IBM Data Science Capstone Project

**Battle of Neighborhoods**

Determining a suitable business location with the power of Data and Science

horizontal line

# Placeholder image

**Problem Description**

Wake county located in North Carolina is one of the fastest growing counties in the USA with large influx of immigrants from other states and nationalities calling it their home.Its major city Raleigh has a high concentration of highly educated workforce with a good mix of young and middle aged population drives the economy of the state and opens up tremendous possibilities for new businesses to thrive.North carolina has been influenced greatly by settlers from New England and Hispanic immigrants.North carolina is also famous for its barbecue and bacon but other cuisines are starting to become popular. This opens up a possibility to create a recommender system for anyone who wants to open a new restaurant that can be successful in the area.We will attempt to determine a good location to open a new restaurant by exploring the neighbourhoods for restaurants distribution within Raleigh using recently learnt data science techniques

The following datasets will be used for our assessment

Raleigh restaurants info using Foursquare

Raleigh Neighbourhoods

1. **https://en.wikipedia.org/wiki/Raleigh,\_North\_Carolina\_neighborhoods**

2. **https://en.wikipedia.org/wiki/Category:Neighborhoods\_in\_Raleigh,\_North\_Carolina**

3. **Restaurant density data from Foursquare API’s**

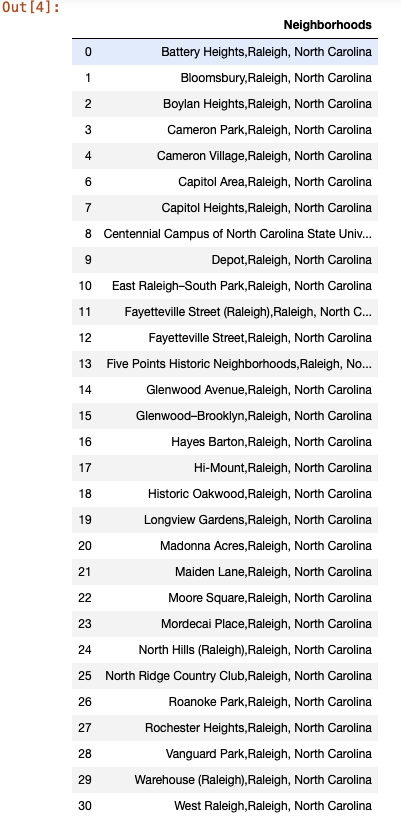
4. **Classification of types of restaurants based on cuisines**

**Data exploration and Presentation**

The data from the above datasets will be read via pandas dataframe and cleaned up to classify Restaurants to Neighborhoods.Then we will be able to cluster the restaurants in the area and as well as correlate the quality of restaurants again based on cuisine to the neighbourhood which would potentially provide us insights on the type of restaurant that can be setup. We will then perform data clustering using KNN on the restaurant data and generate visualisation plots and graphs that will aid in the final decision

**Raleigh Neighborhood Exploration**

We begin by exploring the Raleigh City.We identified that there are 32 distinct neighbourhoods in Raleigh.We sourced this data from wikipedia and obtained the location information.We then removed some overlapping neighbourhoods and cleaned this data for the final set of 27 neighborhoods



### In addition to the above we also sourced data of the 7 major Raleigh area classification listed below.We ran the neighbourhood data using the Nominatim library but this yielded incorrect latitude and longitude information so we had to manually cleaned up this data.Then we created a dataframe using the Neighbourhood,latitude and longitude list data



In [9]:

