DATA SCIENCE PROJECT

ANALYSIS OF NEIGHBORHOODS IN QUEENS, NEW YORK

AVANTIKA KOTHANDARAMAN

TO DETERMINE THE APPROPRIATE LOCATION FOR A RESTAURANT SET-UP

1. Introduction

1.1 BACKGROUND

New York is one of the most popular destinations in the world. There are five boroughs in NY, namely, Manhattan, Brooklyn, Bronx, Queens and Staten Island. Each borough has got its own unique feel and experience. For this project, I have chosen Queens as the borough to explore, analyse and assess the neighborhoods to find out. Queens is the second-most populated borough in NY, after Brooklyn. Queens is also home to the John F Kennedy International Airport and the La Guardia Airport. The cost of living in a place like Queens is very expensive and so while making an investment in projects like these, one needs to be well-informed of as many aspects as possible. Knowledge about the neighborhood in which said restaurant is to be set up by finding out the kind of competition from other similar restaurants and the overall safety of the neighborhood can be acquired by analyzing the data about the location and the crime rates.

In this project, I aim to analyse the neighbourhoods of Queens, NY, to determine the most suitable location for a restaurant that is bound to ensure maximum return of investment and is also safe. This project will benefit aspiring entrepreneurs with ideas of starting a restaurant in Queens. It will be able to solve their dilemma pertaining to the location aspect of their restaurant.

1.2 PROBLEM

Using the location data that we have acquired about Queens, NY, we can analyse neighbourhoods using different Machine Learning algorithms and Data Science techniques to visualise the neighbourhoods, find the list of areas with heavy population and diversity, find other restaurants in close proximity who are potential competitors, etc.,

By using these algorithms, we can arrive at a conclusion and finalise on the most appropriate location(s).

1.3 TARGET AUDIENCE

This project will be of use to aspiring entrepreneurs who wish to set-up restaurants in Queens, that will ensure good profits and popularity. Using this project, one can determine the safest neighborhood and one that can minimise competition and thrive in terms of profits. Other data science enthusiasts with a passion for analysis might also find this project interesting.

2. Data acquisition and cleaning

2.1 DATA SOURCES

For this project, Foursquare data will be used. Foursquare is a company providing location data of different places in the world. By creating an account, we have access to their API with our unique credentials. Using those credentials, we can call their API for accessing their data at any point of time during analysis. This data gives us all possible known locations of establishments and small businesses in Queens, New York. Using that data, we can proceed with our analysis.

For analysing the safety of Queens, we have to acquire the dataset recording the daily crimes in the borough. A free dataset for the whole of New York from the web was acquired from NYC crime.

2.2 DATA CLEANING

This is the process in which we perform the pre-processing of the datasets.

We have to download the datasets and load them onto the notebook we are working on. The New York dataset was loaded as a .json file. After observing the, we see that all the important information in the file is in the 'features' section. We extract that section alone and put it into a dataframe. For location analysis, we need only the names of the boroughs, neighbourhoods and the latitude and longitude coordinates. We loop along the obtained data to fill in the needed content in our dataframe. We perform further cleaning by isolating the content of Queens alone, removing the contents of the other boroughs.

For the crime dataset, we first check its info to see how many fields have null values in them. We drop those rows as they are both unnecessary and will hinder our analysis too. We then change the datatype format of the columns containing date values to the standard Pandas datetime type.

For analysing the neighborhood data, we have acquired the needed data from Foursquare. It gives us information about all the nearby venues and their coordinates.

All the data in both datasets have been pre-processed and cleaned and now are ready for analysis.

2.3 FEATURE SELECTION

We can see after forming the dataframe and cleaning that we have 2088 rows and 7 columns. We now choose the features we really want and those that we do not.

From the Foursquare data, we decide to keep the names of the neighborhood, its latitude and longitude coordinates, along with the names of the venue and its corresponding latitude and longitude coordinates, and the venue category. This will help us accurately locate every single venue in every single neighborhood of Queens. It will also help us accurately locate specific locations in the database, as per our choices, provided we know its latitude and longitude coordinates.

From our city dataset, we included the names of borough, neighborhood, latitude and longitude details. We will then join these two datasets, to perform clustering analysis to identify similar restaurants, i.e, ones that are similar in terms of locality and cuisine.

In the crime dataset, for our analysis, we acquired a dataset with a large number of missing/null values. This is not good, as it will result in improper analysis. It is a wise process to drop the fields containing such null values. We also drop the fields involving time as we do not require the time of crime occurrence in our analysis of data. We also drop the column involving the description of the location of the crime in question. We are removing that column as we do not need a description of the location. All we need are the latitude and longitude coordinates to locate the exact spot. We change the format of the date type fields to datetime type.

