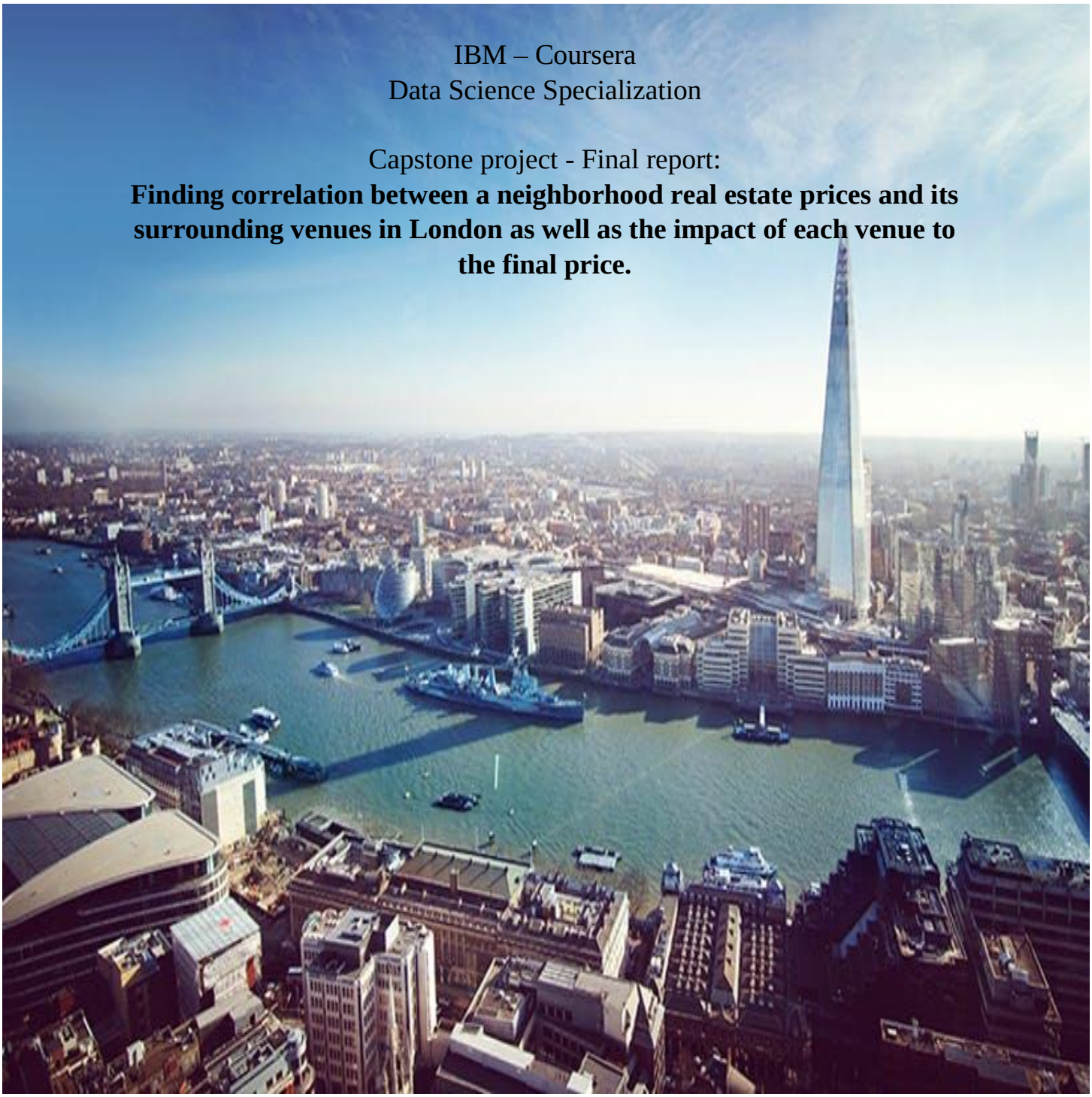
IBM – Coursera
Data Science Specialization

Capstone project - Final report:

Finding correlation between a neighborhood real estate prices and its surrounding venues in London as well as the impact of each venue to the final price.



By: Georgios Trakakis

Contents

Project Summary

London is a major global city and world cultural capital, with strengths in the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism, and transportation. A major settlement for 2,000 years, London is the world's largest financial center and the 5th or 6th richest city in the world. The property investment capital of the UK, London has a diverse array of housing and communities across a giant urban area. You can find different types of long-let investment property in various areas to suit most budgets, and being the world's most-visited city as measured by international arrivals means there is no shortage of short-let opportunities too. Having the largest concentration of higher education institutes in Europe guarantees that London has an active student property rental market as well.

