Introduction

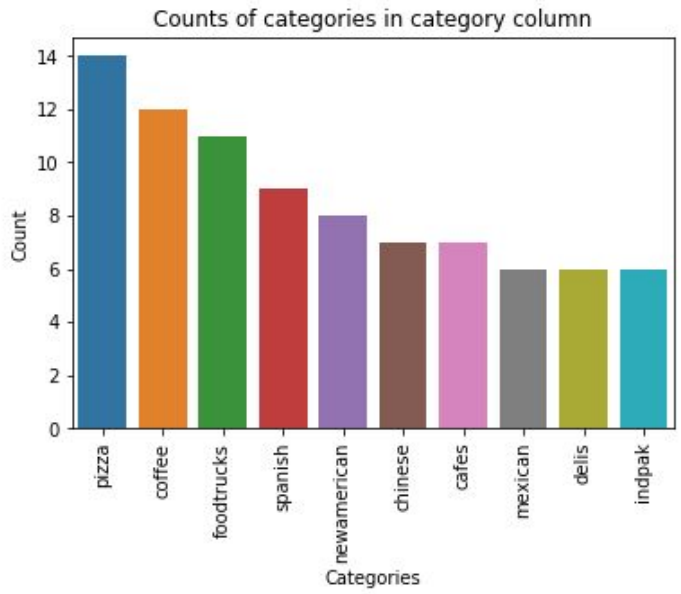
This report's purpose is to analyze and make suggestions on what is the best type and location to put a restaurant in the area of Newark, NJ. Newark, NJ is a city that has a population of 285,154 people, where roughly 50% of the population are african american and 36% have hispanic origins. Of the 36%, most of these people fall in the category of Puerto Rican and South American descent. (US Census 2017) This report will take an in-depth look at segmentation of restaurants in Newark, NJ in order to determine the outcome of what restaurant would be best suited for a place in Newark, NJ.

Data

The data that was collected for this report was allocated from the Yelp API's businesses search. The columns in the data correspond to important characteristics of each restaurant. The columns used are, Name, Reviews, Rating, Latitude, Longitude, and the category alias or Type. The type of restaurant was one hot encoded for purposes of segmentation. Descriptive statistics of each of the columns are listed below.

	reviews	rating	long	lat
count	163.000000	163.000000	163.000000	163.000000
mean	70.055215	3.628834	-74.169887	40.735573
std	103.997819	0.752750	0.004388	0.004281
min	1.000000	1.500000	-74.179360	40.726841
25%	10.000000	3.000000	-74.172988	40.732056
50%	26.000000	3.500000	-74.170701	40.735347
75%	75.000000	4.000000	-74.166612	40.738695
max	737.000000	5.000000	-74.160670	40.744943

Countplot provided for different types of establishments in Newark NJ.



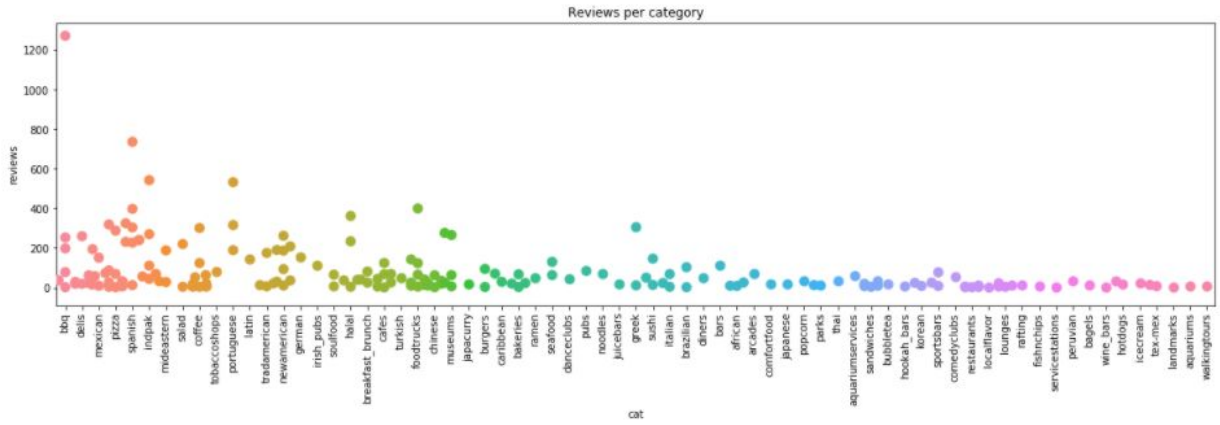
The categorical data will be one-hot-encoded into numerical features and fed into a Kmeans clustering algorithm using sklearn. The algorithm will provide clusters that will give insight into the highest rated restaurants as well as restaurants that have foot traffic (a high number of reviews)

