CEGE0097 Project Report: Site selection for bubble tea stores in London based on the built environment and socioeconomic factors.

Group 12
Jiayi Zhao 22095995
Zimo Guo 22055869
Liting Wang 22046528
Xue Li 22086637

Supervised by Dr James Haworth

Civil, Environmental and Geomatic Engineering
University College London

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter will set out the aims and objectives of the project, explaining the overall intention of the project and specific steps that will be taken to achieve that intention.

1.1 Background and Motivation

Bubble tea is milk tea with bubbles. In the 1990s, bubble tea spread all over the world. Especially in East Asia, such as Singapore, Japan, and Korea[1].

In the business world, the location of a bubble tea shop can really determine its fate. The choice of location can determine the fate of a bubble tea shop[2]. Bubble tea is just an embellishment of life, an embellishment after people have reached a certain level of consumption, not a necessity, so customers will not go out of their way for a cup of bubble tea. Therefore, the choice of location is crucial. Currently, most site selection relies on experience and intuition, which lacks efficiency and accuracy. But this problem can be solved with Geographic Information System(GIS).

The main consumers of bubble tea shops are mainly young people, so in addition to considering the flow of people, whether these people are more young people also needs to be considered, if the flow of people is more than expensive, but no young people, the market is still difficult to open[3]. In addition, it is important to look at the Built environment and Socioeconomic aspects. In this report the built environment consists of transportation, land use, recreation, restaurants, and peer shops (i.e., coffee shops).

Socioeconomic will then be analysed in terms of demographic and cost.

1.2 Aims and Objectives

The aim of the project was to develop a method of selecting a potentially suitable location for a new bubble tea shop that would maximise profits[4].

$$Profit = Revenue – Costs \tag{1.1}$$

In order to achieve this, there were several objectives.

1. Identifying the target group
2. Finding criteria for choosing a suitable site based on the built environment and socioeconomics
3. Sourcing suitable and reliable dataset for analysis
4. Processing and visualizing the dataset using GIS
5. Obtaining the highest scoring site by adjusting the weights of criteria

Chapter 2

Data Description

The data for this study were obtained from various sources, including the UK government [5], London Datastore [6], Transport for London [7] and many other resources. These reputable institutions and platforms provide credible and reliable data. A summary of the data we used can be seen in Table 2.1. The vector data includes feature classes such as points and polygons, while the raster data has a resolution of 177×368.

2.1 Built Environment Related Data

2.1.1 Transportation and Land Use Data

Bus stations and subway stations are more accessible by distance, and the closer the distance, the more suitable. Land use is to determine the appropriate place to build bubble tea shops according to the characteristics of different land use types.

2.1.2 Recreation and Restaurant Data

The cultural infrastructures published by the Greater London Authority(GLA)[8] are the primary inspiration for the recreation infrastructures. The recreation infrastructures contain theatres, parks, pubs, museums, libraries, exhibition halls, and other places. OpenStreetMap[9] is where the information for cafes and restaurants comes from. All three datasets are point data, combined with polygon data from the London boundary[10]

Category		Criteria	Data source	Type	Description
Built-environment	Transportation	Bus stops	https://www.data.london.gov.uk/dataset/tfl-bus-stop-locations-and-routes	Point	Locations of bus stops in London.
		Subway stations	https://tfl.gov.uk/info-for/open-data-users/our-open-data	Point	Locations of subway stations in London.
		Land use	https://www.data.gov.uk/dataset/255d7b61-1a93-4a5c-8d13-2b45952ebf7e/corine-land-cover-2012-for-the-uk-jersey-and-guernsey	Polygon	Land use type.
	Recreation	Cafe	http://www.openstreetmap.org/	Point	Location of cafe in London
		Restaurant	http://www.openstreetmap.org/	Point	Location of restaurants in London.
		Recreation infrastructure	https://www.data.gov.uk/dataset/92a4192d-df25-4695-98a5-139bcd0620c8/cultural-infrastructure-map	Polygon	Spatial distribution of recreations at ward level in London.
Socio-economics	Demographics	Age	https://digimap.edina.ac.uk/	Polygon	Population aged 16 to 34.
		Social grade	https://digimap.edina.ac.uk/	Polygon	Social grade AB.
		Age (rate)	https://digimap.edina.ac.uk/	Polygon	Population aged 16 to 34 (percentage).
		Social grade (rate)	https://digimap.edina.ac.uk/	Polygon	Social grade AB (percentage).
		Student amount	https://digimap.edina.ac.uk/	Polygon	Student amount.
		Unemployment rate	https://digimap.edina.ac.uk/	Polygon	Unemplyment rate (percentage).
	Cost	Rent	https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2022#employee-earnings-data	Polygon	Using median house prices in London represents the rent of the milk tea shops. (unit: £).
		Employee salary	https://data.london.gov.uk/dataset/average-house-prices	Polygon	Using weekly earnings for full-time employees in London represents salary that the milk tea shop owner need to pay. (unit: £)

Table 2.1: Sources and descriptions of data.

for processing and analysis.

The data for recreational infrastructure contains 12 cultural infrastructures in London from 2018 to 2020, with 6,579 features. And London's restaurant dataset consists of the top 100 nearby (with a radius of 2 Km) restaurants to the borough centre, for a total of 3,133 features. The cafes dataset consists of 1,017 features containing all the cafes in London.

2.2 Socioeconomics Related Data

2.2.1 Demographic Data

The 2011 Census data from Digimap is the primary data utilized for the demographic analysis, which can be found at https://digimap.edina.ac.uk/roam/download/society. It contains data on age, economic activity, and social grade used in bubble tea site selection. The data are separated by borough. Household Reference Persons in the United Kingdom between the ages of 16 and 64 are grouped in the 2011 Census data. The estimations are up to date as of March 27, 2011, Census Day [11]. The whole data from this census is accessible in hundreds of separate datasets (or tables), covering the full range of population characteristics and subject areas. The 2011 Census provided the age, economic activity, and social grade information used in this site selection.

Age data offers essential demographic data that all users of census data need. A single year's age data can be used to create intricate profiles and trends that are crucial for local governments and other public bodies in influencing the funding and delivery of services for certain population groups, such as children and the elderly. According to a recent CLSA survey of consumers, 94 per cent of thoseaged between 20 and 29 bought bubble tea in the last three months, reported by Bloomberg [12]. So, age data is considered in bubble tea site selection.

Economic activity is also considered in site selection, especially unemployment information. People who lose their jobs lose their incomes and unemployment is one of the most important factors affecting individual well-being[13]. We assumed that low income could reduce the willingness to purchase drinks like bubble tea. Economic activity is determined by whether or not a person was employed or actively seeking employment in the week before the census. It offers a measurement of whether or not a person was an active participant in the labour market rather than just a basic indicator of whether or not they were currently employed. The data shows if people are employed, actively seeking employment, waiting to start a new job, available to start a new job, or whether they are neither employed nor looking for employment. The number of hours a person works and their type of employment—whether employed or self-employed—are other factors that contribute to the economic activity classification. However, we only focus on this project's employment and unemployment rates. They are the most important effects of income.

The social grade data received by Digimap is largely comprehensive and has reliable sources. The datasets have additional restrictions, though. For example, the Statistical Disclosure. Control over social grade information may have an impact on numerous aspects of source data accuracy. Some counts would indeed be affected since records in the Census database were moved to different geographic locations to prevent personal data from being released from the 2011 Census. Since the record swapping was aimed at those homes with unusual characteristics in small areas, the largest effects would typically be at the lowest geographies.

The City of London is absent, which affects how comprehensive the social grade data is. Even though the most recent data came from 11 years ago, the major goal of the study was to investigate the gap between boroughs, and a significant gap might not have developed in those 11 years. As a result, the information is still useful. In conclusion, most of the data in the 2011 Census dataset is regarded as reliable, and just a small proportion seems to be less reliable.

2.2.2 Cost Data

One of the most important factors leading to a restaurant's success is its location, however, operating costs are also of immeasurable importance. This section describes the rent cost and salary cost data in detail and the reasons we chose them. Here, we assume that we do not buy a physical store to start a milk tea shop, but lease it. In addition, we also assume that housing prices are positively correlated with the rent of milk tea shops. Median wages in the catering industry and industry-wide median wages are also assumed to be positively correlated.

Rent Data

Rental cost is one of the most important factors in determining the profitability of a restaurant [14] as well as a milk tea shop. For data classified by borough or smaller than borough, only residential rent or office rent information can be found, but no rent information for milk tea shops, or even for commercial properties. Therefore, we decided to choose London house price data instead of rent data, since rents and house prices are usually positively correlated [15]. This means that the higher the house price, the higher the rent usually is.

Employee Salary Data

The median weekly gross earnings data of full-time employees in London was used to represent the cost of shopkeepers paying staff wages. It is found that the median annual salary in the service industry is £26.6k [16], and in London it is £41.9k [17]. The latter figure is almost double that of the former, but data for the services sector is difficult to find. Therefore, we can only use the data of all industries as a reference.

2.3 London Boundary Related Data

The London boundaries from the London Datastore[10] used in this project are mainly at three levels, namely Lower Super Output Areas (LSOAs), London Wards and London Boroughs. It contains a range of key GIS boundary files for ESRI and Map Info covering Greater London.

Chapter 3

Exploratory Spatial Data Analysis

Exploratory data analysis (EDA [18]) is a method of analyzing a data set, which is used to explore the properties of the data set so that we can formulate reasonable hypotheses. In this project, we applied this approach and used it on spatial data, that is, using exploratory spatial data analysis (ESDA [19]). Many methods are used in the analysis, such as histograms, kernel density, scatterplots, and geographic representation. This chapter introduces some of the methods we used in the analysis, and the conclusions and some assumptions after the analysis.

3.1 Built Environment

3.1.1 Transportation

Two kinds of transportation data were considered in this section: bus stops and subway stations.

Bus Stops

The city of London boasts approximately 19,000 bus stops distributed throughout the city, as illustrated in Figure 3.1. Notably, there are many bus stops in the central areas of London, particularly around the River Thames. In contrast, the number of bus stops in areas closer to the outskirts is relatively low.

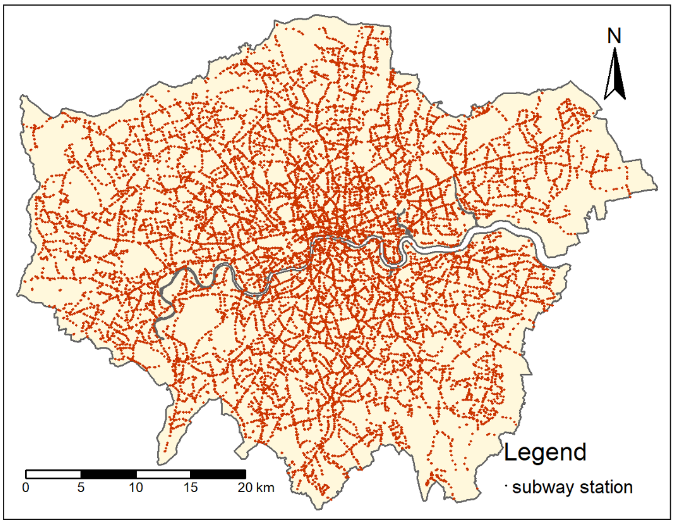

Figure 3.1: Distribution of bus stops.

Furthermore, by calculating the kernel density of bus stops, it can be seen from the results displayed in Figure 3.2 that the bus stops are clustered in the central regions of the city, with the density gradually decreasing as one moves from the centre towards the periphery.

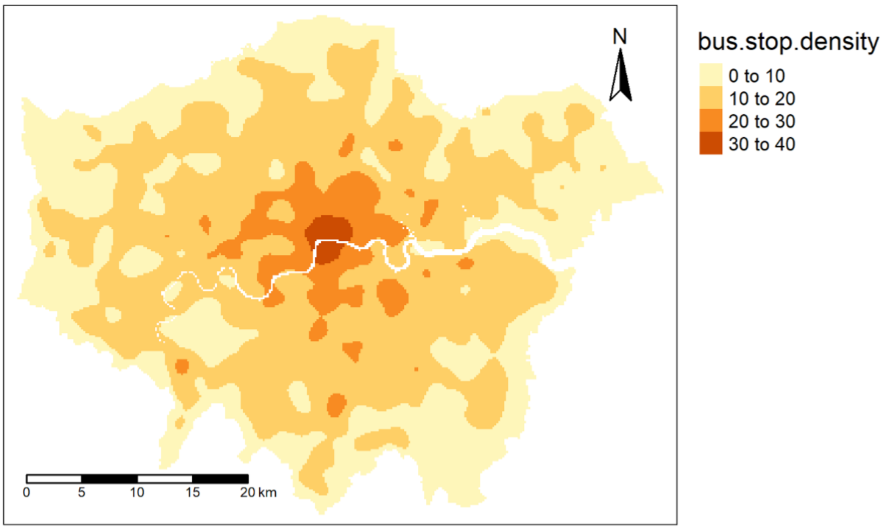

Figure 3.2: Kernel Density of bus stops.

Subway Stations

The London subway system comprises 11 lines that serve 272 stations. As depicted in Figure 3.3, most subway stations are located on the north side of the River Thames. Additionally, several lines run in a northwest-to-southeast direction, which results in a higher concentration of subway stations in the northwest areas of the city.

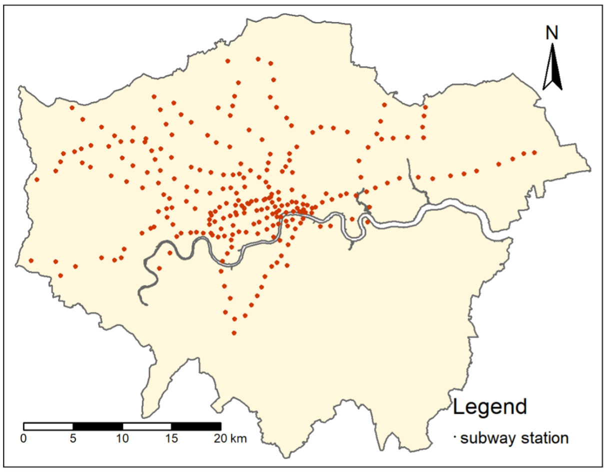

Figure 3.3: Distribution of subway stations.

The kernel density of subway stations, as shown in Figure 3.4, reveals a clustering of stations in the central region of London, indicating that it is more convenient for individuals to access the subway system in this area. As one moves further away from the centre, the distribution of subway stations becomes increasingly sparse.

