Town Recommendation

Dilasha Ghimire

BSc. (Hons) Computing, Softwarica College of IT and E-commerce,

Coventry University

ST5014CEM Data Science for Developers

Mr. Siddhartha Neupane

August 19, 2024

Table of Content

Introduction

Identifying the top 10 best towns in Cornwall and Bristol required an in-depth analysis of various datasets sourced from multiple websites. The goal was to assess factors that contribute to the quality of life in these towns. The entire process was conducted using R, which served as the sole programming language for obtaining, cleaning, exploring, and modeling the data.

The analysis drew on information about house prices, broadband speeds, crime rates, and school quality in both counties. By thoroughly cleaning and sorting the data, we ensured that our findings were both accurate and reliable. Exploring the relationships between these factors allowed us to identify significant trends and patterns that impact the desirability of these towns.

The outcome of this project is expected to be a ranked list of the top 10 best towns in Cornwall and Bristol, based on the combined criteria. This list will provide valuable insights for residents, potential homebuyers, and policymakers, helping them make well-informed decisions about where to live or invest. By understanding these factors, individuals and families can better assess their options, and policymakers can prioritize improvements in areas that will most enhance the quality of life. This comprehensive analysis ultimately aims to benefit the community by highlighting the most desirable places to live.

Data Cleaning

The data for this project was carefully gathered from official UK government websites, ensuring that it was both reliable and authentic. Before diving into the cleaning process, I took the time to review the project requirements to determine which records were essential. As someone new to data manipulation and analysis, there were a few instances where I included records that turned out to be unnecessary. These initial missteps were part of the learning process and later I became more familiar with the data and its needs.

1. House Sales:

When it came to cleaning the house pricing data, I started by considering the requirements for exploratory data analysis (EDA) and linear modeling (LM). To streamline the process, I organized the data into three separate files: bristol_cleaned, cornwall_cleaned, and a combined CSV file created with the bind_rows() function. The original dataset contained 16 columns, but after going through the cleaning process, I was able to narrow it down to 9 columns that included the crucial information needed for analysis.

During this phase, I discovered an error where some towns, such as London, Wells, and Weston-super-Mare, were mistakenly listed as part of the City of Bristol. I corrected this issue by applying a filter to exclude these incorrect entries:

```
all_data %>%
  filter(V14 == "CITY OF BRISTOL") %>%
  # Filter rows where county is Bristol
  select(ID = V1, Housing_Price = V2, Year = V3, Postcode = V4, Type = V5, Town_City = V12, District = V13, County = V14) %>%
  # Select and rename specific columns
  mutate(Year = as.Date(Year, format = "%Y-%m-%d"),
         Short_Postcode = sub(" .*", "", Postcode)) %>%
  # Format Year column to remove time and convert to Date, and create Short_Postcode column
  distinct() %>%
  # Remove duplicate rows
  filter(!(County == "CITY OF BRISTOL" & Town_City %in% c("LONDON", "WELLS", "WESTON-SUPER-MARE"))) %>%
  # Filter out incorrect cities under CITY OF BRISTOL county
  mutate(Postcode = if_else(Postcode == "", "Not available", Postcode),
         Short_Postcode = if_else(Short_Postcode == "", "Not available", Short_Postcode)) %>%
  # Replace empty strings with "Not available" for Postcode and Short_Postcode only
  write_csv("C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Clean Data\\Cleaned Housing Data\\house_pricing_bristol.csv")
# Save the cleaned data
```

Figure 1: Code screenshot

Additionally, I created a separate dataset called towns_dataset that focused specifically on town-related records. This dataset was essential for the linear modeling and recommendation system used later in the analysis. By carefully managing and cleaning the data, I aimed to ensure accurate and insightful results for the project.

```
all_data %>%
  filter(V14 == "CORNWALL") %>%
  # Filter rows where county is Cornwall
  select(ID = V1, Housing_Price = V2, Year = V3, Postcode = V4, Type = V5, Town_City = V12, District = V13, County = V14) %>%
  # Select and rename specific columns
  mutate(Year = as.Date(Year, format = "%Y-%m-%d"),
         Short_Postcode = sub(" .*", "", Postcode)) %>%
  # Format Year column to remove time and convert to Date, and create Short_Postcode column
  distinct() %>%
  # Remove duplicate rows
  mutate(Postcode = if_else(Postcode == "", "Not available", Postcode),
         Short_Postcode = if_else(Short_Postcode == "", "Not available", Short_Postcode)) %>%
  # Replace empty strings with "Not available" for Postcode and Short_Postcode only
  write_csv("C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Clean Data\\Cleaned Housing Data\\house_pricing_cornwall.csv")
# Save the cleaned data
```

