Report on Town Recommendation System

Aashista Karki

B.Sc. Computing Softwarica College of IT and E-commerce

ST5014CEM Data Science for Developers

Siddhartha Neupane

Jan 24 2024

Table of Contents

[Introduction 4](#_Toc157114522)

[Cleaning data 4](#_Toc157114523)

[House pricing 5](#_Toc157114524)

[Population Cleaning data 6](#_Toc157114525)

[Broadband Speed Cleaning data 6](#_Toc157114526)

[Crime rate Cleaning data 7](#_Toc157114527)

[School Cleaning data 8](#_Toc157114528)

[Exploratory data analysis 9](#_Toc157114529)

[-House price data representation 10](#_Toc157114530)

[-Broadband Speed data representation 12](#_Toc157114531)

[- Crime rate data representation 14](#_Toc157114532)

[-School data representation 18](#_Toc157114533)

[Linear modeling 21](#_Toc157114534)

[Recommendation System 28](#_Toc157114535)

[Reflection **Error! Bookmark not defined.**](#_Toc157114536)

[Legal and ethical issues 32](#_Toc157114537)

[Conclusion 33](#_Toc157114538)

[References 33](#_Toc157114539)

[Appendix 34](#_Toc157114540)

Table of Figures

[Figure 1- House price cleaning1 5](#_Toc157115426)

[Figure 2-House price cleaning 2 6](#_Toc157115427)

[Figure 3-Population cleaning data 6](#_Toc157115428)

[Figure 4-Broadband speed cleaning code 7](#_Toc157115429)

[Figure 5-Crime data cleaning code 8](#_Toc157115430)

[Figure 6- School data cleaning code 9](#_Toc157115431)

[Figure 7-Average house prices by county 10](#_Toc157115432)

[Figure 8-Average house prices barchart 11](#_Toc157115433)

[Figure 9-Average house prices line graph 12](#_Toc157115434)

[Figure 10-Average Download speed by COunty 13](#_Toc157115435)

[Figure 11-Average download speed within Kent 13](#_Toc157115436)

[Figure 12-Average Download Speed within surrey 14](#_Toc157115437)

[Figure 13- Drug offence rate by district 15](#_Toc157115438)

[Figure 14-Radar chart 16](#_Toc157115439)

[Figure 15-Robbery rate in march 2022 17](#_Toc157115440)

[Figure 16 Drug offence rate 18](#_Toc157115441)

[Figure 17-Average attainment score both county 19](#_Toc157115442)

[Figure 18-Average attainment score kent line graph 20](#_Toc157115443)

[Figure 19-Average attainment score surrey line graph 21](#_Toc157115444)

[Figure 20-House price vs Average download speed 23](#_Toc157115445)

[Figure 21-House price vs Drug offence rate 24](#_Toc157115446)

[Figure 22-Attainment score vs House prices 25](#_Toc157115447)

[Figure 23-Attainment Score vs Drug offence rate 26](#_Toc157115448)

[Figure 24-Average download speed vs Drug offence Rate 27](#_Toc157115449)

[Figure 25-Average Download speed vs Attainment Score 28](#_Toc157115450)

# **Introduction**

This Individual Coursework endeavors to develop a town recommender system catering to the needs of international students planning a study exchange program in England. The primary goal is to assist in making informed decisions about towns in the counties of Kent and Surrey. Recommendations are based on a detailed analysis of various factors such as educational institutions, cost of living (exemplified by house prices), broadband speed, and safety ratings (illustrated by local crime statistics).

The exclusive use of datasets released by the UK government ensures the reliability of the analysis. These datasets are sourced from reputable entities, primarily accessed through the <https://data.gov.uk/> website, aligning with ethical and legal considerations.

The steps involved in this project is first data gathering using official uk government data, secondly data cleaning using r language, thirdly to plot out various graphs as the need of project and lastly to recommend town based on the top scores.

The report unfolds with a comprehensive Exploratory Data Analysis (EDA) section, using graphical plots and summary statistics to comprehend the distribution of single-variable data, identify outliers, and investigate relationships between variables through plots and correlation coefficients.

The report concludes with a discussion of legal and ethical considerations associated with the data used and the recommendation system developed. Additionally, it reflects on the application of the data mining lifecycle to this problem, summarizes conclusions, and offers recommendations for future improvements or extensions.

