Coursera Capstone
IBM Applied Data Science Capstone

Finding the Optimal Location for a New Museum in Brooklyn

By Abbie Maemoto

July 2020

Introduction

New York is a melting pot of culture, art, and history. This New York reputation is in large part due to the thousands of historic museums located all across New York. These museums are hubs of tourism, attracting people from all walks of life. In these museums, tourists lose track of time, immersing themselves in the worlds of art, fashion, performance, and science. The world-wide attraction to these museums is an important factor in New York's economy. A popular museum can help sustain nearby hotel and restaurant industries. As most museums are non-profits, the popularity of a museum is largely dictated by the number of members a museum can obtain. Therefore, the location of the museum is of utmost importance to its success. Property developers, in the interest of maintaining museum popularity, must be careful in deciding the location of the museum and thus, in this study, we will look at the optimal locations for the opening of New York's newest museum.

Business Problem

The objective of this study is to identify the best locations in the city of Brooklyn, New York to open a new museum. Through implementing data science extraction and machine learning analysis methods, we aim to answer the business question: If a property developer is looking to build a new museum in Brooklyn, NY, where is the optimal location to open the museum?

Target Audience

The target audience of this study are investors and property developers looking to plan the construction and opening of a brand-new museum in New York City. Finding the optimal location for the museum will help obtain valuable investors for the project. As New York is already home to the second-highest number of museums (approx. 2500) in the United States, finding the best possible location will be critical for being able to compete with the existing museums.

Data

To tackle this project, we will need the following information:

a) A list of the neighborhoods in Brooklyn, NY. This will help confine the focus of our study specifically to the region of Brooklyn which is the one of the most visited cities in New York.
b) The coordinates (latitude and longitude) of the neighborhoods found in part a
c) Surrounding venue data that can contribute to a further understanding of whether the neighborhood could foster the development of a new museum. This will be used for clustering later in the project.

The list of neighborhoods was found at https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Brooklyn which consists of 40 neighborhoods. We will utilize data mining techniques to extract the data from this website, supplemented by Python and beautifulsoup packages. Then, using the geocoder package, we will obtain the coordinates of the various neighborhoods. Next, we will use

FourSquare to analyze the venue data for all of the neighborhoods. This project will consist of various data analysis tools, including web scraping, implementing API through FourSquare, data wrangling, K-means clustering, and visualization with the Folium package.

Methodology

First, we found a list of neighborhoods in Brooklyn on a Wikipedia page at https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Brooklyn . We then extracted this neighborhood data with the BeautifulSoup and the pandas python packages, resulting in a list of neighborhood names. In order to get the geographical coordinates of each neighborhood, we used the Geocoder package to translate coordinates into longitude-latitude form. After adding these coordinates to the data frame containing the neighborhood data, we visualized the location of the neighborhoods using the Folium package which allowed us to ensure that our coordinates were indeed neighborhoods in Brooklyn, NY. Then we used the FourSquare API package to get a better understanding of the surrounding neighborhoods, extracting the top 100 venues that were within a 2000 mile radius. In order