The above reasons constitute London one of the most expensive places for someone to live. Demand for housing and real estate continues to be robust despite Brexit uncertainty. While it is clear that the UK Parliament wishes to avoid a no-deal Brexit, there is little consensus about what form a future arrangement with the EU should take.

The main objective of this survey is to examine what factors influence most the price of a real estate assets and which houses will be less inelastic to demand should a No deal Brexit occurs. How can the surroundings venues influence the price of the asset and in what extend? Which factor has the larger influence on the price of the asset? The above survey will help real estate investors to focus on regions that have surrounding venues and amenities that drive the price of the housing up. Therefore, should Brexit occur, their losses will be minimum in terms of asset devaluation.

Data Description

The data that will be used for this project will be acquired from the following sources:

- The Neighborhoods of London and their postcodes using Wikipedia(https://en.wikipedia.org/wiki/List_of_areas_of_London)
- The average prices of properties by neighborhood and by £/sqft in London(https://propertydata.co.uk/cities/london)
- The venues in each Neighborhood in London area(https://developer.foursquare.com/)
- Coordinates of each Neighborhood/postcode using geocoder python library(https://developers.arcgis.com/python/guide/using-the-geocode-function/)

The data will only include the price of each property using £/sqft since we do not want to consider properties that have different attributes (e.g. 2 bathrooms or 3 bedrooms) which will create noise in our survey.

Data Preprocessing Steps

1. Scrape the Neighborhoods html data from Wikipedia using Beautiful Soup library or pandas. read_html method.
2. Scrape the Property prices from html format on the internet to dataframe using Beautiful Soup library or pandas. read_html method.
3. Explore the Venues in each Neighborhood using Foursquare API.
4. Get the latitude and longitude data using Geocoder.Arcgis library.
5. Scrape all the data above and combine them into one dataframe using groupby function and inner joins.

The final dataframe is presented below:

	Neighbourhood	Accessories Store	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store
0	Abbey Wood	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00
1	Acton	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00
2	Addington	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00
3	Addiscombe	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00
4	Albany Park	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.03	0.02	0.01

Figure 1 - Final dataframe

Note here that the dataframe approximately 321 columns-venues and thus it is difficult to fit all the venues in the pdf. For more information, please see the Jupyter Notebook.

Method 1: Clustering

For the clustering method unsupervised K-Means Clustering was used. Before running the clustering in our dataframe the number of cluster should be determined. The number of cluster that will give us the best result will be selected. One way to find the best K cluster number is to compare the clusters with the ground truth, if it is available. However, because k -means is an unsupervised algorithm we usually do not have the ground truth in real word problems. But there is still a way to say how bad each cluster is based on the objective of the K-Means. This value is the average distance between data points within a cluster. Also average of the distances of data points from their cluster centroids can be used as a metric of error for the clustering algorithm.

There are some approaches to address this problem, and one of the techniques that is commonly used is to run the clustering across the different values of K and looking at a metric of accuracy for clustering. This metric can be mean distance between data points and their cluster's centroid, which indicate how dense our clusters are or, to what extent we minimize the error of clustering. Then looking at the change of this metric, we can find the best value of K. But the problem is that with increasing the number of clusters, the distance of centroids to data points will always decrease. This means increasing K will always decrease error. So that value of the metric as a function of K is plotted and the elbow point is determined, where the rate of decrease sharply shifts, this is the right K for clustering. This method is

called the elbow method. Below we can see from the graphs that the best number of cluster is 3, since after 3, the rate of change decreases.

