Softwarica
COLLEGE OF IT & E-COMMERCE

in collaboration with

Coventry University

Report on Property Investment in South and West Yorkshire

Batsal Shrestha

College ID: 230509, Coventry ID: 14809402

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

ST5014CEM Data Science for Developers

Siddhartha Neupane

August 05, 2025

Table of Contents

List of Figures

Introduction

Real estate investors often seek locations that offer a mix of affordability, connectivity, safety, and quality of life. South Yorkshire and West Yorkshire are two counties in the UK known for their relatively affordable housing markets and improving infrastructure. This report is on the comparison of these two counties using data-driven research to guide foreign investors on an informed decision based on key factors determining property investment.

The primary goal is to provide international investors with insights gained through the data, showing which county offers a better environment for property investment. By looking at factors like crime rates, broadband speeds, housing costs, and public amenities like schools, the report brings attention to trends and regional differences that may not be easily noticed.

The analysis uses datasets from the UK government: the House Price dataset from gov.uk (Price Paid Data, 2025), broadband speed data from Ofcom (Ofcom, 2023), crime data from Police UK (Data Downloads|data.police.uk, n.d.), and school performance data from the Department for Education (Compare School Performance, n.d.). Additional population and geographic mapping datasets were provided by the module leader.

This report follows the full data science lifecycle: data collection, cleaning, exploratory data analysis, linear modeling, and investment recommendation. All stages are executed using R, which enables efficient processing, statistical modeling, and clear visual results.

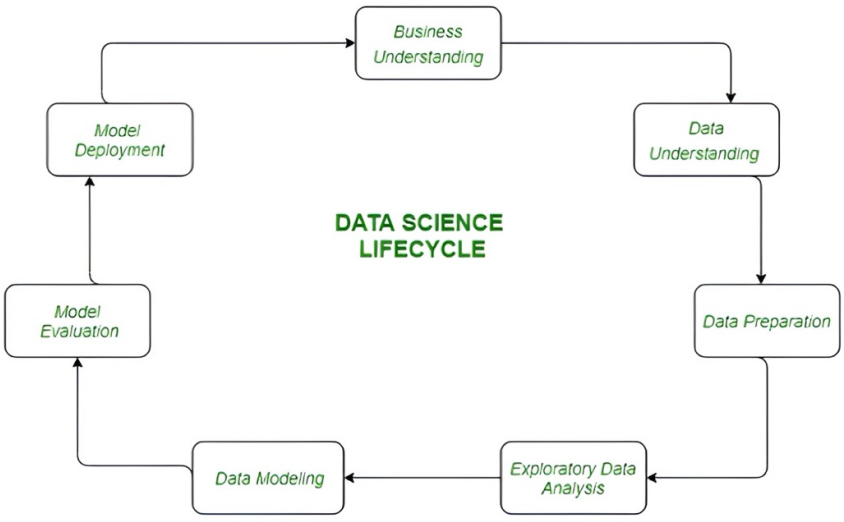

Figure 1: Data Science Lifecycle

Cleaning Data

Cleaning data is the elementary starting point of data science to obtain consistency, precision, and wholeness. Low quality data can affect the results and lead to incorrect conclusions (GeeksforGeeks, 2025). Five core datasets are utilized in this project: house sales, towns and population, broadband speed, crime, and schools, from genuine UK government websites. Each dataset needed specific cleaning to manage missing values, standardize formats, and allow for fair comparison of South and West Yorkshire.

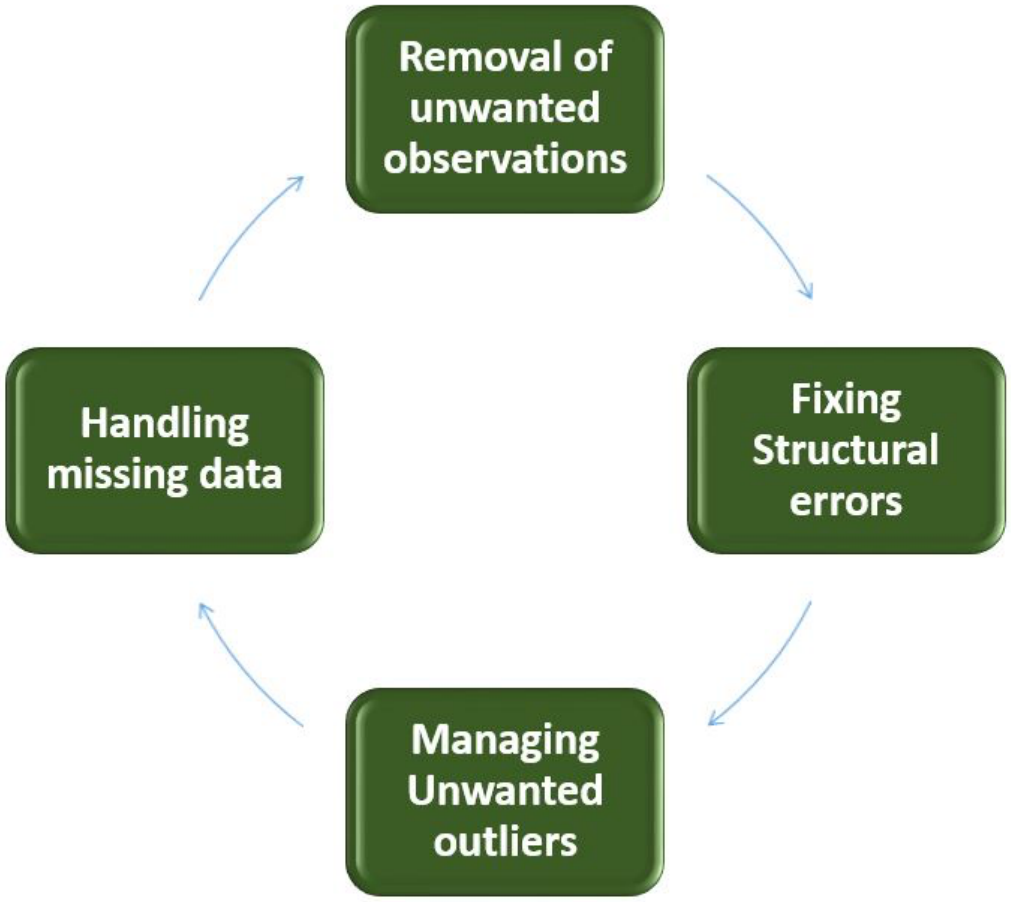

Figure 2: Cleaning Data

House Sales

For the house price dataset from 2021 to 2024, data cleaning included renaming the columns, filtering for Yorkshire counties, standardizing formats, and removing missing records. Dates were parsed, short postcodes derived for aggregation, and LSOA codes merged using government postcode mapping. This ensured consistency, added geographic context, and kept relevant fields for regional comparison.

```r
# Select and clean columns
df = df %>%
  select(Price, Date, Postcode, Town, District, County, Property_Type) %>%
  mutate(
    Price = as.numeric(Price),
    Date = as.Date(Date),
    Postcode = toupper(Postcode),  # Standardize to uppercase but keep spaces
    County = str_to_title(County),  # Convert to Proper Case
    District = str_to_title(District),
    Town = str_to_title(Town)
  ) %>%
  drop_na() %>%
  filter(County %in% c("South Yorkshire", "West Yorkshire")) %>%
  mutate(
    Year = year(Date),
    shortPostcode = str_sub(str_replace_all(Postcode, " ", ""), 1, 4)  # For grouping
  )

```

9:32 (Top Level) R Script

	Price	Date	Postcode	Town	District	County	Property_Type	Year	shortPostcode
1	150000	2021-11-08	LS27 9AL	Leeds	Leeds	West Yorkshire	O	2021	LS27
2	430000	2021-10-29	LS6 1BU	Leeds	Leeds	West Yorkshire	D	2021	LS61
3	135000	2021-10-26	LS10 1LP	Leeds	Leeds	West Yorkshire	F	2021	LS10
4	131000	2021-12-21	HD2 2SN	Huddersfield	Kirklees	West Yorkshire	O	2021	HD22
5	180000	2021-10-26	LS10 1NG	Leeds	Leeds	West Yorkshire	F	2021	LS10
6	65000	2021-12-23	HD4 6DL	Huddersfield	Kirklees	West Yorkshire	O	2021	HD46
7	180000	2021-10-26	LS10 1NW	Leeds	Leeds	West Yorkshire	O	2021	LS10
8	60000	2021-12-20	BD4 7EJ	Bradford	Bradford	West Yorkshire	T	2021	BD47
9	255000	2021-09-29	BD10 0QZ	Bradford	Bradford	West Yorkshire	O	2021	BD10
10	93500	2021-07-16	LS12 5LT	Leeds	Leeds	West Yorkshire	F	2021	LS12
11	6492000	2021-12-17	WF13 2SU	Dewsbury	Kirklees	West Yorkshire	O	2021	WF13
12	300000	2021-12-20	BD4 7HY	Bradford	Bradford	West Yorkshire	O	2021	BD47

Showing 1 to 12 of 241,528 entries, 10 total columns

Towns and Postcodes

Town names from house price data were standardized. Only towns from South and West Yorkshire were kept. Population figures were cleaned and projected for 2020-2024 using annual growth rates. Postcode prefixes were extracted to link population with town data. This enabled consistent, location-based analysis at the town level in the two counties.

```
64  # Remove duplicates, pick first location info per shortPostcode
65  postcode_location = all_house_data %>%
66    distinct(shortPostcode, Town, District, County) %>%
67    group_by(shortPostcode) %>%
68    summarise(
69      Town = first(Town),
70      District = first(District),
71      County = first(County)
72    ) %>%
73    ungroup()
74
75  # Join population and location data
76  population_with_location = pop_data_clean %>%
77    left_join(postcode_location, by = "shortPostcode") %>%
78    drop_na() %>%
79    filter(County %in% c("West Yorkshire", "South Yorkshire")) %>%
80    select(shortPostcode, Town, District, County, Population2020, Population2021, Population2022, Popul
81
86:31  (Top Level)                                                                    R Script
```

	shortPostcode	Town	District	County	Population2020	Population2021	Population2022
1	BD1	Bradford	Bradford	West Yorkshire	4074.795	4096.901	4117.058
2	BD10	Bradford	Bradford	West Yorkshire	27493.183	27642.333	27778.333
3	BD11	Bradford	Leeds	West Yorkshire	12061.861	12127.297	12186.963
4	BD12	Bradford	Bradford	West Yorkshire	17723.659	17819.810	17907.484
5	BD13	Bradford	Bradford	West Yorkshire	26533.970	26677.917	26809.172
6	BD14	Bradford	Bradford	West Yorkshire	9423.229	9474.350	9520.964
7	BD15	Bradford	Bradford	West Yorkshire	16269.438	16357.699	16438.179
8	BD16	Bingley	Bradford	West Yorkshire	27409.265	27557.960	27693.545
9	BD17	Shipley	Bradford	West Yorkshire	19015.357	19118.515	19212.578
10	BD18	Shipley	Bradford	West Yorkshire	28954.840	29111.920	29255.151
11	BD19	Cleckheaton	Kirklees	West Yorkshire	23389.707	23516.596	23632.297
12	BD2	Bradford	Bradford	West Yorkshire	30344.264	30508.882	30658.986

Showing 1 to 12 of 126 entries, 9 total columns

Broadband Speeds

Broadband performance and coverage data were cleaned by standardizing. Columns that were not relevant to the analysis were removed, and only entries from South and West Yorkshire were retained. Variables like download speed and percentage of availability were selected to compare connectivity between regions. This ensured consistency and geographic relevance for further analysis.

```r
# Function to clean broadband coverage data
clean_broadband_coverage = function(coverage_file, postcode_map) {
  df = read_csv(coverage_file, show_col_types = FALSE)
  # Use the pcds column which contains spaces
  df = normalize_postcode(df, "pcds")
  merged = df %>%
    inner_join(postcode_map, by = "Postcode_clean") %>%
    filter(County %in% c("South Yorkshire", "West Yorkshire")) %>%
    select(
      Postcode = pcds,
      District,
      County,
      SFBB_Availability = `SFBB availability (% premises)`,
      FTTP_Availability = `FTTP availability (% premises)`,
      Below_10Mbps = `% of premises unable to receive 10Mbit/s`
    )
  return(merged)
}
```

1:1 (Top Level) R Script

	Postcode	LSOA	District	County	AverageDownloadSpeed_Mbps
1	BD10 0AA	Bradford 017B	Bradford	West Yorkshire	39.7
2	BD10 0AB	Bradford 017B	Bradford	West Yorkshire	72.0
3	BD10 0AE	Bradford 017D	Bradford	West Yorkshire	28.0
4	BD10 0AF	Bradford 027C	Bradford	West Yorkshire	91.9
5	BD10 0AG	Bradford 017D	Bradford	West Yorkshire	11.9
6	BD10 0AH	Bradford 017D	Bradford	West Yorkshire	32.0
7	BD10 0AJ	Bradford 017D	Bradford	West Yorkshire	69.2
8	BD10 0AL	Bradford 017D	Bradford	West Yorkshire	73.0
9	BD10 0AN	Bradford 017D	Bradford	West Yorkshire	53.3
10	BD10 0AP	Bradford 017D	Bradford	West Yorkshire	23.5
11	BD10 0AQ	Bradford 027C	Bradford	West Yorkshire	65.4
12	BD10 0AT	Bradford 027C	Bradford	West Yorkshire	60.2

Showing 1 to 12 of 73,615 entries, 5 total columns

Crime

Monthly crime files for South and West Yorkshire were cleaned by selecting relevant columns, standardizing names, extracting the year and month, and removing invalid data. LSOA names were used to identify districts, matched using postcode mappings for geographic analysis. Cleaned records were then combined across counties and months to maintain accuracy and consistency for comparison.

```
40      # Clean and transform
41      df_cleaned = df %>%
42        select(CrimeID, Month, LSOA.name, Crime.type) %>%
43        mutate(
44          County = county_name,
45          District = word(LSOA.name, 1),
46          LSOA = LSOA.name,
47          Year = as.integer(substr(Month, 1, 4)),
48          Month = as.integer(substr(Month, 6, 7))
49        ) %>%
50        select(CrimeID, Year, Month, County, District, LSOA, CrimeType = Crime.type) %>%
51        filter(
52          !is.na(CrimeID), CrimeID != "",
53          !is.na(CrimeType), CrimeType != "",
54          !is.na(District), District != ""
55        ) %>%
56        left_join(postcode_lsoa, by = c("LSOA" = "LSOA")) %>%
57        filter(!is.na(Postcode))  # Remove rows without postcode
76:20  (Top Level)                                                    R Script
```