Figure 2: Code screenshot

2. Broadband Speeds:

The broadband data was also divided into three separate files to manage the information effectively. The datasets included performance, coverage, and postcode-to-LSOA. Initially, data from these sources were merged into three files for easier handling. From the performance dataset, I extracted average, maximum, and minimum upload and download speeds along with corresponding postcodes. For the coverage dataset, I initially included all premises and all matched premises data but later realized this was not necessary for the final analysis. The postcode-to-LSOA dataset provided valuable area and county information, which was joined with the performance dataset using a left_join() to integrate the broadband performance data with location-specific details.

```
# Filter datasets for Cornwall postcode areas and merge
cornwall_broadband_data = broadband_performance %>%
  # Filter rows where postcode area is either "TR" or "PL" (Cornwall areas)
  filter(postcode.area %in% c("TR", "PL")) %>%
  # Join the filtered broadband_performance dataset with the filtered broadband_coverage dataset
  inner_join(broadband_coverage %>% filter(pca %in% c("TR", "PL")),
             by = c("postcode", "postcode_space" = "pcds")) %>%
  # Select relevant columns for the final dataset
  select(postcode, postcode_space, postcode.area,
         Average.download.speed..Mbit.s., Maximum.download.speed..Mbit.s., Minimum.download.speed..Mbit.s.,
         Average.upload.speed..Mbit.s., Maximum.upload.speed..Mbit.s., Minimum.upload.speed..Mbit.s.,
         Average.data.usage..GB., All.Premises, All.Matched.Premises)

# Join the postcode_to_lsoa dataset with the cornwall_broadband_data
cornwall_broadband_data = cornwall_broadband_data %>%
  left_join(postcode_to_lsoa %>%
              select(pcds, lsoa_area = lsoa11nm, county = ladnm), # Select and rename columns
            by = c("postcode_space" = "pcds")) %>% # Join by matching columns
  filter(county == "Cornwall")
```

Figure 3: Code screenshot

3. Crime Rate:

The crime data presented a challenge due to its division into 36 separate CSV files, for each county. To streamline the cleaning process, a list and function approach was used. This method allowed for efficient processing and consolidation of the multiple files, simplifying the data handling and ensuring consistency across the different datasets.

```
# Define a list of file paths for the crime data CSV files for 2024
file_paths_cornwall_2024 = list(
  "dc-2024-01" = "C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Obtain Data\\Crime Rate Data\\devon-and-cornwall-street\\2024-01\
  "dc-2024-02" = "C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Obtain Data\\Crime Rate Data\\devon-and-cornwall-street\\2024-02\
  "dc-2024-03" = "C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Obtain Data\\Crime Rate Data\\devon-and-cornwall-street\\2024-03\
  "dc-2024-04" = "C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Obtain Data\\Crime Rate Data\\devon-and-cornwall-street\\2024-04\
  "dc-2024-05" = "C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Obtain Data\\Crime Rate Data\\devon-and-cornwall-street\\2024-05\
  "dc-2024-06" = "C:\\Users\\ghimi\\Desktop\\Town-Recommendation\\Obtain Data\\Crime Rate Data\\devon-and-cornwall-street\\2024-06\
)

# Function to read and clean each file for Cornwall data
read_and_clean_cornwall = function(file_path) {
  df = read_csv(file_path)  # Read the CSV file
  # Remove rows with missing 'Crime ID'
  df = df %>% filter(!is.na(`Crime ID`))
  return(df)
}

# Function to rename columns for Cornwall data
rename_columns_cornwall = function(df) {
  df = df %>%
    rename(
      Crime_ID = `Crime ID`,
      Reported_by = `Reported by`,
      Falls_Within = `Falls within`,
      LSOA_code = `LSOA code`,
      LSOA_name = `LSOA name`,
      Crime_type = `Crime type`,
      Last_outcome_category = `Last outcome category`
    )
  return(df)
}
```

Figure 4: Code screenshot

4. **Schools:**

Cleaning the school's data was particularly challenging. Initially, there was confusion about which datasets were relevant for the analysis. After some research and assistance, it became clear that the ks4final and school_information CSV files were the ones needed. These files contained all the necessary data for evaluating school performance and quality. Once identified, these datasets were used to extract and clean the relevant information, providing a solid foundation for the analysis of educational factors.

