# Report on an Individual Data Science Project

Shishir Sharma

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

ST5014CEM Data Science for Developers

Siddhartha Neupane

Jan 25, 2024

Table of Contents

[Report on an Individual Data Science Project 1](#_Toc157119223)

[Table of Figure 3](#_Toc157119224)

[**Introduction** 4](#_Toc157119225)

[Cleaning data 5](#_Toc157119226)

[House pricing data cleaning 6](#_Toc157119227)

[Population data cleaning 7](#_Toc157119228)

[Broadband Speed Data Cleaning 8](#_Toc157119229)

[Crime rate Data Clining 9](#_Toc157119230)

[School Data Cleaning 10](#_Toc157119231)

[Exploratory data analysis 11](#_Toc157119232)

[Data visualization 12](#_Toc157119233)

[Housing Price 13](#_Toc157119234)

[Crime rate 18](#_Toc157119235)

[School 23](#_Toc157119236)

[Linear modelling 25](#_Toc157119237)

[House price vs Download speed 26](#_Toc157119238)

[House prices vs drug offence 27](#_Toc157119239)

[Attainment score vs drug offence rate 28](#_Toc157119240)

[Attainment score vs house prices 29](#_Toc157119241)

[Average download speed vs drug offence rate 30](#_Toc157119242)

[Average download speed vs attainment score 32](#_Toc157119243)

[Recommendation system 33](#_Toc157119244)

[Legal and ethical issues 35](#_Toc157119245)

[Conclusion 36](#_Toc157119246)

[References 36](#_Toc157119247)

[Appendix 37](#_Toc157119248)

# Table of Figure

[Figure 1 data cleaning 6](#_Toc157119249)

[Figure 2 data cleaning 7](#_Toc157119250)

[Figure 3 data cleaning 8](#_Toc157119251)

[Figure 4 data cleaning 9](#_Toc157119252)

[Figure 5 data cleaning 10](#_Toc157119253)

[Figure 6 average house price by counties 12](#_Toc157119254)

[Figure 7 average house price barchart 13](#_Toc157119255)

[Figure 8 average house price line graph 14](#_Toc157119256)

[Figure 9 Average download speed box plot 15](#_Toc157119257)

[Figure 10 Average Download speed Surrey 16](#_Toc157119258)

[Figure 11Average Download speed kent 17](#_Toc157119259)

[Figure 12 Drug offence Box plot 18](#_Toc157119260)

[Figure 13 robbery crime pie chart 19](#_Toc157119261)

[Figure 14 Drug offence line graph 20](#_Toc157119262)

[Figure 15 radar chart 21](#_Toc157119263)

[Figure 16 Attainment score box plot 22](#_Toc157119264)

[Figure 17 Attainment sore line graph kent 23](#_Toc157119265)

[Figure 18 attainment score line graph surrey 24](#_Toc157119266)

[Figure 19 house price vs average download speed 26](#_Toc157119267)

[Figure 20house price vs drug offence rate 27](#_Toc157119268)

[Figure 21 attainment score vs drug offence 28](#_Toc157119269)

[Figure 22 attainment score vs house price 29](#_Toc157119270)

[Figure 23download speed vs drug offence 30](#_Toc157119271)

[Figure 24 download speed vs attainment score 31](#_Toc157119272)

[Figure 25 House recommendation 32](#_Toc157119273)

[Figure 26 housing ranking 32](#_Toc157119274)

[Figure 27 school ranking 33](#_Toc157119275)

[Figure 28 Broadband speed 33](#_Toc157119276)

[Figure 29 crime rating 34](#_Toc157119277)

[Figure 30 Recommended city 34](#_Toc157119278)

### Introduction

the town recommendation system developed in R, meant to carefully assess towns based on numerous essential parameters, including educational institutions, cost of living (exemplified by home prices), internet speed, and safety ratings (expressed by local crime statistics). The inclusion of educational institutions extends beyond a mere acknowledgment of their location; rather, the system combines extensive school data to measure the quality, variety, and performance of schools, colleges, and universities within each town. This data-driven approach ensures that individuals or families with a special emphasis on academic success can make well-informed selections about their possible location of residence. The complete research also goes into the economic landscape by measuring the cost of living, specifically through house costs, providing a sophisticated insight of the financial consequences for prospective inhabitants. Additionally, the rating of broadband speed reflects the rising reliance on digital connectivity in contemporary society, catering to the needs of individuals involved in distant work, online education, and other digital activities. The safety ratings, drawn from local crime statistics, offer an objective perspective on the security and well-being of citizens, addressing concerns linked to personal safety and property protection. By harnessing the capabilities of R, this recommendation system intends to create a robust, flexible, and transparent platform that supports data-driven decision-making in the difficult process of town selection.

In examining Kent and Surrey, the R-based town recommendation system not only evaluates basic variables like cost of living and safety but delves down into granular school data, providing a critical assessment of academic potential within each town. This granular approach integrates standards Attainment score , crime rates, broadband speeds etc. The algorithm also evaluates varied housing possibilities, ensuring a customised perspective on the housing market in these places. By delving deep into specific data points, this system strengthens its potential to deliver brief and crucial recommendations for individuals or families choosing communities in Kent and Surrey, supporting a more informed decision-making process.

Keeping accuracy in mind all the data relased by official United Kingdom government. Ensuring efficient and reliable data source.

Cleaning data

Data cleaning a major part of data science, It's the process of finding and fixing mistakes or clutter in a dataset. This could mean dealing with missing information, getting rid of duplicates, and making sure everything is organized and accurate. The goal is to have clean and reliable data for better decision-making (Data cleaning in R, n.d.).

House pricing data cleaning

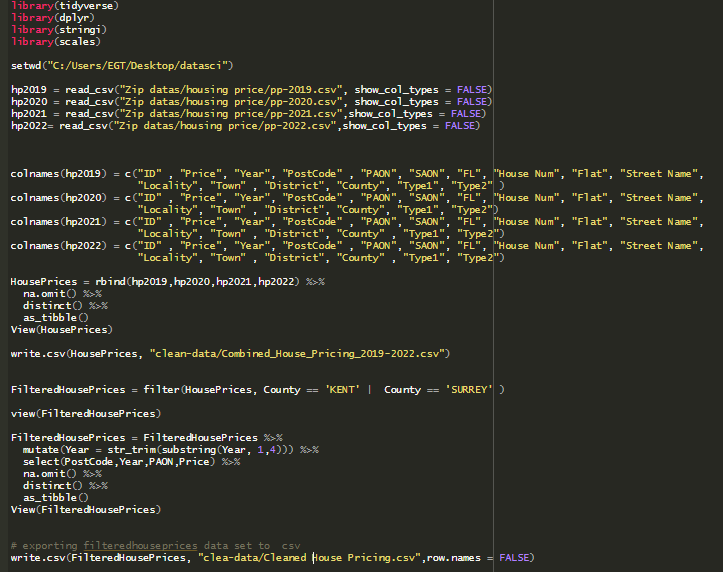
Upon downloading and importing the dataset using the read.csv function in R, the initial step involved filtering the information to focus exclusively on specific counties. Subsequently, the column name containing county information was adjusted for clarity. All individual data points were then consolidated to sift through and extract the pertinent information effectively. Leveraging the filter command, the dataset was refined to exclusively include house prices for kent and surrey. Finally, the amalgamated dataset and the filtered house price data were exported and saved employing the write.csv command for further analysis or utilization.

Figure 1 data cleaning

Population data cleaning

The R code loads a population dataset ("Population2011\_1656567141570.csv") using read\_csv, inspects it with View, and then cleans the data by removing duplicate rows with distinct and eliminating rows with any missing values using na.omit. The cleaned dataset is saved as "Clean\_population.csv" in the "clean-data" directory via write\_csv, and the Postcode column is sorted. This streamlined process ensures a polished population dataset without duplicates or missing values for further use.