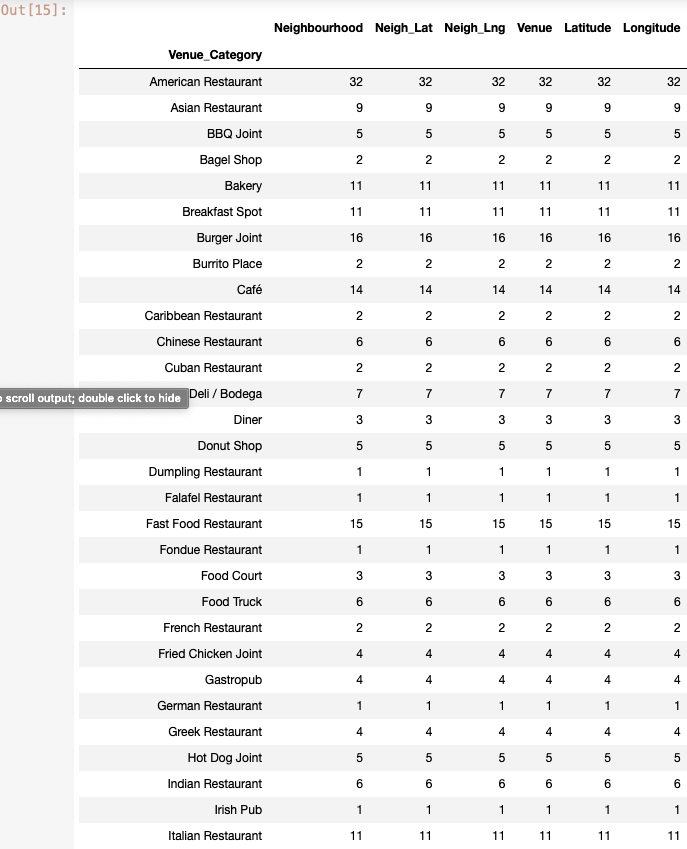
With the neighbourhood dataframe we will use this feed the location data to our foursquare API call to get list of 50 restaurants within 5km radius in each of the neighbourhood

### We used the list data from our Foursquare api call to create a dataframe final\_Ral.From this dataframe

1. we were able to get a count of venue categories by neighbourhood.
2. Most popular cuisine types across multiple neighbourhoods

Looks like North Raleigh has the least number of restaurants.Unsurprisingly American,Mexican and Pizza seem to top the charts.We then get a list of top ten restaurant categories and present a bar graph





We obtain the following data and save it in a list.We will use this for our various analysis going forward

1. Neighbourhood Latitude
2. Neighbourhood Longitude
3. Venue Name
4. Venue Category
5. Venue Latitude
6. Venue Longitude