	CrimeID	Month	Year	CrimeType	Postcode	District
1	71db7d56e1a6a39042a6ea2f195d641b2f3343ec5ae...	5	2022	Burglary	S71 4DB	Barnsley
2	24ec97bea629f5fc24ad858b794093d3afad4d5c157f...	5	2022	Criminal damage and arson	S71 4DB	Barnsley
3	064ccc2986c64d1a9484442bfadad67ce592a192772...	5	2022	Criminal damage and arson	S71 4DB	Barnsley
4	fa8d02a8d62f37ee37231e342b9a7dc60cd0c15e8ba...	5	2022	Criminal damage and arson	S71 4DB	Barnsley
5	8ccfead985514740d7899f199d928e80230475c84c3...	5	2022	Public order	S71 4DB	Barnsley
6	f24b52feb76b4cb6210a7f685c5e1a657861f3c1794e...	5	2022	Public order	S71 4DB	Barnsley
7	54cb5091b403cf30cdfbe2338bcf956dfb2179d8847a...	5	2022	Violence and sexual offences	S71 4DB	Barnsley
8	cb60da2b704273535e3ab6d867d1ed9e21366c7c59...	5	2022	Violence and sexual offences	S71 4DB	Barnsley
9	b0fe29546601abed843a665a5332fc60a42ace47833...	5	2022	Violence and sexual offences	S71 4DB	Barnsley
10	33472d526649bcc5c91d577706282f5eb6071d1dc5c...	5	2022	Violence and sexual offences	S71 4DB	Barnsley
11	bc444e9f2d14f178f6eaef13e79dfb16714aba60d661...	5	2022	Violence and sexual offences	S71 4DB	Barnsley
12	4e70b510452b4fe0ec6a29bf35d5cecbe1c0fa911b99...	5	2022	Violence and sexual offences	S71 4DB	Barnsley

Showing 1 to 12 of 1,261,381 entries, 8 total columns

School

School performance data for KS2 and KS4 was cleaned by selecting key metrics, handling missing values, and converting special strings to numeric. School records were enhanced using postcode mappings to keep only those from South and West Yorkshire. Final datasets excluded incomplete or irrelevant rows, ensuring consistent structure for comparing educational quality across both regions.

```r
# Function to clean KS4 data
clean_ks4_data = function(year_folder, year_label, file_name) {
  file_path = file.path(input_base, year_folder, file_name)
  school_info_path = file.path(input_base, year_folder, "england_school_information.csv")

  ks4_data = suppressWarnings(
    read_csv(file_path, show_col_types = FALSE, guess_max = 10000)
  ) %>%
    mutate(School_URN = as.integer(URN)) %>%
    select(
      School_URN,
      School_Name = SCHNAME,
      Postcode = PCODE,
      ATT8SCR
    ) %>%
    mutate(
      Attainment_8_Score = clean_numeric(ATT8SCR),
      Year = year_label
```

131:20 (Top Level) R Script

	School_URN	School_Name	Attainment_8_Score	Year	Postcode
1	146501	Astrea Academy Dearne	39.0	2021-2022	S63 9EW
2	131749	Barnsley Academy	40.8	2021-2022	S70 3DL
3	130524	Barnsley College	6.0	2021-2022	S70 2YW
4	146320	Darton Academy	51.0	2021-2022	S75 5EF
5	144606	Holy Trinity Catholic and Church of England School	42.8	2021-2022	S71 2LF
6	146455	Horizon Community College	51.5	2021-2022	S70 6PD
7	140979	Kirk Balk Academy	50.6	2021-2022	S74 9HX
8	143907	Netherwood Academy	39.3	2021-2022	S73 8FE
9	139210	Outwood Academy Carlton	43.5	2021-2022	S71 3EW
10	139211	Outwood Academy Shafton	42.5	2021-2022	S72 8RE
11	106653	Penistone Grammar School	60.1	2021-2022	S36 7BX
12	106965	The Robert Ogden School	0.8	2021-2022	S63 0BG

Showing 1 to 12 of 813 entries, 8 total columns

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an important step in understanding the structure, patterns, and relationships within the data (GeeksforGeeks, 2025b). In this project, EDA was performed using R and relevant packages to investigate house prices, broadband speeds, crime rates, and academic performance in South and West Yorkshire. This analysis provided key insights to help make informed property investment decisions based on affordability, safety, connectivity, and education.

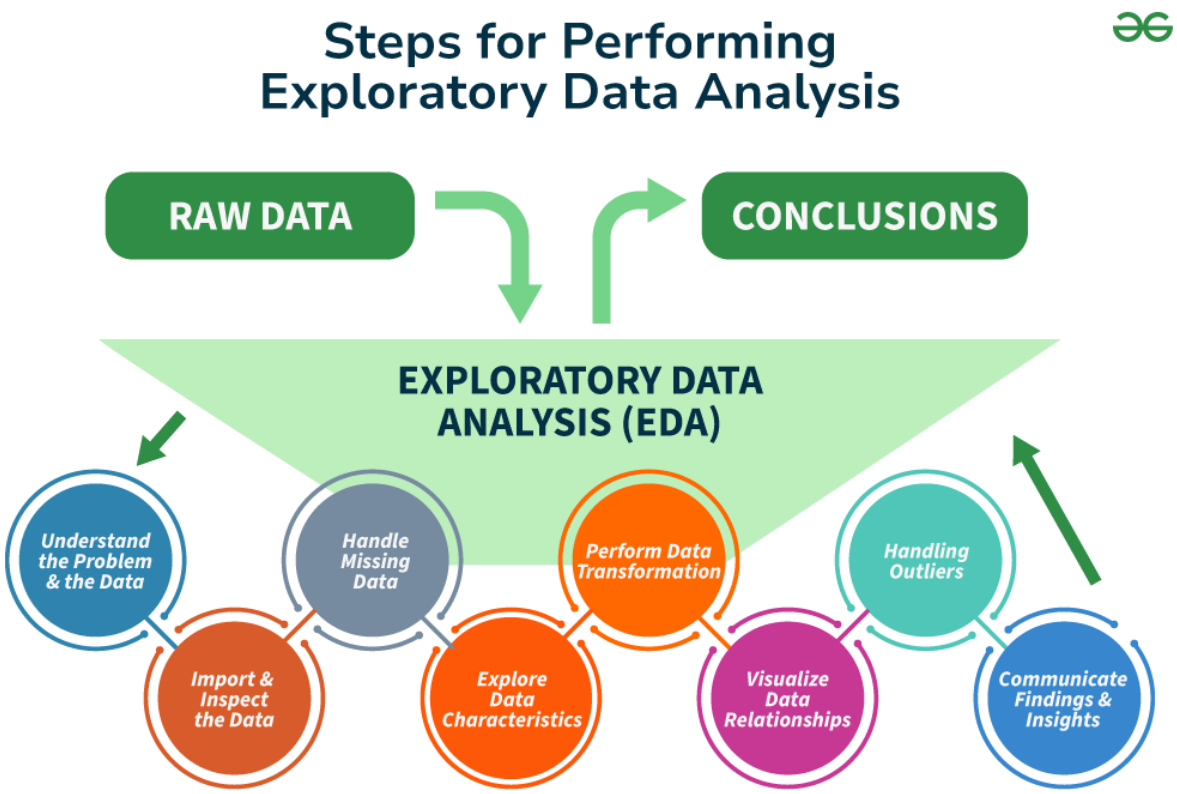

Figure 3: Exploratory Data Analysis

House Prices

This section explores regional differences in housing markets across South and West Yorkshire, focusing on trends, averages, and price distributions to determine affordability and investment possibilities.

Average House Price Trends by District and County

These line graphs show the average house price trends over 2021 to 2024 across districts in South Yorkshire and West Yorkshire.

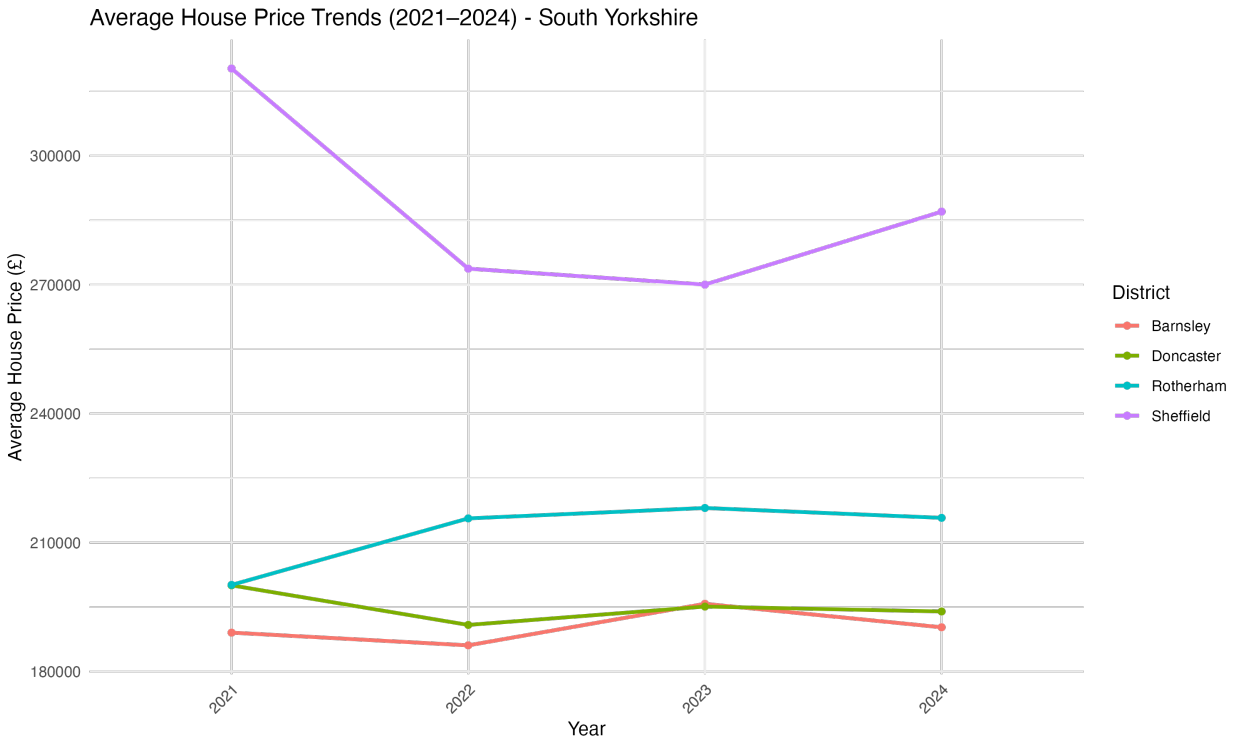

Figure 4: Average House Prices from 2021 to 2024 by District- South

Sheffield leads South Yorkshire but shows recent price declines after peaking in 2021. Barnsley and Doncaster fluctuate moderately, while Rotherham displays steady growth, suggesting mixed but stable market dynamics.

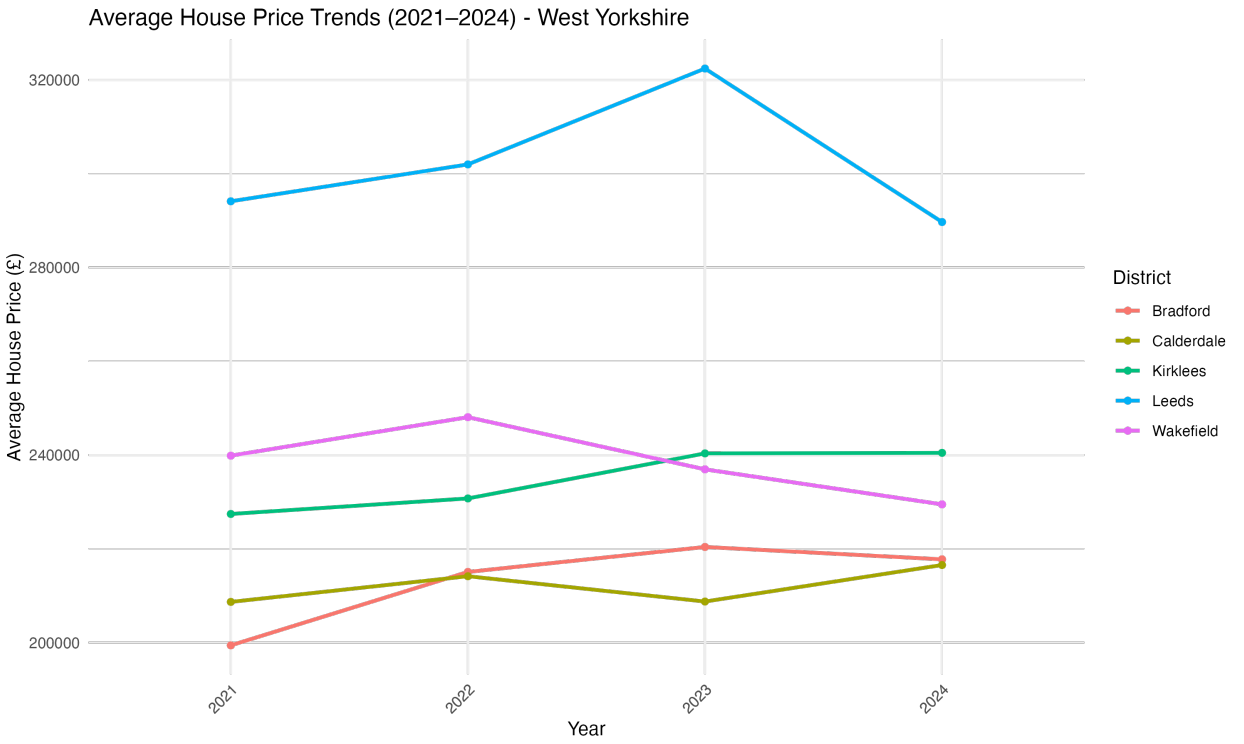

Figure 5: Average House Prices from 2021 to 2024 by District- West

Leeds shows high but volatile prices with a peak in 2023, followed by a slight dip. Kirklees steadily rises, Bradford and Calderdale fluctuate mildly, indicating varied affordability and investment opportunities.

West Yorkshire commands higher, more volatile prices, driven by Leeds. South Yorkshire has lower, steadier trends with Sheffield's recent dip. Both regions offer diverse investment and affordability profiles for buyers.

Average House Prices in 2023 by District and County

These bar graphs show the average house prices in South and West Yorkshire districts for 2023.

