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 INTEREST

This project will be of use to aspiring entrepreneurs who wish to set-up restaurants in Queens, that will ensure good profits and popularity. 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.

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