**Battle of the Neighbourhoods Data Capstone Project**

The purpose of this project is to demonstrate the skills acquired throughout the IBM Data Science course (hosted by Coursera). The task assigned in this final project is to fabricate a problem that could theoretically be solved by using special data; more specifically, a problem that a real entity would need to approach a data scientist to solve.

**Introduction/Business Problem**

I have been tasked by a restaurant chain with locations in many boroughs of New York to find an optimal location for a new venue in southern Manhattan. The restaurants in this chain primarily serve pizza and so the institution would like recommendations on where to place this new venue based on density of competitor pizza places in the area and the ratings of these other venues too. The company’s brand itself is very well known in New York and has recently acquired the funds to set up this new location in the hopes that they can expand their business into Manhattan with more locations in the north to come down the line as well. I intend to analyse this data by clustering it and finding a position in which there are few competitors, but also where the nearest pizza places have a low rating (so that they are not as difficult to compete with).

**Data Collection and Usage**

We will utilise the Foursquare API to acquire our location data on the other pizza places in Manhattan, we shall then cluster these venues and decide where the optimal position for a new venue could be. To decide this optimal location, we will assess the positions and ratings of each existing venue and choose a spot where there is a small number of competitors with low ratings as requested by the client. Because of this, we will be using the ‘search’ and ‘trending’ endpoints to acquire this data from foursquare; in addition, we will also be using the K-means clustering algorithm to classify our data.

The venues themselves will be local pizza places such as “Prince Street Pizza” and “Joe’s Pizza”, we will only be using venues in the southern areas of Manhattan for this as it is the region in which the restaurant chain wishes to set up the new location. Our center point to retrieve this data will be “The Sheen Center for Thought & Culture”, which is a catholic-affiliated performing arts complex with 2 theatres, rehearsal studios & an art gallery. From this location, we will set a large radius for pizza places in Manhattan and call the data with the Foursquare API. The data will be extracted from Foursquare and then cleaned/formatted using pandas in a data frame.

**Methodology and Analysis**

In this section, we will go into detail on how the task was executed, what techniques were implemented, and why these methods were used:

**Stage 1: Data Collection and Cleaning**

Utilizing the Foursquare API allowed us to extract data with the ‘search’ and ‘trending’ endpoints into Python. Two data frames were created from these endpoints using pandas and all unnecessary data was extracted, below is an example of the ‘search’ data. This meant that we could find the positions of all the nearby locations for clustering later on:



Overall, there were 29 pizza places in southern Manhattan, the field names are described as follows:

* **Name:** The publicly established venue name.
* **Lat:** The latitude of the venue.
* **Lng:** The longitude of the venue.
* **Distance:** Distance in meters from our center point that we searched from.

The data frame produced by the ‘trending’ endpoint gave us a similar table but with the amount of foot-traffic each venue received in the past month. This allowed us to see which venues were popular in recent history.

**Why Foursquare, Python and Pandas?**

The decision to use Foursquare and pandas to collect and format this data was due to the quality and up-to-date data provided by the Foursquare API and the simple fact that Python is a language with lots of functionality while remaining very high level relative to its counterparts. The choice of package for data wrangling was pandas simply because of how easy to use and efficient the corresponding functions for use in data frame manipulation are.

**Stage 2: Clustering**

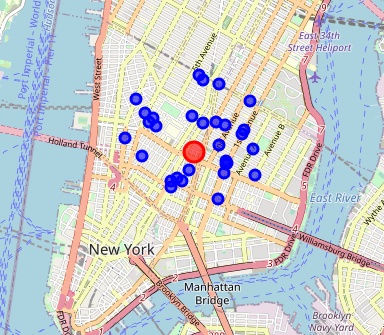
The next stage in the analysis was to employ some sort of machine learning technique to classify our venues so that we can distinguish areas of higher density from areas of lower density. For this endeavour, the k-means algorithm was eventually settled on. The optimal number of clusters was found using the ‘elbow method’, this value was 3. Thus, we had all the pieces to analyse the ‘search’ data frame in full. The data was to be clustered using its foot traffic, thus we can visualize the positions and the popularity of each venue simultaneously. To do this, all that was required was to merge our data frames together so that the positions of each venue corresponded to their popularity.

