Friedrich-Alexander University Erlangen-Nürnberg

Faculty of Sciences

– M. Sc. Data Science–

Project WS 2023/24

Methods of Advanced Data Engineering

Exploring the Correlation Between Demographic Subdivision and Quality of Life in Urban Areas: A Case Study of London (2014-2016)

submitted by

Julian Maibach

Matriculation Number: 23392611

Julian Maibach: *Exploring the Correlation Between Demographic Subdivision and Quality of Life in Urban Areas: A Case Study of London (2014-2016)*, 10. January 2024, Erlangen

Contents

1. Introduction

1.1. Background

Health has always been a fairly extensive discussed topic and with thematically more diverse treatment of its components, sub-areas such as mental health steadily gained general awareness over the last years. With this progress in thinking about personal health, the question about the term 'quality of life' has also arisen rather frequently. Being very situation-dependent and mostly subjective, it is yet a hard to define aspect of every person's life.

1.2. Research Question and Motivation

This project therefore aims to investigate whether there are conclusions that can be drawn out of a possible correlation between demographic strata and their respective quality of life. Therefore, different demographic and quality of life aspects shall be examined using sample data of the City of London in the years 2014–2016. The results of this project should provide fundamental insights about a possible dependency of the private situation on the quality of live on individual demographic groups.

2. Methods

2.1. Data Sources

2.1.1. Datasource1: London Borough Demographics

- Metadata URL: `https://www.kaggle.com/datasets/marshald/london-boroughs/"`
- Data URL: `https://www.kaggle.com/datasets/marshald/london-boroughs/?select=london-borough-profiles-2016+Data+set.csv"`
- Data Type: CSV

The data source profiles demographic data, such as labour market, economy and many more regarding the boroughs of London in the year of 2016. Due to its compact, yet diverse information provided, the data set was the foundation of the data analysis of this project.

2.1.2. Datasource2: London Crime Data

- Metadata URL: `https://www.kaggle.com/datasets/jboysen/london-crime`
- Data URL: `https://www.kaggle.com/datasets/jboysen/london-crime?select=london_crime_by_lsoa.csv"`
- Data Type: CSV

Crime in major metropolitan areas, such as London, occurs in distinct patterns. This data source covers the number of criminal reports by month, borough, and major/minor category from Jan 2008-Dec 2016, though only the data from 2014-2016 is used.

2.1.3. Datasource3: Housing in London

- Metadata URL: `https://www.kaggle.com/datasets/justinas/housing-in-london`
- Data URL: `https://www.kaggle.com/datasets/justinas/housing-in-london?select=housing_in_london_monthly_variables.csv"`
- Data Type: CSV

This data contains information about the housing market of London from the years 1999 until 2019, though only the monthly data from 2014-2016 is used. The data has been extracted from London Datastore. It is released under UK Open Government Licence v2 and v3. The underlining datasets can be found here:

- `https://data.london.gov.uk/dataset/uk-house-price-index`
- `https://data.london.gov.uk/dataset/number-and-density-of-dwellings-by-borough`
- `https://data.london.gov.uk/dataset/subjective-personal-well-being-borough`
- `https://data.london.gov.uk/dataset/household-waste-recycling-rates-borough`
- `https://data.london.gov.uk/dataset/earnings-place-residence-borough`
- `https://data.london.gov.uk/dataset/recorded_crime_summary`
- `https://data.london.gov.uk/dataset/jobs-and-job-density-borough`
- `https://data.london.gov.uk/dataset/ons-mid-year-population-estimates-custom-a`

2.2. Data Pipeline

2.2.1. pipeline.sh

This shell script orchestrates the data pipeline by installing required Python packages and triggering the 'retrieve_data.py' Python script. It ensures the presence of essential dependencies before executing the data pipeline.

2.2.2. retrieve_data.py

This Python script defines functions to connect to Kaggle, check for file existence, download missing files, and process existing files. It also includes data cleaning

procedures and creates tables in the SQLite database. Therefore, the script serves as the core component of the data pipeline.

```
def clean_dataset(df, file_info):
    """
    :param df: The dataframe which is to be cleaned.
    :param file_info: Information about the file which is to
        be processed. Retrievable from the csv_files_info.
       json.
    :return: The cleaned dataframe containing only the
       wanted data.
    """
    important_cols = file_info['new_column_names'].values()

    # Check if the column 'year' or 'date' exists
    # If yes, convert the column to datetime and filter out
       all data before 2013 and past 2016
    if 'year' in df.columns:
        df = df[df['year'] > 2013]
        df = df[df['year'] < 2017]
    if 'date' in df.columns:
        df['date'] = pd.to_datetime(df['date'], format='%Y-%
           m-%d')
        df = df.loc[(df.date.dt.year > 2013) & (df.date.dt.
           year < 2017)]
        df['date'] = df['date'].dt.strftime('%Y/%m/%d')

    # Replace empty strings and wrong entries with NaN
    df.replace('\\-|nan|\\#', np.nan, regex=True, inplace=
       True)
    df.replace('.', np.nan, regex=False, inplace=True)
    df.replace(',', '', inplace=True)
    # Convert columns that contain pound signs to numeric
       values
    columns_with_pound = [col for col in df.columns if any(
       isinstance(val, str) and '£' in val for val in df[col
       ])]
    df[columns_with_pound] = df[columns_with_pound].replace
       ({'£': '', ',': ''}, regex=True)
    df[columns_with_pound] = df[columns_with_pound].apply(pd
       .to_numeric, errors='coerce')
```

```
    # Drop all rows with NaN values
    cleaned_df = df[important_cols].dropna()
    cleaned_df = cleaned_df.astype(file_info['column_types'
        ], errors='ignore')

    return cleaned_df
```

Figure 1.: *Language Python.* The *clean_data* function takes a raw DataFrame and the as arguments. The is received from the *csv_files_info.json* file and contains all information that is needed to process a data source: from retrieval from kaggle to cleaning and reshaping the data frame and exporting it to a SQLite database table.

During the data cleaning, some minor difficulties had to be faced. Within data source 1 (2.1.1), there were some empty or invalid values which had to be eliminated in the first place. The inconsistency in the data led to the fact that some boroughs could not be included in the project since there were too little data available for them. Having a high focus on this data set due to its variety in important variables displayed, the outcome of the results of the project was influenced quite a bit, since the data of the two other data sources (2.1.2 and 2.1.3) had to be condensed in a way, that there was a coherent base for the data visualisation and interpretation. As data source 1 contained mainly data from the year 2016, data from the other data sets were adjusted respectively. For trends end development statistics, a timespan from 2014 until 2016 was set.

2.3. Variables studied

In order to gain insights into possible answers to the research question, it was necessary to define categories for both demographics and quality of life that should be investigated further. That concluded in the following list of variables.

Demographic	**Quality of Life**
Age	Crime
Gender	Transportation and Environment
Income Level	Politics (more socio-political)
Education Level	Health
Employment	
Ethinicity (BAME groups)	
Housing type	

Table 1.: List of categories that should be investigated on further within this project.

Regarding this selection of variables that were chosen to be of interest, the following analysis were planned:

- **Life Satisfaction and Demographics Analysis**
  Insights on a possible correlation between life satisfaction with the average age, as well as with the ethnicity distribution in a borough.
- **Crime Analysis**
  An examination of crime rates per borough and major crime categories provides insights into the security perceptions and safety variations among demographic segments.
- **Housing and Economic Analysis**
  Trends in house prices and gross annual pay across boroughs highlight disparities and potential economic implications for residents.
- **Education and Socio-Economic Factors**
  Proportions of education levels and correlations between educational attainment, income, and employment rates demonstrate socio-economic dynamics across boroughs.
- **Transport and Environmental Analysis**
  Insights into car ownership distribution and relationships between public transport accessibility and greenspace highlight environmental and transportation disparities.
- **Political Analysis**
  The study showcases the distribution of political seats and voter turnout in local elections, shedding light on the political landscape's representation and engagement.
- **Wellbeing Analysis**

The analysis portrays subjective well-being scores across boroughs, offering insights into life satisfaction, happiness, worthwhileness, and anxiety levels.