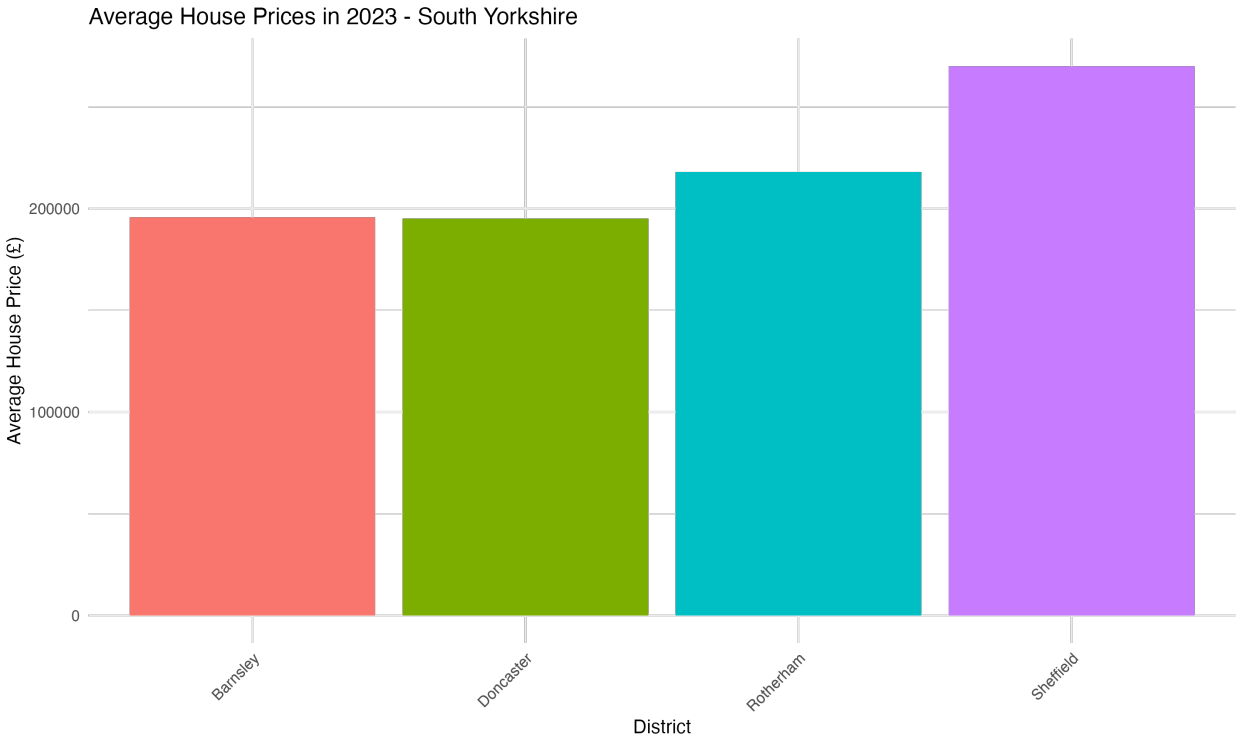

Figure 6: Average House Prices by District (2023)- South

South Yorkshire's most expensive district is Sheffield, while the least expensive districts are Barnsley and Doncaster. Rotherham offers mid-range prices, showing a diverse housing market with options for various budgets and investors.

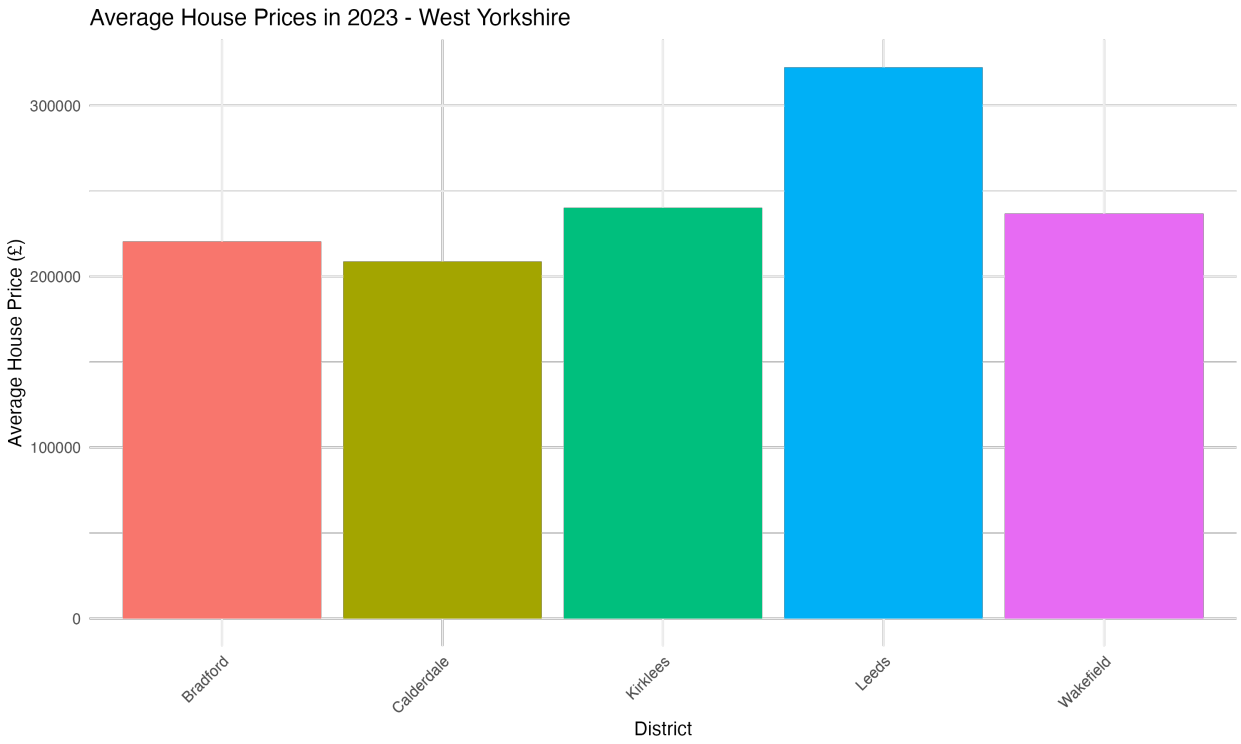

Figure 7: Average House Prices by District (2023)- West

Leeds has the highest prices in West Yorkshire. Kirklees and Wakefield are moderately positioned, and Bradford and Calderdale provide more affordable housing, illustrating varied opportunities throughout the region.

West Yorkshire generally has more expensive housing, particularly in Leeds, than South Yorkshire's more reasonably priced areas. This contrast helps investors weigh potential returns against budget considerations across counties.

House Price Distributions by District in South and West Yorkshire

These boxplots display the difference and distribution of house prices between districts in South Yorkshire and West Yorkshire.

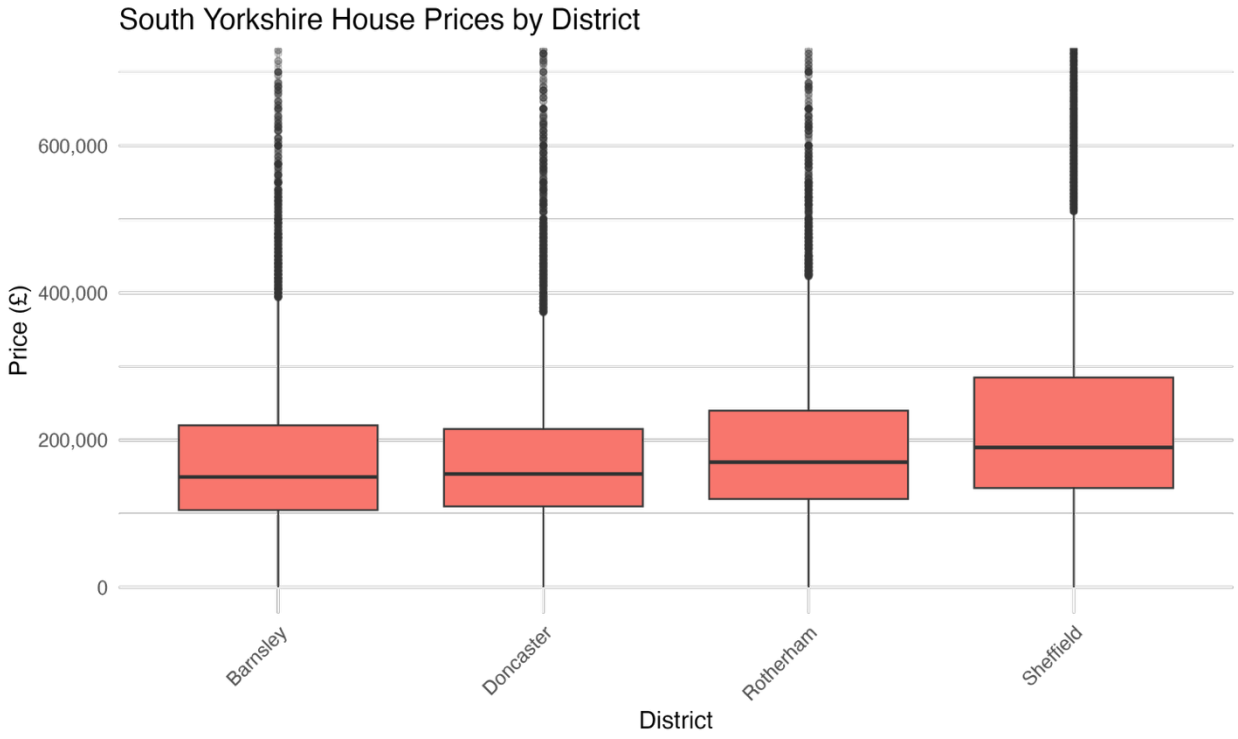

Figure 8: House Price Distribution by District - South

Barnsley and Doncaster have low medians and smaller ranges, demonstrating affordability. Sheffield stands out with high variability and extreme outliers, while Rotherham holds moderate values, indicating a mixed market.

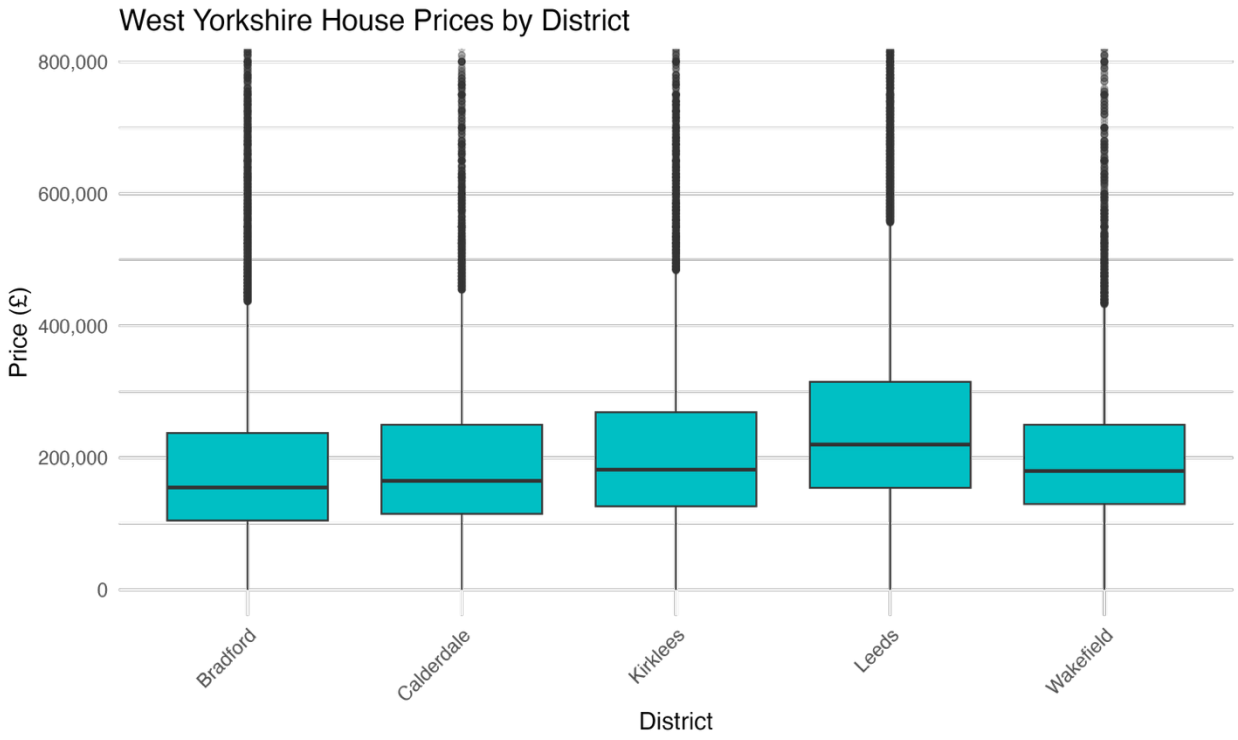

Figure 9: House Price Distribution by District- West

Leeds dominates with the highest median and wide spread, reflecting a premium market. Bradford and Calderdale remain affordable, while Kirklees and Wakefield show moderate pricing along with some high-end outliers.

West Yorkshire, led by Leeds, has higher medians and more price variety than South Yorkshire. South districts are more affordable overall, though both regions reveal high-end outliers across districts.

Broadband Speed

This section explores broadband download speed patterns across South and West Yorkshire to evaluate digital connectivity, a crucial factor for remote work, education, and modern lifestyle suitability.

Broadband Download Speed Distribution by District in South and West Yorkshire

These boxplots compares the distribution of average broadband download speeds across the districts in Yorkshire, both South and West.

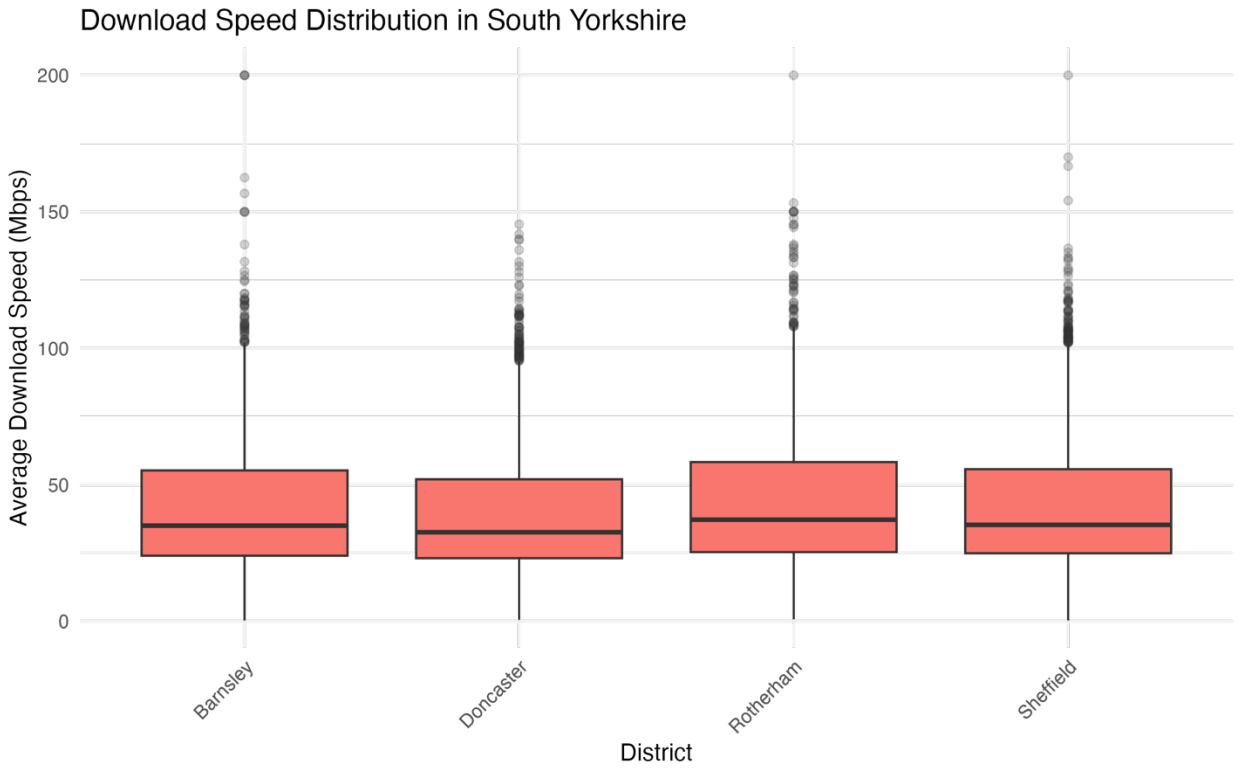

Figure 10: Download Speed Distribution by District- South

South Yorkshire has a range of download speeds, with Rotherham having the highest median. Barnsley and Doncaster are a little behind. Despite outliers, typical speeds remain moderate, reflecting decent but uneven digital connectivity.