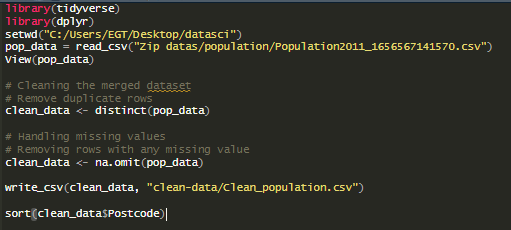


Figure 2 data cleaning

Broadband Speed Data Cleaning

Upon importing broadband performance and coverage datasets using the read\_csv function from the tidyverse and dplyr libraries in R, the datasets were merged based on the shared 'postcode' column using inner\_join. The merged dataset was examined and underwent a quality check to identify and report columns where every value was NA. Following this, duplicate rows were removed, and missing values in numeric columns were imputed with their respective means. The cleaned dataset was then exported to a new CSV file, "Clean Broadband.csv," ensuring a refined and reliable dataset for further analysis or utilization.

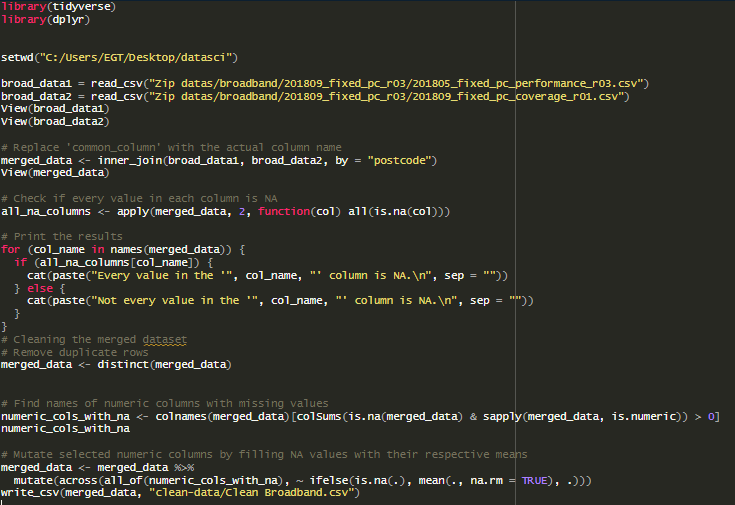


Figure 3 data cleaning

Crime rate Data Clining

Two crime datasets for Kent and Surrey in September 2023 are imported using read\_csv, and the data is merged using rbind. The subsequent data manipulation involves slicing values in the 'LSOA name' and 'Falls within' columns. Specifically, the last five characters are removed from 'LSOA name', and the suffix "Police " is removed from 'Falls within'. A null column labeled 'Context' is then deleted. Finally, the cleaned dataset is saved as "Clean\_crime.csv" in the "clean-data" directory using write\_csv.

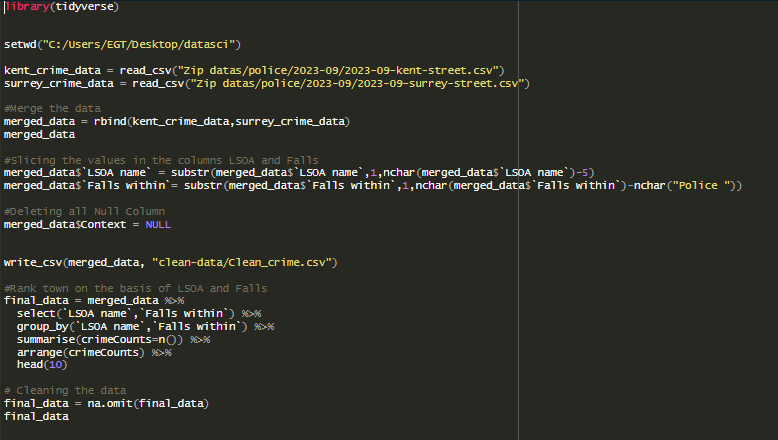


Figure 4 data cleaning

School Data Cleaning

The Imports and processes school data for Kent and Surrey from 2021 to 2023. It selects relevant columns, handles missing values, and merges the datasets. The cleaning steps involve replacing empty strings with NA, removing rows with any NA values, excluding rows with "NE" or "SUPP" in the 'ATT8SCR' column, converting 'ATT8SCR' to numeric format, and filtering out remaining NA values. The refined dataset is saved as "CleanSchooldata.csv" in the "clean-data" directory, providing a concise and standardized school dataset for further analysis or use.

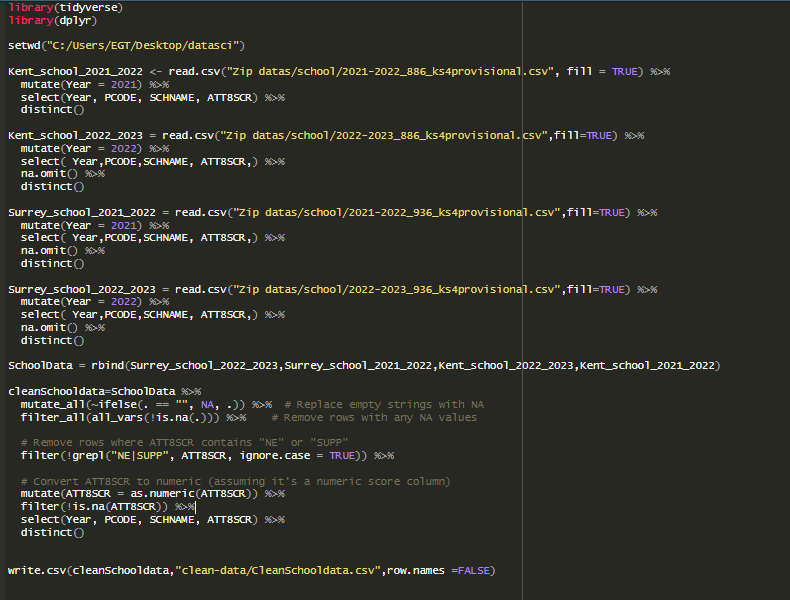


Figure 5 data cleaning

# **Exploratory data analysis**

Exploratory Data Analysis (EDA) is a vital phase in data analysis that focuses on learning and summarizing datasets to identify patterns and trends. The primary purpose is to gather insights and generate hypotheses for further inquiry. EDA involves summarizing data through descriptive statistics and visuals, such as histograms and scatter plots, to provide an overview of the dataset's primary aspects.

Cleaning the data is a crucial aspect of EDA, removing missing values, outliers, and assuring data consistency. Univariate analysis studies individual variables, while bivariate analysis explores interactions between pairs of variables, commonly using scatter plots or correlation matrices. Multivariate analysis digs at relationships among numerous variables concurrently, revealing complex patterns through techniques like clustering or dimensionality reduction.

During EDA, hypotheses are created based on observable patterns, guiding further statistical studies and modeling efforts. Interactive exploration technologies offer dynamic displays, allowing users to examine different elements of the dataset interactively. Effective communication of EDA findings, frequently through concise visualizations and summaries, is critical for disseminating insights to stakeholders and decision-makers. Overall, EDA is an iterative process that influences subsequent data analysis procedures and plays a critical role in extracting meaningful information from datasets.

## Data visualization

Data visualization is the graphical depiction of data, translating intricate data into charts, graphs, or other visual representations. It promotes understanding, highlighting patterns and trends in the data, making it more accessible for analysis and decision-making. Visualizing data is a valuable tool to express insights, identify outliers, and explain information in a clear and straightforward way. Ggplot from tidyverse is commonly used in data visualisation in R.