3. Results

3.1. Life Satisfaction and Demographics Analysis

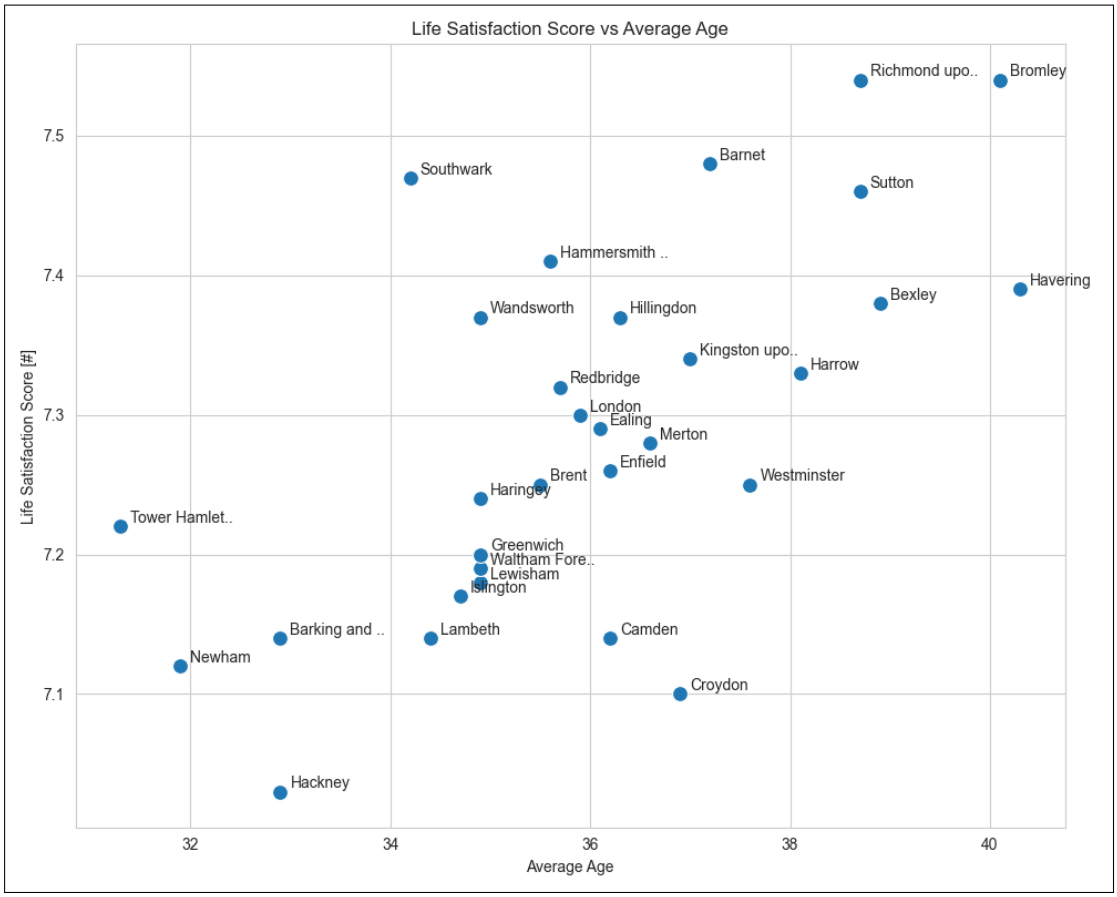


Figure 2.: Life Satisfaction scores versus the average age predominant in a borough (data from the year 2016)

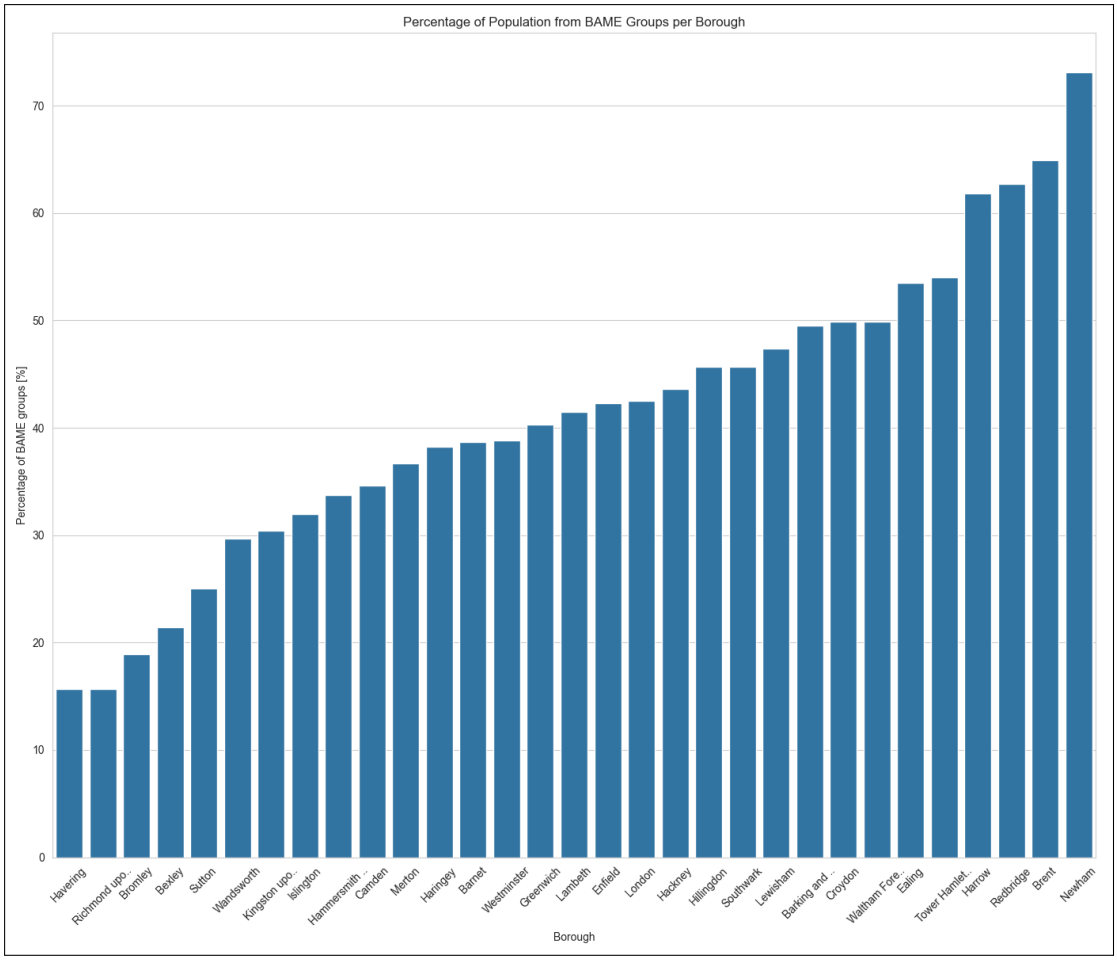


Figure 3.: The percentage of the population in each borough belonging to BAME groups

Bromley, *Richmond upon Thames*, and *Havering*, characterized by higher average ages and higher life satisfaction, also have the lowest percentages of BAME groups. Conversely, areas with lower to medium average ages, low to medium life satisfaction scores, show higher percentages of BAME groups—such as *Newham*, *Hackney*, and *Croydon*.

3.2. Crime Analysis

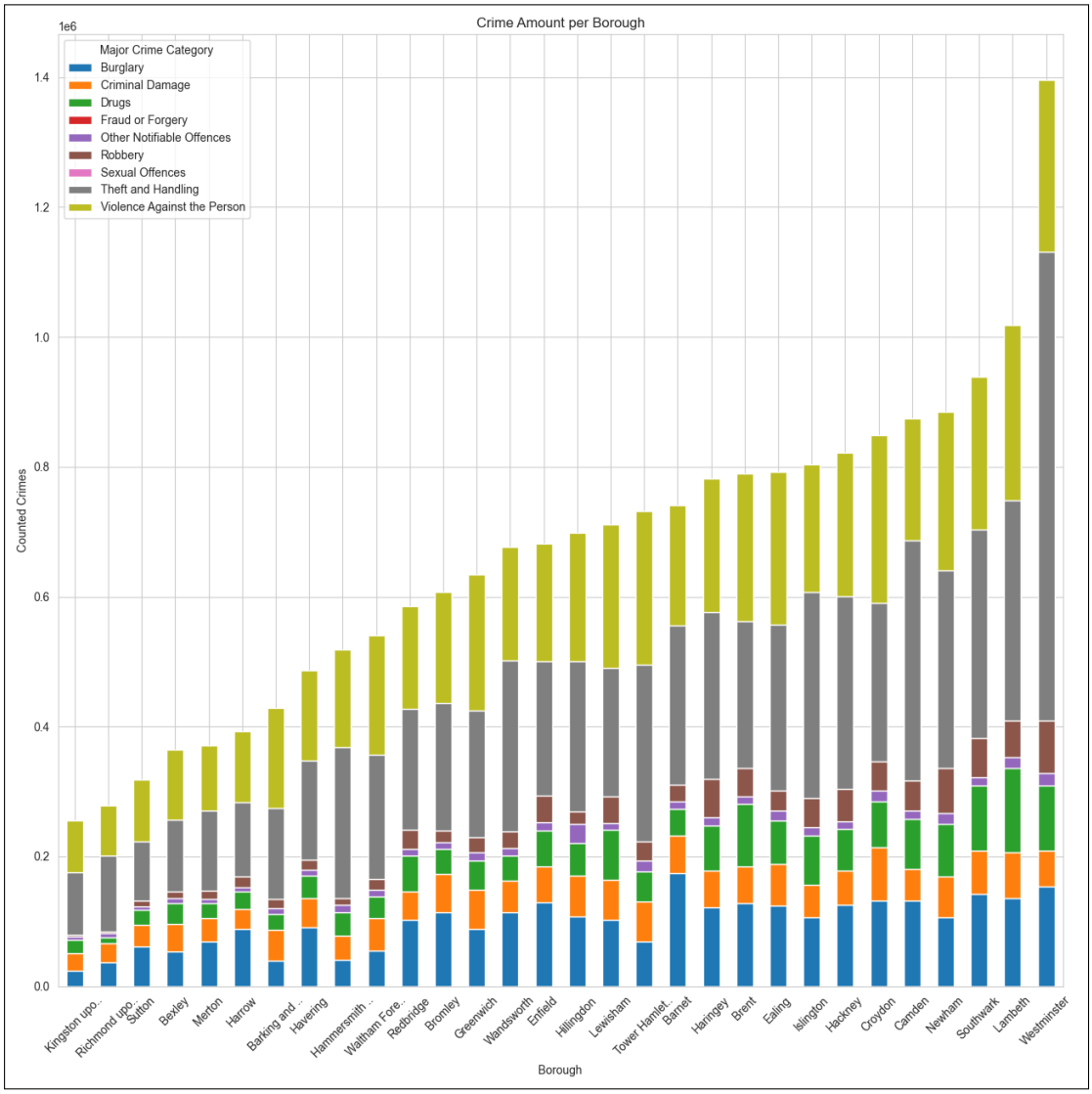


Figure 4.: The total counted crimes per borough from 2014-2016

At a quick glance, there a some aspects that stand out. The borough *Westminster* has by far the most counted crimes, compared to all other boroughs and is followed by *Lambeth* and *Southwark*. The most common major crime categories are *Theft and Handling*, *Violence Against the Person* and *Burglary* (in descending order), which seems to apply to all boroughs. The boroughs with the least counted crimes are *Kingston upon Thames*, *Richmond upon Thames* and *Sutton*.