Methodology

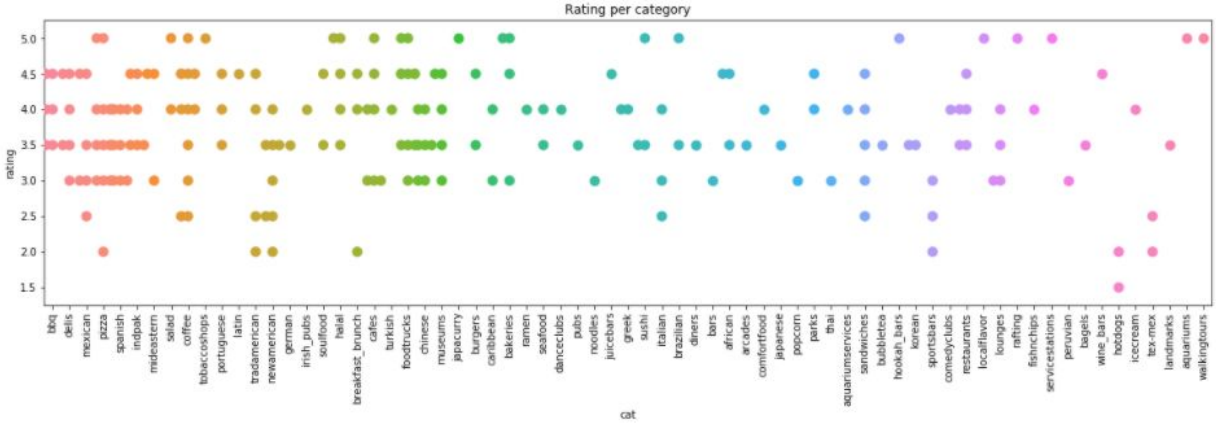
Exploratory Data Analysis

When solving this issue we have taken data from the Foursquare API and analyzed it with a Kmeans clustering algorithm in order to get insights into what makes up a good restaurant in regards to “foot traffic” and rating. Below is an analysis of trends in the data that was collected from the API.

