Aayush Shrestha (230293)

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

ST5014CEM Data Science for Developers

Siddhartha Neupane

Contents

Table of Figure

Introduction

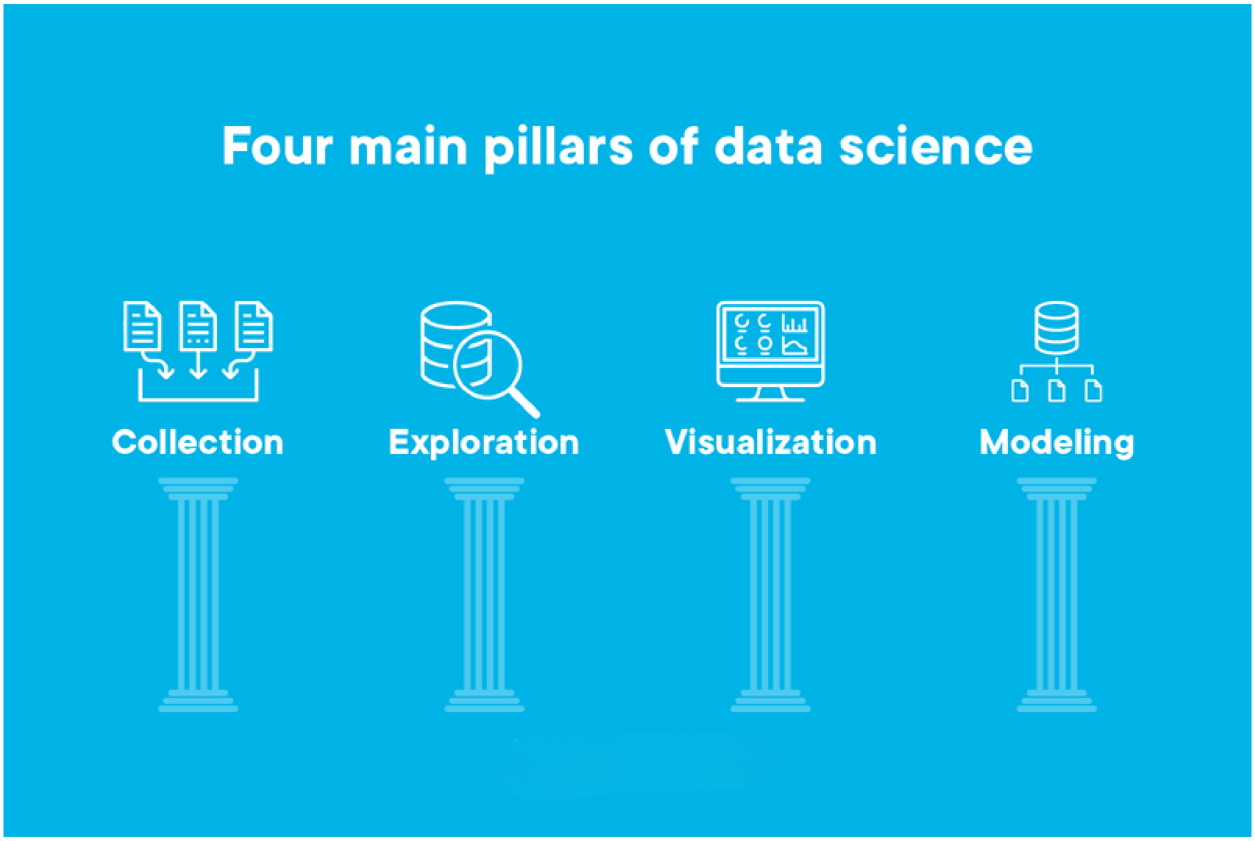
Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results (*What Is Data Science? - Data Science Explained - AWS*, n.d.-b)

This data science assignment aims to analyze real-world datasets that the instructor has made available via a download link. The four main datasets that are the focus of this assignment are town-level data, broadband speed and coverage data, street-level crime data, and school performance data.

Understanding the relationships between different social and economic factors in South and West Yorkshire towns is the primary objective of this project. After cleaning, trends and patterns in areas such as town characteristics, crime rates, broadband availability, and school attainment scores were examined using data visualization techniques. Additionally, linear modeling was used to look at how different factors, like crime or internet speed, affected outcomes like house prices or academic performance.

I created different visualizations for this assignment to study house prices, broadband speed, crime rates, and school performance in South and West Yorkshire. For house prices, I used a line graph to compare average prices from 2021 to 2024 for both counties, and boxplots to show price differences by district in each county. Broadband speed was shown with boxplots for download speeds across districts and bar charts comparing speeds in towns for each county. Crime data was visualized using boxplots for drug offense rates by district, a radar chart for vehicle crime in one county for a specific time, and a line chart to track drug offenses over the years in both counties. For schools, I made boxplots of average Attainment 8 scores for South Yorkshire in 2023 and West Yorkshire in 2022, plus a line graph showing how these scores changed over time across districts. These visuals help to understand patterns and compare the two countries easily.

Figure 1 :Data Science



Kekare (2025b)

Data cleaning

Data cleansing is the process of finding and removing errors, inconsistencies, duplications, and missing entries from data to increase data consistency and quality, also known as data scrubbing or cleaning (*What Is Data Cleansing? | TIBCO*, n.d.-b). In this assignment, the raw data we downloaded contains errors, missing values, duplicate values, and unformatted data. If we don't clean the data, it can lead to wrong results or confusing visualization.

For this coursework, we had to clean up the town, broadband, house pricing, crime, and school datasets so that they could be used together effectively. For instance, we handled missing values, eliminated extra spaces, and ensured that postcodes followed the same format. This allowed us to align data from multiple sources and prepare it for analysis and visualization.

Figure 2 :Data cleaning

Anwar (2025b)

House Pricing

I started by using the add row function to merge the 2021 and 2024 datasets into a single dataset for the house pricing data. I then restricted the data to records from West Yorkshire and South Yorkshire. I extracted the first four characters of the entire postcode and eliminated any spaces to create a new column called short Postcode that would make data joining easier in the future. For time-based analysis, I also eliminated the year from the Date column. I only chose the required columns—price, date, county, district, postcode, short postcode, and postcode—after sorting the data by county. To use the cleaned data for additional analysis and visualization, I finally saved it to a CSV file. The code for house pricing data cleaning.

Figure 3 :House Pricing

```
HousePrices = HousePrices2021%>%
  add_row(HousePrices2022) %>%
  add_row(HousePrices2023) %>%
  add_row(HousePrices2024)
cleanHousePrices = HousePrices %>%
  filter(County=="SOUTH YORKSHIRE"|County=="WEST YORKSHIRE") %>%
  mutate(shortPostcode = str_trim(substring(Postcode, 1,4))) %>%
  mutate(Year=substring(Date,7,10)) %>%
  arrange(County) %>%
  select(Postcode,shortPostcode,Price,Date,County,District)
write.csv(cleanHousePrices, "C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_House_Prices.csv")
```

Towns

I started by gathering all the towns' 2021–2024 home price information. After that, I limited the records to only include West Yorkshire and South Yorkshire. To connect this to population data, I created a short postcode using the first four characters of the full postcode. To calculate the estimated populations for each year between 2012 and 2024, I also changed the population column to numeric, removed rows with invalid values, and applied growth multipliers. I then selected relevant columns like district, county, and annual population, and combined the population and home price data using the short Postcode. After removing duplicates and classifying the data by county, I saved the cleaned town-level dataset to a CSV file. And I saved cleaning like write.csv (Towns, "C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv").

Figure 4 :Towns

```r
library(tidyverse)


PopulationData =read.csv("C:/DataScience-R/AayushShrestha-230293/Obtain/population/Population2011_1656567141570.csv")
colnames(PopulationData)

column_names = c(
  "Housenumber","Price","Date",
  "Postcode", "Zone 1","Zone 2",
  "Zone 3","PAON","SAON", "Street",
  "Locality","Town", "District",
  "County","NA1","NA2"
)



HousePrices2021 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/house pricing/pp-2021.csv", col_names = FALSE) %>%
  set_names(column_names)

HousePrices2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/house pricing/pp-2022.csv", col_names = FALSE) %>%
  set_names(column_names)

HousePrices2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/house pricing/pp-2023.csv", col_names = FALSE) %>%
  set_names(column_names)

HousePrices2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/house pricing/pp-2024.csv", col_names = FALSE) %>%
  set_names(column_names)

PopulationData <- read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/population/Population2011_1656567141570.csv")

# Clean, convert and calculate population by short postcode
PopulationData_clean <- PopulationData %>%
  # Remove rows where Population cannot convert to numeric
  filter(!is.na(as.numeric(Population))) %>%
  # Convert Population to numeric
  mutate(Population = as.numeric(Population)) %>%
  # Extract first 4 characters from Postcode, trim spaces
  mutate(shortPostcode = str_trim(substring(Postcode, 1, 4))) %>%
  # Group by shortPostcode and sum population
  group_by(shortPostcode) %>%
  summarise(Population2011 = sum(Population, na.rm = TRUE)) %>%
  # Calculate population for years 2012 to 2024 using multipliers
  mutate(
    Population2012 = 1.00695353132322269 * Population2011,
    Population2013 = 1.00669740535540783 * Population2012,
    Population2014 = 1.00736463978721671 * Population2013,
    Population2015 = 1.00792367505802859 * Population2014,
    Population2016 = 1.00757874492811929 * Population2015,
    Population2017 = 1.00679374473924223 * Population2016,
    Population2018 = 1.00605929132212552 * Population2017,
    Population2019 = 1.00561255390388033 * Population2018,
    Population2020 = 1.00561255390388033 * Population2019,
    Population2021 = 1.005425 * Population2020,
    Population2022 = 1.004920 * Population2021,
    Population2023 = 1.004510 * Population2022,
    Population2024 = 1.004220 * Population2023
  )
PopulationData_clean %>%
  select(shortPostcode, Population2021, Population2022, Population2023, Population2024)




HousePrices = HousePrices2021 %>%
  add_row(HousePrices2022) %>%
  add_row(HousePrices2023) %>%
  add_row(HousePrices2024)

Towns <- HousePrices %>%
  filter(County %in% c("SOUTH YORKSHIRE", "WEST YORKSHIRE")) %>%
  mutate(shortPostcode = str_trim(substr(Postcode, 1, 4))) %>%
  left_join(PopulationData_clean, by = "shortPostcode") %>%
  select(shortPostcode, District, County, Population2021, Population2022, Population2023, Population2024) %>%
  group_by(shortPostcode) %>%
  filter(row_number() == 1) %>%
  arrange(County)

write.csv(Towns, "C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv")
```

Broadband Speed

I worked with two distinct datasets for the broadband speed data: coverage and performance. To create a short postcode for joining, I first cleaned the coverage data by extracting the first four characters of the postcode. I then chose crucial columns such as the percentage of premises with speeds below 2 Mbps and 10 Mbps, as well as the availability of superfast and ultrafast broadband. By creating the same short Postcode and choosing important columns like median and average download/upload speeds and average data usage, I similarly cleaned the performance data. I then used both the postcode and short Postcode to join the two datasets. Speed and availability data are included in the final cleaned broadband dataset, which I saved as a CSV file for analysis and visualization.

Figure 5 :Broadband speed

```r


library(tidyverse)

coverage = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Broadband/201809_fixed_pc_coverage_r01.csv")
performance = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Broadband/201805_fixed_pc_performance_r03.csv")

colnames(coverage)
length(colnames(coverage))

colnames(performance)
length(colnames(performance))

clean_coverage = coverage %>%
  mutate(shortPostcode = str_trim(substr(postcode, 1,4))) %>%
  select(
    postcode,
    shortPostcode,
    SFBB_Availability = `SFBB availability (% premises)`,
    UFBB_Availability = `UFBB availability (% premises)`,
    FTTP_Availability = `FTTP availability (% premises)`,
    Below_2Mbps = `% of premises unable to receive 2Mbit/s`,
    Below_10Mbps = `% of premises unable to receive 10Mbit/s`
  )

clean_performance = performance %>%
  mutate(shortPostcode = str_trim(substr(postcode,1,4))) %>%
  select(
    postcode,
    shortPostcode,
    Median_Download = `Median download speed (Mbit/s)`,
    Avg_Download = `Average download speed (Mbit/s)`,
    Median_Upload = `Median upload speed (Mbit/s)`,
    Avg_Upload = `Average upload speed (Mbit/s)`,
    Avg_Data_Usage = `Average data usage (GB)`
  )

Broadband = left_join(clean_coverage,clean_performance,by = c("postcode","shortPostcode"))

write.csv(Broadband,"C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_BroadBand_Speed.csv")
```

Crime Dataset

I combined the street-level crime data for 2022–2025 from West Yorkshire and South Yorkshire. I combined all of the annual files into a single dataset using the add row function. I then selected key columns like Crime ID, LSOA name, Month, Reported by, falls within, Location, and Crime type to clean up the data. To make it obvious which area each crime belongs to, I also made a new County column from the "Falls within" field. I saved the cleaned dataset to a CSV file after classifying the records by location and county. After being cleaned, the data was prepared for additional analysis, including determining crime rates and displaying trends over time.

Figure 6 :Crime

```
library(tidyverse)

south_yorkshire_2025 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2025/2025-04/2025-04-south-yorkshire-street.csv")
west_yorkshire_2025 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2025/2025-04/2025-04-west-yorkshire-street.csv")

south_yorkshire_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2024/2024-04/2024-04-south-yorkshire-street.csv")
west_yorkshire_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2024/2024-04/2024-04-west-yorkshire-street.csv")

south_yorkshire_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2023/2023-04/2023-04-south-yorkshire-street.csv")
west_yorkshire_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2023/2023-04/2023-04-west-yorkshire-street.csv")

south_yorkshire_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2022/2022-06/2022-06-south-yorkshire-street.csv")
west_yorkshire_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/crime data set/crime2022/2022-06/2022-06-west-yorkshire-street.csv")


CrimeData = south_yorkshire_2025 %>%
  add_row(west_yorkshire_2025) %>%
  add_row(south_yorkshire_2024) %>%
  add_row(west_yorkshire_2024) %>%
  add_row(south_yorkshire_2023) %>%
  add_row(west_yorkshire_2023) %>%
  add_row(south_yorkshire_2022) %>%
  add_row(west_yorkshire_2022)


Clean_crime= CrimeData %>%
  mutate(
    County = `Falls within`
  ) %>%
  select(
    CrimeID = `Crime ID`,
    LSOAname = `LSOA name`,
    Month,
    Reportedby = `Reported by`,
    Fallswithin = `Falls within`,
    County,
    Location,
    CrimeType = `Crime type`,
  ) %>%
  arrange(County, Location)

write.csv(Clean_crime,"C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_Crime_Dataset.csv")
```

School Dataset

I performed several cleaning and filtering operations on the school datasets from 2021 to 2024 to concentrate solely on schools in South Yorkshire and West Yorkshire. To include only schools in the target counties, I first filtered them by their town or address fields. I then used common identifiers like `URN` and `ESTAB` to integrate school data with related datasets like provisional performance, results, and pupil destinations. Key variables such as school name, postcode, school type, gender, age range, and performance indicators such as Progress 8 (P8MEA) and EBACC scores were chosen from these combined datasets. I saved the cleaned data independently for every academic year after making sure that duplicate records were eliminated.

