Strategic Location for Establishing an Asian Restaurant in Seattle

1) Business Problem and Background

Before starting your restaurant, we need to thoroughly think through the budget and initial expenses. This can help you to experience a smooth and successful launch whether you are starting from scratch or buying an existing one. Many restaurants fail to pay equal consideration to such factors and thereby results in the end of restaurant business.

The potential success of a restaurant depends on numerous factors such as brand value, customer fidelity, demand rate and quality of food. Apart from these, Location undoubtedly plays a significant factor in this decision-making process to make the business profitable and competitive in the existing market condition. A client seeks to establish a franchised Asian restaurant in Seattle, Washington. In this project, we are seeking to find the optimal and strategic area of neighborhood based on the aforementioned factors to open this new promising business. The foundation of our reasoning will be the median income of the households by neighborhood, the number of Asians living in each neighborhood and the underlying market competition. Prospective entrepreneurs seeking to set up a new restaurant specializing in a particular niche can find the insights derived through this project compelling and can indeed reap significant benefits.

2) Data Extraction and Understanding

The dataset that we need for this project is not readily available as a coma separated file (csv) or in a structured database. As a result, we would be utilizing Data scraping techniques to extract the required data to proceed with further analysis. Web scraping deals with extracting data from a website automatically with the help of web crawlers. Web crawlers are scripts that connect to World Wide Web using HTTP protocols and allows them to fetch data in an automated manner. The data about the Seattle neighborhoods were scraped from the Wikipedia page (https://en.wikipedia.org/wiki/List\_of\_neighborhoods\_in\_Seattle). It contains the neighborhood names and the greater district name to which they belong. Geopy library was used for geocoding to obtain the latitudes and longitudes of each neighborhood. Furthermore, the median household income by neighborhood, neighborhood population and the percentage of Asians living in each neighborhood were scraped from another website and added to the main data frame. The restaurant data for each neighborhood was derived using the Foursquare API and was integrated with the data frame. By using the FOURSQUARE API, /venues/explore endpoint, the number of Asian restaurants were found in each neighborhood. By using this parameter, it will be useful to judge the level of competition that will be faced by our new restaurant. This data was extracted from multiple websites and placed in a single data frame.

The data consists of the following attributes:

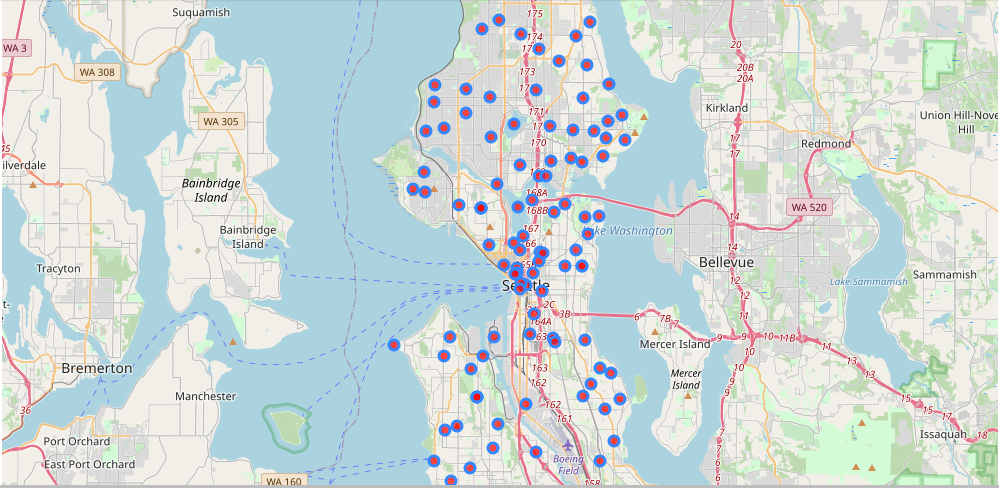
1. Neighborhood name - Unique name for each neighborhood of Seattle city.
2. Larger Neighborhood name - the district within which each neighborhood comes.
3. Latitude of each neighborhood.
4. Longitude of each neighborhood.
5. Median household income - Median household income of households living in each neighborhood.
6. Population - Population of each neighborhood.
7. Asians Percent - Denotes the percentage of Asians living in each neighborhood.
8. Number of Asian Restaurants – Denotes the number of Asian restaurants in each neighborhood.

