Wine Business Concept & Venues Analysis of Brooklyn

Luis Henriques

May 2021

1. Introduction

1.1 Background

New York City also known as, "The Big Apple" or "The City that Never Sleeps", is the most populous city in the United States.

With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2), New York City is also the most densely populated major city in the United States and also one of the world's most populous megacities.

New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. New York City is composed of five boroughs, each of which is a county of the State of New York. The five boroughs are Brooklyn, Queens, Manhattan, the Bronx and Staten Island.

Brooklyn, the most populous borough, is known for its cultural, social, and ethnic diversity, an independent art scene, distinct neighborhoods, and a distinctive architectural heritage. Downtown Brooklyn is the largest central core neighborhood in the Outer Boroughs.

The borough has a long beachfront shoreline including Coney Island, established in the 1870s, as one of the earliest amusement grounds in the U.S. Marine Park and Prospect Park are the two largest parks in Brooklyn. The last decade, Brooklyn has evolved into a thriving hub of entrepreneurship and high technology startup firms and of postmodern art and design and one of the most desire borough to start a new business.

On the other side of the Atlantic Ocean, we can find a small country in size, but a huge country in terms of culture, world history, and a country vast in natural wealth. This country is **Portugal**. Portugal is the westernmost country and one of the eldest nations in Europe. It is comprised of two parts, a continental part, and two autonomous regions: the archipelagos of the Azores and Madeira. For such a small country it is very diverse in geography from north to south and the capital city boasts 300 days of sunshine per year.

Portugal has been Europe's best-kept secret for tourism until recent years when it was awarded the World Travel Award for Europe's Leading Destination. With nearly 1800km of coastline, Portugal is a world-class destination for beach lovers, it has 17 UNESCO World Heritage Sites and is known for its world-class gastronomy including wine. To many wine experts, Portugal is the last frontier of wine in Western Europe; there is still so much to be tasted and explored. For serious wine lovers it's an exciting country to explore since most of the grape varieties are native to the country and not found anywhere else in the world.

Due to its growing popularity, the biggest challenge is finding Portuguese wines outside of Portugal and especially in the USA. One might find a wine or two in a specialty store but you most certainly won't find a dedicated section in your local wine shop.

1.2 Business Description

Recently, I was approached by an entrepreneur who lived in New York for 15 years and in the last 2 years has lived in Portugal. From early on, she fell in love with Portugal and one of the best things the country can offer, Portuguese wine. Based on her recent passion, this entrepreneur decided to start a wine export business from small Portuguese producers to the United States of America directly to the final consumer. After the successful launch of the export business, this entrepreneur saw an opportunity to open a Portuguese wine store in the borough of Brooklyn in order to make the exclusive wines available to the inhabitants of New York.

With my data scientist skills I will try to **find a place to locate the new wine store**, preferably that is **close to Portuguese restaurants** and located in a **neighborhood where the average base salary is high**, due to the exclusivity of the wines, and without any **wine stores in the vicinity**.

2. Data acquisition and cleaning

2.1 Data Sources

Based on the problem at hand, I need to describe what are our main requisites that will impact our decision and present to potential investors:

- The wine store should be located in a neighborhood within one of the **highest household income** / **highest population density**
- If possible, should be **near a Portuguese restaurant(s)** in the vicinity
- The **distance** to other wine stores or wine bars.

For our analysis to be carried out successfully, it is necessary to obtain data from several sources. The data was downloaded from:

- **2014 New York City Neighborhood Names** dataset from Spatial Data Repository of NYU. This dataset has all neighborhood's name and respective coordinates
- **Foursquare API** to verify the location of each venue and respective description (restaurant and bar) in our chosen neighborhood
- **Average Household Income** of Brooklyn borough from a New York Real State website (https://ny.curbed.com/)

2.2 Data Cleaning

Data downloaded or scraped from multiple sources were combined into three different tables and at the end were merged only into one.

All the information regarding the New York neighborhoods locations were in a shapefile format. It was need to extract the information from there and transform into a dataframe. Initially it was possible to verify that the dataset had information regarding the five boroughs: Brooklyn, Queens, Manhattan, Bronx and Staten Island. For our analysis, it was only maintain Brooklyn.