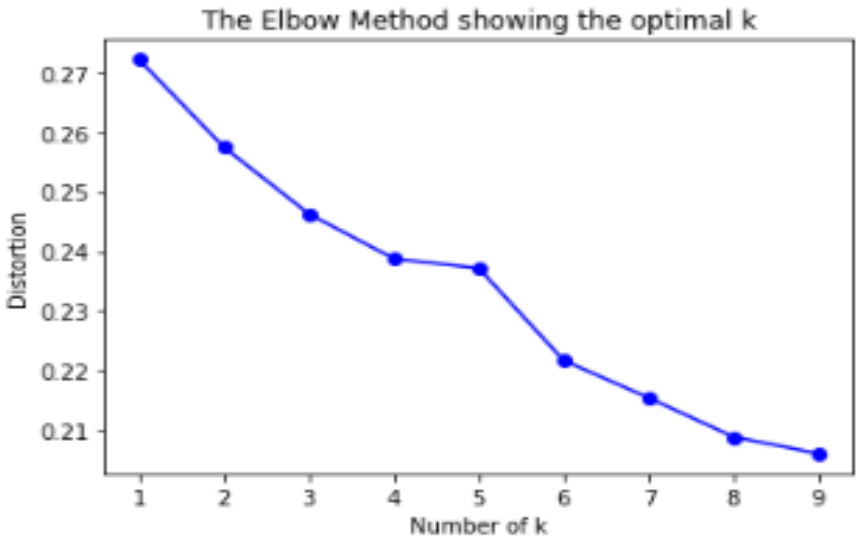


Figure 2 – Elbow Method

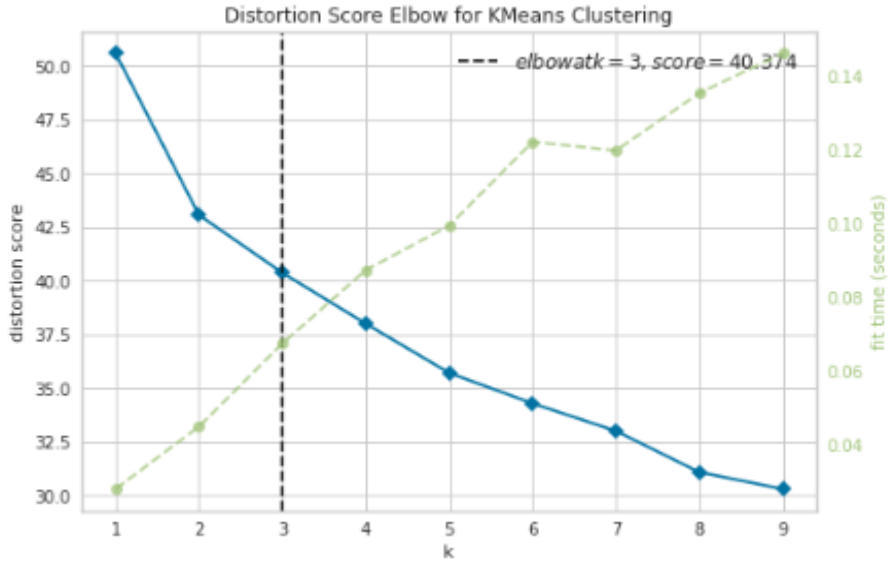


Figure 3 – Elbow Method

Method 1: Results

Finally, the required dataframes were merged. The final dataframe included the following: Neighborhood, Borough, Postcode, £/sqft, Latitude, Longitude, Cluster Labels and the 10 most common venues for all the neighborhoods. In order to label each price, prices histogram was visualized as seen below:

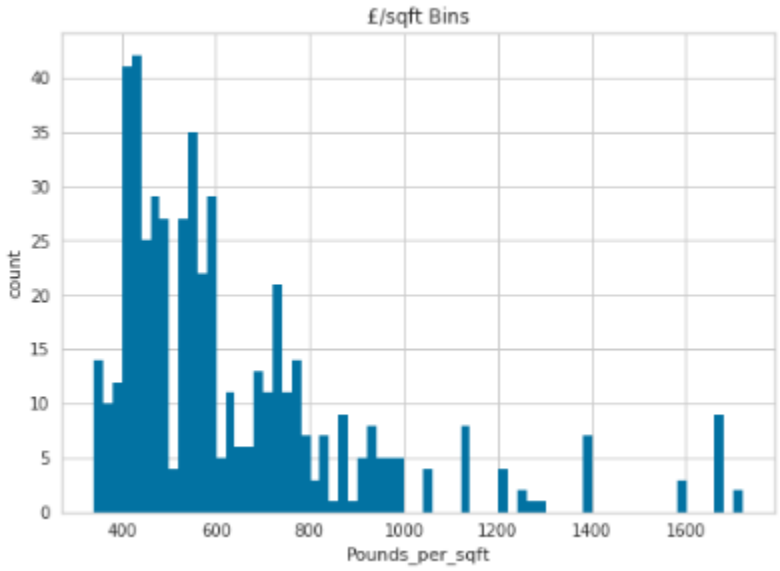


Figure 3 – Prices Histogram

Each cluster was labeled accordingly:

1. Low Price (340-600) £/sqft
2. Medium Price (600-800) £/sqft
3. Above Medium Price (800-1000) £/sqft
4. High Price (1000-1200) £/sqft
5. Luxury Price (1200-and above) £/sqft

The above clusters were added in the final dataframe in order to help us in our analysis.