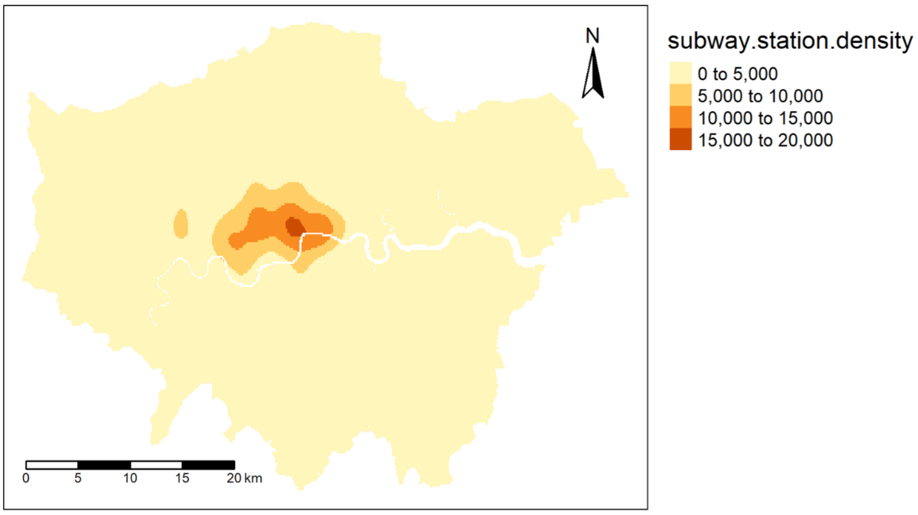

Figure 3.4: Kernel Density of subway stations.

3.1.2 Land Use

The land use in London is displayed in Figure 3.9. The data reveals that most of London's area is characterised by discontinuous urban land use. Additionally, it can be observed that industrial and commercial units are situated in proximity to estuaries and water bodies. Furthermore, non-irrigated arable land can be found in the central regions of the city. Lastly, the main airports in London are situated in the western parts of the city.

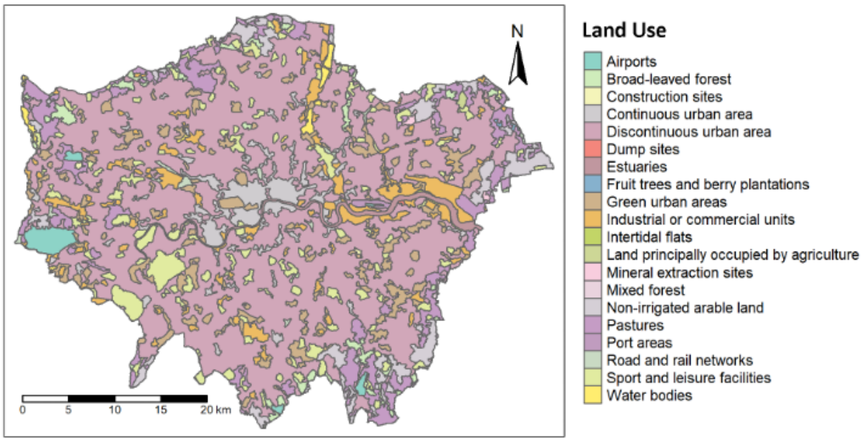

Figure 3.5: Kernel Density of subway stations.

3.1.3 Recreation

Around 6579 recreational infrastructures are spread out across London in the City of London. The map below demonstrates that there are less recreational infrastructures the further away from the suburbs you are. Libraries are more equitably placed closer to the city centre than taverns, which are distributed more centrally.

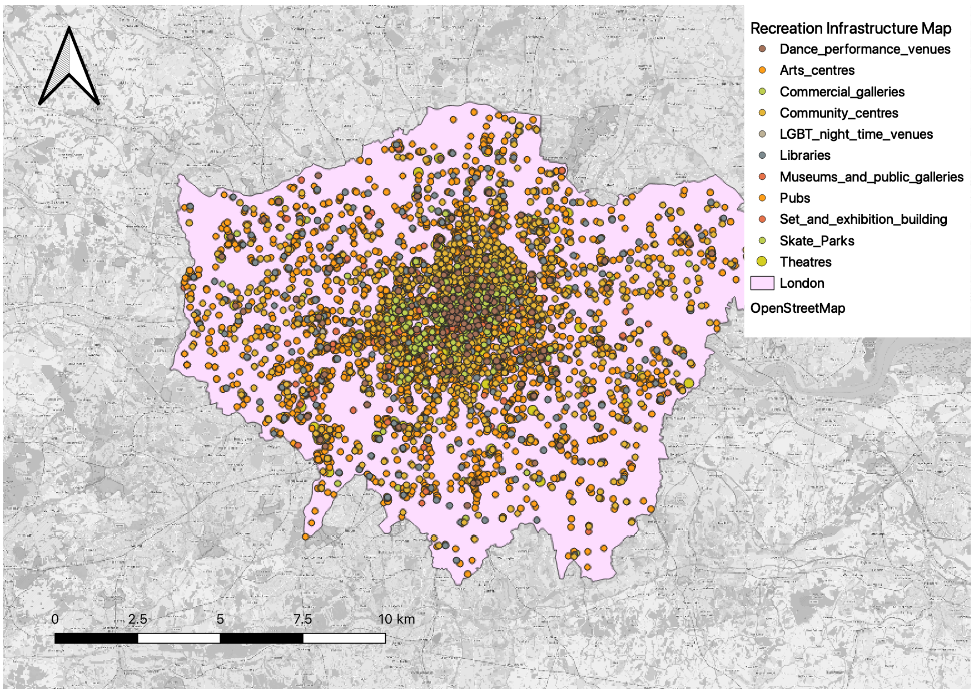

Figure 3.6: Distribution of recreation infrastructures in London

3.1.4 Restaurant

London's restaurants dataset consists of top 100 nearby (with radius of 2 Kilometers) restaurants to borough centre. Most restaurants are located on the east and west sides of the Thames as shown in Figure 3.7.

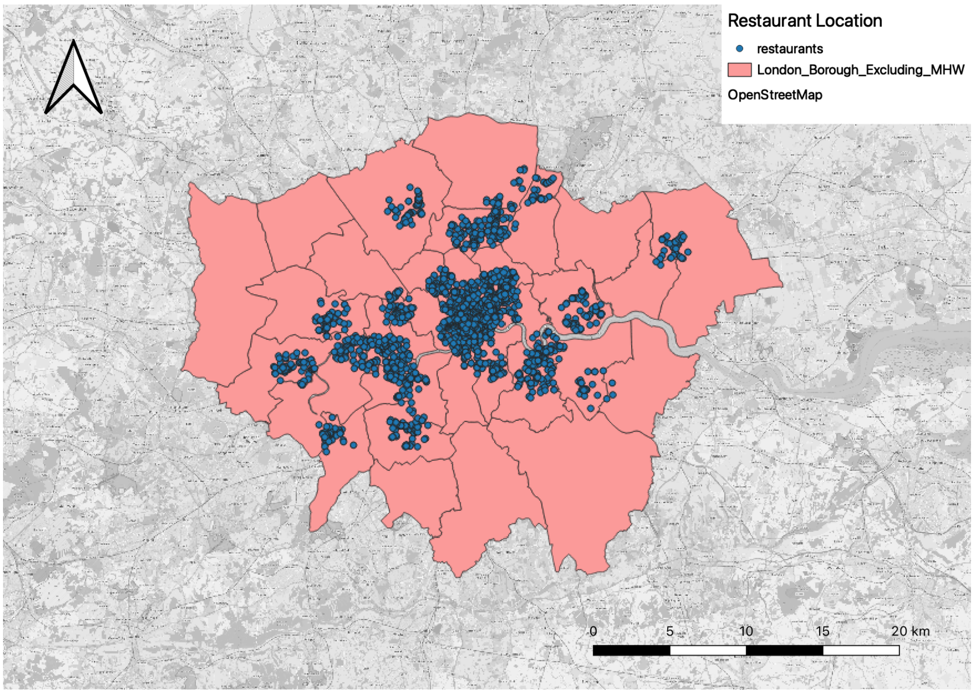

Figure 3.7: Distribution of restaurants in London

The restaurant heatmap in Figure 3.8 shows its clustering in the central areas of the London borough, indicating that these areas are more likely to attract footfall. As people move further away from the centre of the borough, the distribution of restaurants becomes increasingly sparse.

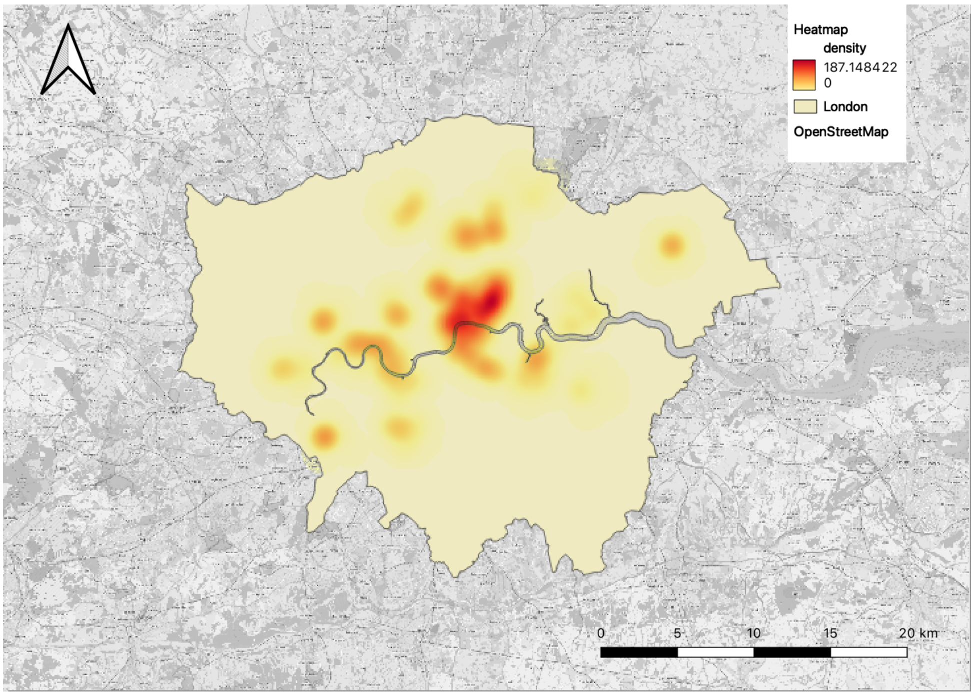

Figure 3.8: Heatmap of restaurants in London

3.1.5 Cafe

Figure 3.9 depicts the location of cafes in London. According to the data, cafes are primarily located in the city's centre, with the density drastically dropping as one approaches toward the outskirts of city.

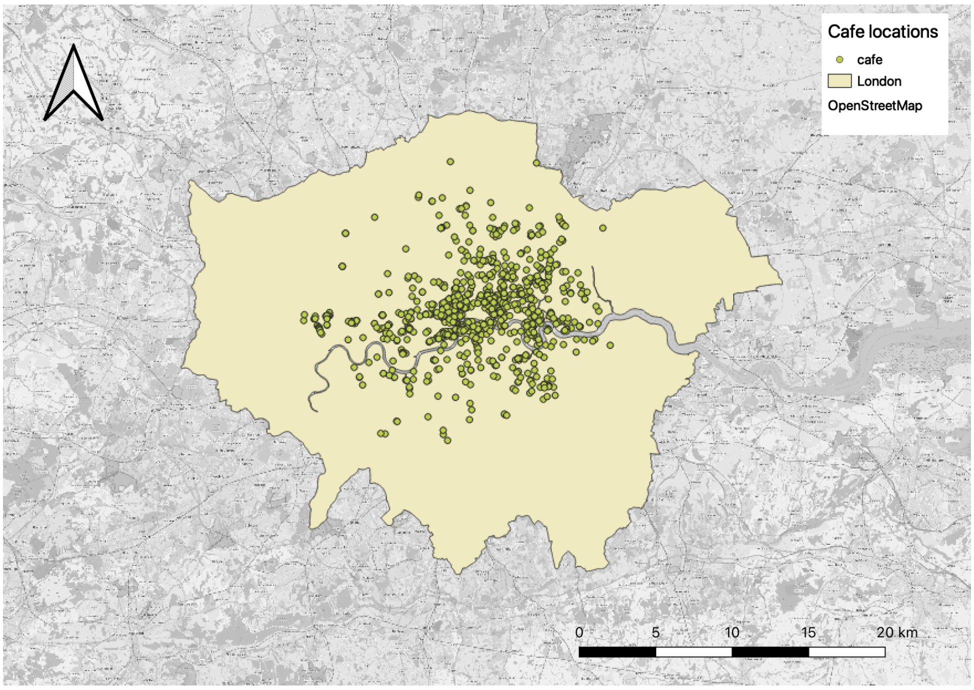

Figure 3.9: Distribution of cafe in London

3.2 Socioeconomics

3.2.1 Demographic Data

For the age, economic activity, and social grade data, firstly we loaded the data in R and simply plotted the data to check the completeness of the dataset. As it was shown in Figure 3.10, the City of London is absent from the dataset and the corresponding area is white due to the missing data.

