**BATTLE OF NEIGHBORHOOD FINAL REPORT**

**Introduction/Business Problem**

**How to Choose the Best Location in London to start an Indian Restaurant?**

Of course, food and facility are significant to the success of a restaurant, but the location can be just as crucial, especially in the early years. Not all locations are suitable for opening a restaurant. It all depends on the style of restaurant, the neighborhood, the targeted customers, amenities, competitors. If these things are clear in one’s mind then they will be able to choose the best location to open up their business with good results. There are many things to keep in mind while opening a restaurant of your choice. The following points need to be considered before the business idea is put into action:

* The style of the restaurant

Does it focus on a specific cuisine?

How do you want the service to be: self-service or table service?

How should be the ambience: casual, fine-dining, or high-end?

* The target customers

To start an Indian restaurant in London, one needs to know the pulse of customers: age group, their tastes, the amount they are willing to spend, are we focusing just on just Indians, or other Asians

* Competition

Another main point to consider is the competitors nearby. How well they are doing their business. What kind of restaurants they are running. Their positives and negatives. How are they attracting their customers? It will be highly competitive if there are similar restaurants in the nearby locations. It is better to look for area where there is less choices and high demand.

* Menu

It is the heart of the restaurant. It is where everything else can be compromised. India being a hub for variety of food from north to south, you should be able to please customers from every part of India.

* Location

Selecting the right location is a challenging tasks of all. Many factors needs to be considered and compromised before you fix the location, like:

* Accessibility
* Parking facility: If the restaurant is located at place where the public transport is less assessable, you have to check the parking facilities. If customers have to walk a fair distance to get to your restaurant, they may opt to go somewhere else "more convenient."
* Cost
* Building type and size

**Data**

British Indians form the largest ethno-national group in London with a population of around 542,857 or 6.6% of the population, making them the single largest visible ethnic minority population in the country.

The London boroughs are the 32 local authority districts that make up Greater London. The London boroughs have populations of around 150,000 to 300,000. Inner London boroughs tend to be smaller, in both population and area, and more densely populated than Outer London boroughs. This will be focus for solving the restaurant problems.