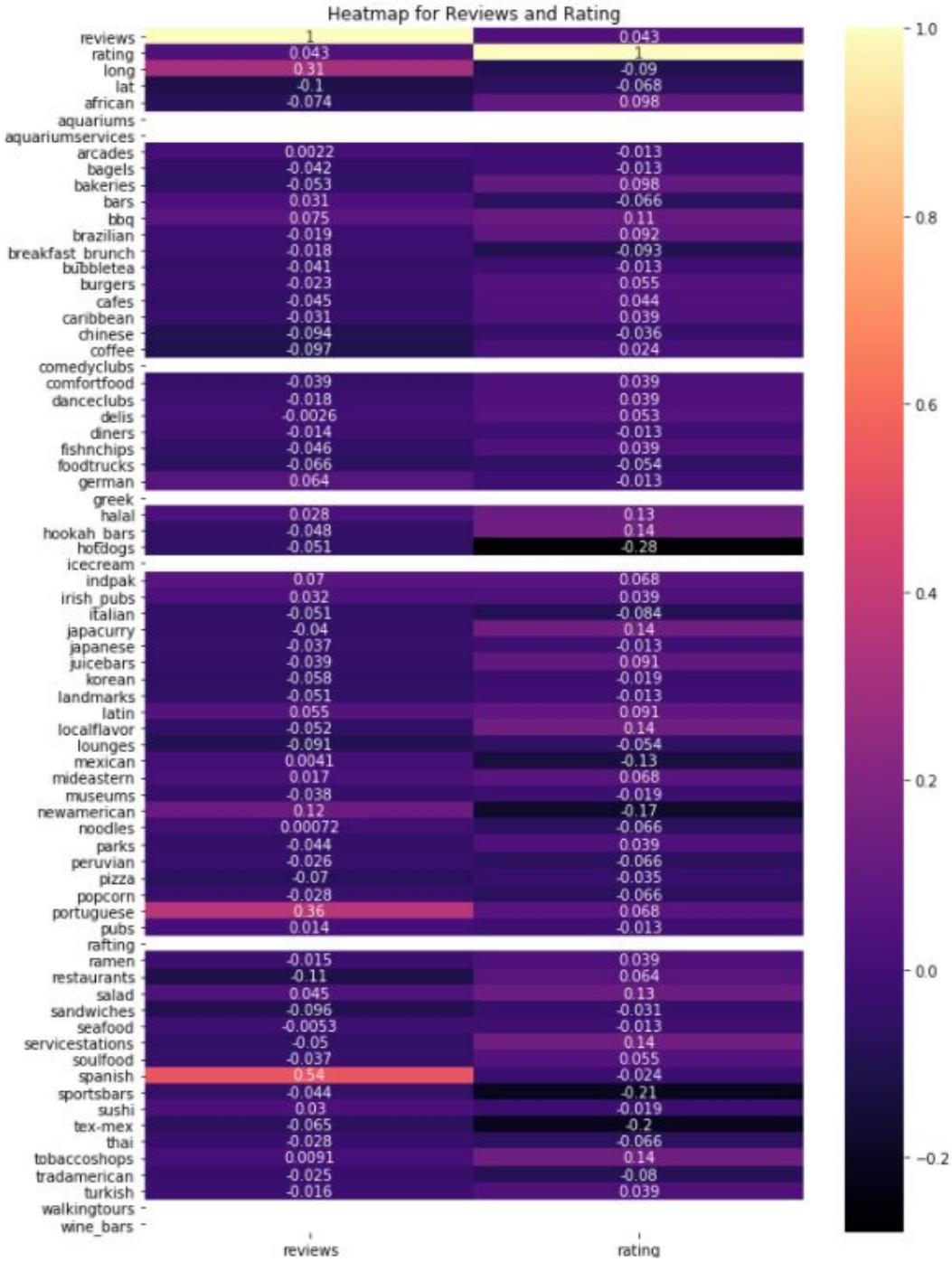
There are no visible favorites when looking at only reviews when looking at groups of restaurants. The graph below shows that some single restaurants have a high amount of ratings but there is no particular category that has a cluster above 300. However this graph does show many restaurants that are the only one of their kind that have very low reviews. This could be either because they have low foot traffic or they are brand new to the area.



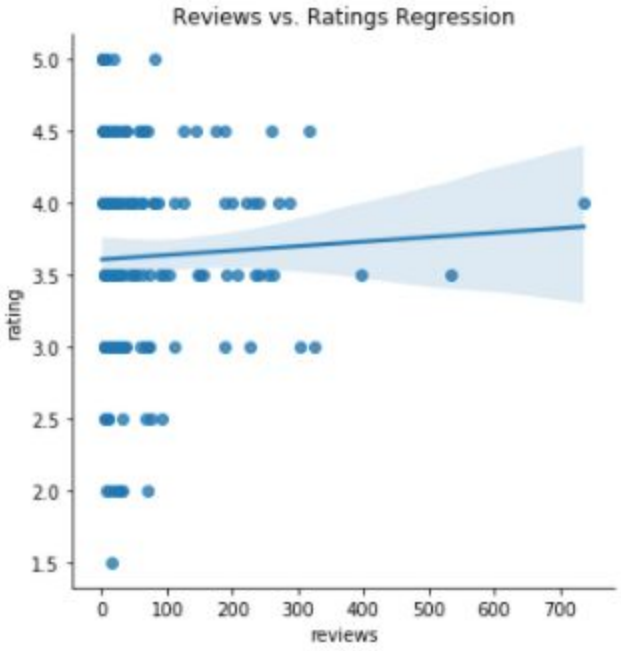
The same categories when analyzed for rating show a similar trend of many ratings in the 3.5 to 4.5 area while some singular restaurants fall either very low(<3) or very high ratings(5). These ratings can be affected by how long the store has been open. If someone writes a review for a new restaurant that is less than three it could be impacted far more than an established restaurant that has many high ratings to start.