3.3. Housing and Economic Analysis

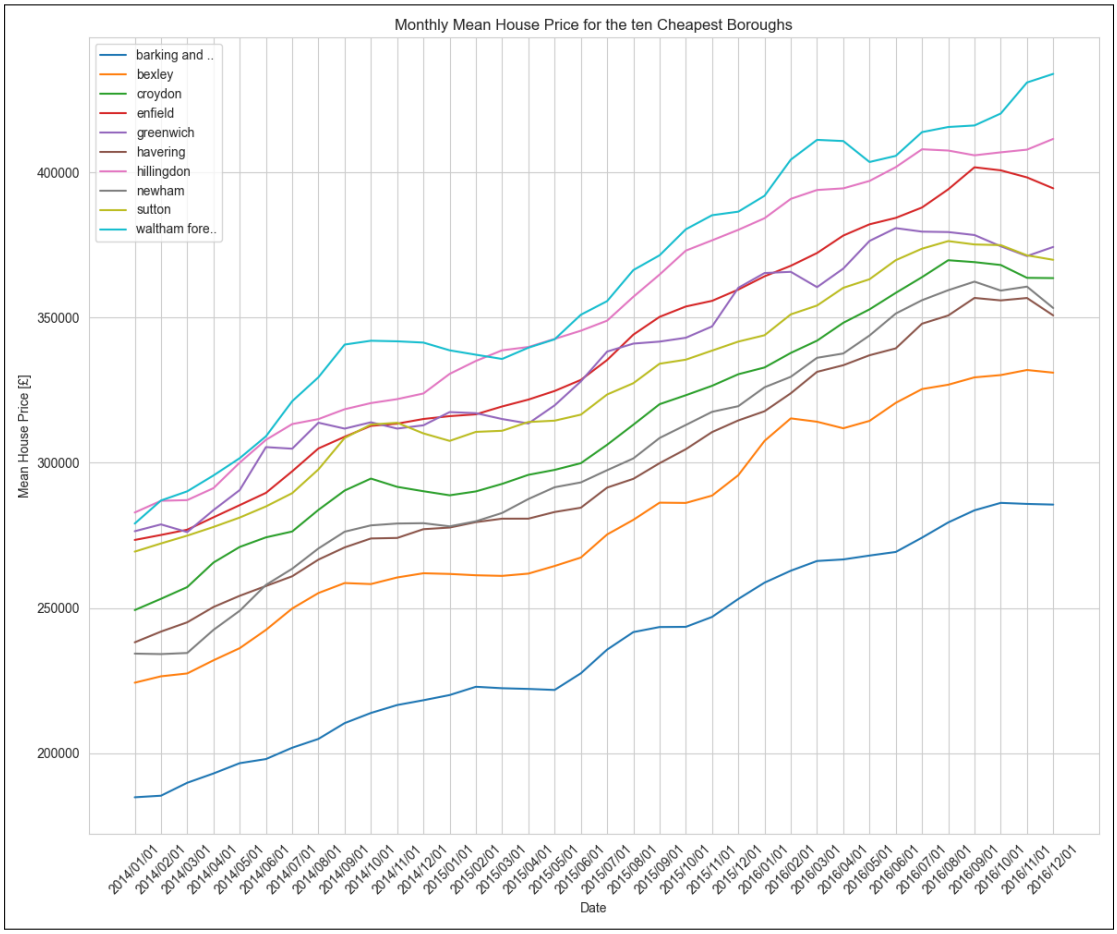


Figure 5.: The monthly mean house price for the ten cheapest boroughs from 2014-2016

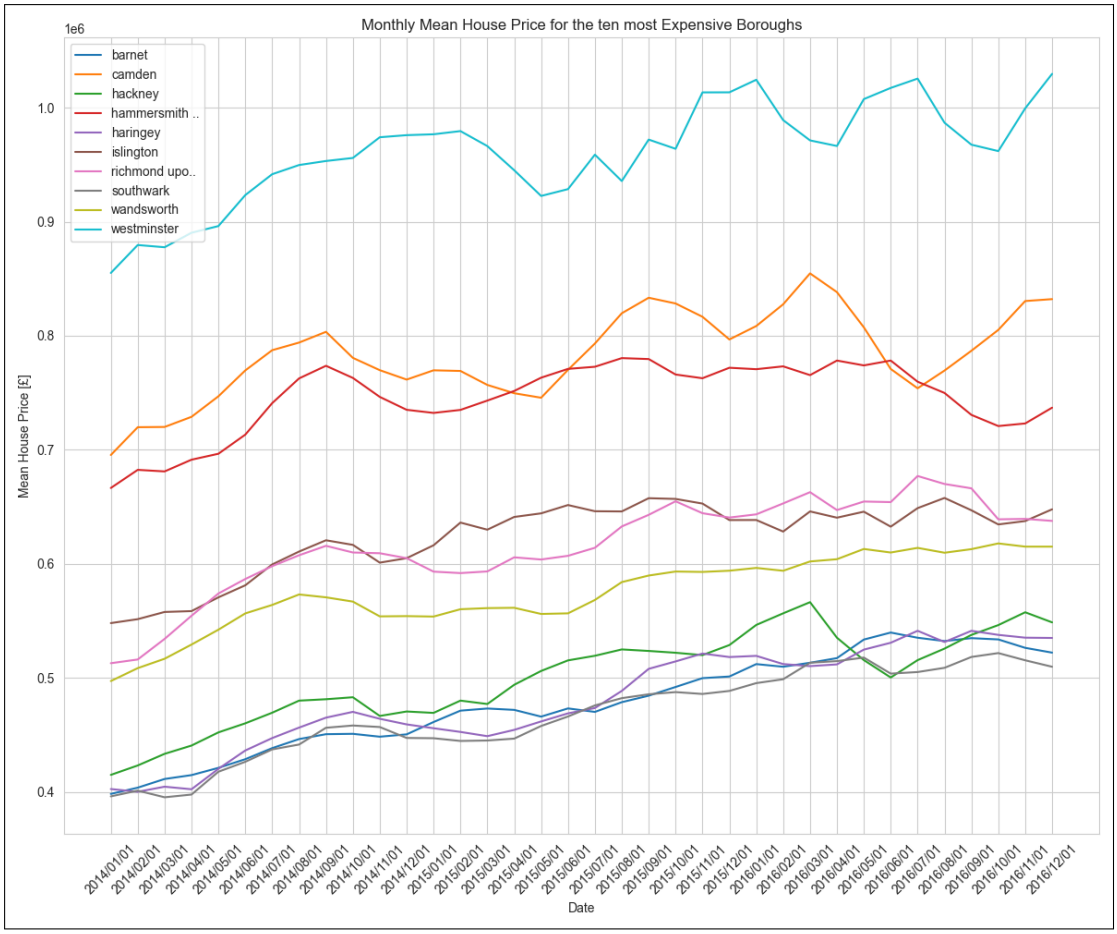


Figure 6.: The monthly mean house price for the ten most expensive boroughs from 2014-2016

Regarding the mean house prices during the years from 2014 to 2016, the cheapest boroughs were *Barking and Dagenham* and *Bexley* and the most expensive ones were *Westminster* and *Camden*. Both plots show a significant growth in the mean house price during these two years, whereas especially the most expensive house prices fluctuated a lot more than the cheaper ones.