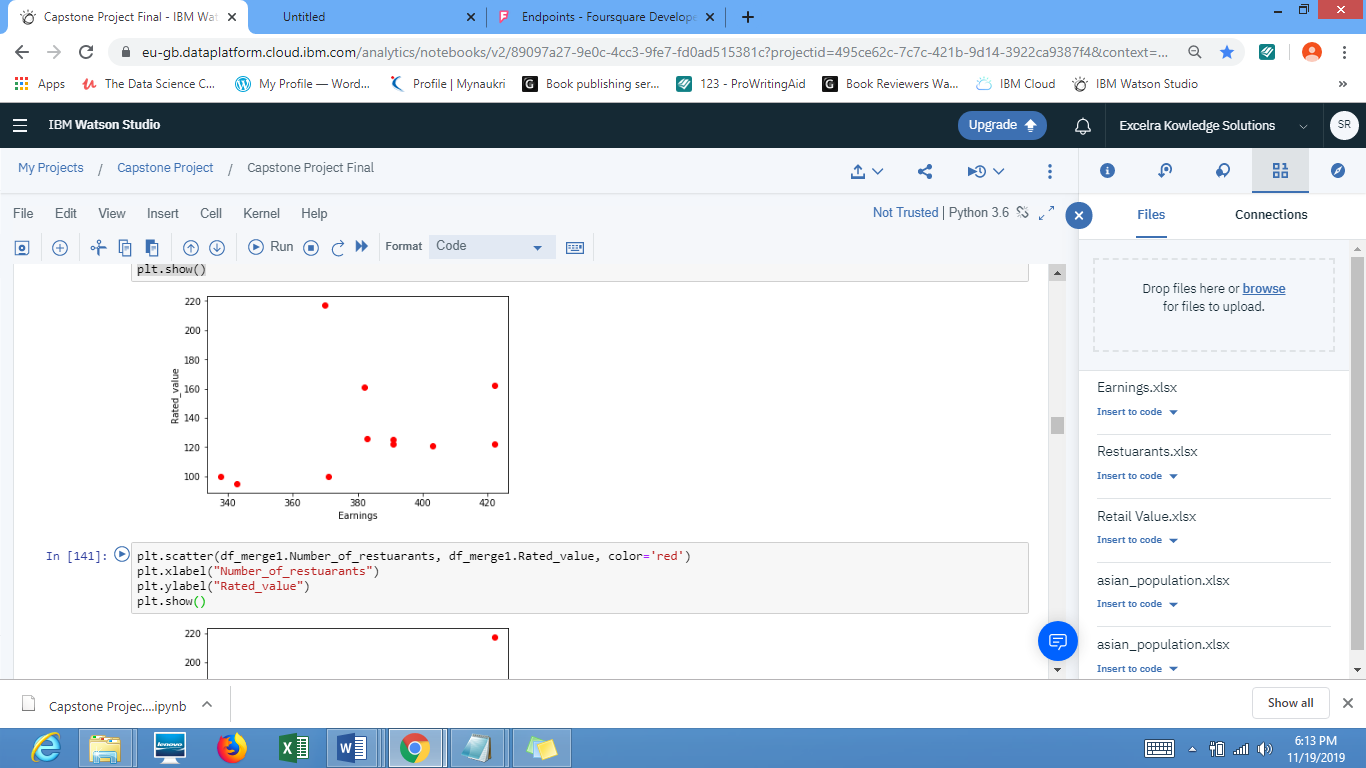
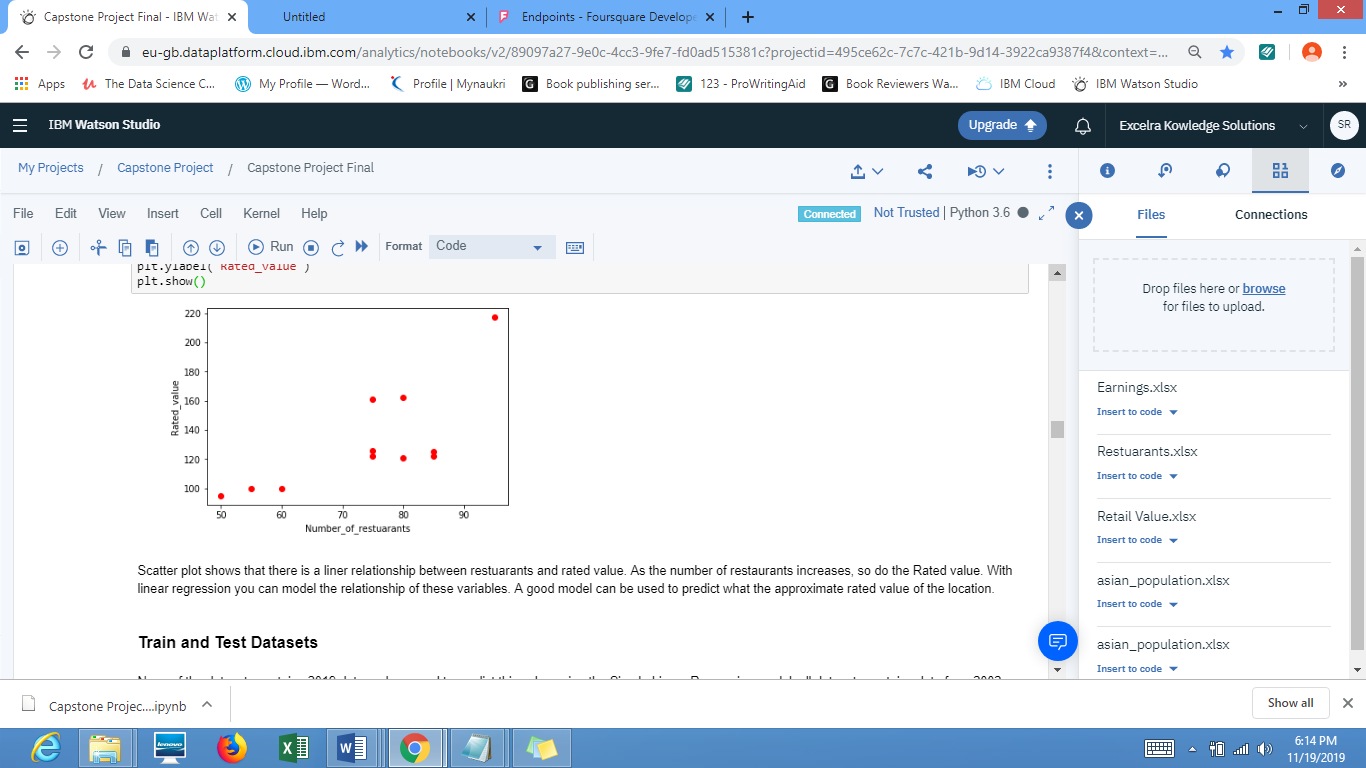
To solve the problem we need datasets based on:

