Coursera Capstone

IBM Applied Data Science Capstone

New Shopping Mall Opening in Brooklyn, United States of America

Done by: Hasan Ali Mohamed Taresh

April, 2020

Section 1: Introduction

Nowadays, shopping malls are considered one of the main targets to be visited, despite your needs. You can do grocery shopping, dine at restaurants, watch movies, shop for clothes from various fashion outlets, and much more. In 2017, shopping malls accounted for 8% of retailing space in the United States. As a result, there are many shopping malls in the United States, and many more are being built these days. Opening shopping malls allows property developers to earn consistent rental income. Hence, opening a new shopping mall requires serious consideration and deep study, putting in mind certain factors. One of the main factors to consider is the location of the shopping mall, which could determine whether the mall will be a success or not.

1.1 Business Problem

The objective of this project is to analyze and choose the best locations in Brooklyn (a borough of New York City, USA) to open a new shopping mall. With the power of Data Science methodologies, along with machine learning techniques such as clustering, this project aims to provide answers for the following business problem: If a property developer is planning to open a new shopping mall in Brooklyn, in New York City, USA, where is the best recommended area that should be chosen to build the mall on?

1.2 Targeted Audience

This project is certainly useful to property developers along with investors that are interested in opening or investing in a new shopping mall in Brooklyn.

Section 2: Data Description

In order to solve the problem described in the previous chapter, we need to gather the following data:

- A list of neighborhoods in Brooklyn.
- The latitude and longitude coordinates of each neighborhood in Brooklyn.
- Venue data, which is related to shopping malls.

2.1 Data Sources

1. Wikipedia:

We are going to obtain a list of neighborhoods of Brooklyn from the following page: https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Brooklyn. After extracting the list (using web scraping techniques), we are going to obtain the geographical coordinates for each neighborhood using Python Geocoder package.

2. Foursquare API:

We will use Foursquare API to retrieve venue details for each neighborhood extracted from the Wikipedia page. It is good to mention here that we are going to focus on the "Shopping Mall" category from various categories that Foursquare will provide for us, in order to solve the problem described.

Section 3: Methodology

First of all, we need to retrieve a list of neighborhoods in Brooklyn. Luckily, the list we seek is accessible already in the Wikipedia page mentioned in section 2.1. We will do web scraping utilizing Python requests along with BeautifulSoap packages to extract the list. In any case, the list will contain only names of neighborhoods available in Brooklyn. So, we need to get the geographical coordinates of each neighborhood in the form of latitude and longitude in order to use Foursquare API. To do so, we will use the Geocoder package that will convert addresses from names-only to coordinates. Afterwards, we are going to populate the data into a pandas Data Frame, and then visualize the neighborhoods list in map using Folium package.

The next step is to use Foursquare API to obtain the top 100^{th} venues that are within a radius of 10000 meters. In order to make API calls with Foursquare, we have already registered in Foursquare website as a developer account, to obtain the Foursquare ID and secret key. We make API calls to Foursquare passing the ID and secret key available for us, passing by the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON

format and we will extract the venue name, venue category, venue latitude and longitude. With the data available, we can check how many venues were returned for each neighborhood and examine how many unique categories can be created from all the returned venues. Then, we will analyze each neighborhood by grouping the rows using neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Shopping Plaza” data, we will filter the “Shopping Plaza” as venue category for the neighborhoods.

