

- 1) Introduction
- 2) Data Explanation and Exploration
- 3) Statistical Analysis and Interpretation
- 4) Conclusions
- 5) References

# Final\_Project

Code ▾

Olivia Nguyen, Dionne Phan

2024-11-19

Show

```
##  
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':  
##  
## filter, lag
```

```
## The following objects are masked from 'package:base':  
##  
## intersect, setdiff, setequal, union
```

Show

```
## Warning: package 'caret' was built under R version 4.4.2
```

```
## Loading required package: lattice
```

Show

```
##  
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':  
##  
## src, summarize
```

```
## The following objects are masked from 'package:base':  
##  
## format.pval, units
```

Show

```
##  
## Attaching package: 'xtable'
```

```
## The following objects are masked from 'package:Hmisc':  
##  
## label, label<-
```

# 1) Introduction

Education is a cornerstone for social mobility and economic development. However, regional inequalities persist in England, creating barriers to equal opportunities. Studies show that regions like London outperform others, such as the North East and coastal areas, in both early years and later educational stages. For instance, London has the highest rate of HE progression, while rural and coastal areas face challenges like limited resources, fewer role models, and economic constraints (Playford et al., 2023; Dunatchik et al., 2018).

City size also plays a crucial role. Smaller towns often produce a higher share of graduates relative to their population but struggle to retain this talent, while large urban centers, such as London and Manchester, attract graduates but exhibit broader disparities within their population (ONS, 2024). Despite these findings, the interaction between regional and urban characteristics and their combined impact on educational attainment remains underexplored.

Educational attainment shows a strong North-South divide in England, including London having major improvements within early years development and HE progression compared to the South West and North East (Dunatchik et al., 2018). City size also plays a critical role because smaller towns usually have higher numbers of graduates in comparison to their population. However, most of these graduates often move to cities like London and Manchester for better prospects, which adds to the discrepancy in talent retention rates within regions. Moreover, socioeconomic indicators such as income deprivation, job density, and proximity to universities are strong predictors that determine educational inequalities across regions and further heighten the disadvantaged position of certain areas, according to ONS (2024) and Playford et al. (2023).

This study tries to fill in some of the gaps in prior research by exploring **longitudinal effects of attainment disparities from early years, Key Stage 2, through to HE progression**, investigating **how urbanization patterns, including university presence and job opportunities, affect their decision in Higher Education and how these things might mitigate or exacerbate these inequalities**. Thus, the study shall try to reach deeper insights of the influence that good early educational attainment, geographic and socio-economic settings have on school performance. Specifically, this study asks the following questions:

- 1. Do early academic achievements influence students' ability to maintain their advantage in later stages, and how do teenage or earlier successes impact their decision to pursue higher education at age 19?*
- 2. How does urbanization shape students' decisions to pursue higher education and affect educational inequalities?*

Better understanding of the predictive relationship between early performance and later outcomes allows for proper guidance of children in the appropriate direction right from the foundation years of a child by both the parents and educators. Recognition of long-lasting effects stemming from the earliest accomplishments of early years could secure the resources and motivation children need for a bright future. Early intervention can act as the stepping stone, which helps the child develop the necessary capabilities, confidence, and motivation to succeed in later periods of education.

This insight also allows parents, educators, and policymakers to play a pivotal role in addressing significant educational disparities across regions and city sizes in England. By exploring how early academic achievements, urbanization patterns, and socio-economic factors influence educational attainment, the study aims to provide a deeper understanding of the factors contributing to inequalities. The findings can be used to inform targeted interventions to ensure equal opportunities for all students, which will ultimately promote social mobility and reduce disparities across regions.

In this analysis, we will first use descriptive statistics to compare the differences in the proportion of students in the region who have Key Stage 2, Key Stage 4, and Level 6 qualifications or higher across all regions, using statistical measures such as mean, median, and standard deviation. After that, we will separate the data by region to examine if there are any significant differences. We will also compare these statistics under different urbanization conditions. Then, we will use regression analysis and hypothesis testing to determine which factors contribute to educational attainment.

## 2) Data Explanation and Exploration

### 2.1 Dataset Overview:

We constructed our database from UK's Office for National Statistics (ONS), which is the recognized national statistical institute of the United Kingdom. In particular, the ONS collects information through survey and non-survey sources to generate official statistics relevant to the United Kingdom at the national, regional, and local levels for the economy, population, and society. The data for the particular set "Educational Attainment of Young People in English Towns", collected in July 2023, was directly collected from individuals, households, public bodies, and businesses by the Office of National Statistics, with a view on minimizing the burden on respondents.

The dataset we are analyzing contains 1,104 observations and 31 variables, with each row representing a town or city in England. This data comes from the 2011 Census and school records from the 2012–2013 period. It provides detailed insights into various aspects of these towns and cities, including demographics, geography, economy, and education. The variables within the dataset range from population size and coastal

status to income levels, job density, the presence of universities, and educational outcomes at different ages.

## 2.2 Key Variables:

The dataset contains variables relevant to geographic regions, socioeconomic conditions, and educational outcomes. Key variables is included in following table:

Show

| Variable_Name   | Units                | Data_Type | Description  |
|---|----------------------|-----------|--|
| town11cd  | N/A                  | character | Town/city geography code (2011)  |
| size_flag   | N/A                  | character | Size category of the built-up area or built-up area subdivision based on resident population (from Census 2011)                          |
| job_density_flag  | N/A                  | character | Variable used to describe towns as working, residential or mixed.  |
| rgn11nm   | N/A                  | character | English region name  |
| income_flag   | N/A                  | character | Variable used to describe towns as lower income deprivation, mid income deprivation or higher income deprivation                         |
| university_flag   | N/A                  | character | Variable used to describe whether the town/city has a university   |
| key_stage_2_attainment_school_year_2007_to_2008                   | Proportion (%)       | numeric   | Proportion of pupils that achieved level 4 or above (expected level) in key stage 2 in English and Maths in the 2007 to 2008 school year |
| key_stage_4_attainment_school_year_2012_to_2013                   | Proportion (%)       | numeric   | Proportion of pupils that achieved 5 GCSE or more, including English and Maths, with grades A*-C in the 2012 to 2013 school year         |
| activity_at_age_19_full_time_higher_education                     | Proportion (%)       | numeric   | Proportion of the town/city's 2012/13 key stage 4 cohort in full time higher education at the age 19.                                    |
| highest_level_qualification_achieved_by_age_22_less_than_level_1  | Proportion (%)       | numeric   | Proportion of the town/city's 2012/13 key stage 4 cohort with less than a Level 1 qualification at age 22.                               |
| highest_level_qualification_achieved_by_age_22_level_1_to_level_2 | Proportion (%)       | numeric   | Proportion of the town/city's 2012/13 key stage 4 cohort with a level 1 or level 2 qualification at age 22.                              |
| highest_level_qualification_achieved_by_age_22_level_3_to_level_5 | Proportion (%)       | numeric   | Proportion of the town/city's 2012/13 key stage 4 cohort with level 3, level 4 or level 5 qualification at age 22.                       |
| highest_level_qualification_achieved_by_age_22_level_6_or_above   | Proportion (%)       | numeric   | Proportion of the town/city's 2012/13 key stage 4 cohort with level 6 or above qualification at age 22.                                  |
| highest_level_qualification_achieved_b_age_22_average_score       | Average Score (0-10) | numeric   | Town/city highest qualification average score based on highest levels of qualifications achieved of the 2012/13 KS4 cohort.              |
| education_score   | Score ((-10)-10)     | numeric   | Town/city education score based on attainment levels of the 2012/13 Key stage 4 cohort.  |

Introduce Educational System in England included in Project

Key Stage 2 (KS2) is for pupils aged 7–11, assessed through SATs in English and Maths at the end of Year 6. Key Stage 4 (KS4) covers ages 14–16, where students take GCSE exams in core and optional subjects at the end of Year 11, determining their progression to further education. A Level 6 qualification is equivalent to a bachelor's degree, indicating advanced knowledge and skills in a specific field, typically pursued after completing school education.

## 2.3 Data Cleaning:

Because original dataset contains 31 variable but we do not use all of them. So, we will create a new dataframe called *english\_education* selecting variable that we will use for analysis.

[Show](#)

### 2.3.1 Converting all categorical variables to factors for consistency in analysis.

In this step, we will create a new variable called *small\_towns* based on the *size\_flag*. If the size flag indicates small towns, the *small\_towns* variable will be set to 0; otherwise, it will be set to 1. Additionally, we will introduce a new variable called *university*. If the town has a university, the *university* variable will be set to 1; otherwise, it will be set to 0.

[Show](#)

### 2.3.2 Handling missing data:

[Show](#)

```
##                                town11cd
##                                4
##                                size_flag
##                                0
##                                job_density_flag
##                                4
##                                rgn11nm
##                                2
##                                income_flag
##                                4
##                                university_flag
##                                4
##                                key_stage_2_attainment_school_year_2007_to_2008
##                                0
##                                key_stage_4_attainment_school_year_2012_to_2013
##                                0
##                                activity_at_age_19_full_time_higher_education
##                                1
##                                highest_level_qualification_achieved_by_age_22_less_than_level_1
##                                859
##                                highest_level_qualification_achieved_by_age_22_level_1_to_level_2
##                                48
##                                highest_level_qualification_achieved_by_age_22_level_3_to_level_5
##                                1
##                                highest_level_qualification_achieved_by_age_22_level_6_or_above
##                                231
##                                highest_level_qualification_achieved_b_age_22_average_score
##                                0
##                                education_score
##                                0
##                                small_towns
##                                0
##                                university
##                                4
```

For our analysis, we will exclude variables with significant missing data. For example, the variables *highest\_level\_qualification\_achieved\_by\_age\_22\_less\_than\_level\_1* and *activity\_at\_age\_19\_out\_of\_work* have 859 and 462 missing values, respectively. As such, we will not use these variables to answer our research questions.

For variables with minimal missing data, we will handle them using the *na.rm = TRUE* argument during calculations to ensure accurate results without removing entire rows.

Because in this analysis, we want to focus on small, medium, big towns and city, so we will filter out other remaining observation in *size\_flag*.  
Because these remaining observations occupy really small so it does not affect analysis.