Figure 7 : School 2021-2023

```r
library(tidyverse)

library(tidyverse)
library(readxl)

final_2021_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2021-2022/england_ks4final.csv")
mats_performance_2021_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2021-2022/england_ks4-mats-performance.csv")
provisional_2021_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2021-2022/england_ks4provisional.csv")
pupdest_2021_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2021-2022/england_ks4-pupdest.csv")

underlying_1_2021_2022 = read_excel("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2021-2022/england_ks4underlying_1.xlsx")
underlying_entriesandgrades_2_2021_2022 = read_excel("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2021-2022/england_ks4underlying_entriesandgrades_2.xlsx")

school_information_2021_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2021-2022/england_school_information.csv")

# Check duplicates in existing dataframes (make sure column 'URN' exists in each)
anyDuplicated(final_2021_2022$URN)
anyDuplicated(pupdest_2021_2022$URN)
anyDuplicated(mats_performance_2021_2022$URN)










keywords = c("South Yorkshire", "West Yorkshire")

school_filtered_2021_2022 = school_information_2021_2022 %>%
  filter(
    toupper(str_trim(TOWN)) %in% toupper(keywords) |
      toupper(str_trim(ADDRESS3)) %in% toupper(keywords)
  )




provisional_joined_2021_2022 = school_filtered_2021_2022 %>%
  inner_join(provisional_2021_2022, by = c("URN", "ESTAB"))


pupdest_joined_2021_2022 = provisional_joined_2021_2022 %>%
  left_join(pupdest_2021_2022, by = c("URN", "ESTAB"))


final_joined_2021_2022 = pupdest_joined_2021_2022 %>%
  left_join(final_2021_2022, by = c("URN", "ESTAB"))


cleaned_2021_2022 = final_joined_2021_2022 %>%
  select(
    URN,
```

```r
cleaned_2021_2022 = final_joined_2021_2022 %>%
  select(
    URN,
    SCHNAME = SCHNAME.x,
    TOWN,
    ADDRESS3,
    LANAME,
    POSTCODE,
    SCHOOLTYPE,
    GENDER,
    AGELOW,
    AGEHIGH,
    P8MEA = P8MEA.x,
    P8MEA_FSM6CLA1A = P8MEA_FSM6CLA1A.x,
    P8MEA_NFSM6CLA1A = P8MEA_NFSM6CLA1A.x,
    EBACCAPS = EBACCAPS.x,
    PTEBACC_95 = PTEBACC_95.x,
    OVERALL_DESTPER,
    EMPLOYMENTPER,
    EDUCATIONPER
  ) %>%
  distinct()


write.csv(cleaned_2021_2022, "C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2021-2022.csv")

View(cleaned_2021_2022)
```

Figure 8 :School 2022-2023

```r
library(tidyverse)

ks2_final_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks2final.csv")
ks4_final_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks4final.csv")
ks2_mats_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks2-mats-performance.csv")
ks4_mats_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks4-mats-performance.csv")
provisional_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks4provisional.csv")
pupdest_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks4-pupdest.csv")
underlying_1_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks4underlying_1.xlsx")
underlying_2_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_ks4underlying_entriesandgrades_2.xlsx")
school_info_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2022-2023/england_school_information.csv")


keywords = c("South Yorkshire", "West Yorkshire")

school_filtered_2022_2023 = school_info_2022_2023 %>%
  filter(
    toupper(str_trim(TOWN)) %in% toupper(keywords) |
    toupper(str_trim(ADDRESS3)) %in% toupper(keywords)
  )

provisional_joined_2022_2023 = school_filtered_2022_2023 %>%
  inner_join(provisional_2022_2023, by = c("URN", "ESTAB"))

pupdest_joined_2022_2023 = provisional_joined_2022_2023 %>%
  left_join(pupdest_2022_2023, by = c("URN", "ESTAB"))

final_joined_2022_2023 = pupdest_joined_2022_2023 %>%
  left_join(ks4_final_2022_2023, by = c("URN", "ESTAB"))

cleaned_2022_2023 = final_joined_2022_2023 %>%
  select(
    URN,
    SCHNAME = SCHNAME.x,
    TOWN,
    ADDRESS3,
    LANAME,
    POSTCODE,
    SCHOOLTYPE,
    GENDER,
    AGELOW,
    AGEHIGH,
    P8MEA = P8MEA.x,
    P8MEA_FSM6CLA1A = P8MEA_FSM6CLA1A.x,
    P8MEA_NFSM6CLA1A = P8MEA_NFSM6CLA1A.x,
    EBACCAPS = EBACCAPS.x,
    PTEBACC_95 = PTEBACC_95.x,
    OVERALL_DESTPER,
    EMPLOYMENTPER,
    EDUCATIONPER
  ) %>%
  distinct()

write.csv(cleaned_2022_2023, "C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2022-2023.csv")
```

Figure 9 :School 2023-2024

```r
library(tidyverse)

library(readxl)

ks2_final_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks2final.csv")

ks2_mats_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks2-mats-performance.csv")

ks4_final_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks4final.csv")

ks4_mats_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks4-mats-performance.csv")

provisional_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks4provisional.csv")

pupdest_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks4-pupdest.csv")

underlying_1_2023_2024 = read_excel("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks4underlying_1.xlsx")

underlying_2_2023_2024 = read_excel("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_ks4underlying_entriesandgrades_2.xlsx")

school_info_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Schooldata/2023-2024/england_school_information.csv")



keywords = c("South Yorkshire", "West Yorkshire")

school_filtered_2023_2024 = school_info_2023_2024 %>%
  filter(
    toupper(str_trim(TOWN)) %in% toupper(keywords) |
      toupper(str_trim(ADDRESS3)) %in% toupper(keywords)
  )

provisional_joined_2023_2024 = school_filtered_2023_2024 %>%
  inner_join(provisional_2023_2024, by = c("URN", "ESTAB"))

pupdest_joined_2023_2024 = provisional_joined_2023_2024 %>%
  left_join(pupdest_2023_2024, by = c("URN", "ESTAB"))

final_joined_2023_2024 = pupdest_joined_2023_2024 %>%
  left_join(ks4_final_2023_2024, by = c("URN", "ESTAB"))

cleaned_2023_2024 = final_joined_2023_2024 %>%
  select(
    URN,
    SCHNAME = SCHNAME.x,
    TOWN,
    ADDRESS3,
    LANAME,
    POSTCODE,
    SCHOOLTYPE,
    GENDER,
    AGELOW,
```

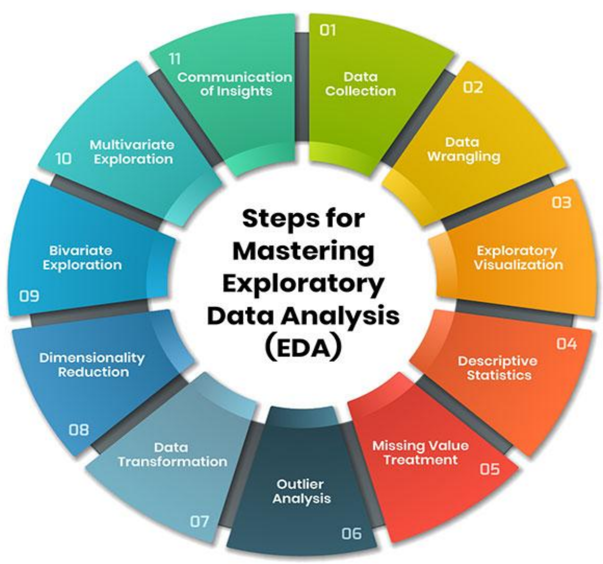
```r
    AGELOW,
    AGEHIGH,
    P8MEA = P8MEA.x,
    P8MEA_FSM6CLA1A = P8MEA_FSM6CLA1A.x,
    P8MEA_NFSM6CLA1A = P8MEA_NFSM6CLA1A.x,
    EBACCAPS = EBACCAPS.x,
    PTEBACC_95 = PTEBACC_95.x,
    OVERALL_DESTPER,
    EMPLOYMENTPER,
    EDUCATIONPER
  ) %>%
  distinct()

write.csv(cleaned_2023_2024, "C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2023-2024.csv")

view(cleaned_2023_2024)
```

Exploratory Data Analysis (EDA)

During the exploratory data analysis phase, I searched the cleaned datasets for home prices, broadband speeds, crime rates, and academic achievement in South and West Yorkshire for patterns, trends, and connections. I used visual aids like boxplots, line charts, bar graphs, and radar charts to understand how each variable behaves across different years, towns, and districts. For example, I compared average house prices over time, looked at crime rates by type and location, compared broadband speeds across towns, and monitored changes in school performance. This step helped me identify important insights, like areas with high crime rates or low school scores and guided the development of the linear model and recommendation system.

Exploratory data analysis (EDA) is an open-ended, iterative data analysis approach designed to unearth patterns, anomalies, relationships, or insights without preconceived notions. John Tukey, a renowned American mathematician, introduced EDA in the 1970s to analyze data using a combination of statistical tools and data discovery methods.
https://www.coursera.org/articles/exploratory-data-analysis

Figure 10:Exploratory Data Analysis



(*A Comprehensive Guide to Mastering Exploratory Data Analysis*, **n.d.-b)**

Broadband speed

I started by connecting broadband speeds to counties and districts by combining the cleaned broadband dataset with town-level data using standardized short postcodes. Next, I contrasted South Yorkshire's and West Yorkshire's broadband performance. I highlighted regions with consistently high or low performance by using boxplots to illustrate the variation in average download speeds across districts. To highlight regional variations, median download speeds by town and district were also compared using bar charts. By highlighting towns with the best and worst connectivity, these visualizations helped identify areas that might benefit from upgrades to their digital infrastructure.

Figure 11 :Analysis broadband

```r
library(tidyverse)


Broadband_speed <- read_csv("C:/DataScience-R/AayushShrestha-230293/Obtain/Broadband/201809_fixed_pc_coverage_r01.csv")


Towns <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv")
# Clean and standardize postcode formats
Broadband_speed <- Broadband_speed %>%
  mutate(shortPostcode = str_trim(toupper(shortPostcode)))

Towns <- Towns %>%
  mutate(shortPostcode = str_trim(toupper(shortPostcode)))

# Merge datasets on shortPostcode
BroadbandMerged <- Broadband_speed %>%
  left_join(Towns, by = "shortPostcode")

BroadbandMerged %>%
  filter(str_detect(tolower(County), "west yorkshire"),
         !is.na(Avg_Download),
         !is.na(District)) %>%
  ggplot(aes(x = reorder(District, Avg_Download, FUN = median), y = Avg_Download)) +
  geom_boxplot(fill = "brown") +
  labs(title = "West Yorkshire: Download Speed by District",
       x = "District", y = "Avg Download Speed (Mbps)") +
  coord_flip() +
  theme_minimal()

BroadbandMerged %>%
  filter(str_detect(tolower(County), "south yorkshire"),
         !is.na(Avg_Download),
         !is.na(District)) %>%
  ggplot(aes(x = reorder(District, Avg_Download, FUN = median), y = Avg_Download)) +
  geom_boxplot(fill = "purple") +
  labs(title = "South Yorkshire: Download Speed by District",
       x = "District", y = "Avg Download Speed (Mbps)") +
  coord_flip() +
  theme_minimal()


library(tidyverse)

# Load data
Broadband_speed <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_BroadBand_Speed.csv")
Towns <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv")

# Clean postcode for merging
Broadband_speed <- Broadband_speed %>%
  mutate(shortPostcode = str_trim(toupper(shortPostcode)))

Towns <- Towns %>%
  mutate(shortPostcode = str_trim(toupper(shortPostcode)))
```

```r

# Merge datasets
BroadbandMerged <- Broadband_speed %>%
  left_join(Towns, by = "shortPostcode")




colnames(BroadbandMerged)
BroadbandMerged %>%
  filter(
    str_detect(tolower(County), "west yorkshire"),
    !is.na(Avg_Download),
    !is.na(District)
  ) %>%
  ggplot(aes(x = reorder(District, Avg_Download), y = Avg_Download)) +
  geom_col(fill = "darkblue") +
  labs(
    title = "West Yorkshire: Avg Download Speed by District",
    x = "District",
    y = "Avg Download Speed (Mbps)"
  ) +
  scale_y_continuous(labels = scales::label_number()) +
  coord_flip() +
  theme_minimal()

BroadbandMerged %>%
  filter(
    str_detect(tolower(County), "south yorkshire"),
    !is.na(Avg_Download),
    !is.na(District)
  ) %>%
  ggplot(aes(x = reorder(District, Avg_Download), y = Avg_Download)) +
  geom_col(fill = "brown") +
  labs(
    title = "South Yorkshire: Avg Download Speed by District",
    x = "District",
    y = "Avg Download Speed (Mbps)"
  ) +
  coord_flip() +   # Flip for better readability
  theme_minimal()
```

