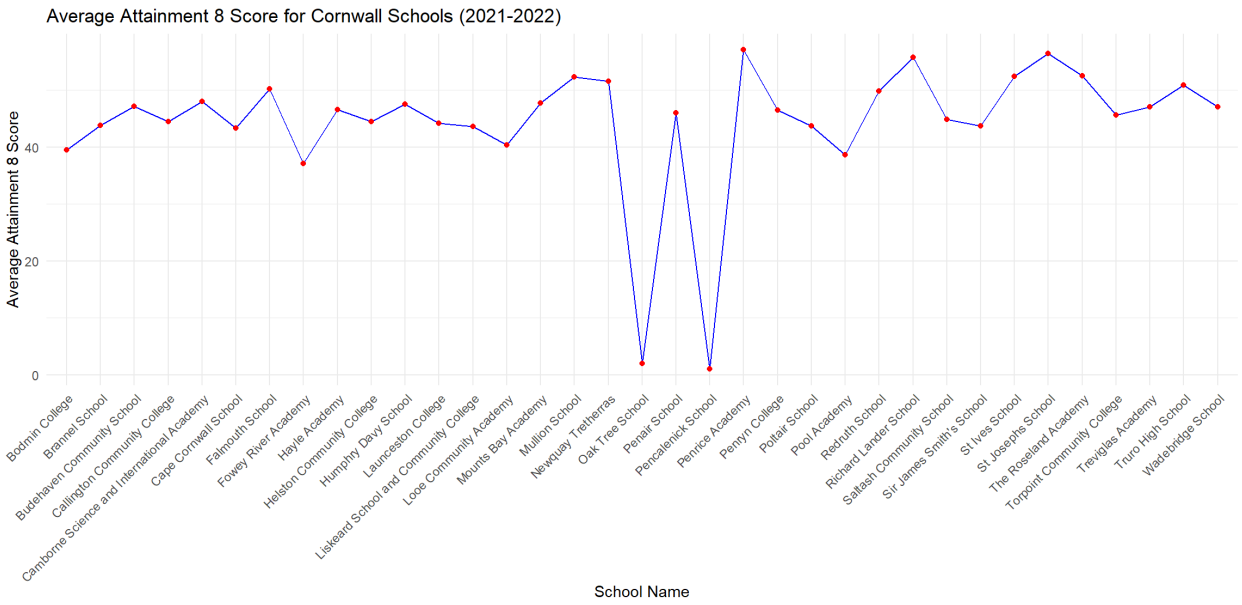



*Figure 19: Cornwall average attainment 8 score in academic year 2021-2022*

# Linear Modelling

## House prices vs Download Speed:

Datasets on house pricing and broadband speeds were merged based on common postcodes. A linear model was applied to examine the relationship, and a scatter plot was generated to visualize this connection. The analysis aimed to identify whether faster internet speeds are associated with higher house prices, providing insights for infrastructure investments and market dynamics.