Show

## 2.4. Exploratory Data Analysis

### 2.4.1 Summary statistics for key variables.

Show

```

##      town11cd          size_flag      job_density_flag      rgn11nm
## Length:1100      Length:1100      Length:1100      Length:1100
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
## income_flag      university_flag
## Length:1100      Length:1100
## Class :character  Class :character
## Mode  :character  Mode  :character
##
##
##
## key_stage_2_attainment_school_year_2007_to_2008
## Min.      :28.09
## 1st Qu.:68.67
## Median :74.20
## Mean      :74.06
## 3rd Qu.:80.09
## Max.      :98.63
##
## key_stage_4_attainment_school_year_2012_to_2013
## Min.      :33.33
## 1st Qu.:54.41
## Median :60.77
## Mean      :61.28
## 3rd Qu.:67.61
## Max.      :92.86
##
## activity_at_age_19_full_time_higher_education
## Min.      : 7.874
## 1st Qu.:26.258
## Median :32.306
## Mean      :33.495
## 3rd Qu.:39.367
## Max.      :73.446
## NA's      :1
## highest_level_qualification_achieved_by_age_22_less_than_level_1
## Min.      :0.90
## 1st Qu.:2.00
## Median :2.60
## Mean      :2.78
## 3rd Qu.:3.40
## Max.      :7.70
## NA's      :859
## highest_level_qualification_achieved_by_age_22_level_1_to_level_2
## Min.      : 6.50
## 1st Qu.:21.20
## Median :26.50
## Mean      :26.76
## 3rd Qu.:31.90
## Max.      :51.70
## NA's      :48
## highest_level_qualification_achieved_by_age_22_level_3_to_level_5
## Min.      :23.00
## 1st Qu.:39.95
## Median :43.80
## Mean      :43.88
## 3rd Qu.:47.70
## Max.      :66.10
## NA's      :1
## highest_level_qualification_achieved_by_age_22_level_6_or_above
## Min.      :12.10
## 1st Qu.:23.50
## Median :28.00

```

```
## Mean :29.53
## 3rd Qu.:34.50
## Max. :61.40
## NA's :231
## highest_level_qualification_achieved_b_age_22_average_score
## Min. :2.567
## 1st Qu.:3.262
## Median :3.479
## Mean :3.513
## 3rd Qu.:3.724
## Max. :5.143
##
## education_score small_towns university
## Min. :-10.028015 Min. :0.0000 Min. :0.00000
## 1st Qu.: -2.583122 1st Qu.:0.0000 1st Qu.:0.00000
## Median : -0.295358 Median :1.0000 Median :0.00000
## Mean : -0.003958 Mean :0.6018 Mean :0.05636
## 3rd Qu.: 2.241820 3rd Qu.:1.0000 3rd Qu.:0.00000
## Max. : 11.871505 Max. :1.0000 Max. :1.00000
##
```

- Key Stage 2 attainment: The proportion of pupils achieving the expected level in Key Stage 2 ranges from 28.09% to 98.63%, with a median of 74.20% and a mean of 74.06%.
- Key Stage 4 attainment: The proportion of pupils achieving 5 GCSEs A\*-C including English and Maths ranges from 33.33% to 92.86%, with a median of 60.77% and a mean of 61.28%.
- Qualifications by age 22 Level 6 or above: The proportion of individuals achieving Level 6 or above by age 22 ranges from 12.10% to 61.40%, with a mean of 29.53%.
- Educational score: The educational score varies widely, with a mean of -0.00396, indicating that, on average, towns slightly underperform in education compared to the national benchmark.
- Full-time Higher Education: The proportion in full-time higher education at age 19 ranges from 7.87% to 73.45%, with a mean of 33.50%.

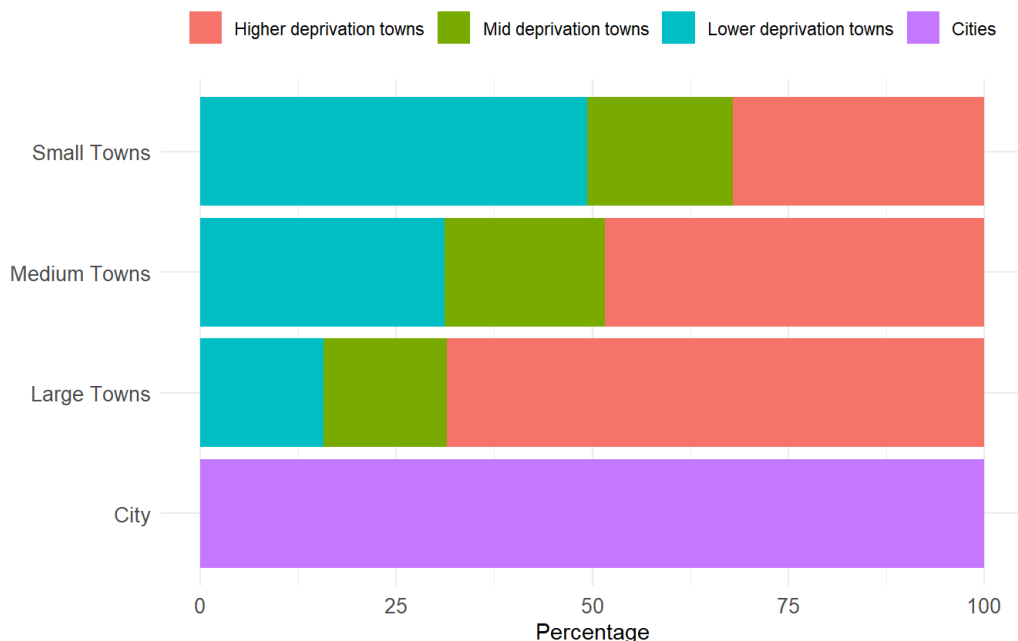
## 2.4.2 How does the difference in educational attainment vary across regions and city sizes in England?

Before addressing the main question, I would like to first highlight a key finding from previous studies using the same dataset. They discovered that small towns with lower levels of deprivation tend to have better educational outcomes compared to large cities. (Office for National Statistics analysis using Longitudinal Education Outcomes (LEO) from the Department for Education (DfE), 2023)

Show

### Small towns are less likely to be classed as having high income

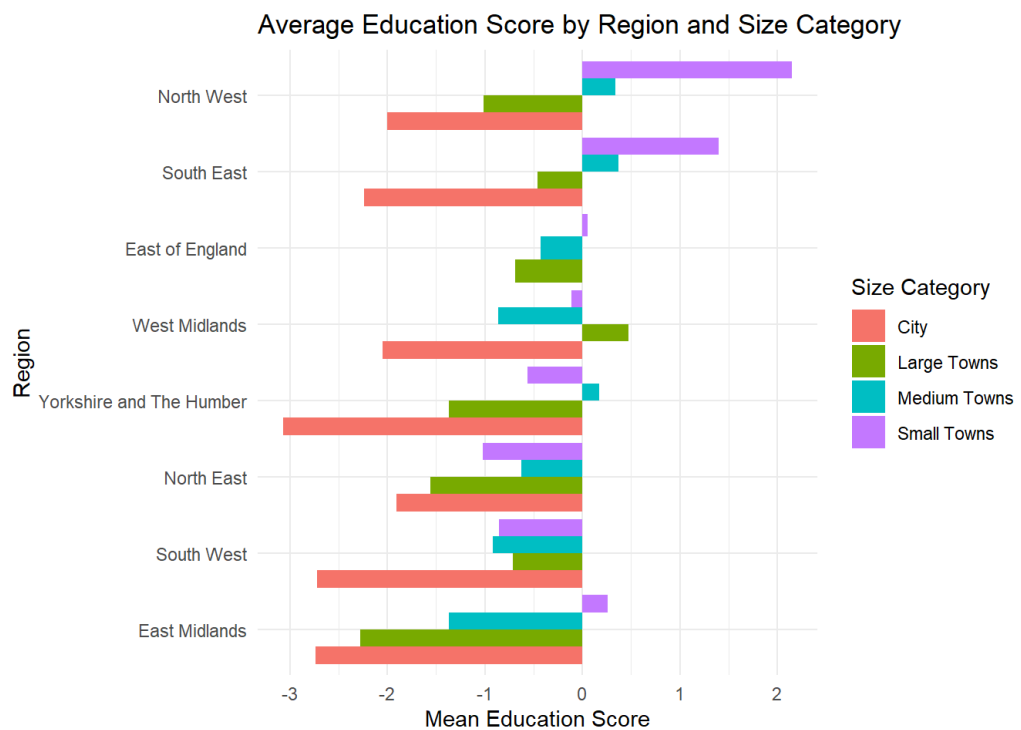
Income deprivation group by town size in England



The horizontal bar graph illustrates the distribution of income deprivation groups across different town sizes in England. It shows that small towns have a larger proportion of lower deprivation towns, indicating a relatively low rate of income deprivation. In contrast, large towns have a significant portion classified as higher deprivation towns which shows greater challenges with income inequality. Overall, the chart shows that smaller towns are less likely to face severe income deprivation compared to larger urban areas.

Score in different areas

Show



The graph is a horizontal bar chart that illustrates the mean education scores across different regions in England, grouped by four size categories: city, large towns, medium towns, and small towns. Each region has four bars, each corresponding to a size category. As we can see, city generally exhibit negative education scores, while small towns tend to have positive or neutral scores in certain regions like the North West . There is notable diversity across regions, with Yorkshire and The Humber showing predominantly negative scores across all size categories (only medium towns has positive value), whereas the South East and East Midlands have a broader mix of scores across the categories.