### 

### 

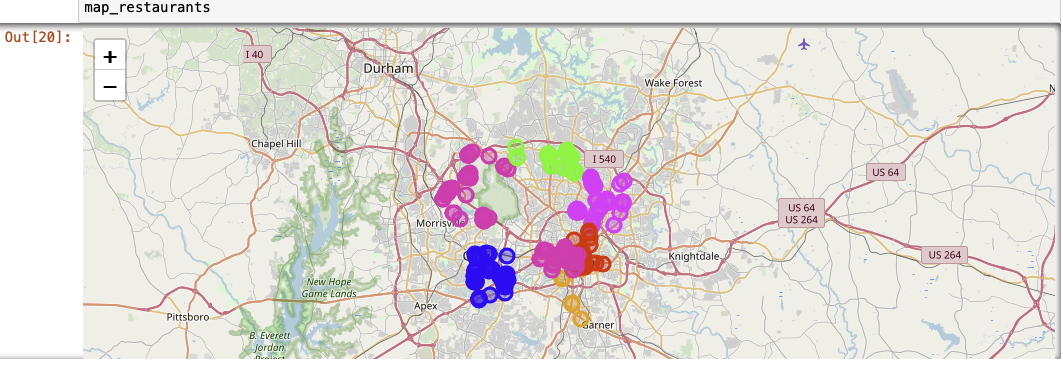
### 

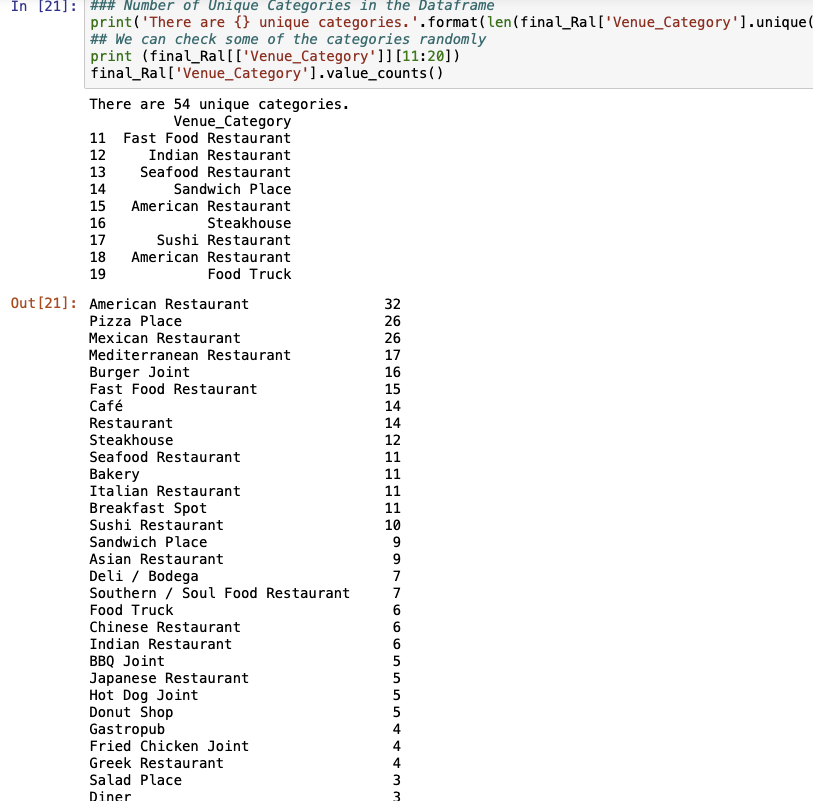
### We used the list data from our Foursquare api call to create a dataframe final\_Ral.From this dataframe

1. we were able to get a count of venue categories by neighbourhood.
2. Most popular cuisine types across multiple neighbourhoods

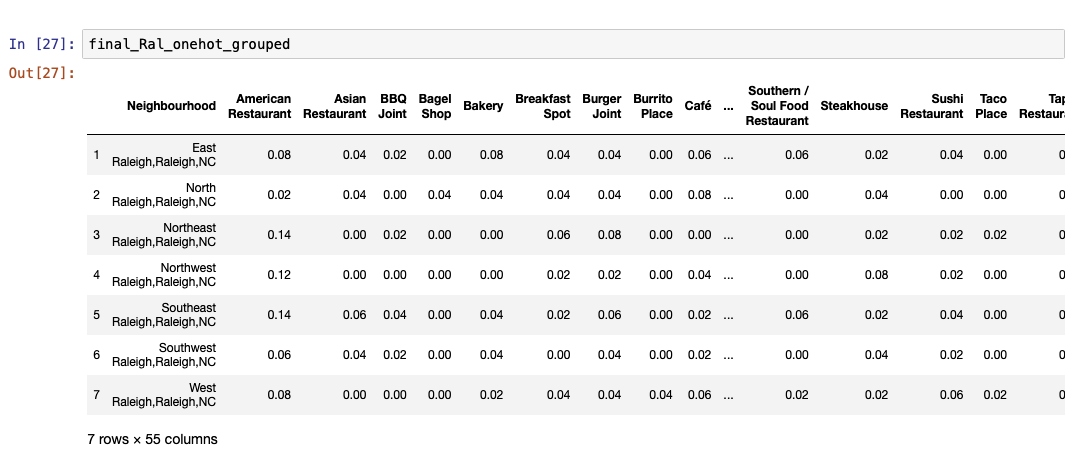
Looks like North Raleigh has the least number of restaurants.Unsurprisingly American,Mexican and Pizza seem to top the charts.We then get a list of top ten restaurant categories and present a bar graph

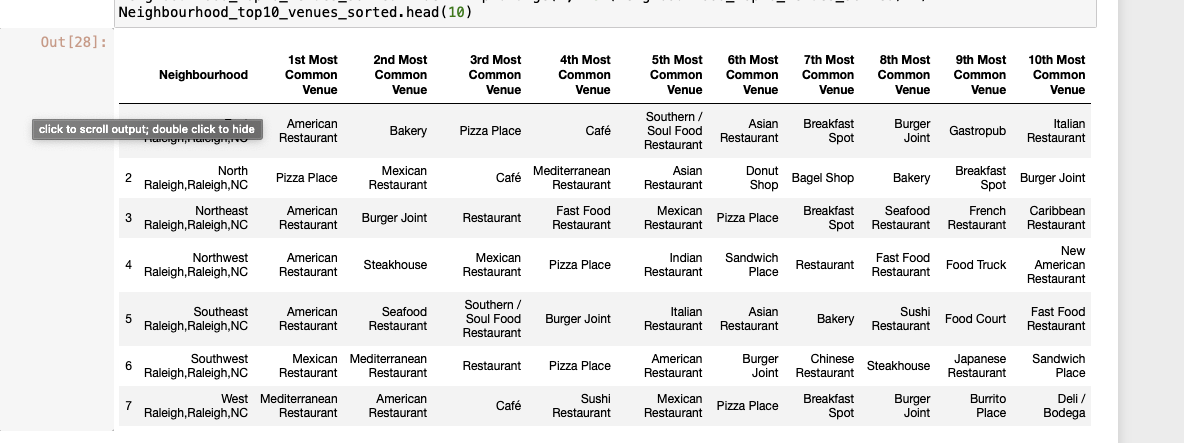
Now using Folium we will visualize the restaurants in the various Raleigh neighbourhoods





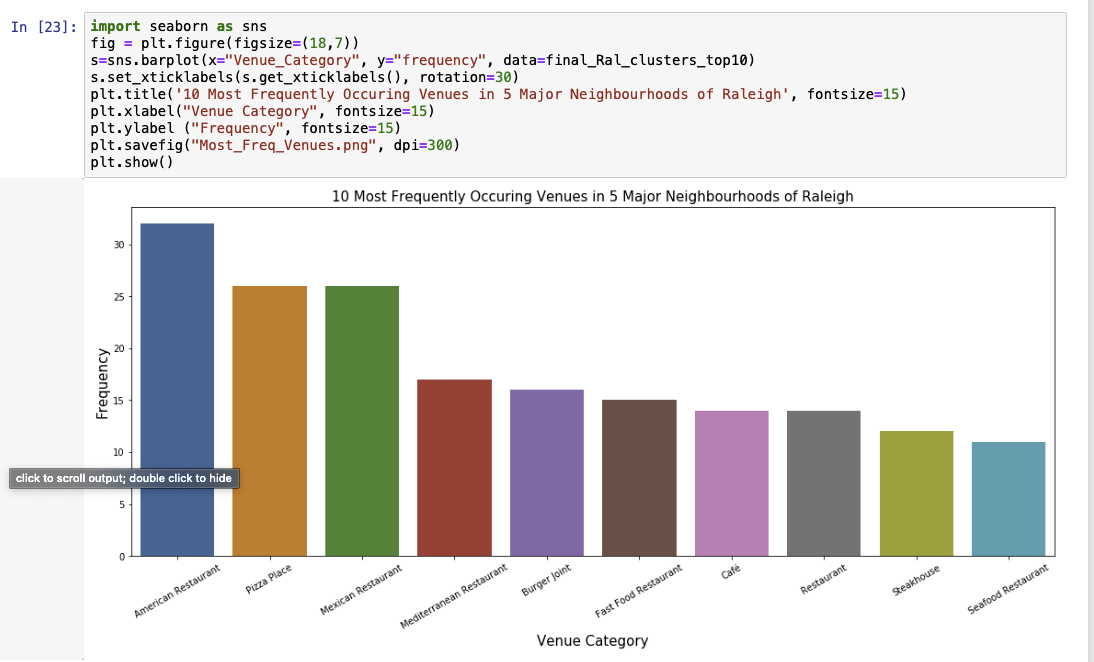
We use one hot encoder to perform “binarization” of the venue category and include it as a feature to train the model.We finally obtain then the top 5 restaurants in each neighbourhood by frequency