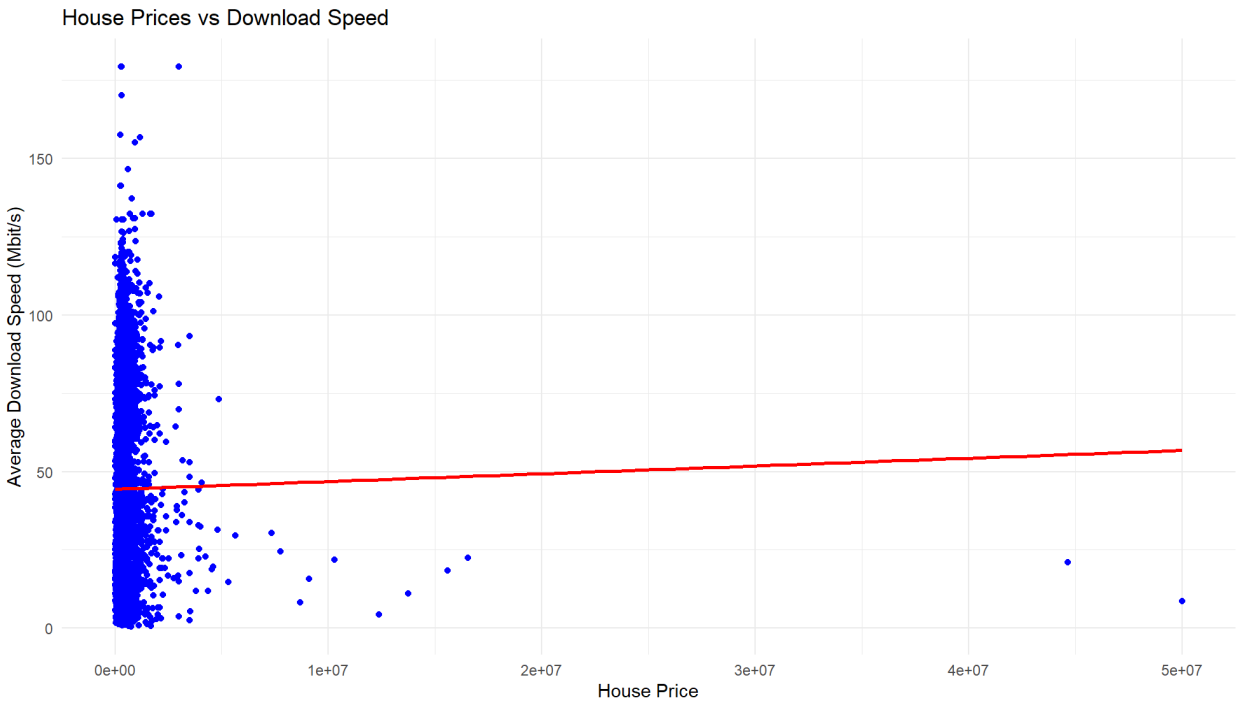



*Figure 20: House prices vs Download Speed*

**House price vs Drug Rate (2022):**

The study in 2023 looked into the correlation between house prices and drug-related crime rates. Initially, two sets of data were used, one focusing on house prices and the other on drug-related crimes. The data was refined to include only relevant information - house sales in 2023 and drug offenses reported in the same year. By merging the datasets, each area was represented by both house price and drug crime data. A linear model was then used to analyze the relationship between drug offenses and house prices. The findings were summarized and visually represented in a scatter plot, which aimed to illustrate any potential pattern indicating that higher drug crime rates are associated with lower house prices in 2023. This insight could be valuable for homeowners, buyers, and policymakers.