Cleaning the data is absolutely vital in any data analysis process. If the data isn't cleaned properly, everything that follows—whether it's exploratory data analysis (EDA), linear modeling (LM), or any other type of analysis—can end up being inaccurate. It's like building a house on a shaky foundation; if the data isn't reliable from the start, the results won't be either. Properly cleaning the data ensures that the analysis is based on accurate and relevant information, which is crucial for drawing meaningful and correct conclusions. In short, getting the cleaning right is key to ensuring that the whole analysis process produces valuable and trustworthy insights.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is like taking a first look at a new book; it helps you get a sense of the story before diving deeper. By examining and summarizing data, EDA reveals key patterns, trends, and relationships. It's essential for spotting issues like missing values or outliers early on, which can otherwise skew results. EDA also helps in forming initial hypotheses and choosing the right methods for further analysis. Essentially, it's about getting to know your data and ensuring you're building on a solid foundation, so your conclusions are accurate and meaningful.

1. **House Prices:**

- Average House Price in Year 2022 (Boxplot)

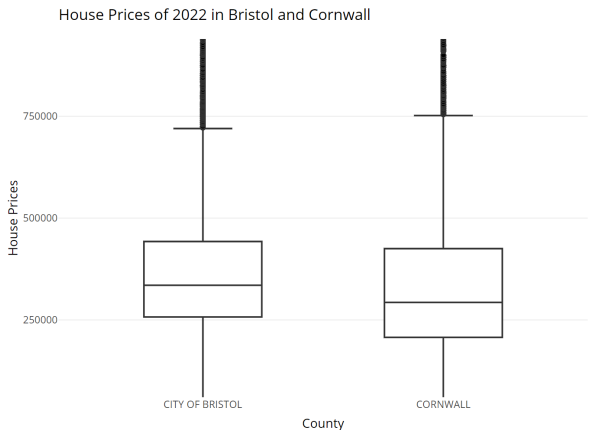

Figure 5: EDA

The boxplot for average house prices in 2022 shows that there isn't a significant difference between the two counties. However, the boxplot for Cornwall is slightly lower than that of Bristol. This indicates that house prices in Cornwall are generally more affordable compared to Bristol. The visual representation suggests that, on average, housing in Bristol is comparatively more expensive than in Cornwall.

- Average House Price in Year 2022 (Bar Chart)

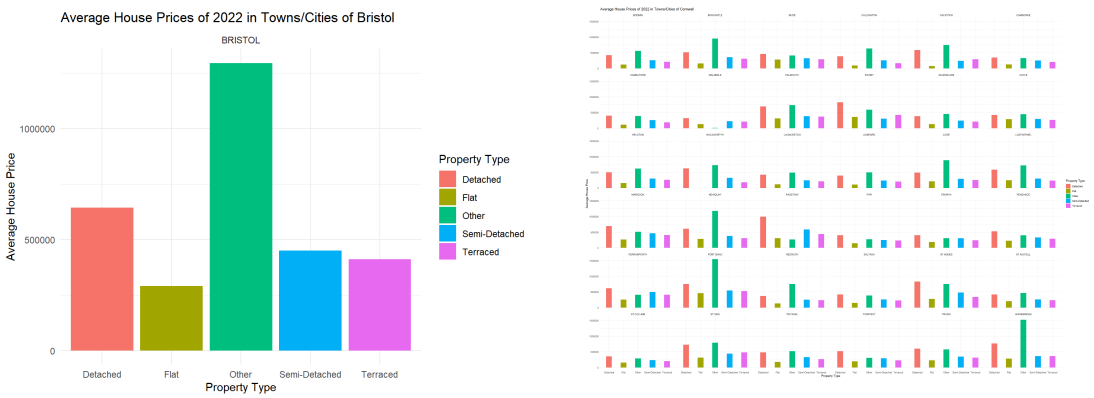

Figure 6: EDA

The bar chart depicting average house prices in 2022 for both counties confirms that Bristol, being both a city and a county, has higher average house prices compared to Cornwall. The chart also highlights that the dataset includes 36 towns and cities from Cornwall. Among these, St Ives and Wadebridge stand out with the highest average house prices, while Delabole has the lowest. This visualization reinforces the conclusion that housing in Bristol is generally more expensive, and provides insights into the varying house prices within Cornwall.