2.4.3 Descriptive Statistics and Visualizations

Longitudinal effects of early attainment gaps

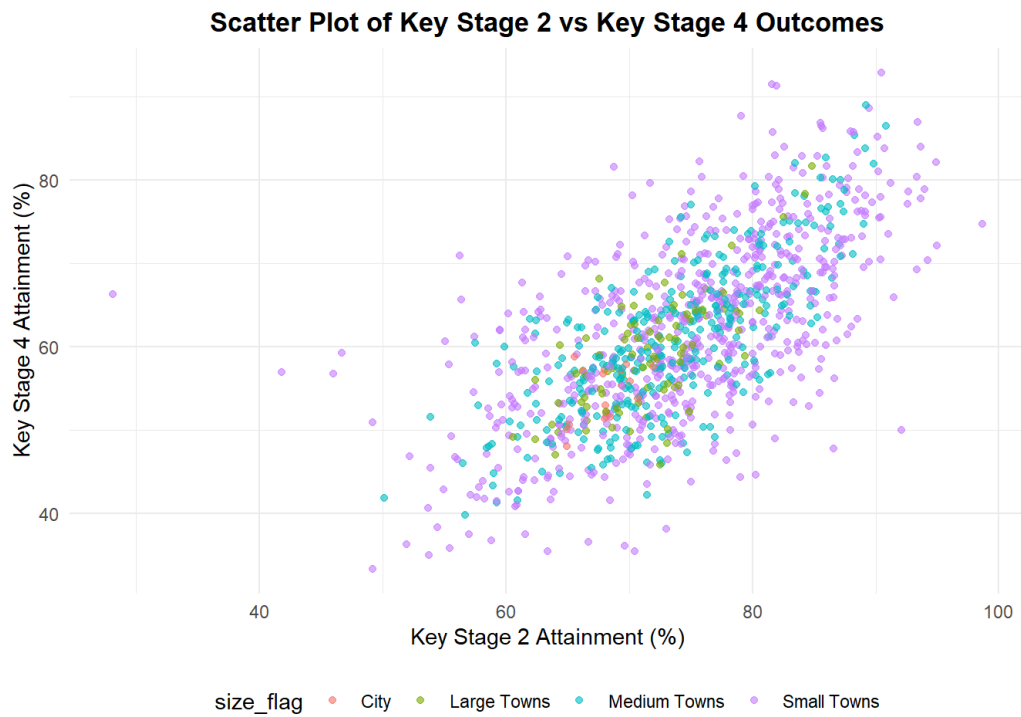
What is the relationship between Key Stage 2 attainment and Key Stage 4 outcomes in different towns?

Show

| mean_key_stage_2 | median_key_stage_2 | sd_key_stage_2 | mean_key_stage_4 | median_key_stage_4 | sd_key_stage_4 |
|------------------|--------------------|----------------|------------------|--------------------|----------------|
| 74.06304         | 74.20279           | 8.519216       | 61.27973         | 60.76897           | 10.11955       |

As observed, the average of 74.06% and the median of 74.2% are high at Key Stage 2, although the relatively low standard deviation of 8.52 depicts that a lot of the towns and cities have rather similar results at the Key Stage 2 level. On the contrary, in key stage 4, the mean (61.28%) and the median (60.77%) are less than in Key Stage 2, therefore fewer children generally reach the required standard of attainment as measured by GCSE results. It therefore has a high standard deviation (10.12), suggesting great variation in attainment across large areas.

Show



This scatter plot visualizes the relationship between Key Stage 2 Attainment (%) and Key Stage 4 Attainment (%) for students, with the data points categorized based on the `size_flag` variable. As we can see in the picture, there appears to be a positive correlation between Key Stage 2 and Key Stage 4 Attainment. This indicates that people having good attainment at Key Stage 2 tend to have higher attainment in Key Stage 4 in both small towns and larger towns.

*How do Key Stage 2 and Key Stage 4 outcomes correlate with the likelihood of achieving Level 6+ qualifications by age 22?*

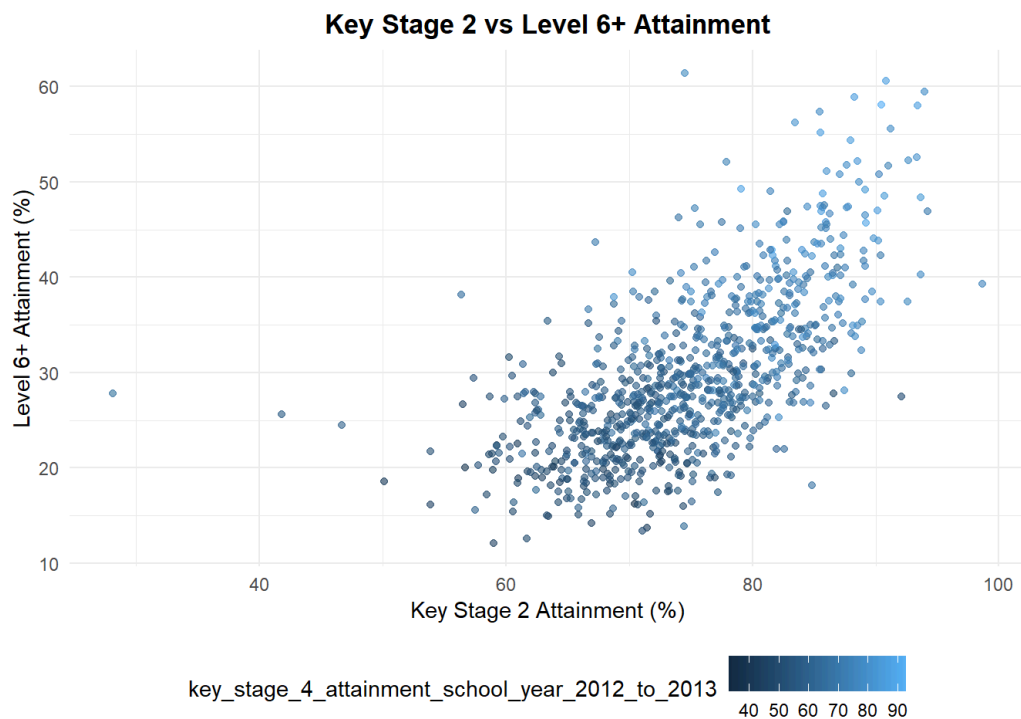
For better understanding the relationship between academic achievement in early stages and the likelihood of achieving Level 6+ (bachelor's degree or higher) by age 22, we will calculate mean, median for Key Stage 2, Key Stage 4, and Level 6+

Show

| mean_key_stage_2 | median_key_stage_2 | mean_key_stage_4 | median_key_stage_4 | mean_level_6_plus | median_level_6_plus |
|------------------|--------------------|------------------|--------------------|-------------------|---------------------|
| 74.06304         | 74.20279           | 61.27973         | 60.76897           | 29.52647          | 28                  |

As we can see in the results table, the percentage of individuals with higher qualifications tends to decrease. For example, the mean percentages for Key Stage 2 and Key Stage 4 are relatively high, at 74% and 61.3%, respectively. However, the percentage of people with Level 6 qualifications is significantly lower, at just around 30%, with the median only reaching 28%.

Show



This scatter plot explores the relationship between Key Stage 2 Attainment (%) and Level 6+ Attainment (%), where data points are shaded by the Key Stage 4 Attainment (2012–2013) percentage. The x-axis shows Key Stage 2 attainment, while the y-axis represents Level 6+ attainment. There is a positive correlation, in that the town has the higher the Key Stage 2 attainment, the higher the Level 6+ attainment. Moreover, the shading in the figure suggests that the higher values of Key Stage 4 attainment are more frequent for students whose Key Stage 6+ scores are higher, which strengthens a broader association between early and advanced attainment.

How do Key Stage 2 and Key Stage 4 outcomes correlate with the likelihood of achieving Level 6+ qualifications by age 22 in different area?

Show

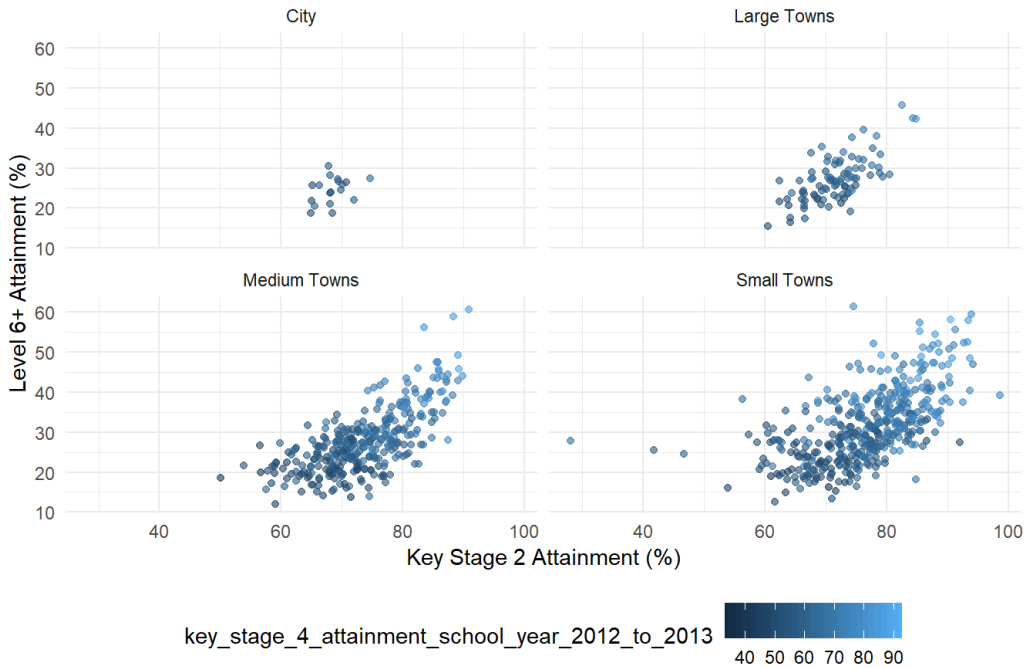
| size_flag    | mean_key_stage_2 | median_key_stage_2 | mean_key_stage_4 | median_key_stage_4 | mean_level_6_plus | median_level_6_plus |
|--------------|------------------|--------------------|------------------|--------------------|-------------------|---------------------|
| City         | 68.39655         | 68.12193           | 54.84553         | 55.55830           | 24.36111          | 25.05               |
| Large Towns  | 71.64647         | 71.70051           | 59.10339         | 58.96552           | 27.11011          | 26.80               |
| Medium Towns | 73.51211         | 73.27434           | 60.74205         | 59.89583           | 27.75485          | 26.80               |
| Small Towns  | 74.81746         | 75.43860           | 62.01611         | 61.93064           | 31.59282          | 30.50               |

The table provides a detailed comparison of educational attainment metrics across areas categorized by population size (size\_flag): City, Large Towns, Medium Towns, and Small Towns. For each area, the data includes the mean and median scores for Key Stage 2 (ages 7–11, school year 2007–2008), Key Stage 4 (ages 14–16, school year 2012–2013), and Level 6+ qualifications (highest qualification achieved by age 22).

The results show a consistent trend of higher educational outcomes in smaller towns compared to larger ones. For instance, Small Towns have the highest average Key Stage 2 score of 74.82 and a median of 75.44, compared with Cities, which had an average of 68.40 and a median of 68.12. This trend continues into Key Stage 4, where Small Towns attain an average of 62.02 compared to the Cities' 54.85. At Level 6+, the gap is even more extreme, with Small Towns achieving the highest mean of 31.59 and median of 30.50, compared to the Cities' mean of 24.36 and median of 25.05. This may further indicate that students from small towns are either surrounded by enabling educational environments or other beneficial factors for academic success in comparison with urban areas.