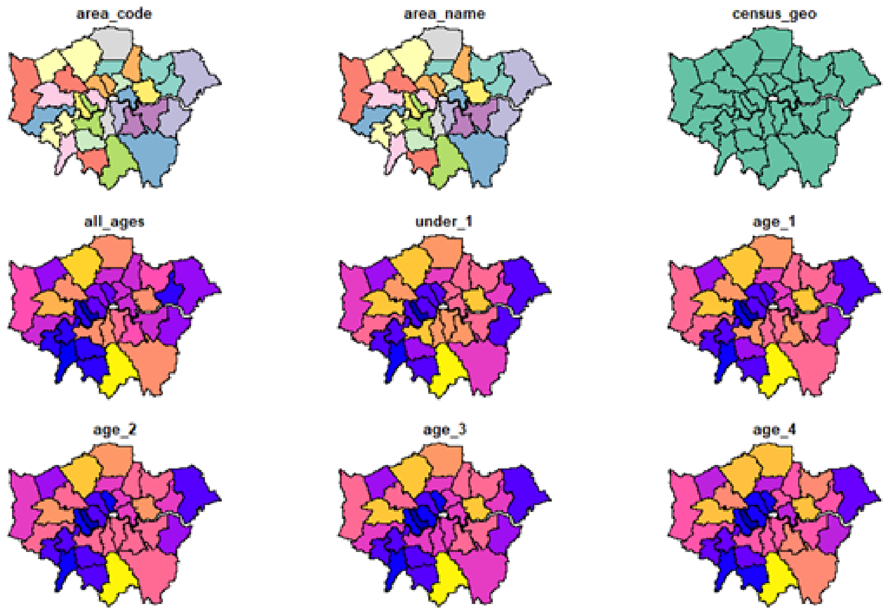

Figure 3.10: Data stored in R

The Summary function in R was used to present detailed statistical information on age, economic activity, and social grade data. For example, we can obtain age statistical data between 16 and 34 in Figure 3.11. The maximum value is three times the minimum value, representing that the distributions of young age significantly vary in different boroughs.

```
  age_16_to_
 Min.   : 44836
 1st Qu.: 65619
 Median : 82476
 Mean   : 82356
 3rd Qu.: 99369
 Max.   :122818
```

Figure 3.11: Age statistical data between 16 and 34

We can also examine a histogram of the ages between 16 and 34 in Figure 3.12:

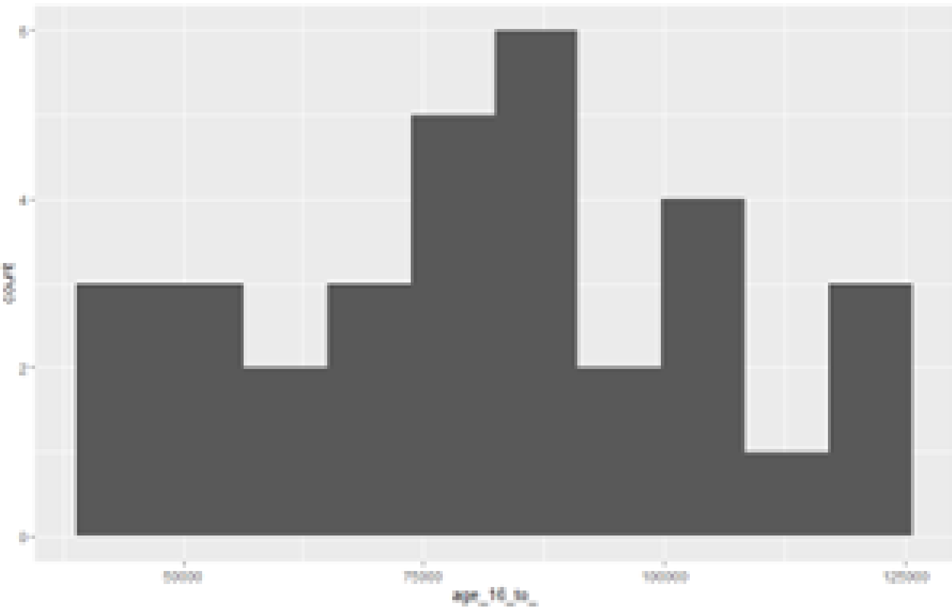

Figure 3.12: Histogram of the ages between 16 and 34

By drawing the age data in map by R, we can find the borough where both the proportion of young people and the number of young people have relatively large values. It seems that Tower Hamlets is a potential borough suitable for bubble tea stores in Figure 3.13 3.14.

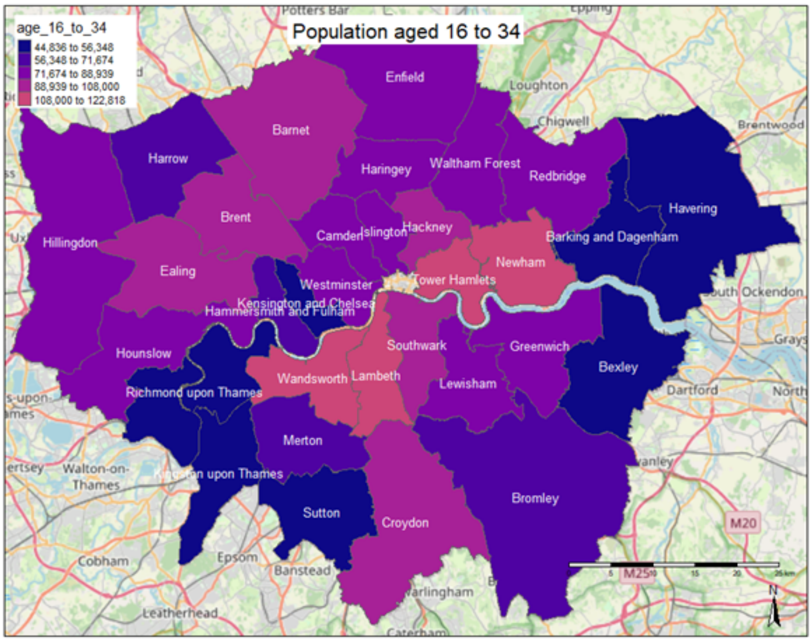

Figure 3.13: Population aged between 16 and 34

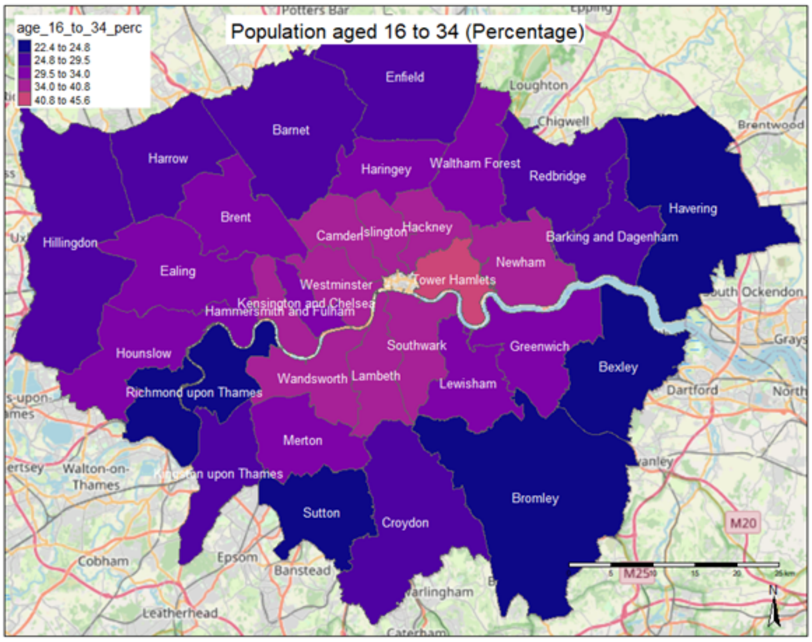

Figure 3.14: Population aged between 16 and 34 (percentage)

Using the same approach, we could also calculate the statistical information for economic activity data in Figure 3.15 3.16. We focus on the number of students and the unemployed percentage which influence potential customers.

```
    student
 Min.   : 6454
 1st Qu.:10932
 Median :15577
 Mean   :14912
 3rd Qu.:17349
 Max.   :24959
```

Figure 3.15: Student statistical data

```
   unemploy_1
 Min.   :3.000
 1st Qu.:4.375
 Median :5.050
 Mean   :5.138
 3rd Qu.:6.025
 Max.   :7.300
```

Figure 3.16: Unemployment statistical data

We can also examine a histogram of the number of students in Figure 3.17:

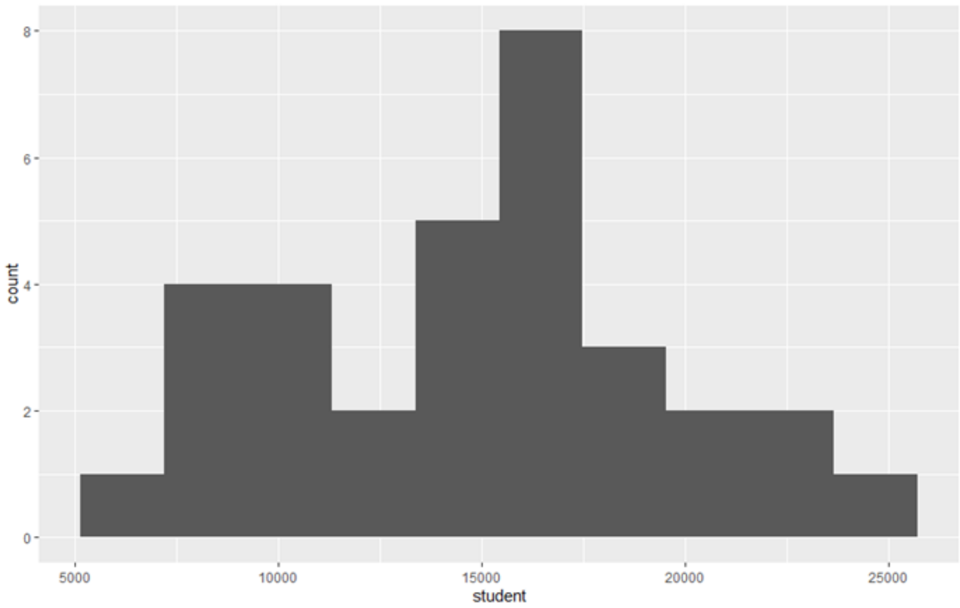

Figure 3.17: Histogram of student data

The borough with more students and a lower unemployed percentage would be primarily selected. The map can not only demonstrate an overview of the dataset but help us to find out the potential borough fitting our requirements for bubble tea site selection in Figure 3.18 3.19.

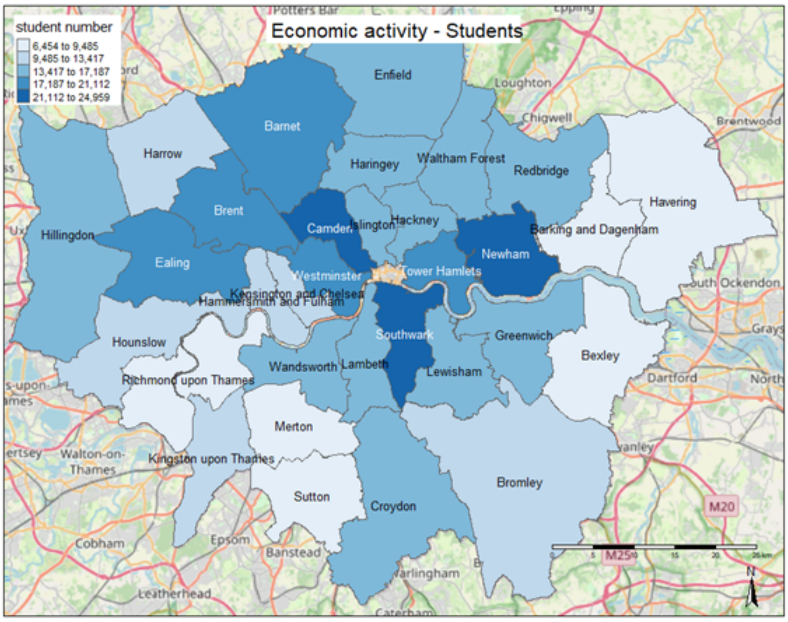

Figure 3.18: Economic activity - Students

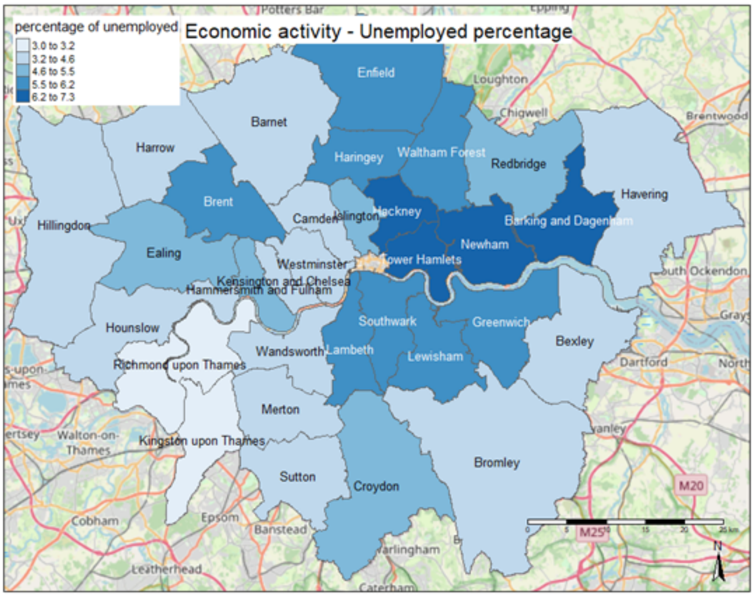

Figure 3.19: Economic activity - Unemployed percentage

For the social grade data, the UK Office for National Statistics (ONS) created the socioeconomic classification known as Approximated Social Grade using an algorithm created by the MRS Census & Geodemographics Group, which has six categories: A, B, C1, C2, D, and E [20]. We mainly focused on social grade A and B, which corresponds to the people with higher income. The details are shown in Figure 3.20 3.21.