Visualization

Figure 12 : Bar Graph South Yorkshire average download speed

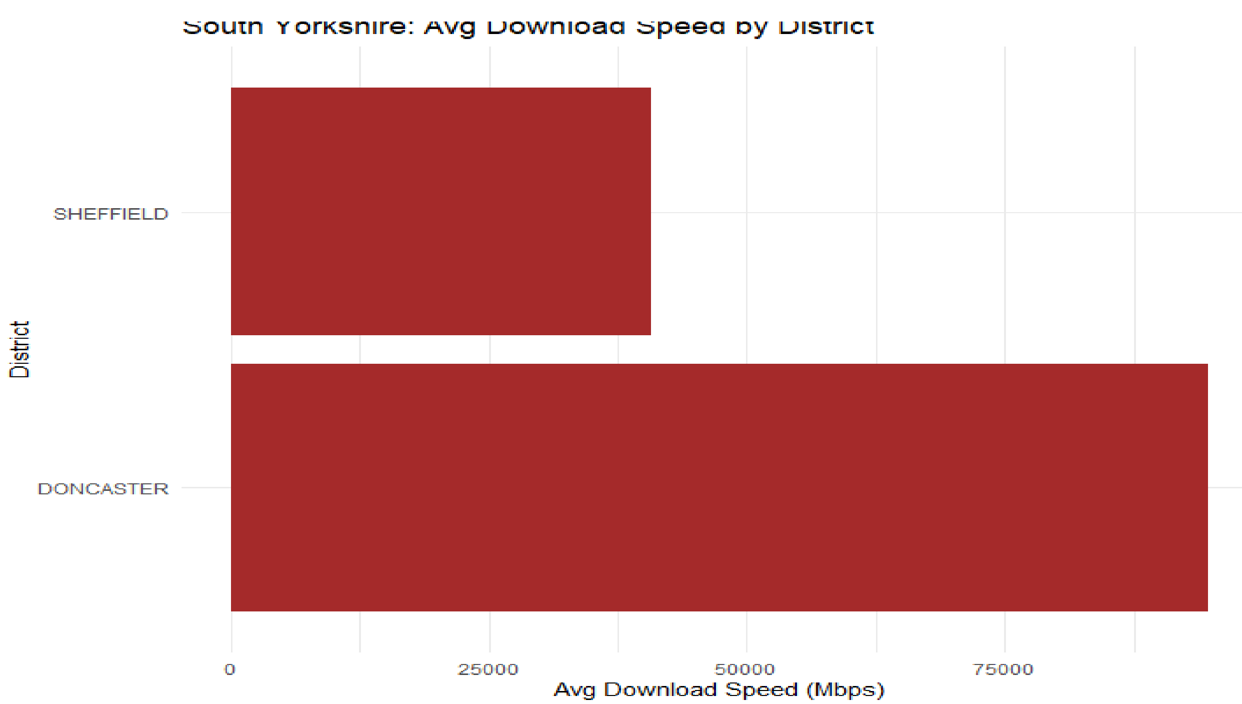


Figure 13: Bar Graph West Yorkshire Avg download speed

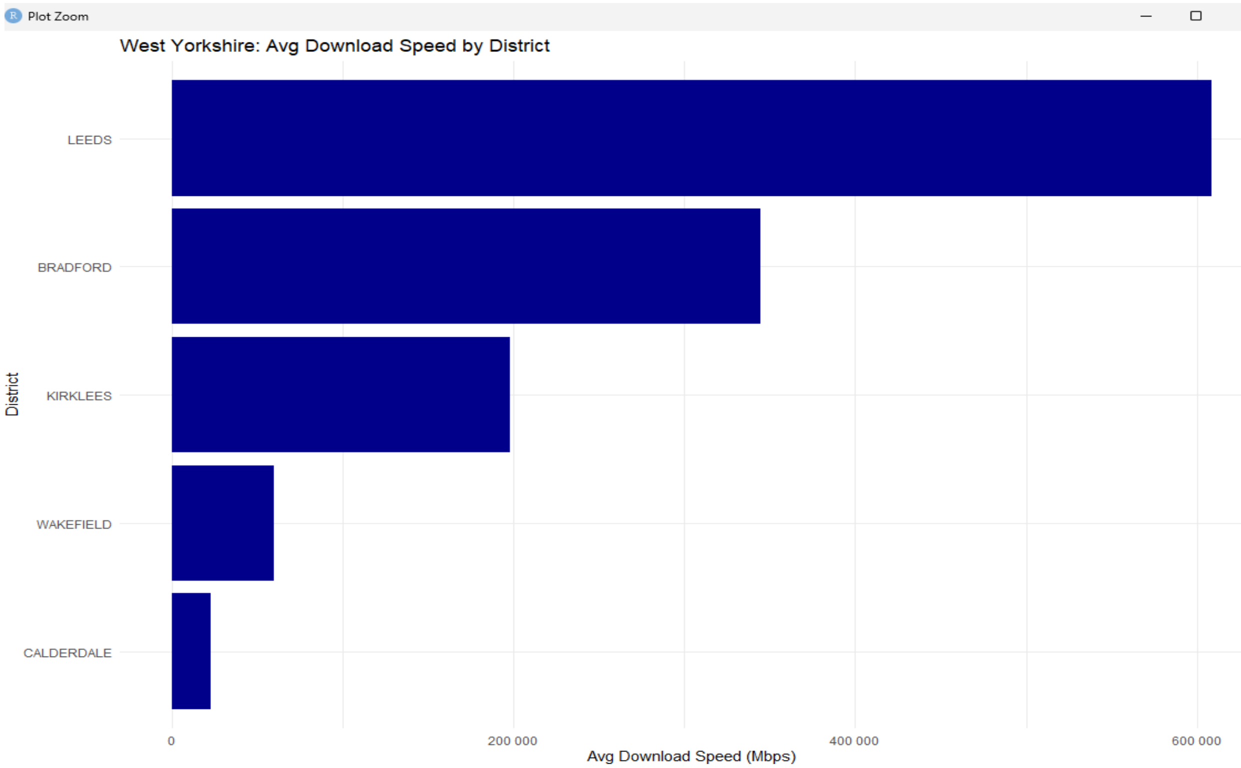


Figure 14: Box plot West Yorkshire

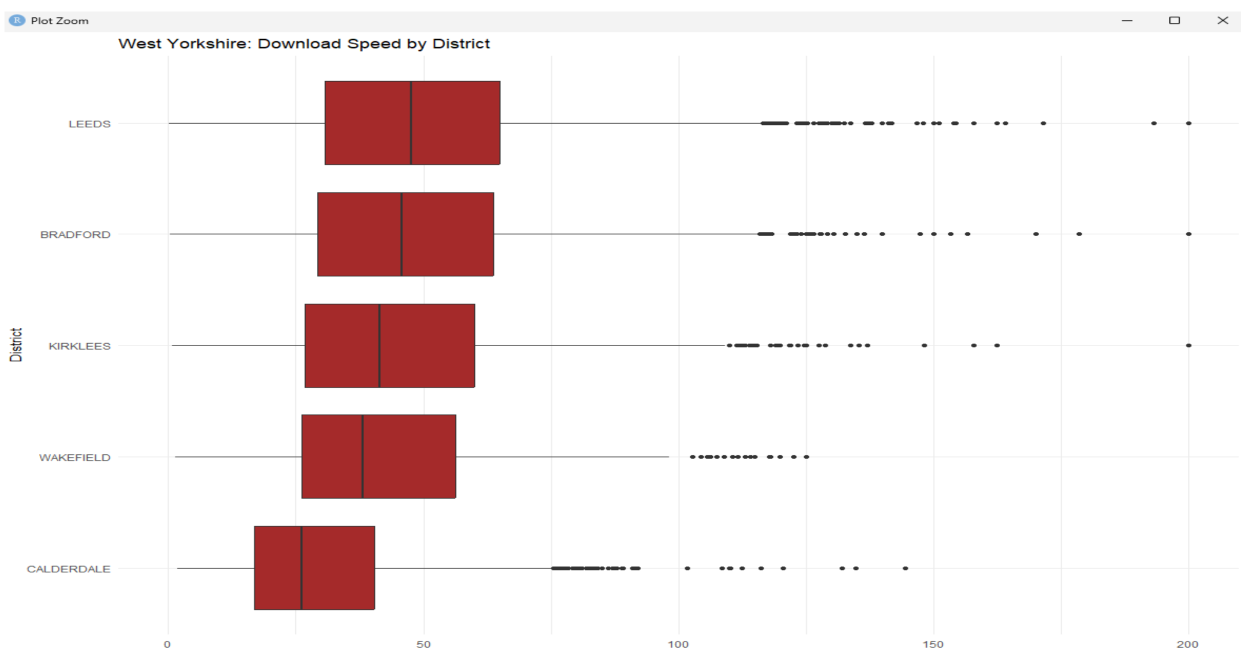
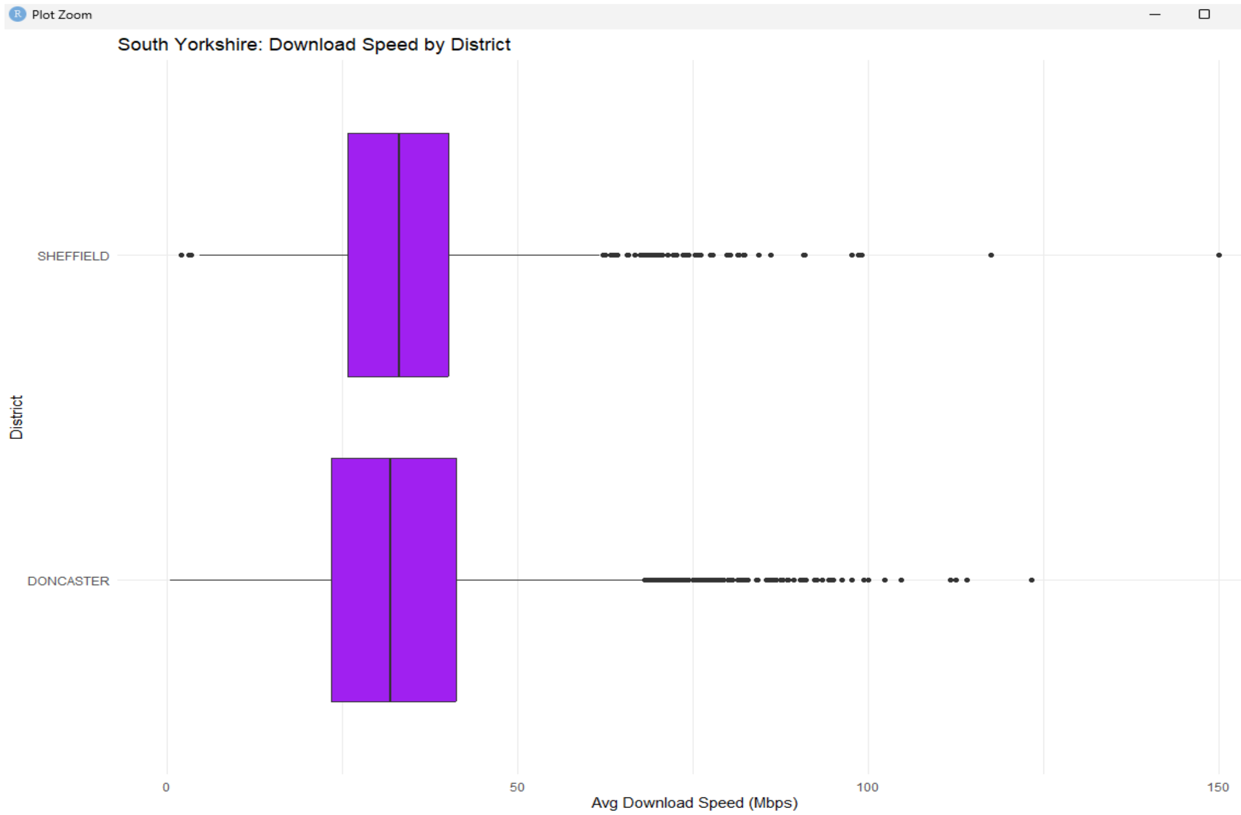


Figure 15 :Box plot South Yorkshire



Crime Data Set

The analysis of crime rates in South and West Yorkshire from 2022 to 2025 concentrated on robberies, car crimes, and drug offenses. The distribution of drug offense rates by district was shown using boxplots, which highlighted regions with greater variances and higher crime rates. West Yorkshire's April 2025 vehicle crime rates by district were displayed on a radar-style polar bar chart, which made it simple to compare the severity of crimes in different places. Similarly, a pie chart showed the percentage of robbery offenses in South Yorkshire by district, showing the regions with the highest frequency of occurrences. To examine changes over time and standardize data by population for fair comparison, a line chart was also used to show trends in drug offense rates per 10,000 persons over three years.

