Coursera Capstone

Project - The Battle of Neighborhoods

INTRODUCTION

In this final project we need to solve the issue of the Italian restaurant owner Gianni, that wants to expand his business in New York and open a second restaurant, the only issue is that he has no idea on where to open it.

Gianni has a medium-high price level restaurant in the Manhattan area, it is a pretty successful restaurant with a FourSquare review score of 9.6 out of 10, due to his perfect mix between traditional cooking and modern style.

Using the power of data we will help Gianni to find the best neighborhood to open the new restaurant.

DATA USED

To be able to achieve our final result we needed to use data from multiple sources:

❏ The geocoder library to separate in neighborhoods the city of New York.

❏ The FourSquare data to be able to retrieve the italian restaurant venues and info about those venues in a radius of a specific neighborhood.

METHODOLOGY - INTRO

The approach that we take to solve this issue was based on two main factors:

❏ The Tier of the restaurants.
❏ The Review score of the restaurants.

The Tier value is given from Foursquare to make a separation of the different restaurants based on the restaurant price range.It goes from Tier 1(the cheapest) to Tier 4(the most expensive).

Since what we want is a mid-high range restaurant we collected data regarding Tier 3 and Tier 4 restaurants.

METHODOLOGY - DATA ACQUISITION

To be able to collect and separate the data based on our needs we called the FourSquare api's using the python language to collect all the info regarding Italian restaurant venues in a radius of 1 km.

We later sorted this data inside data frames using the *pandas* library.

Once all the data(name, location, id) regarding Italian restaurants located in a 1 km radius from each NY city neighborhood was collected, we were able to get also,with another call to FourSquare api's, the Tier of the restaurant and the Review score associated to it.

We were able to select only the restaurant that we were interested in (Tier 3 and above) and to prepare our final dataframe to clusterize them using a k-means Machine Learning algorithm

METHODOLOGY - DATA ANALYSIS

We grouped all our Italian restaurant venues by neighborhood and by taking the mean of Tier and Review of each neighborhood

Using the k-means Machine Learning algorithm, thanks to the *sklearn* python library, we were able to clusterize all the restaurants under 4 different classes.

In the next slide there will be a visual representation, achieved by using the *folium* library, of the Italian Restaurant venues clusterized in 4 different classes based on the Tier and Review