# **Cleaning data**

Cleaning data is the process of detecting and rectifying errors in datasets, including addressing issues like missing values, duplicates, and inconsistencies. This practice is essential for bolstering the accuracy and comprehensiveness of data, thereby improving the trustworthiness of analyses and decision-making. Employing techniques like exploratory analysis aids in honing the quality of data. The importance of clean data lies in its ability to predict errors and the formation of deceptive conclusions, underscoring its pivotal role in generating meaningful insights.

The obtained data will be then cleaned to be used for plotting , summary, graphs etc. Various commands such as omit, distinct, mean etc will be used to clean the data to make it duplicate free and to remove null values and make the data clean. After the data is cleaned it is written in a new csv file using write.csv command.

## House pricing cleaning data

Upon acquiring the dataset, new names were assigned to the columns. Subsequently, the data for each individual year underwent filtering to eliminate any unnecessary variables. Finally, the data from all the years were merged into a unified tibble for comprehensive analysis.

Figure 1- House price cleaning1

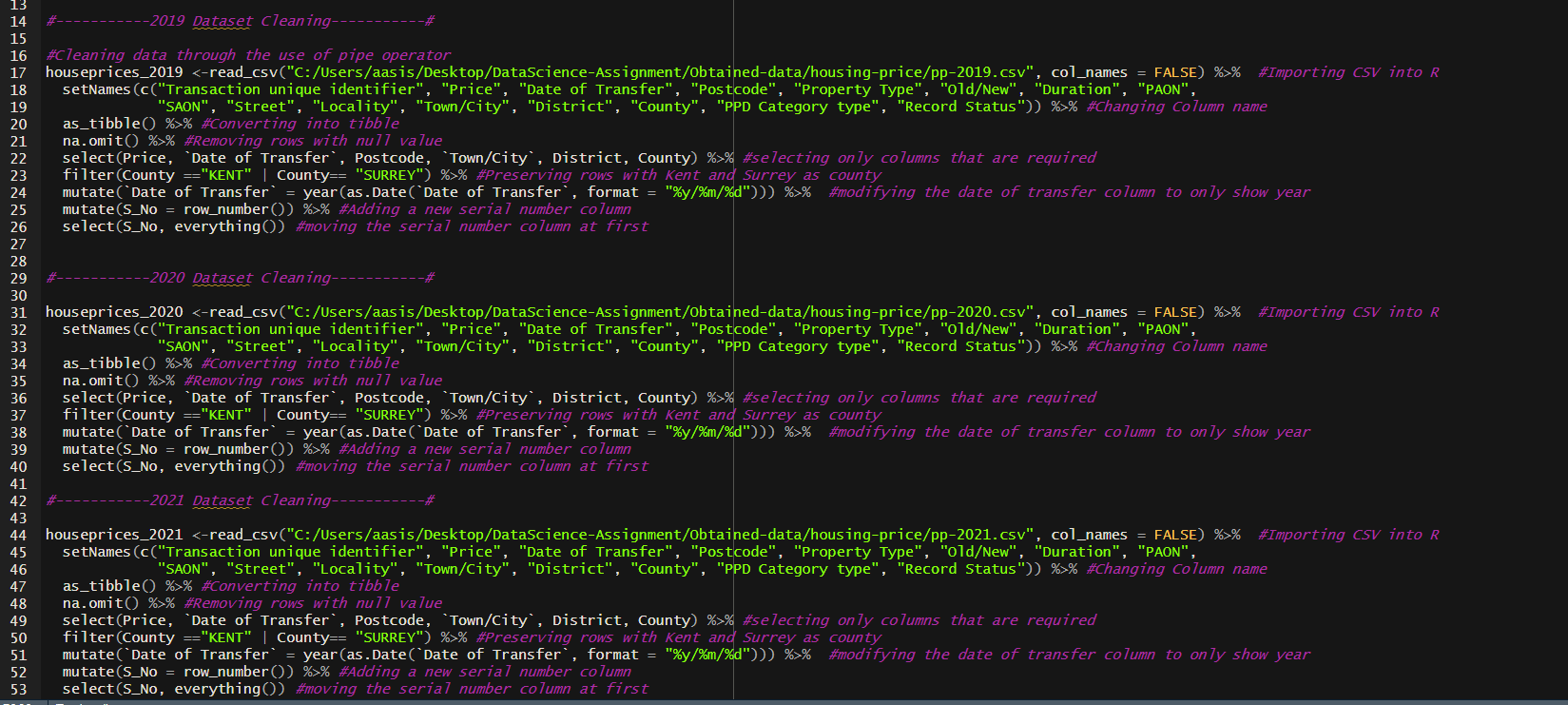


Figure 2-House price cleaning 2

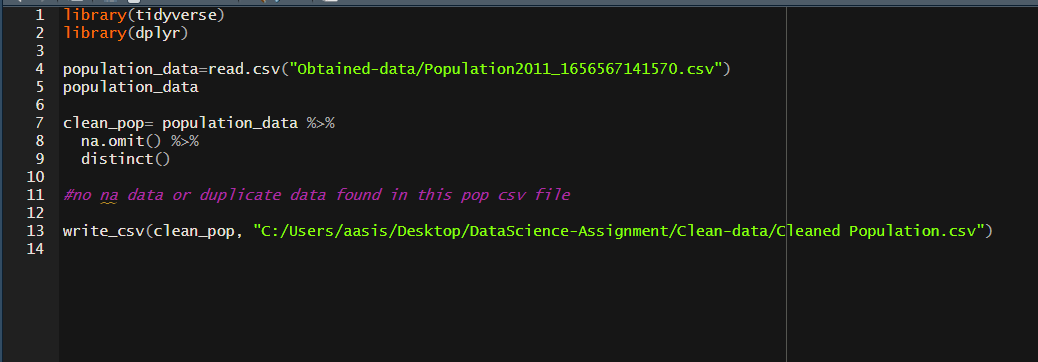
A screen shot of a computer

Description automatically generated

## Population Cleaning data

After obtaining the population data it was imported into r studio. The na values as well as duplicated values were to be removed using omit and distinct commands. But the population data was already clean so it was written into a csv file.

Figure 3-Population cleaning data



## Broadband Speed Cleaning data

This R code begins by loading essential libraries for data manipulation and setting the working directory. It then imports a cleaned dataset linking postcodes to LSOA codes. The code proceeds to read a broadband speed dataset, selecting relevant columns, renaming them, and joining the data with the cleaned postcode-to-LSOA dataset. Additional data manipulations include selecting specific columns, handling missing values, and adding a serial number column. The cleaned broadband speed dataset is finally saved as a CSV file. Overall, the code ensures data consistency, prepares it for analysis, and stores the cleaned dataset for further use.