The information regarding the Average Household Income of Brooklyn had some neighborhood names that didn't match with the neighborhood's name with the respective coordinates in our previous dataframe. Since I know that at the end I should merge both dataframe, it was needed to do some modifications: 10 neighborhoods of 65 were dropped because it did not have the respective location and it was necessary to change the name of 7 neighborhoods.

2.3 Feature Selection

After data cleaning, there were 65 samples and 10 features from the neighborhood's location dataframe and 65 samples and 6 features from the Average Household Income dataframe.

From the neighborhood's location dataframe I maintained the following features:

- **Name**: Name of the neighborhood;
- **Borough**: Name of the borough;
- **Long**: Neighborhood's longitude;
- **Lat**: Neighborhood's latitude.

From the Average Household Income dataframe I maintained the following features:

- **Neighborhood**: Name of the neighborhood;
- **Borough**: Name of the borough;
- **Median Household Income:** Median household income refers to the income level earned by a given household where half of the households in the geographic area of interest earn more and half earn less.

After merged the two dataframe we got the following one as we can see in the Figure 1 and also represented in the map (Figure 2).

	Name	Borough	Long	Lat	Median Household Income ($)
0	Bath Beach	Brooklyn	-73.998752	40.599519	55193
1	Bay Ridge	Brooklyn	-74.030621	40.625801	57980
2	Bedford Stuyvesant	Brooklyn	-73.941785	40.687232	37343
3	Bensonhurst	Brooklyn	-73.995180	40.611009	46137
4	Bergen Beach	Brooklyn	-73.898556	40.615150	72158