Show

## Key Stage 2 vs Level 6+ Attainment



The graph highlights clear differences in Level 6+ Attainment and Key Stage 2 Attainment across various town sizes. In cities, even towns with Key Stage 2 Attainment around 70% only achieve 20–30% Level 6+ Attainment, indicating limited progression to higher levels. Similarly, in large towns, while Key Stage 2 Attainment ranges between 60–80%, the proportion achieving 5 GCSEs or more (including English and Maths) with grades A–C is slightly lower at 50–70%, and Level 6+ Attainment remains mostly around 20–30%.

In contrast, medium and small towns show a strong positive correlation between Key Stage 2 Attainment and Level 6+ Attainment. Towns with higher proportions of students meeting expected levels in Key Stage 2 also demonstrate higher proportions achieving GCSE success and Level 6+ Attainment. This suggests that smaller towns are more effective in supporting students to reach higher attainment levels compared to larger urban areas.

### Impact of urbanization patterns in educational attainment

*What are the differences in Level 6+ attainment between towns with and without universities?*

First, we will explore `highest_level_qualification_achieved_by_age_22_level_6_or_above` with just 1 variable which is `university`

Show

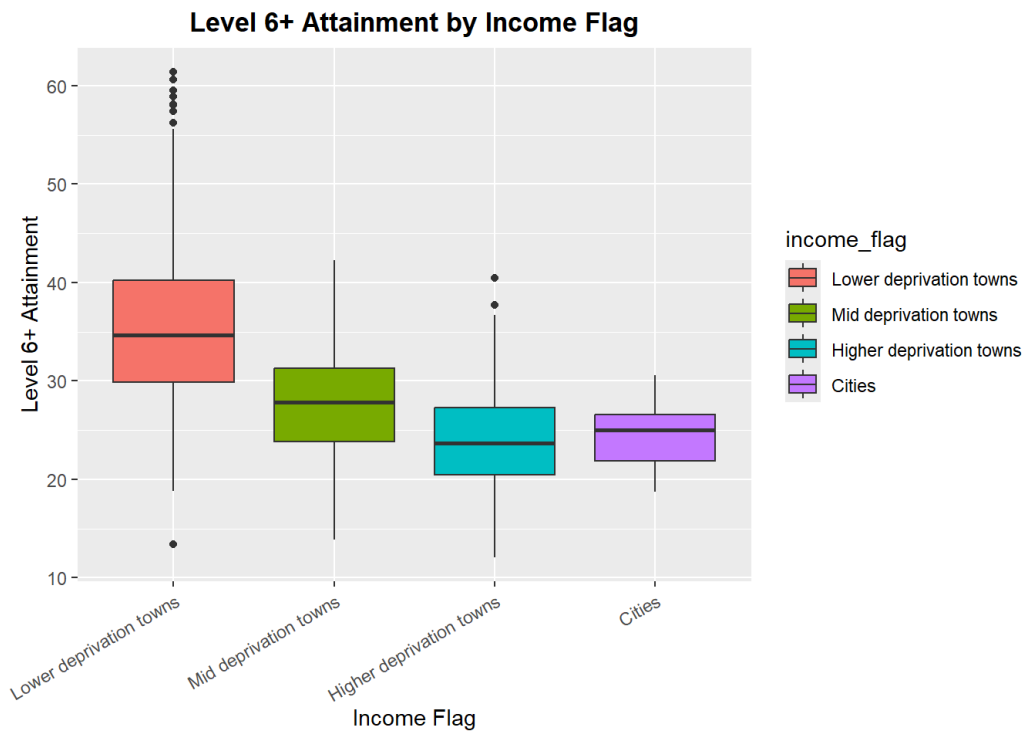
| university | mean_level6_plus | median_level6_plus | sd_level6_plus |
|------------|------------------|--------------------|----------------|
| 0          | 29.62092         | 28.1               | 8.636798       |
| 1          | 28.27541         | 27.3               | 6.234278       |

We will compare towns with and without universities, focusing on the proportion of students achieving Level 6 qualifications or above, which indicates higher education completion. As we expected, towns with universities would have higher educational attainment. However, towns without universities have a slightly higher mean proportion of students completing Level 6 qualifications (29.62%) compared to towns with universities (28.28%). The median values follow a similar trend, with non-university towns having a median of 28.1%, slightly higher than 27.3% in university towns. However, the standard deviation is higher in non-university towns (8.63%) than in university towns (6.23%), suggesting more variability in higher education outcomes among towns lacking universities.

Since this table summarizes attainment across all urbanization levels, it may cause a contradiction with what we expect. This counterintuitive result could be explained by students in non-university towns commuting to nearby cities or regions to pursue higher education, while towns with universities might attract more diverse populations, including individuals pursuing non-academic career paths. These findings highlight that the presence of a university does not automatically correlate with higher rates of local higher education attainment, and urbanization levels likely play a significant role in shaping these outcomes.

We will explore `income_flag` with `highest_level_qualification_achieved_by_age_22_level_6_or_above`.

Show



The box plot shows the distribution of “Level 6+ Attainment” across various income categories. It shows that “Lower deprivation towns” tend to have the highest median level 6+ qualification, with a relatively tight interquartile range (IQR). In contrast, “Higher deprivation towns” have a lower median and exhibit a wider spread, suggesting greater variability in educational attainment. “Cities” and “Mid deprivation towns” both show moderate medians but differ in their IQRs, with “Cities” having a smaller range, indicating consistency in attainment levels. Overall, the data suggests that areas with lower deprivation have higher educational attainment levels, while higher deprivation towns struggle more in this regard.

We then explore in more detail by exploring `highest_level_qualification_achieved_by_age_22_level_6_or_above` with `university` (university presence) and `income_flag` (income levels).

Show

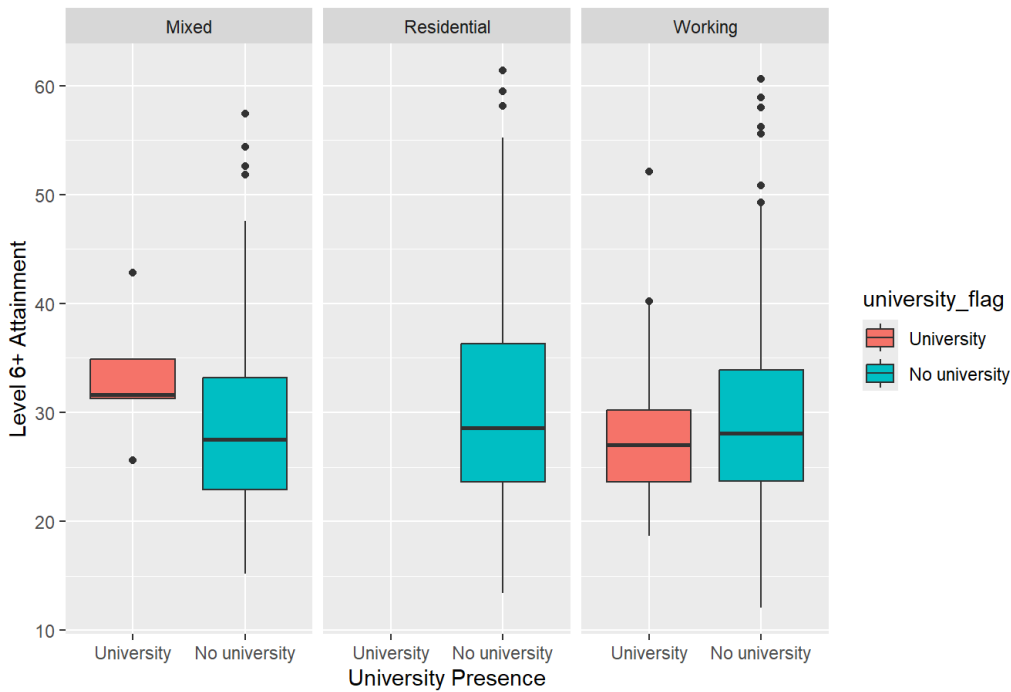
| university | income_flag              | mean_level6 | median_level6 | sd_level6 |
|------------|--------------------------|-------------|---------------|-----------|
| 0          | Lower deprivation towns  | 35.73000    | 34.75         | 8.387857  |
| 0          | Mid deprivation towns    | 27.94295    | 27.80         | 5.843780  |
| 0          | Higher deprivation towns | 23.89342    | 23.70         | 4.926981  |
| 1          | Lower deprivation towns  | 34.51875    | 33.35         | 7.231341  |
| 1          | Mid deprivation towns    | 28.53333    | 28.45         | 3.398484  |
| 1          | Higher deprivation towns | 26.10667    | 25.40         | 4.333172  |

The data shows how Level 6+ educational attainment differs between towns with and without universities based on income levels. In lower deprivation towns, those without universities have a slightly higher average attainment (35.73%) compared to those with universities (34.52%). On the other hand, in higher deprivation towns, having a university seems to help, with an average attainment of 26.11% compared to 23.89% in towns without universities. For mid-deprivation towns, areas without universities have an average of 27.94%, which falls between the lower and higher deprivation levels. The variability in attainment is greatest in lower deprivation towns, but towns with universities show slightly more consistent results. This suggests that university presence may help stabilize educational outcomes, especially in wealthier areas.

We will then explore this in more detail by exploring `highest_level_qualification_achieved_by_age_22_level_6_or_above` with `university_flag` (university presence) and `job_density_flag` (urbanization levels).

Show

### Level 6+ Attainment by University Presence and Urbanization



We see that in “Mixed” towns, those has universities show higher Level 6+ attainment with a narrower range, while towns without universities exhibit lower attainment with a wider spread. In “Residential” towns, without universities have a larger range and more outliers which shows greater variability in education attainment than “Mixed” towns. In “Working” towns, the median attainment is slightly higher in towns without universities compare to towns with universities, but these also show a broader distribution and more outliers compared to towns with universities. We can say that in “Mixed” towns have better proportion of students achieving Level 6 qualifications or above than other urbanization levels, so the score might vary by different factors.