For example, the Queens dataset can be used to choose a particular location/neighborhood, say, Elmhurst, and we can find out any number of nearby venues to that location that we want, by specifying the radius of search and the limit, i.e., the number of venues. Say we want to search for 70 nearby venues to Elmhurst for a radius of 500, we can do so merely by calling our Foursquare API using our credentials and by specifying the parameters. We can even repeat the same process of finding 70 locations for instance, through a radius of 100, just by entering a 'for' loop in our code. Since, our focus here is only on restaurants, we can apply another filter to give us the data of nearby restaurants in 70 nearby locations alone. Thus forms our required dataset.

Now, after pre-processing, cleaning and feature selection of the different datasets, we are ready to begin our analysis.

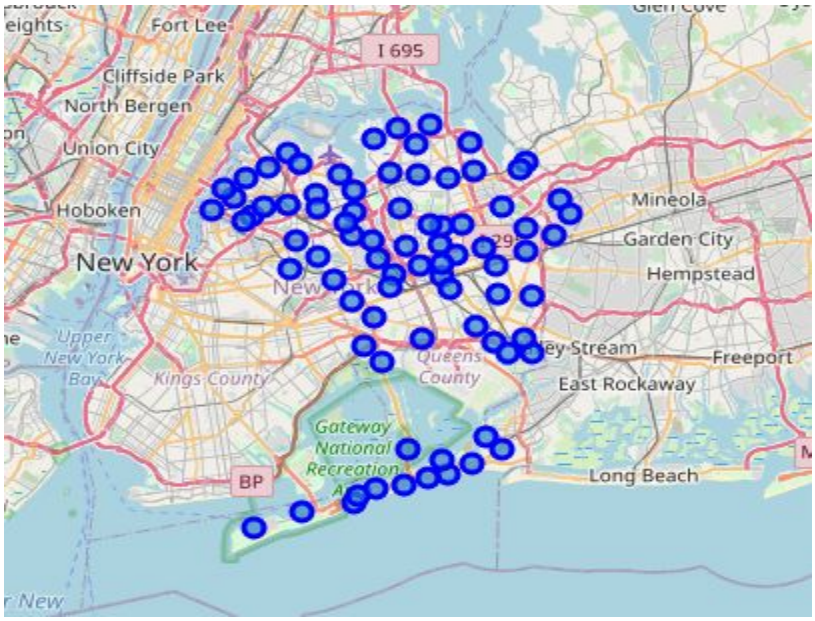
3. Exploratory data analysis

Our first focus in this analysis is to find out and investigate the neighboring areas and filter them to give us the list of restaurants to assess competitiveness. We make use of certain machine learning models for this purpose.

Before beginning our major analysis, first, it is important to verify that the data we collected has that of all five boroughs. We use the .shape feature to verify said concern. Now, we filter the dataset, and extract the data of Queens alone from the whole New York borough, neighborhood, latitude and longitude dataset. Now we can analyse our target region further.

As we are mostly interested in location data, we will acquire the location coordinates of New York as a whole, and Queens alone later. Proper visualization of our regions of interest will enable us to understand the locations and get used to it. Using the acquired data about the location, we use Folium maps to plot our regions and observe. This usage of Folium is possible only after installing the necessary packages and importing the folium library. It is highly recommended.

Visualization of the Queens borough using Folium:



This visualization enables us to see the exact location, border and position of the Queens borough in New York. At the end of this analysis, it is one or more of the above plotted neighborhoods that we will conclude our analysis with. We can notice that Queens is not just one joint area, but rather has an extension at the bottom. We will definitely have to include that in our consideration too.

Acquiring data about nearby venues from Queens:

Now, this is the part of our analysis where we make use of Foursquare. Foursquare is a location data providing site. Just by entering the coordinates of the area in question, we can acquire details about nearby locations, namely, restaurants, bars, gyms, libraries, absolutely any venue!.

Now by entering the coordinates of Queens, and since I wish to see the locations of 100 nearby venues from my coordinates for a radius of 500. Using my credentials from Foursqaure, I made a call to their API to acquire the needed data, which was formatted into a dataframe.

By analysing the acquired dataframe, I inferred that we were able to acquire 269 nearby venues and their location coordinates.

Extracting the restaurants data from the dataframe

For our analysis, we do not need data about anything but restaurants. So will acquire the data about restaurants alone for our analysis.

These restaurants will now be grouped by the neighborhood, and will be classified in their dataframe as per the cuisine style.

Our data is now ready for some modelling.