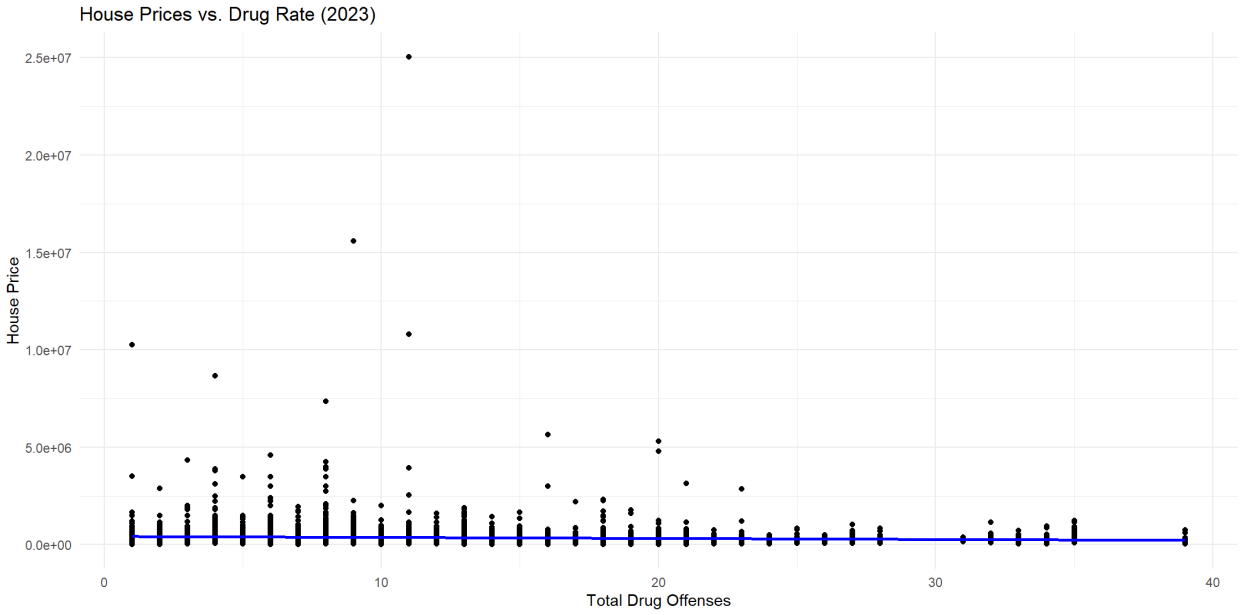



*Figure 21: House price vs Drug Rate (2022)*

**Attainment 8 score vs House Price (2022):**

The connection between house prices and school performance was examined by comparing average house prices and Attainment 8 scores in various towns and counties during 2022. The house price and school data were organized, merged into a single table, and analyzed to explore the relationship between house prices and academic performance. The results were summarized and visualized in a graph to determine if higher house prices are associated with better academic performance, providing insights for homeowners, educators, and policymakers.