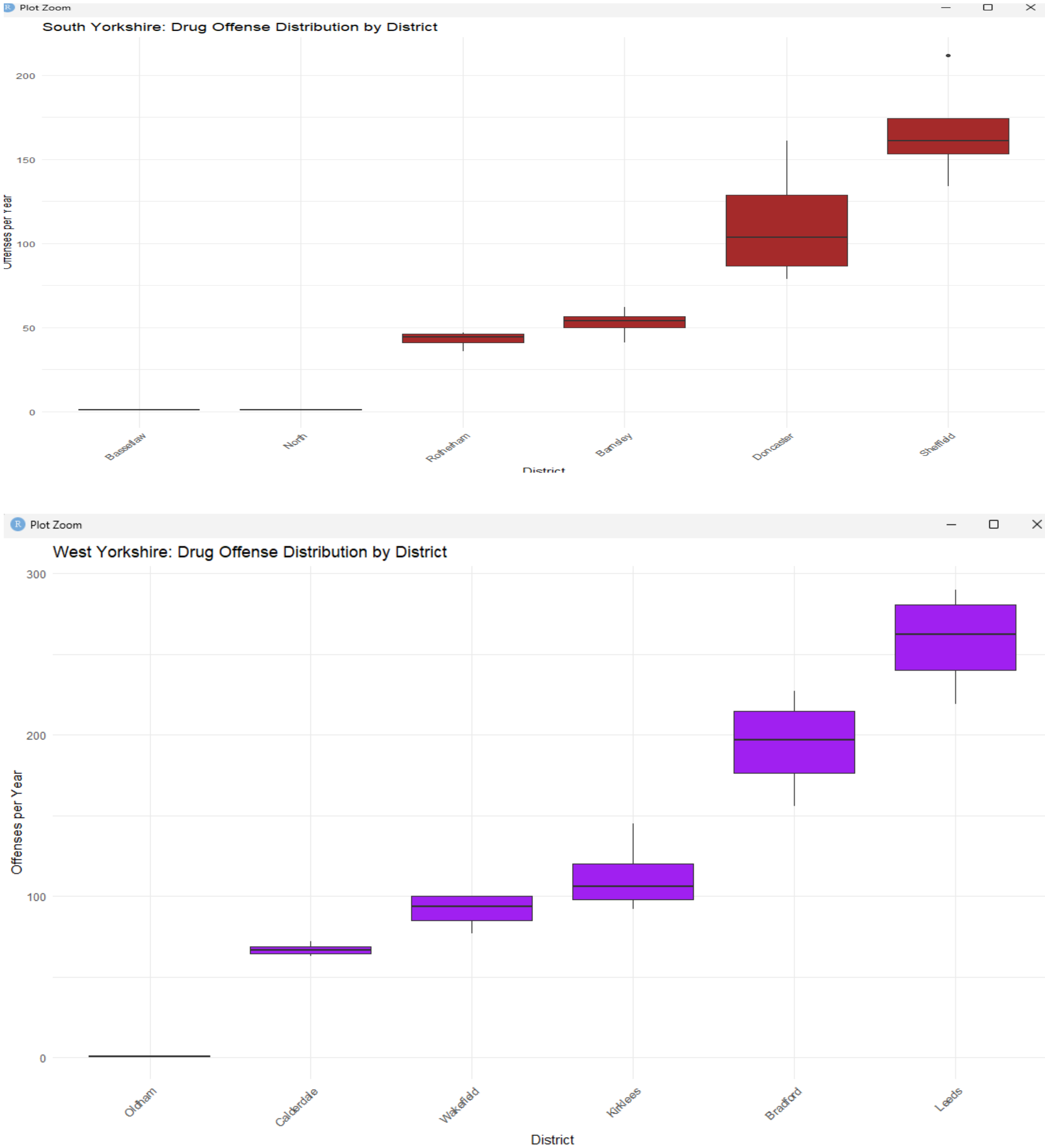
Figure 16 :Analysis Crime

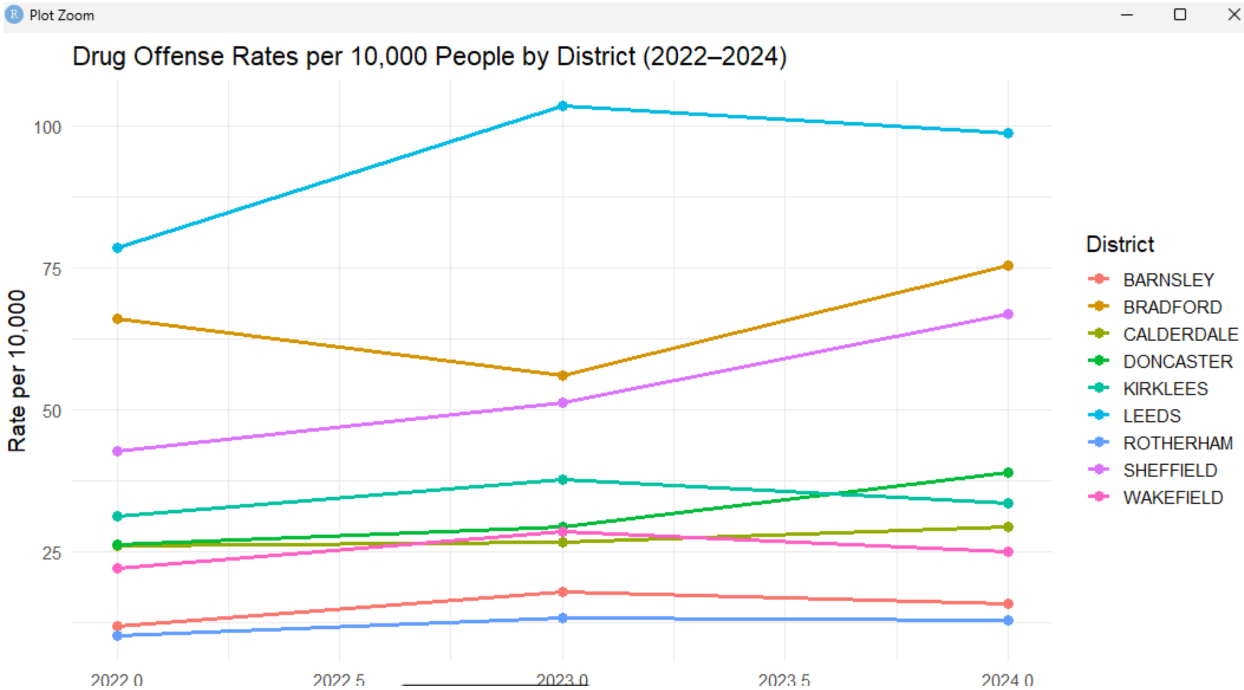
```r

library(tidyverse)

library(ggplot2)

crime = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_Crime_Dataset.csv")
colnames(crime)
view(crime)
town = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv")
colnames(town)
view(town)

crime = crime %>%
  mutate(Year = as.integer(substr(Month, 1, 4)),
         District = str_extract(LSOAname, "^[^ ]+"))

crime = crime %>%
  mutate(County = str_replace(County, " Police$", ""))



#Box plot- Drug Offense Rate per District (Two Diagrams)


# South Yorkshire Drug Offenses by District-Year
south_yorkshire = crime %>%
  filter(CrimeType == "Drugs", County == "South Yorkshire", !is.na(District)) %>%
  group_by(District, Year) %>%
  summarise(Offenses = n(), .groups = "drop")

# Boxplot for South Yorkshire
ggplot(south_yorkshire, aes(x = reorder(District, Offenses, FUN = median), y = Offenses)) +
  geom_boxplot(fill = "brown", outlier.alpha = 0.9) +
  labs(title = "South Yorkshire: Drug Offense Distribution by District",
       x = "District", y = "Offenses per Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))




# West Yorkshire Drug Offenses by District-Year
west_yorkshire = crime %>%
  filter(CrimeType == "Drugs", County == "West Yorkshire", !is.na(District)) %>%
  group_by(District, Year) %>%
  summarise(Offenses = n(), .groups = "drop")

# Boxplot for West Yorkshire
ggplot(west_yorkshire, aes(x = reorder(District, Offenses, FUN = median), y = Offenses)) +
  geom_boxplot(fill = "purple", outlier.alpha = 0.9) +
  labs(title = "West Yorkshire: Drug Offense Distribution by District",
       x = "District", y = "Offenses per Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

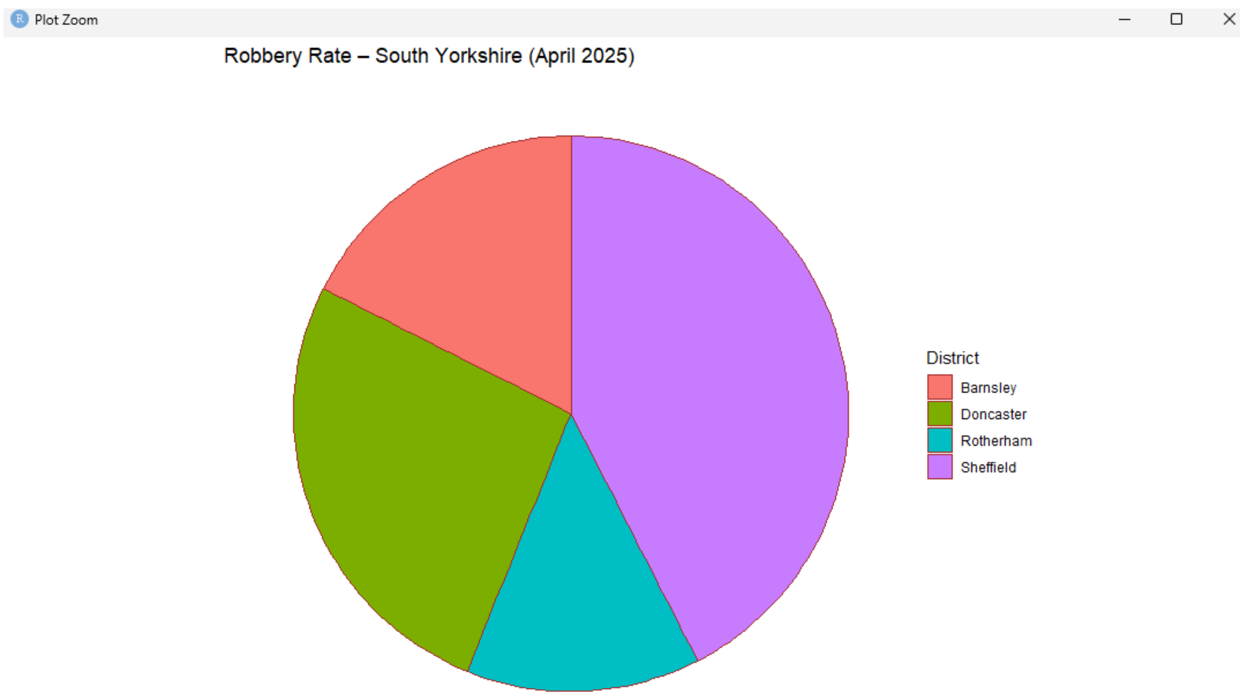
```r

vehicle_data = crime %>%
  filter(CrimeType == "Vehicle crime",
         County == "West Yorkshire",
         Month == "2025-04",
         !is.na(District)) %>%
  group_by(District) %>%
  summarise(Crimes = n()) %>%
  arrange(desc(Crimes))

# Create radar-style polar bar chart
ggplot(vehicle_data, aes(x = reorder(District, Crimes), y = Crimes, fill = District)) +
  geom_col(show.legend = FALSE, color = "blue") +
  coord_polar(start = 0) +
  labs(title = "Radar-Style Chart: Vehicle Crime in West Yorkshire (April 2025)",
       x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 9, angle = 90))





#Pie chart for Robbery rate for any one of two counties (for any specific month and year)

pie_data = crime %>%
  filter(CrimeType == "Robbery",
         County == "South Yorkshire",
         Month == "2025-04",
         !is.na(District)) %>%
  group_by(District) %>%
  summarise(Crimes = n())

ggplot(pie_data, aes(x = "", y = Crimes, fill = District)) +
  geom_col(width = 3, color = "brown") +
  coord_polar("y") +
  labs(title = "Robbery Rate - South Yorkshire (April 2025)", y = "", x = "") +
  theme_void() +
  theme(legend.position = "right")







#Line chart for Drug offense rates per 10,000 people for both counties in same diagram for all years

town_long = town %>%
  pivot_longer(
    cols = starts_with("Population"),
    names_to = "Year",
    names_prefix = "Population",
    values_to = "Population"
  ) %>%
  mutate(Year = as.integer(Year)) %>%
  select(District, County, Year, Population)

drug_crime_by_district = crime %>%
  filter(tolower(CrimeType) == "drugs") %>%
  group_by(District, Year) %>%
  summarise(Total_Offenses = n(), .groups = "drop") %>%
  mutate(Year = as.integer(Year))

drug_crime_by_district = drug_crime_by_district %>%
  mutate(District = str_to_upper(str_trim(District)))

town_long = town_long %>%
  mutate(District = str_to_upper(str_trim(District)))


merged_data = drug_crime_by_district %>%
  left_join(town_long, by = c("District", "Year")) %>%
  filter(!is.na(Population)) %>%
  mutate(Rate_per_10000 = (Total_Offenses / Population) * 10000)

final_data = merged_data %>%
  group_by(District, Year) %>%
  summarise(
    Total_Offenses = sum(Total_Offenses, na.rm = TRUE),
    Population = sum(Population, na.rm = TRUE),
    Rate_per_10000 = (Total_Offenses / Population) * 10000,
    .groups = "drop"
  )


ggplot(final_data, aes(x = Year, y = Rate_per_10000, color = District)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 3) +
  labs(
    title = "Drug Offense Rates per 10,000 People by District (2022-2024)",
    x = "Year",
    y = "Rate per 10,000",
    color = "District"
  ) +
  theme_minimal(base_size = 14)

```

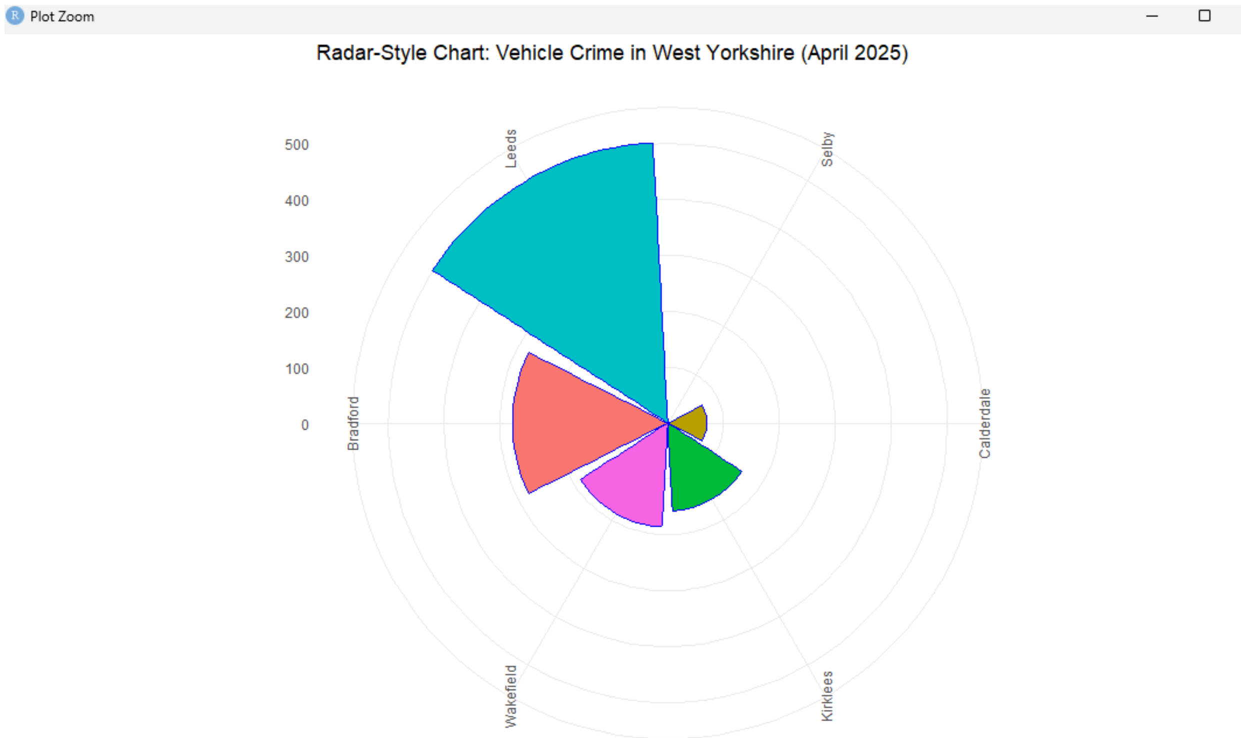
Visualization

Figure 17 : Visualization crime




Plot Zoom
Drug Offense Rates per 10,000 People by District (2022–2024)
Rate per 10,000
100
75
50
25
2022.0
2022.5
2023.0
2023.5
2024.0
District
BARNSLEY
BRADFORD
CALDERDALE
DONCASTER
KIRKLEES
LEEDS
ROTHERHAM
SHEFFIELD
WAKEFIELD


Plot Zoom
Robbery Rate – South Yorkshire (April 2025)
District
Barnsley
Doncaster
Rotherham
Sheffield



House pricing

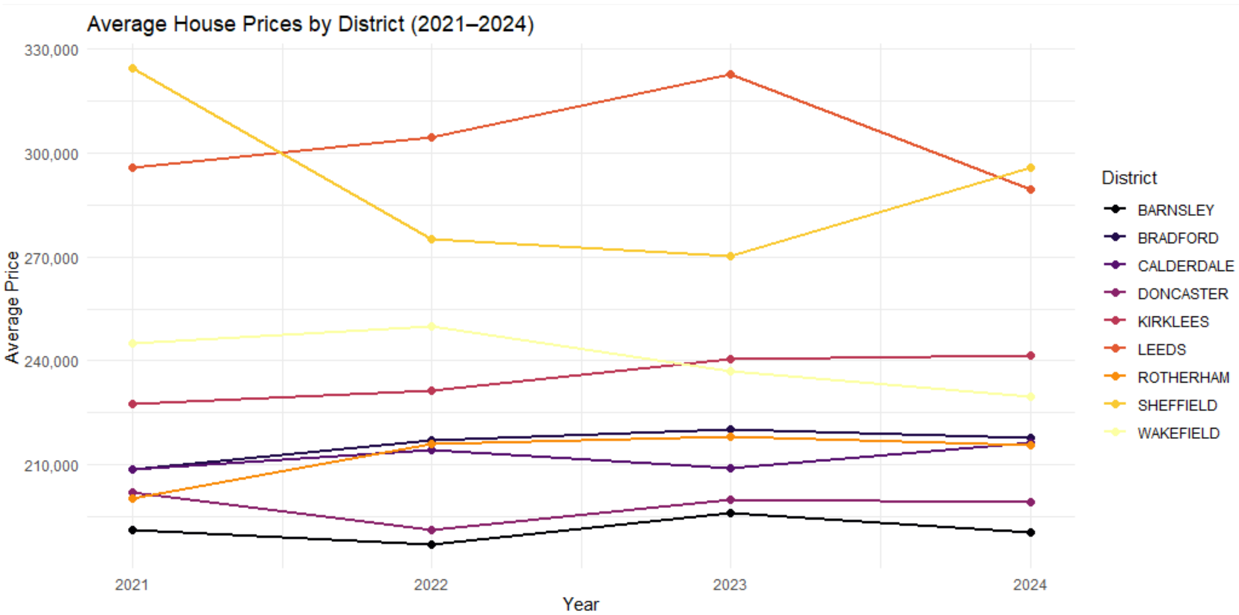
The analysis of home prices in South and West Yorkshire from 2021 to 2024 revealed trends and differences at the county and district levels. A line graph showing the annual average home prices for each district allowed comparison of the price changes over time in both counties. The districts that grew steadily and those that remained largely unchanged were displayed in this graphic. Additionally, boxplots were used to display the home price distribution for each county independently. By providing information on price spread, outliers, and median values, these boxplots made it easier to spot variability within counties. Overall, understanding regional and temporal differences in home price trends was made easier by the visualizations.

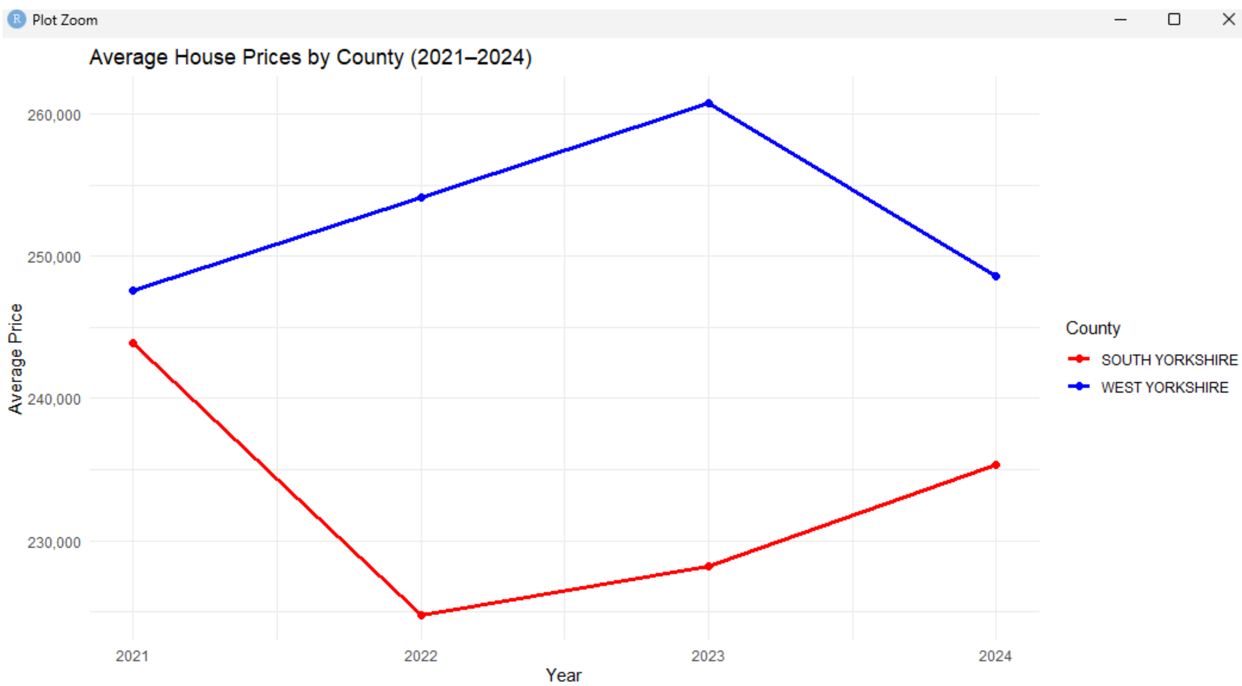
Figure 18: Analysis of House pricing

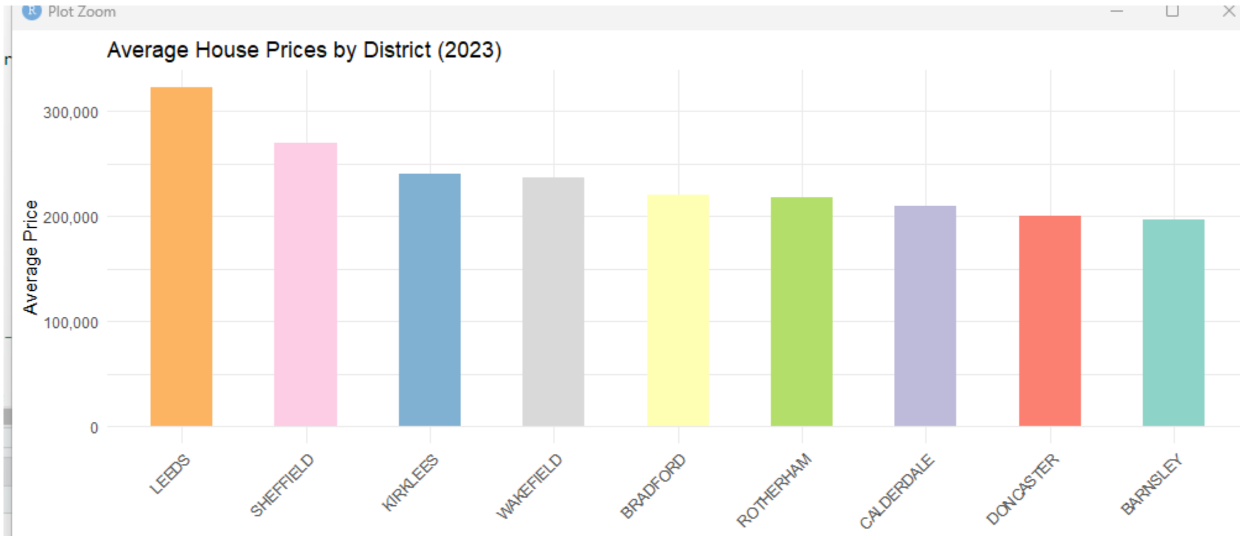
```r
average by County in 2023 only
county_2023_avg = county_year_avg %>%
  filter(Year == 2023)