Top of Form

Figure 4-Broadband speed cleaning code

A screen shot of a computer screen

Description automatically generated

## Crime rate Cleaning data

This combines crime data from Kent and Surrey, loads the datasets using the tidyverse and dplyr libraries, merges them into a single dataset (merged data), and displays the result. Subsequently, it slices values in the 'LSOA name' and 'Falls within' columns, removing excess characters. The 'Context' column is then deleted to remove any null values in this column. Finally, the cleaned dataset is saved as a CSV file named "Cleaned Crime Data.csv" at the specified path.

Figure 5-Crime data cleaning code

A screenshot of a computer program

Description automatically generated

## School Cleaning data

This R code processes school data from Kent and Surrey. It reads CSV files, each representing a specific region and academic year, loads the necessary libraries (tidyverse, dplyr), and manipulates the data. The code combines the datasets into a single tibble (**combine\_data**) using **rbind**. Then, it cleans the data by handling empty strings, removing rows with any missing values, filtering out rows with "NE" or "SUPP" in the 'ATT8SCR' column, converting 'ATT8SCR' to numeric, and selecting specific columns. Finally, the cleaned school data is saved as a CSV file named "Cleaned School Data.csv" in the specified directory.

Figure 6- School data cleaning code

A screenshot of a computer program

Description automatically generated

# Exploratory data analysis

Exploratory Data Analysis (EDA) is telling a story using pictures and graphs. It's about digging into data to find interesting insights without jumping to conclusions. EDA makes sure that the questions we ask about the data make sense and that the answers match what we already know.

In the beginning, we decide whether to use visuals or other methods to explore the data. Then, we figure out if we want to focus on one thing or look at multiple aspects together. EDA acts like a spotlight, helping us spot unusual patterns in the data and giving us guidance on how to analyze it effectively. We start by looking at individual parts, then explore relationships between two things, and finally, we check how different factors interact with each other (Patil, What is exploratory data analysis? 2022).