4. Modeling

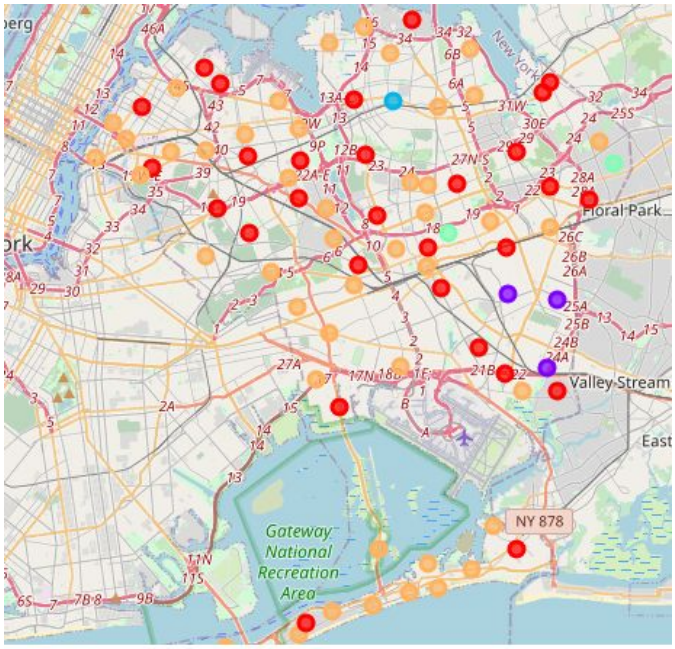
Clustering machine learning algorithm for restaurants type classification

Now, we have to analyse the location data that we have formed a dataframe with. For modeling the data, we are not going to perform any predictions. So a regression or classification type of modeling will prove inefficient for our analysis. Our main aim is to group all similar restaurants together. By grouping similar restaurants together, we can find out what neighborhoods in Queens serve the same cuisine. Suppose, a particular neighborhood has a high number of Italian restaurants, setting up another Italian restaurant in the same neighborhood would result in competition and our restaurant may not be able to live up to its expectations. We will use k-means clustering to group similar restaurants together as one, and then we can see which neighborhoods they come under to assess cuisine popularity in that neighborhood.

I then performed a count process to see how many of those venues in specific are there. Now the next step involves the filtering of the venue data to give us information about the restaurants alone, since that is our area of interest.

I have grouped the top 5 restaurants in every neighborhood into one dataframe, as it is impractical to analyse all the numerous restaurants in Queens.

Before proceeding further with our k-means clustering, one-hot encoding of data is essential to properly plot the locations on the map. After this, we can use the .mean method to find the frequency of occurrence of each restaurant to find how popular each cuisine style is in that neighborhood. For example, by doing this, we can find out the most common cuisine styles in a neighborhood. After applying the algorithm, we get the following graph, with similar restaurant cuisines grouped as one cluster.



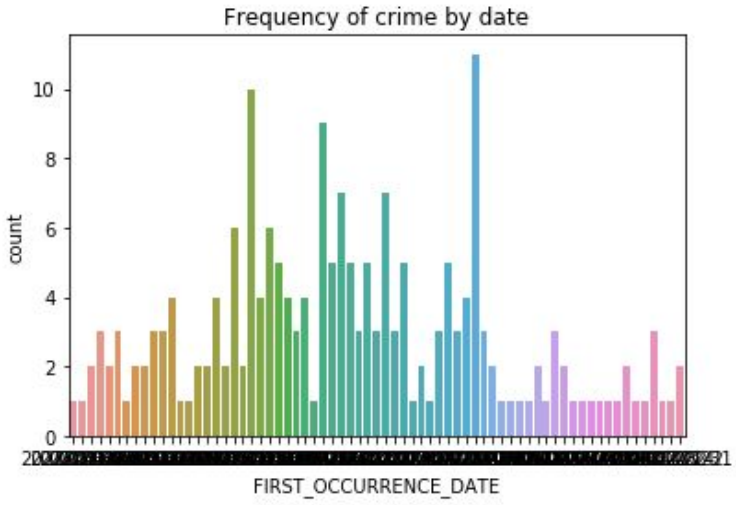
I have chosen to divide the overall count of restaurants in every neighborhood into 5 divisions, and they will be sorted by their similarity in cuisines.

For example, we will be able to visualise the majority of Asian restaurants in one clusters amongst other cuisines, which are comparatively less popular. We can then see which neighborhood it is that falls under the category with a high number of South-East Asian restaurants, and we can arrive at deductions.