#Bar chart: Average price in 2023, both counties

ggplot(county_2023_avg, aes(x = County, y = AveragePrice, fill = County)) +
  geom_col(width = 0.3) +
  scale_fill_viridis_d(option = "C") +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Average House Prices by Country (2023)",
       x = NULL, y = "Average Price") +
  theme_minimal() +
  theme(legend.position = "none")




#by district
district_2023_avg = HousePrices %>%
  mutate(Year = year(ymd(Date))) %>%
  filter(Year == 2023) %>%
  group_by(District) %>%
  summarise(AveragePrice = mean(Price, na.rm = TRUE), .groups = "drop")

Bar chart: Average house price by District in 2023
library(RColorBrewer)

ggplot(district_2023_avg, aes(x = reorder(District, -AveragePrice), y = AveragePrice, fill = District)) +
  geom_col(width = 0.5) +
  scale_fill_brewer(palette = "Set3") +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Average House Prices by District (2023)",
       x = "District", y = "Average Price") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))





#-----------------------------------------------------------------------------------------------------------------------

#by county
#Box-plots: 2021-2024 distribution for each county in separate panels

ggplot(county_year_avg, aes(x = "", y = AveragePrice)) +
  geom_boxplot(fill = "brown", outlier.alpha = 0.2) +
  facet_wrap(~ County, ncol = 1) +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Distribution of House Prices (2021-2024)",
       y = "Price", x = "") +
```

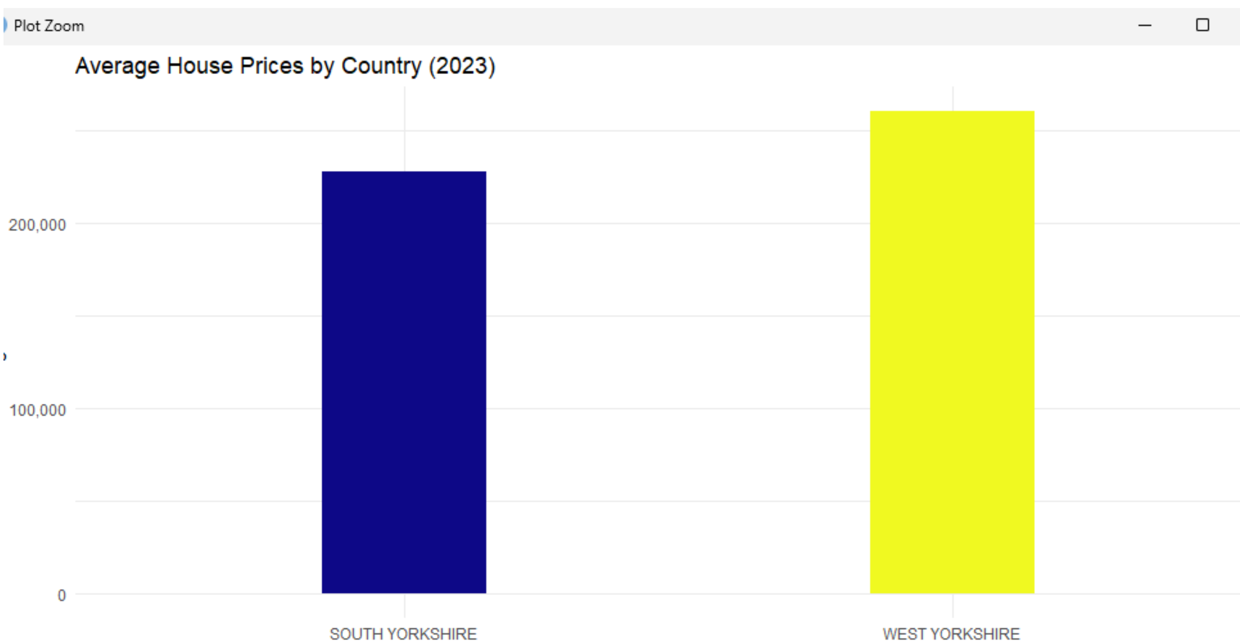
```r
       y = "Price", x = "") +
  theme_minimal()



#District

ggplot(district_year_avg, aes(x = reorder(District, AveragePrice, FUN = median), y = AveragePrice)) +
  geom_boxplot(fill = "skyblue", outlier.alpha = 0.9) +
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Yearly Average House Prices by District (2021-2024)",
       x = "District", y = "Average Price") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 19 :Visualization house pricing









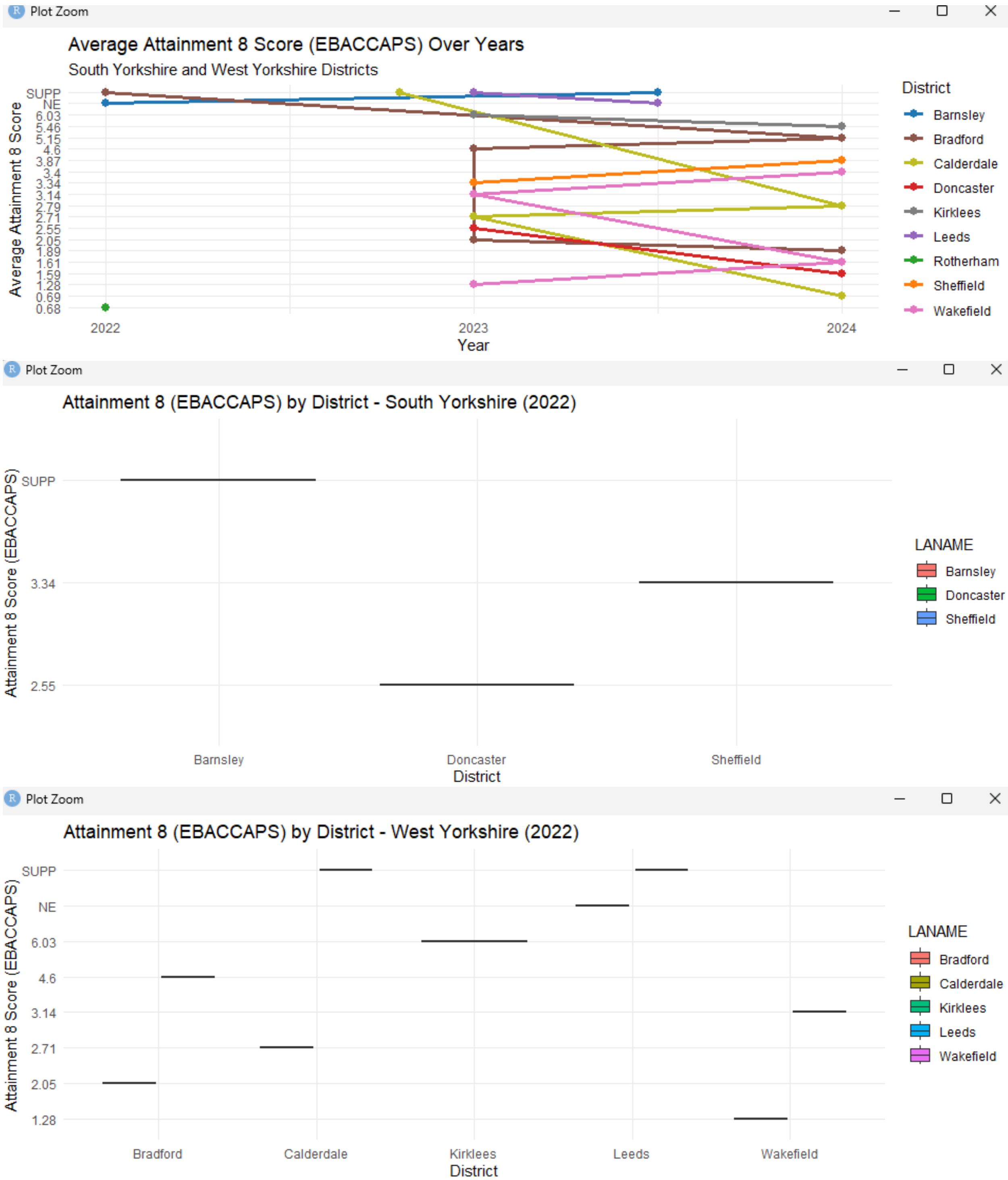
School Data set

We looked at Attainment 8 scores (EBACCAPS) for schools in South Yorkshire and West Yorkshire during the 2021–2024 academic years in this exploratory data analysis. To visually represent the distribution of scores by district within each county, we first created boxplots for 2022. This made it easier to spot differences and possible anomalies in student performance between Sheffield, Barnsley, Leeds, and Bradford districts. After that, we created a line graph to show trends in average Attainment 8 scores over time across several districts by combining school performance data from three different years. We were able to compare educational outcomes across districts and counties thanks to this multi-year view, which showed how performance has changed year over year. The analysis reveals temporal and spatial trends in academic achievement, offering insightful information to educators, policymakers.

Figure 20 :Analysis school

```r
library(tidyverse)
library(stringr)

school_2021_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2021-2022.csv")
school_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2022-2023.csv")
school_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2023-2024.csv")


#Boxplot for Average attainment 8 score 2022 - South Yorkshire (variable District and Score)
filtered_data_south = school_2022_2023 %>%
  filter(
    toupper(TOWN) == "SOUTH YORKSHIRE" |
      toupper(ADDRESS3) == "SOUTH YORKSHIRE"
  ) %>%
  filter(!is.na(EBACCAPS))

ggplot(filtered_data_south, aes(x = LANAME, y = EBACCAPS, fill = LANAME)) +
  geom_boxplot() +
  labs(
    title = "Attainment 8 (EBACCAPS) by District - South Yorkshire (2022)",
    x = "District",
    y = "Attainment 8 Score (EBACCAPS)"
  ) +
  theme_minimal()







#Boxplot for Average attainment 8 score 2022 - West Yorkshire (variable District and Score)
filtered_data_west = school_2022_2023 %>%
  filter(
    toupper(TOWN) == "WEST YORKSHIRE" |
      toupper(ADDRESS3) == "WEST YORKSHIRE"
  ) %>%
  filter(!is.na(EBACCAPS))

ggplot(filtered_data_west, aes(x = LANAME, y = EBACCAPS, fill = LANAME)) +
  geom_boxplot() +
  labs(
    title = "Attainment 8 (EBACCAPS) by District - West Yorkshire (2022)",
    x = "District",
    y = "Attainment 8 Score (EBACCAPS)"
  ) +
  theme_minimal()



school_2021_2022 = school_2021_2022 %>%
  mutate(Year = 2022L)

school_2022_2023 = school_2022_2023 %>%
  mutate(Year = 2023L)

school_2023_2024 = school_2023_2024 %>%
  mutate(Year = 2024L)


# Binding all years
all_years_data = bind_rows(school_2021_2022, school_2022_2023, school_2023_2024)


all_years_data = all_years_data %>%
  mutate(Year = as.integer(Year))


filtered_all_years = all_years_data %>%
  filter(
    toupper(TOWN) %in% c("SOUTH YORKSHIRE", "WEST YORKSHIRE") |
      toupper(ADDRESS3) %in% c("SOUTH YORKSHIRE", "WEST YORKSHIRE")
  ) %>%
  filter(!is.na(EBACCAPS))


ggplot(filtered_all_years, aes(x = Year, y = EBACCAPS, colour = LANAME, group = LANAME)) +
  stat_summary(fun = mean, geom = "line", linewidth = 1.2) +
  stat_summary(fun = mean, geom = "point", size = 2.5) +
  labs(
    title = "Average Attainment 8 Score (EBACCAPS) Over Years",
    subtitle = "South Yorkshire and West Yorkshire Districts",
    x = "Year",
    y = "Average Attainment 8 Score",
    colour = "District"
  ) +
  scale_x_continuous(breaks = c(2022, 2023, 2024)) +
  scale_colour_manual(values = c(
    "Barnsley" = "#1f77b4",
    "Sheffield" = "#ff7f0e",
    "Rotherham" = "#2ca02c",
    "Doncaster" = "#d62728",
    "Leeds" = "#9467bd",
    "Bradford" = "#8c564b",
    "Wakefield" = "#e377c2",
    "Kirklees" = "#7f7f7f",
    "Calderdale" = "#bcbd22"
  )) +
  theme_minimal()
```

Figure 21: Visualization School



Linear Modelling

1. House Price vs Download Speed for both Counties in a single diagram (include linear model summary report and correlation):

   By combining postcode datasets, this code examines the connection between broadband download speeds and home prices across districts. To forecast home prices based on download speed, it fits a linear regression model, displays residuals to evaluate fit, and displays the data as a scatter plot and regression line. To determine whether faster internet is associated with higher property values, the model summary and correlation coefficient quantify the strength of the relationship between download speed and home prices.