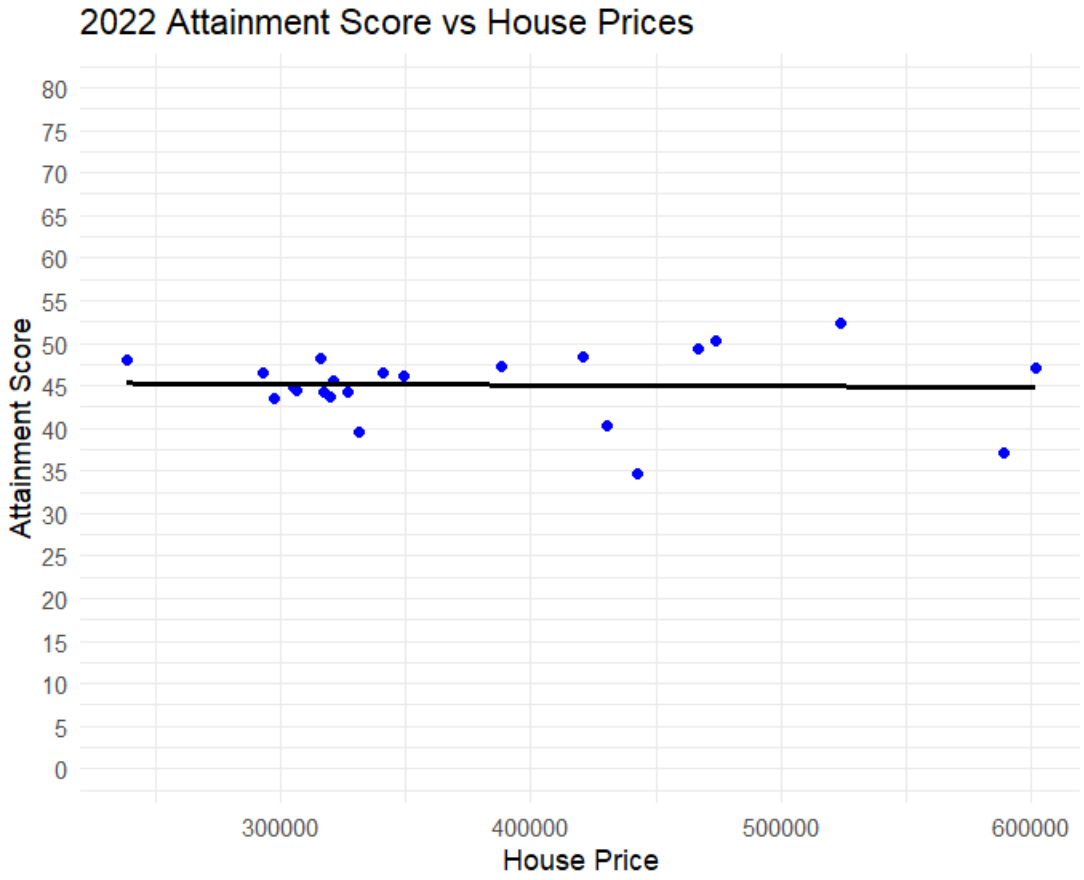



*Figure 22: Attainment 8 score vs House Price (2022)*

**Average Download Speed vs Drug Offense Rate (2022):**

The data was gathered from broadband speed records, crime statistics, and population information. First, the average download speed for each county was calculated. Then, the crime data was filtered to focus only on drug-related offenses, and the drug offense rate was calculated by dividing the number of drug crimes by the population in each area. The data was then combined into a single dataset, linking the average download speeds with the drug offense rates for each town or city. A linear model was created to see if there was any correlation between these two variables, meaning whether higher or lower download speeds were associated with higher or lower rates of drug offenses. The results were visualized in a graph, where each point represents a town or city. The graph includes a green line showing the general trend between drug offense rates and average download speeds. This analysis helps to understand if there is any significant connection between internet speed and crime rates in a given area.

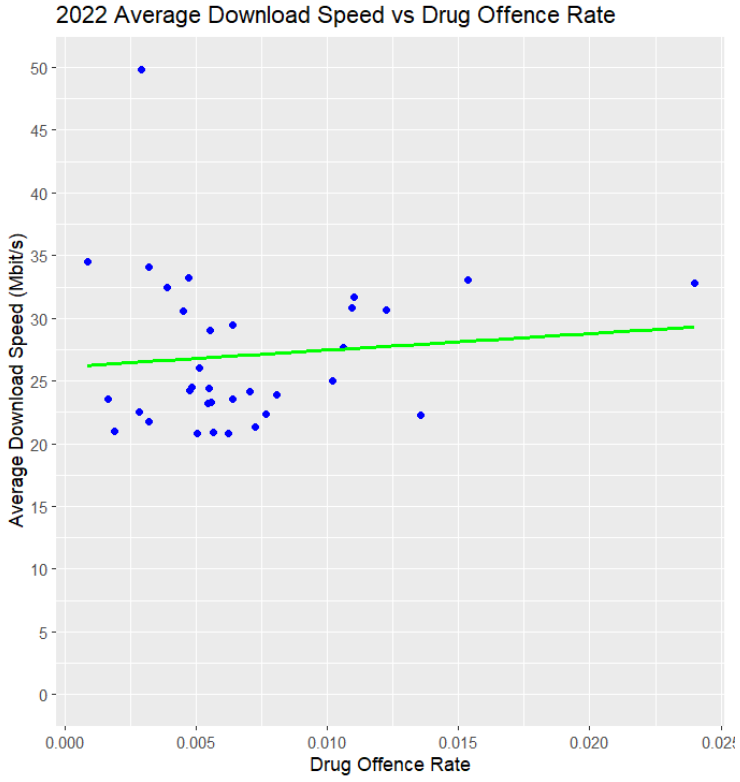



*Figure 23: Average Download Speed vs Drug Offense Rate (2022)*

**Attainment 8 score vs Drug Offense Rate (2022):**

Initially, the average Attainment 8 scores were computed for each town based on school data. Subsequently, the drug crime data was analyzed to calculate the rate of drug offenses per population for each town. These two datasets were then combined to assess whether there is a correlation between the rate of drug offenses in an area and the academic performance of students. A linear model was developed to predict the potential impact of changes in the drug offense rate on the Attainment 8 scores, and the relationship was visually depicted in a graph to demonstrate whether higher drug offense rates are associated with increased or decreased academic performance.