- Average House Price From (2020 - 2023) – (Line Chart)

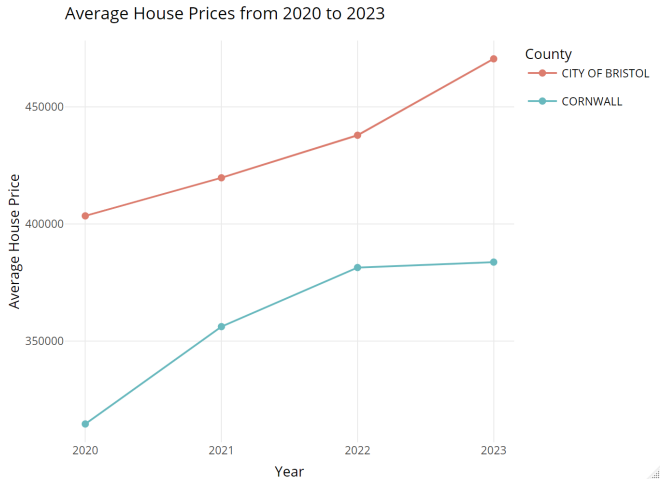

Figure 7: EDA

The line chart showing average house prices from 2020 to 2023 highlights that Bristol, as a single county, consistently has higher average house prices compared to Cornwall. Given that Cornwall encompasses 36 towns and cities, the overall average house price is naturally lower.

2. Broadband Speed:

- Average Download Speeds in Both Counties (Boxplot)

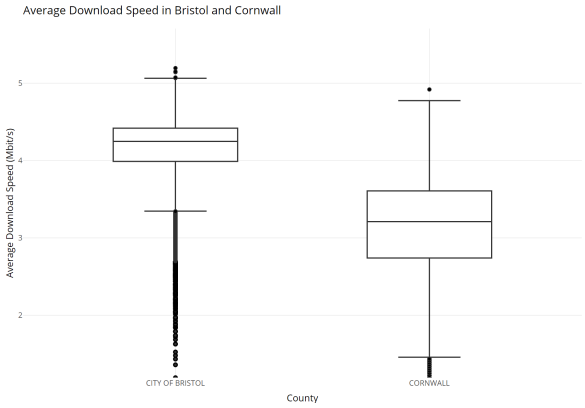

Figure 8: EDA

The boxplot of average download speeds in both counties shows that the City of Bristol has a higher average download speed compared to Cornwall. However, Cornwall's interquartile range is larger, indicating greater variability in download speeds. This means that while Bristol generally offers faster speeds on average, most of Cornwall's download speeds are clustered closer to the median, suggesting that a significant portion of Cornwall's data falls within a narrower range around the median.

- Average and Maximum Download Speeds in Both Counties (Bar Chart)

Figure 9: EDA

The bar chart comparing average and maximum download speeds in both counties reveals distinct patterns. In Bristol, both the average and maximum download speeds are predominantly in the upper half of the chart, with most average speeds exceeding 4 Mbit/s. This suggests that

Bristol generally offers higher and more consistent download speeds. In Cornwall, while there is noticeable fluctuation in the average download speeds across different areas, the maximum speeds are largely consistent, with only three LSOA areas deviating from the norm. This consistency in maximum speeds indicates a stable level of service throughout most parts of Cornwall, despite the variations in average speeds.

3. Crime Rate:

- Drug offense rate in both counties towns or districts in the year 2022 (Box- plot)

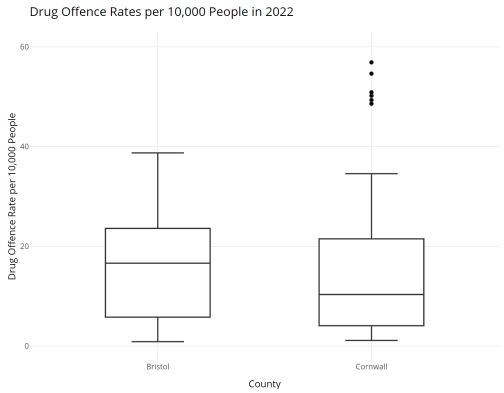

Figure 10: EDA

The box plot illustrating the drug offense rate in towns or districts of both counties for the year 2022 shows that the rates are quite similar overall. However, Bristol has a slightly higher median drug offense rate compared to Cornwall. Additionally, Cornwall exhibits a prominent outlier, indicating that while most areas in Cornwall have lower or moderate offense rates, there is at least one area with a significantly higher rate.