*How does urbanization shape students' decisions to pursue higher education?*

Show

| university | mean_higher_education | median_higher_education | sd_higher_education |
|------------|-----------------------|-------------------------|---------------------|
| 0          | 33.39998              | 32.14286                | 10.58008            |
| 1          | 35.08643              | 33.76488                | 7.52437             |

The data shows that towns without universities have a slightly lower mean percentage of students participating in full-time higher education at age 19 (33.40%) compared to towns with universities (35.09%). However, the variation (standard deviation) is higher in towns without universities (10.58%) than in those with universities (7.52%), indicating that participation rates are more consistent in university towns. The median values also reflect this pattern, with university towns having a higher median (33.76%) compared to non-university towns (32.14%). This suggests that the presence of a university may contribute to slightly higher and more stable participation in higher education.

Show

Descriptive statistics for full-time higher education at age 19 due to the presence of universities and urbanization levels

| university | job_density_flag | mean_higher_ed | median_higher_ed | sd_higher_ed |
|------------|------------------|----------------|------------------|--------------|
| 0          | Mixed            | 32.45799       | 30.95975         | 9.922206     |
| 0          | Residential      | 32.89819       | 31.44310         | 11.265088    |
| 0          | Working          | 34.41611       | 33.33333         | 10.105759    |
| 1          | Mixed            | 41.32482       | 40.29785         | 6.636414     |
| 1          | Working          | 34.53920       | 32.96703         | 7.399035     |

The statistics demonstrate the influence of university presence and urbanization on students' decisions to pursue full-time higher education at age 19. Towns with universities tend to have higher mean participation rates in higher education, particularly in “Mixed” job-density areas, which show the highest mean of 41.32% and a relatively low standard deviation of 6.64%. This suggests a consistent trend of students in these areas

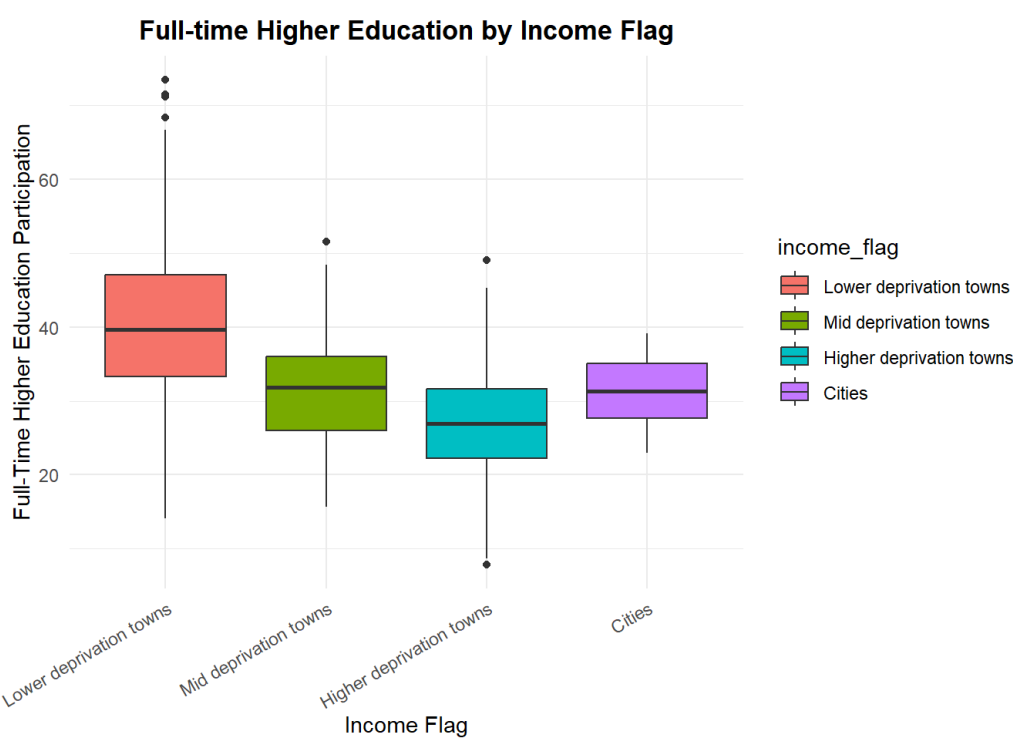
opting for higher education, potentially due to the balance of residential and employment opportunities coupled with access to university resources.

However, towns without universities exhibit lower mean participation rates. Among these, “Working” areas have the highest mean of 34.42%, while “Mixed” areas report the lowest mean of 32.46%. The standard deviations in these towns are higher, with “Working” and “Residential” areas showing variability of 10.11% and 11.27%, respectively. This indicates a more inconsistent trend in students’ pursuit of higher education in towns lacking universities.

These results highlight that the combination of university presence and a favorable mix of urbanization factors, such as residential opportunities and local employment, significantly influences educational participation at age 19.

We will explore `income_flag` with `activity_at_age_19_full_time_higher_education`.

Show



The box plot illustrates the distribution of “Full-Time Higher Education Participation” across different income categories. The data shows that “Lower deprivation towns” have the highest median participation rate, with a relatively narrow interquartile range (IQR) which show the consistency in the participation levels within this group. However, “Higher deprivation towns” exhibit the lowest median and the largest spread, highlights significant changes in participation rates. “Mid deprivation towns” and “Cities” fall between these two extremes, with moderate medians and varying degrees of spread. Outliers are particularly clear in the “Lower deprivation” towns, indicating a few towns with significantly higher participation rates compared to the rest. Therefore, we can say that areas with lower deprivation tend to have higher and more consistent participation in full-time higher education, whereas higher deprivation areas face challenges to pursue higher education.

### 2.5 Ethical considerations related to the dataset

When working with the provided dataset, several ethical considerations must be addressed to ensure responsible analysis. While the data does not include personal identifiers, geographic codes and detailed town-level variables (e.g., `town11cd`) could lead to indirect identification, particularly in small or rural populations. When such areas need it, aggregation for regional levels should be conducted. It is expected to avoid the stigmatizing or ranking of an area solely on deprivation indicators. Another example of sensitive variables for socioeconomic status is to refrain from framing them in context that perpetuates stereotypes about particularly disadvantaged regions.

Another sensitive aspect would be the question of data bias. Coastal and rural towns may face unique challenges that are not fully captured, potentially underestimating inequalities. Furthermore, the dataset’s focus on income deprivation and job density could overlook cultural or historical factors influencing educational outcomes. It is crucial to interpret findings within a broader context and acknowledge limitations in representativeness, particularly for underrepresented areas. The transparency of data provenance and the adherence to legal standards, such as GDPR, ensure that the research process does not lose trust and integrity. These measures help ensure that the analysis is ethical, accurate, and respectful to the communities represented.

## 3) Statistical Analysis and Interpretation

### 3.1 The statistical methods

To examine whether strong academic performance in early years leads to better outcomes later in life and to explore what other factors, beyond early academic success, shape students' decisions to pursue higher education, we will use various statistical techniques. A multiple linear regression model was used to model how variables such as town size, university status, income flag, and job density increase the probability of the student population attaining level 6 or above qualifications or pursuing higher education at age 19. This approach has been deemed valid with its ability to assess multiple predictors along with their respective independent impacts on educational outcome, meanwhile delivering a holistic approach on its effects.

In addition, hypothesis testing, such as two-sample t-tests, was conducted to test whether town size impacted higher participation and qualification rates in higher education convertibly. ANOVA, on the other hand, explored whether predictors such as university presence and job density had an impact on educational outcomes, and correlation analysis was used to assess the strength of relationships between continuous variables. These methods also suited group differences analysis and concomitant associations, so that data interpretations are stable and statistically founded.

### 3.2 Results and Interpretation

#### Longitudinal effects of early attainment gaps

*How does early educational attainment influence post-18 education?*

Show

```
##
## Call:
## lm(formula = activity_at_age_19_full_time_higher_education ~
##     key_stage_2_attainment_school_year_2007_to_2008 + key_stage_4_attainment_school_year_2012_to_2013,
##     data = model_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.3553  -4.2837   0.0767   3.8357  23.3054
##
## Coefficients:
##                                Estimate Std. Error t value
## (Intercept)                   -24.98635     1.96496  -12.716
## key_stage_2_attainment_school_year_2007_to_2008    0.25219     0.03535   7.134
## key_stage_4_attainment_school_year_2012_to_2013    0.66280     0.02944  22.510
##                                Pr(>|t|)
## (Intercept)                   < 2e-16 ***
## key_stage_2_attainment_school_year_2007_to_2008  2.06e-12 ***
## key_stage_4_attainment_school_year_2012_to_2013  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.171 on 866 degrees of freedom
## Multiple R-squared:  0.6209, Adjusted R-squared:  0.62
## F-statistic: 709.1 on 2 and 866 DF,  p-value: < 2.2e-16
```