Method 1 Conclusion

Cluster 1: Cluster 1 includes real estate assets with low to medium prices. Houses are surrounded by restaurants, cafes, coffee shops and pubs
Cluster 2: Cluster 2 includes real estate assets with low prices. Most common venues are Grocery Stores and since the Neighborhoods are not in a central location, asset prices are low.
Cluster 3: Cluster 3 includes most of the luxury real estate prices. Most common venues are Hotels, Museums, Pubs, French/Italian luxury restaurants and exhibits.

Method 2: Multiple Linear Regression

The rationale behind the Multiple Linear Regression model is that the real estate price of an asset is dependent on the venues nearby. The Multiple Linear regression uses the price_per_square_feet as the dependent(y) variable and the venue occurrences(X) as the independent variables. Before we run the regression y values should be normalized using the StandardScaler method. Then the data are split into train and test data in order to train our model and calculate the Regression metrics.

Multiple Linear Regression Metrics:
R2-score: -1.1213311866657404e+26
Mean Squared Error: 1.2918370416523627e+26

Method 2 Conclusion

It is clear that the multiple linear regression cannot be used to predict the prices of the real estate assets based on the venues nearby. R2-score: -1.1213311866657404e+26 is negative and thus our model is not a good predictor of the asset price.

Method 3: PCR Regression

PCR is a regression technique which is based on Principle Component Analysis. One of the most important applications of PCA is for speeding up machine learning algorithms. Firstly, we will perform PCA on the features set

to obtain the principle components. Then select a subset for the next step. Secondly we will use regression on the subset of principal components to get a list of coefficient correlations. PCA will reduce the dimension of our dataframe and thus we will be able to run the regression with the values that have the most contribution and make the largest impact in our dataframe. PCR technique will not include the correlation between the independent variables in our dataframe. Again, the data are split into train and test data in order to train our model and calculate the PCR Regression metrics.

PCR Regression Metrics:
R2 score: 0.7278400791689091
MSE: 0.31354364452148487

Method 3 Conclusion

The result is much better than that of the Multiple Linear Regression.R^2 is really high which means that 0.72% of the variation in our dependent variable is explained by our independent variables. Below you can see the coefficients with the most positive, negative and neutral influence along with the relevant venues.

```
Coefficients with the most positive influence: [0.06645841 0.06289679 0.05487673 0.0500742  0.04599851 0.0
4562024
 0.0447385  0.04433566 0.04427189 0.04244077]

Venue types with most positive effect: ['Modern European Restaurant' 'Salad Place' 'Juice Bar' 'Gift Shop'
 'Recording Studio' 'Kids Store' 'Peruvian Restaurant' 'Souvenir Shop'
 'Speakeasy' 'Cricket Ground']

Coefficients with the most negative influence: [-0.02311182 -0.02029252 -0.02013216 -0.01918561 -0.0189422
9 -0.01813977
 -0.01791299 -0.01740341 -0.01718586 -0.01710827]

Venue types with most negative effect: ['Bus Stop' 'North Indian Restaurant' 'Convenience Store' 'Train St
ation'
 'Monument / Landmark' 'Liquor Store' 'Supermarket' 'Art Museum' 'Park'
 'Church']

Coefficients with the least impact: [ 3.37594959e-06  3.37594959e-06 -5.31084390e-05  5.85648461e-05
 -1.46830424e-04 -2.70443051e-04 -3.45769000e-04 -3.95806513e-04
 -4.40554362e-04 -4.86574733e-04]

Coefficients that have the least impact: ['Shop & Service' 'Caucasian Restaurant' 'Pet Store'
 'Portuguese Restaurant' 'Entertainment Service' 'Dry Cleaner'
 'Recreation Center' 'Music Venue' 'Pub' 'Breakfast Spot']
```

We clearly observe that venues that have the most positive influence in increasing real estate asset price are 'Modern European Restaurant' 'Salad Place' 'Juice Bar' 'Gift Shop' ‘Recording Studio' 'Kids Store' 'Peruvian Restaurant' 'Souvenir Shop' 'Speakeasy' 'Cricket Ground'

Project Conclusion:

Overall both clustering and PCR Regression methods can be used to predict approximately a neighborhood's average house price per sqft.