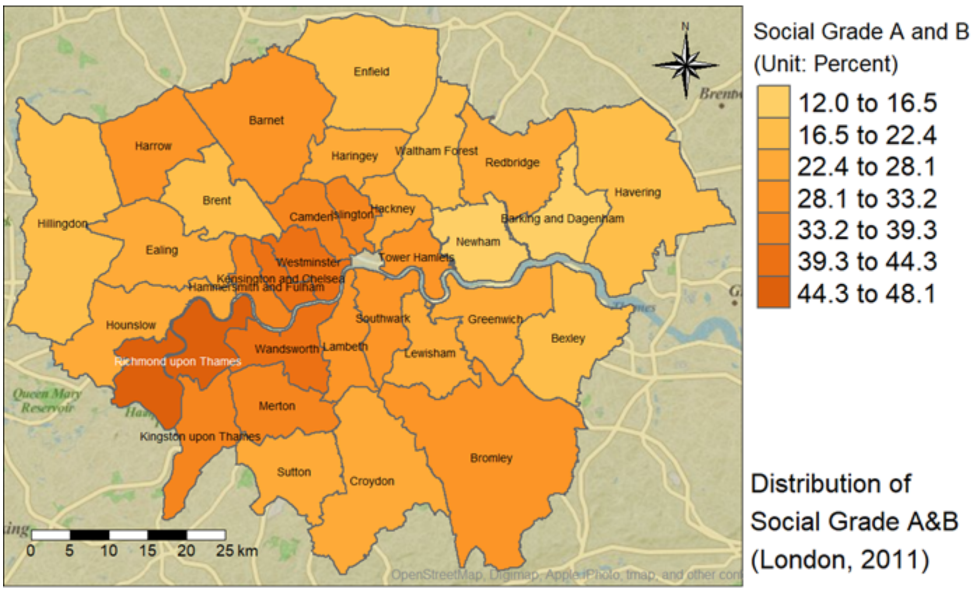

Figure 3.20: Distribution of Social Grade A&B in London at borough level (Percent)

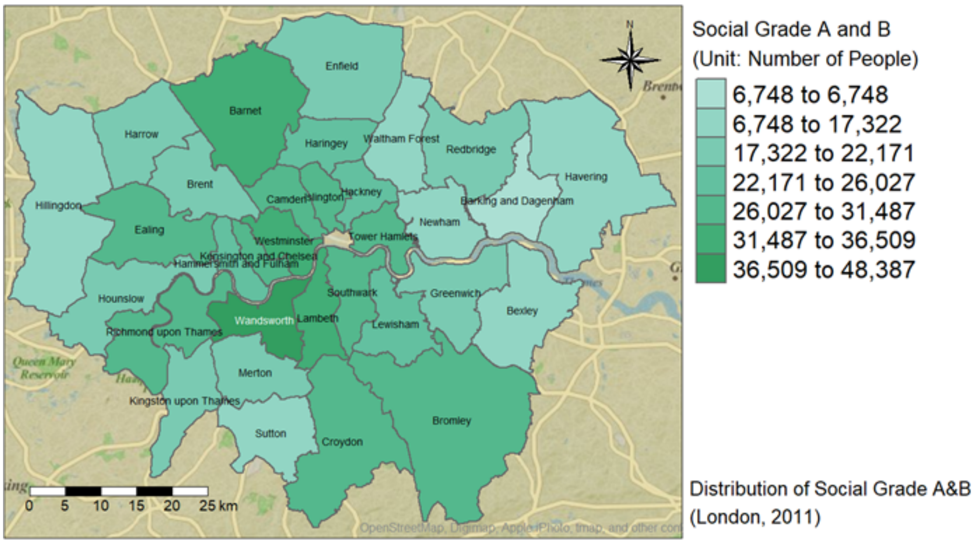

Figure 3.21: Distribution of Social Grade A & B in London at borough level (Number)

3.2.2 Cost Data

Salary

The median gross weekly earnings for data are used to represents the level of salaries. The following are the characteristics of the earning data, derived from the ESDA in Section 4.4.

- Median gross weekly earnings are non-normally distributed, with two boroughs with the highest two median earnings having a big gap from other boroughs.
- Higher earnings appeared in central areas of London. Besides, there are more boroughs with high income in the south than in the north.
- The data are at a borough level with 33 records, which is highly limited to represent the actual situation of earnings in a smaller area (e.g., salary data in Oxford Street).

Rent

The median house prices data in London is used as the indicator to judge the rent level of an area. Also in Section 4.4, the properties of it are summarized as follows:

- The distribution for the median house prices in London is positively skewed, with a few houses having high prices.

- Most of the places with high housing prices are in the mid-west and north-west of London, and there are scattered areas with high housing prices in other places.
- There may be outliers within the areas, but LSOA is a relatively suitable level. If needed, point data of house prices may be useful to find out outliers.

Chapter 4

Methodology

This chapter discusses the process for analyzing data, and the methodologies used in it.

4.1 Transportation and Land Use

(Contributor: Liting Wang 22046528)

4.1.1 Construction of Indicator System

Economic theory suggests that proximity to demand, transportation infrastructure, and other factors are essential considerations in site selection [21]. Among these, market demand is a crucial determinant in site selection. A location with strong market demand can drive high traffic to a business, leading to increased profitability. Therefore, transportation infrastructure and land use type are included as indicators in the site selection process. The way of measuring these criteria can be seen in Table 4.1.

Category	Criteria	Measurement
Buit-environment	Bus stops	Proximity to bus stops(m)
	Subway stations	Proximity to rivers(m)
	Land use	Land use type

Table 4.1: Indicator system for transportation and land use data.

4.1.2 Single Factor Analysis

In the single-factor analysis section of this study, various evaluation methods were employed for different types of data, but the overall process followed similar steps. Firstly, the data was geographically aligned and cropped to meet the scope of the study. Secondly, the land use data was converted into a raster format with a resolution of 1.6m $\times$ 1.6m. Lastly, each factor was assigned a suitability score ranging from 1 to 10, with 1 being the least suitable and 10 being the most convenient. The grading criteria varied depending on the aspect being evaluated. For example, proximity was used as an indicator of grading for transportation infrastructure, while different land use types themselves served as indicators of grading for land use data.

Transportation Infrastructure (Bus Stops and Subway Stations)

Accessibility to transportation infrastructure is vital in determining a location's suitability, as it increases the ease of reach for individuals. Therefore, the closer a site is to transportation infrastructure, the higher its suitability score. To measure this, the study first calculated the shortest Euclidean distance from any point within the study area to the nearest bus stop and subway station. Then, using the 'terra' package, the distance from the centre of each raster cell in the land use dataset was calculated to the nearest bus stop and subway station. This created a new raster the same size as the original land use dataset, containing the calculated distances. The next step involved reclassifying the values of this new raster and determining which values should be assigned to each score from 1 to 10. There are several methods for reclassification, such as equal interval, quantile, natural breaks (Jenks), and standard deviation. In this study, the equal interval method was chosen to divide the range of attribute values into subranges of equivalent size. It allows for the specification of the number of intervals and automatically determines the classification interval based on the scope of values. This method reclassified the distance between bus stops and subway stations.

Land use	Reclassification
Airports	1
Broad-leaved forest	2
Construction sites	3
Continuous urban area	5
Discontinuous urban area	5
Dump sites	Restricted
Estuaries	Restricted
Fruit trees and berry plantations	2
Green urban areas	6
Industrial or commercial units	10
Intertidal flats	Restricted
Land principally occupied by agriculture	2
Mineral extraction sites	Restricted
Mixed forest	2
Non-irrigated arable land	3
Pastures	Restricted
Port areas	8
Road and rail networks	Restricted
Sport and leisure facilities	10
Water bodies	Restricted

Table 4.2: Land use reclassification.

Land Use

The original land use data have 20 categories. Referring to the characteristics of different land use types, the reclassification scores are as in Table 4.2. After determining the reclassification scores, assign the scores to each cell in raster data.

4.2 Recreation, restaurant and cafe

(Contributor: Xue Li 22086637)

4.2.1 Data Acquisition

The raw data for entertainment infrastructure consists of 12 cultural infrastructures in London from 2018 to 2020, with a total of 6,579 features. And the restaurant dataset for London consists of the top 100 restaurants near the centre of the borough (2km radius), with 3133 features. There are 1250 rows of data, and 5 attributes. Yet of these data, only the name and location data are required.

4.2.2 Data Preprocessing

The attributes that were not useful for this project were removed and the remaining data was filtered out.
The filtered data in the source data that belonged to the recreational infrastructure and then put it together in an XLS file. This was then converted to a CSV file and imported into QGIS software adding layers corresponding to the point data.

4.2.3 Geographic Data Analysis of Recreation

The location of the bubble tea shop needs to be close to places where people gather. Such as theatres, cinemas, parks, and other entertainment venues, or near large factories and institutions, which can attract pedestrians on the one hand, and make it easy for customers to remember the location of the shop on the other. Therefore, the closer a location is to recreational infrastructure, the higher its suitability score. As shown in Figure1, the distribution of recreational infrastructure at ward level was obtained by calculating the sum of the number of points in each area. The resulting vector type data was then rasterised to obtain a new raster.

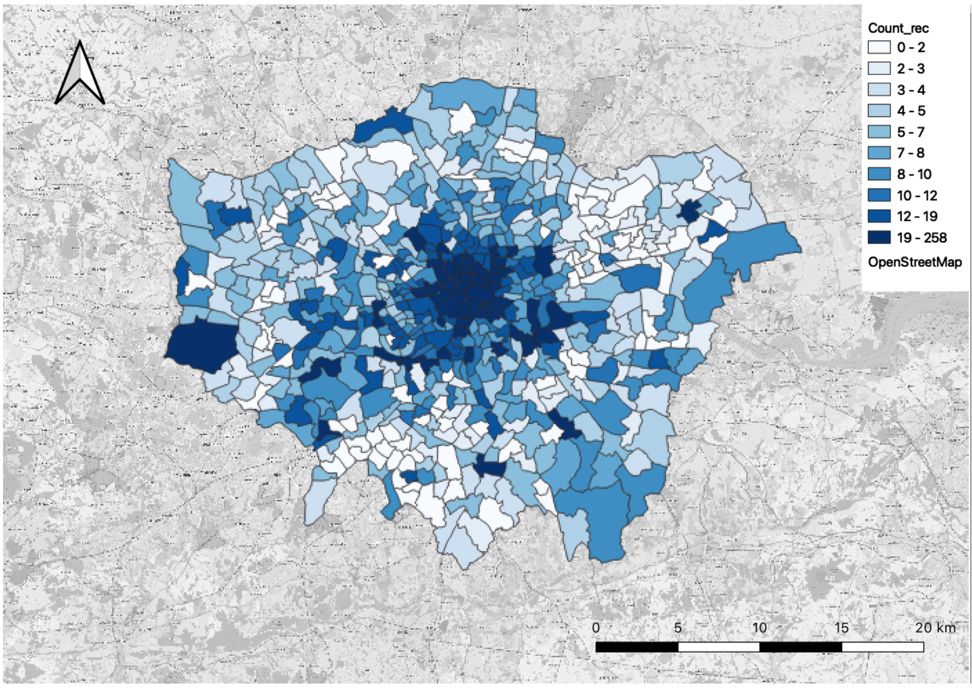

Figure 4.1: Distribution of recreation infrastructures at ward level in London

The values in this new raster will then be reclassified, and it will be decided which values should be assigned to each score between 1 and 10. The equal interval method, the quantile method, the natural break method (Jenks), and the standard deviation method are only a few of the reclassification techniques available. In this study, we used the equal interval approach, which creates equal-sized subranges from the range of attribute values. The categorization intervals are automatically determined based on the range of values and the number of intervals can be set.

4.2.4 Geographic Data Analysis of Restaurant

Proximity to restaurants scores high because a bubble tea shop located near a famous chain or brand name shop can use their branding to attract some customers. The research initially determined the shortest Euclidean distance between any site in the study region and the closest bus stop and metro station to gauge this. Then, the distance between the center of each raster in the land use dataset and the closest restaurant was computed using the 'terra' software program. The computed distances were then added to a new raster that was the same size as the original land use dataset. Finally, the equal interval approach was used to reclassify the values in this new raster.

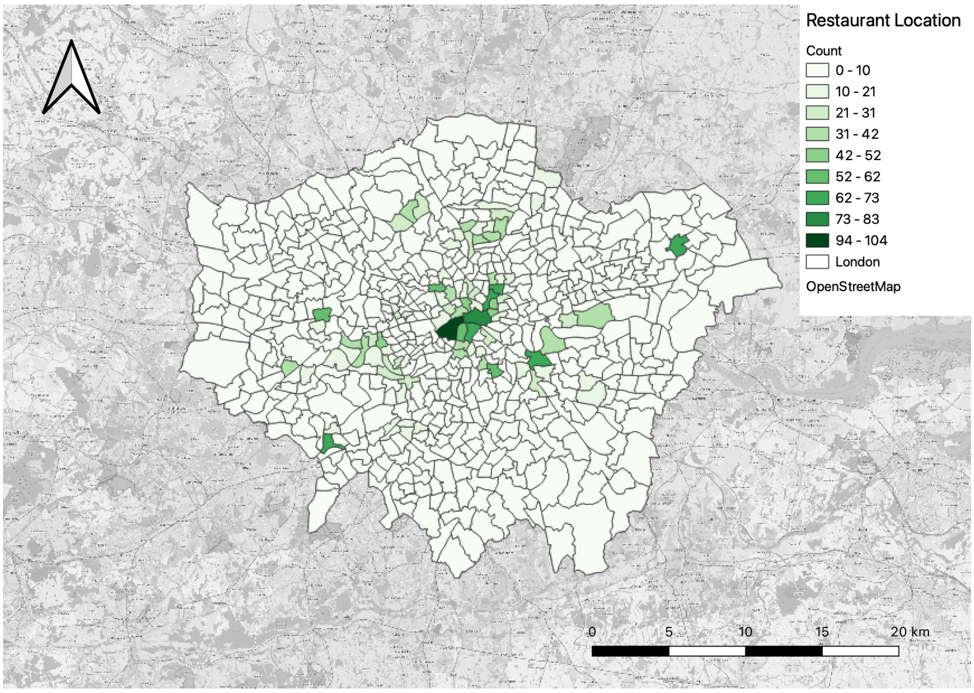

Figure 4.2: Distribution of restaurants at ward level in London

4.2.5 Geographic Data Analysis of Cafe

It is vital to gather information about the coffee shop to ensure that it is kept as far away from some of the rival establishments as feasible. Before the points were calculated in the polygon, the cafe had to be filtered out of the source data since the data types for the London boundary and the cafe are different. This required that the data be reprojected to complete the data conversion. The rasterize was then downgraded from 10 to 1 since the closer a bubble tea business was to a cafe, the worse off it operated, therefore the closer it was, the lower the score should be.

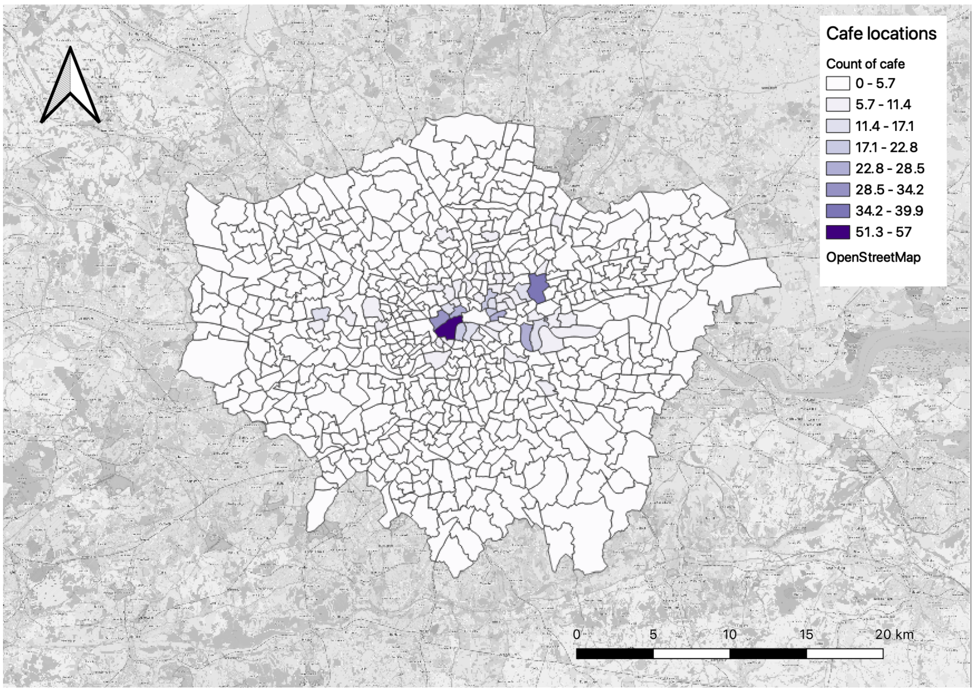

Figure 4.3: Distribution of cafes at ward level in London

4.3 Demographics

(Contributor: Zimo Guo 22055869)

4.3.1 Pre-processing

- First of all, a rectangle was drawn on the Digimap interactive UI to choose the proximate area of Greater London and obtain age, economic activity, and social grade data in Geodatabase format.