## Housing Price

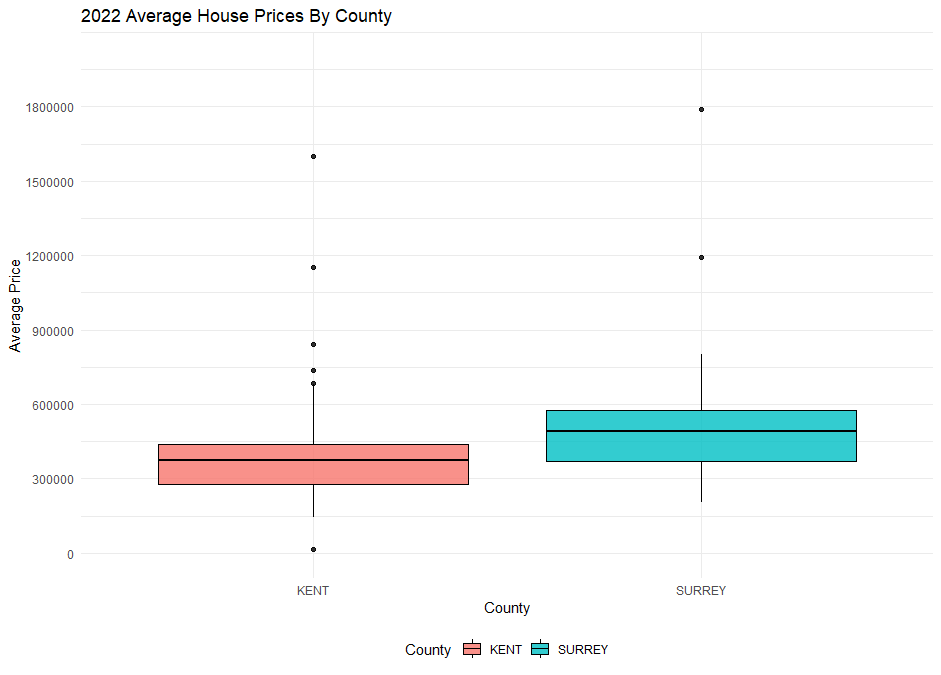


Figure 6 average house price by counties

The boxplot illustrates the distribution of average house prices in Kent and Surrey counties. The x-axis represents the two counties, Kent and Surrey, while the y-axis depicts the house prices. Each boxplot consists of a box that represents the interquartile range (IQR), with the median marked by a line inside the box. Whiskers extend from the box to indicate the data range, excluding outliers. The chart provides a visual comparison of the central tendency and variability in house prices between Kent and Surrey, offering insights into the distribution and potential differences in average property values across the two counties.

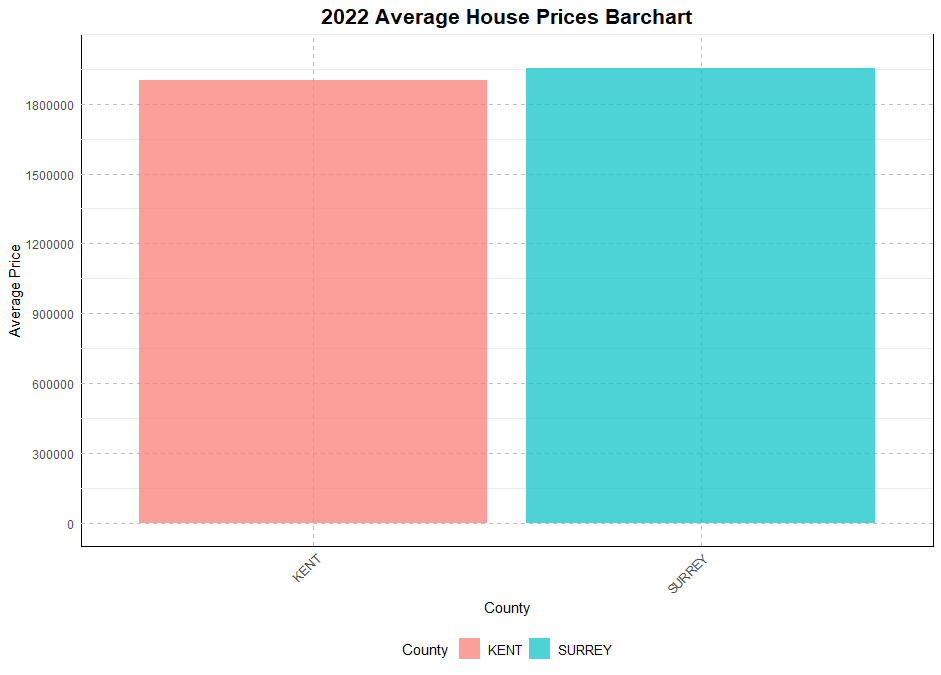


Figure 7 average house price barchart

The bar chart illustrates the average house prices in Kent and Surrey counties. Each bar represents one county, with the height indicating the respective average price. This straightforward visual comparison highlights any differences or similarities in average property values between Kent and Surrey.

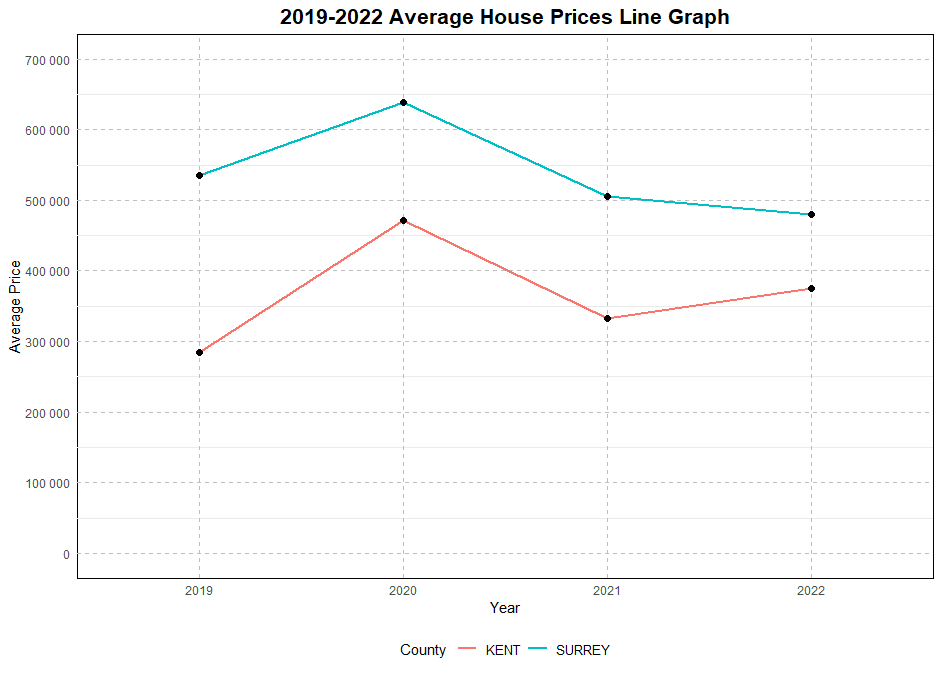


Figure 8 average house price line graph

The line chart depicts the trend in average house prices from 2019 to 2022. Over this period, the x-axis represents the years (2019, 2020, 2021, and 2022), while the y-axis displays the corresponding average house prices. The line connects the data points for each year, providing a visual representation of the price fluctuations over time. This type of chart allows for a quick observation of the overall trend in average prices across the specified years, enabling insights into whether prices have increased, decreased, or remained relatively stable during this period.  
Broadband speed

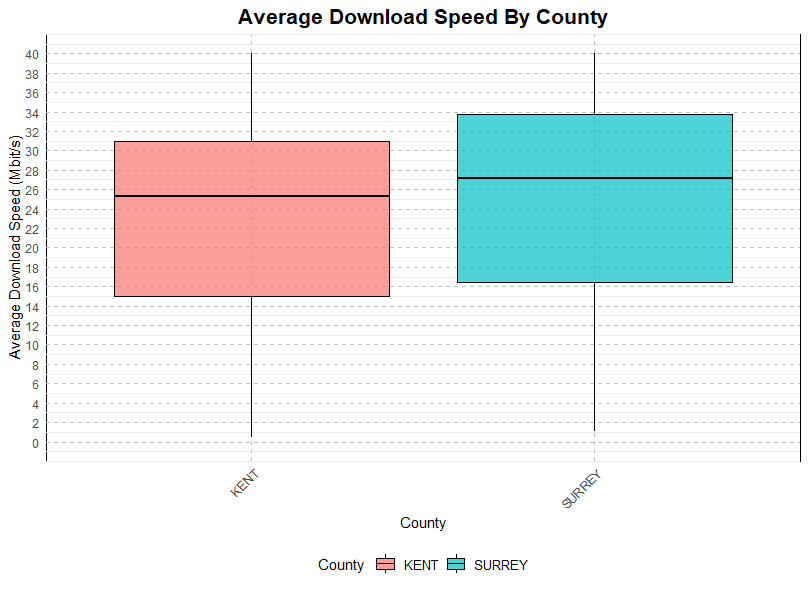


Figure 9 Average download speed box plot

The boxplot compares the average download speeds in Kent and Surrey. The x-axis distinguishes the two counties, while the y-axis represents average download speeds. Each boxplot summarizes the distribution, showing the median, interquartile range, and potential outliers. This visual comparison provides a quick overview of the typical download speeds and variability between Kent and Surrey.