The heatmap below shows correlations between all the variables as well as the reviews and ratings columns. This shows that there are high correlations between reviews and the three categories, Spanish, Portuguese and longitude. People with Hispanic heritage make up 36.4% of Newark's population so it is a good indicator that we should see some correlation between these two values. According to a 2011-2015 American Community survey 7.68% of Newark's population speaks Portuguese. The correlation comes as no surprise since there is a high percentage of people with Portuguese descent that live in Newark, NJ.

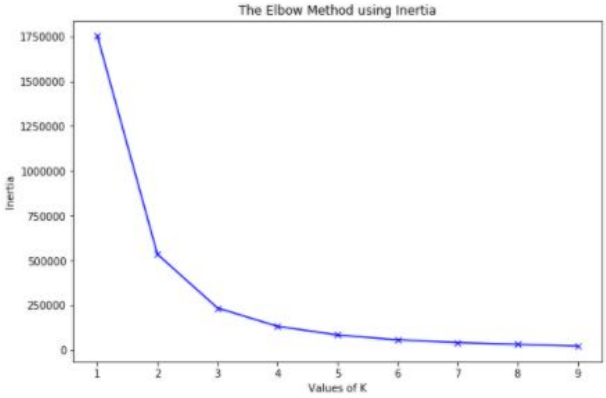


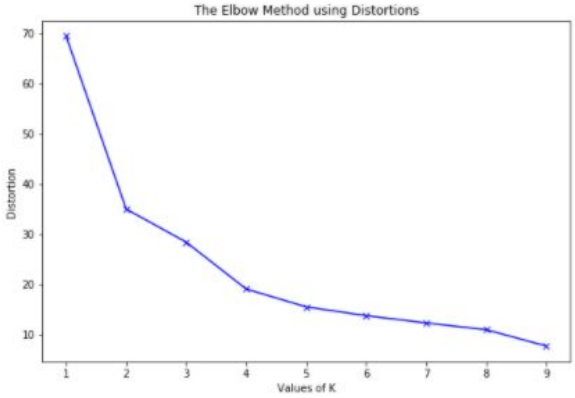
The graph below shows a trendline of reviews and ratings. There seems to be no trend for restaurants that have many reviews being higher than other restaurants that have fewer reviews.



Modeling

In order to model this data we will be using the K-means clustering algorithm. This algorithm will use 74 dimensional space using the variables longitude, latitude, rating, reviews, and the category data converted into one hot encoded columns. In order to choose a good value for K, or how many clusters we want, we need to view the inertia and distortions of the data points. According to the graphs below two is one of the best choices for the number of clusters.