- Vehicle Crime Rate per 10000 people in the Specific month of your choice in year 2022 (Radar Chart)

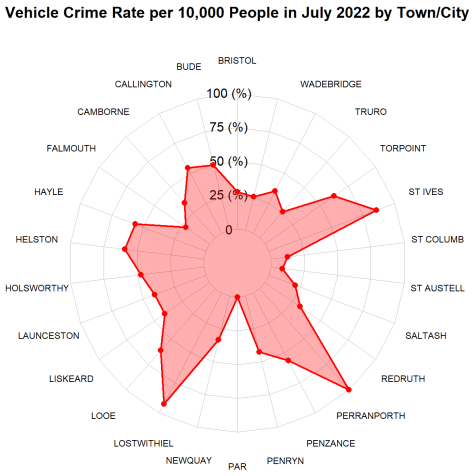

Figure 11: EDA

The radar chart displaying the vehicle crime rate per 10,000 people in July 2022 reveals notable variations across different towns. Perranporth, Lostwithiel, and St Ives stand out as the most unsafe towns in terms of vehicle crime during this month. These towns exhibit significantly higher crime rates compared to others, indicating a higher incidence of vehicle-related offenses.

- Robbery crime rate per 10000 people in the specific month of your choice in year 2022 (Pie Chart)

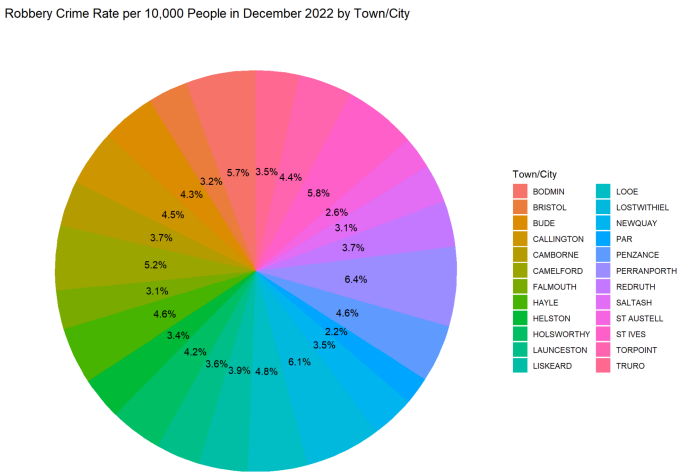

Figure 12: EDA

- Drug offense rate per 10000 people in both counties (Line Chart)

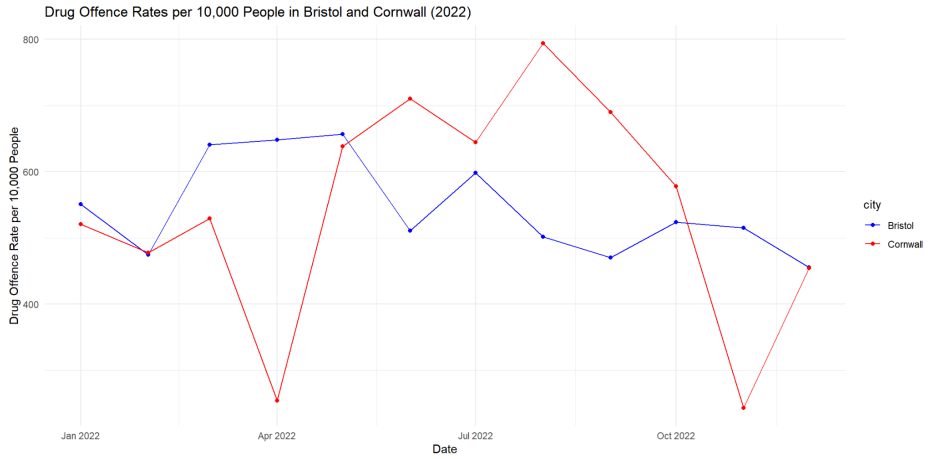

Figure 13: EDA

The line chart depicting the drug offense rate per 10,000 people in both counties throughout 2022 shows that both Bristol and Cornwall started at the same rate in January and ended at the same point in December. However, their trends in between were quite different. Cornwall experienced a lot of fluctuation, reaching its lowest point in April and peaking between July and October. In contrast, Bristol's drug offense rate remained comparatively steady throughout the year, showing fewer dramatic changes compared to Cornwall's more volatile pattern.