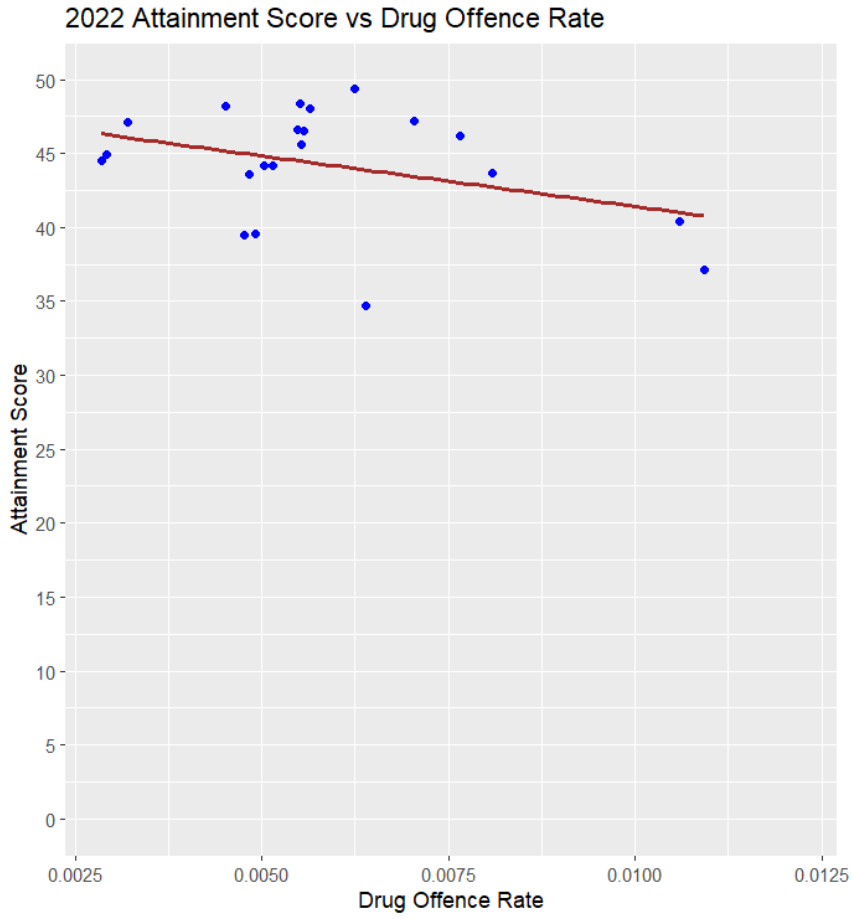



*Figure 24: Attainment 8 score vs Drug Offense Rate (2022)*

**Average Download Speed vs Attainment 8 score 2022:**

The data was organized by town and county to figure out the typical download speed and average Attainment 8 scores for each area. Then, these sets of data were combined to examine the potential connection between internet speed and student achievement. A model was made to guess the average download speed based on the Attainment 8 score, and the connection was displayed in a graph. This graph indicates whether better academic performance is linked with faster internet speeds in different towns.