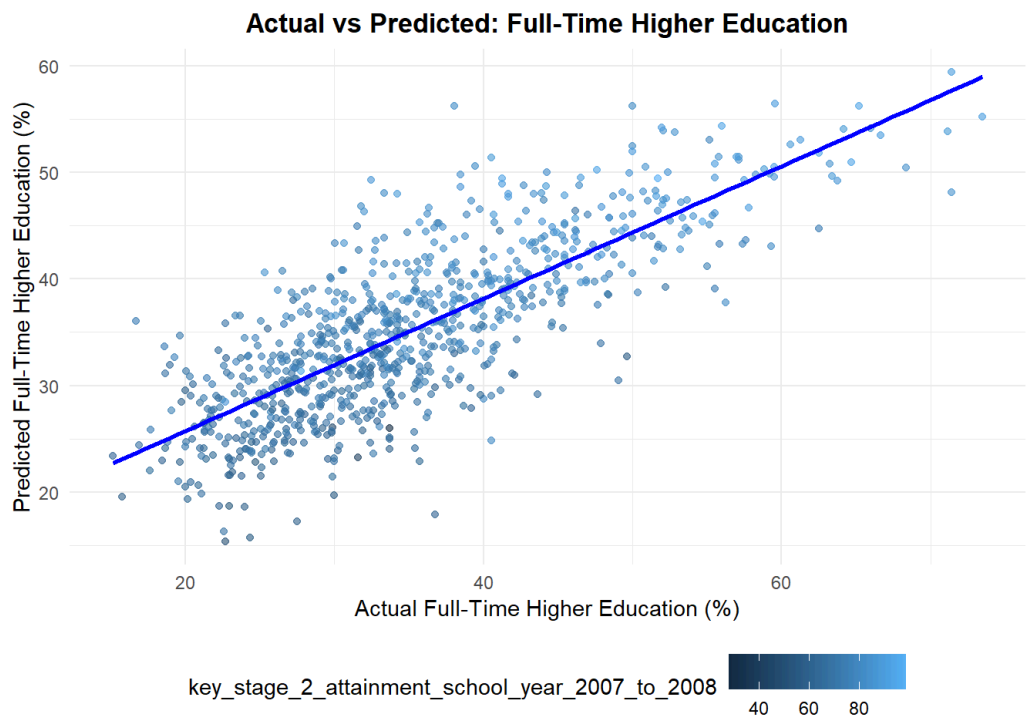
The linear regression results indicate that both Key Stage 2 and Key Stage 4 attainments significantly predict the likelihood of entering full-time higher education at age 19. Specifically, when both the proportion of Key Stage 2 and Key Stage 4 attainment are at 0%, the proportion of the town/city's 2012/13 Key Stage 4 cohort in full-time higher education at age 19 will be -23.78%. The coefficient for Key Stage 2 attainment suggests that for every 1% increase in the proportion of students achieving Level 4 or above in Key Stage 2 English and Maths in 2007-2008, the proportion of students entering full-time higher education at age 19 will increase by 0.25%, with a highly significant p-value (< 2e-16), holding Key Stage 4 attainment constant. Similarly, for every 1% increase in the proportion of students achieving 5 GCSEs or more, including English and Maths, with grades A\*-C in 2012-2013, the proportion of students entering full-time higher education at age 19 will increase by 0.63%, holding Key Stage 2 attainment constant, with a very significant p-value (< 2e-16). These results suggest that students who perform better at both Key Stage 2 and Key Stage 4 are more likely to enter full-time higher education.

The residual standard error of 6.68% indicates that the predictions made by the model deviate from the actual proportion of students in full-time higher education by about 6.68%. The adjusted R-squared value of 58.54% indicates that approximately 58.54% of the variability in the likelihood of entering higher education is explained by Key Stage 2 and Key Stage 4 attainment, which represents a reasonably good model fit.

Show

| activity_at_age_19_full_time_higher_education | predicted_full_time_education |
|---|-------------------------------|
| 41.84100                                      | 39.57442                      |
| 35.48387                                      | 30.22973                      |
| 31.64557                                      | 42.72041                      |
| 25.53191                                      | 31.01117                      |
| 38.04348                                      | 56.19175                      |
| 23.66864                                      | 36.57304                      |

Show



Full-time higher education is predicted from Key Stage 2 and Key Stage 4 attainment by a linear regression model, which indicates a reasonable fit. The R-squared value is 0.6209 and the Adjusted R-squared is 0.62. This means that about 62% of the actual values of full-time higher education are explained by the predictor variables Key Stage 2 and Key Stage 4 attainment. While the model captures the lion's share of the variance, the remaining 38% is unexplained, suggesting that other factors not included in the model may also affect higher education outcomes. The scatter plot reinforces this, in which a clear positive relationship between the predicted and actual values is evident, although with some noticeable deviations, indicating areas where the model's predictions are less accurate.

*How does early educational attainment influence later educational attainment?*

Show

```
##
## Call:
## lm(formula = highest_level_qualification_achieved_by_age_22_level_6_or_above ~
##     key_stage_2_attainment_school_year_2007_to_2008 + key_stage_4_attainment_school_year_2012_to_2013,
##     data = model_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.366  -3.450  -0.110   3.136  24.213
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept)    -23.60919     1.64029  -14.393
## key_stage_2_attainment_school_year_2007_to_2008    0.26768     0.02951   9.071
## key_stage_4_attainment_school_year_2012_to_2013    0.53202     0.02458  21.645
##              Pr(>|t|)
## (Intercept)    <2e-16 ***
## key_stage_2_attainment_school_year_2007_to_2008    <2e-16 ***
## key_stage_4_attainment_school_year_2012_to_2013    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.152 on 866 degrees of freedom
## Multiple R-squared:  0.633, Adjusted R-squared:  0.6322
## F-statistic: 747 on 2 and 866 DF, p-value: < 2.2e-16
```

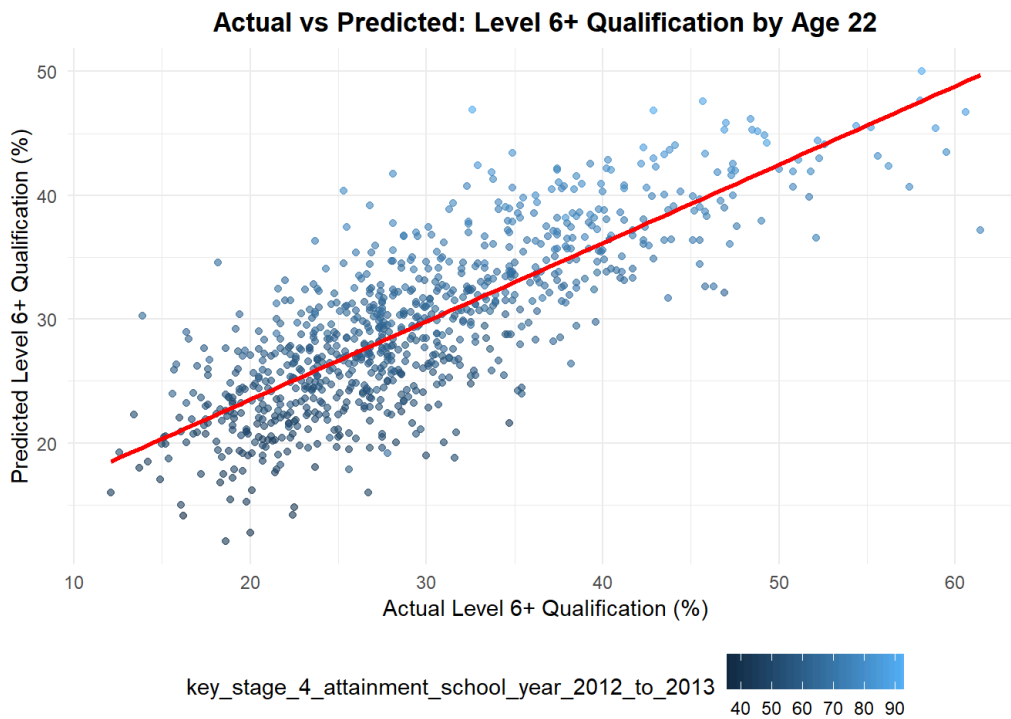
The results of the linear regression model show that both Key Stage 2 and Key Stage 4 attainment significantly predict the likelihood of achieving Level 6+ qualifications by age 22. Specifically, when both Key Stage 2 and Key Stage 4 attainments are at 0%, the predicted likelihood of achieving Level 6+ qualifications is -23.61%. For every 1% increase in the proportion of students achieving Level 4 or above in Key Stage 2, the likelihood of achieving Level 6+ qualifications increases by 0.27%, with a highly significant p-value (< 2e-16). Similarly, for every 1% increase in students achieving 5 or more GCSEs with grades A\*-C in Key Stage 4, the likelihood of achieving Level 6+ qualifications increases by 0.53%, also with a significant p-value (< 2e-16).

The residual standard error of 5.15% indicates that the model's predictions deviate from the actual results by about 5.15%. The adjusted R-squared value of 0.6322 suggests that approximately 63.22% of the variability in achieving Level 6+ qualifications is explained by Key Stage 2 and Key Stage 4 attainment, indicating a good model fit. These findings underscore the importance of early academic performance in shaping future educational success, highlighting that improvements in Key Stage 2 and Key Stage 4 can significantly increase the likelihood of students attaining Level 6+ qualifications by age 22.

Show

| highest_level_qualification_achieved_by_age_22_level_6_or_above | predicted_level_6_plus_education |
|---|----------------------------------|
| 33.3  | 32.71875                         |
| 32.5  | 25.37258                         |
| 28.1  | 35.55705                         |
| 23.6  | 25.97603                         |
| 32.6  | 46.89694                         |
| 23.7  | 30.56899                         |

Show



The above scatter plot series show the actual Level 6+ qualifications achieved at age 22 against the estimated based on attainment in Key Stage 2 and attainment in Key Stage 4. The red line of this trend shows a fairly encouraging line of best fit—a fairly positive correlation in that direction between actual and predictions, percentage-wise. This graph reflects the nature of the table, whereby predictions are all over the place from the actual attainment. For example, the attainment percentages for lower attainment percentages, say 23.6 and 23.7, have higher predicted values than the actual achievements, according to the plot, while at the same time the model overestimates at some points, say at 32.6%. The plot highlights the discrepancies in the model's predictions concerning Level 6+ qualifications, mostly following a positive trend, though there are various discrepancies which bring out the complex nature of factors affecting educational performance.

#### Impact of urbanization patterns in educational attainment

Show

*Does urbanization factors mitigate educational inequalities?*

Show

```
##
## Welch Two Sample t-test
##
## data: small_towns$activity_at_age_19_full_time_higher_education and large_towns$activity_at_age_19_full_time_higher_education
## t = 1.0042, df = 1058.3, p-value = 0.3155
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.5892933 1.8246476
## sample estimates:
## mean of x mean of y
## 33.74130 33.12362
```

In this two sample t-test, the null hypothesis states that there is no difference in the mean proportion of the 2012/13 key stage 4 cohort in full-time higher education at age 19 between small towns and large towns. Conversely, the alternative hypothesis says that there is a difference in the mean students pursuing full-time higher education at age 19 between small towns and large towns.

The t-test compares the proportion of students from small and large towns who went into full-time higher education by age 19. The result shows no significant difference between small towns (33.74%) and large towns (33.12%). The p-value of 0.3155 is much higher than 0.05 which we think that the difference we see is likely due to chance rather than any real effect. The confidence interval also suggests that there is no strong evidence that students from small and large towns differ in how many go on to full-time higher education at age 19.

Show

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## university_flag      1    166   166.4    1.536 0.2155
## job_density_flag      2    582   291.1    2.687 0.0685 .
## university_flag:job_density_flag      1    340   340.2    3.140 0.0767 .
## Residuals          1094 118499   108.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

The test shows that the presence of universities and `job_density_flag` have limited impact on the likelihood of students pursuing full-time higher education at age 19. In this test, we will use a threshold of 0.05. The effect of university presence is not statistically significant since  $p = 0.2155$ , which shows that whether or not a university is present in a town does not strongly influence higher education participation at age 19. Urbanization levels has  $p = 0.0685$ , saying a slight influence of urbanization on higher education participation, but it is not strong enough to be affected. Also, the effect of universities may vary depending on urbanization by look at the interaction between university presence and urbanization levels which has  $p = 0.0767$ . Overall, when both university presence and job density are considered together, their individual effects on full-time higher education participation at age 19 are less clear than when each is examined separately.