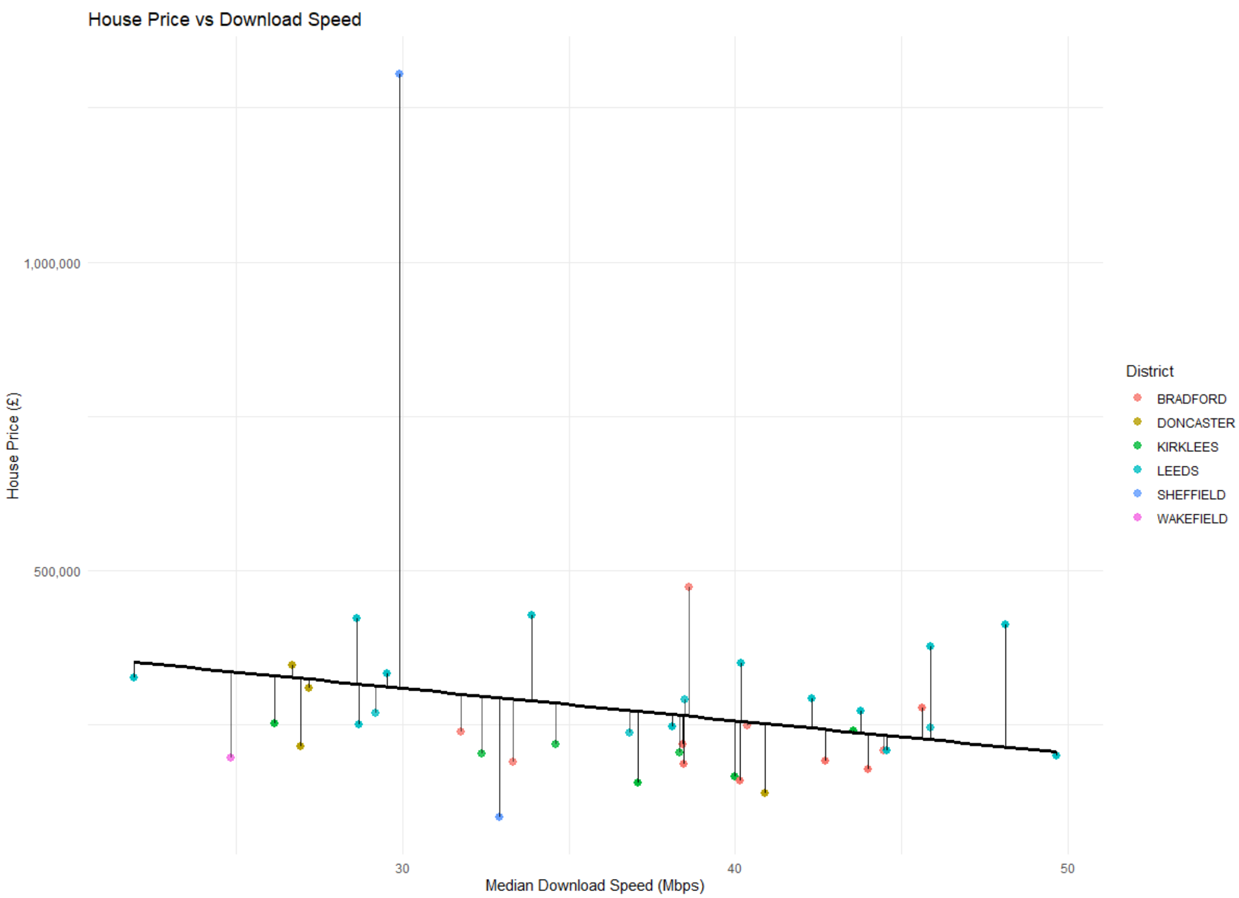
   Code:

Figure 22 : Code house price vs download speed

```r
#----------------------------------House Price vs Download Speed for both Counties in single diagram (include linear model summary report and correlation)--
library(tidyverse)
library(scales)

HousePrices = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_House_Prices.csv") %>%
  select(shortPostcode, Price)

BroadBandSpeed = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_BroadBand_Speed.csv") %>%
  select(shortPostcode, Median_Download)

Town <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv") %>%
  select(shortPostcode, District, County)



# 1. Aggregate or deduplicate each dataset by postcode
HousePrices_unique <- HousePrices %>%
  group_by(shortPostcode) %>%
  summarise(Price = mean(Price, na.rm = TRUE), .groups = "drop")

BroadBandSpeed_unique <- BroadBandSpeed %>%
  group_by(shortPostcode) %>%
  summarise(Median_Download = mean(Median_Download, na.rm = TRUE), .groups = "drop")

Town_unique <- Town %>%
  distinct(shortPostcode, .keep_all = TRUE)

# 2. Join cleaned datasets
CombinedData <- HousePrices_unique %>%
  inner_join(BroadBandSpeed_unique, by = "shortPostcode") %>%
  inner_join(Town_unique, by = "shortPostcode") %>%
  filter(!is.na(Price) & !is.na(Median_Download) & !is.na(District))

# 3. Sample 100 rows from CombinedData
SampleData <- CombinedData %>%
  sample_n(100, replace = TRUE)



SampleModel = lm(Price ~ Median_Download, data = SampleData)


SampleData = SampleData %>%
  mutate(
    Predicted = predict(SampleModel),
    Residual = Price - Predicted
  )

class(SampleData$Town)
colnames(SampleData)


ggplot(SampleData, aes(x = Median_Download, y = Price)) +
  geom_point(aes(color = District), size = 2.5, alpha = 0.8) +  # color by District
  geom_smooth(method = "lm", se = FALSE, color = "black", size = 1.2) +  # regression line
```

```
  geom_segment(aes(xend = Median_Download, yend = Predicted), color = "black", alpha = 0.6) +  # residuals
  labs(
    title = "House Price vs Download Speed",
    x = "Median Download Speed (Mbps)",
    y = "House Price (£)",
    colour = "District"
  ) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()



FullModel = lm(Price ~ Median_Download, data = CombinedData)


correlation = cor(CombinedData$Price, CombinedData$Median_Download, use = "complete.obs")
cat("Correlation between Price and Download Speed:", correlation, "\n")

summary(FullModel)
```

Figure 23 :House price vs download speed



2. House price vs Drug rates (2023) per 10000 people for both counties in a single diagram (include linear model summary report and correlation):

   This study investigates the relationship between drug crime rates and home prices in South and West Yorkshire in 2023. The code determines the drug crime rate per 10,000 people after cleaning and merging data on drug crimes, home prices, and population by county. The effect of the crime rate on home values is then evaluated using a linear regression model. With a trend line and residuals, the resulting plot illustrates the relationship, and the correlation score provides insight into how strong it is. This sheds light on whether lower property values are linked to higher crime rates.

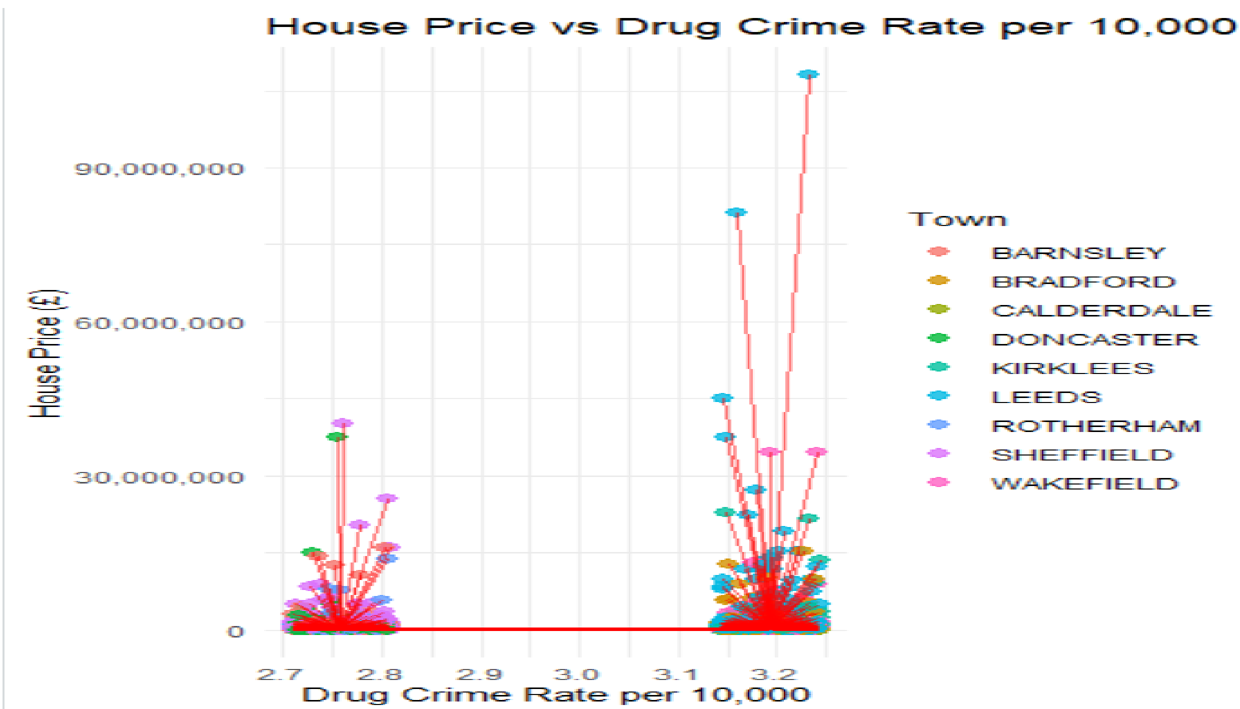
Code:

Figure 24 :Code house price vs drug rates

```r
library(tidyverse)
library(lubridate)
library(scales)
library(stringr)

# House prices
HousePrices <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_House_Prices.csv") %>%
  mutate(
    Year = year(ymd(Date)),
    County = str_trim(str_to_upper(County)),
    shortPostcode = str_trim(str_to_upper(shortPostcode))
  ) %>%
  filter(Year == 2023) %>%
  select(shortPostcode, Price, County)

# Drug crimes
Crime <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_Crime_Dataset.csv") %>%
  mutate(
    Year = as.integer(substr(Month, 1, 4)),
    County = str_replace(County, " Police$", ""),
    County = str_trim(str_to_upper(County))
  ) %>%
  filter(Year == 2023, CrimeType == "Drugs") %>%
  group_by(County) %>%
  summarise(DrugCrimes = n(), .groups = "drop")

# County population
Population <- tibble(
  County = c("SOUTH YORKSHIRE", "WEST YORKSHIRE"),
  Population = c(1417000, 2342000)
)

# Crime rate
CrimeRate <- inner_join(Crime, Population, by = "County") %>%
  mutate(DrugRatePer10k = DrugCrimes / Population * 10000)

Town <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv") %>%
  mutate(
    shortPostcode = str_trim(str_to_upper(shortPostcode)),
    County = str_trim(str_to_upper(County)),
    Town = District  # ✅ Rename District to Town for clarity
  ) %>%
  select(shortPostcode, Town, County)

# Join house prices with towns
HousePrices_Town <- inner_join(HousePrices, Town, by = c("shortPostcode", "County"))

# Join with crime rate
CombinedData <- inner_join(HousePrices_Town, CrimeRate, by = "County") %>%
  filter(!is.na(Price), !is.na(DrugRatePer10k), !is.na(Town))  # ✅ Confirm Town exists

# Add jitter
CombinedData <- CombinedData %>%
  mutate(DrugRatePer10k_jitter = DrugRatePer10k + runif(n(), -0.05, 0.05))
```

```r
# Add jitter
CombinedData <- CombinedData %>%
  mutate(DrugRatePer10k_jitter = DrugRatePer10k + runif(n(), -0.05, 0.05))

# Fit model
FullModel <- lm(Price ~ DrugRatePer10k, data = CombinedData)

# Add predictions
CombinedData <- CombinedData %>%
  mutate(
    Predicted = predict(FullModel),
    Residual = Price - Predicted
  )

# Final plot
ggplot(CombinedData, aes(x = DrugRatePer10k_jitter, y = Price)) +
  geom_point(aes(color = Town), size = 2.5, alpha = 0.8) +
  geom_smooth(aes(x = DrugRatePer10k), method = "lm", se = FALSE, color = "red", size = 1.2) +
  geom_segment(aes(xend = DrugRatePer10k, yend = Predicted), color = "red", alpha = 0.6) +
  labs(
    title = "House Price vs Drug Crime Rate per 10,000 (2023)",
    x = "Drug Crime Rate per 10,000",
    y = "House Price (£)",
    colour = "Town"
  ) +
  scale_y_continuous(labels = comma) +
  theme_minimal() +
  theme(legend.position = "right")



cat("\n--- Linear Model Summary Report ---\n")
print(summary(FullModel))


correlation = cor(CombinedData$Price, CombinedData$DrugRatePer10k, use = "complete.obs")
cat("\n--- Correlation Analysis ---\n")
cat("Correlation between House Price and Drug Crime Rate per 10,000:", correlation, "\n")
```

Figure 25 :House prices vs drug rates



3. Attainment 8 score vs House Price for both counties in a single diagram (include linear model summary report and correlation)

This study investigated the possible correlation between Attainment 8 scores, which measure academic achievement, and home prices in South and West Yorkshire towns. Clean house price data from 2023 was combined with town locations and school performance data from 2021–2024. A linear regression model (lm) was fitted with Attainment 8 as the predictor and House Price as the outcome. A jitter was applied to the scores to reduce point overlap in the scatter plot. The plot shows segments, a regression line, and data points by town to illustrate prediction errors (residuals). The strength of a relationship is gauged by the correlation coefficient. This modeling aids in assessing the potential effects of educational quality on local property values.