While sometimes we use tables with numbers like averages, most people prefer pictures for better understanding. Different types of visuals and tools help us track and understand the data. If one way isn't clear, we try another to get a fuller picture of what the data is telling us. EDA lets us uncover hidden insights in the data, making it an exciting journey of discovery.

## House price data representation

A box plot to visually compare the average house prices in Kent and Surrey for the year 2020. It starts by grouping cleaned house price data by town, district, county, and date of transfer, calculating the average price for each group. The code then filters this data to include only records from 2020. Using the ggplot library, it creates a box plot, with the x-axis representing counties (Kent and Surrey), the y-axis representing average prices, and the plot filled by county for clarity. The chart is titled "2020 Average House Prices By County Box Plot" and is customized to show prices up to 2,000,000 with breaks at intervals of 300,000 on the y-axis. The resulting visualization provides insights into the distribution of average house prices in the specified counties for the specified year.

Figure 7-Average house prices by county

A graph of a house price

Description automatically generated

A bar chart to visually compare the average house prices in Kent and Surrey for the year 2020. It filters the grouped house price data to include only records from 2020, then uses ggplot to plot a bar chart with counties on the x-axis, average prices on the y-axis, and bars filled by county. The chart is titled "2020 Average House Prices Barchart" and is customized to display prices up to 2,000,000 with breaks at intervals of 300,000 on the y-axis. The resulting visualization provides a clear comparison of average house prices between the two specified counties for the specified year.

Figure 8-Average house prices barchart

A graph showing a number of houses

Description automatically generated

This graph groups cleaned house prices by county and year, calculating the average price for each group. It then creates a line graph to visualize average house prices from 2019 to 2022. The data is filtered to include only records from these years, and the graph compares county prices over this period. The resulting line graph is titled "2019-2022 Average House Prices Line Graph," with the x-axis representing years, the y-axis representing average prices, and different colors indicating different counties. Points on the lines highlight specific data points, and the graph is customized with specific limits, breaks, and labels for better clarity.

Figure 9-Average house prices line graph

A graph of a line graph

Description automatically generated

## Broadband Speed data representation

Ggplot was used to plot various graphs of broadband speed of both kent in surrey. At first average download speed of both county was compared in a box plot where surrey had more speed. Following that average download speed within kent was put in a bar chart. Also pie chart of robbery crimes committed is also plotted.