**Why k-means?**

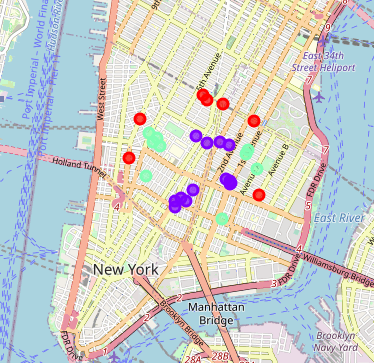
This was an unsupervised machine learning problem, considering we have few points to work with and we are using numerical data to classify points into categorical values, a strong choice was to utilize the k-means algorithm. In addition to this, k-means works well with geographical data in problems like this, and so it seemed to be the most optimal.

**Stage 3: Visualization**

After formatting the special data for each competing venue and clustering them together, we could then plot a map of them to visualize it:

**Before Clustering**

The red marker is our center point and the blue markers are our competing venues. From the plot, we can see that many of the venues are closer to the center of the island, there is a lot of space for a location on the outskirts of Manhattan. However, it is unwise to make conclusions without considering all the variables; thus, to simply claim a good position for our new location should be on the outer edges of the island is hasty.

**After Clustering**

The colour that a venue is assigned corresponds to its cluster, the data seems to have produced layers, where the outer layer forms the red cluster, purple is in the middle, and green is somewhere in between. Each cluster shows a level of popularity for that set of venues, where red is the lowest, green is in the middle, and purple is the highest. Thus, we can now make our desired conclusions.

**Results and Recommendations**

**Results**

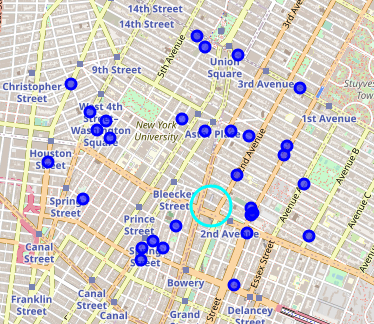
After merging the two data frames and comparing the position of each venue to its popularity, we found that venues closer to the center of Manhattan were more popular than those on the outskirts. We also found that (possibly as a result of this), that there was a higher density of competitors in the center layer of our set of clusters. These results imply that the closer a venue is to the center, the higher chance it has of receiving more attention. This is not surprising as we can expect more people to live near the center of the island than on the outer edges, thus it is easier for any given resident to walk to their nearest pizza place as they are likely to live close by to it.

In contrast to these regions, the opposite trend was portrayed by the venues further away from the center. That is, we saw that the lowest rated venues and least amount of foot traffic was present in the outer areas. Because of this result, we can rectify a misconception that arose in “Stage 3” where it seemed as though a good area to place our new location would be far away form the others on the outer edges because it wouldn’t have any competition. The reason this is false, is because we now know that venues which are situated away from the busy residential areas don’t received much attention or good ratings.

**Recommendations**

**Option 1: The Center and Close to Other Venues**

We know that the center venues receive more attention than the others, thus it may be a good idea to place our new venue nearer the center of the southern region of Manhattan. However, a flaw in this option is that the restaurant would have to compete and perform well in relation to its neighbours. In addition, property value in this area is a lot higher than it is on the outskirts, and so the restaurant would have higher a pressure to make a return on this investment compared to if we placed it on the outer areas. The specific area to place this location that this option refers to is seen below:



**Option 2: The Center but Further from Other Venues**

We have established a strong negative connection to placing our venues on the outer edges, thus our second option is similar to the first one, but with a slight twist. Here, we wish to place our venue in the center, but find the exact area where it will be as far from other exist competitors as possible. Thus, we are in an area of high traffic, but we have as little competition as the space we are working with allows. The downside to this option is that once again the venue will be expensive to set up initially and it would still have to perform well; as well as this, it would not be as close to the most popular venues as it could be (like in the first option), thus the time it would take for the brand to be recognized in Manhattan would be a little longer. The area that the venue would be placed in this option is shown below:

