**Battle of the Neighborhoods Report**

By: Renxiang Tang

**Introduction**

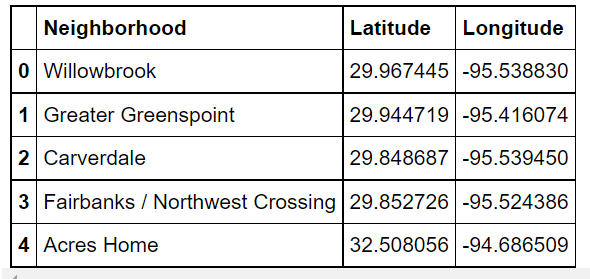
For this Capstone Project, I will be leveraging the Foursquare location data in order to determine optimal places to open a Restaurant in the Houston Area. I will be looking at specific neighborhoods in the area to determine whether a neighborhood is viable by comparing the variety of venues available, as well as the amount of competition for Restaurants. This will serve as a great demonstration on the use of the Foursquare API in order to solve business problems.

**Data**

Using the Foursquare API, I will be able to obtain information on various venues available in the Houston area. With this information, I hope to be able to group neighborhoods into clusters with similar venue makeup, and analyzing which clusters are able to sustain a new Restaurant. For breaking down the Houston Area into individual neighborhoods, we can webscrape information using websites such as Wikipedia.

**Methodology**

Webscraping for location information, will be handled with existing python packages to issue a request to relevant Wikipedia pages and formatting the data using BeautifulSoup. The resulting information will be assigned to a data frame like the following.



Using this information, we will leverage the Foursquare API to retrieve information on venues around the neighborhoods. For this project, we retrieved information on 544 venues. We then group the venues into their respective neighborhoods and find out the most popular venues available for each neighborhood. The resulting data frame looks something like this.

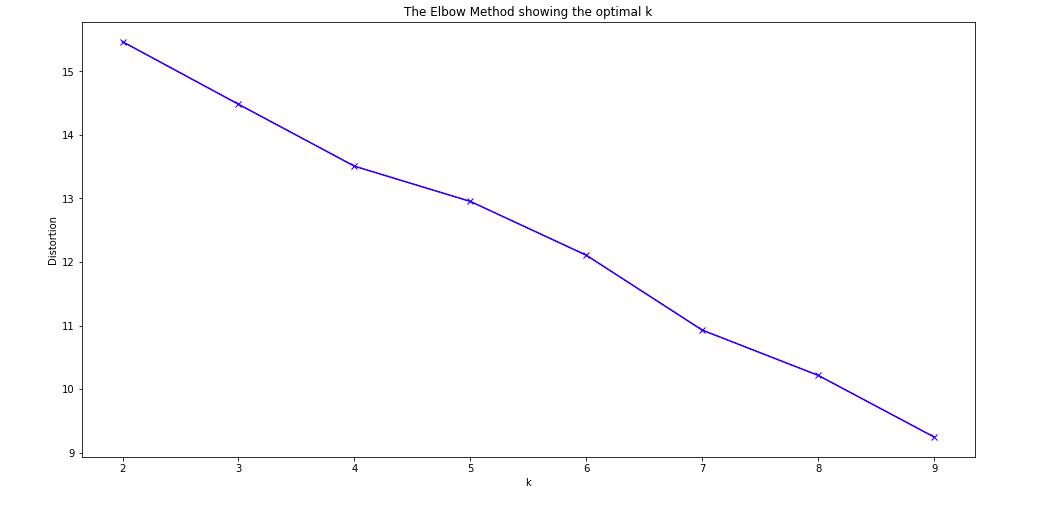


With this we can move on to analyze the data.