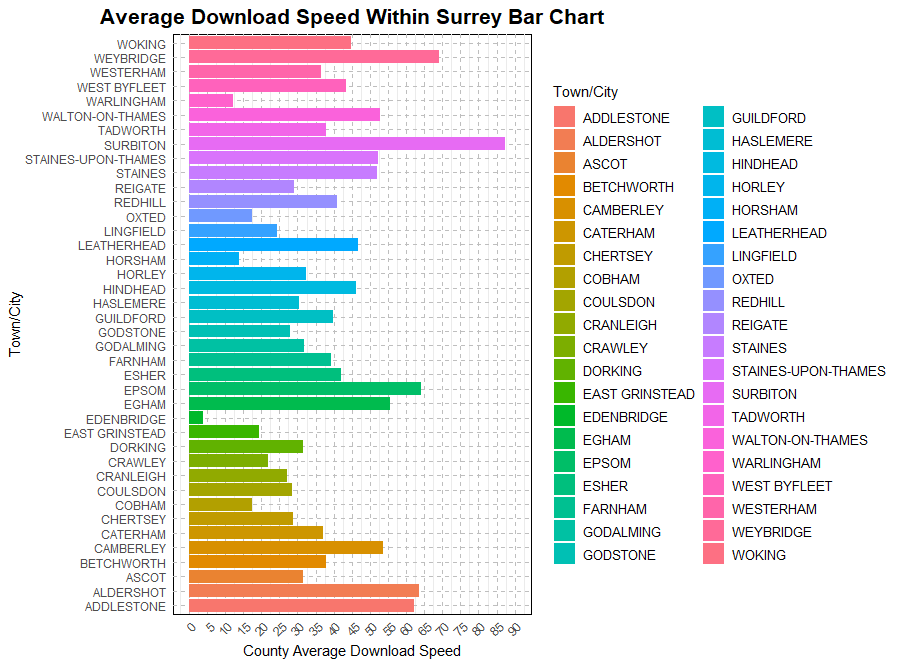


Figure 10 Average Download speed Surrey

A bar chart displaying the average download speeds in Surrey for different towns/cities would have the y-axis representing the towns/cities and the x-axis indicating the download speeds. Each bar would correspond to a specific town or city, with the height of the bar indicating the average download speed for that location. This visual representation provides a quick comparison of download speeds across different areas in Surrey, offering insights into variations in internet performance within the region.

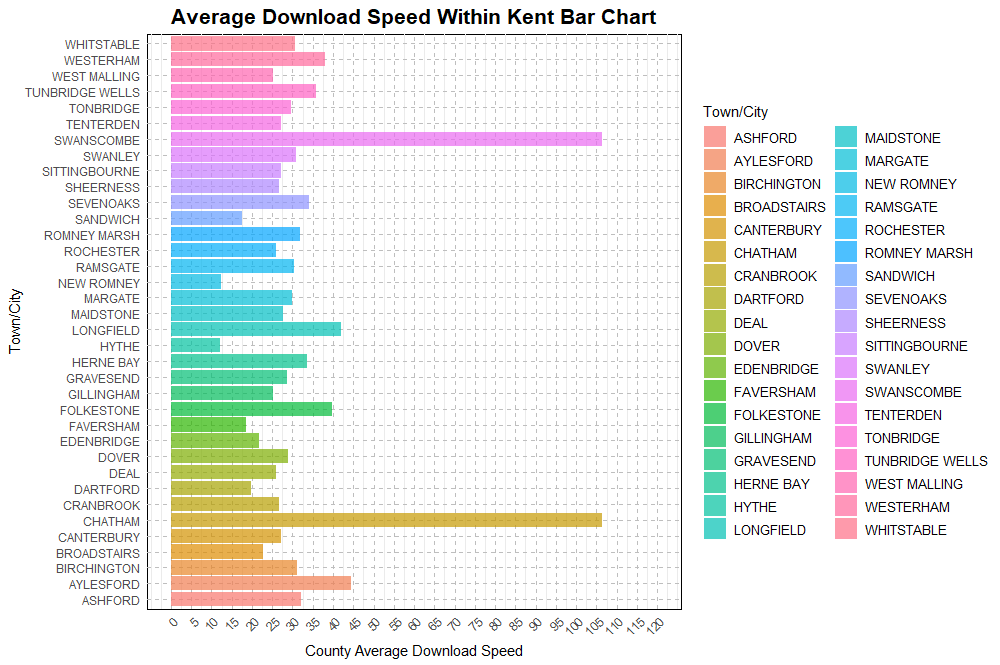


Figure 11Average Download speed kent

The bar chart illustrating the average download speeds in Kent for various towns/cities would feature the towns/cities on the y-axis and the corresponding average download speeds on the x-axis. Each bar would represent a specific town or city, with the height of the bar indicating the average download speed for that location. This visual representation facilitates a straightforward comparison of download speeds across different areas in Kent, providing insights into the variations in internet performance within the county.

Crime rate

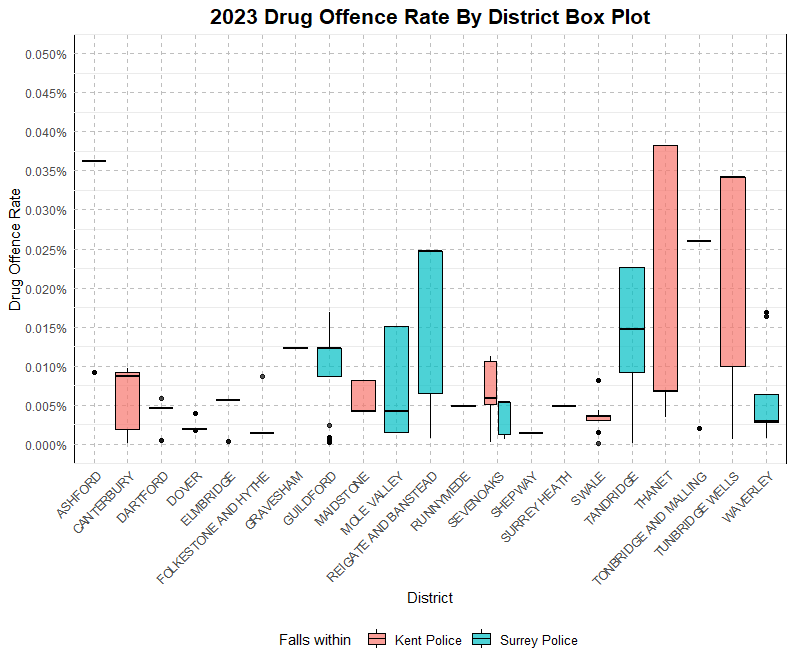


Figure 12 Drug offence Box plot

The box plot for 2023 drug offenses by district visually summarizes the distribution of these offenses. Each box represents the interquartile range, with the median marked inside. Whiskers extend to show the data range, and any outliers beyond the whiskers are individually marked. This provides a quick comparison of drug offenses across districts, offering insights into central tendency and potential outliers.

