IBM Data Science Specialization

Applied Capstone Project

Coursera

**Final Report**

Effect of surrounding venues on prices of rental apartments

Zohair Hashmi

[Hashmi.zohair@gmail.com](mailto:Hashmi.zohair@gmail.com)

# **Table of Contents**

|  |  |  |
| --- | --- | --- |
| I. | Introduction | 3 |
| II. | Data description | 4 |
| III. | Methodology   1. Data Collection, Cleaning and Processing 2. Visualization 3. Principal Component Regression | 4 |
| IV. | Results | 6 |
| V. | Discussion | 7 |
| VI. | Conclusion | 7 |
|  |  |  |
|  |  |  |

# Introduction

This report is for the final project of the Coursera – IBM Data Science Professional Certification. The certification spans over 9 different courses, hosted on the Coursera platform. The final project requires making use of the Foursquare API for getting important neighborhood data and integrating it with a problem or issue that we intend to solve.

For this project, my main goal is to explore the neighborhoods of London and observe how the average rent for a 2-bedroom apartment varies in accordance with the venues surrounding it in a given range of a kilometer.

As most families look for a 2-bedroom apartment, we are keeping our analysis limited to this size only. When renting an apartment, it is very important to understand the demographics of the neighborhood before doing so. The two factors that are likely to influence their decision the most are:

1. The monthly rent of the apartment.
2. The nearby venues in the apartment.

For our case, the target audience are families looking for renting a 2 bedroom apartment, people willing to put their owned apartments on rents and the real estate organizations who can decide what location to build their apartments around where they can maximize the rents.

# Data Description

The greater London has a total of 33 neighborhoods in the inner and outer London. Each neighborhood has a different infrastructure and different number of venues that define its demographics. These demographics in turn influence the price of purchasing or renting an apartment or a house. The valuation office agency keeps an annual record of monthly rents paid across the entire UK, based on different size of the apartment. The data for 2-bedroom apartments is extracted from the excel file available on the government of UK’s website.

To sketch the map and show the distribution of neighborhoods across London, we use the geojson file of the greater London available for free on the website data.cdrc.ac.uk.

Therefore, our data sources are:

1. Government of UK’s website. <https://www.gov.uk/government/organisations/valuation-office-agency>
2. FourSquare API which provides the surrounding venues of a given coordinates.
3. Open source of CDRC for geojson files of Great Britain. Link: <https://data.cdrc.ac.uk/dataset/..../greaterlondon.json>

# Methodology

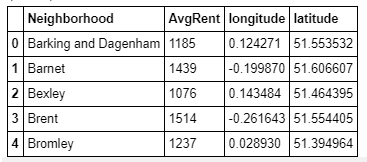
The assumption is that real estate price is dependent on the surrounding venue. Thus, regression techniques will be used to analyze the dataset. The regressors will be the occurrences of venue types. And the dependent variable will be standardized average prices. At the end, a regression model will be obtained. Along with a coefficients list which describes how each venue type may be related to the increase or decrease of a neighborhood’s real estate average price around the mean.

## Data Collection, Cleaning and Processing

The process of data collection and cleaning is as followed:

* Download the excel file “Publications\_AllTables\_200619” from the source mentioned previously in data description paragraph. Create a new excel file using the data of 2 bedroom apartments only and uploading it IBM’s cloud storage.
* Remove unnecessary columns and rows, such as the total count of rents in each neighborhood and the total data of inner and outer London.
* To have neighborhood coordinates, we download the directly read the csv file of Libraries in London from London govt’s database.
* Again, remove unnecessary columns and group the data according to the neighborhoods and their latitude and longitude coordinates.
* Merge both datasets into one single dataframe.

The resulting dataframe looks like this:



## Visualization

In order to observe the distribution of each neighborhood in London, we use maps from the folium library to sketch a map of London and place markers over each neighborhood. Since a geojson file is not available on the internet that distributes the greater London into its small constituencies, the spread of average rent was not demonstrated using choropleth maps.

## Principal Component Regression (PCR)

PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression. (Wikipedia, n.d.)

PCR employs the power of PCA, which can convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. As the result, the number of features is reduced while keeping most of the characteristic of the dataset.   
Then PCR use Linear Regression on the converted set to return a coefficient list, just like in normal Regression techniques.

The R2 Square is not very close to 1 but it is a significant improvement over Linear Regression, which is why PCR is applied to understand the correlation between surrounding venues and the avg rent in neighborhoods.

The result is displayed in the following figure:



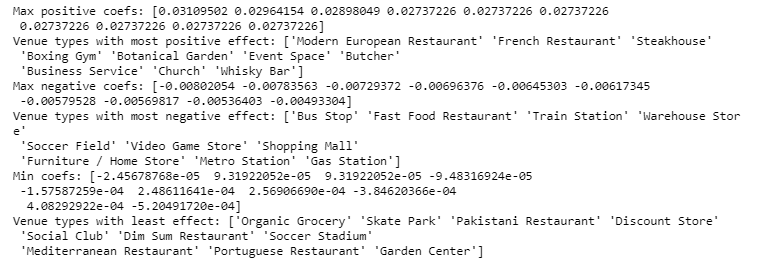
# Results

The score obtained through PCR is not the perfect score and it is advisable to use a more sophisticated method. But with the given score, a good insight is achieved as to how nearby venues affect the rental prices of apartments.

The poor model is due to:

* Higher number of features and a low number of samples.
* Machine learning technique used not suitable enough.
* Other factors that influence the price but not taken into consideration.

The concluding result included indication of the fact that restaurants, gym and gardens play an important role in raising the price of rents and that those apartments located near bus stops and train stations are not fancied by families to be much luxurious and thus have a lower rent.



# Discussion

It is observed that:

* Usually the required data is not available and requires further research to make solid conclusions.
* Combining data from different sources require that inconsistencies are identified and removed.
* Geojson files for UK are not available in much variety so mostly it requires conversion of shape format files into geojson which can be done using the ogr libraries of python.
* An area of research is now open to look for suitable techniques to construct a model that is worthwhile.

# Conclusion

The project aims at using data of coordinates and average rent of apartments available on different platforms along with the foursquare API to build a model that indicates the correlation between occurrences of different venues and the average prices of renting an apartment.

PCR was used but there are better techniques available that can be investigated and worked upon. This project is inspired by the work done by Toan Thien Le in the year 2018. I have tried to implement his model of New York City on my model of London and observe the results.

Thank you for going through my project. Please feel free to suggest me improvements that can be made.