Predicting London Rent Using Geospatial Data

Assignment for the IBM Capstone Project

1 Introduction

1.1 Background

London is a huge city with a population of almost 9 million people. The money which flows through it has immense pulling power, attracting businesses, investors, and young professionals like myself looking to enter the workforce. With so much capital and wealth at the center of it, the demand for accommodation around the city has driven housing prices significantly higher than anywhere else in the UK. London, therefore, is an economic microcosm where higher wages and greater professional opportunity is matched by inflated prices for property, which might cost half as much anywhere else in the country. The churn of accommodation in London is faster as well, with renters moving through properties rapidly and often across large distances.

Each district of London has a unique character defined by its people, its natural surroundings, and the businesses and services within it. Rent prices also vary a significant amount within the city, with flats in the highly gentrified inner-city being notoriously costly and flats in the suburbs around London markedly less. The diverse nature of London living raises interesting questions about how well one can predict it might cost to live in a certain area, even before delving into property specifics.

1.2 Project aim

The aim of this project is to explore whether mean London rental prices can be predicted using only geospatial data and information about surrounding venues. For simplicity, this investigation will only encompass prices of studio flats as a barometer for trends of other sorts of accommodation. The geographical scope will be the London metropolitan area which includes Greater London and its commuting belt. This area will be sub-divided by postcode district - the greatest level of granularity for which data is available. Supervised learning regression techniques will be used to build models using a training set of the data, and subsequently assessed with a test set.

1.3 Interest

Young professionals moving around London and first-time renters are the primary groups of stakeholders who would benefit from a predictive analysis involving venue data. By understanding how their preferred surroundings impacts their rent, they will be able to assess the likelihood of finding a place in a certain part of the city which aligns with their lifestyle and budget.

By extension, landlords and housing agents would be interested in whether a certain density of nearby venues affects the average price of a property in the area. They would therefore be able to match the needs of their potential customers more precisely and secure more business.

2 Data

2.1 Variables and Data Sources

The predictive variables for the investigation will include:

- Venue data, such as the number of venues and types of venues in each district;
- Geographical data, such as the location, population, size, and distance from city centre of each district

The target variable for the investigation will be the average monthly rent of a studio flat.

2.1.1 Venue Data

Venue data will be gathered using the Foursquare API.

These data will help establish how the number of certain types of venues may correlate to the cost of living in London. For example, we might expect a district with a large number of expensive wine bars to reflect affluence of its inhabitants which could correspond to higher property prices. Alternatively, we might expect areas with a higher venue density to reflect a more urban environment, and therefore more expensive properties in a closer proximity to workplaces.

2.1.2 Geographical Data

Geographical data will be gathered using census data from the Office of National Statistics. The latest data available for postcode district geographies is from August 2019. Since population censuses are only conducted every ten years at the district level, the best available statistics from 2011 are a little out of date. However, for this experimental purpose they should serve reasonably well.

These data will help establish how fundamental geographical aspects of districts may correlate to the cost of living in them. For example, it is reasonable to assume that there is a relationship between rent and distance from city centre, but what does this relationship look like and how much weight does this factor have when other factors are considered in conjunction? Also, will population density indicate anything about how expensive a district might be to live in?

2.1.3 Rental Data

Rental data will be gathered from the open-source statistics provided by the Valuation Office Agency. The latest data available on the UK government website for the private rental market in London dates between July 2018 to June 2019.