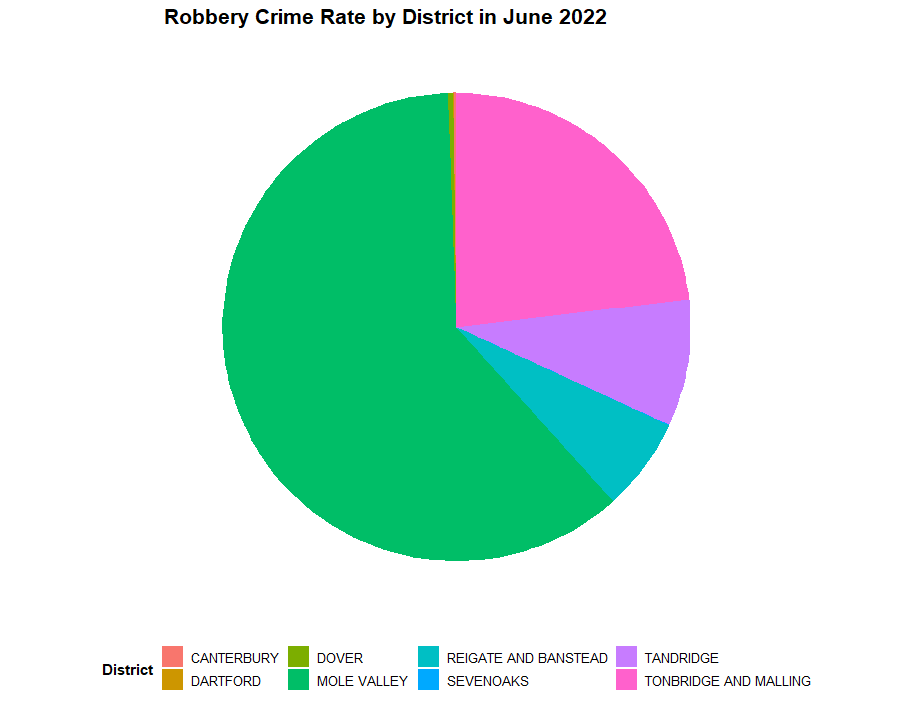


Figure 13 robbery crime pie chart

The pie chart for robbery crime rates in June 2022 succinctly illustrates the proportional distribution of such crimes. Each slice of the pie represents a specific category or location, showcasing the relative contribution of each to the overall robbery crime rate for that month. This visual representation offers a quick and concise overview of how different areas or categories contribute to the total robbery incidents in June 2022.

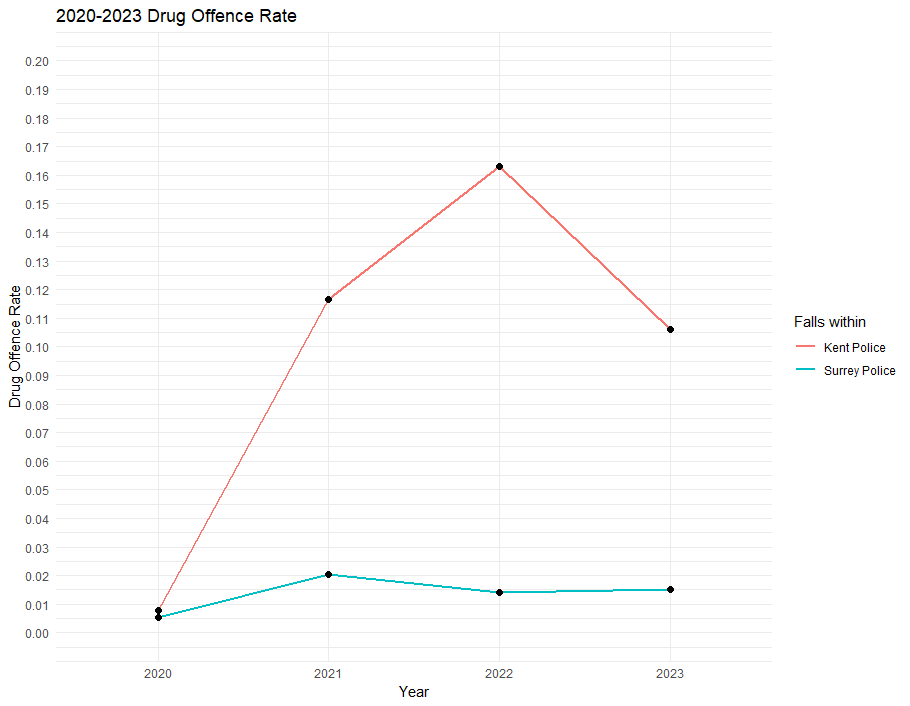


Figure 14 Drug offence line graph

The line graph for drug offense rates in Kent and Surrey from 2020 to 2023 provides a concise depiction of the trends over this period. The x-axis represents the years, while the y-axis displays the respective drug offense rates. The lines for Kent and Surrey showcase the fluctuation in drug offenses over the specified timeframe, offering a quick visual comparison of trends in these two regions.

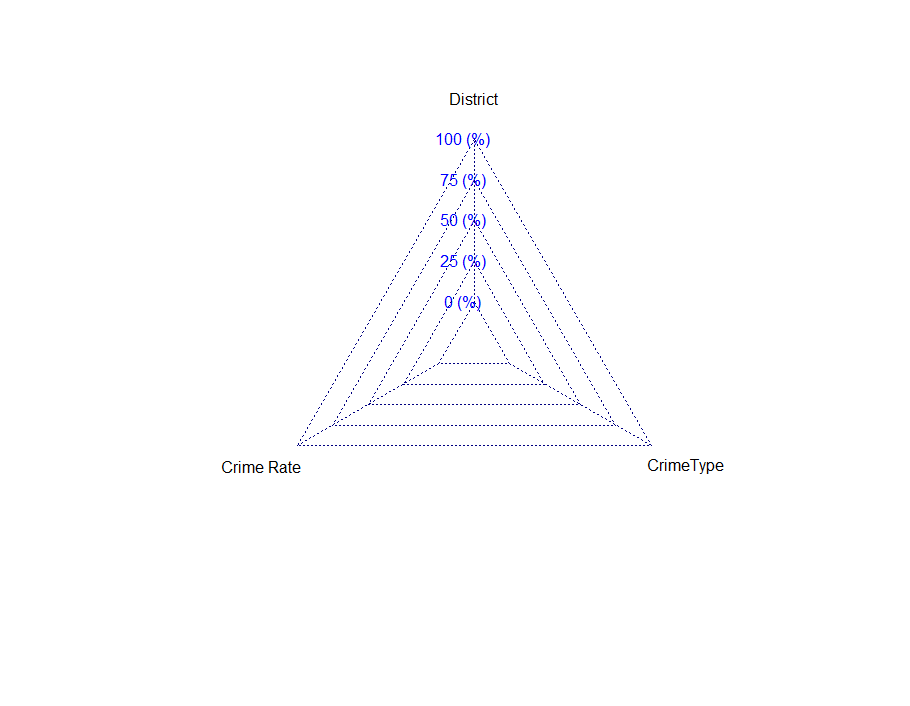


Figure 15 radar chart

School

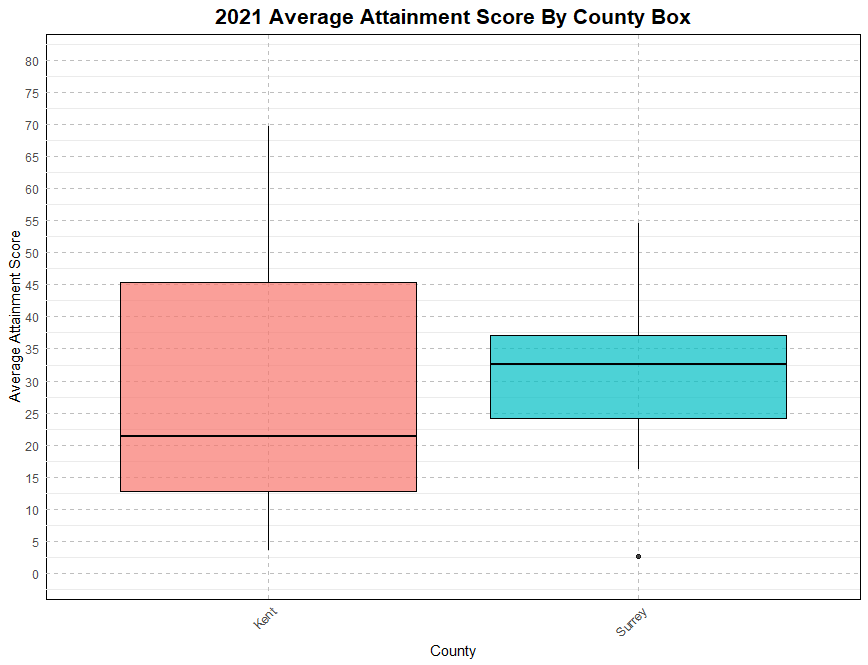


Figure 16 Attainment score box plot

The box plot for average attainment scores in Kent and Surrey succinctly displays the distribution of scores in these counties.