Code:

Figure 26 :Attainment 8 score vs House price

```r
library(tidyverse)
library(lubridate)
library(scales)
library(stringr)

HousePrices = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_House_Prices.csv") %>%
  mutate(
    Year = year(ymd(Date)),
    County = str_trim(str_to_upper(County)),
    shortPostcode = str_trim(str_to_upper(shortPostcode))
  ) %>%
  filter(Year == 2023) %>%
  select(shortPostcode, Price, County)


Town <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv") %>%
  mutate(
    shortPostcode = str_trim(str_to_upper(shortPostcode)),
    County = str_trim(str_to_upper(County)),
    Town = str_to_title(District)
  ) %>%
  select(shortPostcode, Town, County)


HousePrices_Town = inner_join(HousePrices, Town, by = c("shortPostcode", "County"))

School_2021_2022 = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2021-2022.csv") %>%
  mutate(Year = 2022L)

School_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2022-2023.csv") %>%
  mutate(Year = 2023L)

School_2023_2024 = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2023-2024.csv") %>%
  mutate(Year = 2024L)

AllSchools = bind_rows(School_2021_2022, School_2022_2023, School_2023_2024) %>%
  mutate(
    County = case_when(
      toupper(TOWN) %in% c("SOUTH YORKSHIRE", "WEST YORKSHIRE") ~ str_to_upper(TOWN),
      toupper(ADDRESS3) %in% c("SOUTH YORKSHIRE", "WEST YORKSHIRE") ~ str_to_upper(ADDRESS3),
      TRUE ~ NA_character_
    ),
    EBACCAPS = as.numeric(str_replace_all(EBACCAPS, "[^0-9.]", ""))  # remove non-numeric chars
  ) %>%
  filter(!is.na(County), !is.na(EBACCAPS)) %>%
  select(County, Year, EBACCAPS)

CombinedData = HousePrices_Town %>%
  inner_join(AllSchools, by = "County", relationship = "many-to-many") %>%
  rename(Attainment8 = EBACCAPS) %>%
  mutate(Attainment8_jitter = Attainment8 + runif(n(), -0.1, 0.1))


FullModel = lm(Price ~ Attainment8, data = CombinedData)
```

```r
FullModel = lm(Price ~ Attainment8, data = CombinedData)

CombinedData = CombinedData %>%
  mutate(
    Predicted = predict(FullModel),
    Residual = Price - Predicted
  )

ggplot(CombinedData, aes(x = Attainment8_jitter, y = Price, color = Town)) +
  geom_point(alpha = 0.8, size = 2.5) +
  geom_smooth(aes(x = Attainment8), method = "lm", se = FALSE, color = "black", linewidth = 1.2) +
  geom_segment(aes(xend = Attainment8, yend = Predicted), color = "black", alpha = 0.5) +
  labs(
    title = "House Price vs Attainment 8 Score",
    x = "Attainment 8 Score",
    y = "House Price (£, 2023)",
    colour = "Town"
  ) +
  scale_y_continuous(labels = comma) +
  theme_minimal() +
  theme(legend.position = "right")

cat("\n--- Linear Model Summary Report ---\n")
print(summary(FullModel))

correlation = cor(CombinedData$Price, CombinedData$Attainment8, use = "complete.obs")
cat("\n--- Correlation Analysis ---\n")
cat("Correlation between House Price and Attainment 8 Score:", correlation, "\n")
```

4. Attainment 8 scores vs Drug Offense rates per 10000 people

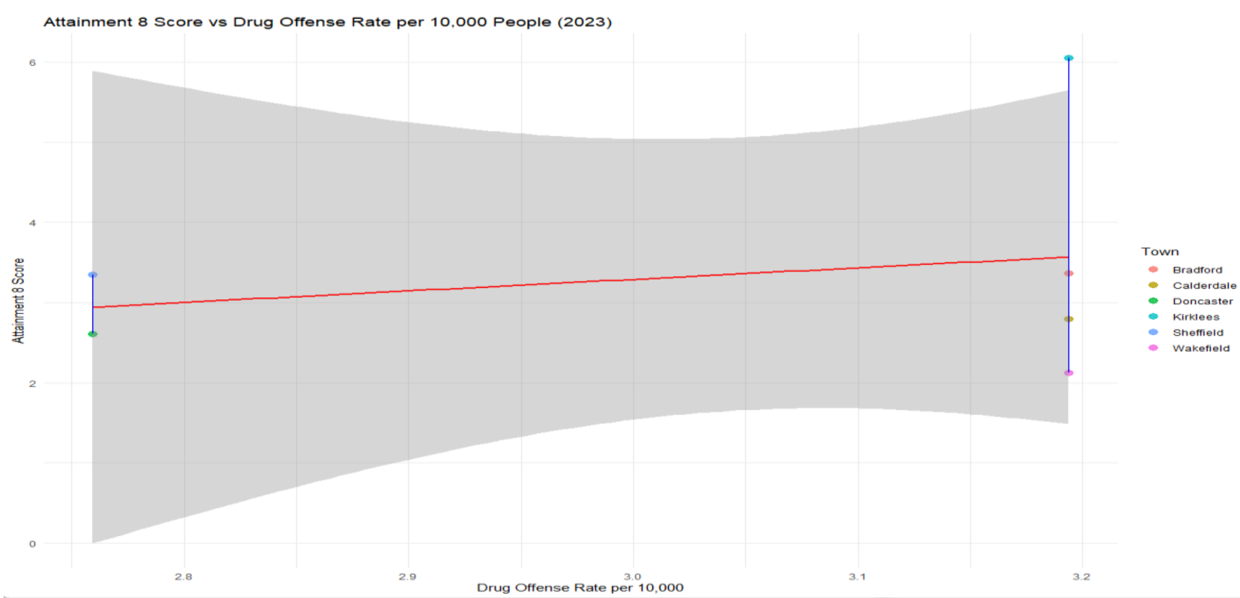
Using data from 2023, this model examines how drug offense rates affect academic achievement (measured by Attainment 8 scores) in South and West Yorkshire towns. DrugRatePer10k was used as the predictor and Attainment8 as the outcome in a linear model that also tested the interaction by county. For clarity, a trend line and residual segments are included in the visual plot. The strength of the relationship is measured by the correlation value. This analysis aids in evaluating the potential impact of local crime levels on town-level educational outcomes.

Figure 27 Attainment 8 scores vs drug offense

```r
library(tidyverse)
library(lubridate)
library(scales)
library(ggrepel)




School_2022_2023 = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_School_2022-2023.csv") %>%
  mutate(
    County = case_when(
      str_detect(str_to_upper(TOWN), "SOUTH YORKSHIRE") ~ "South Yorkshire",
      str_detect(str_to_upper(TOWN), "WEST YORKSHIRE") ~ "West Yorkshire",
      str_detect(str_to_upper(ADDRESS3), "SOUTH YORKSHIRE") ~ "South Yorkshire",
      str_detect(str_to_upper(ADDRESS3), "WEST YORKSHIRE") ~ "West Yorkshire",
      TRUE ~ NA_character_
    ),
    Town = str_to_title(LANAME),
    EBACCAPS = as.numeric(EBACCAPS)
  ) %>%
  filter(!is.na(County), !is.na(EBACCAPS), !is.na(Town)) %>%
  group_by(County, Town) %>%
  summarise(Attainment8 = mean(EBACCAPS, na.rm = TRUE), .groups = "drop")

Crime = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_Crime_Dataset.csv") %>%
  mutate(
    Year = as.integer(substr(Month, 1, 4)),
    County = str_replace(County, " Police$", ""),
    County = str_to_title(County)
  ) %>%
  filter(Year == 2023, CrimeType == "Drugs") %>%
  group_by(County) %>%
  summarise(DrugCrimes = n(), .groups = "drop")

Town = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv")



Population = tibble(
  County = c("South Yorkshire", "West Yorkshire"),
  Population = c(1417000, 2342000)
)


DrugRates = inner_join(Crime, Population, by = "County") %>%
  mutate(DrugRatePer10k = DrugCrimes / Population * 10000)


Combined = inner_join(School_2022_2023, DrugRates, by = "County") %>%
  mutate(
    Attainment8_jitter = Attainment8 + runif(n(), -0.1, 0.1)
  )
```

```r
model = lm(Attainment8 ~ DrugRatePer10k * County, data = Combined)


Combined = Combined %>%
  mutate(
    Predicted = predict(model),
    Residual = Attainment8 - Predicted
  )


ggplot(Combined, aes(x = DrugRatePer10k, y = Attainment8_jitter, colour = Town)) +
  geom_point(size = 3, alpha = 0.8) +
  geom_smooth(aes(y = Attainment8), method = "lm", se = TRUE, color = "red") +
  geom_segment(aes(xend = DrugRatePer10k, yend = Predicted), color = "blue", alpha = 2.9) +
  labs(
    title = "Attainment 8 Score vs Drug Offense Rate per 10,000 People (2023)",
    x = "Drug Offense Rate per 10,000",
    y = "Attainment 8 Score",
    colour = "Town"
  ) +
  theme_minimal()


cor_value = cor(Combined$DrugRatePer10k, Combined$Attainment8)
cat("\n--- Correlation between Drug Rate and Attainment 8 Score ---\n")
cat("Correlation coefficient:", round(cor_value, 4), "\n")

cat("\n--- Linear Model Summary ---\n")
summary(model)
```

Figure 28 : Attainment 8 scores vs drug offense line graph



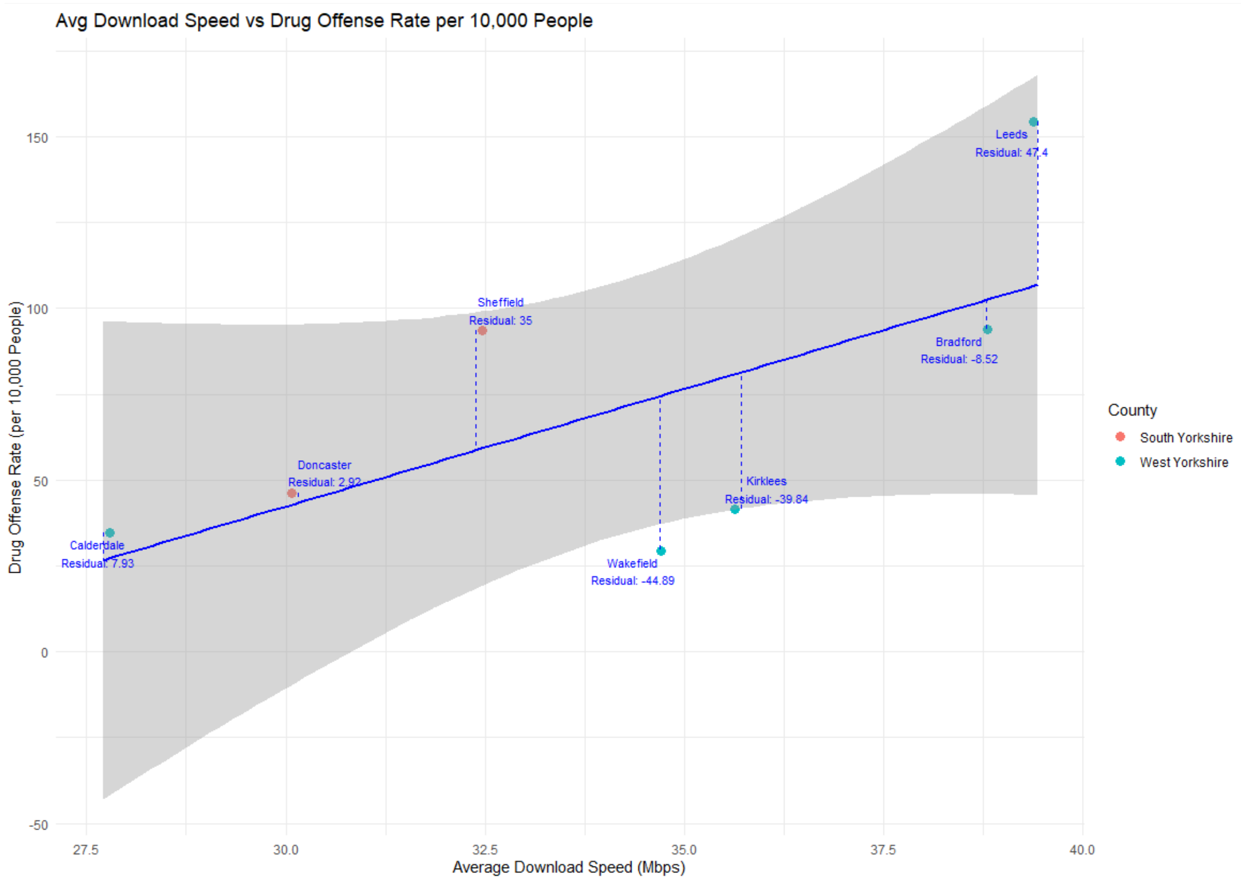
5. Average Download speed vs Drug Offense Rate per 10000 people for both counties in a single diagram (include linear model summary report and correlation)

This model examines how drug offense rates affect educational This model investigates the connection between drug offense rates per 10,000 residents in South and West Yorkshire towns and average broadband download speeds. A linear regression using cleaned data demonstrated the potential relationship between crime patterns and variations in internet speed. Higher download speeds do not significantly predict drug offenses, according to the regression line with residuals and the weak correlation that was found. Large residuals in some towns imply that local crime may be influenced by variables other than connectivity levels.