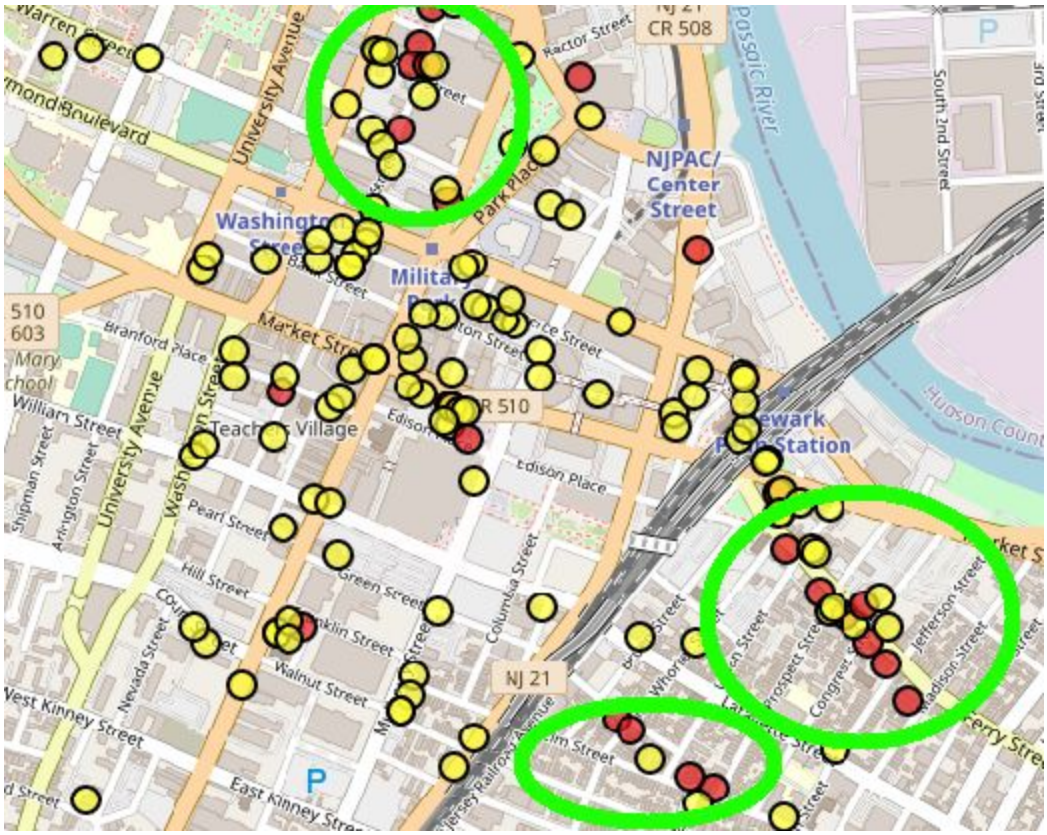


Results

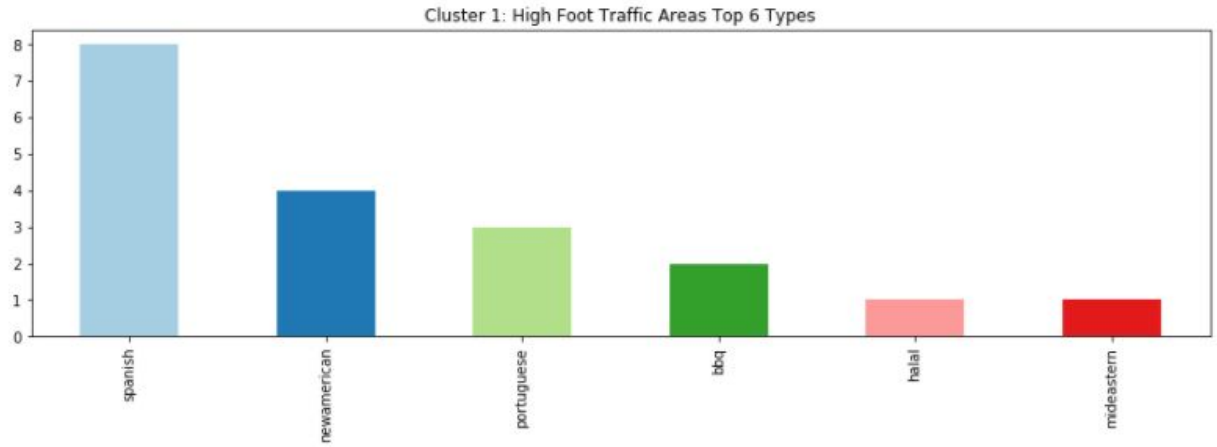
In breaking down the clusters we can see that there are 139 restaurants in cluster zero and 24 restaurants in cluster one. There is also a difference in rating from cluster zero to cluster one of 0.143. The most noticeable difference between the two clusters is the reviews left for cluster zero and cluster one. On average there is a difference of 244 reviews between cluster zero restaurants and cluster one restaurants. Upon visual inspection of the data there are three sections that stand out for cluster one. Restaurants in cluster one seem to be clustered around Ferry Street, Elm Street, and Bleeker Street.