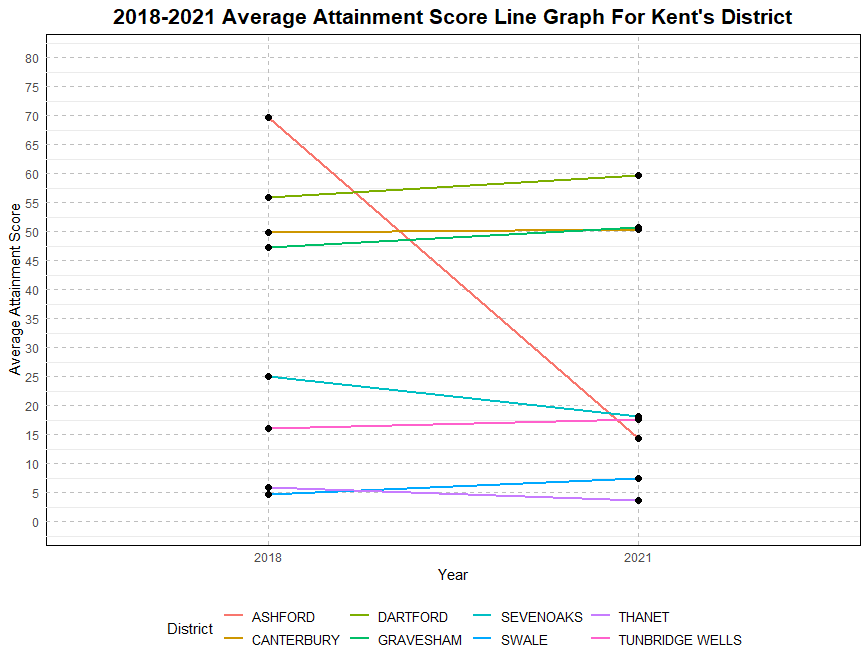


Figure 17 Attainment sore line graph kent

The line graph depicting attainment scores in Kent from 2018 to 2021 concisely illustrates the trend over this period.

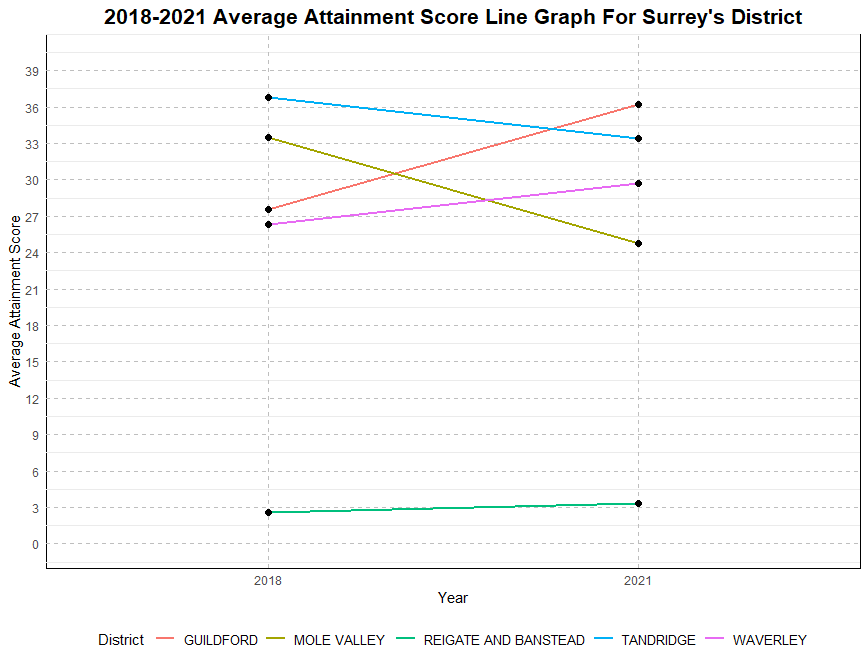


Figure 18 attainment score line graph surrey

The line graph depicting attainment scores in Surrey from 2018 to 2021 concisely illustrates the trend over this period.

Linear modelling

Linear modeling is a statistical technique used to evaluate the connection between a dependent variable and one or more independent variables. It presupposes a linear relationship, stating this as a simple equation where changes in the independent variables correspond to proportional changes in the dependent variable. The model involves estimating coefficients for the intercept and slopes that best suit the observed data (What Is a Linear Model?, n.d.). Commonly utilised in sectors such as economics and social sciences, linear modeling provides insights into the strength and direction of correlations between variables (Faraway, n.d.). While it presupposes linearity and may not capture complicated connections, it serves as a core tool for understanding and forecasting relationships in diverse domains.

## House price vs Download speed

The short linear model for 2020 house prices vs. average download speed in Kent and Surrey suggests a modest negative correlation. The line slopes slightly downwards, indicating that, on average, as the download speed increases, there is a slight decrease in house prices. This simple linear relationship provides a quick insight into the potential association between 2020 house prices and average download speed in the regions of Kent and Surrey.

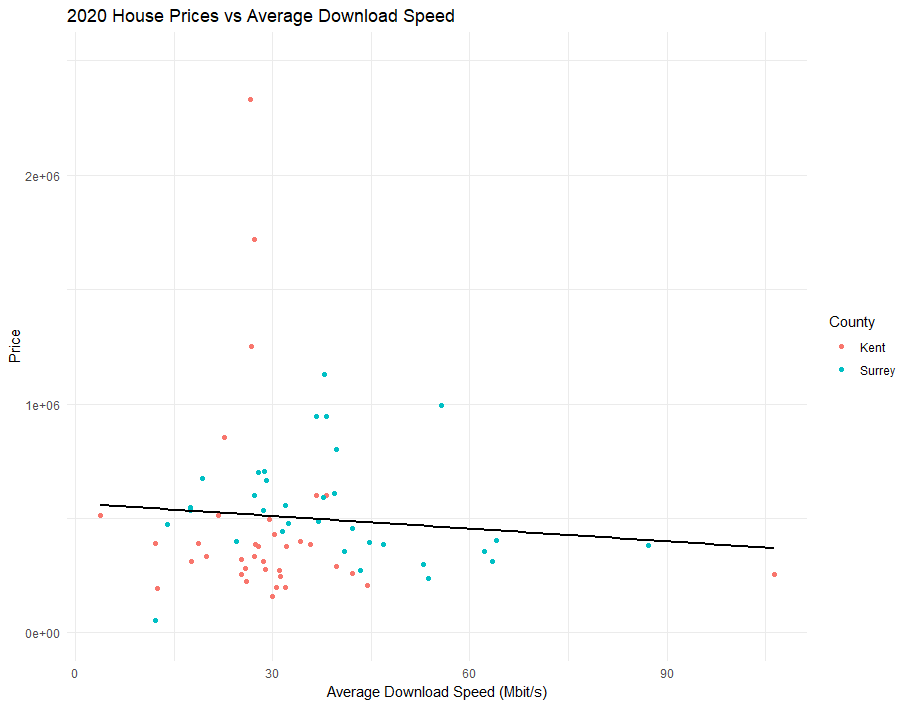


Figure 19 house price vs average download speed

## House prices vs drug offence

The brief linear model for 2020 house prices vs. drug offence rates in Kent and Surrey indicates a slight negative correlation. The downward slope of the line suggests that, on average, as drug offence rates increase, there is a slight decrease in house prices. This simple linear relationship provides a concise overview of the potential association between 2020 house prices and drug offence rates in the regions of Kent and Surrey.