Figure 10-Average Download speed by COunty

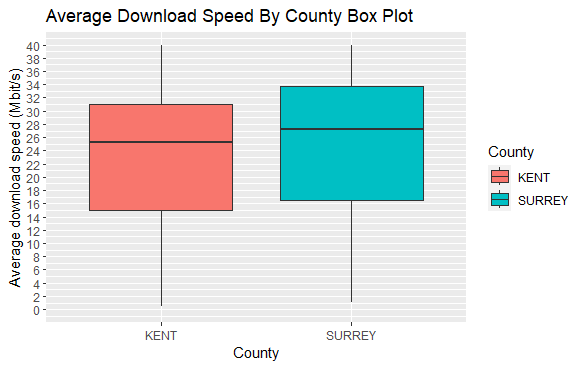


Figure 11-Average download speed within Kent

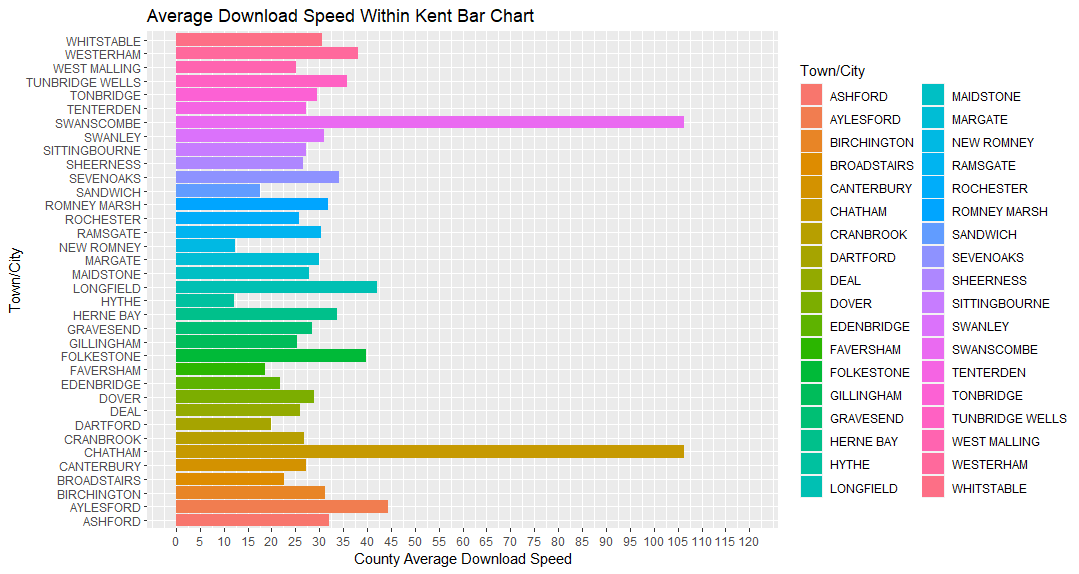


Figure 12-Average Download Speed within surrey

A graph of speed in various colors

Description automatically generated

## - Crime rate data representation

Ggplot and fmsb library was used to plot various graphs of crime rate of both Kent and surrey. At first Drug offence rate by district is plotted in a box plot. Following that a radar chart of vehicle offence rate per 10k is plotted. Also average download speed of surrey was visualized in a bar chart along with appropriate legend and colors. At last line chart of drug offence rate of both kent and surrey is plotted.

Figure 13- Drug offence rate by district

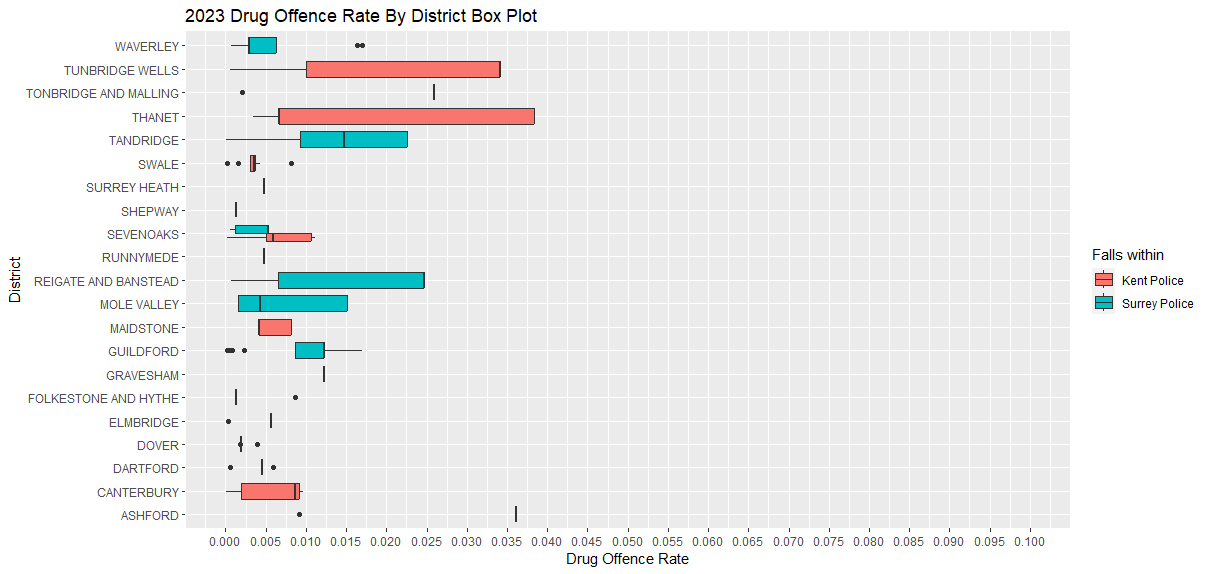


Figure 14-Radar chart

A diagram of a triangle

Description automatically generated

Figure 15-Robbery rate in march 2022

A purple circle with colorful stripes

Description automatically generated

Figure 16 Drug offence rate

A graph of a drug offense rate

Description automatically generated

## School data representation

In this section average attainment score is compared between both kent and surrey in a box plot. Then average attainment score line graph is drawn of kent county following that it is also drawn for surrey county.