3) Exploratory Data Analysis

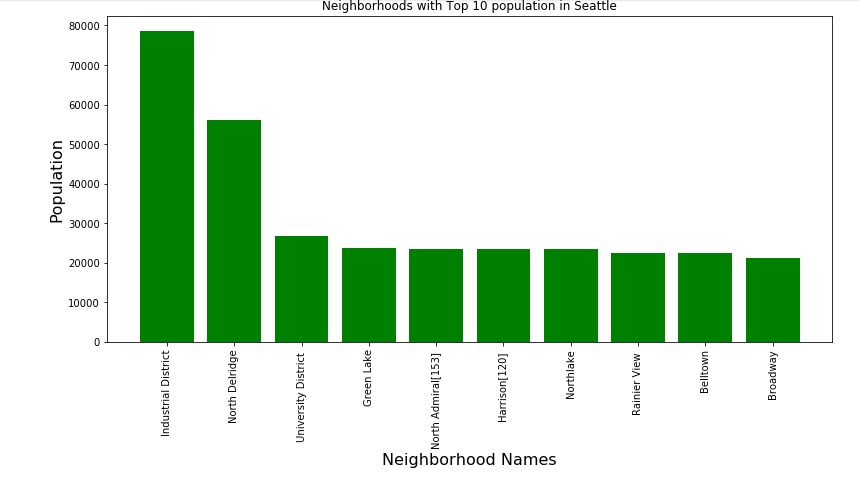
Exploratory data analysis is the preliminary step in data analysis. It is used to summarize the main characteristics of data, gain better understanding of data set and to unleash the various relationships in the underlying data. It can also be used to determine features that have a significant impact on the machine learning algorithm in which the predictive model will be deployed. It refers to the process of performing critical investigations on data so as to discover patterns, to search anomalies, to test hypothesis and to check assumptions with the help of statistical analysis.

It is good to understand the data first and try to get as many insights from it as possible. Exploratory Data analysis is all about making sense in data before getting our hands dirty on it. Statistical analysis is used to carry out analysis such as mean, median and mode in the data set. The pandas data frame has various functions such as describe(), info() and head() are used to explore the initial patterns in the data.

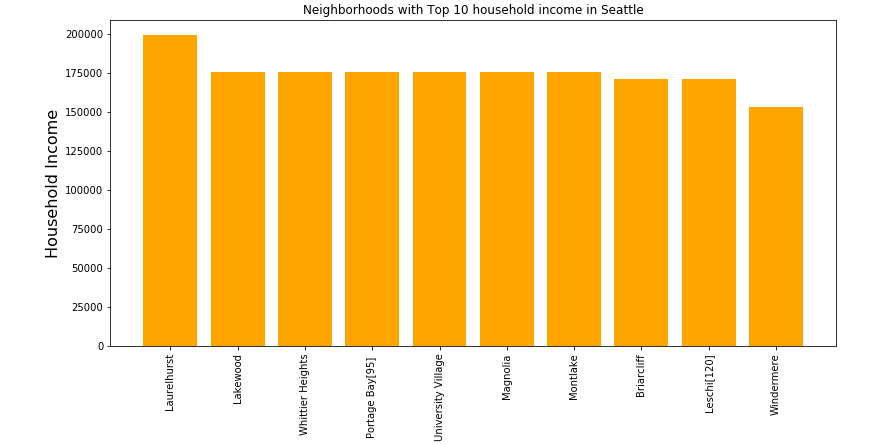
In our analysis, we would be using the Leaflet map to visualize the different neighborhoods in Seattle area.



Distribution of Population by Neighborhood Top 10.



Distribution Of Household Income by Neighborhood Top 10.



4) Data Pre-processing.

When we think about data, we usually think about large datasets, usually comprised about a number of rows and columns. While its likely a scenario, its definitely not the case. Most of the times the data is available in unstructured formats like videos, audios, images, text, speech recordings and all.

As the data for this project was scraped from multiple websites and joined together to create a parent data frame, the quality of data should be our priority. These quality should be handled adequately before modelling the data with a machine learning algorithm.

1. Missing values : It is very normal to have missing values in our data since they are collected from a variety of different sources and the data collection techniques may not be reliable. For whatever reason it might be, the missing values must be taken into consideration before proceeding with the predictive analysis.

There are multiple ways to handle missing data. Either we can drop the rows that contain the missing values or we can estimate the missing values using the aggregate functions available. In our analysis we will be estimating the missing values by using the average value of the column for that value. As a result we won’t be needed to drop the rows and the entire dataset can be used for modelling.