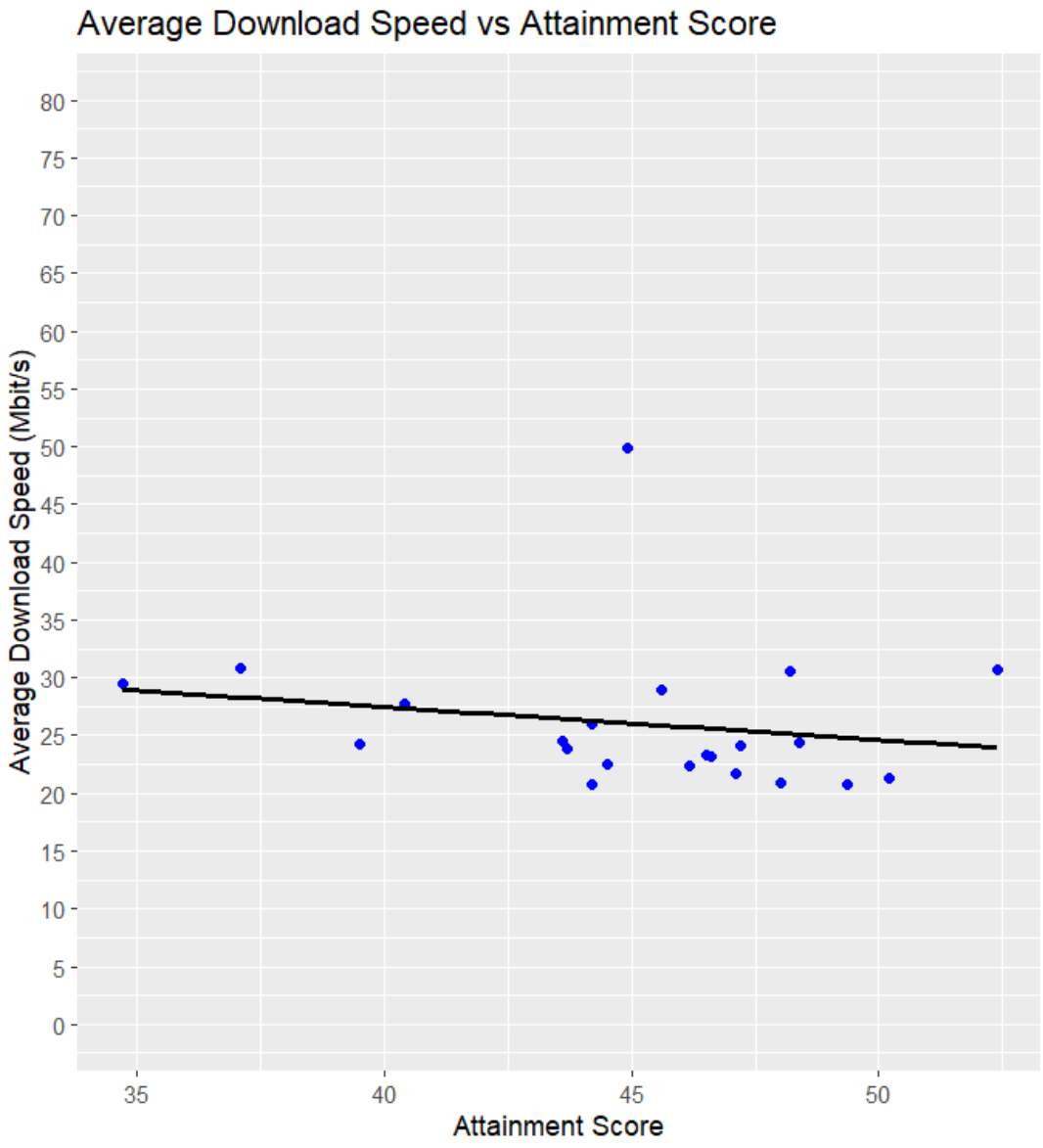



*Figure 25: Average Download Speed vs Attainment 8 score 2022*

## Legal and ethical issues

**Legal issues:**

In the analysis of property investment, it is necessary to address several legal issues to ensure compliance with regulations and protect the interests of all parties involved. Data privacy and protection are of utmost importance. Adherence to laws such as the UK's Data Protection Act 2018 and GDPR is required to safeguard personal data. This entails anonymizing or pseudonymizing data to prevent the identification of individuals and obtaining proper consent if personal data is used. Data accuracy and reliability are also crucial, thus using accurate and trustworthy data from authorized sources is necessary to avoid legal repercussions associated with misinformation or data manipulation. Additionally, it is essential to respect intellectual property rights by ensuring that all datasets, software, and proprietary information are used in accordance with their licenses and providing appropriate attribution. Compliance with local regulations governing real estate analysis and investment practices is also essential. Failure to adhere to these legal requirements can lead to significant legal consequences and undermine the credibility of the analysis.