**Analysis**

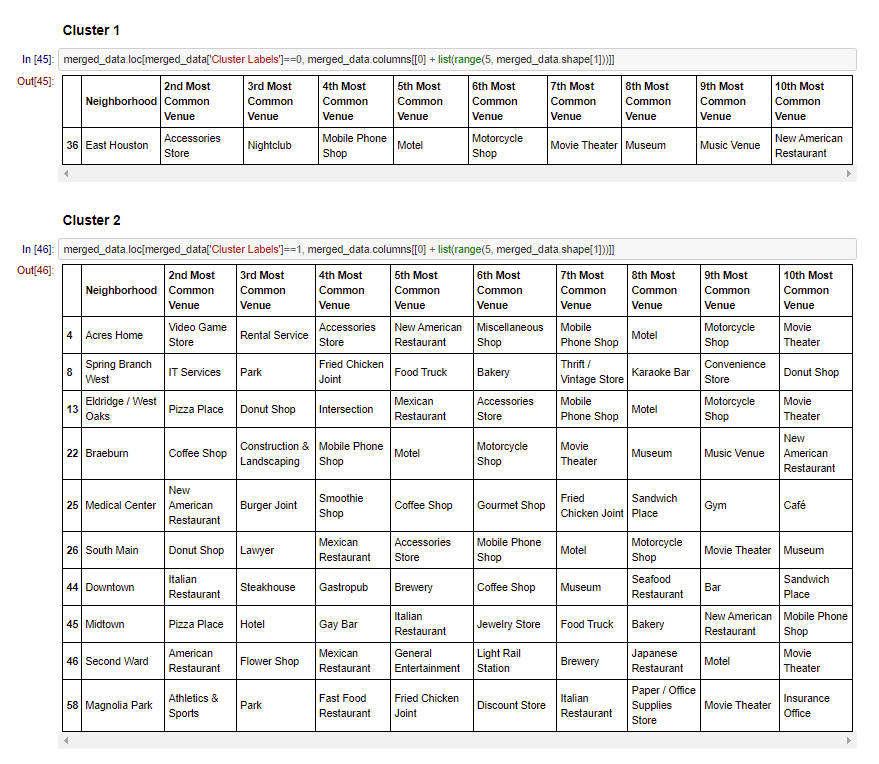
By manipulating the data so that each venues have its own field and then normalizing the values, we can then use the resulting data frame and apply a k-means clustering algorithm to reveal any patterns.

In order to do that, we ill first try and find the optimal k value for the algorithm. We will use the elbow method and iterate the algorithm through a range of k values. The results are as followed.

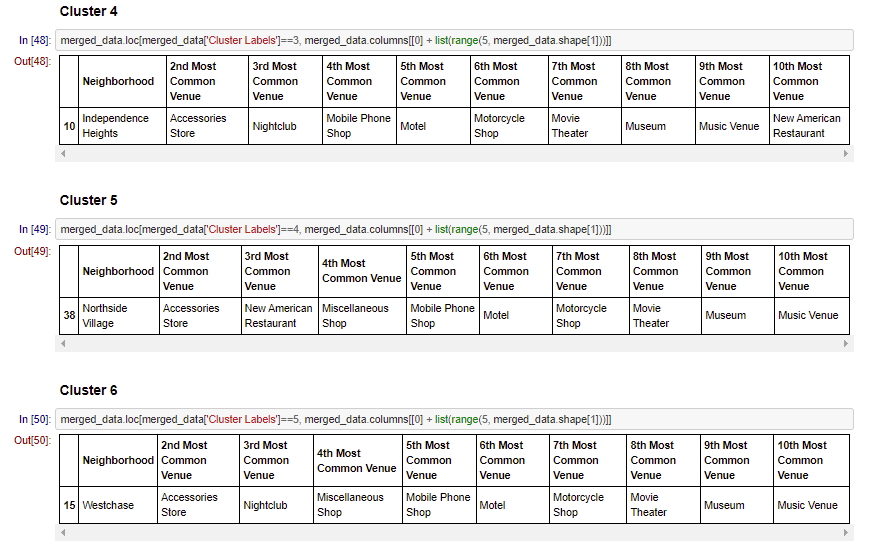


This actually show that our data may not be able to create good cluster. This may be due to the data set not being large enough. However, we will be using k=6 to further analyze the neighborhoods.

With that, the following clusters were created.







Now by analyzing these clusters, we can see that neighborhoods belonging to clusters 2 and 3, may have too much competition to start a new restaurant, where as neighborhoods such as Northside Village, or East Houston offers limited competition while still containing a variety of venues to have enough traffic for sustaining a business.

**Results and Discussion**

Based on the results, my opinion is to make is to start a new restaurant in locations such as East Houston or Westchase. The primary reason being that based on the popular venues, there are less competition for restaurant businesses. This coupled with a variety of venues to induce traffic can aid in sustaining a business.

However, this conclusion is flawed due to the fact that we were only able to retrieve a limited number of venues using the Foursquare API. There were too many unique values which made the clustering algorithm not be able to produce efficient groupings for effective analysis. Our analysis is very much limited by the data which we are able to gather. With that being said, the proof of concept is here and stakeholders can leverage this method to aid them in many business decisions.

**Conclusion**

The purpose of this project is to demonstrate the value of leveraging the Foursquare API to gather data in order to produce informed business decision, such as opening a new restaurant. Using Foursquare, we obtained information on venues for multiple places and using available python packages, analyzed the neighborhoods to determine an optimal location.

However, this type of analysis is limited by the quantity and quality of our data. In this specific example, We can see that we could not obtain enough data to produce high quality clusters of neighborhoods. A solution to this problem can be to manipulate the available data to narrow down unique category tags to help out the clustering algorithm.