Figure 17-Average attainment score both county

A graph of a graph with a red and blue box

Description automatically generated with medium confidence

Figure 18-Average attainment score kent line graph

A graph with different colored lines

Description automatically generated

Figure 19-Average attainment score surrey line graph

A graph with different colored lines

Description automatically generated

# Linear modeling

Leveraging the principles of linear regression, a statistical technique employed to identify connections between variables, streamlines the analysis of diverse datasets without repetitive efforts. In this approach, data points are linked by a straight line, and the objective is to determine regression coefficients that minimize errors, ensuring the best possible fit. Linear regression comes in two main forms: simple linear regression, dealing with a single variable, and multiple linear regression, which involves several independent variables and is more intricate (Mishra, Linear Modeling 2021).

The analysis encompasses a broad spectrum of information, including housing prices, crime rates (such as drug-related, robbery, and vehicle-related crimes), and significant educational metrics. By amalgamating these diverse datasets, a comprehensive comparison is facilitated, aiding in the identification of countries that have performed exceptionally well.

House price vs Average Download Speed

This begins by importing cleaned house prices and broadband speed datasets. It then groups the house prices and broadband speed data by town and county, calculating the average price and download speed for each group. The two datasets are joined into a single table. The code proceeds to create a linear model predicting house prices based on average download speed. The summary of the linear model is displayed, providing insights into the relationship between house prices and broadband speed. Additionally, a graphical representation of the linear model is generated using ggplot, with data points colored differently for Kent and Surrey. The resulting plot visualizes the 2020 house prices against average download speed, including a linear regression line.

Top of Form

Figure 20-House price vs Average download speed

A graph with red and blue dots

Description automatically generated

House price vs Drug rate

This R code begins by grouping cleaned house prices for the year 2021 by town and county, calculating the average price for each group. It then modifies the crime dataset to focus on drug-related offenses, creating a new dataset showing drug offense rates for each town and county in 2021. The two datasets are joined into a single table, and a linear model is created to predict house prices based on drug offense rates. The summary of the linear model is displayed, offering insights into the relationship between house prices and drug offenses. A graphical representation of the linear model is generated using ggplot, with data points colored differently for Kent and Surrey. The resulting plot visualizes 2021 house prices against drug offense rates, including a linear regression line. Finally, null values are removed from the joined dataset.

Top of Form

Figure 21-House price vs Drug offence rate

A graph with a green line

Description automatically generated

Attainment 8 Score vs House price

First group cleaned house prices for the year 2021 by town and county, calculating the average price for each group. It then groups school data by town and county for the same year, calculating the average attainment score for each group. The two datasets are joined into a single table based on the town, converting town names to lowercase for consistency. A linear model is created to predict average attainment scores based on average house prices. The summary of the linear model is displayed, providing insights into the relationship between attainment scores and house prices. A graphical representation of the linear model is generated using ggplot, with data points colored differently for Kent and Surrey. The resulting plot visualizes the 2021 attainment scores against house prices, including a linear regression line. Null values are removed from the joined dataset.