cluster	
0	139
1	24

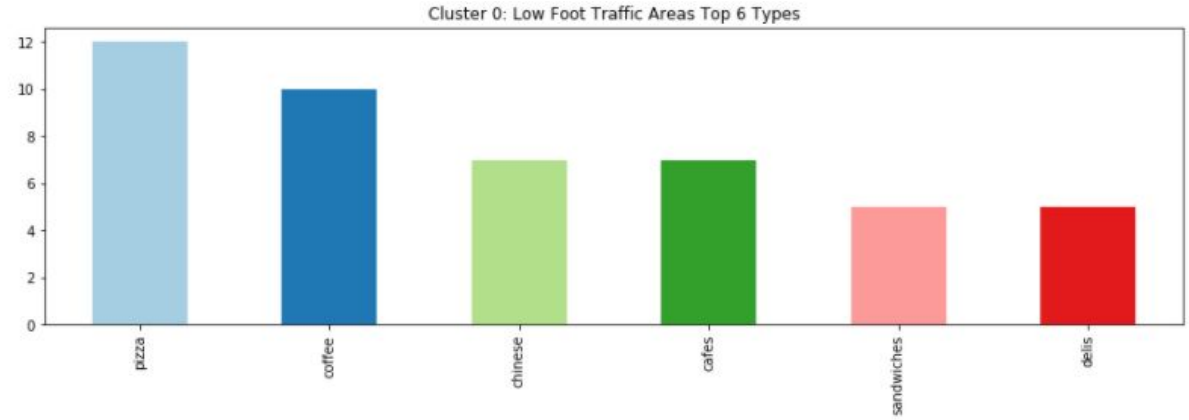
cluster	reviews	rating
0	34.136691	3.607914
1	278.083333	3.750000



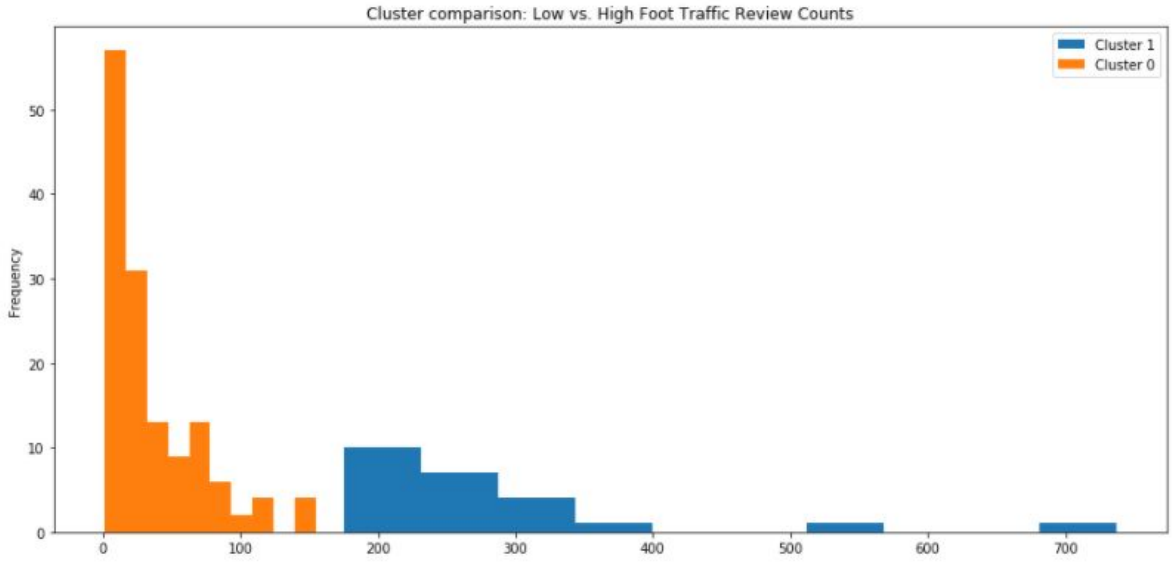
In cluster one there were a number of Spanish, American, and Portuguese restaurants according to the graph below. This matches up well with the heatmap that was shown in the analysis phase of this project. These clusters of cluster one restaurants are labeled as high foot traffic areas because of the high amount of ratings left at these types of restaurants.