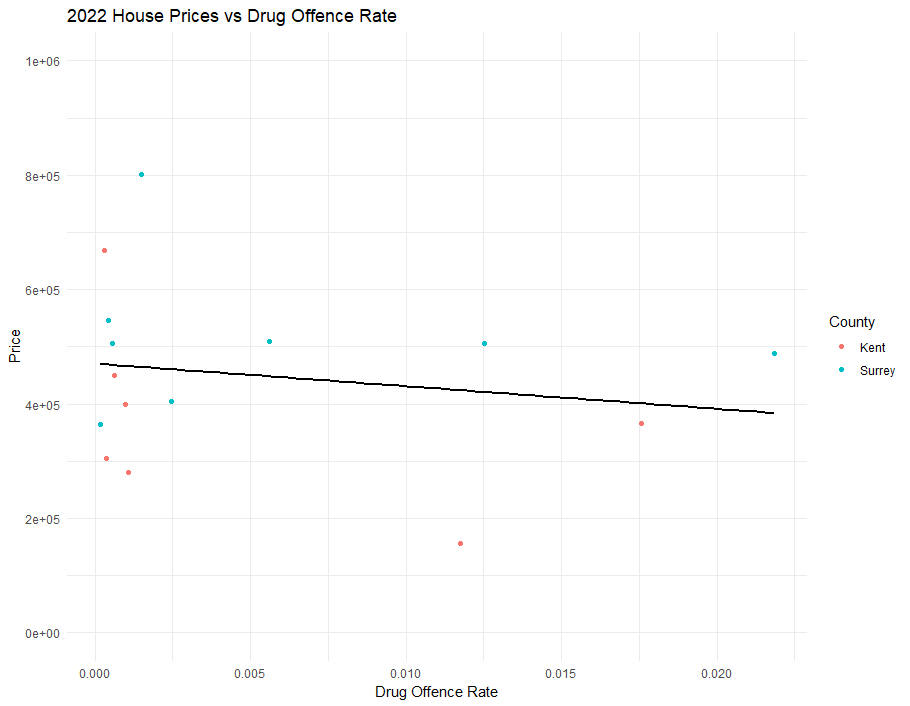


Figure 20house price vs drug offence rate

## Attainment score vs drug offence rate

The succinct linear model for attainment scores vs. drug offence rates in Kent and Surrey reveals a negative correlation (Picardo, n.d.). The downward slope of the line implies that, on average, as drug offence rates increase, there is a decrease in attainment scores. This concise linear relationship offers a quick insight into the potential association between attainment scores and drug offence rates in the regions of Kent and Surrey.

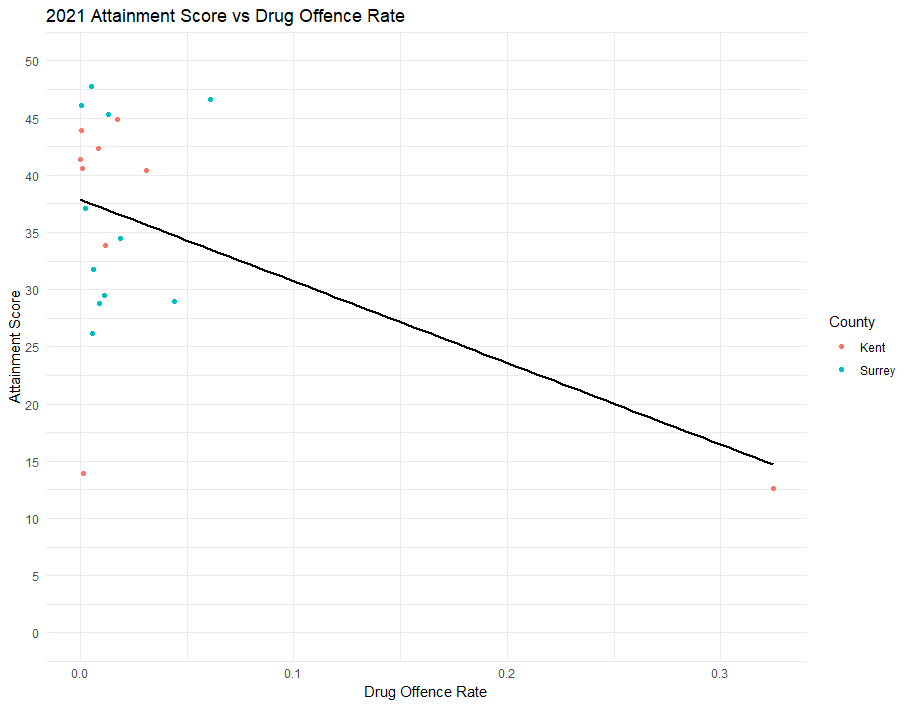


Figure 21 attainment score vs drug offence

## Attainment score vs house prices

The brief linear model for attainment scores vs. house pricing rates in Kent and Surrey suggests a positive correlation. The upward slope of the line indicates that, on average, as house pricing rates increase, there is an upward trend in attainment scores. This concise linear relationship provides a quick insight into the potential association between attainment scores and house pricing rates in the regions of Kent and Surrey.

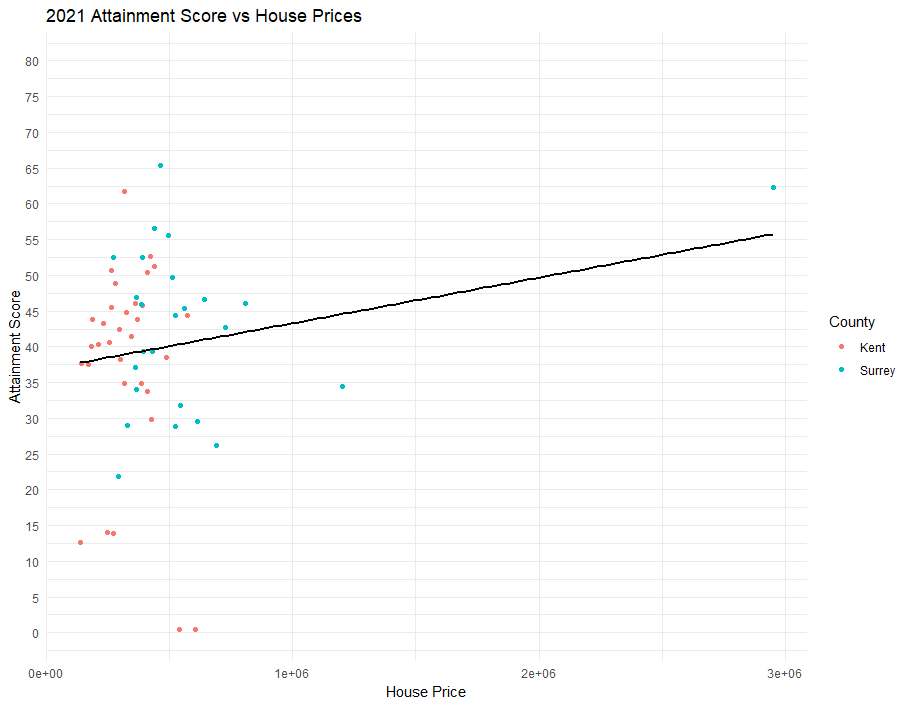


Figure 22 attainment score vs house price

## Average download speed vs drug offence rate

The concise linear model for download speed vs. drug offence rates in Kent and Surrey suggests a slight positive correlation. The upward slope of the line indicates that, on average, as download speeds increase, there is a minor upward trend in drug offence rates. This brief linear relationship provides a quick insight into the potential association between download speed and drug offence rates in the regions of Kent and Surrey.