1. Dimensionality Reduction

Most real world datasets have a large number of features. For example, consider an image processing problem, we might have to deal with thousands of features, also called as dimensions. As the name suggests, dimensionality reduction aims to reduce the number of features - but not simply by selecting a sample of features from the feature-set, which is something else — Feature Subset Selection or simply Feature Selection.

Conceptually, dimension refers to the number of geometric planes the dataset lies in, which could be high so much so that it cannot be visualized with pen and paper. More the number of such planes, more is the complexity of the dataset.

What dimension reduction essentially does is that it maps the dataset to a lower-dimensional space, which may very well be to a number of planes which can now be visualized, say 2D. The basic objective of techniques which are used for this purpose is to reduce the dimensionality of a dataset by creating new features which are a combination of the old features. In other words, the higher-dimensional feature-space is mapped to a lower-dimensional feature-space.

In our project we achieve dimensionality reduction by combining the population and the percentage of Asians in each neighborhood by calculating the Asian population and using it for our analysis. As a result it aims to understand the number of our target customers present in each neighborhood, thereby giving us an important parameter for modelling capturing the essence of both the features in our data. Hence the population and percent of Asians columns of our data will not be used for analysis.

1. Feature Scaling/ Standardization

The different features in our dataset belong to different ranges. These differences in ranges can pose a serious threat to our analysis since there is a possibility that one feature might dominate our analysis than the rest. To overcome these problems, Standardization techniques are used to scale the features so that they all fall within a same range. As a result, all the features contribute equally to our predictive analysis.

In our analysis, we used the sklearn.preprocessing.StandardScaler class to perform feature scaling and get the data ready for our modelling.

5) Predictive Modelling

Predictive modelling entails the use of machine learning algorithms to analyze the given dataset and anticipate the outcomes for new data. In this project, we would be using the unsupervised machine learning algorithms to divide the neighborhoods in Seattle in a cluster such that the neighborhoods having similarity are placed in same cluster. This technique aims to create clusters such that all the members in a cluster have less intracluster distances and more intercluster distances.

We will be implementing the K Means clustering algorithm to cluster our neighborhood data.

**Kmeans** algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k means algorithm works is as follows:

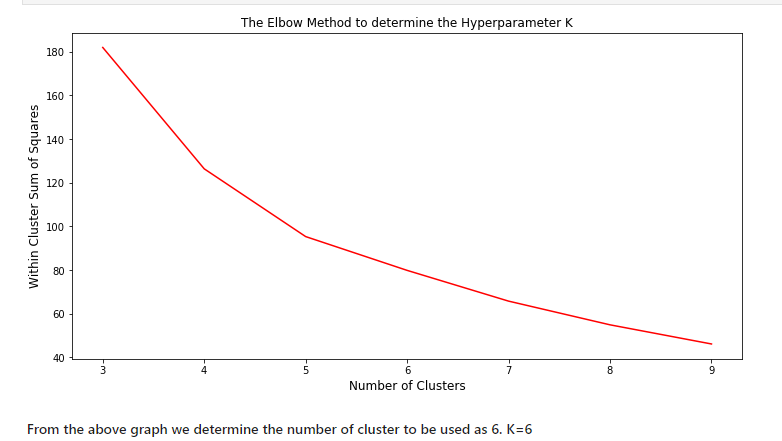
1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K*data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing. Compute the sum of the squared distance between data points and all centroids. Assign each data point to the closest cluster (centroid). Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Since clustering algorithms including k means use distance-based measurements to determine the similarity between data points, it’s recommended to standardize the data to have a mean of zero and a standard deviation of one since almost always the features in any dataset would have different units of measurements such as age vs income.

Given k means iterative nature and the random initialization of centroids at the start of the algorithm, different initializations may lead to different clusters since k means algorithm may stuck in a local optimum and may not converge to global optimum. Therefore, it’s recommended to run the algorithm using different initializations of centroids and pick the results of the run that that yielded the lower sum of squared distance.