- ArcGIS Pro was used to load this source data in order to extract the information for the Greater London region. In order to generate new shape files, the 'select' tool was used to precisely select the boroughs of Greater London using SQL queries. The borough name is represented by an attribute of the source data called "area name" that was utilized for matching in the Where clause of queries.

- Finally, age, economic activity, and social grade statistics for Greater London's boroughs are contained in three new files after extraction.

4.3.2 Exploratory Spatial Data Analysis

- Detailed statistical data on age and economic activity were presented using R's Summary function, and the histogram was created using ggplot and geom histogram.

- The map for Exploratory Spatial Data Analysis was then produced by coding using the R mapping tool tmap (thematic map). Data was first read from the shape file using the st_read() function, and then it was displayed on the basemap, which has a straightforward and clear design.

- Jenks Natural was used to classify the data using the function tm_polygons(). Jenks classification is used to classify demographic data in a way that is more relevant by attempting to reduce each class's average deviation from the class mean while maximizing each class's variance from the means of the other classes. Each borough name on the map is represented by the function tm_text(). The relevant components were added to the map using tm_scale_bar() and tm_compass().

- To make each component look better, the size, colour, and position were modified using tm_layout().

4.3.3 Multi-criteria Decision Analysis

- We require the values of the individual raster layers to be within the same range in order to aggregate them in a weighted sum. Here, we will give each raster cell an appropriateness score of 1:10 depending on its value.

- The age, economic activity, and social grade were stored as sf object at the previous step. Thus, the sf objects should be converted into terra SpatVector objects using 'vect' function. Functions 'classIntervals' and 'cbind' were used to generate matrices to reclassify the old data. The matrix contains the ranges of new classification by specifying three columns "from-to-becomes".

- Then, function 'st_rasterize' was used to convert sf objects to star objects, and function 'rast' was used to convert the star objects to SpatRaster objects which were used to reclassify the data.

- Finally, the function 'classify' was used to generate the score of data between 0 and 10. Part of the code are shown in Figure 4.4.

```
slopeInt <- classIntervals(as.numeric(unlist(values(age_terra))), n=10, style="equal")
slopeBrks <- cbind(slopeInt$brks[10:1], slopeInt$brks[11:2], 10:1)
age_star = st_rasterize(age["age_16_to1"])
age_raster = rast(age_star)
age_16_to1 <- classify(age_raster, slopeBrks, include.lowest=TRUE)
values(age_16_to1)
tm_shape(age_16_to1/2.5)+
  tm_raster()+
  tm_layout(legend.bg.color="white",
            title="Population aged 16 to 34 (Percentage)",
            title.position = c("center", "top"),
            title.color = "black",
            title.bg.color = "white")


##Economic activity - Students
eco_terra <- vect(eco["student"])
values(eco_terra)
slopeInt <- classIntervals(as.numeric(unlist(values(eco_terra))), n=10, style="equal")
slopeBrks <- cbind(slopeInt$brks[10:1], slopeInt$brks[11:2], 10:1)
eco_star = st_rasterize(eco["student"])
eco_raster = rast(eco_star)
student <- classify(eco_raster, slopeBrks, include.lowest=TRUE)
values(student)
tm_shape(student)+
  tm_raster()+
  tm_layout(legend.bg.color="white",
            title="Economic activity - Students",
            title.position = c("center", "top"),
            title.color = "black",
            title.bg.color = "white")
```


Figure 4.4: Part of the code for data reclassification

4.4 Cost

(Contributor: Jiayi Zhao 22095995)

4.4.1 Data Acquisition

The raw data of earnings is from the ONS website [22]. The data has three attributes (i.e., "Description", "Code" and "Median") and 441 records as shown in Figure 4.5. Not all the records are expected as we only focus on London. Therefore, data from London need to be extracted into a new .csv file.

```
> head(salary)
        Description      Code Median
1    United Kingdom K02000001    640
2     Great Britain K03000001    642
3 England and Wales K04000001    642
4           England E92000001    646
5        North East  E12000001    575
6     Darlington UA E06000005    582
> summary(salary)
 Description           Code              Median
 Length:411         Length:411         Length:411
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

Figure 4.5: Sample records and a summary of median gross weekly earnings for full-time employees for all local authorities by place of work.

Raw data on house prices are retrieved from London Datastore [23]. The first six records and a summary of the data are in Figure 4.6. There are 12015 rows of data, and 5 attributes (i.e., “Code”, “Area”, “Year”, “Measure” and “Length”). Among them “Code” is a spatial attribute. The data at the LSOAs level ranged from the year 1995 to 2017. However, only data from 2017 are what we need.

```
> head(price)
       Code                Area                 Year Measure   Value
1 E09000001      City of London Year ending Dec 1995  Median 105,000
2 E09000002 Barking and Dagenham Year ending Dec 1995  Median  49,000
3 E09000003              Barnet Year ending Dec 1995  Median  85,125
4 E09000004              Bexley Year ending Dec 1995  Median  62,000
5 E09000005               Brent Year ending Dec 1995  Median  68,000
6 E09000006             Bromley Year ending Dec 1995  Median  76,625
> summary(price)
     Code               Area               Year              Measure
 Length:12015       Length:12015       Length:12015       Length:12015
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
    Value
 Length:12015
 Class :character
 Mode  :character
```

Figure 4.6: Sample records and a summary of median gross weekly earnings for full-time employees for all local authorities by place of work from 1995 to 2017.

4.4.2 Data Preprocessing

Earnings data from London extracted were now what we expect as shown in Figure 4.7. It has 33 records with each one describing the earning data for one borough in London. It can be seen from Figure 2 that no significant difference between the mean and median. As for “GSS_CODE”, it can be used to find the location for associated earning values.

```
> head(salary)
   GSS_CODE Earning
1 E09000001    1089
2 E09000002     616
3 E09000003     626
4 E09000004     661
5 E09000005     804
6 E09000006     717
> summary(salary)
   GSS_CODE            Earning
 Length:33          Min.   : 596.0
 Class :character   1st Qu.: 661.0
 Mode  :character   Median : 708.0
                    Mean   : 734.9
                    3rd Qu.: 804.0
                    Max.   :1089.0
```

Figure 4.7: Sample records and a summary of earnings data for full-time employees for London by place of work.

House price data from 2017 are extracted and some records and a summary are generated in Figure 4.8. Records are down to 4835 and only location data and price values have remained. It can be seen that the maximum value is about more than 10 times larger than both the median and the mean. Additionally, the mean is about 100k larger than the median. It is likely that extremely high special values raise the average.