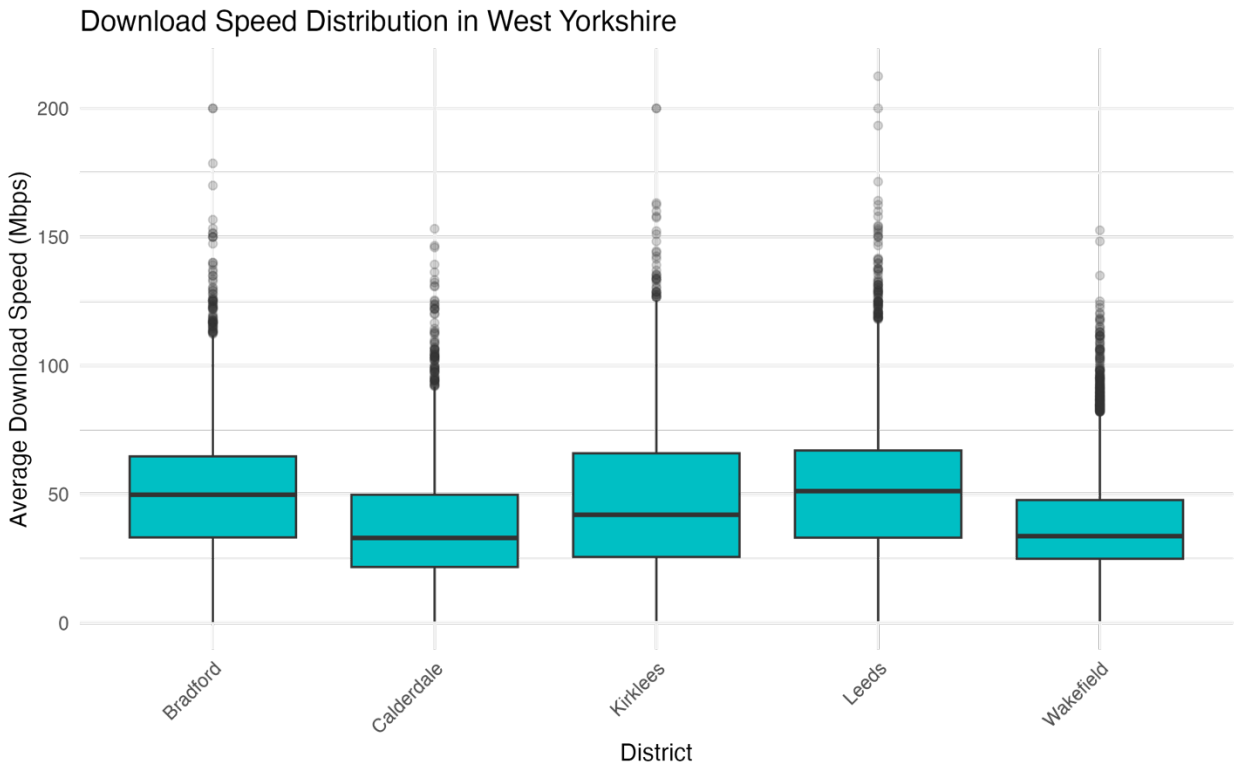

Figure 11: Download Speed Distribution by District- West

Leeds and Bradford lead in broadband speeds, offering strong digital infrastructure. Kirklees also shows strong performance, whereas Calderdale and Wakefield show lower medians. The region overall supports fast, reliable internet connectivity.

West Yorkshire outperforms the South in broadband speed, particularly in Leeds and Bradford. South Yorkshire offers fair connectivity, but slightly lower medians may affect remote work potential and digital service access.

Average Broadband Download Speed in South and West Yorkshire Towns

These bar charts show average broadband speeds in selected towns in South and West Yorkshire.

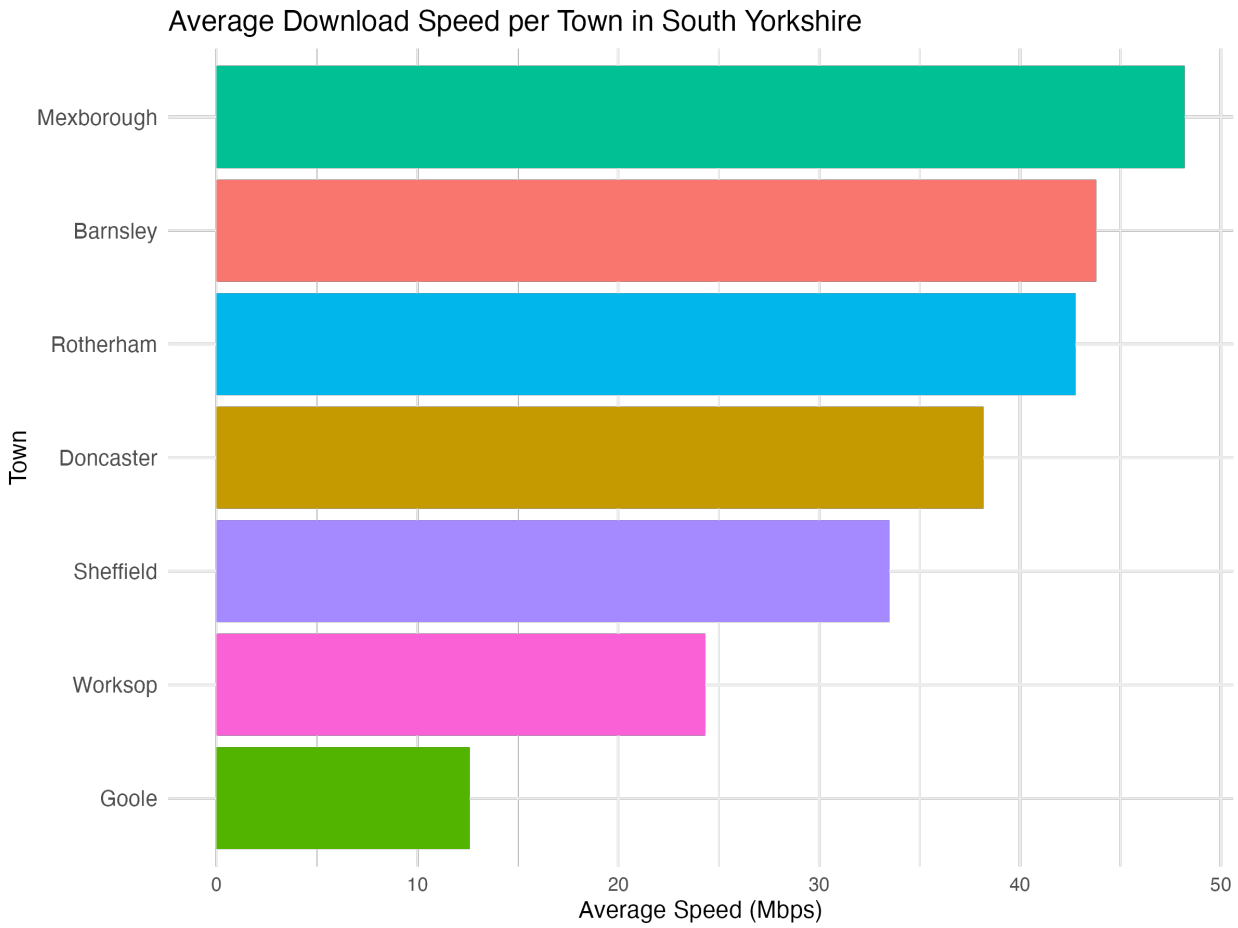

Figure 12: Download Speed by Town- South

The majority of South Yorkshire towns have speeds between 30 and 50 Mbps, with Mexborough setting the standard. Goole and Worksop trail behind, highlighting regional disparities and potential limitations in digital service access.

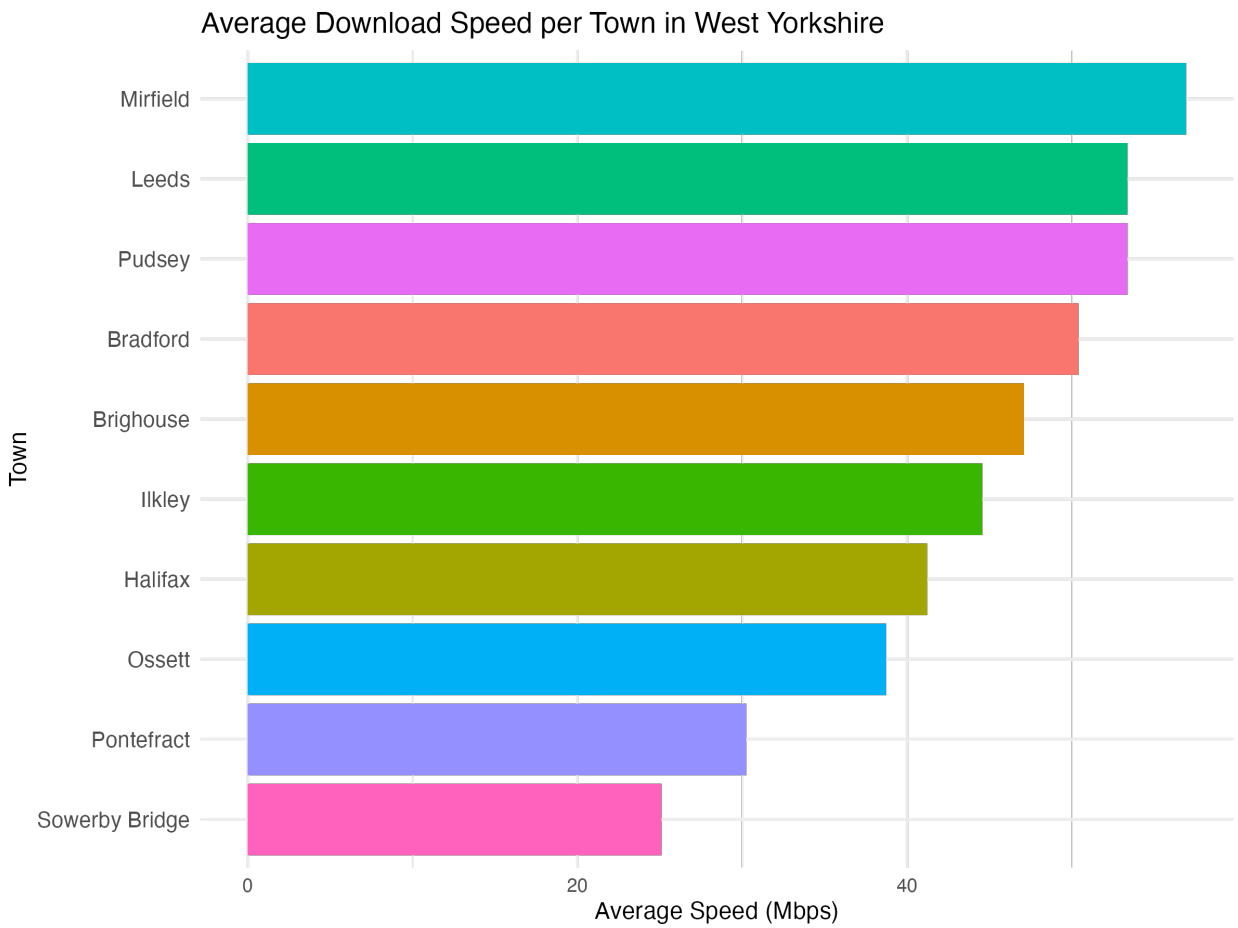

Figure 13: Download Speed by Town- West

West Yorkshire towns offer wider speed variation, from Sowerby Bridge to Mirfield. Towns like Leeds and Pudsey suggest strong digital readiness for modern lifestyles.

Overall, West Yorkshire towns generally perform better than those in the South based on broadband speed, offering greater range and peaks. This positions West as more attractive for technically proficient residents and remote workers looking for fast and reliable connectivity.

Crime Rate

This section examines trends in drug crime, robbery, and vehicle crime to assess public safety across South and West Yorkshire, which is key for investors prioritizing secure, low-risk property locations.

Drug Offense Rate Distribution by District and County

These boxplots show the annual drug offense rates per 10,000 people for each county's districts.

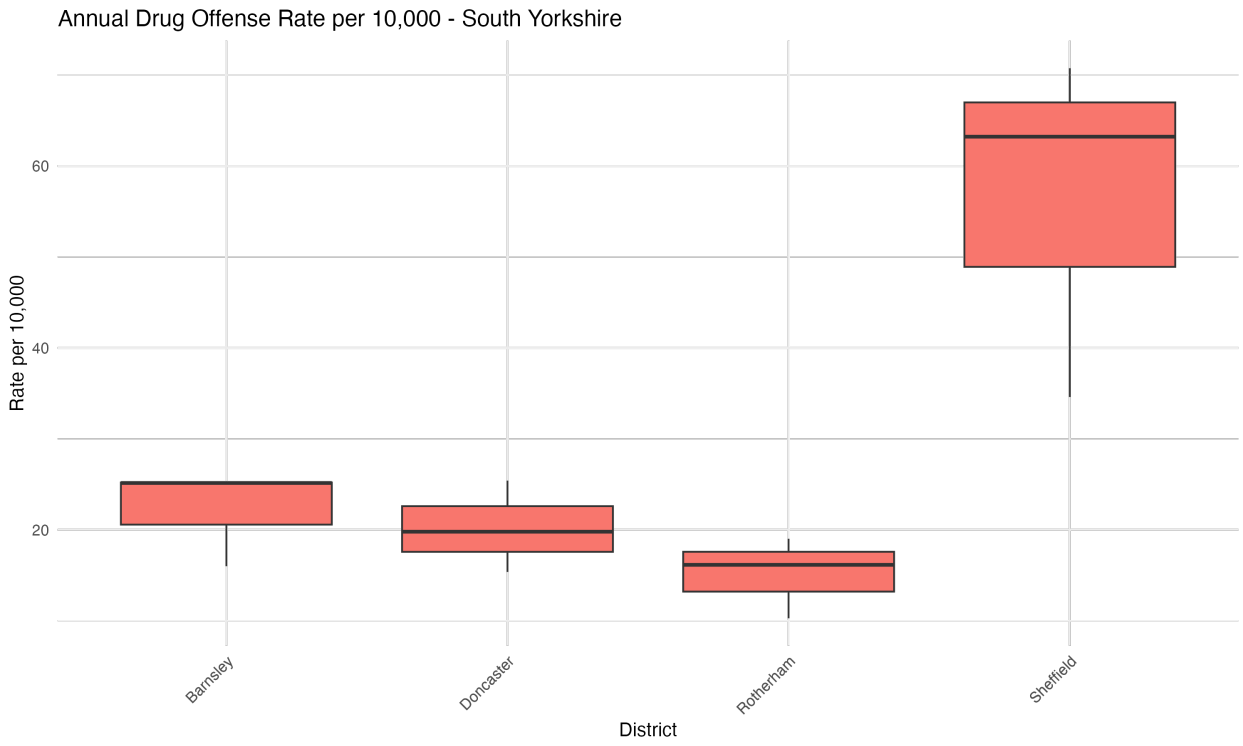

Figure 14: Drug Offence Rate Distribution by District- South

Sheffield reports the highest and most variable drug offense rates, indicating possible hotspots. The other districts, like Rotherham and Barnsley, have considerably lower medians, making them more stable and safer areas overall.