4. Schools:

- Average Attainment 8 score in the year 2021-2022 academic year for Both Counties (Boxplot)

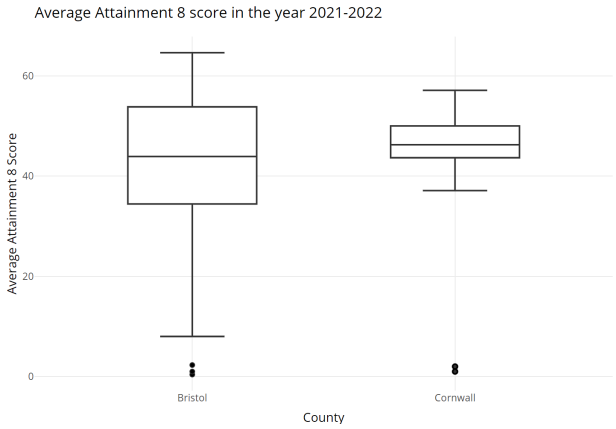

Figure 14: EDA

The boxplot of the average Attainment 8 scores for the 2021-2022 academic year in both counties reveals distinct patterns. Cornwall has a smaller interquartile range (IQR), indicating that the scores are more tightly clustered around the median. In contrast, Bristol exhibits a substantially larger IQR, showing a wider spread of scores. Both counties have outliers below the lower fence, indicating that some schools or areas within each county scored significantly lower than the majority.

- Bristol average attainment 8 score in academic year 2021-2022 (Line Chart)

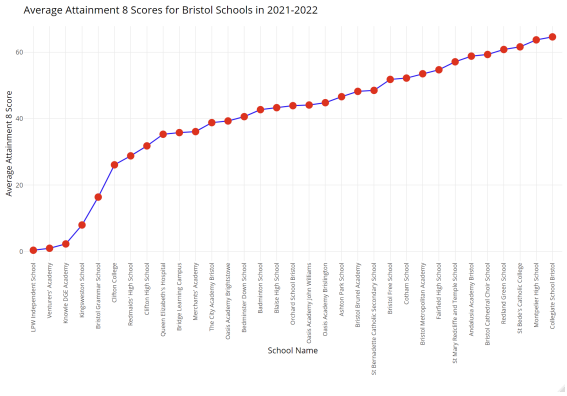

Figure 15: EDA

- Cornwall average attainment 8 score in academic year 2021-2022 (Line Chart)

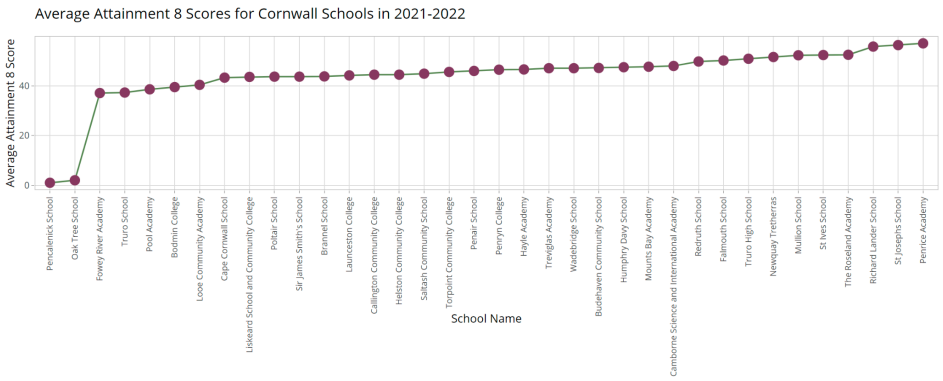

Figure 16: EDA

The line chart depicting Bristol's average Attainment 8 scores for the 2021-2022 academic year shows a generally linear trend. This indicates that the scores vary across schools, suggesting that performance is not uniform and some schools perform better or worse than others.