```
> head(price)
   LSOA11CD  Value
1 E01000001 935000
2 E01000002 849950
3 E01000003 760000
4 E01000005     NA
5 E01032739 577750
6 E01032740 880000
> summary(price)
   LSOA11CD             Value
 Length:4835        Min.   : 136250
 Class :character   1st Qu.: 368000
 Mode  :character   Median : 466000
                    Mean   : 554354
                    3rd Qu.: 612875
                    Max.   :7775000
                    NA's   :292
```

Figure 4.8: Sample records and a summary of earnings data for full-time employees for London by place of work.

4.4.3 Data Analysis

The histograms in Figure 4.9 show the distribution of earnings. They are not symmetrically distributed and there is a big gap between the highest two values and other values. Therefore, the median can better identify the central position. In addition, quantile breaks will be a better solution to describe variations in earnings.

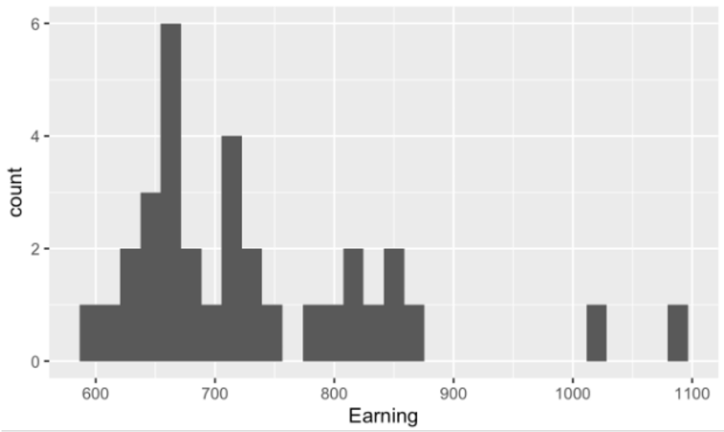

Figure 4.9: Histogram of earnings.

The distribution of median house prices is in Figure 4.10. It is observed that there are very few houses with prices higher than 2 million (£). In order to view most of the data more clearly, we find out histograms for houses that are cheaper than 2 million (in Figure 4.10). In Figure 4.11, data are positively skewed, and we can learn that the median, 466000, is a better representation of the central tendency than the mean, 554354. Therefore, quantile is again an effective method to describe the variations in house prices.

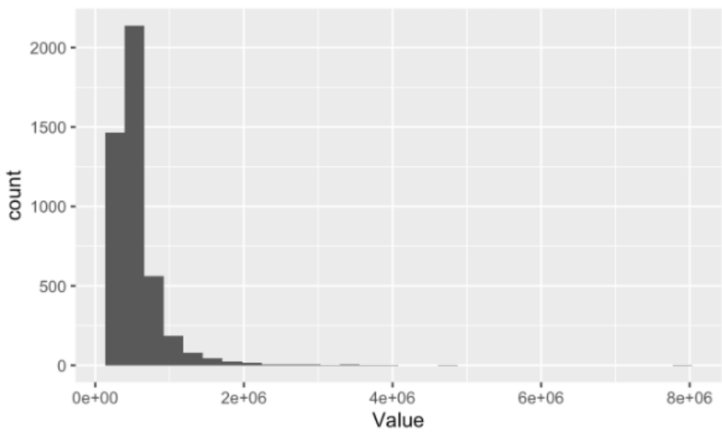

Figure 4.10: Histogram of house prices.

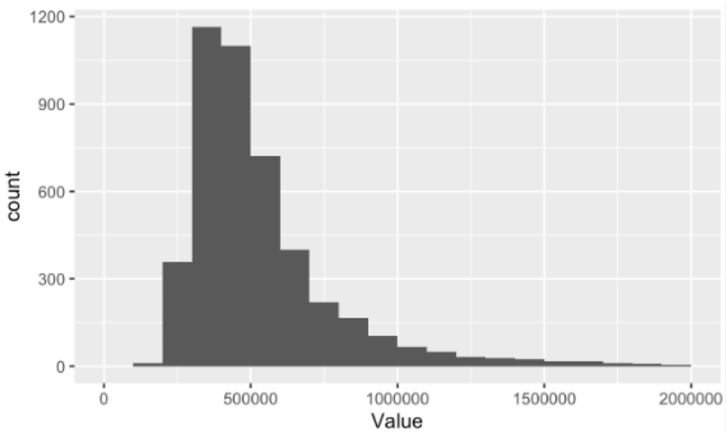

Figure 4.11: Histogram of house prices cheaper than £2 million.

4.4.4 Geographic Data Analysis

To explore how data change geographically, earning data were associated with its borough through code data (i.e., E09000001), and visualized in the map displayed in Figure 4.12. Data were graduated into ten categories by quantile breaks. From this map, we can learn that the centre of London has higher income levels than other places. However, this map is not very informative, since there are only 33 records for boroughs. Finer information (e.g., earning data for an area of a specific postcode) cannot be delivered from this map.

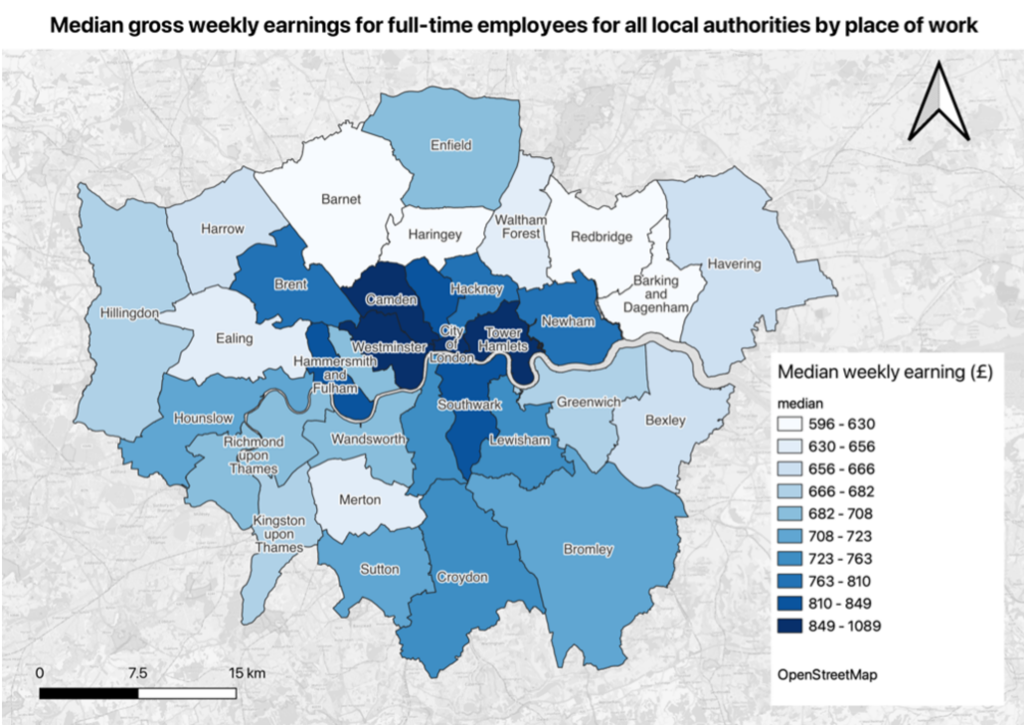

Figure 4.12: Map of median gross weekly earnings for full-time employees for all local authorities by place of work.

Compared with income data, house price information classified at the LSOAs level can provide more detailed information at a finer level. As we can see from Figure 4.13, houses with higher house prices are clustered in the west-central and northwest areas of London. There are far fewer LSOAs with high house prices elsewhere.

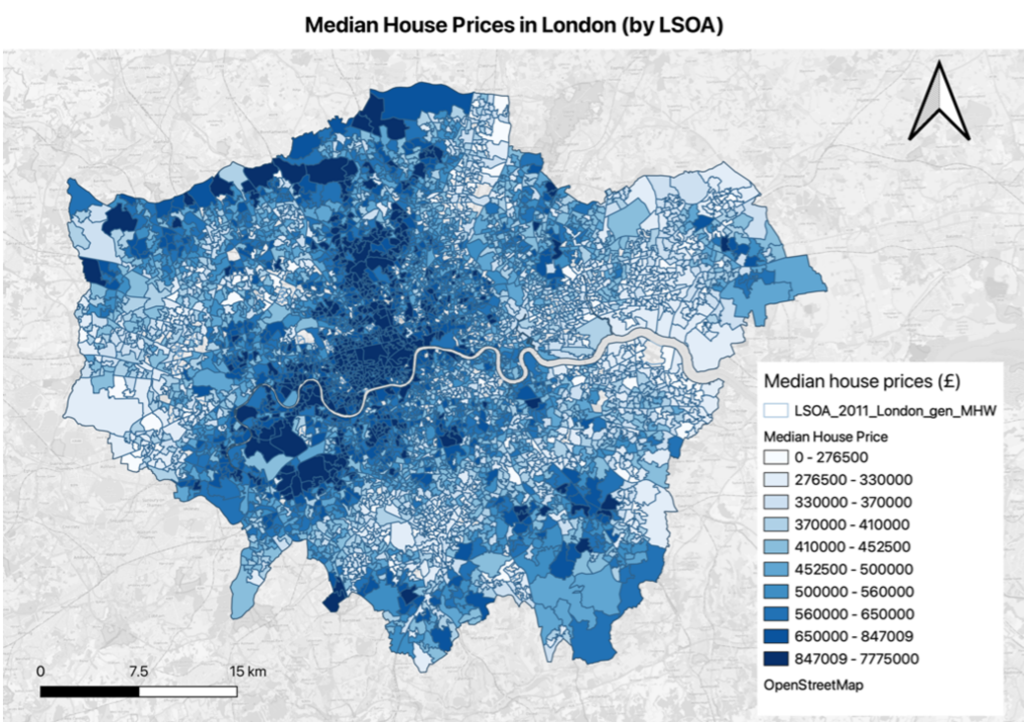

Figure 4.13: Map of median earnings data for full-time employees for London by place of work.

4.4.5 Reclassify Datasets

In order to combine our raster layers in the end, their value must be reclassified to the same common measurement scale (i.e., 1-10 in this case).

Salary

In this study, areas with higher earnings are assigned lower suitability score, as it indicates higher wage the shopkeeper need to pay, which means higher cost. As previously said, the distribution for earnings is skewed, so we use quantile to reclassify the data. The result of the reclassification is in Figure 5.13.

Rent

Similarly, since the distribution of house prices is also skewed, they are also reclassified using quantile as shown in Figure 5.14. A higher house price will have a lower score since it means higher rent.

Chapter 5

Results

5.1 Single-factor Analysis

5.1.1 Transportation Infrastructure

(Contributor: Liting Wang 22046528)

Bus Stops

The wide distribution of bus stops throughout London is reflected in the results of the suitability analysis, as shown in Figure 5.1. The entire area of London is found to have a score of 8 to 10, except for the locations of the bus stops themselves.

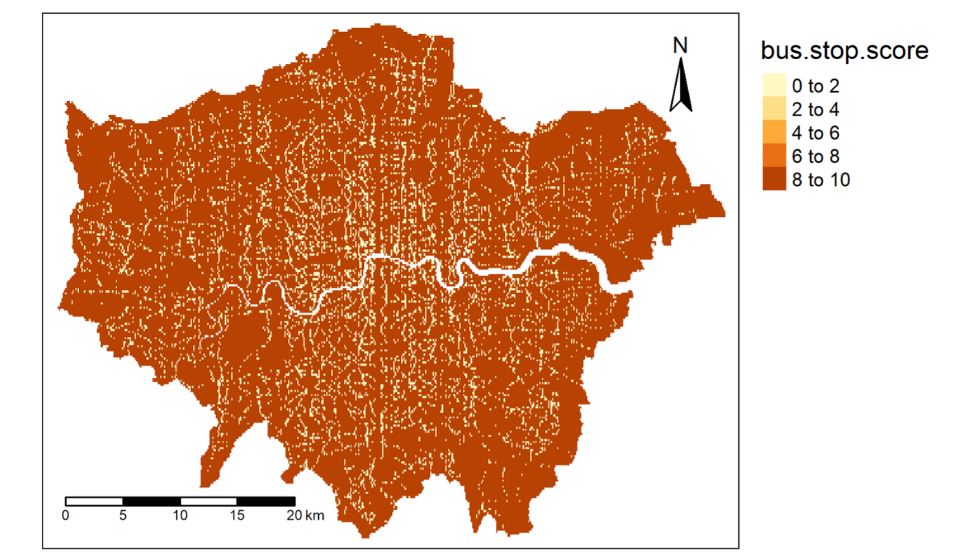

Figure 5.1: Bus stops suitability score.

Subway Stations

The results of the suitability score analysis for subway stations, as depicted in Figure 5.2, indicates that the northwest areas of London have the highest scores, suitable for bubble tea stores. The scores decrease as one moves southeast. The south of Bromley has the lowest scores, ranging from 2 to 4. Notably, a small area in the north and southwest has scores of 6 to 8, while the surrounding areas have scores of 8 to 10.

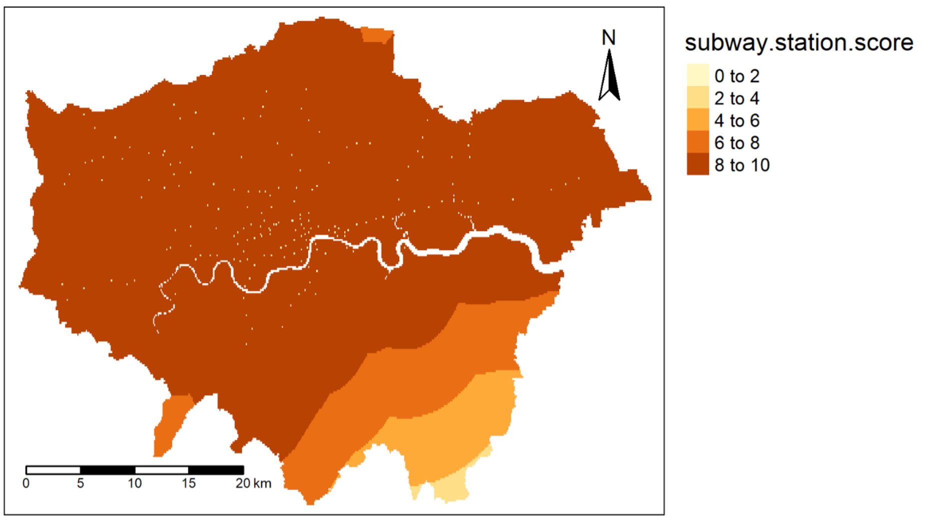

Figure 5.2: Subway stations suitability score.

5.1.2 Land Use

(Contributor: Liting Wang 22046528)

Figure 5.3 illustrates the spatial distribution of the land use suitability score. As seen in the figure, the majority of land in London receives a suitability score of 4-6. The land with the highest suitability score is dispersed throughout the city, with some areas located near the River Thames and others in the southwest region. The white rasters in the figure also represent restricted locations, such as dumpsites, roads, railway networks, and water bodies.

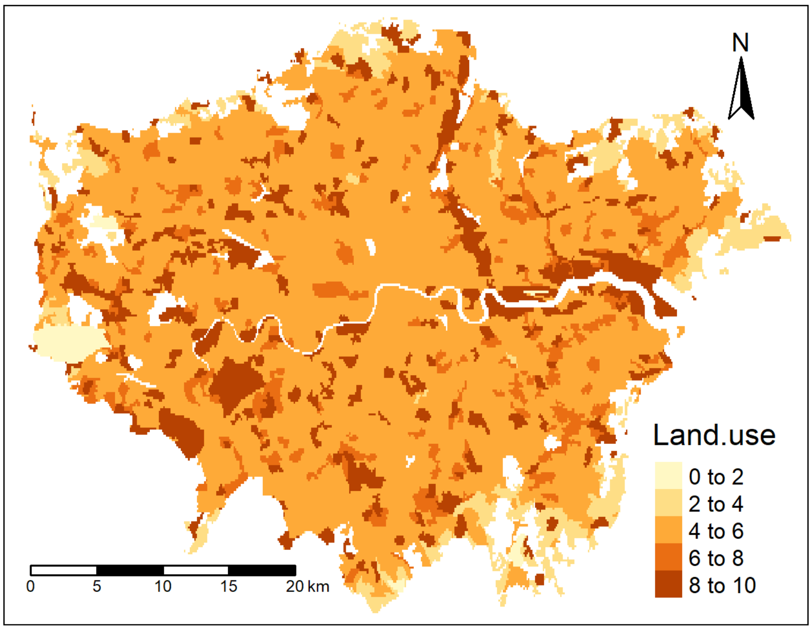

Figure 5.3: Bus stops suitability score.

5.1.3 Recreation

(Contributor: Xue Li 22086637)

The appropriateness analysis's findings are consistent with London's extensive distribution of recreational facilities, as shown in Figure 5.4. The areas with the highest scores are mostly in central and south-west London. The lowest scores were found in the south of Southwark, ranging from 1 to 3.

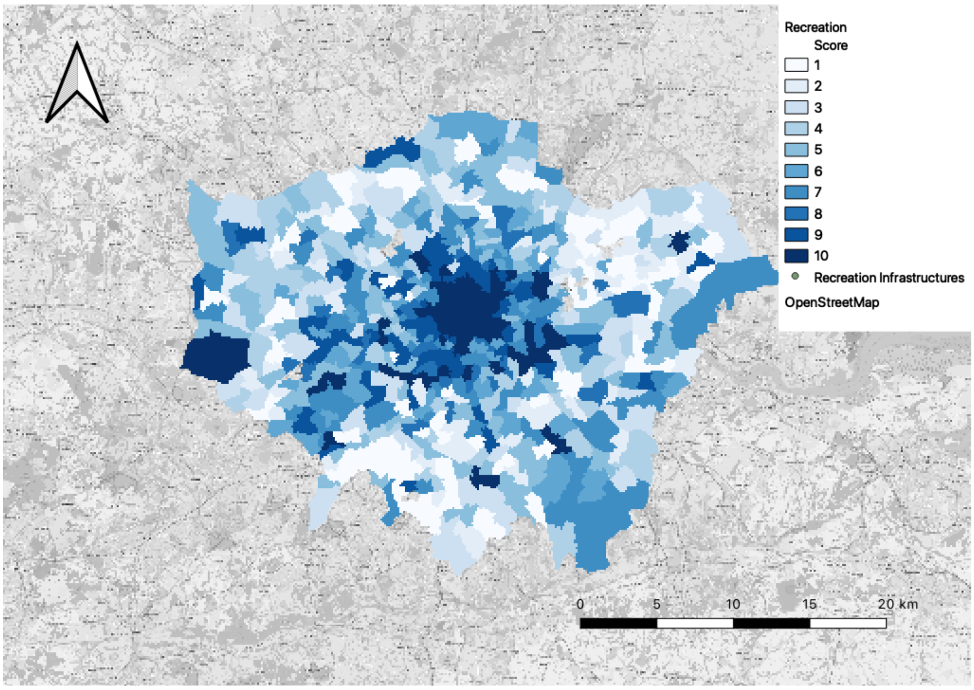

Figure 5.4: Recreation infrastructures suitability score

5.1.4 Restaurant

(Contributor: Xue Li 22086637)

The results of the appropriateness analysis reflect the fact that recreational infrastructure is widely dispersed throughout London, as illustrated in Figure 5.5. The areas with the highest scores are mostly in central and south-west London.

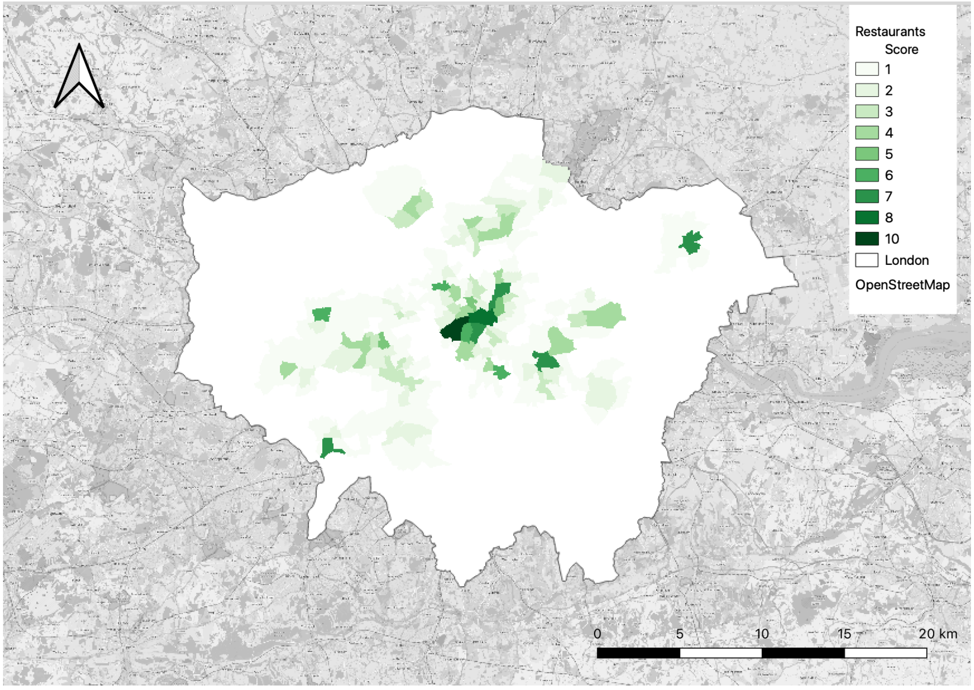

Figure 5.5: Restaurants suitability score

5.1.5 Cafe

(Contributor: Xue Li 22086637)

The appropriateness score study for coffee reveals, as depicted in Figure 5.6, that the more central parts of London score lower and that there is a relative decline in the suitability of building a bubble tea store. As one approaches the periphery, the scores rise. This demonstrates that the most central sectors are more competitive and may be located a little further afield.