Figure 22-Attainment score vs House prices

A graph of a house price

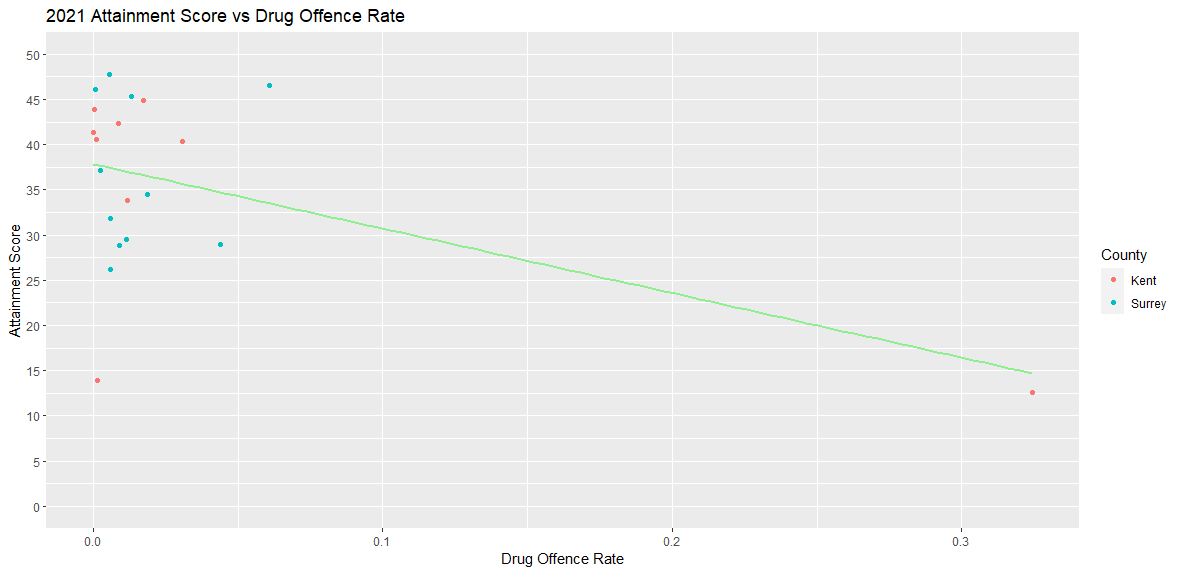
Description automatically generated

Attainment 8 score vs drug rate

This begins by grouping school data for the year 2021 by town and county, calculating the average attainment score for each group. It then modifies the crime dataset to focus on drug-related offenses, creating a new dataset showing drug offense rates for each town and county in 2021. The two datasets are joined into a single table based on the town, converting town names to lowercase for consistency. A linear model is created to predict average attainment scores based on drug offense rates. The summary of the linear model is displayed, providing insights into the relationship between attainment scores and drug offenses. A graphical representation of the linear model is generated using ggplot, with data points colored differently for Kent and Surrey. The resulting plot visualizes the 2021 attainment scores against drug offense rates, including a linear regression line.

Top of Form

Figure 23-Attainment Score vs Drug offence rate



Average Download Speed vs Drug rate

This begins by grouping broadband speed data by town and county, calculating the average download speed for each group. It then modifies the crime dataset to focus on drug-related offenses, creating a new dataset showing drug offense rates for each town and county in 2022. The two datasets are joined into a single table based on the town. A linear model is created to predict average download speeds based on drug offense rates. The summary of the linear model is displayed, providing insights into the relationship between download speeds and drug offenses. A graphical representation of the linear model is generated using ggplot, with data points colored differently for Kent and Surrey. The resulting plot visualizes the 2022 average download speeds against drug offense rates, including a linear regression line.

Figure 24-Average download speed vs Drug offence Rate

A graph showing a line

Description automatically generated

Average Download Speed vs Attainment Score

This begins by grouping broadband speed data by town and county, converting town names to lowercase for consistency, and calculating the average download speed for each group. It then groups school data by town and county, converting town names to lowercase, and finding the average attainment score for each group. The two datasets are joined into a single table based on the town. A linear model is created to predict average download speeds based on attainment scores. The summary of the linear model is displayed, providing insights into the relationship between download speeds and attainment scores. A graphical representation of the linear model is generated using ggplot, with data points colored differently for Kent and Surrey. The resulting plot visualizes the average download speeds against attainment scores, including a linear regression line.

Figure 25-Average Download speed vs Attainment Score

A graph with red and blue dots

Description automatically generated

# **Recommendation System**

House ranking

Margate came in first place from kent county with an average price of 125226. Also higher score is given to more affordable house.

Figure 26-House ranking

A screenshot of a computer program

Description automatically generated

Broadband Ranking

The town of Chatham from kent county came in first with the average download speed of 106 .

Figure 27-Broadband speed ranking

A screenshot of a computer

Description automatically generated

Crime ranking

Orpington town in surrey county came in 1st place with the least amount of crimes count.

A screenshot of a computer program

Description automatically generated

School ranking

Dartford town from kent came in 1st place with highest average attainment score.

A screenshot of a computer program

Description automatically generated

Overall ranking



# **Legal and ethical issues**

In the modern era, the utilization of data has resulted in a plethora of legal, ethical, and social dilemmas. One of the most crucial concerns is the protection of data. The collection, storage, and sharing of personal information can lead to infringement of individual rights. To safeguard privacy, it is imperative to implement measures against unauthorized access and misuse.