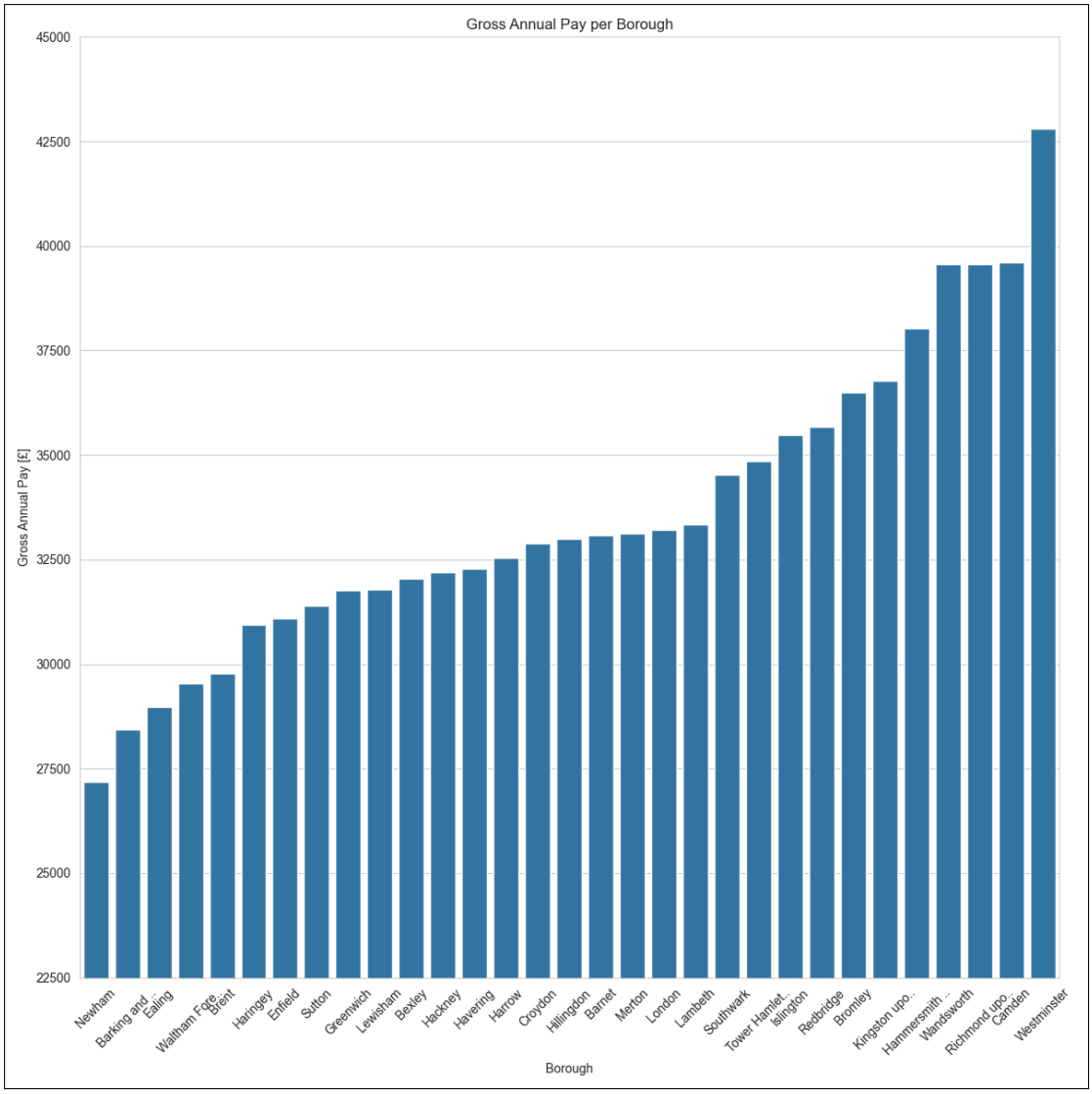


Figure 7.: The gross annual pay per borough in the year 2016

We see that *Westminster*, *Camden*, *Richmond upon Thames*, and *Wandsworth* are characterized by higher levels of annual pay, while *Newham*, *Barking and Dagenham*, along with *Ealing*, are associated with comparatively lower annual pay levels.

3.4. Education and Socio-Economic Factors

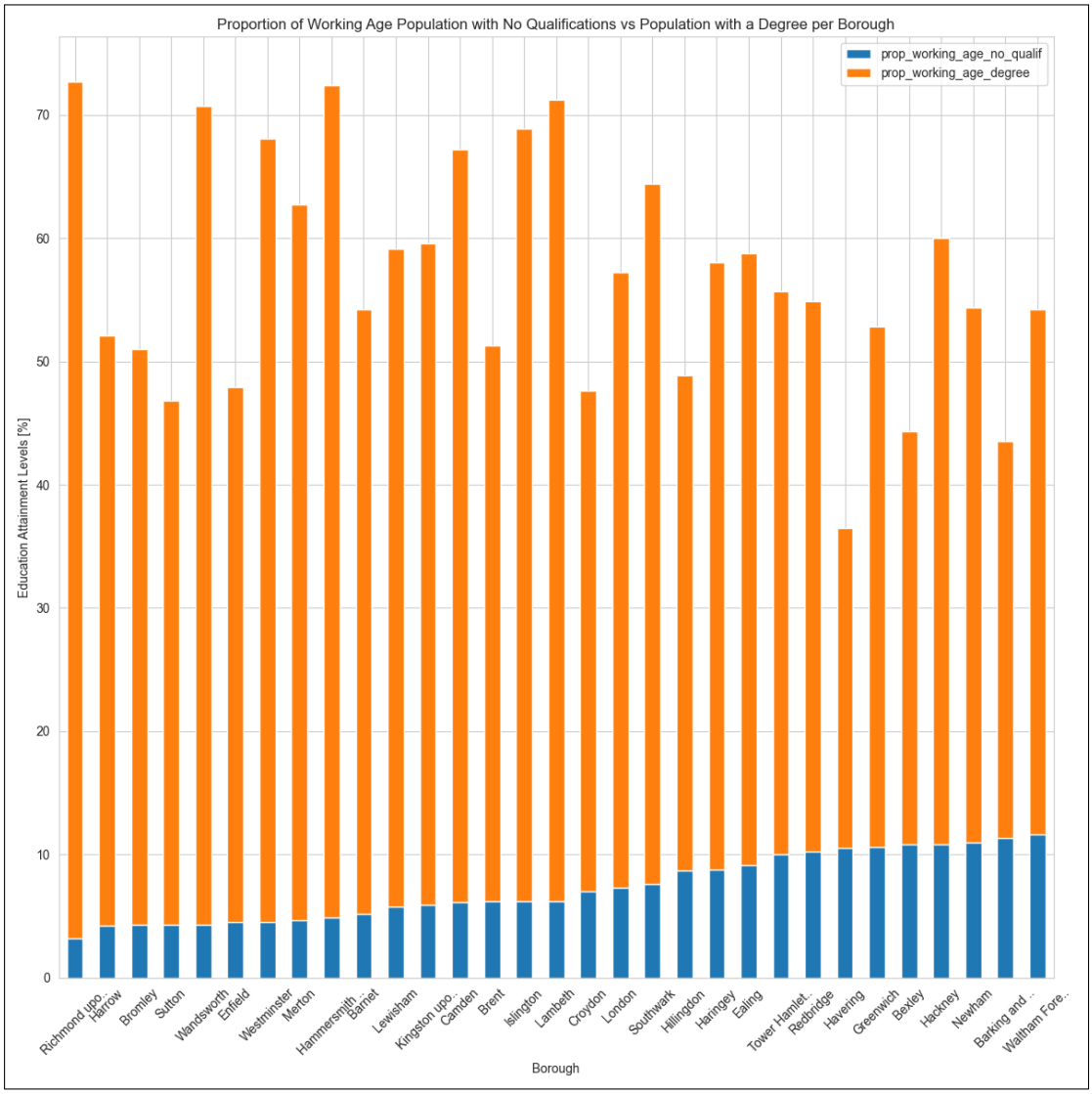


Figure 8.: The proportion of people in the year 2016 that are in their working age and either have no qualification at all or a degree or equivalent and above. The statistics takes measure across all boroughs.

The distribution of education levels across the people in the working age leads us to the conclusion, that *Richmond upon Thames* has the least uneducated and simultaneously the highest amount of people with a degree and above in the group of working age people. Further, it becomes apparent that in *Havering* the total amount of people in their working age is significantly lower compared to the other boroughs.

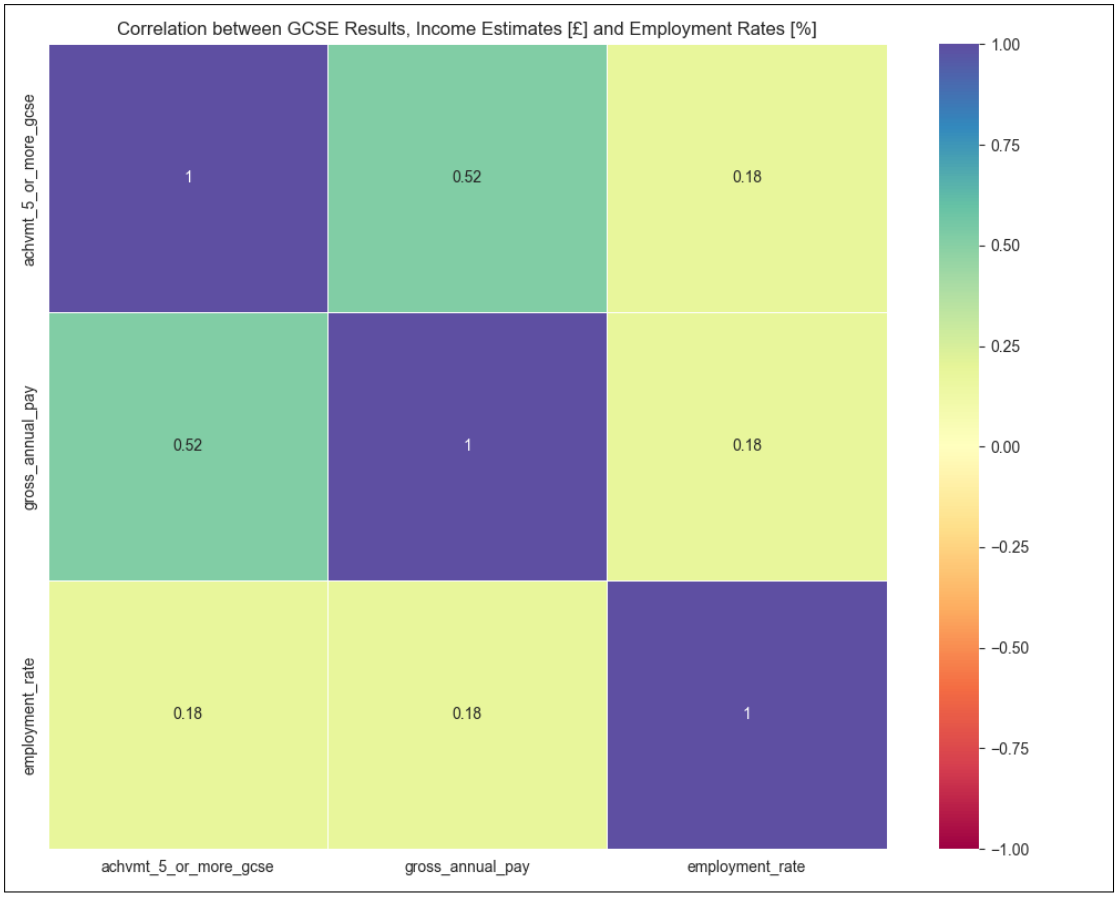


Figure 9.: The correlation between the pupils achievements, the income estimates and employment rates for all boroughs (data from 2016).