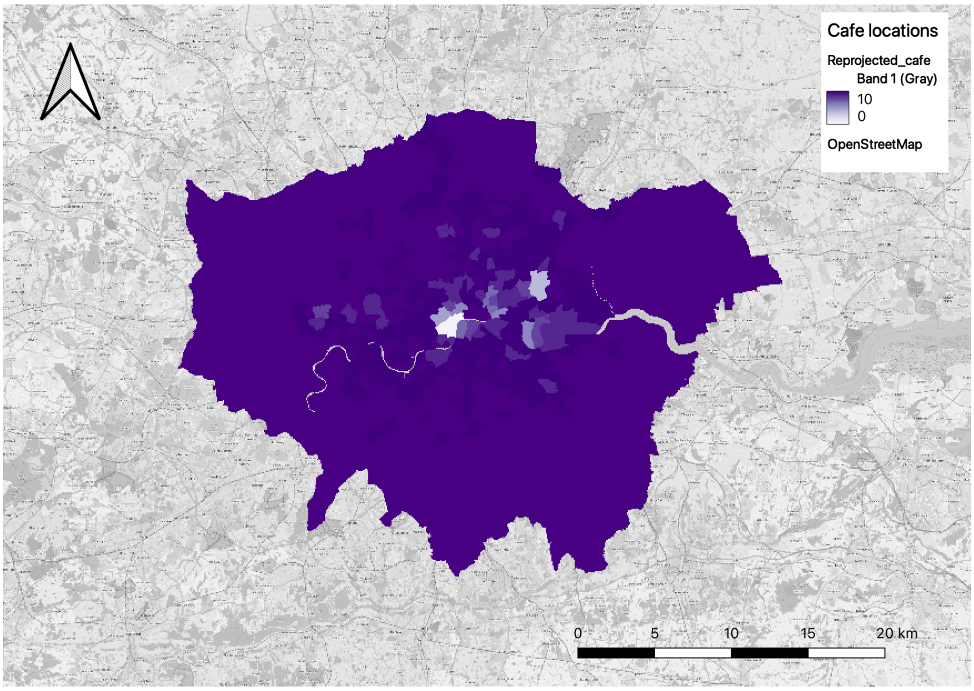

Figure 5.6: Cafes suitability score

5.2 Demographics data

(Contributor: Zimo Guo 22055869)

5.2.1 Young people

The figure 5.7 and 5.8 shows the spatial distribution of young people. As seen in Figure 5.7, a large number of young people are living near the River Thame and others in the northwest region. The southwest region has the lowest scores, ranging from 0 to 2. However, the proportion of young people shows a different distribution in figure 5.8. Most of the area in London receives a suitability score under 4. And Bromley has the highest score which is greater than 8.

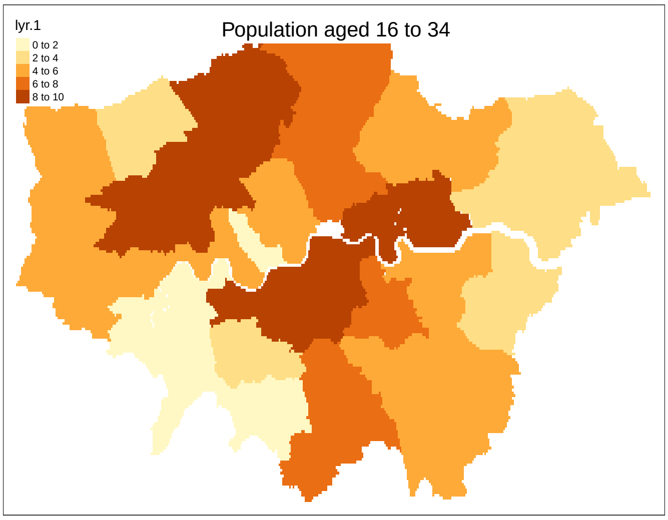

Figure 5.7: Population aged 16 to 34 score

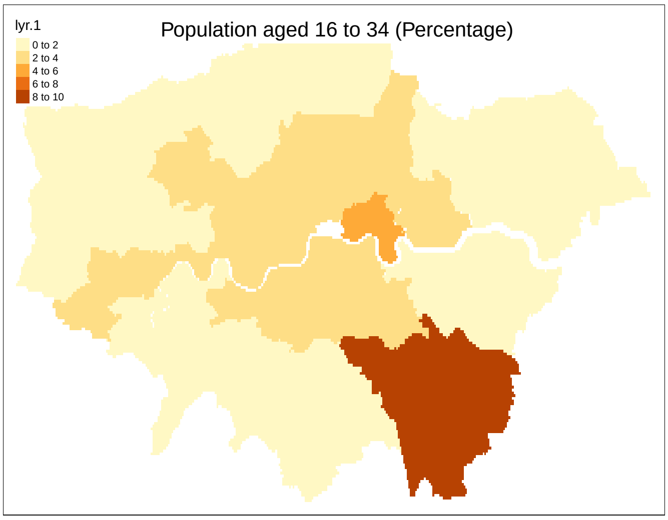

Figure 5.8: Population aged 16 to 34 score (based on rate)

5.2.2 Student

The figure 5.9 shows the spatial distribution of students. As seen in the figure 5.9, a large number of students are living near the River Thame and others in the northwest region, which shows a similar distribution to figure 5.7, due to the fact that a large number of young people are students. The southwest region also has the lowest scores, ranging from 0 to 2, which is also similar to the young people data.

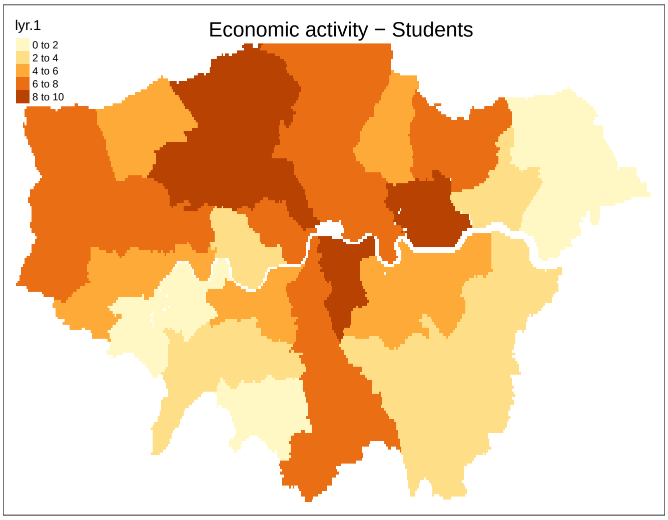

Figure 5.9: Student score

5.2.3 Unemployment

Figure 5.10 shows the spatial distribution of unemployed people. The higher score indicates fewer unemployed people. The highest score was shown in the southwest and southeast region. The east area nearby Newham and Hackney represents a low suitability score. In conclusion, the mean score of the north area is lower than the south area, and the east area is lower than the west area.

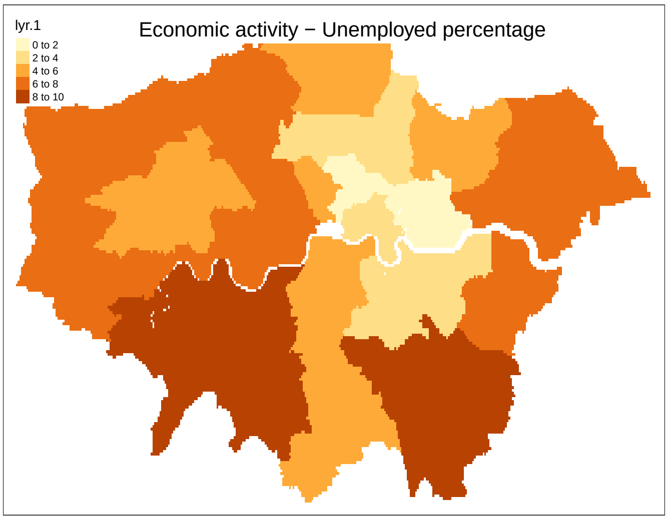

Figure 5.10: Unemployed score

5.2.4 Social grade

The figure 5.11 and 5.12 shows the spatial distribution of social grade (A and B level). The higher value indicates the level of consumption in different borough. As seen in the figure 5.11, the highest score was represented in the western area of London. The western region also shows a high score in figure 5.12. Additionally, most of the area nearby the River Thame receives a higher suitability score than other places.

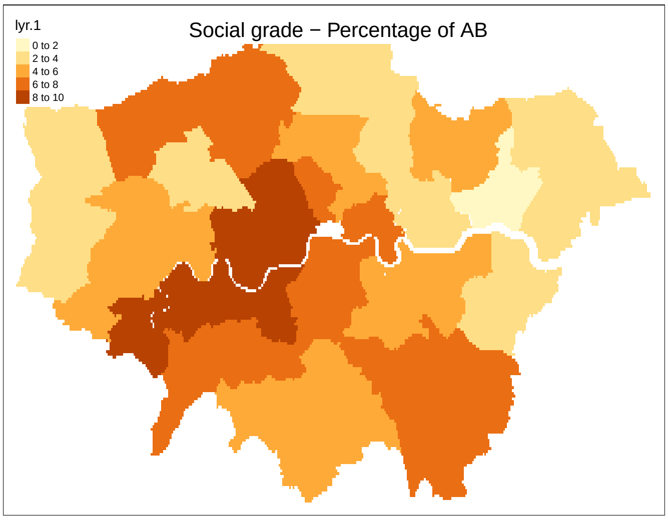

Figure 5.11: Social grade rank (based on rate)

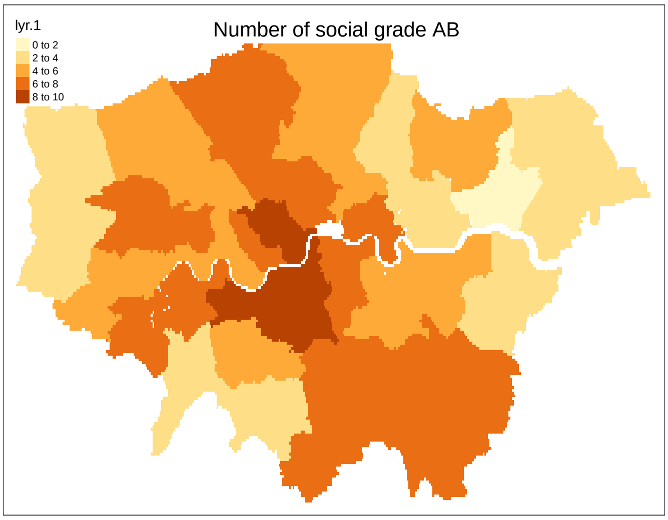

Figure 5.12: Social grade rank

5.3 Cost data

(Contributor: Jiayi Zhao 22095995)

This section illustration results of reclassification discussed in Section 4.4.5.

5.3.1 Salary

The resulting suitability scores for earning data are shown in Figure 5.13. Center areas of London were given the lowest scores, which are negatively correlated with earnings. In contrast, northern London has more high-scoring places. In addition, Merton is the only borough with a score of 9-10 in the south of the Thames.

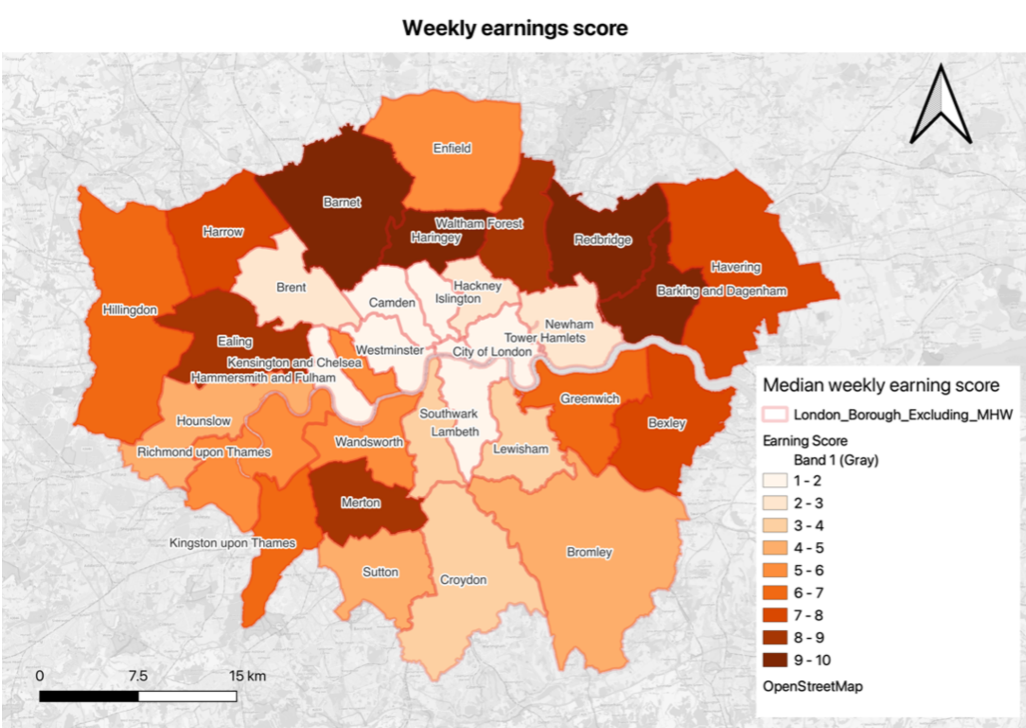

Figure 5.13: Scores for earning data.

5.3.2 Rent

Suitability scores for house prices in Figure 5.14. Scores are mainly high in west and east London, with a small number of high scores in the north and south. Besides, cores and housing prices are negatively correlated.

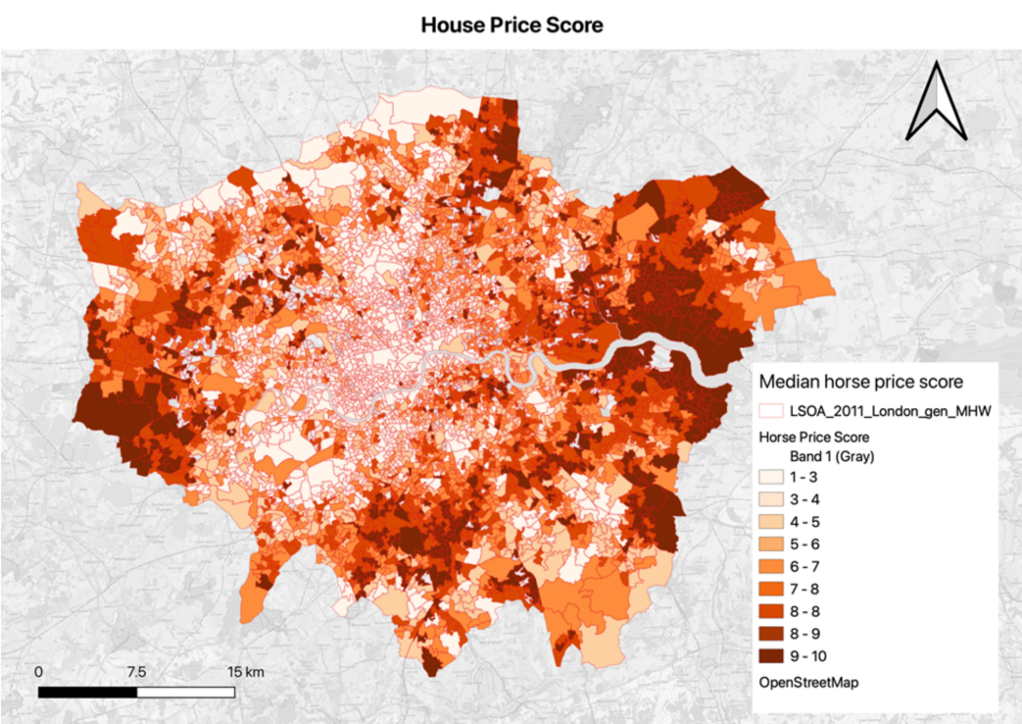

Figure 5.14: Scores for house prices data.

Chapter 6

Discussion and Conclusions

6.1 Weights of Criteria

In this section, we discuss how we weigh each criterion. The decisions of final weights for each of our criteria can be found in Figure 6.1. We mainly referenced analysis in [14] to decide the weights for each score layer when combining them. In the literature, we found that economic, transportation, and commercial areas weigh approximately 0.25. Therefore, we have assigned equal weights of 0.33333 to each category. To simplify calculations, we have rounded the socioeconomic weight to 0.3, resulting in a total weight of 0.7 for the built environment.

Category		Criteria	Weight
Built-environment	Transportation	Bus stops	0.16
		Subway stations	0.16
	Lan use	Land use	0.08
	Recreation	Recreation infrastructure	0.1
		Restaurant	0.1
		Café	0.1
Socio-economic	Demographics	Age	0.0143
		Social grade	0.0143
		Age (rate)	0.0143
		Social grade (rate)	0.0143
		Student amount	0.0143
		Unemplyment rate	0.0143
	Cost	Rent of the shop	0.2
		Employee salary	0.0142

Figure 6.1: Weights for criteria.

6.1.1 Socioeconomics

According to [14], it has been established that rent plays a significant role in determining the overall socioeconomic status of an individual or community. As such, with a total weight of socioeconomic 0.3, we have assigned a weight of 0.2 to the rent criterion in our analysis. The remaining socioeconomic criteria have been assigned a weight of approximately 0.014 each, which is equivalent to 0.1 divided by the total number of criteria (i.e., 7). This allocation of weights allows for a more comprehensive and equitable evaluation of the socioeconomic status of the population under study.

6.1.2 Built Environment

Literature suggests that economic, transportation, and commercial areas weigh 0.25. Therefore, we have assigned similar weights to these categories, with a 0.3 assigned to recreation. As noted in the literature, the importance of the environment is relatively low at 0.1. Thus, we have assigned a weight of 0.08 to land use. The remaining weight is divided equally between bus stops and subway stations, each receiving a weight of 0.16.

6.2 Methodology of Combining Results

- First of all, all results have different CRS, so that the CRS should be converted as the same value before combining results.

- Additionally, all results have different origins, extent and cell sizes, so the resample function should be used to transfer data in the same format.

- A weighted overlay should be used to carry out the reclassified layers. The weights have been decided based on the relative literature review.

- The final map displays the sites with a suitability score of greater or equal to 9. Based on the standards we developed, these sites can be regarded as appropriate for bubble tea site selection. Part of the code was shown in Figure 6.2.