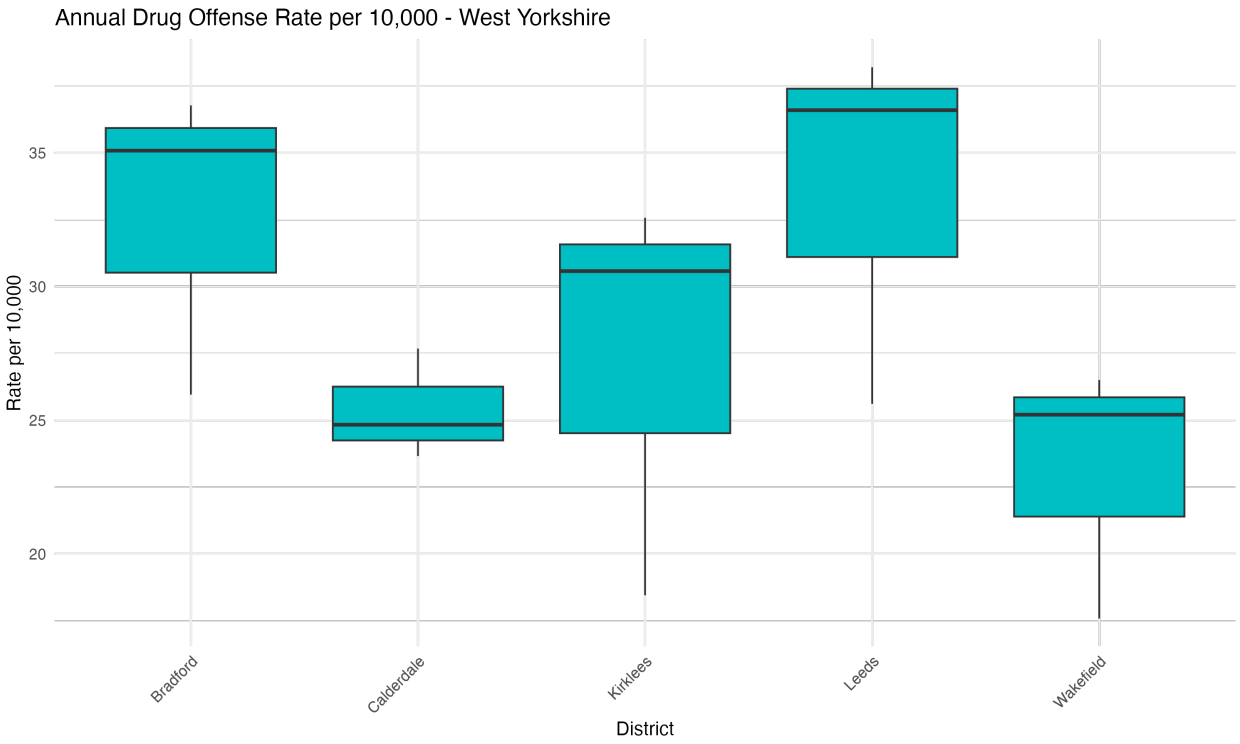

Figure 15: Drug Offence Rate Distribution by District- West

Offense rates are more balanced in West Yorkshire, with the highest offending being Leeds and Bradford. Calderdale remains notably low. Overall, this district shows moderate risk with fewer outliers.

South Yorkshire show stronger differences, with Sheffield showing higher rates. West Yorkshire appears more balanced, Leeds and Bradford remain a concern. Crime-aware investors may prefer lower-rate districts for rental stability.

Vehicle Crime Rate by District in South Yorkshire May 2022

This radar chart illustrates the vehicle crime rates per 10,000 people in South Yorkshire districts for May 2022.

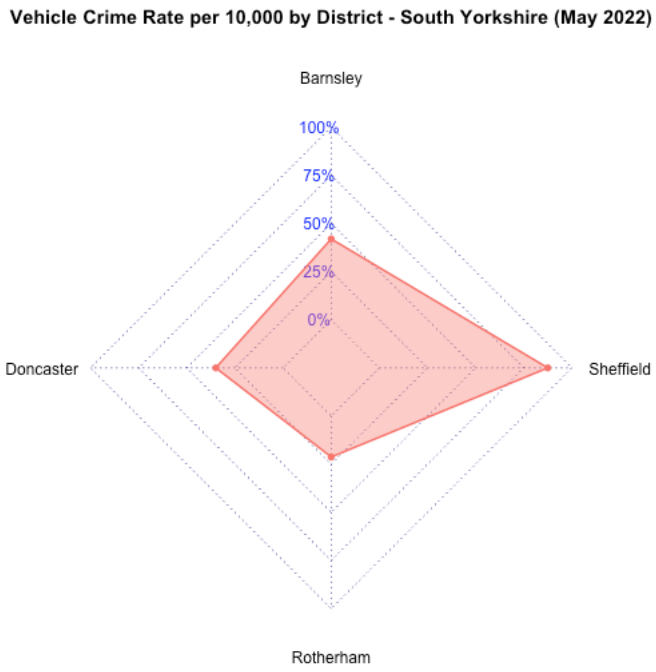

Figure 16: Vehicle Crime by District

Sheffield leads vehicle crime rates, with more than twice Barnsley and four times Rotherham. Doncaster stands mid-range. These figures suggest Sheffield faces localized vehicle crime concentration. For property investors, areas like Rotherham and Doncaster offer reduced crime risk, which may result in fewer insurance issues and greater peace of mind for residents.

Robbery Rate by District in West Yorkshire May 2022

This pie chart displays the proportion of robberies in West Yorkshire districts in May 2022 per 10,000 people.

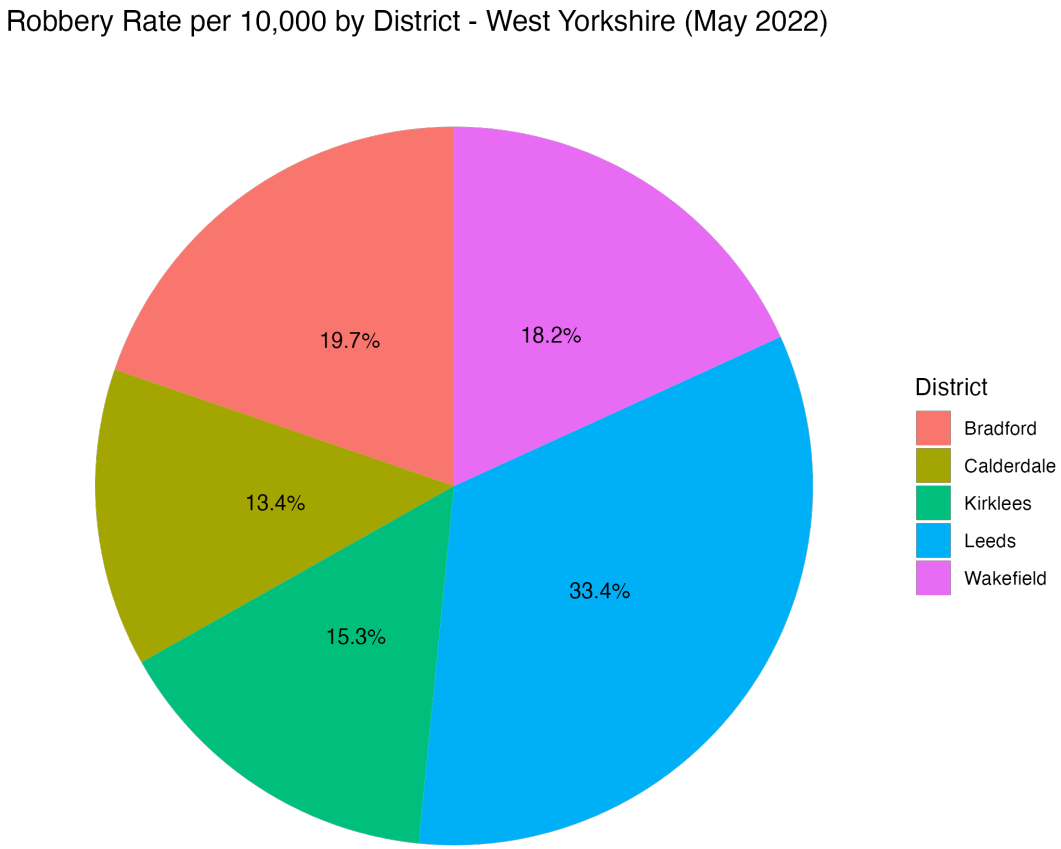

Figure 17: Robbery Rate by District

Leeds shows the highest robbery rate, with over twice Calderdale's. Bradford and Wakefield follow with moderate levels. This pattern reflects urban-driven risk in Leeds, affecting desirability. Lower rates in Calderdale and Kirklees show safer alternatives for families and conservative investors. Crime patterns influence not just safety, but long-term rental appeal.

Drug Offense Rate Trends by County from 2022 to 2024

This line graph displays drug offense levels per 10,000 people in South and West Yorkshire from 2022 to 2024.

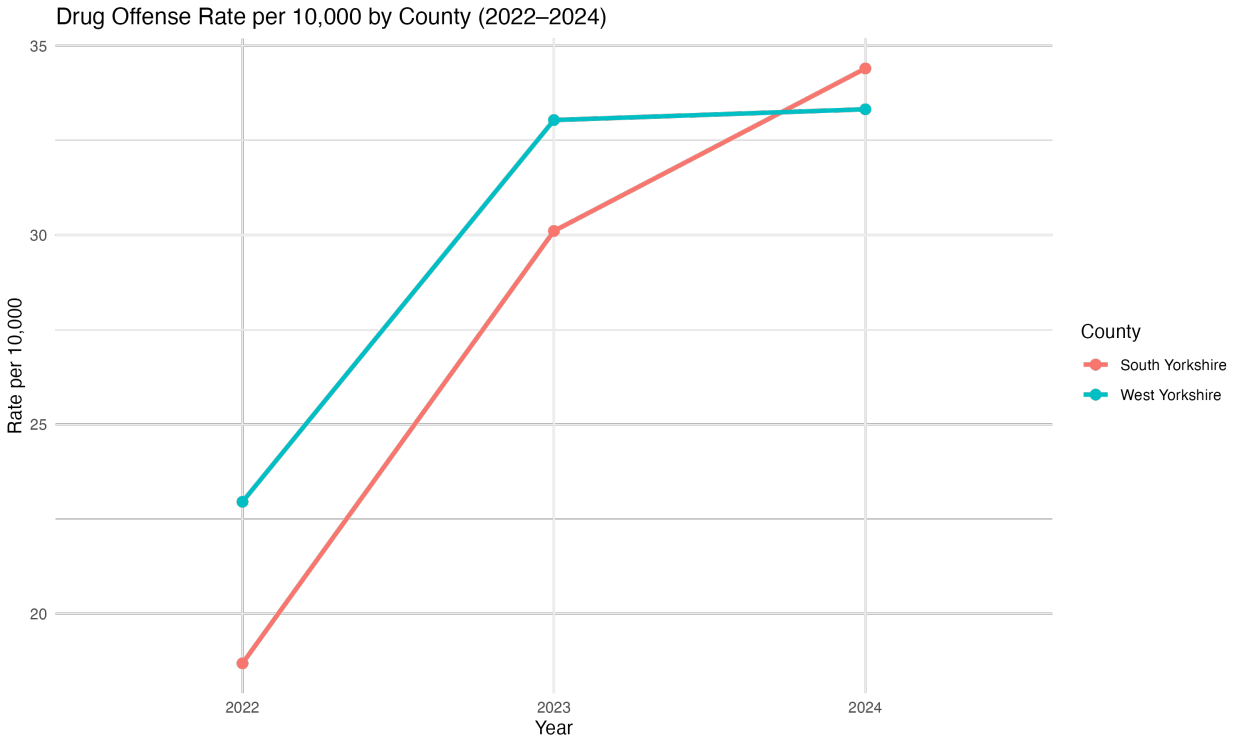

Figure 18: Drug Offense Trends (2022-2024)

Drug offenses increased in both counties. However, South Yorkshire's rise from 2022 to 2024 was faster than West Yorkshire's. While West Yorkshire has stabilized, South Yorkshire's significant increase may indicate policy changes or worsening drug issues. West Yorkshire appears more stable, while South Yorkshire demands localized research. This trend matters for crime-conscious buyers and those seeking lower-risk property investments.

Schools

This section looks at education performance across districts in South and West Yorkshire. It highlights important factors for families and property investors focused on education.

Attainment 8 Score Distribution in 2022 by District in South Yorkshire

This boxplot displays attainment 8 scores for South Yorkshire districts during the 2022 academic year.

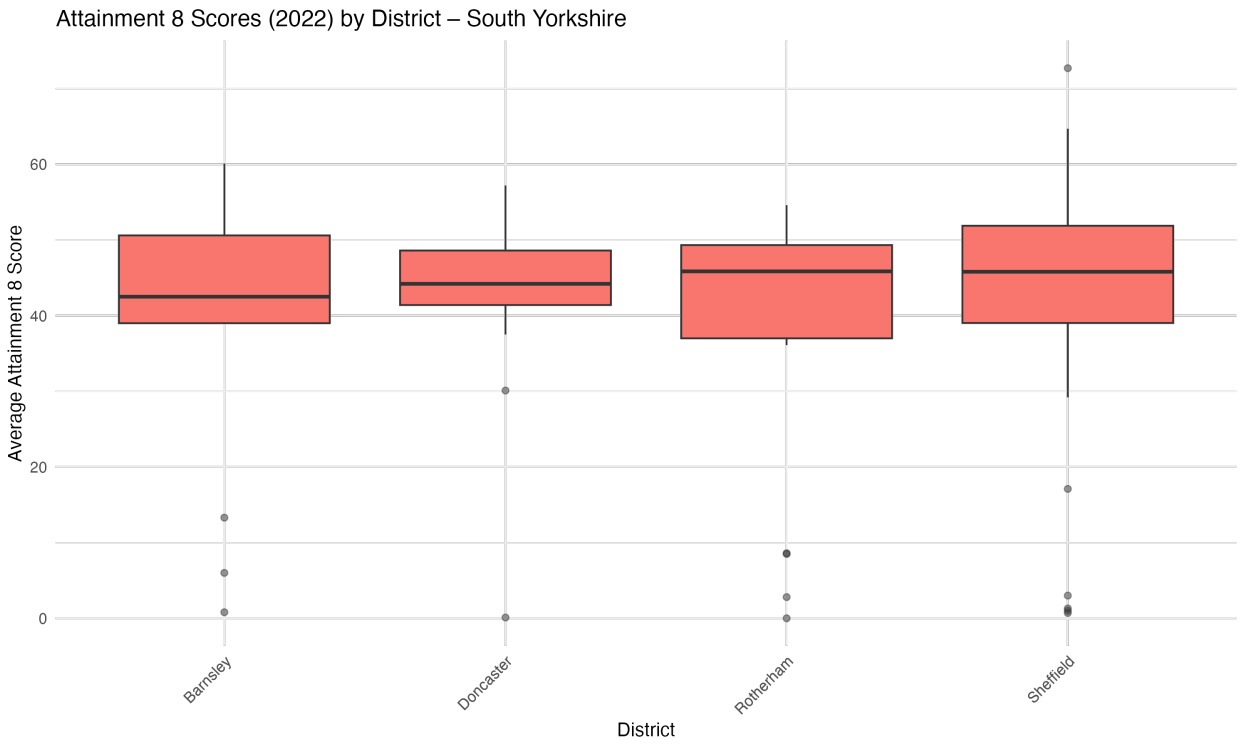

Figure 19: Attainment Scores (2022)- South

Sheffield and Doncaster are leading South Yorkshire in median attainment scores, indicative of better scholarly achievement. However, large variation throughout all districts indicates struggling schools. Barnsley and Rotherham have lower means and wider spreads, indicating deeper inequality. For families prioritizing education, Sheffield and Doncaster appear more favorable investment areas due to their stronger school performance and overall academic reliability.