In Cornwall, the line chart reveals a different pattern. There are two schools with scores at the bottom, significantly lower than the rest. The majority of schools, however, cluster closely along the same line, indicating similar performance levels.

Linear Modelling

The linear modeling results reveal significant insights into the data. The positive correlation between Average Download Speed and Attainment 8 Score indicates that better internet connectivity is associated with improved educational performance, suggesting that schools with higher-quality internet access may offer better learning opportunities. In contrast, the strong negative correlation between Attainment 8 Score and House Price implies that higher property values are linked to lower educational attainment, possibly reflecting socio-economic factors affecting educational resources. Other negative correlations suggest that higher drug offense rates might be associated with lower average download speeds or poorer educational outcomes, highlighting broader socio-economic trends. Overall, these relationships underscore how various factors interact and influence each other within the dataset.

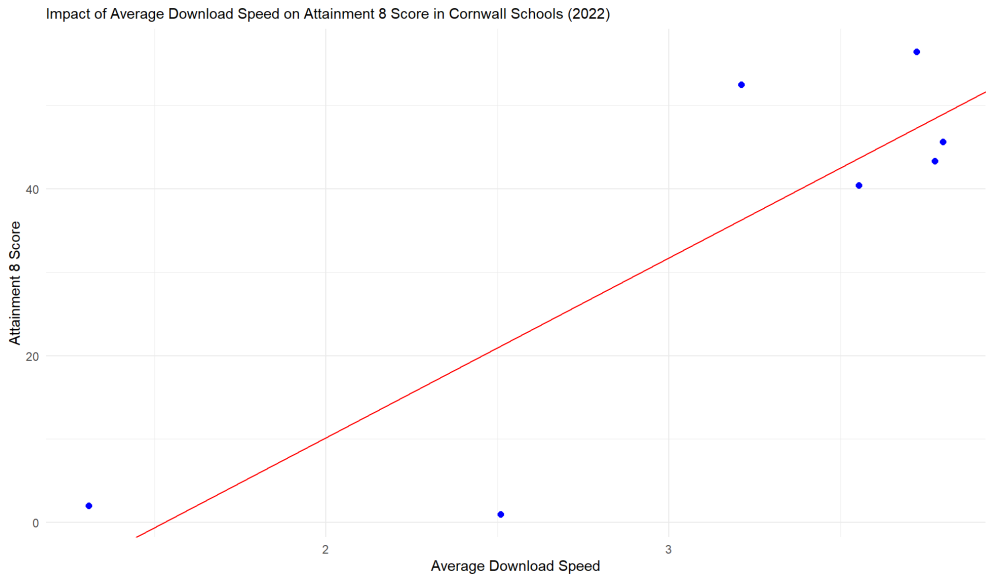

Figure 17: Download Speed vs Attainment 8 Score

```
> # Calculate correlation coefficient for Cornwall
> correlation_cornwall = cornwallAnalysis %>%
+   summarise(corCoeff = cor(ATT8SCR, `Average.download.speed..Mbit.s.`))
> print(correlation_cornwall)
# A tibble: 1 × 1
  corCoeff
     <dbl>
1    0.858
>
> # Linear modeling for Cornwall
> cornwallModel = lm(ATT8SCR ~ `Average.download.speed..Mbit.s.`, data = cornwallAnalysis)
> summary(cornwallModel)

Call:
lm(formula = ATT8SCR ~ Average.download.speed..Mbit.s., data = cornwallAnalysis)

Residuals:
      1       2       3       4       5       6       7
 -5.175  -3.276   6.717 -20.187  16.191   9.093  -3.364

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -32.931     18.664  -1.764   0.1379
Average.download.speed..Mbit.s.   21.564      5.761   3.743   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.01 on 5 degrees of freedom
Multiple R-squared:  0.737,     Adjusted R-squared:  0.6844
F-statistic: 14.01 on 1 and 5 DF,  p-value: 0.01339
```