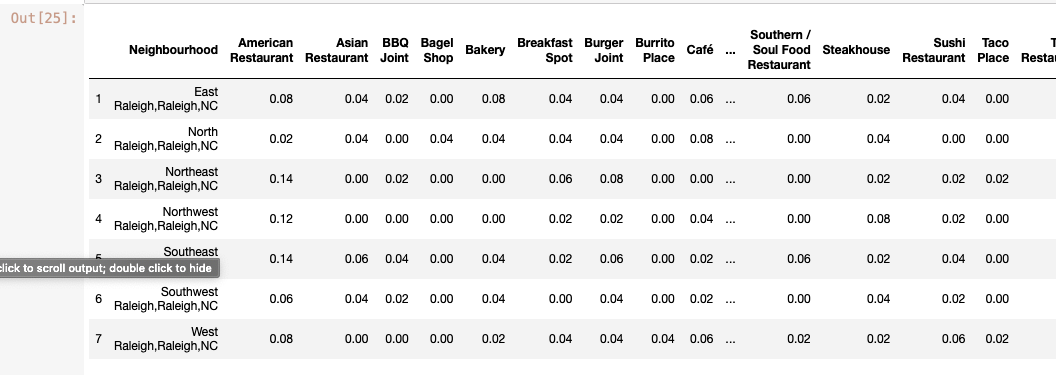


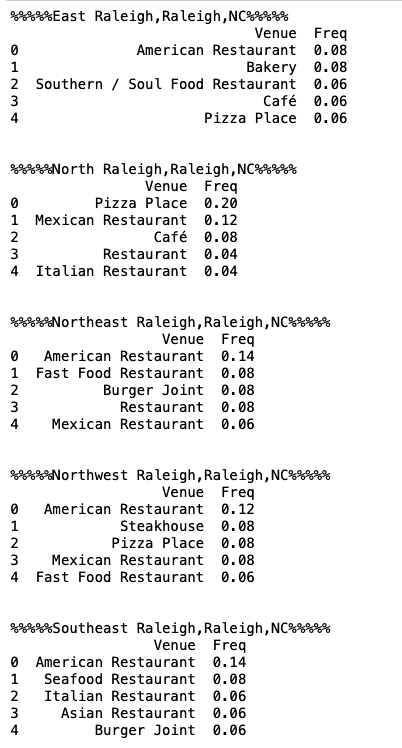
Finally the 10 most common venues across each Neighbourhood is listed and we will merge this dataframe with our kluster labels for our final dataset



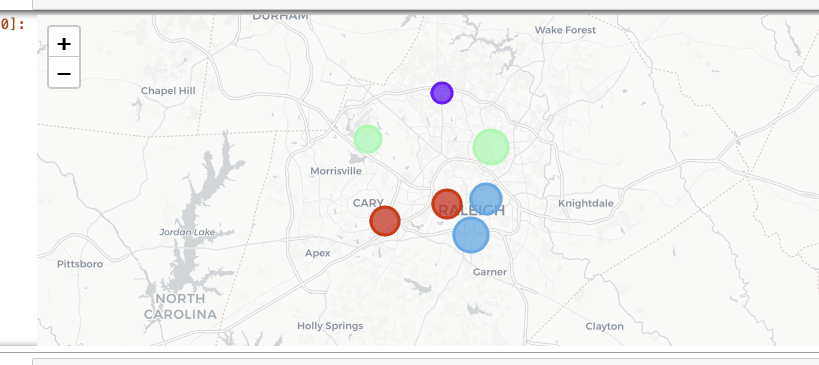


We will cluster these 7 neighbourhoods based on the venue categories and use K-Means clustering. So our expectation would be based on the similarities of venue categories, these districts will be clustered.





We will finally plot the clusters in a leaflet map using Folium library as below



**Observations**

From the various exploratory data analysis and clustering operations performed on our dataset ,cluster 1 has the lowest number of neighbourhoods (North Raleigh) which incidentally has the lowest concentration of restaurants .This could mean a potential location to open a restaurant in this area.The other insight that we could gather is that southern Raleigh seem to be oversaturated with lot of restaurants and our advice to new potential business owners is to avoid this area if you plan to open American or Mexican restaurants.This however is a small study based on the learnings from the course.In the future we can expand our analysis by getting trending venues,restaurant ratings information to dissect the data further to provide any new businesses revealing insights on the type(breakfast,sit-down,fast food) ,cuisine(American,Indian,Mexican..) ,the ideal price point that will make it a successful venture