Attainment 8 Score Distribution in 2022 by District in West Yorkshire

This boxplot displays attainment 8 scores for West Yorkshire districts during the 2022 academic year.

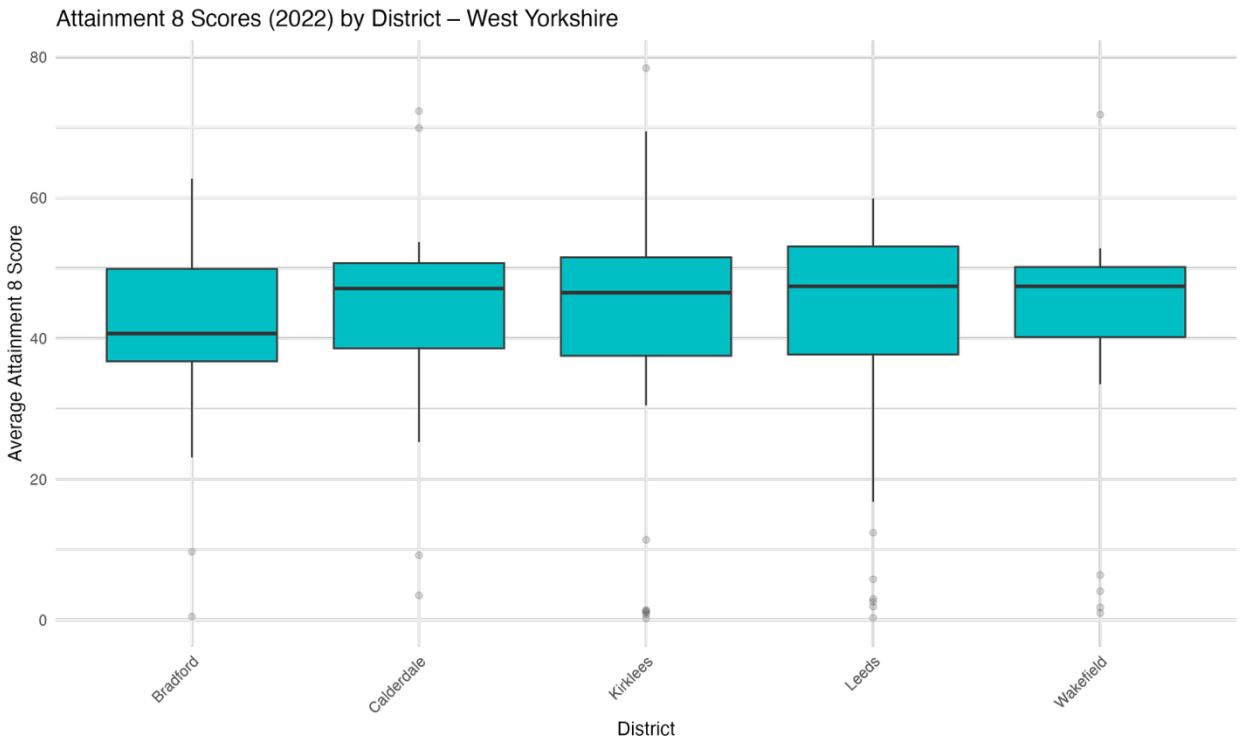

Figure 20: Attainment Scores (2022)- West

Leeds, Calderdale, and Wakefield show higher median Attainment 8 scores, outperforming many South Yorkshire districts. Despite outliers, their score distributions are more favorable. While underperforming schools exist, West Yorkshire demonstrates stronger and more consistent academic performance. This reliability makes it attractive to investors and families who prioritize educational quality when choosing long-term residential areas for stability and growth.

Attainment 8 Score Trends from 2021 to 2024 by County

This line graph displays average Attainment 8 scores from 2021 to 2024 for districts within both counties.

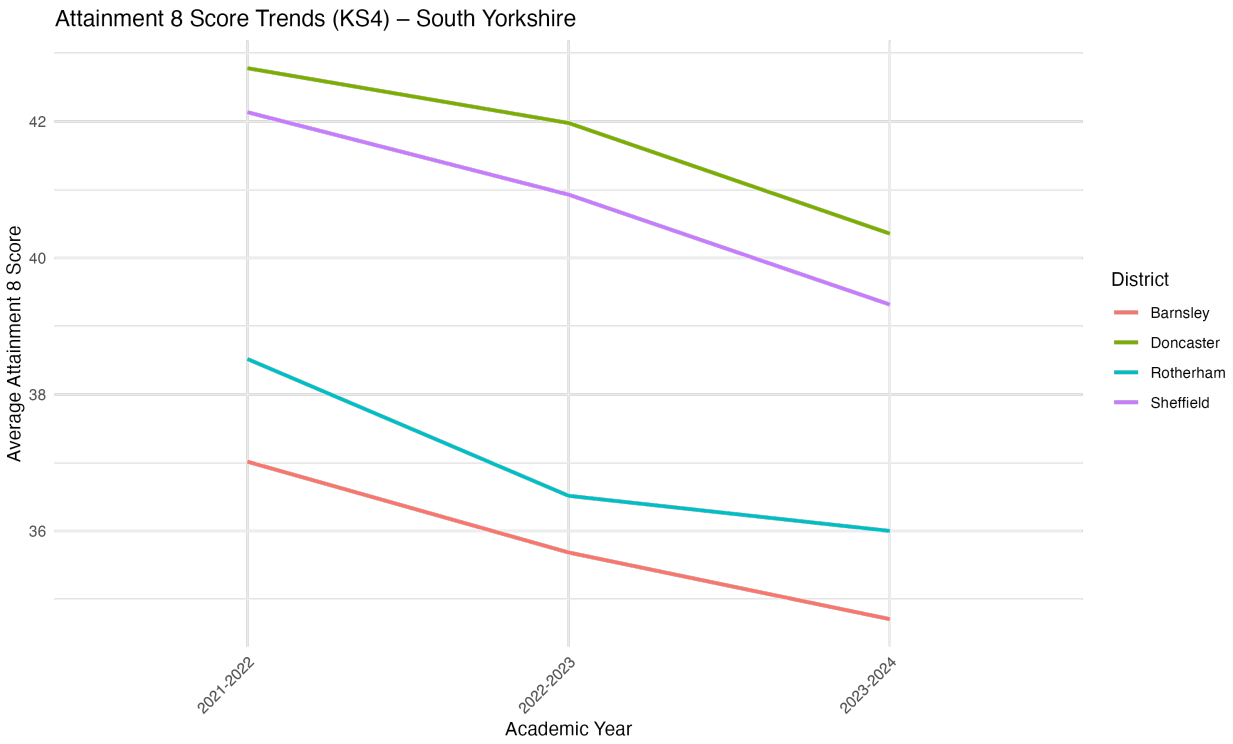

Figure 21: Attainment Scores Trend by Year- South

Most South Yorkshire districts have seen slight drops in scores, with Barnsley and Rotherham facing larger declines. Sheffield retains a relatively stable score, highlighting the county's diverse educational outcomes.

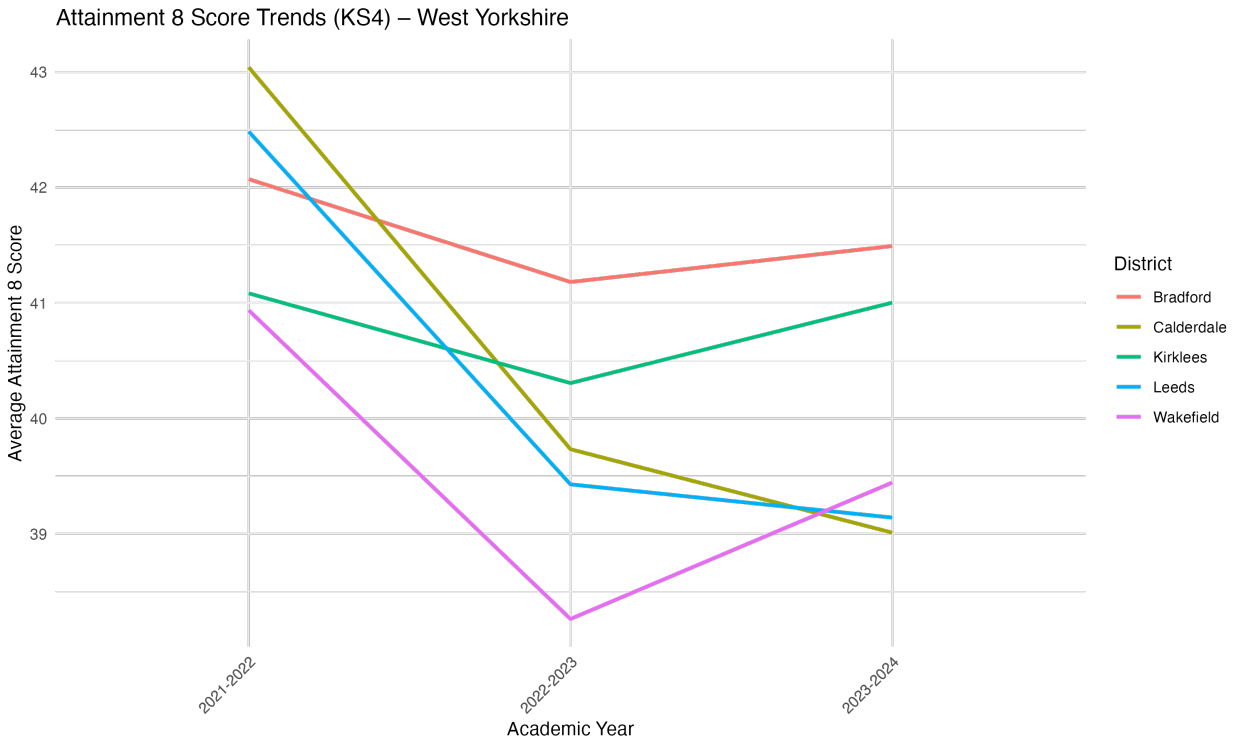

Figure 22: Attainment Scores Trend by Year- West

West Yorkshire districts are generally stable, especially Bradford and Kirklees. Leeds shows modest decline but stays above average, reflecting more resilient education performance across the county.

West Yorkshire reflects more consistent patterns of attainment, while South Yorkshire exhibits greater variation and declines. Families and investors valuing education may find West Yorkshire's stability more appealing for long-term prospects.

Linear Modeling

Linear modeling is a statistical method used to find the relationship between two variables by fitting a linear equation to the data (Beers, 2025). In this analysis, datasets for South Yorkshire and West Yorkshire were aggregated by county and short postcode, then merged to enable comparison across regions. Using R and the tidyverse packages, linear regression models, correlation coefficients, and scatterplots with regression lines were created to explore connections between broadband speed, house price, crime rate, and academic performances.

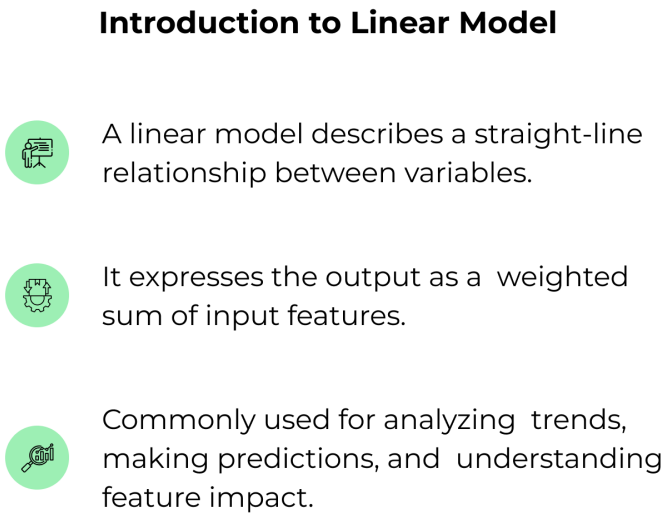

Figure 23: Linear Model

Broadband Speed and House Prices

This model looks at the relationship between average broadband speed and house prices in South and West Yorkshire districts.

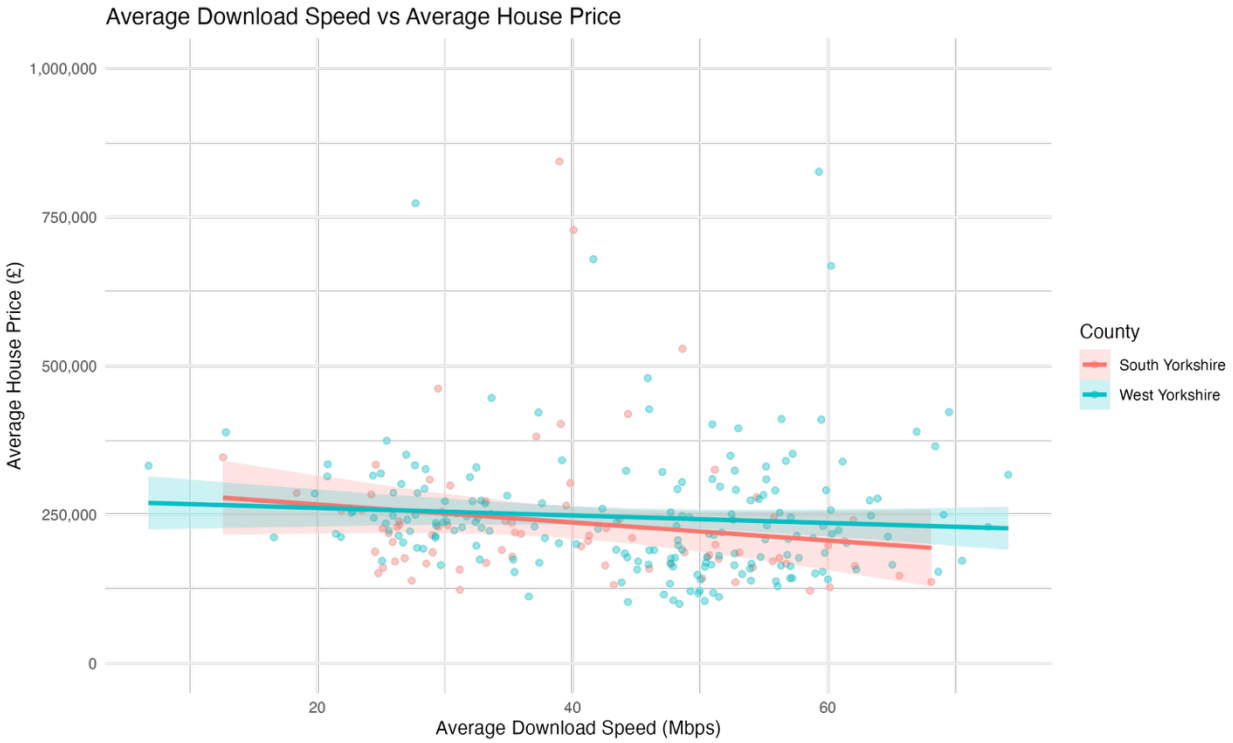

Figure 24: House Price vs Download Speed

```
--- Linear Model Summary: South Yorkshire ---
> print(summary(lm_south))

Call:
lm(formula = Avg_House_Price ~ Avg_Download_Speed, data = filter(merged_data,
    County == "South Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-191445 -100852  -57100  -10598 2763169

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         458999     123100   3.729 0.000366 ***
Avg_Download_Speed    -4699       2965  -1.585 0.117119
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 340400 on 77 degrees of freedom
Multiple R-squared:  0.03159,   Adjusted R-squared:  0.01901
F-statistic: 2.511 on 1 and 77 DF,  p-value: 0.1171
```

```
--- Linear Model Summary: West Yorkshire ---
> print(summary(lm_west))

Call:
lm(formula = Avg_House_Price ~ Avg_Download_Speed, data = filter(merged_data,
    County == "West Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-143490  -66843  -20332   42657  590058

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       273937.8    26132.0  10.483   <2e-16 ***
Avg_Download_Speed   -633.5      560.2  -1.131     0.26
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105800 on 188 degrees of freedom
Multiple R-squared:  0.006754,  Adjusted R-squared:  0.001471
F-statistic: 1.278 on 1 and 188 DF,  p-value: 0.2596
```

```
--- Correlations ---
> cat("South Yorkshire Correlation:", round(cor_south, 3), "\n")
South Yorkshire Correlation: -0.178
> cat("West Yorkshire Correlation:", round(cor_west, 3), "\n")
West Yorkshire Correlation: -0.082
>
```