Figure 18: Download Speed vs Attainment 8 Score

Recommendation System

The recommendation system developed for this project evaluates and ranks towns across Bristol and Cornwall based on a composite score derived from key factors: crime rates, school quality, and house prices. The system merges multiple datasets, including housing, broadband, crime, and educational performance, to create a comprehensive view of each town. The final score is calculated by normalizing and weighting these factors, with crime rates and house prices being inversely related to the score, while school quality is directly proportional. The resulting recommendations identify the top towns that offer a balance of low crime, high-quality schools, and affordable housing. In this case, towns like Bristol and various locations in Cornwall, including Helston, Liskeard, and Saltash, emerged as top recommendations with high scores. This approach ensures that the recommended towns provide optimal living conditions based on the combined criteria, helping individuals make informed decisions about where to live or invest.

```r
# Combine Bristol datasets
combined_bristol = reduce(list(house_data_bristol, broadband_bristol, crime_bristol, schools_bristol), full_join, by = c("Town", "Postcode"))

# Combine Cornwall datasets
combined_cornwall = reduce(list(house_data_cornwall, broadband_cornwall, crime_cornwall, schools_cornwall), full_join, by = c("Town", "Postcode"))

# Combine Bristol and Cornwall data
combined_data = bind_rows(combined_bristol, combined_cornwall)

# Function to calculate a final score based on crime rate, school quality, and house prices
calculate_final_score = function(data) {
  data %>%
    mutate(
      Housing_Price = ifelse(is.na(Housing_Price), median(data$Housing_Price, na.rm = TRUE), Housing_Price),
      crime_rate = ifelse(is.na(Crime_ID), 0, 1),
      school_quality = ifelse(is.na(OFSTEDRATING), 0, as.numeric(OFSTEDRATING)),
      crime_rate_score = 10 * (1 - rescale(crime_rate, to = c(0, 1))),
      school_quality_score = 10 * rescale(school_quality, to = c(0, 1)),
      house_price_score = 10 * (1 - rescale(Housing_Price, to = c(0, 1))),
      final_score = 0.4 * crime_rate_score + 0.3 * school_quality_score + 0.3 * house_price_score
    )
}
```

Figure 19: Town Recommendation Function Code

	ID	Postcode	Town	County	Final Score
1	{E073986C-14E7-2134-E053-6C04A8C0233B}	BS3 1AA	BRISTOL	CITY OF BRISTOL	8.500000
2	{DBA933FA-63B6-669D-E053-6B04A8C0AD56}	TR13 9EB	HELSTON	CORNWALL	8.500000
3	{CD5A9DCC-A270-310A-E053-6C04A8C00A1F}	PL14 4NH	LISKEARD	CORNWALL	8.500000
4	{D707E536-6700-0AD9-E053-6B04A8C067CC}	Not available	BUDE	CORNWALL	8.500000
5	{C18F412B-9C8B-81A6-E053-6B04A8C0AD18}	PL12 4QG	SALTASH	CORNWALL	8.499997
6	{0B853950-BCB7-69A5-E063-4704A8C07DAC}	TR11 5PA	FALMOUTH	CORNWALL	8.499997
7	{EC7AD09A-92DE-9200-E053-6C04A8C0E306}	PL26 7TQ	ST AUSTELL	CORNWALL	8.499993
8	{D93B27B1-C11F-3100-E053-6C04A8C08887}	TR19 6EL	PENZANCE	CORNWALL	8.499993
9	{1A0C5C63-EAD4-7CBE-E063-4804A8C06C96}	PL30 3HX	BODMIN	CORNWALL	8.499993
10	{B0A9D11B-FEC0-4C1F-E053-6C04A8C0D716}	TR16 6QZ	REDRUTH	CORNWALL	8.499988

Figure 20: Top 10 Recommended Towns

Conclusion

The recommendation system effectively demonstrates the application of data integration, normalization, and modeling techniques to derive actionable insights. By combining datasets on house prices, broadband speeds, crime rates, and school quality, and then employing linear modeling and normalization methods, the system provides a comprehensive assessment of living conditions in Bristol and Cornwall. The final scores, which aggregate these factors with weighted importance, facilitate a clear ranking of towns based on overall desirability. The top recommendations, including towns like Helston, Liskeard, and Saltash, highlight how data-driven approaches can uncover patterns and guide decision-making. This process underscores the importance of rigorous data cleaning, integration, and analysis in generating meaningful results that inform real-world decisions.

Appendix

1. **Github link**

   https://github.com/dilasha-ghimire/Town-Recommendation

2. **Drive link**

   https://drive.google.com/drive/folders/1gntT6LINO9KLwQMYmp1bWAs1hd1eQNNu?usp=drive_link