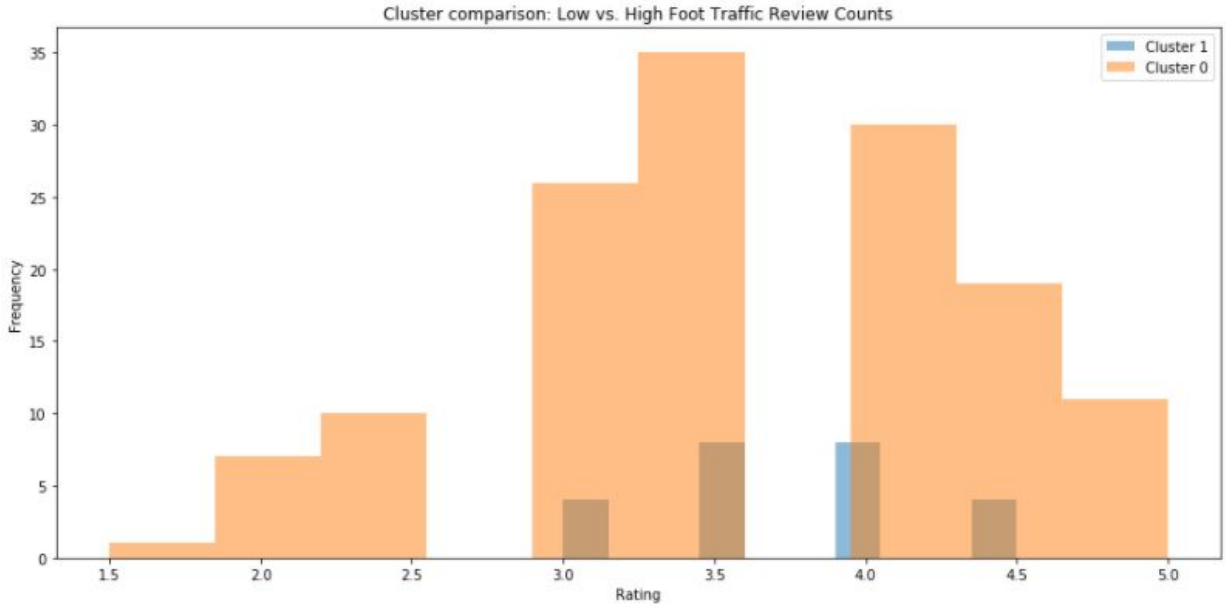
In cluster zero there were many instances of pizza, coffee, and chinese restaurants. These were located in areas where cluster zero restaurants were high frequency. These areas are classified as low foot traffic areas because of the low number of reviews left at these types of restaurants.



In reviewing the distributions of each cluster for their reviews it is clear that cluster one has many more reviews than cluster zero. They are both right skewed distributions with a mean closer to the start of each range. This means that cluster one should be the model choice for any new restaurant that wants to be created in Newark NJ.



In reviewing the distributions for each cluster in the category of rating it is important to note that cluster zero is the only cluster that has ratings below three and five exactly. Cluster one looks particularly normal with a mean in between 3.5 and 4.0. Cluster zero is a slightly left skewed distribution with a mean that looks close to 3.5 or slightly above. Given this, we can say that modeling a new restaurant from cluster one will possibly generate more of a conservative approach when it comes to rating of a restaurant.



Discussion

From the analysis above the recommendations for a new restaurant in the city of Newark, New Jersey should be one of the categories that is most popular in cluster one. These restaurants

are Spanish, American, Portuguese, BBQ, Halal, and Mid-Eastern. The placement of this restaurant should also be placed in a high foot traffic area. The areas where foot traffic is the highest according to the analysis is on or around Ferry Street, Elm Street, and Bleeker Street in Newark, NJ. The investor should not create a restaurant in cluster zero areas and should refrain from starting a restaurant that specializes in pizza, coffee, Chinese, and deli sandwiches.

Conclusion

In this report we aimed to conclude which type of restaurant and its placement would be appropriate for Newark, NJ. Using a K-means clustering algorithm we were able to conclude that the restaurant should be one of the following types:

- Spanish
- American
- Portuguese
- BBQ
- Halal
- Mid-Eastern

These restaurants have the most reviews on average in the area of Newark, NJ. Upon visual inspection of these restaurants in the high foot traffic cluster it was discovered that the placement of the restaurant should be somewhere on Ferry Street, Elm Street, or Bleeker Street. This report is only data collected from foursquare and could benefit from more data from other sources like traffic data and crime rates in specific parts of the city.