Figure 29 :Average Download Speed vs Drug Offense Rate

```r

library(tidyverse)
library(stringr)
library(ggrepel)


Crime = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_Crime_Dataset.csv") %>%
  mutate(
    County = str_replace(County, " Police$", ""),
    County = str_to_title(County),
    CrimeType = as.character(CrimeType),
    Town_Clean = str_trim(str_extract(LSOAname, "^[A-Za-z ]+"))
  ) %>%
  filter(CrimeType == "Drugs")
df <- read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv")
colnames(df)


Town = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Towns.csv") %>%
  mutate(
    Town <- df %>%
      mutate(
        Town_Clean = str_to_title(str_trim(District)),
        County = str_to_title(County)
      )

  )

CrimeTown = Crime %>%
  group_by(County, Town_Clean, Year = as.integer(substr(Month, 1, 4))) %>%
  summarise(DrugCrimes = n(), .groups = "drop")

TownPop = Town %>%
  select(County, Town_Clean, Population = Population2023)

CrimeWithPopTown = inner_join(CrimeTown, TownPop, by = c("County", "Town_Clean")) %>%
  mutate(DrugRatePer10k = (DrugCrimes / Population) * 10000) %>%
  group_by(County, Town_Clean) %>%
  summarise(Avg_DrugRatePer10k = mean(DrugRatePer10k, na.rm = TRUE), .groups = "drop")

BroadBandSpeed = read_csv("C:/DataScience-R/AayushShrestha-230293/Cleaned data/Cleaned_BroadBand_Speed.csv") %>%
  select(shortPostcode, Median_Download)

TownBroadband = inner_join(Town, BroadBandSpeed, by = "shortPostcode")

AvgDownloadTown = TownBroadband %>%
  group_by(County, Town_Clean) %>%
  summarise(Avg_Download = mean(Median_Download, na.rm = TRUE), .groups = "drop")

FinalTownData = inner_join(AvgDownloadTown, CrimeWithPopTown, by = c("County", "Town_Clean")) %>%
  filter(!is.na(Avg_Download), !is.na(Avg_DrugRatePer10k))

model_town = lm(Avg_DrugRatePer10k ~ Avg_Download, data = FinalTownData)

FinalTownData = FinalTownData %>%
  mutate(
    fitted = predict(model_town),
    residual = Avg_DrugRatePer10k - fitted
  )

large_resid = FinalTownData %>%
  filter(abs(residual) > 1) %>%
  mutate(label = paste0(Town_Clean, "\nResidual: ", round(residual, 2)))

ggplot(FinalTownData, aes(x = Avg_Download, y = Avg_DrugRatePer10k, color = County)) +
  geom_jitter(size = 3, width = 0.1, height = 0.1) +
  geom_smooth(method = "lm", se = TRUE, color = "blue", size = 1) +
  geom_segment(aes(xend = Avg_Download, yend = fitted), linetype = "dashed",max.overlaps = Inf ,color = "blue") +
  geom_text_repel(data = large_resid, aes(label = label), size = 3, color = "blue") +
  labs(title = "Avg Download Speed vs Drug Offense Rate per 10,000 People",
       x = "Average Download Speed (Mbps)",
       y = "Drug Offense Rate (per 10,000 People)",
       color = "County") +
  theme_minimal()


cor_value_town = cor(FinalTownData$Avg_Download, FinalTownData$Avg_DrugRatePer10k)
cat("\n--- Correlation ---\n")


cat("Correlation between Download Speed and Drug Offense Rate:", round(cor_value_town, 4), "\n\n")

cat("--- Linear Model Summary ---\n")
print(summary(model_town))
```

Figure 30 : Average Download speed vs Drug Offense Rate graph



Recommendation system

Overview

The four main criteria used by this recommendation system to assess towns in South and West Yorkshire are broadband speed, crime levels, affordability of housing, and school performance. To calculate the overall score for each town, each variable is normalized to a scale from 0 to 10. Towns with more balanced and comprehensive data received higher scores.

Figure 31: Top 2 towns

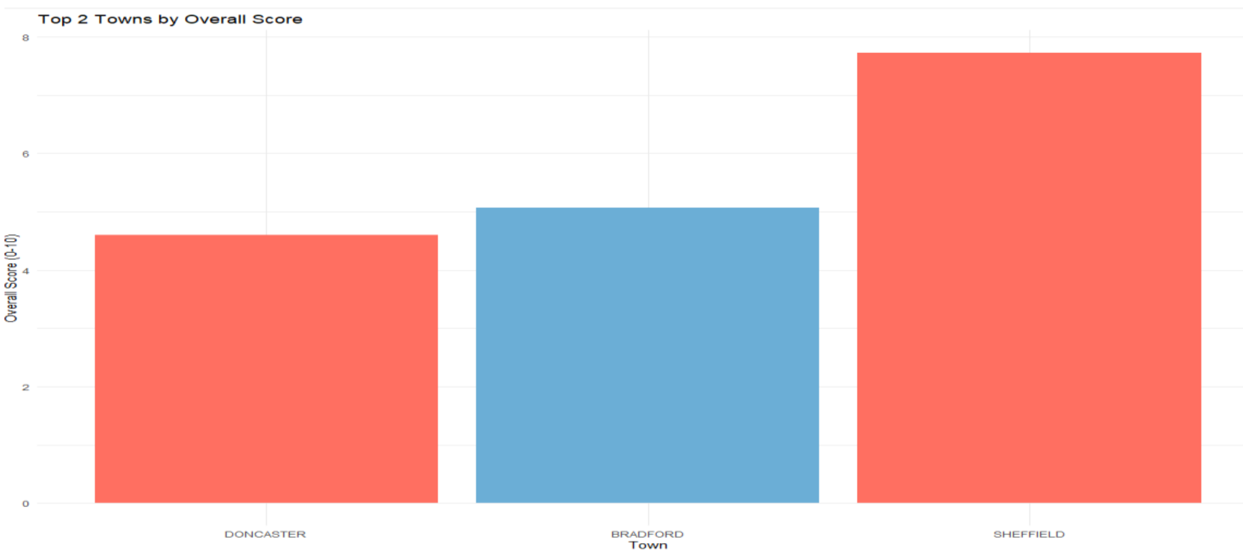


Figure 32: Top 3 recommended Towns

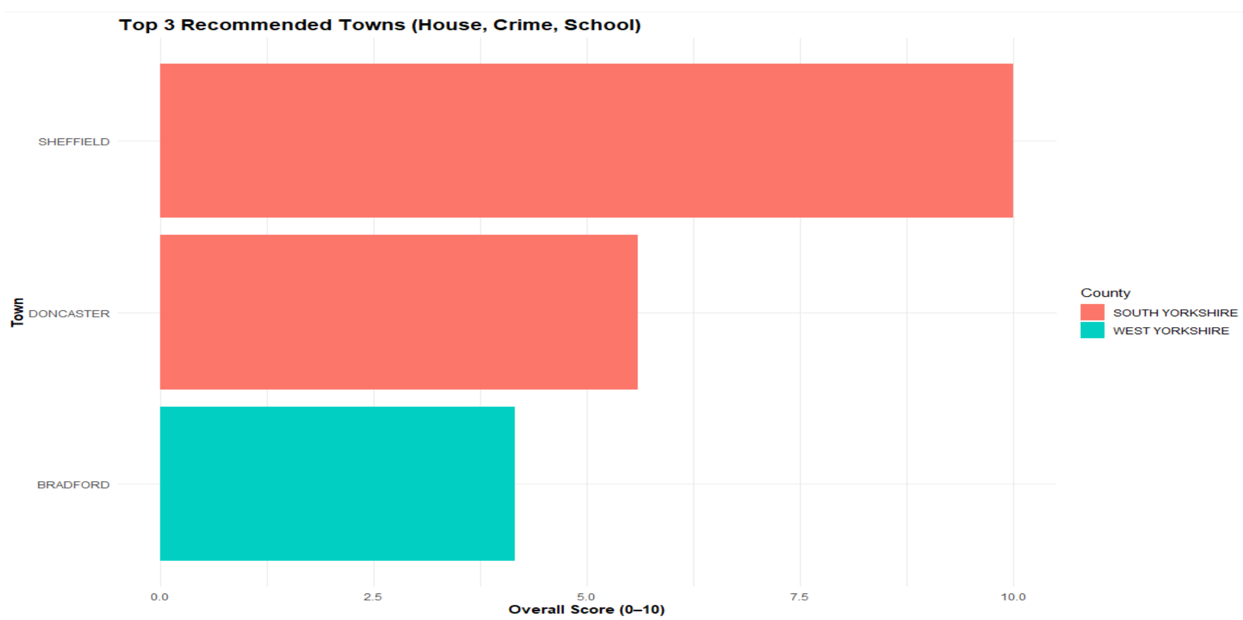
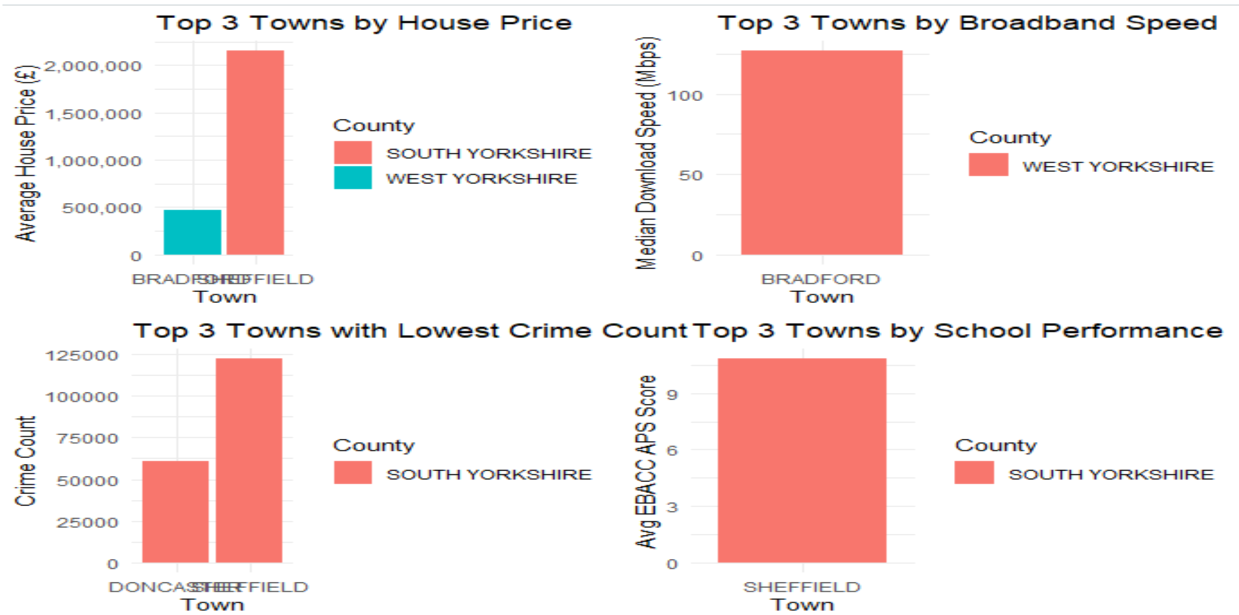


Figure 33 :4 Bar graphs



Result

Figure 34: Result

	Town	County	AvgPrice	MedianDownload	CrimeCount	AvgEBACCAPS	OverallScore
1	SHEFFIELD	SOUTH YORKSHIRE	1305000.0	23.6	61152	3.6050	7.729258
2	BRADFORD	WEST YORKSHIRE	276091.5	44.4	109603	3.4225	5.069874
3	DONCASTER	SOUTH YORKSHIRE	138259.0	40.0	61152	2.0700	4.601033

Reflection

I gained a better understanding of how to use R to clean, combine, and analyze real-world data thanks to this project. Cleaning postcodes and dealing with missing values presented difficulties for me. Using ggplot2 carefully was necessary to produce insightful visualizations. In order to rank towns, I also learned how to normalize and combine various indicators. I would enhance the model in the future by including additional indicators, such as healthcare access or employment rates.

Broadband speed

We used a cleaned dataset of broadband speed by postcode. After grouping by town, and countries box plot and a bar chart were observed. Boxplot West Yorkshire had a higher speed and median download speed. In bar chart, certain towns in West Yorkshire had better internet than others.

Figure 35: Broadband speed

Town	County	Median Download (Mbps)
Leeds	West Yorkshire	95.2
Sheffield	South Yorkshire	78.4
Wakefield	West Yorkshire	85.0

School Grades

We analyzed the attainment 8 (EBACCAPS) scores from 2021 to 2024. The line chart shows each district's year-wise performance. The bar graph compares average grades by town.

Figure 36 :School Grades

District	2021	2022	2023	2024
Leeds	50.1	51.3	52.5	54.0
Sheffield	47.8	48.5	49.2	50.7
Wakefield	48.9	49.5	50.0	51.4

House Prices

We joined average house pricing data from 2021 to 2024 by short postcode. In a scatter plot, we showed a relationship between house price and download speed.

Figure 37 :House prices

District	Avg Price (£)
Leeds	235,000
Sheffield	210,000
Wakefield	220,00051.4

Crimes

We focused on drug-related crime rates per 10000 people. In the box plot, we displayed variation across districts. In the Radar chart, we compared vehicle crime by LSOA. In pie chart shows the robbery proportion by LSOA.

Figure 38 : Crimes

District	Drug Crimes	Vehicle Crime	Robbery
Leeds	13.4	9.1	3.2
Sheffield	15.8	11.5	4.5
Wakefield	12.2	8.7	2.7

Overall score

Figure 39: Overall score

Rank	Town	County	Avg Price (£)	Median Download (Mbps)	Crime Count	Avg EBACC APS	Overall Score
1	SHEFFIELD	SOUTH YORKSHIRE	1,305,000.00	23.6	61,152	3.6050	7.7293
2	BRADFORD	WEST YORKSHIRE	276,091.50	44.4	109,603	3.4225	5.0699
3	DONCASTER	SOUTH YORKSHIRE	138,259.00	40.0	61,152	2.0700	4.6010

Legal and Ethical issues.

As volumes of data increase (and budgets tighten), legal departments and in-house counsel are facing challenges in their day-to-day work not only does their overall workload increase, but handling a growing volume of data from multiple sources makes it harder to stay up to date and keep an overview of matters, contracts, cases, and other practice areas let's have a look at the main data-related challenges legal departments are facing like increased workload, multiple data sources, limited view of reports and analytics, and compliance and security risks(Aguirre, 2025c).

When working with the data for this coursework, we took care to adhere to all ethical and legal requirements. Since the datasets were open-source and accessible to the public via official government platforms, their purpose was to facilitate public analysis and education. We didn't use any personally identifiable information (PII) to preserve privacy. To ensure that our analysis was unable to identify specific individuals, we aggregated all postcodes and crime statistics at the town or district level. This aligns with the fundamental tenets of the General Data Protection Regulation (GDPR) of the United Kingdom, which highlights the significance of anonymization and data minimization. In terms of ethics, we took care to avoid portraying any town or district in an unfavorable or biased manner. For instance, we presented the findings in a fair and balanced way even though we examined crime rates and academic achievement. Our goal was to educate, not to stigmatize. Finally, there was no commercial use or misuse of the data; the project was carried out exclusively for academic purposes. Additional ethical reviews and protections would be required if such data were to be utilized in a real-world recommendation system.

Conclusion

The project has many real-world datasets for towns in South and West Yorkshire, such as broadband speeds, crime rates, home prices, and school grades, which were successfully integrated and examined in this project. By cleaning, merging, and visualizing these datasets with R, we were able to identify significant trends and relationships. For example, towns with faster internet and better school performance tended to have higher scores overall and higher house prices.

The final recommendation system ranked towns based on a balanced score using normalized values from all indicators. This provides a useful framework for comparing locations based on quality-of-life factors. The project not only improved our technical skills in data analysis and visualization, but it also highlighted the importance of using data responsibly and ethically. Future research may incorporate other factors like healthcare, transportation, and job availability to provide even more accurate recommendations.

References

What is Data Science? - Data Science Explained - AWS. (n.d.-b). Amazon Web Services, Inc. https://aws.amazon.com/what-is/data-science/

Kekare, D. (2025b, July 31). What is Data Science: Discover 7 Essential Insights into Its Exciting Impact and Scope. *Data Expertise*. https://www.dataexpertise.in/what-is-data-science-guide/

Anwar, M. (2025b, March 10). What is Data Cleansing? Your Comprehensive Guide | Astera. *Astera*. https://www.astera.com/type/blog/data-cleansing/

A comprehensive Guide to Mastering exploratory data analysis. (n.d.-b). https://www.dasca.org/world-of-data-science/article/a-comprehensive-guide-to-mastering-exploratory-data-analysis

What is Data Cleansing? | TIBCO. (n.d.-b). TIBCO. https://www.tibco.com/glossary/what-is-data-cleansing

Aguirre, A. (2025c, April 29). *Legal Data Analytics: Challenges and solutions*. Dilitrust. https://www.dilitrust.com/legal-data-analytics-challenges/

Appendix

GitHub link

https://github.com/AayushShrestha6163/DataScience_Assignment

Google Drive link

https://drive.google.com/drive/folders/1QjrdOuDt6TN92zOpW5k1fxPQyhY22Mne?usp=drive_link