Here, we will examine `university_flag` and `job_density_flag` separately.

Show

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## university_flag      1    166   166.4    1.528 0.217
## Residuals          1097 119422   108.9
## 1 observation deleted due to missingness
```

The ANOVA test results say that university presence has a statistically significant effect on the proportion of the town/city's 2012/13 Key Stage 4 cohort in sustained further education at the age of 19. The p-value of 0.000739 show a significant difference in the proportion of students in sustained further education between towns with and without a university. Specifically, towns with a university tend to have a higher proportion of students continuing their education at age 19. Since p-value less than 0.05, we can reject the null hypothesis that university presence has no impact on further education participation, indicating that university presence positively influences educational outcomes.

Show

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## job_density_flag      2    704   352.0    3.245 0.0394 *
## Residuals          1096 118884   108.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

The ANOVA test results show that job density significantly affects the proportion of the town/city's 2012/13 Key Stage 4 cohort in sustained further education at the age of 19. The p-value of 0.00176 indicate that there are significant differences in educational outcomes across towns with different job density levels. This suggests that towns with varying job densities (such as residential, mixed, or working areas) have differing proportions of students continuing their education at age 19. Since p-value less than 0.05, we can reject the null hypothesis that job density has no impact on further education participation.

Show

```
##              [,1]
## income_flag_num    -0.56939792
## job_density_flag_num 0.07085441
## university_flag_num 0.03730054
```

The results show how different factors such as `income_flag`, `job_density_flag`, and `university_flag` are related to the likelihood of students pursuing full-time higher education at age 19. The strongest relationship is with the income flag, which has a correlation of -0.569. This suggests that towns with higher income deprivation tend to have fewer people attending full-time higher education at age 19. Job density has a very weak positive correlation of 0.071, meaning that whether a town is more focused on jobs doesn't seem to have a strong impact on the likelihood of individuals being in full-time higher education at age 19. Also, the presence of a university in the town shows a very weak positive correlation of 0.037 show that having a university doesn't significantly affect the proportion of people in full-time higher education when they were 19.

Show

```
##
## Call:
## lm(formula = activity_at_age_19_full_time_higher_education ~
##     income_flag + job_density_flag + university_flag, data = model2_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.243  -5.327  -0.417   4.307  31.095
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      41.6803     0.6855  60.801 < 2e-16 ***
## income_flagMid deprivation towns    -9.3072     0.7552 -12.324 < 2e-16 ***
## income_flagHigher deprivation towns -13.4447     0.6063 -22.175 < 2e-16 ***
## income_flagCities      -12.6986     2.2562  -5.628 2.46e-08 ***
## job_density_flagResidential     0.5169     0.7637   0.677   0.499
## job_density_flagWorking     0.6714     0.7204   0.932   0.352
## university_flagUniversity     1.7473     1.2714   1.374   0.170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.938 on 862 degrees of freedom
## Multiple R-squared:  0.3756, Adjusted R-squared:  0.3713
## F-statistic: 86.43 on 6 and 862 DF,  p-value: < 2.2e-16
```

The coefficient for “income\_flag” shows a significant negative relationship between income deprivation and pursue full-time higher education at age 19. The towns with mid-income deprivation have a decrease of about 9.3%, higher income deprivation towns decrease of 13.4%, and cities with higher deprivation experience a 12.7% lower enrollment rate. Income in 3 groups: mid-deprivation, higher-deprivation, and cities, are highly statistically significant (p-values < 2e-16). On the other hand, job density flags (Residential and Working) do not significantly impact full-time higher education enrollment for people when they were 19, with p-values of 0.499 and 0.352 compare to significant level of 0.05, respectively. Similarly, having a university in the town shows a weak positive relationship, but this is not statistically significant (p-value = 0.170). Overall, income deprivation has a strong negative effect on full-time higher education participation 19, while job density and university presence have minimal influence in this model.

*Does urbanization levels influence the attainment of level 6+ by age 22?*

Show

```
##
## Welch Two Sample t-test
##
## data:  small_towns$highest_level_qualification_achieved_by_age_22_level_6_or_above and large_towns$highest_level_qualification_achieved_by_age_22_level_6_or_above
## t = 7.3356, df = 830.51, p-value = 5.252e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.009586 5.208557
## sample estimates:
## mean of x mean of y
## 31.59282 27.48375
```

In this two sample t-test, the null hypothesis states that there is no difference in the mean proportion of individuals achieving a Level 6+ qualification between small towns and large towns. Conversely, the alternative hypothesis says that there is a difference in the average proportion of individuals achieving this qualification level between small and large towns. The results indicate a statistically significant difference between the two groups with the p-value = 5.252e-13 compare to significant level of 0.05 which shows strong evidence to reject the null hypothesis that there is no difference in means. The 95% confidence interval for the difference in means is [3.01, 5.21], suggesting that, on average, small towns have 3.01% to 5.21% more individuals achieving Level 6 qualifications or above than large towns. The sample means show that small towns have a higher proportion of individuals with these qualifications (31.59%) compared to large towns (27.48%). This suggests that small towns may provide an environment more supportive to achieving higher educational attainment.

Show

```
##
##          Df Sum Sq Mean Sq F value Pr(>F)
## university_flag      1    103   102.68   1.427  0.233
## job_density_flag      2    188    93.77   1.303  0.272
## university_flag:job_density_flag  1    157   156.81   2.179  0.140
## Residuals          864   62185    71.97
## 231 observations deleted due to missingness
```

We use ANOVA test to test the effects of university presence, urbanization levels, and their interaction on Level 6+ attainment by age 22. In this test, we will use a threshold of 0.05. The main effect of university presence has a p-value of 0.233, showing no statistically significant difference in Level 6+ attainment between towns with and without universities. Similarly, the main effect of urbanization levels have a p-value of 0.272, showing no significant differences in attainment across “Mixed,” “Residential,” and “Working” towns. The interaction between university presence and urbanization levels has a p-value of 0.140, suggesting that the combined effect of these factors on attainment is not statistically significant either. Overall, these results imply that neither university presence, job density, nor their interaction significantly influence Level 6+ attainment.

Show

```
##          [,1]
## income_flag_num      -0.611415943
## job_density_flag_num -0.005002295
## university_flag_num  -0.040489966
```

The results show how different factors such as `income_flag`, `job_density_flag`, and `university_flag` are related to the proportion of people achieving a Level 6 qualification or above by age 22. The strongest relationship is with the income category, showing a correlation of -0.611. This means that areas with higher levels of income deprivation tend to have fewer people reaching higher education levels. On the other hand, the correlation between job density and higher education by age 22 is extremely weak which is -0.005, suggesting that whether a town is more working-focused or residential doesn't significantly impact education levels. Similarly, the presence of a university in a town also shows a weak negative correlation (-0.040), indicating that having a university nearby doesn't strongly affect the proportion of people achieving high educational qualifications. Therefore, income inequality appears to play a significant role in pursue higher education by age 22.

Show

```
##
## Call:
## lm(formula = highest_level_qualification_achieved_by_age_22_level_6_or_above ~
##     income_flag + job_density_flag + university_flag, data = model2_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3816  -4.5738  -0.4617   3.6148  25.6184
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      35.56832    0.57387  61.979 < 2e-16 ***
## income_flagMid deprivation towns    -7.69449    0.63223 -12.170 < 2e-16 ***
## income_flagHigher deprivation towns -11.66829    0.50755 -22.990 < 2e-16 ***
## income_flagCities      -11.77090    1.88873  -6.232 7.19e-10 ***
## job_density_flagResidential      0.21332    0.63935  0.334  0.739
## job_density_flagWorking      0.01139    0.60308  0.019  0.985
## university_flagUniversity      0.55293    1.06432  0.520  0.604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.645 on 862 degrees of freedom
## Multiple R-squared:  0.3922, Adjusted R-squared:  0.388
## F-statistic: 92.72 on 6 and 862 DF, p-value: < 2.2e-16
```

In this regression model, we explore the relationship between income levels, job density, and university presence on the highest level of qualification achieved by age 22. Based on the result, towns with mid or higher income deprivation, as well as cities, tend to have significantly lower levels of qualifications compared to towns with lower income deprivation (reference group). Specifically, the coefficient for mid-deprivation towns is -7.69, for higher-deprivation towns is -11.67, and for cities is -11.77, all showing negative associations with qualification levels, with p-values well below 0.05, indicating strong statistical significance.

However, job density and university presence do not show a significant effect on qualification levels. The coefficients for both job density categories (residential and working) and university presence are very close to zero, with high p-values, which say these factors don't affect educational outcomes for level 6 or above at age 22. Overall, income deprivation might strongly influences educational outcomes, but, job density and university presence may have less of an impact.

#### Final Model

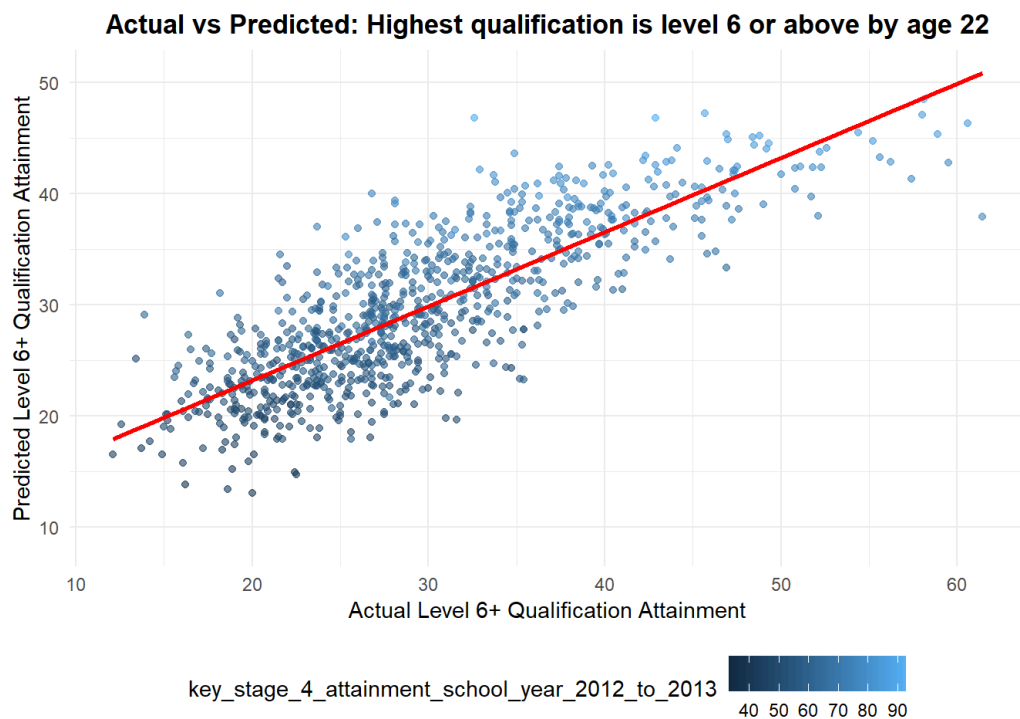
Show