3.5. Transport and Environmental Analysis

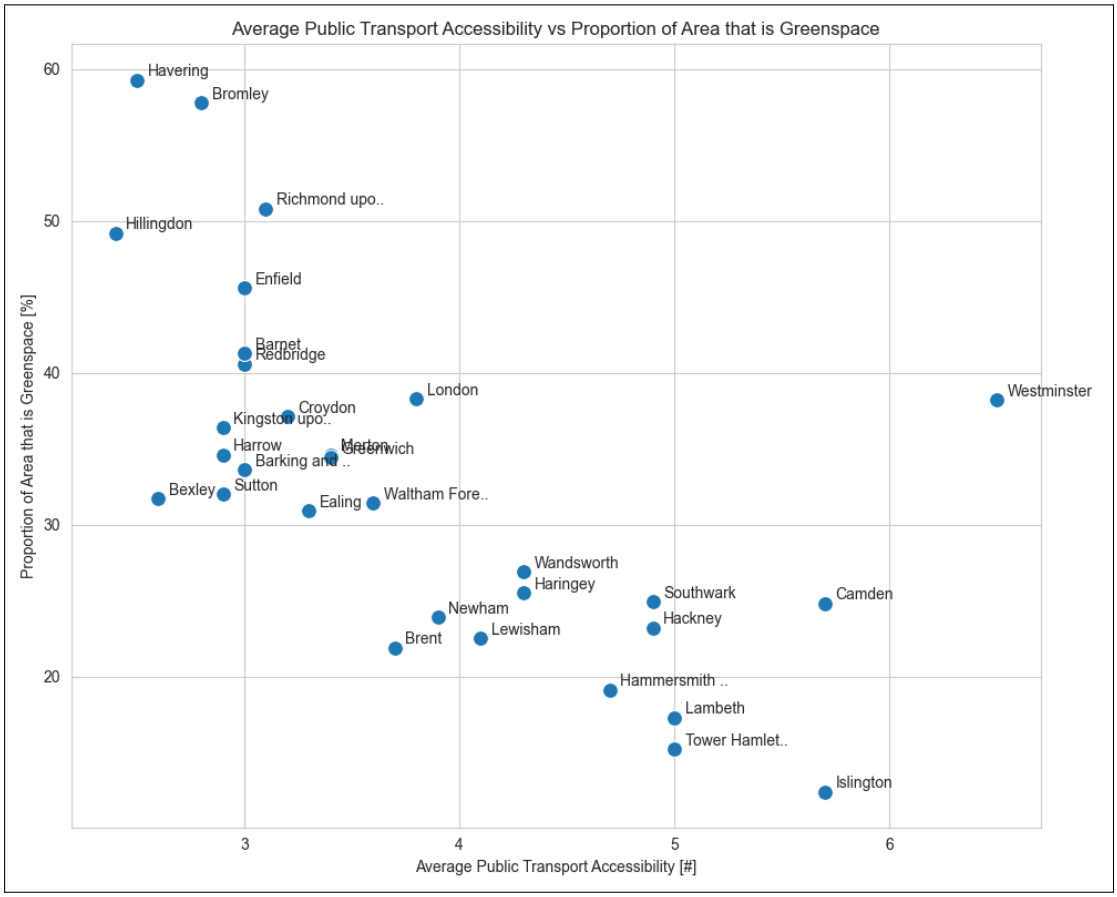


Figure 10.: The percentage of area that is greenspace versus the average public transport accesibility across the boroughs in the year 2016.

It seems that lack of area that is greenspace leads to a higher average public transport accessibility. Only *Westminster* seems to be out of line with this tendency, having a proportion of nearly 40% greenspace and a public transport accessibility close to 7.

3.6. Political Analysis

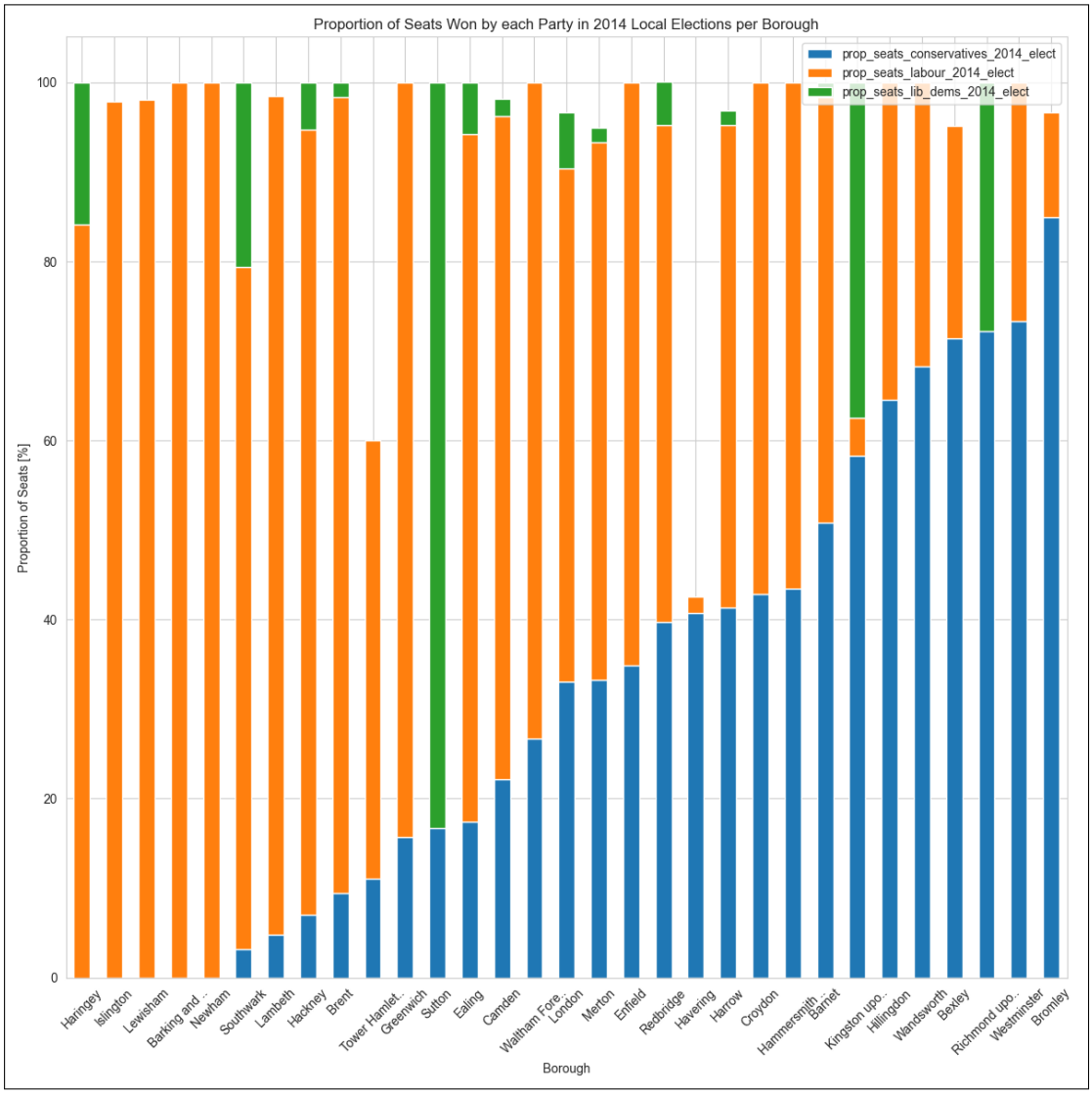


Figure 11.: The proportion of seats won by each party in the 2014 local elections per borough