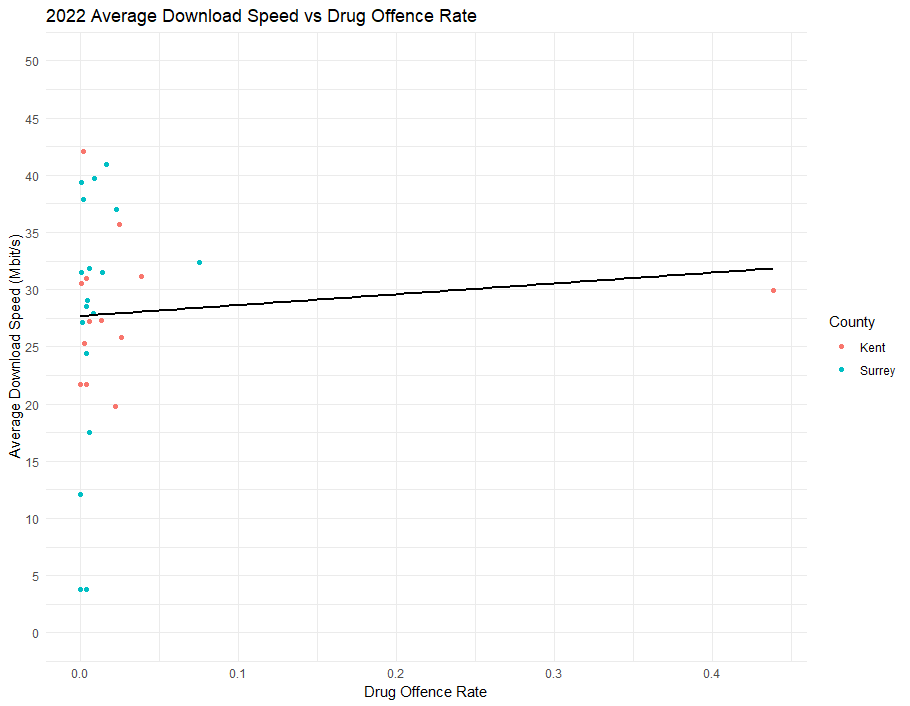


Figure 23download speed vs drug offence

## Average download speed vs attainment score

The succinct linear model for download speed vs. attainment scores in Kent and Surrey suggests a stable or slightly positive correlation. The relatively straight or slightly upward slope of the line indicates that, on average, as download speeds increase, there may be a modest increase in attainment scores. This brief linear relationship provides a quick insight into the potential association between download speed and attainment scores in the regions of Kent and Surrey.

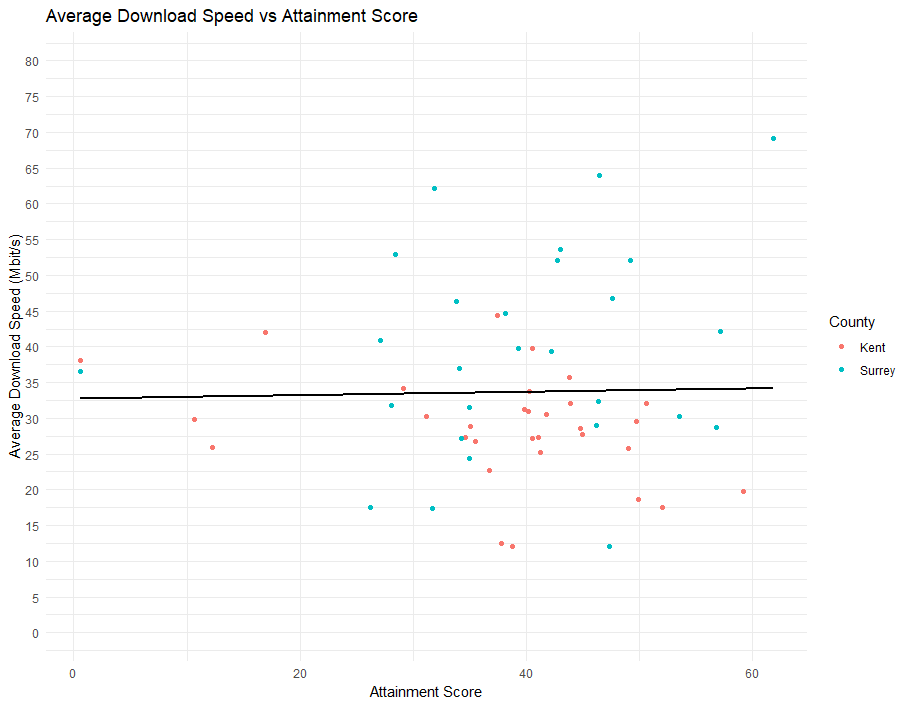


Figure 24 download speed vs attainment score

Recommendation system

recommendation system centered on average scores, employing the R programming language for a concise and effective approach. Utilizing average scores as a key metric allows for a streamlined evaluation of various factors. The system's simplicity and clarity in presenting average scores enable users to quickly assess and compare towns. This recommendation system, rooted in R, provides a straightforward yet robust tool for individuals seeking data-driven insights to inform their decision-making process in selecting optimal living environments.

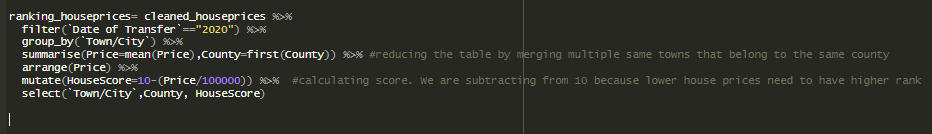


Figure 25 House recommendation

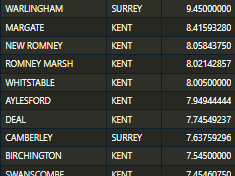


Figure 26 housing ranking

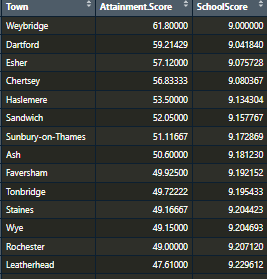


Figure 27 school ranking

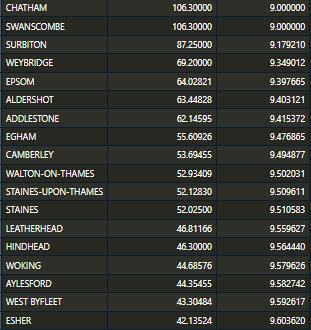


Figure 28 Broadband speed

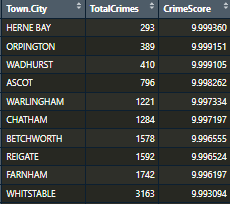


Figure 29 crime rating



Figure 30 Recommended city

# **Legal and ethical issues**

In the evolving field of data science, navigating legal and ethical considerations is paramount. One key problem revolves around data privacy and protection, demanding compliance with regulations to avoid illegal access and misuse of personal information. This legal duty corresponds with ethical imperatives to build comprehensive mechanisms, assuring appropriate data collecting and sharing methods.

Informed consent is an another idea that carries both legal and ethical relevance. Legally, gaining unambiguous consent is necessary, with noncompliance incurring legal implications and privacy legislation breaches. Ethically, securing informed consent demonstrates a commitment to transparency, empowering individuals to comprehend and regulate the usage of their personal data.

Algorithmic bias and fairness constitute multiple difficulties with legal and ethical elements (Awan, n.d.). Legal frameworks restrict discrimination based on protected qualities, obligating enterprises to rectify biases in algorithms. Ethically, maintaining justice in algorithmic decision-making is crucial to prevent propagating societal disparities. Openness and transparency, mandated by some rules and ethically accepted, involve explicit disclosure about data practices, creating trust among stakeholders. In summary, responsible data science means balancing legal duties while respecting ethical standards to prioritize individual rights and society well-being.

# **Conclusion**

In summary, this analysis illustrates the usefulness of applying linear modeling in R for town recommendations. The incorporation of data visualization promotes the accessibility of complicated information, giving consumers with a clear knowledge of crucial elements. The system's complete analysis, including school data, helps to a comprehensive decision-making process. In essence, the combination of linear modeling, data visualization, and recommendation algorithms in R gives an academic yet user-friendly approach to town selection, supporting educated decisions for persons seeking optimal living situations.

# References

Awan, A. A. (n.d.). *What is Algorithmic Bias?* Retrieved from datacamp: https://www.datacamp.com/blog/what-is-algorithmic-bias

*Data cleaning in R*. (n.d.). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/data-cleaning-in-r/

Faraway, J. (n.d.). *Linear models*. Retrieved from sience direct: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/linear-models

Picardo, E. (n.d.). *Negative Correlation: How it Works, Examples And FAQ*. Retrieved from investopedia: https://www.investopedia.com/terms/n/negative-correlation.asp#:~:text=Negative%20correlation%20is%20a%20relationship,indicates%20a%20perfect%20positive%20correlation.

*What Is a Linear Model?* (n.d.). Retrieved from Math works: https://www.mathworks.com/discovery/linear-model.html#:~:text=What%20Is%20a%20Linear%20Model,%2C%20financial%2C%20and%20biological%20data.

Appendix  
https://github.com/shiX0/data-science-