```
##
## Call:
## lm(formula = highest_level_qualification_achieved_by_age_22_level_6_or_above ~
##     key_stage_2_attainment_school_year_2007_to_2008 + key_stage_4_attainment_school_year_2012_to_2013 +
##     income_flag + job_density_flag, data = english_education)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1873  -3.3742  -0.1296   2.9590  23.5171
##
## Coefficients:
##                                Estimate Std. Error t value
## (Intercept)                   -11.58704     2.09299  -5.536
## key_stage_2_attainment_school_year_2007_to_2008    0.20099     0.02938   6.842
## key_stage_4_attainment_school_year_2012_to_2013    0.45919     0.02488  18.459
## income_flagMid deprivation towns                   -2.81535     0.50218  -5.606
## income_flagHigher deprivation towns                 -4.29965     0.46935  -9.161
## income_flagCities                                -2.93904     1.24148  -2.367
## job_density_flagResidential                       -0.76507     0.47409  -1.614
## job_density_flagWorking                           -0.04696     0.44296  -0.106
##
##                                Pr(>|t|)
## (Intercept)                   4.11e-08 ***
## key_stage_2_attainment_school_year_2007_to_2008  1.49e-11 ***
## key_stage_4_attainment_school_year_2012_to_2013  < 2e-16 ***
## income_flagMid deprivation towns                 2.78e-08 ***
## income_flagHigher deprivation towns               < 2e-16 ***
## income_flagCities                                0.0181 *
## job_density_flagResidential                       0.1069
## job_density_flagWorking                           0.9156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.917 on 861 degrees of freedom
## (231 observations deleted due to missingness)
## Multiple R-squared:  0.6677, Adjusted R-squared:  0.665
## F-statistic: 247.1 on 7 and 861 DF, p-value: < 2.2e-16
```

The R-square value indicates that the linear regression model explains about 66.8% of the variance in the proportion of people reaching Level 6 or above qualifications by age 22. Both Key Stage 2 attainment, 2007-2008, and Key Stage 4 attainment, 2012-2013, are statistically significant predictors,  $p < 0.001$ , and higher attainment levels at these stages are positively associated with the likelihood of reaching Level 6 qualifications. More specifically, for every 1 per cent increase in attainment in Key Stage 2 there is a 0.20 per cent increase in Level 6 qualifications while the attainment in Key Stage 4 predicts 0.46 percent increase.

Among the categorical predictors, respondents in Lower deprivation towns have significantly higher rates of Level 6 attainment ( $p = 0.018$ ) compared to the reference category, Cities, with an estimated effect of 2.94%. However, no other categories of income or job density are significantly different from one another. The residual standard error, 4.917, indicates a moderate variability around the fitted values that is not accounted for by the model, but the model itself is highly significant,  $p < 2.2e-16$ , and thus meaningful.

Show



The Actual vs Predicted plot shows that most data points align well with the red line, indicating the model predicts accurately for the majority of cases. However, deviations at higher values suggest potential outliers or limitations in capturing extreme cases. While the linear trend seems reasonable, further analysis is needed to confirm the model's effectiveness and identify improvement areas.

### 3.3 Discussion

This study gives insight into how early academic success influences long-term educational outcomes and how these are shaped by urbanization along with associated inequalities. As far as the first research inquiry is concerned, the findings certify that those who do well at Key Stage 2 (KT2) have greater likelihoods of doing well at Key Stage 4 (KT4). Furthermore, those who do well at both levels have odds of coming out with at least Level 6 or above by age 22 that are significantly greater. This highlights the crucial importance of early academic success in being able to carry through these advantages to the protest stages of education. Such early achievements provide the momentum needed for some students to develop skills and confidence that will aid them in later years.

However, this advantage varies depending on the size and level of deprivation of the town. Even for equal achievements at KT2 and KT4, it is indeed important to consider that small and medium towns that entail lower deprivation present far higher proportions of students achieving Level 6 qualifications than larger towns and cities. This suggests that factors beyond academic performance — such as a supportive community environment, lower socioeconomic barriers, or better alignment between local resources and student needs — play a role in long-term success. Also, these early achievements influence pursuing full-time higher education at the age of 19. However, large towns show that there persist disparities, wherein the proportions of out-of-high-school learners going on to higher education exceed by far the proportions actually attaining Level 6, thus indicating substantially high dropout rates.

On the second research question, urbanization has a complex effect on educational choices and outcomes. Surprisingly outstripped of expectations: the towns with universities do not perform much better than those without universities in terms of level 6 qualification rates by age 22. University influence shows a strong positive link for participation in full-time higher education by age 19, particularly looking at the “Mixed” towns. Perhaps these are the towns that offer the right balance of residential and employment opportunities for entering higher education. Urbanization factors like deprivation levels also play a role while job density is not statistically significant. Throughout these lower deprived areas, educational attainment and participation rates seem to be greater, pointing toward the central task of addressing socioeconomic inequalities so as to mitigate educational disparity.

Notably, urban settings have different participation and qualification rates in higher education. A huge drop of those commencing their studies at age 19 completing Level 6 qualifications by age 22 occurs in larger towns, with the average figure dropping from 33.12% to 27.48%. However, small towns see a much smaller fall-over rates, from 33.74% to 31.6%. One might think that some urbanization would cause a retention challenge such as a more significant drop-out rate in much larger towns. Factors such as economic pressures, inadequate support systems, or a mismatch between aspirations and existing opportunities might provide explanations for the unevenness.

This complex of factors underscores the intersection of early academic achievement, urbanization, and socioeconomic factors in shaping the educational experience of students. While success during the elementary years can provide a sound basis, regional and urban traits can increase or decrease long-run outcomes. These structural inequalities have to be addressed to let the students, irrespective of their geographical and socioeconomic background, realize their potential.

## 4) Conclusions

### Key Takeaways:

This analysis has shown that early academic achievements at Key Stage 2 and Key Stage 4 emerge as crucial predictors of long-term educational success, and high performance at these stages is strongly associated with achieving Level 6 qualifications or above by the age of 22. Students in small and medium towns with lower deprivation demonstrate much higher attainment rates of Level 6 qualifications compared to those in large towns and cities, even if achievements earlier in the educational process are not differing. Urbanization has a multifaceted influence: while the presence of a university appears to encourage age 19 higher education participation, it does not necessarily confer any real advantages in completion rates of qualifications by 22. “Mixed” towns, those with a good balance of residential opportunities and employment, appear to achieve better results in terms of participation in higher education. Finally, larger towns appear to face a greater gap between higher education participation by age 19 and qualification completion by age 22, suggesting retention problems that are less pronounced for smaller towns. These demonstrate thereby the relevance of both academic and environmental influences on educational performance.

### Limitations:

The study has several limitations that may affect the interpretation of its findings. First, this data mostly comprises historical information from limited periods, which by all means hinders the capture of the authentic dynamics of the education or socioeconomic landscape in their contemporary form. Missing variable value matrices would have affected that significantly on the basis of an analytical strength and generalizability standpoint. Furthermore, aspects such as culture or historical perspectives which are quite vital in affecting educational performance were excluded. Also, an analysis using town-level data might mask variations at the individual level or unique contextual effects.

### Next Steps:

Future research could take advantage of longitudinal data to measure how educational outcomes change over time, especially in constant socioeconomic status or under specific policy interventions. Including other variables such as parental education, school funding, or students' aspirations would benefit an understanding of attainment factors in-depth. Qualitative work on the reasons for large town dropouts could give an additional concourse into some difficulties students have to meet. Additionally, examining urbanization's effects on non-educational outcomes, such as employment or migration patterns, would offer a broader understanding of its role in shaping life trajectories.

### Ethical Considerations:

The remaining aspects of the analysis can be ascribed to the ethical consideration of not using individual-level identifiers and aggregating town-level data to guarantee privacy. No towns were demeaned, nor would this report, for instance, in sync with the deprivation ranking of the same or render any stigma on underperforming places. The possibility was noted of indirect identification for smaller towns by virtue of their size or population, so further studies should aggregate data as may be necessary to guard anonymity. Researchers also should sensitively interpret findings to ensure future framing of their publications does not get interpreted as class- reinforcements for inequality or their ramifications for marginalized groups as unintended consequences. Credibility and integrity within the processes were infused by transparency of methods while taking data protection regulations into heed.

## 5) References

- Dataset: Campos, C., Joyce, M., & McNulty, N. (2023). Educational attainment of young people in English towns data. Office for National Statistics. Retrieved from <https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/datasets/educationalattainmentofyoungpeopleinenglishtownsdata> (<https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/datasets/educationalattainmentofyoungpeopleinenglishtownsdata>)
- Playford, C. J., Mountford-Zimdars, A., & Benham-Clarke, S. (2023). Coast and city, it matters where you live: How geography shapes progression to higher education in England. *Social Sciences*, 12(11), 610. <https://doi.org/10.3390/socsci12110610> (<https://doi.org/10.3390/socsci12110610>)
- Dunatchik, A., Wishart, R., Cartagena-Farias, J., & Smith, N. (2018, August). Regional differences in attainment in the early years (Research Report). NatCen Social Research.
- Centre for Subnational Analysis. (2024, March 15). Geographical mobility of young people across English towns and cities: Cohort analysis using the Longitudinal Education Outcomes dataset exploring the link between education attainment and geographical mobility of young people: 2007 to 2019. Office for National Statistics.