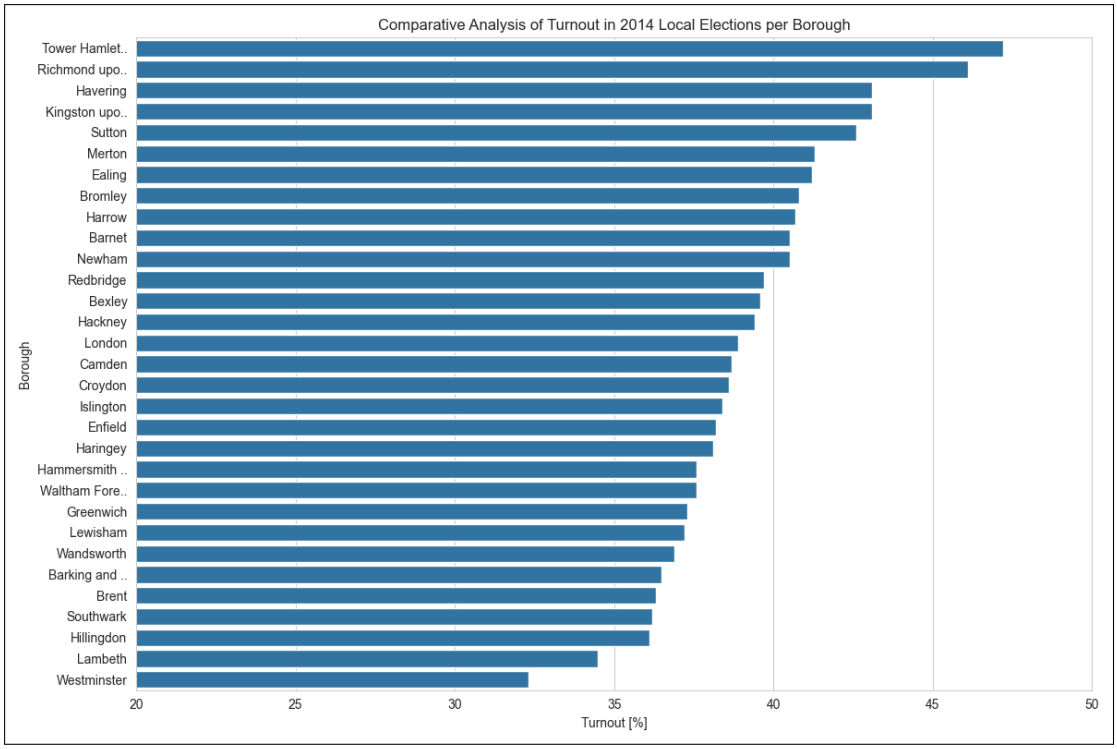
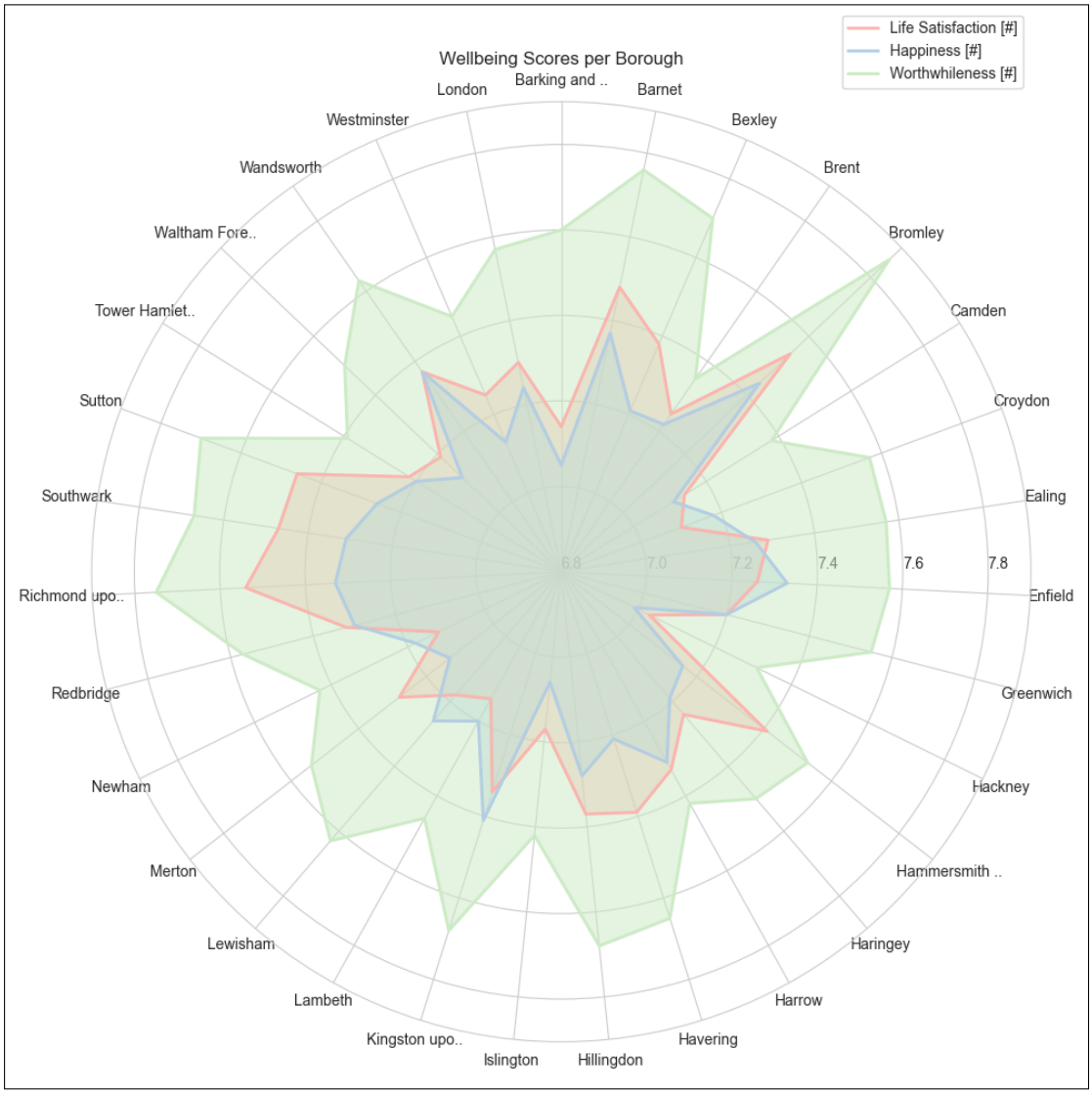


Figure 12.: The comparative turnout in the 2014 local elections across all boroughs

Regarding the local election in 2014 there are some things that catch the spectators attention. *Westminster* was recorded with the lowest voter turnout, despite having a high percentage of Democratic seats. *Bromley* and *Richmond upon Thames* both have experienced a high voter turnout and also boasted a high percentage of Democratic seats. The borough *Sutton* stands out as the only borough with a notably high percentage of Liberal Democrats seats and also a high voter turnout. *Lewisham* and *Barking and Dagenham* showed a predominant preference for the Labour party with exclusively Labour seats. However, these areas experienced low voter turnout.

3.7. Wellbeing Analysis



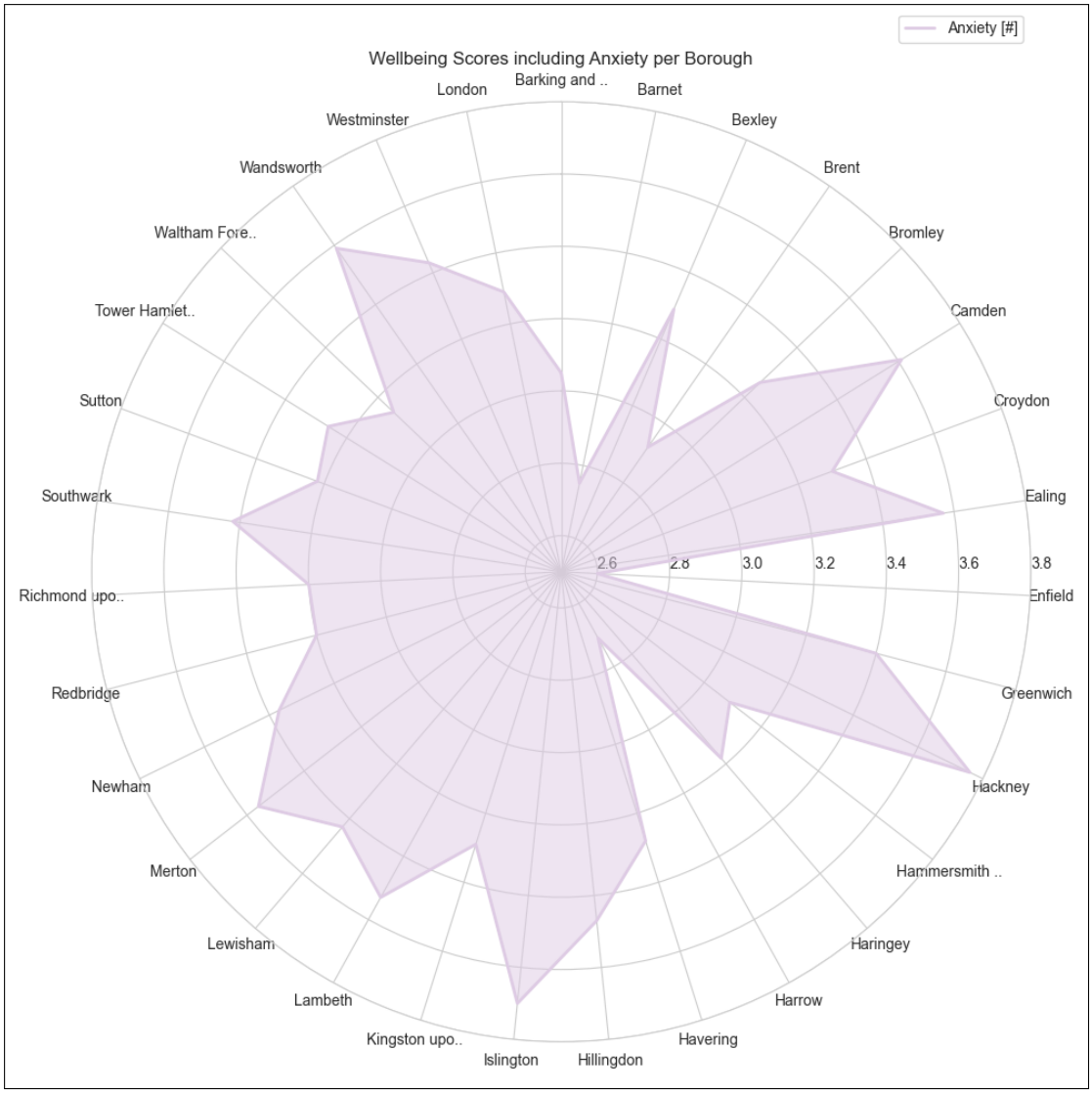


Figure 13.: The wellbeing scores per borough featuring the life satisfaction, happiness, worthwileness and anxiety

Enfield portrays moderate levels of life satisfaction, happiness, and feelings of worthwhileness. Despite not having the highest levels of these positive indicators, Enfield stands out for having the lowest anxiety levels among the compared regions. *Bromley* emerges with the highest levels of life satisfaction, happiness, and feelings of worthwhileness. While these indicators are notably high, Bromley also experiences a moderate level of anxiety, suggesting a fairly positive overall mental well-being but not without some level of anxiousness. *Hackney* reflects the opposite trend compared to *Bromley*. It exhibits the lowest levels of life satisfaction, happiness, and feelings of worthwhileness among the listed areas. Additionally, Hackney records the highest anxiety levels, indicating a notably less positive mental well-being compared to Enfield and Bromley.

4. Discussion

4.1. Interpretation of Results

1. **Life Satisfaction 3.1**
   The data reflects potential associations between demographics and life satisfaction. Areas with higher average ages and lower proportions of BAME groups, such as *Bromley*, *Richmond upon Thames*, and *Havering*, tend to report higher life satisfaction. Conversely, areas like *Newham*, *Hackney*, and *Croydon*, with lower average ages and higher proportions of BAME groups, tend to report lower life satisfaction scores.

2. **Crime 3.2**
   Boroughs like *Westminster*, *Lambeth*, and *Southwark* reported higher crime rates, primarily involving theft, violence against individuals, and burglary. In contrast, *Kingston upon Thames*, *Richmond upon Thames*, and *Sutton* had lower reported crimes. Crime rates can significantly impact the perceived safety and overall well-being of residents in different boroughs, thereby affecting their quality of life.