* The target audience: Age, Gender, Marital status, Employment status, Income
* Existing Restaurants in each boroughs
* Rateable Value Statistics for Commercial and Industrial Floorspace
* Tourism spend estimates as well as domestic spend estimates

These data can be freely obtained from London Datastore (<https://data.london.gov.uk/>) in csv and xls format. And Foursquare location data is also used.

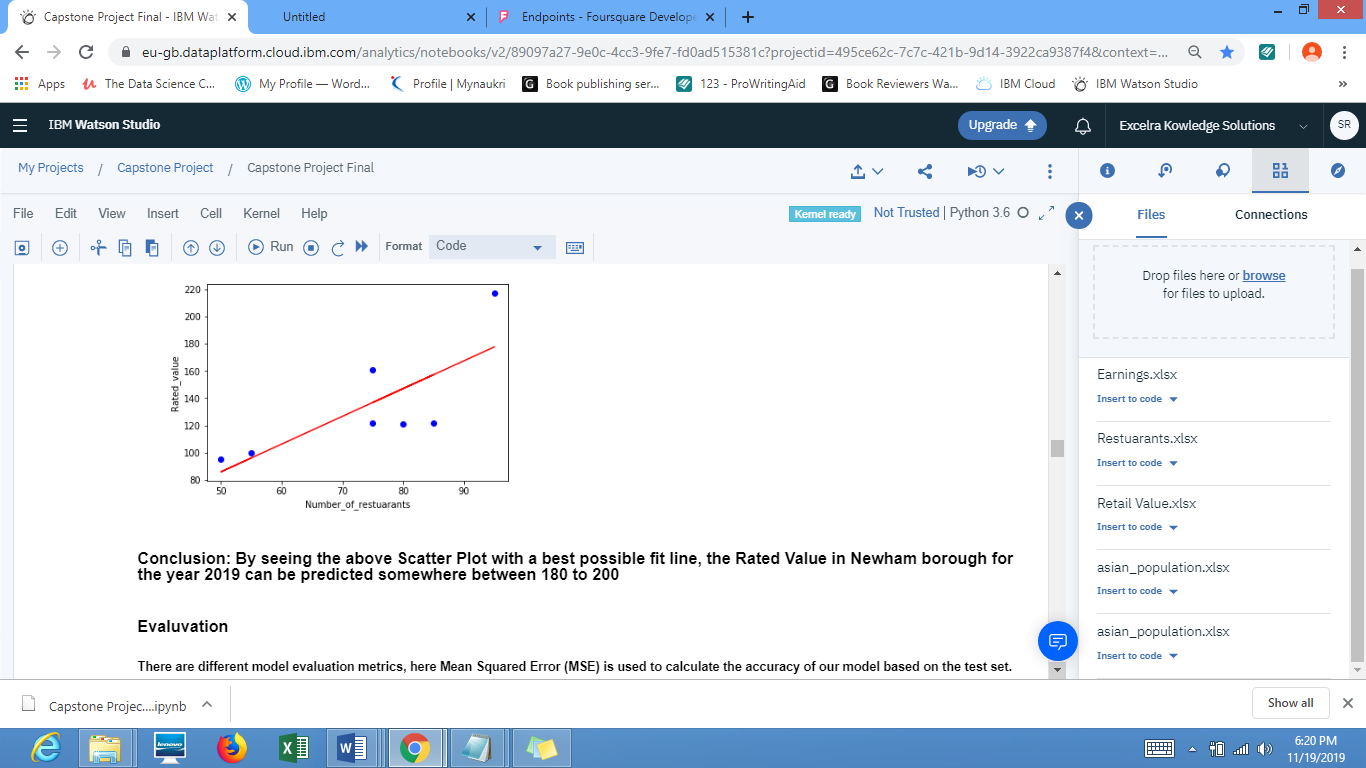
**Methodology**

The required libraries are imported including pandas and numpy. Data’s from London datastore based on ethnicity, weekly earnings, number of restaurants, and retail value of rated value per square meter is collected and cleaned.

Simple linear Regression analysis was the used to predict the rated value for the year 2019 with earnings and number of restaurants as independent variables. 

Based on the scatter plot it is concluded that there is a relation between rated value and number of restuarants, even though there is some clattering of data which needs to be cleaned up.

Then data were divided into train and test data with 2202 to 2012 results. After separting the data the test data is used to evaluvate the model.



There are different model evaluation metrics, here Mean Squared Error (MSE) is used to calculate the accuracy of our model based on the test set.

Mean absolute error: It is the mean of the absolute value of the errors. This is the easiest of the metrics to understand since it’s just average error

Mean Squared Error (MSE): It is the mean of the squared error. It’s more popular than Mean absolute error because the focus is geared more towards large errors

R-squared: It represents how close the data are to the fitted regression line. The higher the R-squared, the better the model fits your data. Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse).

Next step is to segment and cluster the data. To explore the data the latitude and longitude coordinates of all Boroughs in London from a Wikipedia link is taken (<https://en.wikipedia.org/wiki/List_of_London_boroughs>)

Imported request.get, BeautifulSoup to read and explore the data.

Get the Latitude and Longitude of London City using geopy library. Create a map of London with Boroughs superimposed on top using Folium library. Create a DataFrame from the Areas and Link list, and merge these two.

Create geo co-ordinates for all the areas in the Newham borough and merge the Areas(Neighborhoods) and Geocodes DataFrames.