Ethical considerations are also of utmost importance. Striking a balance between utilizing data for gaining insights and respecting individuals' autonomy is a challenging task. When data is used to alter people's behavior, it exacerbates the ethical dilemma. This brings to light issues of transparency and potential manipulation.   
(*5 Principles of Data Ethics for Business*, 2021)

The aspect of safety is another crucial aspect in this scenario. Data breaches can result in theft of personal information, financial harm, and damage to reputation. Ensuring data security is not only a legal requirement, but also a moral obligation to prevent harm to individuals.

# **Conclusion**

In brief, the study aimed to assist a friend in selecting a suitable location in the UK for relocation. Key factors such as housing costs, broadband speed, school quality, and crime rates were examined to gather crucial information. The data underwent meticulous cleaning and was visually presented through graphs and charts for clarity. A specialized model was employed to explore correlations among various factors, leading to the creation of a ranking based on the analysis results.

# **References**

Elgabry, O. (2019, March 2). *The Ultimate Guide to data cleaning*. Medium. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

*R tutorial: Learn R programming language tutorial - javatpoint*. [www.javatpoint.com](http://www.javatpoint.com).

Police Data. (2022). Police Data Download. https://data.police.uk/data/

Registry, H. L. (2023, July 28). *Price paid data*. GOV.UK. <https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads>

*5 Principles of Data Ethics for Business*. (2021, March 16). Business Insights Blog. <https://online.hbs.edu/blog/post/data-ethics>

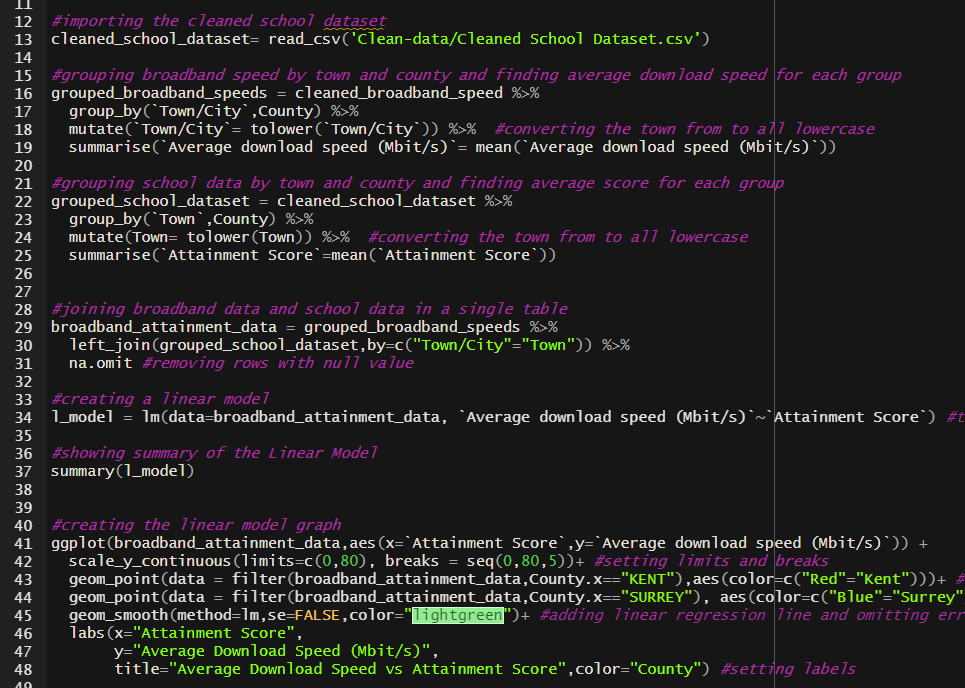
Mishra, P. (2021, December 19). *Linear Modeling*. GeeksforGeeks. <https://www.geeksforgeeks.org/how-to-use-lm-function-in-r-to-fit-linear-models/>

Ofcom. (2023, March 16). *Connected nations 2018: Data downloads*. Ofcom. https://www.ofcom.org.uk/research-and-data/multi-sector-research/infrastructure-research/connected-nations-2018/data-downloads

**Github Link**

**https://github.com/aasiskrk/Datasci-assign**

# **Appendix**



A screenshot of a computer screen

Description automatically generated

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