Figure 1 – Merged Dataframe

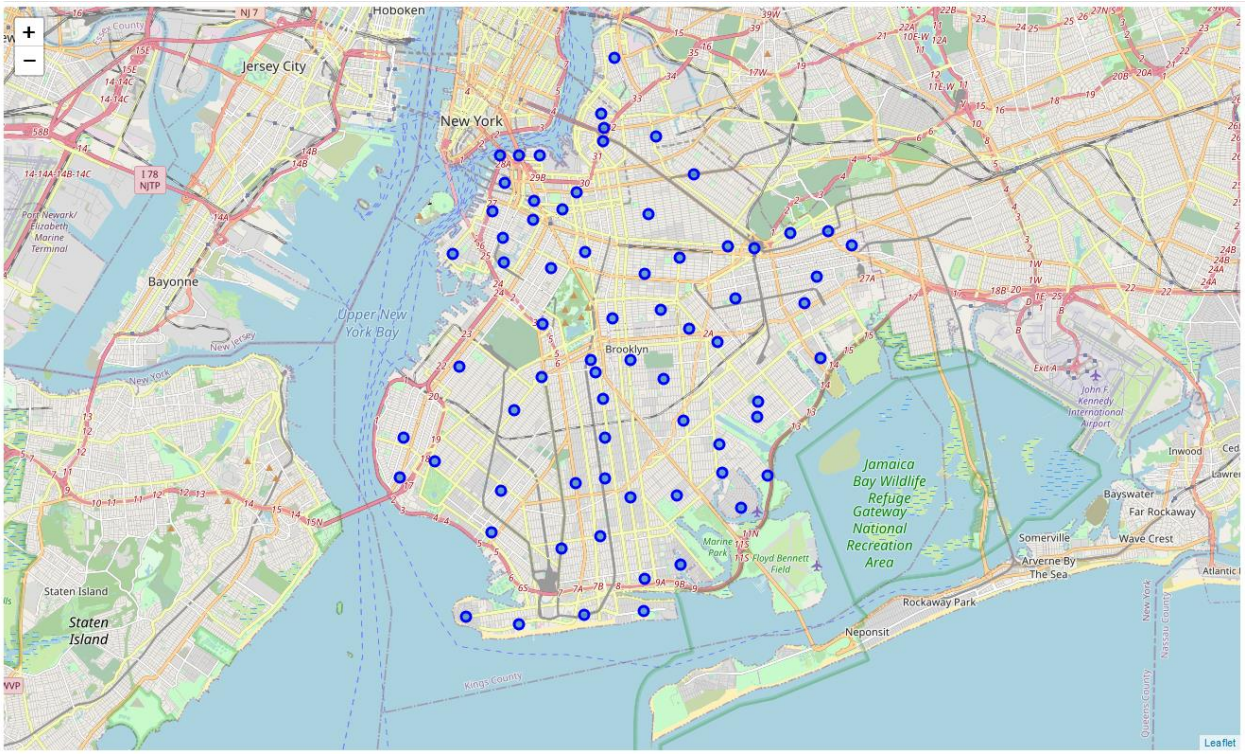


Figure 2 – Neighborhoods representation in a Folium Map

2.4 Foursquare API

In order to find a perfect location for the client Wine Bar we need to check all the local venues 500 meters around of the neighborhood's center. To do that we need to use the Foursquare API. We got 2342 samples (venues). See next figure.

	Neighborhood	Neighborhood Lat.	Neighborhood Long.	Venue	Venue Lat.	Venue Long.	Venue Category
0	Bath Beach	40.599519	-73.998752	Bensonhurst Park	40.597065	-73.998340	Park
1	Bath Beach	40.599519	-73.998752	Bay Parkway Water Front	40.595941	-74.000917	Surf Spot
2	Bath Beach	40.599519	-73.998752	Five Guys	40.595236	-74.000225	Burger Joint
3	Bath Beach	40.599519	-73.998752	Lutzina Bar&Lounge	40.600807	-74.000578	Hookah Bar
4	Bath Beach	40.599519	-73.998752	Pino's Ristorante	40.600955	-74.000806	Italian Restaurant

Figure 3 – Local Venues

I also could observe what the most common venues for each neighborhood are as we can see in Figure 4. It is a good way to do a fast check and see what neighborhoods I can avoid or select.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bath Beach	Chinese Restaurant	Pharmacy	Gas Station	Pizza Place	Fast Food Restaurant
1	Bay Ridge	Spa	Pizza Place	Italian Restaurant	American Restaurant	Pharmacy
2	Bedford Stuyvesant	Coffee Shop	Bar	Playground	Pizza Place	Deli / Bodega
3	Bensonhurst	Chinese Restaurant	Ice Cream Shop	Italian Restaurant	Donut Shop	Sushi Restaurant
4	Bergen Beach	Harbor / Marina	Playground	Baseball Field	Athletics & Sports	Park

Figure 4 – Common venues

4. Methodology

In this project, I will direct my effort on neighborhoods of Brooklyn that have Portuguese restaurants nearby and have a low wine bar or wine store density.

In first step we have collected the required **data: location and type (category) of every venue within 500m from each neighborhood.** I didn't find any Portuguese restaurants. So my main focus it will be in Wine Stores and Wine Bars.

Second step in our analysis will be calculation and exploration of '**Wine bars/stores density**' across different areas of Brooklyn - we will use **heatmaps** to identify a few promising areas.

I will also use K-means for cluster segmentation in order to provide a better insight for our client.

4.1 K-Means

The k-means algorithm searches for a predetermined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:

- The "cluster center" is the arithmetic mean of all the points belonging to the cluster.
- Each point is closer to its own cluster center than to other cluster centers. Those two assumptions are the basis of the k-means model.

In order to identify in how many clusters we can do our segmentation we have to check with the *"elbow method"* and *"silhouette analysis"*. See Figure 5.