The regression plot shows a weak negative correlation between average download speed and house prices in the two counties. In South Yorkshire, higher broadband speed has a weak correlation with lower house prices (r = -0.178, p = 0.117), while West Yorkshire has an even weaker inverse relationship (r = -0.082, p = 0.260). Linear models confirm no statistically significant association.

Drug Offense Rate and House Prices

This model explores the linear relationship between drug offense rates per 10,000 people and average house prices in South and West Yorkshire districts.

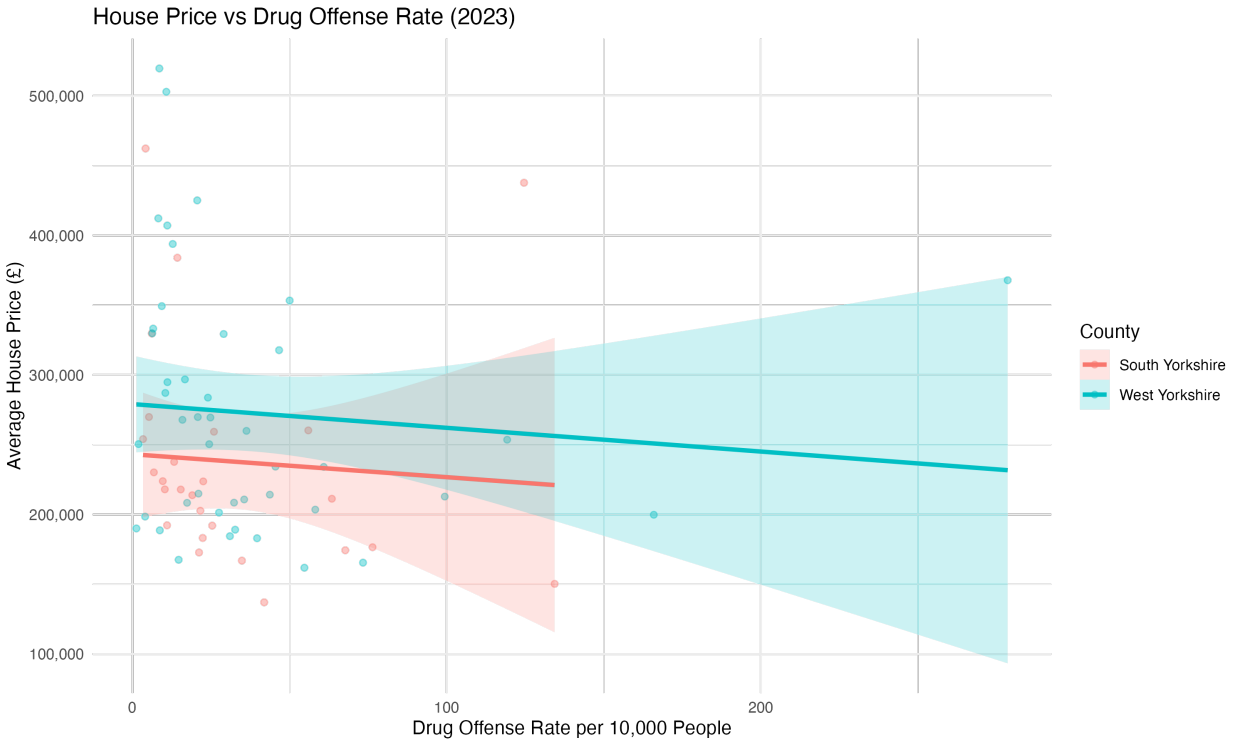

Figure 25: House Price vs Drug Rate (2023)

```
--- Linear Model Summary: South Yorkshire ---
> print(summary(lm_south))

Call:
lm(formula = Avg_Price ~ Drug_Rate_per_10000, data = filter(merged_data,
    County == "South Yorkshire"))

Residuals:
   Min     1Q Median     3Q    Max
-99230 -52824 -22104  18141 219771

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       243172.7    22712.6  10.706 1.27e-10 ***
Drug_Rate_per_10000   -164.4      477.5  -0.344    0.734
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83500 on 24 degrees of freedom
Multiple R-squared:  0.004912,  Adjusted R-squared:  -0.03655
F-statistic: 0.1185 on 1 and 24 DF,  p-value: 0.7337
```

```
--- Linear Model Summary: West Yorkshire ---
> print(summary(lm_west))

Call:
lm(formula = Avg_Price ~ Drug_Rate_per_10000, data = filter(merged_data,
    County == "West Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-109076  -66234  -18783   52626  241984

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         279101.2    17234.4  16.194   <2e-16 ***
Drug_Rate_per_10000   -170.0      279.4  -0.608    0.546
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90050 on 42 degrees of freedom
Multiple R-squared:  0.008738,  Adjusted R-squared:  -0.01486
F-statistic: 0.3702 on 1 and 42 DF,  p-value: 0.5462
```

```
--- Correlations ---
> cat("South Yorkshire Correlation:", round(cor_south, 3), "\n")
South Yorkshire Correlation: -0.07
> cat("West Yorkshire Correlation:", round(cor_west, 3), "\n")
West Yorkshire Correlation: -0.093
>
```

The scatterplot with regression lines suggests a very weak negative relationship between drug offense rates and mean house prices in the two counties. In South Yorkshire, the correlation is minimal (r = -0.070, p = 0.734), and West Yorkshire has an equally weak trend (r = -0.093, p = 0.546). Linear regression results indicate no statistically significant association between the two variables.

Attainment 8 Scores and House Prices

This model assesses the relationship between average school Attainment 8 scores and house prices across districts in South and West Yorkshire.

Figure 26: Attainment vs House Price

```
--- Linear Model Summary: South Yorkshire ---
> print(summary(lm_south))

Call:
lm(formula = Avg_Price ~ Avg_Attainment8, data = filter(merged_sampled,
    County == "South Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-106243  -37701  -15746   20592  324330

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       139023      65054   2.137   0.0383 *
Avg_Attainment8     1826       1528   1.195   0.2386
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76990 on 43 degrees of freedom
Multiple R-squared:  0.03215,   Adjusted R-squared:  0.00964
F-statistic: 1.428 on 1 and 43 DF,  p-value: 0.2386
```

```
--- Linear Model Summary: West Yorkshire ---
> print(summary(lm_west))

Call:
lm(formula = Avg_Price ~ Avg_Attainment8, data = filter(merged_sampled,
    County == "West Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-170887  -66376  -20362   50837  790131

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     239331.8    23213.7  10.310   <2e-16 ***
Avg_Attainment8    108.2      543.7   0.199    0.842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 106600 on 258 degrees of freedom
Multiple R-squared:  0.0001534, Adjusted R-squared:  -0.003722
F-statistic: 0.03957 on 1 and 258 DF,  p-value: 0.8425
```

```
--- Correlations ---
> cat("South Yorkshire Correlation:", round(cor_south, 3), "\n")
South Yorkshire Correlation: 0.179
> cat("West Yorkshire Correlation:", round(cor_west, 3), "\n")
West Yorkshire Correlation: 0.012
>
```

The scatterplots and regression lines reveal almost no correlation in West Yorkshire (r = 0.012, p = 0.84), and a very slight positive correlation in South Yorkshire (r = 0.179, p = 0.24). Linear models indicate these associations are not significant, showing little direct connection between housing prices in these areas and school performance.

Attainment 8 Scores and Drug Offense Rate

This model looks at the relationship between drug offense rates per 10,000 people in South and West Yorkshire and average Attainment 8 school scores.

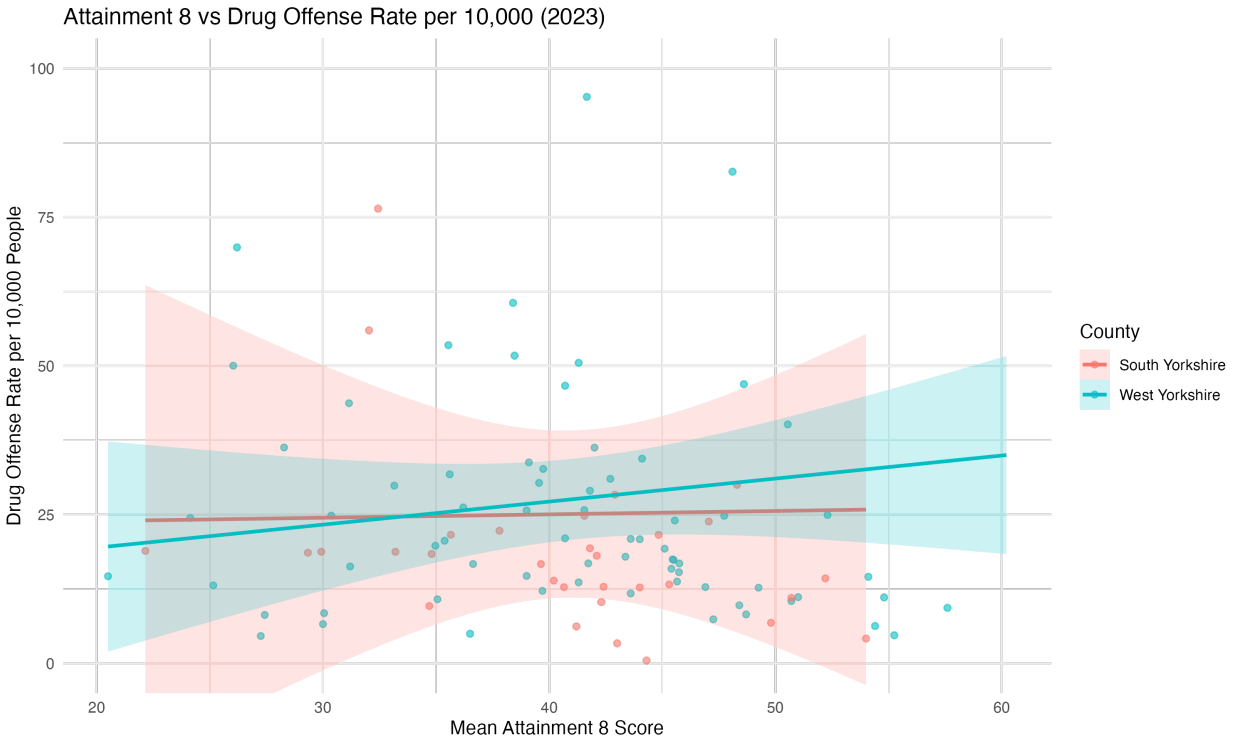

Figure 27: Attainment vs Drug Rate (2023)

```
South Yorkshire Linear Model Summary:
> print(summary(model_south))

Call:
lm(formula = drug_rate_per_10k ~ mean_attainment8, data = filter(combined_data,
    County == "South Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-24.822 -13.045  -6.716  -3.048 193.286

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.7739    40.1131   0.568    0.574
mean_attainment8  0.0567     0.9672   0.059    0.954

Residual standard error: 38.82 on 30 degrees of freedom
Multiple R-squared:  0.0001145, Adjusted R-squared:  -0.03321
F-statistic: 0.003436 on 1 and 30 DF,  p-value: 0.9536
```

```
West Yorkshire Linear Model Summary:
> print(summary(model_west))

Call:
lm(formula = drug_rate_per_10k ~ mean_attainment8, data = filter(combined_data,
    County == "West Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-28.373 -14.892  -7.497   5.613 177.470

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      11.6997    16.6623   0.702    0.485
mean_attainment8  0.3872     0.3974   0.974    0.333

Residual standard error: 28.7 on 69 degrees of freedom
Multiple R-squared:  0.01357,   Adjusted R-squared:  -0.0007284
F-statistic: 0.949 on 1 and 69 DF,  p-value: 0.3334
```

```
Correlation coefficients:
> cat("South Yorkshire:", round(cor_south, 3), "\n")
South Yorkshire: 0.011
> cat("West Yorkshire:", round(cor_west, 3), "\n")
West Yorkshire: 0.116
>
```

The scatterplot with regression lines shows a near-zero relationship between drug offense rates and school performance. South Yorkshire exhibits near zero relationship (r = 0.011, p = 0.954), while West Yorkshire has a very weak positive but statistically insignificant relationship (r = 0.116, p = 0.333). Linear models indicate no significant relationship in either county, suggesting that school achievement has little to no direct connection to drug offenses in both regions.

Average Download Speed and Drug Offense Rate

This model explores the relationship between average broadband download speed and drug offense rates per 10,000 people in South Yorkshire and West Yorkshire.