```r

##overall
weighted_overlay <- ceiling((age_16_to_*0.0143)+(age_16_to1*0.0143)+(student*0.0143)+
                             (unemploy_1*0.0143)+(social_g_1*0.0143)+(social_g_5*0.0143)+
                             (median1_10_Resample*0.0142)+(houseprice_Resample*0.2)+
                             (landuseResample*0.08)+(busResample*0.16) +(subwayResample*0.16)+
                             (classified_Resample*0.3))
tm_shape(weighted_overlay)+
  tm_raster()
hist(values(weighted_overlay), main="", xlab="Suitability score")
suitable_sites <- weighted_overlay>=9
tm_shape(suitable_sites)+
  tm_raster()


```

Figure 6.2: Code to implement raster combination

6.3 Discussion of Results

The results of the weighted overlay analysis is shown in Figure 6.3 , and the histogram of suitability scores is illustrated in Figure 6.4, indicate that the areas with the highest frequency of occurrence are those with a suitability score range of 6-7. This suggests that the area with the highest degree of suitability for the specified criteria is the largest in London. Additionally, as determined by the frequency of occurrence, the second and third largest areas are those with suitability scores in the 5-6 and 7-8 ranges, respectively. Conversely, the areas with suitability scores in the 8-9 and 4-5 ranges exhibit a decreasing trend in size. Furthermore, the smallest area identified by the frequency of occurrence is that with a suitability score range of 9-10.

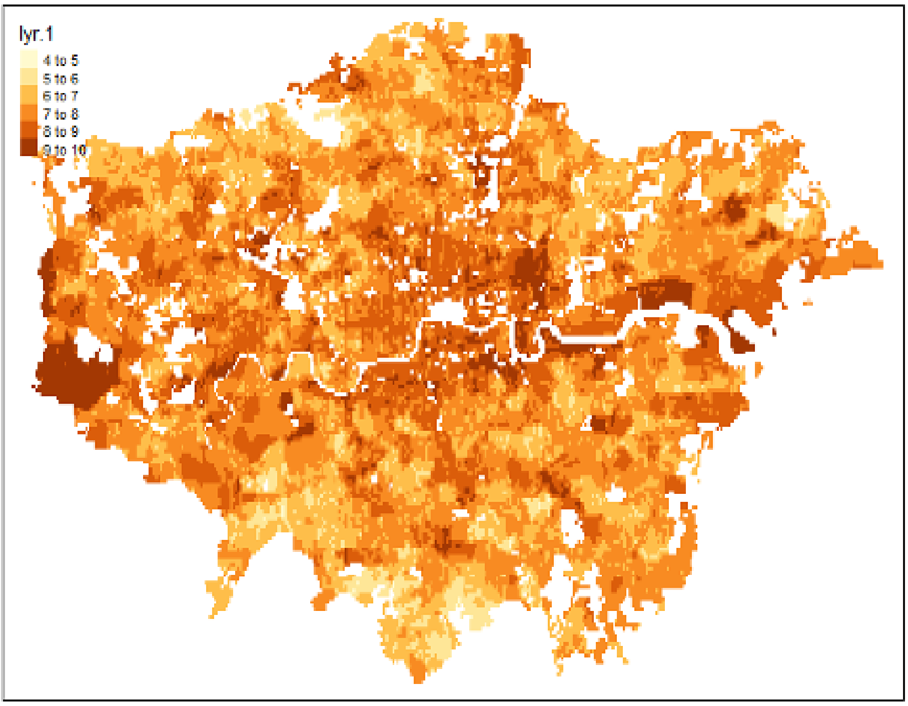

Figure 6.3: Result of the weighted overlay.

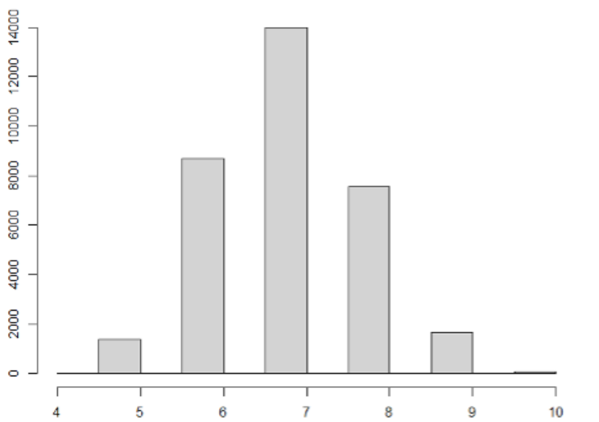

Figure 6.4: Histograms of suitability scores.

The results of the suitability analysis are presented in Figure 6.5, which illustrates the locations with a suitability score of 9 or higher. These sites can be considered optimal locations for developing a bubble tea shop, as they meet the criteria established in the analysis. As shown in the figure, the majority of these suitable sites are concentrated in the central region of London, particularly in the vicinity of the River Thames. It is worth noting that there is a relatively independent high-scoring area in the westernmost part of London. The reason for its appearance may be that the rent is very low when the rent is an important reference indicator. However, it should be noted that a small number of relevant sites are also present in other areas of London, such as the northern and southern regions. These areas possess favourable commercial conditions, convenient transportation, and high pedestrian traffic, making them suitable for a bubble tea store. This map can be a valuable reference for future site selection decisions for bubble tea shops.

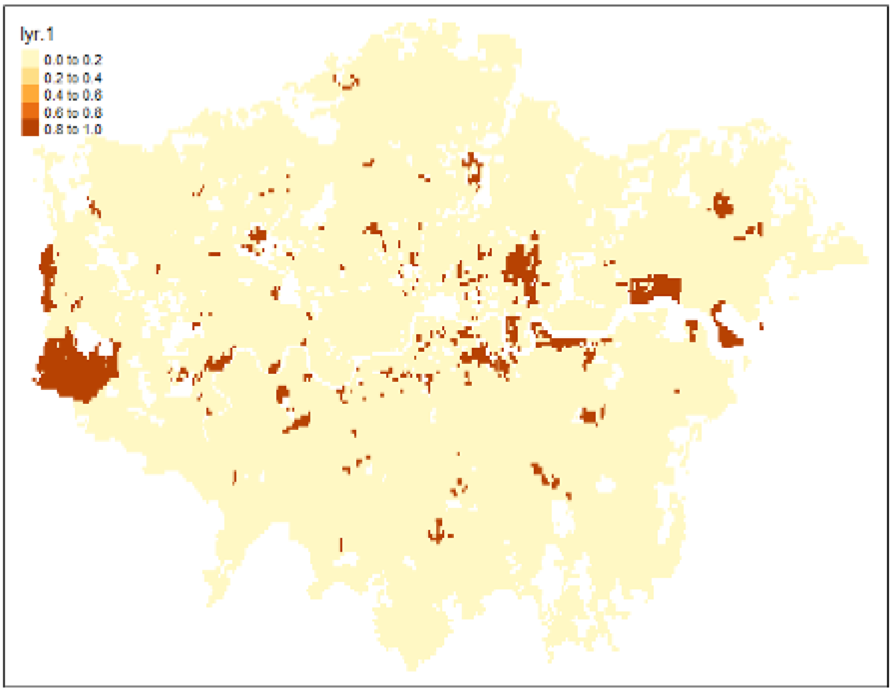

Figure 6.5: Sites with suitability greater or equal to 9.

Bibliography

[1] Bubble tea definition. https://en.wikipedia.org/wiki/Bubble_tea. Accessed December 24, 2022.

[2] Xuguang Bao and Lizhen Wang. A business site selection system using co-locations and ontologies. In *2018 IEEE Intl Conf on Parallel Distributed Processing with Applications, Ubiquitous Computing Communications, Big Data Cloud Computing, Social Computing Networking, Sustainable Computing Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, pages 1034–1035, 2018.

[3] Shirley Dinh and A Li. Living in a boba bubble, 2019.

[4] Profit formula. https://www.investopedia.com/terms/p/profit.asp. Accessed December 24, 2022.

[5] Gov.uk. https://www.gov.uk/. Accessed January 24, 2023.

[6] London datastore. https://data.london.gov.uk/. Accessed January 24, 2023.

[7] Transport for london. https://tfl.gov.uk/. Accessed January 24, 2023.

[8] Cultural infrastructure. https://www.data.gov.uk/dataset/92a4192d-df25-4695-98a5-139bcd0620c8/cultural-infrastructure-map. Accessed December 24, 2019.

[9] Open street map. https://www.openstreetmap.org/. Accessed December 24, 2022.

[10] Statistical gis boundary files for london. https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london. Accessed December 24, 2022.

[11] Digimap. Society data download. `https://digimap.edina.ac.uk/`. Accessed December 6, 2022.

[12] Tara Purswani. Survey: Millennials, gen z big fans of bubble tea. `https://foodinstitute.com/focus/survey-millennials-gen-z-big-fans-of-bubble-tea/`. Accessed December 13, 2022.

[13] Unemployment, (re)employment and wellbeing. `https://whatworkswellbeing.org/resources/unemployment-reemployment-and-wellbeing/`. Accessed December 23, 2022.

[14] Gwo-Hshiung Tzeng, Mei-Hwa Teng, June-Jye Chen, and Serafim Opricovic. Multi-criteria selection for a restaurant location in taipei. *International journal of hospitality management*, 21(2):171–187, 2002.

[15] Joshua Gallin. The long-run relationship between house prices and rents. *Real Estate Economics*, 36(4):635–658, 2008.

[16] 2023 Talent.com. Catering average salary in united kingdom, 2023. `https://uk.talent.com/salary?job=catering`. Accessed January 26, 2023.

[17] plumplot.co.uk. London median salary comparison. `https://www.plumplot.co.uk/London-salary-and-unemployment.html`. Accessed January 26, 2023.

[18] John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[19] Robert Haining, Stephen Wise, and Jingsheng Ma. Exploratory spatial data analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):457–469, 1998.

[20] Social grade a, b, c1, c2, d, e. `https://www.ukgeographics.co.uk/blog/social-grade-a-b-c1-c2-d-e`. Accessed December 23, 2022.

[21] Sharmistha Bagchi-Sen. The dynamics of industrial location: The factory, the firm and the production system, 2001.

[22] Office for national Statistics. Employee earnings in the uk: 2021. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2021. Accessed January 26, 2023.

[23] London Datastore. Average house prices by borough, ward, msoa lsoa. https://data.london.gov.uk/dataset/average-house-prices. Accessed January 26, 2023.