3. **Housing and Economic 3.3**
   Variations in mean house prices from *Westminster* and *Camden* as the most expensive to *Barking and Dagenham* and *Bexley* as the least expensive indicate disparities in housing affordability across boroughs. Additionally, differences in annual pay levels between *Westminster*, *Camden*, *Richmond upon Thames*, *Wandsworth*, and *Newham*, *Barking and Dagenham*, and *Ealing* highlight income inequality. These economic discrepancies can significantly impact residents' standards of living and overall quality of life.

4. **Education and Socio-Economy 3.4**
   Regarding education and income, a strong correlation (0.52) exists between gross annual pay and academic achievement (GCSE grades). Higher pay appears closely linked to better GCSE grades, while the relationship between achieving higher GCSE grades and the employment rate, though present

(0.18), is weaker compared to the association between pay and academic achievement. Areas like *Richmond upon Thames* displayed higher educational attainment among the working-age population, potentially correlating with better job opportunities and higher quality of life. Conversely, lower qualification levels in other areas might contribute to employment challenges and lesser opportunities, impacting overall well-being.

5. **Transport and Environment 3.5**
   The negative correlation between greenspace proportion and public transport accessibility, except for *Westminster*, suggests potential trade-offs between green areas and transportation convenience. These aspects, intertwined with access to parks, pollution levels, and environmental factors, might influence residents' quality of life perceptions, particularly related to health and well-being.

6. **Politics 3.6**
   The political landscape in different boroughs reveals varying voter turnout, party preferences, and seat distributions. Areas like *Westminster* showed high Democratic seats despite low turnout, whereas *Bromley* and *Richmond upon Thames* experienced both high voter turnout and a significant Democratic presence. Additionally, *Sutton* stood out with a notably high percentage of Liberal Democrat seats. These political trends suggest diverse civic engagements and political preferences across boroughs, potentially influencing residents' perceptions of quality of life.

7. **Wellbeing 3.7**
   In terms of mental well-being indicators, *Enfield*, showing moderate positive indicators, records the lowest anxiety levels. *Bromley*, with the highest positive indicators, experiences moderate anxiety. Conversely, *Hackney*, with the lowest positive mental well-being indicators, reports the highest anxiety levels. This correlation suggests that higher levels of life satisfaction, happiness, and feelings of worthwhileness might be associated with lower anxiety levels across these areas, emphasizing the potential relationship between mental well-being indicators and anxiety within these regions.

4.2. Key Takeaways

The data strongly implies that demographic subdivisions within an urban area significantly correlate with the general quality of life. Higher average ages, lower proportions of BAME groups, higher education levels, higher income, and positive

mental well-being indicators appear to align with higher life satisfaction, lower anxiety levels, and potentially better academic achievements.

Areas characterized by older populations, lower ethnic diversity, higher education levels, and higher incomes tend to exhibit higher life satisfaction, while regions with younger populations, higher proportions of BAME groups, lower education levels, and lower incomes tend to report lower life satisfaction scores. Moreover, there seems to be a relationship between mental well-being indicators such as happiness, worthwhileness, and anxiety levels, suggesting an interplay between these factors and overall quality of life in different urban areas.

Therefore, the research question, "To what extent does a demographic subdivision of an urban area correlate with the general quality of life?" is supported by the data, indicating that demographic factors play a substantial role in shaping the overall quality of life within urban areas.

4.3. Limitations and Future Work

4.3.1. Limitations

1. **Data Availability and Quality**: The analysis heavily relies on available data, which might vary in completeness and accuracy across different sources. Incomplete or biased datasets could impact the accuracy of correlations and findings.

2. **Temporal Factors**: The data might represent specific time frames, and changes over time may not be adequately captured. Longitudinal studies could provide a clearer understanding of trends and changes in quality of life indicators.

3. **Complexity of Factors**: Quality of life is multifaceted and influenced by various interconnected factors. This analysis considers several aspects, but other critical elements such as healthcare, community cohesion, cultural amenities, etc., might also significantly impact quality of life.

4. **Geographical Factors**: The analysis is borough-centric; however, smaller geographical units or different spatial analyses might offer more nuanced insights into localized variations in quality of life.

4.3.2. Future Work

1. **Qualitative Research**: Complementing quantitative data with qualitative studies such as surveys, interviews, or focus groups could offer deeper insights into residents' perceptions and experiences related to quality of life.
2. **Comparative Analysis**: Comparing London's boroughs with other urban areas or international cities could provide a broader perspective on urban quality of life dynamics and potentially identify best practices.

A. Appendix

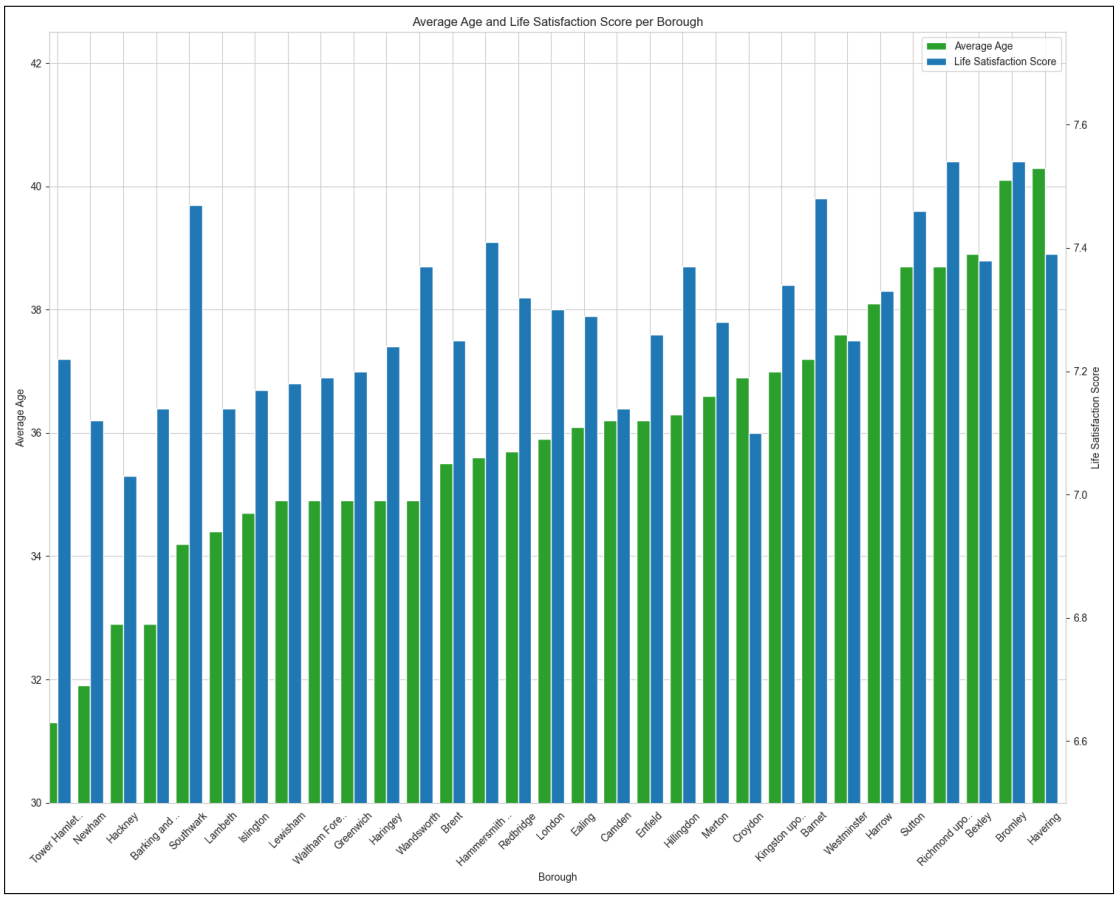
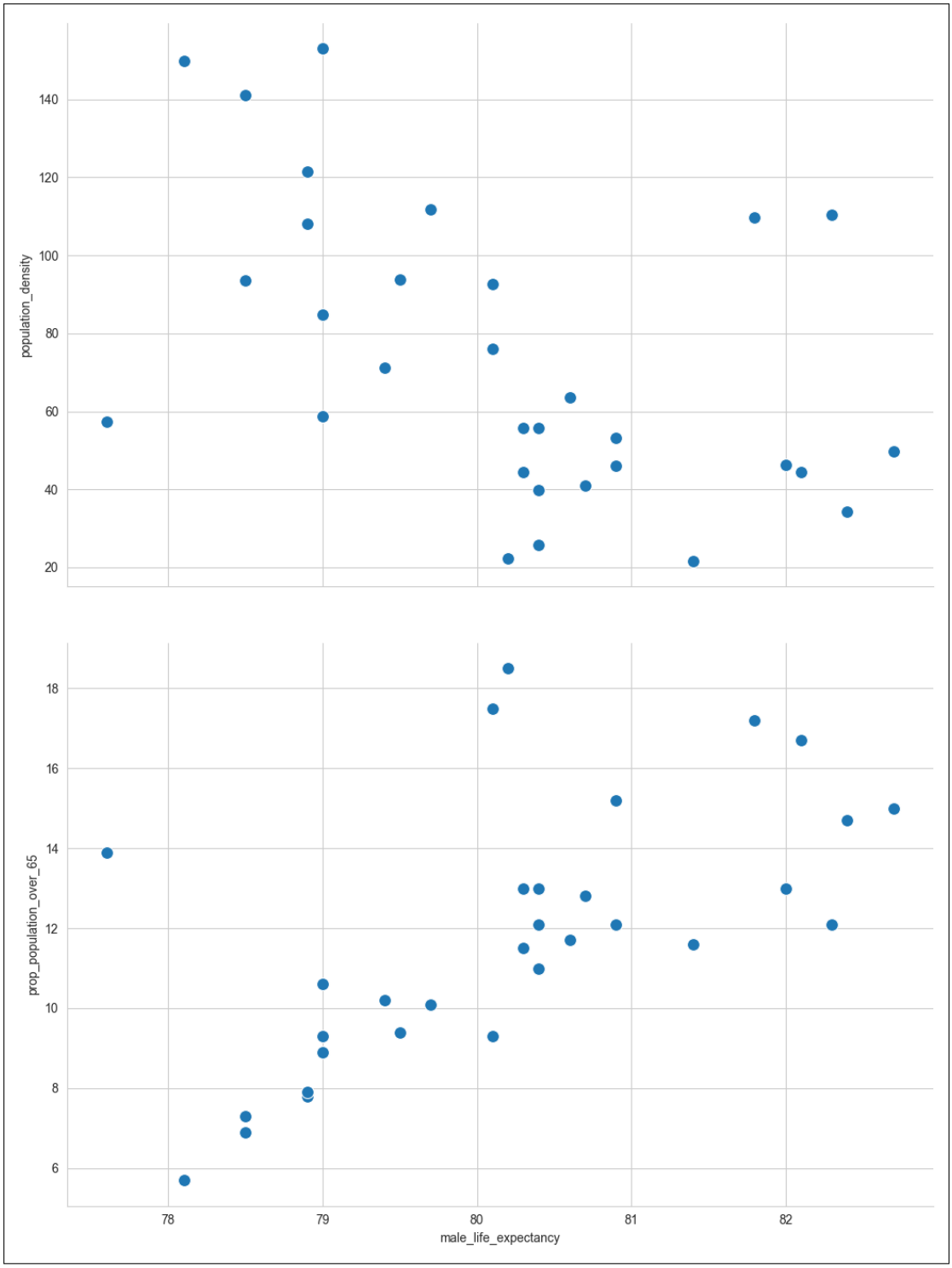