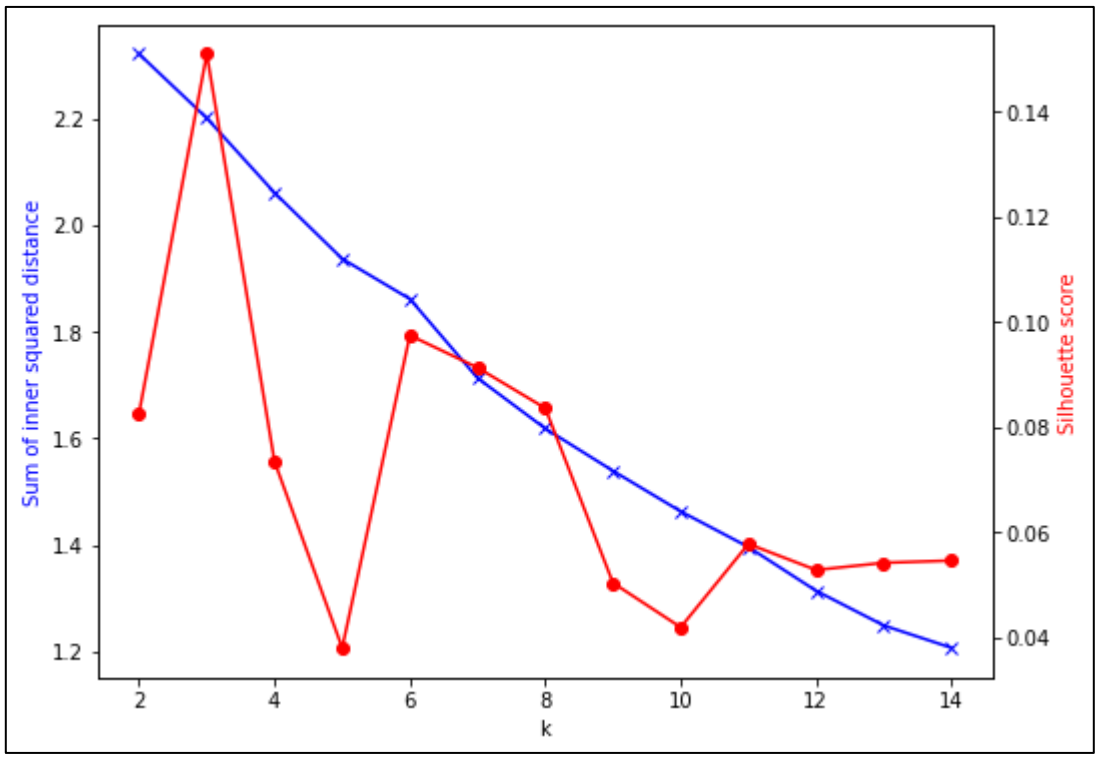
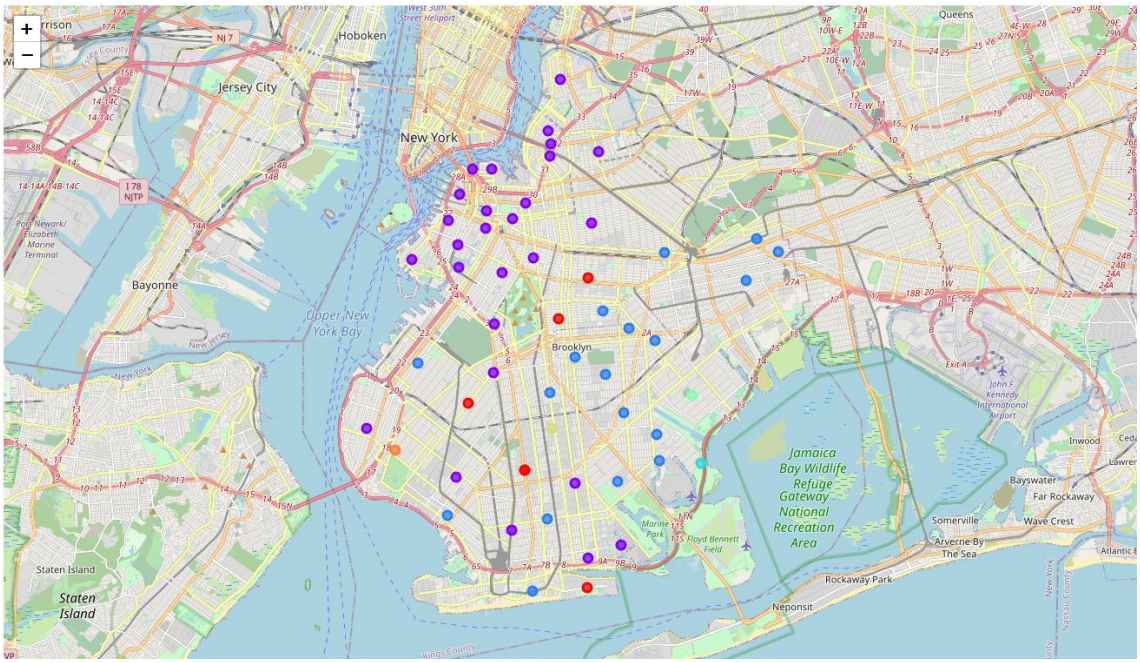
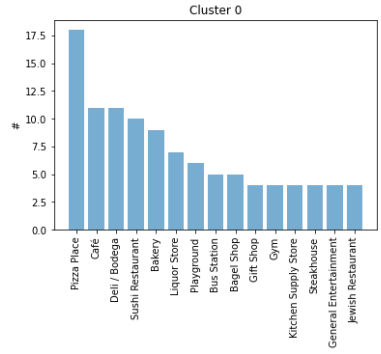