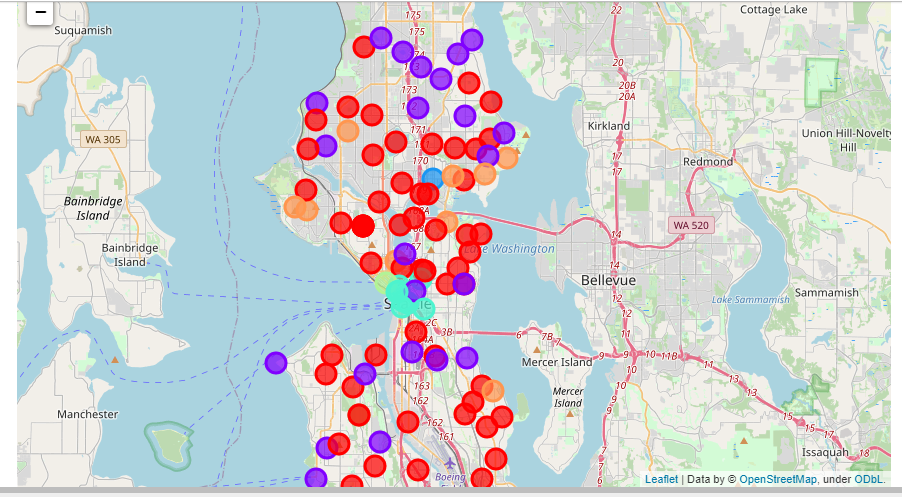
Contrary to supervised learning where we have the ground truth to evaluate the model’s performance, clustering analysis doesn’t have a solid evaluation metric that we can use to evaluate the outcome of different clustering algorithms. Moreover, since k means requires k as an input and doesn’t learn it from data, there is no right answer in terms of the number of clusters that we should have in any problem. Sometimes domain knowledge and intuition may help but usually that is not the case. In the cluster-predict methodology, we can evaluate how well the models are performing based on different K clusters since clusters are used in the downstream modeling.

**Elbow** method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters’ centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow. We’ll use the geyser dataset and evaluate SSE for different values of k and see where the curve might form an elbow and flatten out.



We use the sklearn.cluster.Kmeans to cluster the neighborhoods of Seattle.

Visualizing the Clusters formed in Seattle.



6) Cluster Summary

In summary, neighborhoods in **Cluster 0** possess the medium relative spending power as their median household incomes are significantly higher than the other clusters. The fair number of Southeast Asians and Asian restaurants indicates an adequate demand for Southeast Asian cuisine; and customers that are capable of consuming the meals more regularly than the others.

For **Cluster 1**, although the cluster possesses high spending power in residents, there are less number of Asians living in these neighborhoods. As a result not much of Asian restaurants are available in this region. Therefore, this cluster will be the least desirable option for establishing a new restaurant.

**Cluster 2** is only promising in terms of its great population of Southeast Asians, or target customers, in the region. However, its low spending power and the high number of competitors of similar niche may indicate a high barrier of entry — its demographic is likely to demand less of the pricier menu of our client’s compared to the cheaper options available from the competitors.

**Cluster 3**, though favorable concerning its demographic and optimal number of competitors, its low spending power would unfortunately demonstrate a lack in consumption.

With regards to **Cluster 4,** the residents possess above average spending power as indicated by their average incomes values. There exists a reasonable amount of Asians residing in these neighborhoods and in contrast, the number of Asian restaurants is relatively small. As a result, this cluster is the **perfect** for an entrepreneur who wishes to start an Asian restaurant business in Seattle.

**Cluster 5** only consists of one neighborhood as a result of its significantly high number of competitors and the distribution of Southeast Asians. This shows promise despite its high number of competitors. If the client’s franchised restaurant is priory well established, neighborhoods in ‘Cluster Label 5’ might also be up consideration as the restaurant would have the competitive advantage of brand loyalty to combat the density of competitors.

7) Conclusion

In this study, I have labeled the neighborhoods corresponding to their characteristics — spending power, percentage of target customers, and the number of competitors. The most promising group of neighborhoods for opening an Asian Restaurant, with a niche in Southeast Asian cuisine, appears to be ‘Cluster Label 4’. The higher spending power of the neighbourhoods in this cluster allows them to readily afford the slightly up scaled prices of the client’s Asian restaurant menu. The average distribution of the percentage of target customers — the Southeast Asian demographic — indicates a relatively reasonable demand for the Asian cuisine. The number of competitors is not significant yet adequate enough to be a good indicator of demand for Asian cuisine.

Our client could more specifically consider Industrial District as a location of establishment for optimal results. However, whenever there is a shift in the dynamic of business demands, we could always target different clusters of neighborhoods. Case in point, if the client has plans to expand a well-established franchised restaurant, neighborhoods in ‘Cluster Label 5’ would be the optimal location; this is under the assumption that the aforementioned restaurant would have a competitive advantage of brand loyalty against the high number of competitors in that neighborhoods.