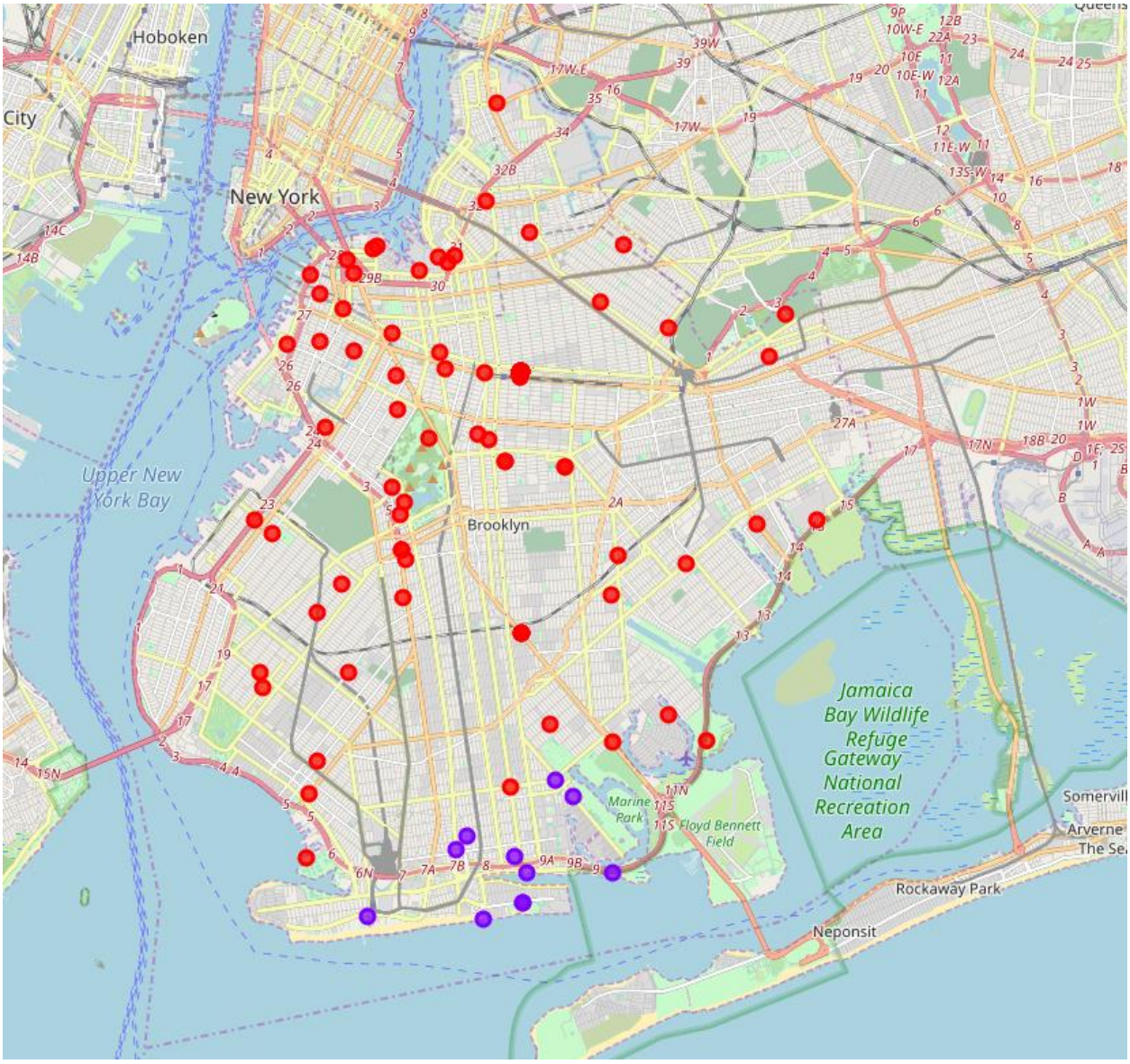
Last but not least, we are going to perform clustering on the data by using the k-means clustering technique. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 2 clusters based on their frequency of occurrence for “Shopping Plaza” category. The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Section 4: Results

The results obtained from the k-means clustering shows that we are able to categorize the list of neighborhoods we have into 2 clusters (as a maximum), based on the frequency of occurrence for “Shopping Plaza” category, as follows:

1. Cluster 0: Neighborhoods with no existence of shopping malls.
2. Cluster 1: Neighborhoods with low to moderate number of shopping malls.

We visualized the results of the clustering in the map below, with cluster 0 in red, and cluster 1 in purple.



Section 5: Discussion

After observation of data, we noticed the following:

1. Most of shopping malls are concentrated on cluster 1, which is in the lower side of Brooklyn close to the beach.
2. On the other hands, the upper part of Brooklyn has no shopping malls in the neighborhoods.
3. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls.
4. Meanwhile, shopping malls in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of shopping malls.

Therefore, as an outcome of this project, we recommend property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 0, with little to no competition at all. Lastly, property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of shopping malls and suffering from intense competition.

Section 6: Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 2 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders, i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section (1.1), the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.