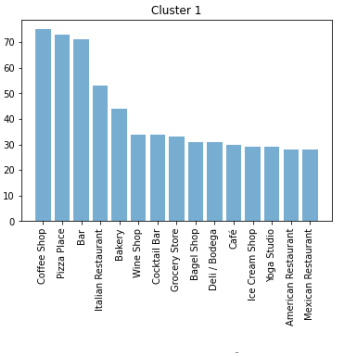
Figure 5 – Elbow method and silhouette score

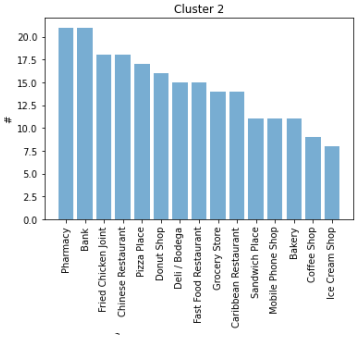
After seeing the result it was decided to choose $k = 7$. Cluster representation:

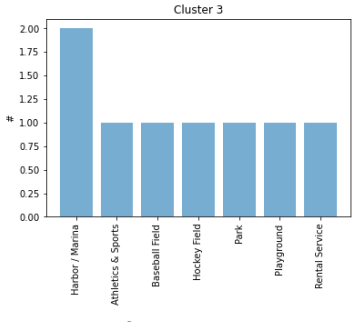


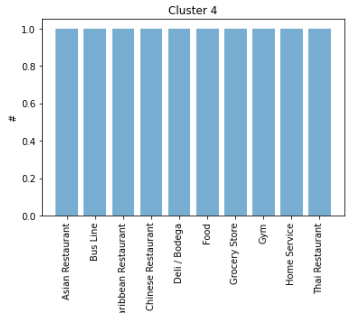
Next, we can check the quantity of the venues for each cluster.

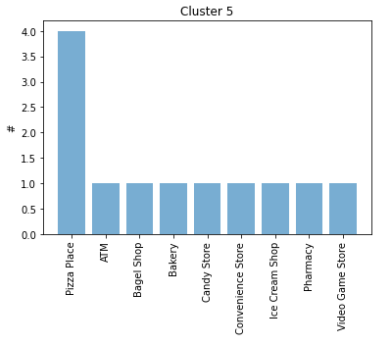












Now we can check which are the neighborhoods with the top 10 household income and check the respective location under a heatmap that represents the wine bar and wine stores density.

	Neighborhood	Borough	Long	Lat	Median Household Income ($)
0	Cobble Hill	Brooklyn	-73.998562	40.687920	105398
1	Brooklyn Heights	Brooklyn	-73.993782	40.695864	105398
2	Gowanus	Brooklyn	-73.994441	40.673932	101784
3	Park Slope	Brooklyn	-73.977050	40.672321	101784
4	Carroll Gardens	Brooklyn	-73.994654	40.680541	85496
5	Red Hook	Brooklyn	-74.012759	40.676254	85496
6	Dumbo	Brooklyn	-73.988753	40.703177	84945
7	Vinegar Hill	Brooklyn	-73.981116	40.703322	84945
8	Downtown	Brooklyn	-73.983463	40.690844	84945
9	Boerum Hill	Brooklyn	-73.983748	40.685683	84945