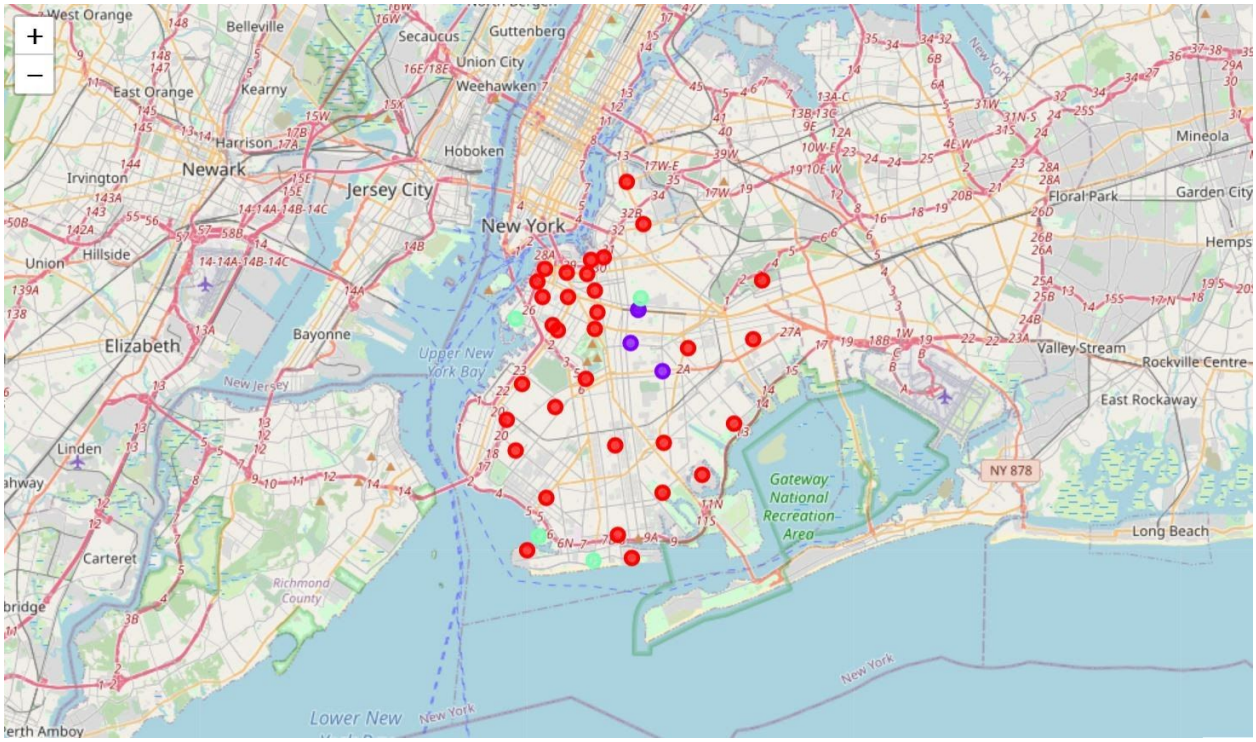
to do this, we registered for a FourSquare Developer account and obtained a Client ID and secret key in order to use the API. Looping through the Python coordinates, FourSquare returned surrounding venue data in the form of a JSON file; we then extracted data like the venue name, venue category, and venue latitude and longitude. We then analyzed the number of venues returned for each neighborhood and isolated the unique categories from all of the Brooklyn neighborhoods, prepping the data for clustering. Since we analyzed “museums”, we filtered the key word “Museum” as the venue category for the neighborhoods. Finally, we performed clustering on the data by k-means clustering. Based on the frequency of the keyword “Museum”, we chose to cluster the neighborhoods into 3 clusters which allowed us to identify relative concentrations of museums in each cluster. This data will help us ultimately decide which neighborhoods are suitable for the development of a new museum.

Results

The results from the k-means clustering helped us visualize the following patterns in the relative concentrations of museums within the three clusters:

- Cluster 0 had virtually no museums
- Cluster 1 had a high concentration of museums.
- Cluster 2 had a moderate concentration of museums.

On the map below, cluster 0 is marked by the red markers, cluster 1 is marked by the purple markers, and cluster 2 is marked by the green markers.



Discussion

As visualized in the “Results” section, most of the museums are concentrated in the central neighborhoods with the highest number in cluster 1. You can also see that there is a moderate concentration of museums along the coastal neighborhoods as depicted by the green markers. On the other hand, there are virtually no museums in the north-western neighborhoods, making these neighborhoods prime for the development of a new museum. This is beneficial as the new museum will be facing very little competition with other museums and can perhaps bring about new industries in the neighborhood if the museum thrives.

Conclusion

In this project, we tackled the business problem of finding the optimal location of a brand-new museum in Brooklyn, New York. Through data mining and machine learning algorithms, we were able to cluster Brooklyn neighborhoods into 3 groups and analyze which cluster could foster the most optimal conditions for the new museum. To definitively answer our original business question, the best place for opening a new museum would be in the north-western cities of Brooklyn where there is little to no competition with other museums, allowing the new museum to capitalize upon all the surrounding industries.