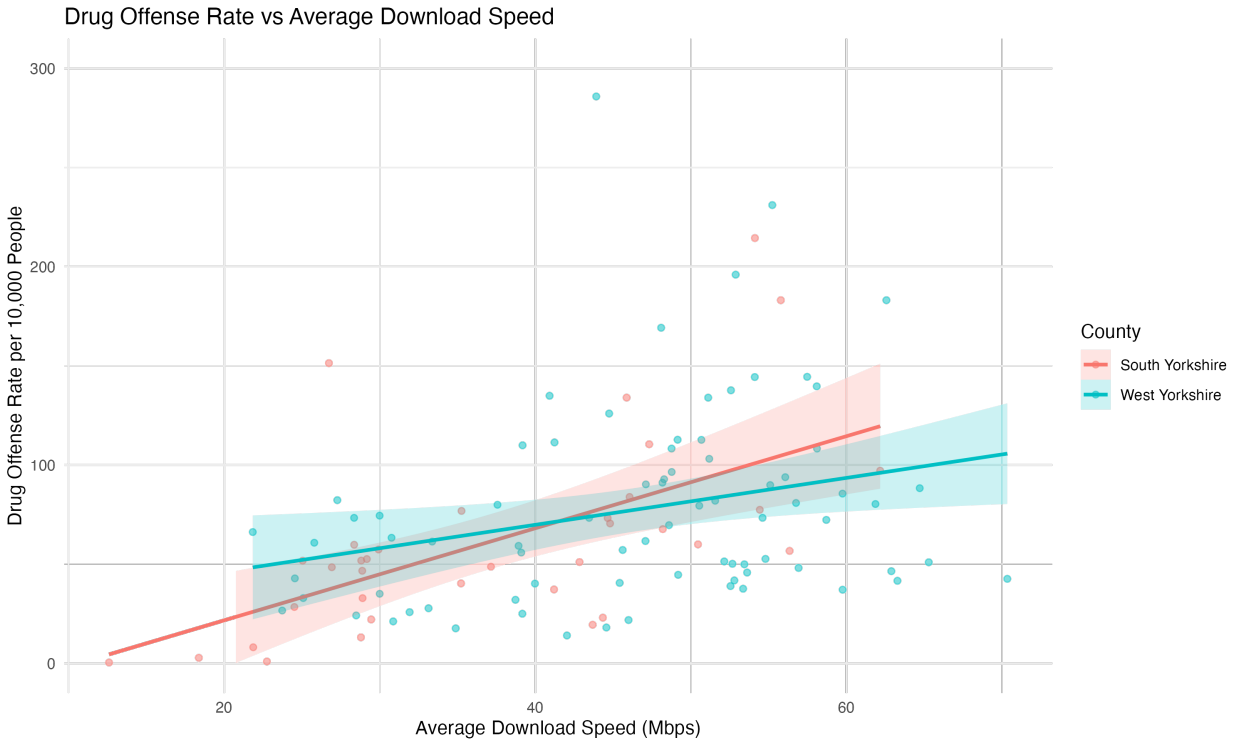

Figure 28: Download Speed vs Drug Rate

```
South Yorkshire Linear Model Summary:
> print(summary(model_south))

Call:
lm(formula = DrugRatePer10000 ~ AvgDownloadSpeed, data = filter(combined_data,
    County == "South Yorkshire"))

Residuals:
   Min     1Q Median     3Q    Max
-91.18 -38.45 -14.77   6.16 580.92

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -57.086     59.258  -0.963    0.342
AvgDownloadSpeed   3.637      1.503   2.420    0.021 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.7 on 34 degrees of freedom
Multiple R-squared:  0.147,     Adjusted R-squared:  0.1219
F-statistic: 5.857 on 1 and 34 DF,  p-value: 0.021
```

```
West Yorkshire Linear Model Summary:
> print(summary(model_west))

Call:
lm(formula = DrugRatePer10000 ~ AvgDownloadSpeed, data = filter(combined_data,
    County == "West Yorkshire"))

Residuals:
   Min     1Q Median     3Q    Max
-354.3 -177.6  -53.8   25.2 3215.0

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       608.078    225.355   2.698   0.0085 **
AvgDownloadSpeed   -9.570      4.731  -2.023   0.0465 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 495.9 on 80 degrees of freedom
Multiple R-squared:  0.04865,   Adjusted R-squared:  0.03676
F-statistic: 4.091 on 1 and 80 DF,  p-value: 0.04645
```

```
Correlation Coefficients:
> cat("South Yorkshire:", round(cor_south, 3), "\n")
South Yorkshire: 0.383
> cat("West Yorkshire:", round(cor_west, 3), "\n")
West Yorkshire: -0.221
>
```

The scatterplot with regression lines indicates opposite trends between the two counties. In South Yorkshire, there is a positive correlation of modest strength between download speeds and increasing rates of drug offenses ($r = 0.383$, $p = 0.021$). In West Yorkshire, there is a weak negative relationship ($r = -0.221$, $p = 0.046$), suggesting areas with higher internet speeds have lower drug crime rates. Both correlations are significant but have low explanatory value.

Average Download Speed and Attainment 8 Scores

This graph illustrates the relationship between average broadband download speed and Attainment 8 school scores in South Yorkshire and West Yorkshire.

Figure 29: Download Speed vs Attainment

```
South Yorkshire Linear Model Summary:
> print(summary(model_south))

Call:
lm(formula = Avg_Attainment8 ~ Avg_Download_Speed, data = filter(combined_data,
    County == "South Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-33.067  -4.423   0.795   7.191  15.310

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        56.9712     4.5507   12.52  < 2e-16 ***
Avg_Download_Speed  -0.4340     0.1069   -4.06 0.000166 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.714 on 52 degrees of freedom
Multiple R-squared:  0.2407,    Adjusted R-squared:  0.2261
F-statistic: 16.49 on 1 and 52 DF,  p-value: 0.0001655
```

```
West Yorkshire Linear Model Summary:
> print(summary(model_west))

Call:
lm(formula = Avg_Attainment8 ~ Avg_Download_Speed, data = filter(combined_data,
    County == "West Yorkshire"))

Residuals:
    Min      1Q  Median      3Q     Max
-39.844  -3.796   1.890   8.527  31.484

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       44.80200    4.97783    9.00 8.11e-15 ***
Avg_Download_Speed -0.09314    0.10585   -0.88    0.381
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.43 on 109 degrees of freedom
Multiple R-squared:  0.007054,  Adjusted R-squared:  -0.002056
F-statistic: 0.7743 on 1 and 109 DF,  p-value: 0.3808
```

```
Correlation Coefficients:
> cat("South Yorkshire:", round(cor_south, 3), "\n")
South Yorkshire: -0.491
> cat("West Yorkshire:", round(cor_west, 3), "\n")
West Yorkshire: -0.084
>
```

The model reveals a negative correlation in South Yorkshire ($r = -0.491$, $p < 0.001$), indicating higher download speeds are associated with slightly lower attainment scores, possibly reflecting complex socio-economic factors. In West Yorkshire, there is no correlation ($r = -0.084$, $p = 0.38$), suggesting that factors other than connectivity affect school performance.

Recommendation System

Overview

A basic recommendation system was built to guide property investment in South and West Yorkshire using UK government datasets. Data on housing prices, internet speed, crime rates, and school performance were cleaned and merged by short postcode. Key indicators were scored from 0 to 10, and an aggregate score was calculated. Towns were then ranked for the purpose of identifying the top 10 areas with the best combination of affordability, connectivity, safety, and school quality.

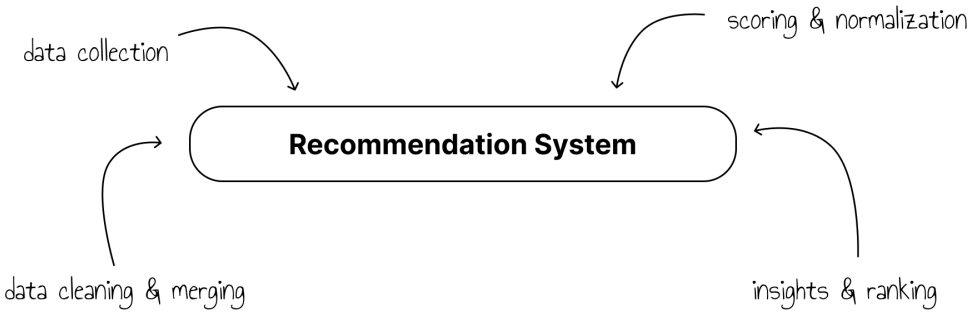

Figure 30: Recommendation System

Results

The recommendation system ranked towns by affordability, broadband speed, safety, and school quality. The following table illustrates the top towns and their scores, exhibiting a balance among these primary factors to consider when making property investment decisions.

```
78  # Calculate scores (0–10 scale)
79  scored_towns = town_scores_data %>%
80    mutate(
81      PriceScore = 10 * (1 - (rank(AvgPrice) / n())),
82      BroadbandScore = 10 * (rank(AvgDownloadSpeed) / n()),
83      CrimeScore = 10 * (1 - (rank(CrimeRate) / n())),
84      SchoolScore = 10 * (rank(AvgAttainment) / n()),
85      OverallScore = round((PriceScore + BroadbandScore + CrimeScore + SchoolScore) / 4, 2)
86    )
87
88  # Aggregate and rank recommended towns
89  recommended_towns = scored_towns %>%
90    group_by(Town, County) %>%
91    summarise(
92      AffordabilityScore = mean(PriceScore, na.rm = TRUE),
93      BroadbandScore = mean(BroadbandScore, na.rm = TRUE),
94      SafetyScore = mean(CrimeScore, na.rm = TRUE),
95      SchoolQualityScore = mean(SchoolScore, na.rm = TRUE),
96      OverallScore = mean(OverallScore, na.rm = TRUE)
97    ) %>%
98    ungroup() %>%
99    arrange(desc(OverallScore)) %>%
97:8  (Top Level)                                                    R Script
```

	Rank	Town	County	AffordabilityScore	SafetyScore	BroadbandScore	SchoolQualityScore	OverallScore
1	1	Mexborough	South Yorkshire	9.35	5.32	7.10	6.13	6.98
2	2	Heckmondwike	West Yorkshire	8.71	6.61	7.58	3.55	6.61
3	3	Bingley	West Yorkshire	3.87	6.13	5.97	8.06	6.01
4	4	Ilkley	West Yorkshire	0.16	8.71	4.84	9.84	5.89
5	5	Barnsley	South Yorkshire	7.15	5.73	4.33	5.51	5.68
6	6	Castleford	West Yorkshire	6.77	3.55	5.81	6.45	5.65
7	7	Pudsey	West Yorkshire	3.71	8.39	7.74	2.10	5.48
8	8	Rotherham	South Yorkshire	7.53	5.22	5.56	2.89	5.30
9	9	Otley	West Yorkshire	1.94	9.03	0.65	9.52	5.28
10	10	Doncaster	South Yorkshire	7.82	1.94	3.55	7.50	5.20

Showing 1 to 10 of 10 entries, 8 total columns

Reflection

This section summarizes key insights from analyzing towns in South and West Yorkshire. By examining broadband speeds, school performance, house prices, crime rates, and overall scores, it highlights elements that influence cost, connectivity, safety, and academic performance. These observations help make data-driven real estate investment choices based on lifestyle, education, and budget considerations.

Broadband Speeds

Scores for broadband speed varied widely among towns. Pudsey, Heckmondwike, and Mexborough performed well, offering relatively faster download speeds, which is vital for remote work and digital connectivity. However, towns like Otley and Doncaster scored lower, indicating limited internet performance. This suggests that while some affordable areas offer good broadband, others lag behind, highlighting the necessity for investors to consider connectivity in addition to other aspects of quality of life when making choices.

School Grades

School quality scores showed high educational attainment in towns like Ilkley, Otley, and Bingley, which had high mean Attainment 8 scores. However, more affordable towns such as Heckmondwike and Rotherham scored significantly lower in this area. This contrast illustrates a trade-off between affordability and academic excellence. For investors prioritizing family appeal or long-term rental demand, proximity to high-performing schools could be a decisive factor, even at the cost of higher property prices.

House Prices

The affordability analysis revealed that towns with much lower average home prices, like Mexborough, Heckmondwike, and Doncaster, appeal to investors on a budget. More desirable towns such as Ilkley and Otley, in contrast, scored poorly on affordability due to higher property prices. This highlights a clear trade-off: premium towns have better amenities and schools, but at a premium price, so investors need to balance budget considerations against longer-term value.

Crimes

Crime rate analysis showed significant differences among towns. The areas with the greatest safety scores were Ilkley, Otley, and Pudsey, suggesting comparatively low crime rates and providing prospective investors with peace of mind. Towns with lower safety ratings, such as Doncaster and Castleford, on the other hand, can be less desirable. This highlights that while affordability may attract investors, safety is still a crucial element affecting long-term investment stability and quality of life for residents.

Overall Score

The overall scores show a mixed set of strengths across South and West Yorkshire towns. Mexborough ranks highest with strong affordability and broadband, while Ilkley and Otley excel in school quality and safety despite lower affordability. No town dominates in all categories, reflecting the multifaceted nature of investment decisions. This balanced approach allows international investors to align property choices with certain goals, such as affordability, education, or lifestyle quality.

Legal and Ethical Issues

In conducting this analysis, legal and ethical considerations were carefully addressed. All datasets were sourced from official UK government sources, showing compliance with data licensing and copyright restrictions. No personally identifiable information was included, upholding privacy and data protection standards according to the UK Data Protection Act and GDPR (Government Digital Service, 2025). Ethically, care was taken to avoid bias in data interpretation, especially when comparing regions. Data cleaning and modeling were done transparently to prevent misinterpretation. Visualizations and summaries were designed to inform, not to stigmatize or mislead stakeholders about any specific area. Moreover, ethical responsibility includes recognizing the limitations of data, not overgeneralizing, and not making final property investment decisions solely based on these results. This project is meant to support decision-making with factual, objective analysis while ensuring fairness, transparency, and respect for the communities represented in the data.

Conclusion

This project applied the complete data science lifecycle to evaluate property investment potential across South Yorkshire and West Yorkshire. Using only UK government-published datasets, key indicators such as housing affordability, internet speed, safety, and academic performance were thoroughly examined. Throughout the project, data was collected, cleaned, visualized, and modeled in R to ensure transparency and reproducibility. Exploratory analysis showed differences in average house prices, broadband speeds, crime rates, and school performance across regions. Linear modelling provided insights into potential relationships among these variables. A recommendation system ranked towns based on combined criteria, giving a structured overview of investment suitability. While the results support data-informed decision-making, they must be interpreted with consideration for data limitations and external contextual factors. The study highlights the power of open data and statistical methods in answering real questions and enables objective comparisons of geographic areas using measurable indicators.

References

Price paid data. (2025, July 28). GOV.UK. https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads

Ofcom. (2023, September 29). Data downloads. Ofcom. https://www.ofcom.org.uk/phones-and-broadband/coverage-and-speeds/data-downloads

Data downloads|data.police.uk. (n.d.). https://data.police.uk/data/

Compare school performance. (n.d.). Compare School and College Performance in England. https://www.compare-school-performance.service.gov.uk/download-data

GeeksforGeeks. (2025, July 23). What is Data Cleaning? GeeksforGeeks. https://www.geeksforgeeks.org/data-analysis/what-is-data-cleaning/

GeeksforGeeks. (2025b, July 28). What is Exploratory Data Analysis? GeeksforGeeks. https://www.geeksforgeeks.org/data-analysis/what-is-exploratory-data-analysis/

Beers, B. (2025, May 31). Regression: definition, analysis, calculation, and example. Investopedia. https://www.investopedia.com/terms/r/regression.asp

Government Digital Service. (2025, January 22). Data protection. GOV.UK. https://www.gov.uk/data-protection

Appendix

GitHub: https://github.com/shresthabatsal/Coursework_BatsalShrestha_230509

Datasets: https://drive.google.com/drive/folders/1qg4DGWxGcSPD_PFUCbGWTbnM_y10p4q4?usp=sharing