ITALIAN RESTAURANTS CLUSTERIZED MAP
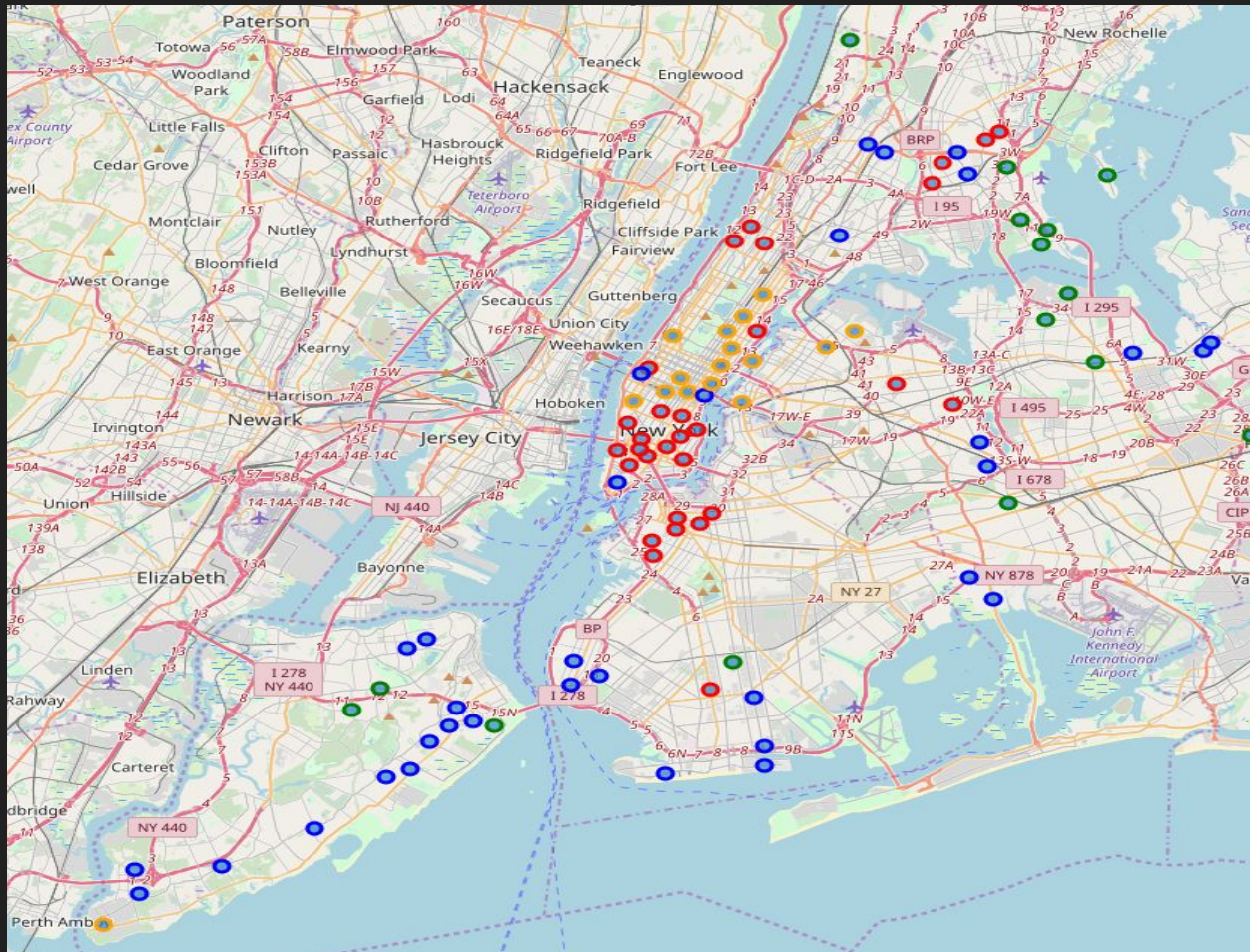
🔵 cluster 0

🔴 cluster 1

🟢 cluster 2

🟡 cluster 3

The k-means algorithm is generating this 4 clusters based on the similarity between the average restaurants Tiers and reviews in the Neighborhoods.
We will move our attention on the cluster 2 the one that is collecting the venues with high Tier and low Reviews.
We will now restrict the results on the cluster 2 neighborhoods and make a further analysis to restrict the result to find the best candidates

ITALIAN RESTAURANTS FINAL MAP

Here we can see an additional separation to highlight the neighborhood in the cluster 0 which have very high price tier and very low review score.

From this visualization we can see that there are two of them that are standing out from the others:
the Arrochar(1) and Richmond Hill(2) neighborhoods.

RESULTS AND DISCUSSION

Thanks to this final clusterization we can clearly see that the neighborhoods of Arrochar and Richmond Hill are the better candidates for the location of our new italian restaurant since the gap between Tier and Review score is the highest.

Having this two neighborhood very highly priced Italian restaurants with very low review, they should welcome the very high standards of Gianni`s restaurant.

Based on this result the better option for the location of the next Gianni`s restaurant is the Richmond Hill neighborhood in Queens area since as we can see in the first map the neighborhoods around Arrochar will offer many more high quality alternatives in comparison of the neighborhoods around Richmond Hill

CONCLUSION

Based on the collected data from FourSquare and on our analysis of them we are confident to suggest the neighborhood of Richmond Hill as a location for Gianni`s next restaurant due to two main reason that we discovered analyzing the data:

- ❏ the lack of a high tier italian restaurant with good reviews but only of some high tiers italian restaurant with low ones
- ❏ the shortage of high tier italian restaurants with good reviews in the nearby neighborhoods

In this link you can find the jupyter notebook containing all the analysis performed.

Notebook