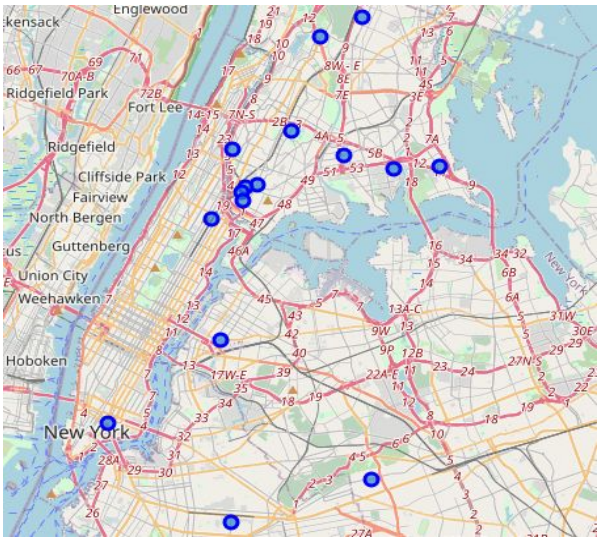
Exploring the crime dataset

In this particular dataset, I have explored the general and basic information about crime in Queens. First, the fields containing null values were dropped from the dataset to make our analysis easier. Unnecessary columns were also dropped.

Now, an analysis was done to determine the time of the year that saw maximum crimes. The observed data was observed using the matplotlib library of plots.



We can now see that maximum crimes were reported during the middle of the year. Another more important analysis was done on the crime dataset: finding locations that saw comparatively higher crime rates. For this purpose too, I have used Folium maps. Since it is impractical to assess every single crime type such as arson, burglary, theft, assault, harassment, and so on, I have performed my analysis only for cases of assault to narrow down our crime search. I am making the assumption that it is enough to assess the most common type of crime alone, to assess localities in general.



We can now infer that the Northern parts of Queens have a higher crime rate for just one type of crime than the other parts. We can now infer the Southern regions are comparatively safer than the northern ones.

5. Results:

We can now understand that each cluster has been grouped together on the basis of the popularity and the cuisine served in it. We were also able to see the most common restaurants in each cluster. The crime dataset tells us that the Northern part of Queens has a higher crime rate when compared to the South.

6. Discussion section- Observations and Inferences:

Analysis of each cluster

It is now clear that clusters 0 and 4 are highly dense clusters. This is due to the fact that these clusters belong to extremely popular cuisines thereby resulting in a greater number of establishments catering to the public.

The most popular cuisines in cluster 0 are Italian and South-East Asian. With this information we can infer that there is a high Italian/South-East Asian population in neighbourhoods of this cluster. However, at the same time, we can also infer that if we set up similar restaurants in these neighbourhoods, we will be exposed to greater competition, and will really have to up our game.

The most popular cuisines in cluster 1 are Caribbean and South-East Asian. The Caribbean cuisine does not appear to be very popular as we have very few restaurants in that cuisine. We can also infer that the neighbourhoods in this cluster have a relatively high Caribbean population.

The most popular cuisines in cluster 2 are Korean and Japanese which can be grouped under South-East Asian cuisine. Again, we can see that the Korean cuisine is not very popular, and that can be attributed to a relatively lower Korean population.

The most popular cuisines in cluster 3 are Indian and other vegetarian restaurants in general. Again, this is not a very high number of occurence, but we can infer that there is a relatively higher Indian population in the neighborhoods of this particular cluster.

Cluster 4 is a very dense cluster. It comprises a combination of Mexican restaurants, fast food joints and also has a considerable number of South-East Asian restaurants. Fast food joints include barbeque joints, burger joints, and smaller entities like food trucks.

Analysis of crime data

We can infer from the crime data that most crimes occur in the Northern region of Queens and that the Southern regions are relatively safer. I made this inference on the assumption that it was enough to assess just one crime type with a maximum occurrence, that is assault.

7. Conclusions and Suggestions:

Just by seeing the density of clusters 0 and 4, we can conclude that the most popular cuisines in Queens are Italian and South-East Asian. By this, we can also infer an aspect of

the demographic situation at Queens, which is that Queens has a relatively higher Italian and Asian population. Should an aspiring owner choose to stick to the norm and open a restaurant of one of these cuisines, he/she would fit into the demographic, but would face great competition from other similar establishments.

If the desired cuisine is Caribbean, then neighborhoods in cluster 1 would prove profitable. Those neighbourhoods are St Albans, Laurelton and Cambria Heights.

If the desired cuisine is Korean, then the neighborhood for that would be Murray Hill.

The neighborhood called Floral Park is famous for its Indian and vegetarian restaurants.

If one wishes to set up a restaurant of Italian/South-east Asian cuisine, any neighborhood is preferable. Under confusion, one can decide using the crime data. Clearly the southern regions have a lesser crime rate. In that aspect, St Albans and Laurelton are the clear winners.

The aspiring owner can choose a cuisine which already has popularity in a neighborhood, but he/she should be wary of the competition it poses. My suggestion would be to choose a safe neighborhood with decent popularity of preferred cuisine and that should be the best way to go!

****