**Ethical issues:**

The integrity and trustworthiness of property investment analysis are influenced not only by legal considerations but also by ethical issues. Managing bias and ensuring fairness is important to guarantee impartiality and comprehensive representation of all relevant factors. It is necessary to implement measures to identify and address biases to uphold ethical standards. Transparency and disclosure are vital as well; clearly communicating methodologies, data sources, and limitations helps prevent misleading information and supports informed decision-making. It is crucial to avoid conflicts of interest to maintain objectivity; analysts must ensure that their recommendations are solely based on data and not influenced by personal or financial interests. Lastly, considering the impact on communities is important, recommendations should support ethical and sustainable investment practices that positively affect local residents and economies. Addressing these ethical concerns fosters trust and ensures that the analysis is conducted with integrity and responsibility.

## Reflection

The analysis aimed to present a comprehensive overview of property investment opportunities in Bristol and Cornwall, considering various factors such as urban and rural characteristics, safety levels, and internet connectivity. Reliable UK government data was used and processed with R to ensure accuracy and reliability. The careful selection and handling of data were undertaken to provide evidence-based recommendations reflecting the true conditions of each area. The findings from this analysis were intended to assist potential investors in making informed decisions by considering factors such as affordability, safety, and quality of life. The report serves as a valuable tool for those navigating property investment in these areas and aims to contribute to a more informed and strategic approach to investment.

## Result

The final ranking of towns and cities is based on a comprehensive evaluation of house prices, internet download speeds, crime rates, and school performance. This provides valuable insights for potential property investors. The top-ranked town, Camborne in Cornwall, offers a favorable combination of lower house prices, good internet speeds, low crime rates, and high school attainment scores. Following closely are Saltash and St Austell, both also in Cornwall, which exhibit similarly advantageous characteristics. The top position of Camborne reflects its relatively affordable housing market, decent internet speed, low crime rates, and commendable school performance. Saltash and St Austell follow, providing slightly varied but competitive scores across the evaluated criteria. Penryn, Launceston, and Redruth also rank highly, indicating that these towns offer a strong mix of affordability, safety, and quality of life. The analysis shows that Cornwall generally offers more favorable conditions for property investment compared to Bristol. Bristol, although a significant city with a higher score in internet speed, ranks lower overall due to relatively higher house prices and other factors. This comprehensive ranking helps investors identify the best locations to consider, emphasizing the importance of balancing affordability, connectivity, safety, and educational opportunities in making informed property investment decisions.

## Conclusion

Throughout this report, a thorough comparison of educational performance between schools in Bristol and Cornwall in the 2021-2022 academic year was carried out. Various statistical methods and visual tools, such as box plots and line charts, were used to uncover key differences and trends in student achievement across the two regions. The analysis revealed variations in Attainment 8 scores, indicating potential influences from unique factors in each county, such as socioeconomic conditions, resource availability, or educational practices. The examination of individual schools highlighted areas where targeted interventions might be needed to bridge gaps in student performance. The visual representation of these data provides a clear way for stakeholders to identify critical areas of concern. This report is a valuable resource for educators, policymakers, and other stakeholders interested in improving educational outcomes in Bristol and Cornwall, offering insights that can guide decisions related to resource allocation, targeted educational programs, and strategies aimed at boosting student achievement. The findings underscore the importance of ongoing monitoring and evaluation to ensure educational quality and equity for all students. This work contributes to the broader goal of educational equity, providing data-driven evidence to support efforts that ensure every student receives the high-quality education they deserve.

## References


- Mali, K. (2024, August 11). *Linear Regression: A Comprehensive guide*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/
- UK Data Service. (2023, July 27). *Students - UK Data Service*. https://ukdataservice.ac.uk/learning-hub/students/
- Maki, G. (2020, October 27). *QIP creates accessible web design for Department of Education Contest*. https://www.qi-partners.com/2020-10-27-qip-creates-accessible-web-design-for-department-of-education-contest/
- Seandavi. (n.d.). *teaching/resources at main · seandavi/teaching*. GitHub. https://github.com/seandavi/teaching/tree/main/resources
- Statistics Solutions. (2024, April 17). *What is Linear Regression? - Statistics Solutions*. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/#:~:text=Linear%20regression%20stands%20as%20a,utilized%20form%20of%20predictive%20analysis.

## Appendix

- **GitHub Link:** NirajanMahato/Data-Science-Assignment (github.com)