Figure 14.: The average age and the life satisfaction presented for each borough


population_density
140
120
100
80
60
40
20
prop_population_over_65
18
16
14
12
10
8
6
78
79
80
81
82
male_life_expectancy

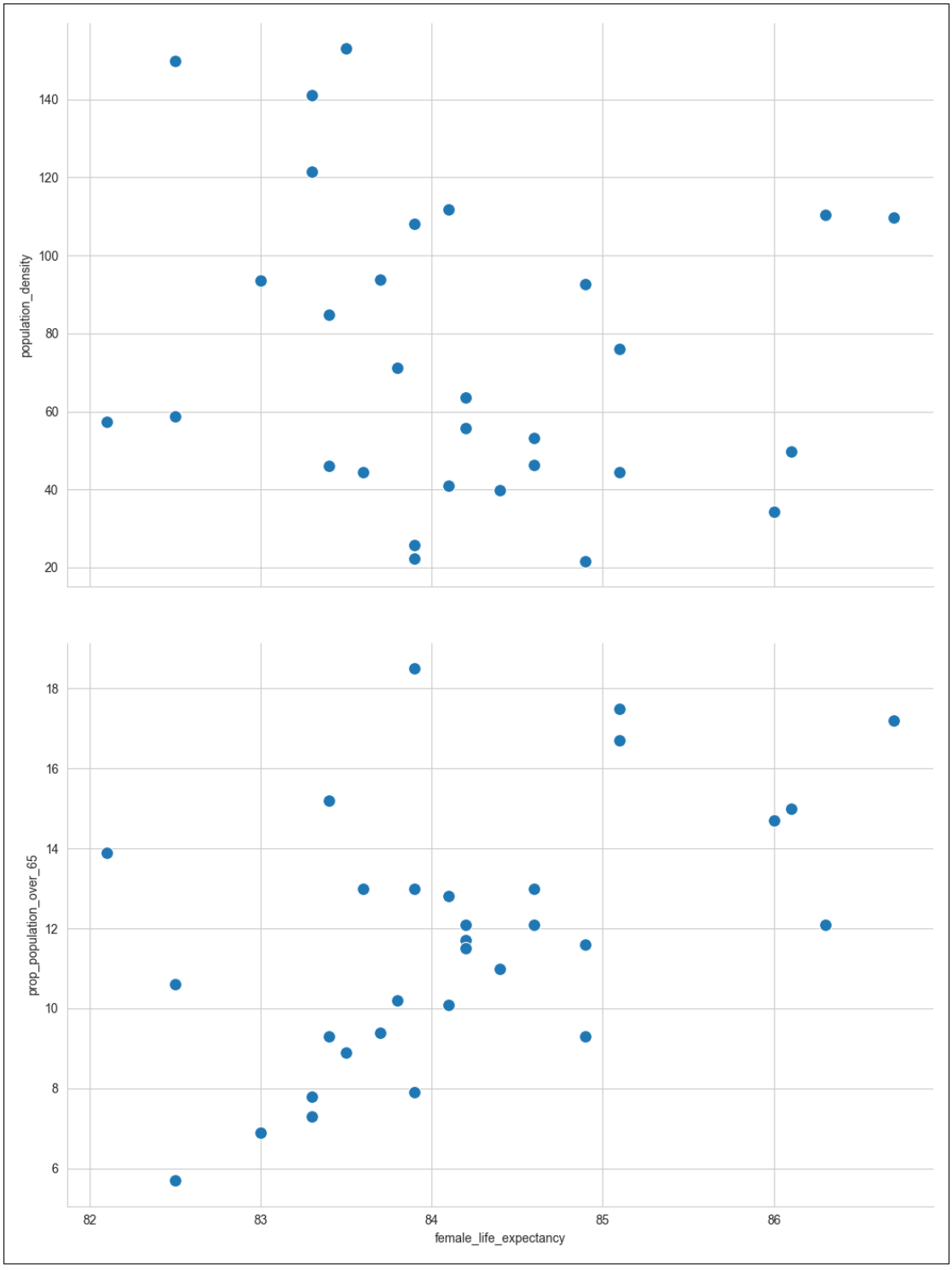


Figure 15.: The average male and female life expectancy versus the population density and the proportion of population over 65 in the boroughs in the year 2016

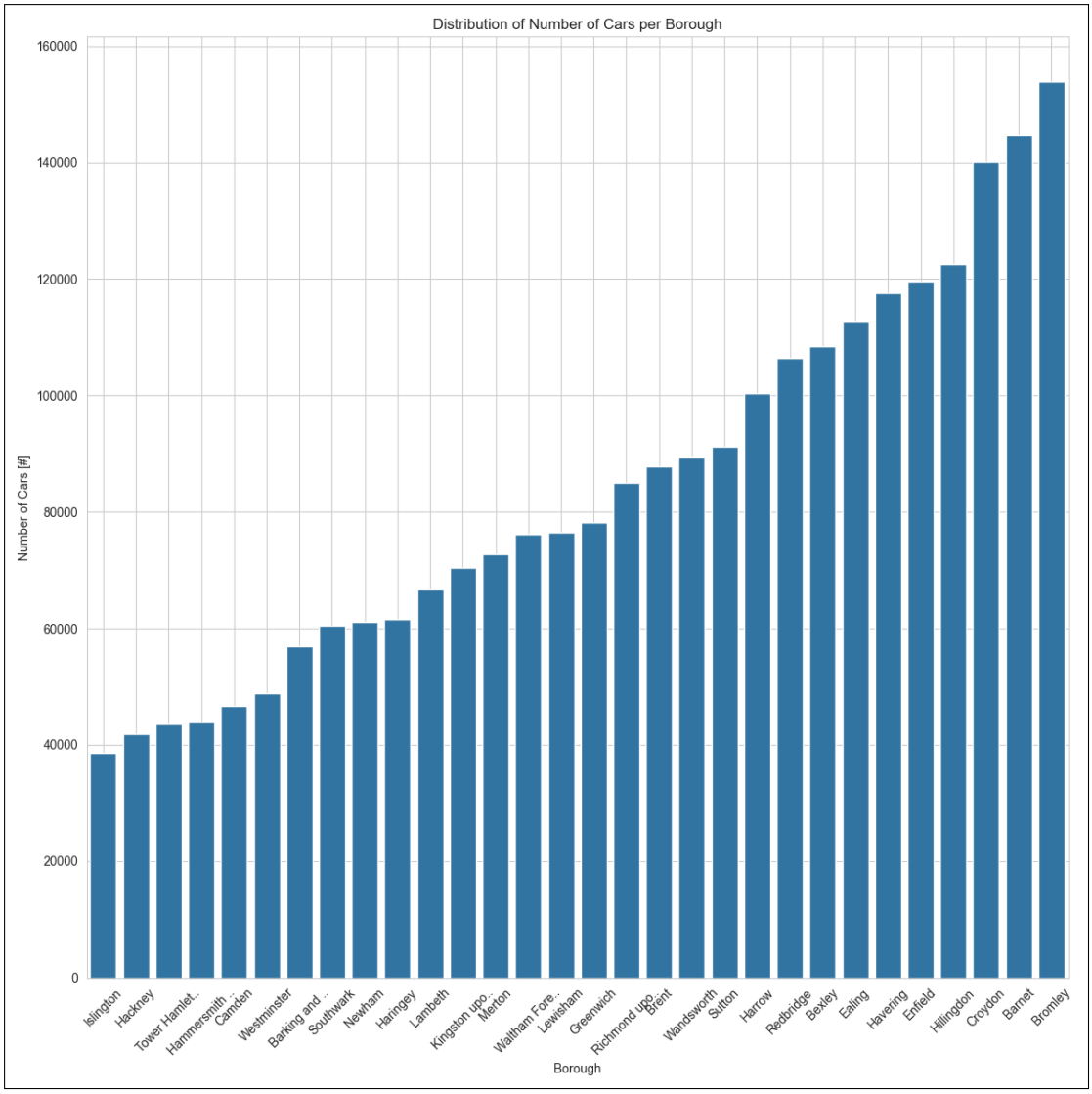


Figure 16.: The distribution of the number of cars per borough

List of Figures