Figure 6 – Top 10 household income neighborhood

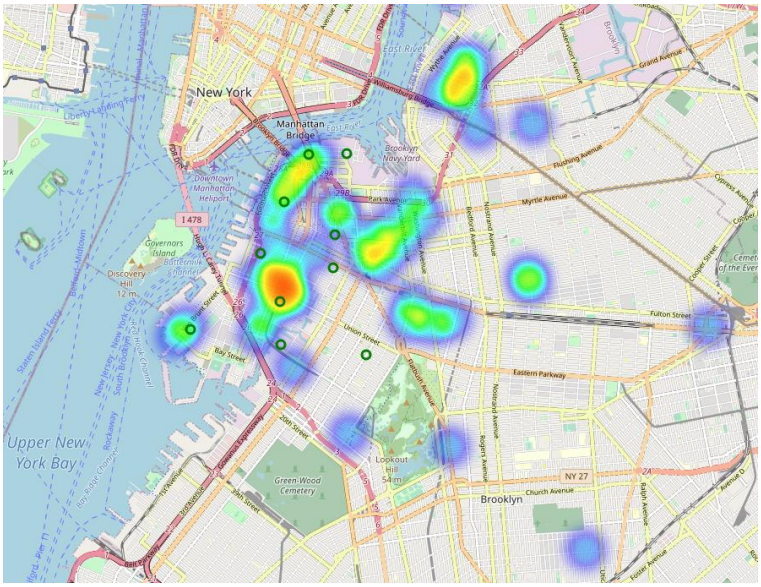


Figure 7 – Neighborhood location vs Wine Store / Wine bar density

It seems that we only have one neighbourhood without any Wine bar or wine store in the surroundings: **Park Slope**

According to several websites:

- *Park Slope is considered one of New York City's most desirable neighborhoods. In 2010, it was ranked number 1 in New York by New York Magazine, citing its quality public schools, dining, nightlife, shopping, access to public transit, green space, safety, and creative capital, among other aspects.* [Wikipedia]
- *After so many trips to New York (>50), we wanted to do something else and Park Slope is absolutely the best neighborhood to stay. It is relaxed, people are friendly and extremely helpful and. you can find everything you need and if you are desperate for big city live, it only takes 30 minutes and you are back in downtown Manhattan.* [TripAdvisor comment]
- *Family-friendly neighborhood hub in the heart of Brooklyn. Close to major railways and subways and Barclay's Center. Offers great shopping and dining options. The safety level is better than other areas of Brooklyn but could use some improvement at nighttime.* [Niche.com]

5. Results and discussion

One of the main objectives of our study was to find an optimal place to open our Wine Store close to Portuguese Restaurant. However, it was found that there isn't any Portuguese Restaurant in Brooklyn.

After using K-Means algorithm for clustering our data, it was choosen to divide into 7 clusters, based on Elbow method and Silhouete Score. The most notorious clusters, with more venues, were Cluster 0, 1 and 2. Cluster 0 is categorized for having mostly Pizza & Café venues and some necessary facilities (Banks, Pharmacies). This cluster is composed of 5 neighborhoods. Cluster 1 is the biggest cluster with 27 neighborhoods. Almost all neighborhoods of this cluster are located up north of Brooklyn. Coffee Shops, Pizza places, Italian Restaurants and bars are predominant in this cluster. As we can see too, the top 10 Neighborhoods with the highest income are from Cluster 1.

Another aspect that was important for our client it was the distance for another Wine venue. Based on the venues discovered by FourSquare API, we got a total of 23 Wine bars and 36 Wine Shops. Mostly are also located up North of Brooklyn.

It was important too, choose a neighborhood with an high income. Based on these important characteristics for our client at the end we chose Park Slope.

6. Conclusion

The purpose of this project was to identify a neighborhood with a Portuguese in the vicinity, with low wine venues around and the population should have a high income in order to aid our client to choose a best place to open her new Portuguese Wine bar. Using Foursquare data and clustering the venues using K-means we only identified one neighborhood that met this conditions. It was Park Slope.

However, it's necessary to do a further analysis to identify the best place in Park Slope. Final decision for an optimal location for the wine bar will be made by the client based on specific characteristics of locations based in the neighborhood, taking into consideration additional factors like attractiveness of each location (proximity to park or water), proximity to other Mediterranean restaurants (because it was not possible to find a Portuguese restaurant), real estate availability, prices...