Foursquare API is used to explore the neighborhoods in London City. Explore function to get the most common venue categories in each neighborhood and group the neighborhoods into clusters. To cluster the neighborhoods K-means Clustering algorithm is used. getNearbyVenues() function is used to give the venue details like venue name, venue latitude, venue longitude, venue category along with neighborhood name, latitude and longitude for each neighborhood.

One-Hot encoding is applied on the venue category data, so that the analysis of the data will be easy in grouping the neighborhoods based on the frequency of occurrence of each venue category

After grouping the neighborhoods based on the occurrence of the venue category frequency, the top 10 venues of each neighborhood are displayed as a dataframe.

K-means Clustering algorithm is applied to the result to segment the data into 5 Clusters and all these 5 clusters are visualized in a map using the Folium library and finally the 5 clusters are examined to determine the refined venue categories that differentiate each cluster.

**Results**

After exploring the Asian population datasets the Newham borough is the borough with highest asian population, including Indians and the price is lesser compared to other boroughs using the rated value datasets. The Newham borough has 123 existing restaurants and taking this as a independent variable (X variable) predicted the rated value per sqm(dependant variable) between 180 to 200 using the Linear Regression model. The MAE and R-Squared with the test data was calculated, but the values were not so good.

In the Segmenting and Clustering section, the neighborhoods of Newham borough are explored, and the top 10 venues of each neighborhood are listed. The neighborhoods are Clustered into 5 clusters using K-means algorithm and their most common neighborhoods are identified. After applying the K-means algorithm the 5 neighborhoods Manor Park, Beckston, Maryland, Custom House and Plaistow are the best locations to open an Indian restaurant.

**Discussion**

Due to constriction of data the results are not so good, it could have been better. But overall using the data especially clustering gave a result which was satisfactory. From this best 5 places were able to be identified to kick start an Indian Restaurant such as Manor Park, Beckston, Maryland, Custom House and Plaistow

**Conclusion**:

Better datasets can be used to improve the results and datasets with more recent data will give more accuracy to the problem solution.