**IBM – COURSERA**

**DATA SCIENCE SPECIALIZATION**

**CAPSTONE PROJECT – FINAL REPORT The Battle of the Neighborhoods**

****

**~ Nandini Gupta**

**Table of content:**

**Contents**

**INTRODUCTION** **3**

**BUSINESS PROBLEM** **4**

**TARGET AUDIENCE** **4**

**SOLUTION DESIGN APPROACH** **5**

**SUCCESS CRITERIA** **5**

**DATA** **6**

**METHODOLOGY** **6**

**ANALYTIC APPROACH** **6**

**DATA EXPLORATORY ANALYSIS** **7**

**RESULTS** **14**

**DISCUSSION** **14**

**CONCLUSION** **15**

**INTRODUCTION**

**The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over a land area of about 302.6 square miles (784 km2),**

**It is diverse and is the financial capital of USA. It is multicultural. It provides lot of business opportunities and business friendly environment. It has attracted many different players into the market. It is a global hub of business and commerce. The city is a major center for banking and**

**finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States. This also means that the market is highly competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analyzed carefully. The insights derived from analysis will give good understanding of the business environment, which help in strategically**

**targeting the market. This will help in reduction of risk and better control on the Return on Investment**

**New York is also the most densely populated major city in the United States. Located at the southern tip of the state of New York. A global power city, New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.**

**NY is split up into five** **boroughs: the**

**Bronx, Brooklyn, Manhattan, Queens,**

**and Staten Island. Each borough has the**

**same boundaries as a county of the state.**

****

**BUSINESS PROBLEM**

**The City of New York is famous for it’s excellent cuisine. It's food culture includes an array of international cuisines influenced by the city's immigrant history. Italian & Indian restaurants have become so popular in the United States now it seems that there is one on every corner, not only in major cities but also in smaller cities. One of my friends who is thinking of starting a restaurant in the NY neighborhood, consulted with me to get some analysis done with the all-possible data available. Manhattan being the costliest place, it was decided to compare rest of the boroughs and pick one of the most suitable neighborhoods with in the shortlisted boroughs. Based on the data analysis, it is expected to logically conclude which restaurant type (Italian Or Indian) and its recommended location. All the choices to be rationalized with the data analysis**

* + **it helps to distinguish the selections, securing long-term success. Overall Problem Statement can be broken into the following**

**Exploring the Boroughs in NY and narrow down to one.**

**Explore the Venues in the neighborhoods across that specific Borough**

**Narrow down to handful of neighborhoods and then deep dive into the current Restaurants & Hotels landscape across those.**

**Venue clustering by filtered neighborhoods and analyze the best choice of the restaurant and the best fit location.**

**TARGET AUDIENCE**

**Any Business Entrepreneurs or Companies who would like to start a Restaurant business in NewYork. The objective is to narrow down to best possible, affordable neighborhood to start a restaurant. The model also look at picking a type of restaurants from multiple choices like Italian Vs Indian. The Solution is expected to rationalize the choices backed up with data and its analysis. For this project, all boroughs except Manhattan being considered due to high cost.**

**SOLUTION DESIGN APPROACH**

**Solution is approached in seven steps as listed below**

**STEP 1: Pull all the boroughs & the respective neighborhood details of the New York data using newyork\_data.json.['newyork\_data.json' - https://cocl.us/new\_york\_dataset]**

**STEP 2: Deep Dive into the shortlisted Borough from Step 1 Using FourSquare APIs**

**STEP 3: Explore Venues across the neighborhoods in that Borough & Narrow down to handful of it based on larger number of Venues Vs less number of Restaurants +Hotels**

**STEP 4: Deep Dive into the shortlisted neighborhoods using, Word Cloud, Means of frequency of each category of Restaurants & identifying the Top5 Common Restaurants/Hotels**

**STEP 5: Clustering the neighborhood using K-means & identifying the locations on the Map.**

**STEP 6: Concluding the Choices of Restaurants & Locations basis of the data analysis in Step**

**SUCCESS CRITERIA**

**The success criteria of this project will be a good recommendation of borough/neighborhood for the choice of a restaurant, to the Stakeholder from the Target Audience. All choices and recommendations should be rationalized with the data analysis and inferences made.**

**DATA**

**One City will be analyzed in this project : NewYork USA .**

**Data sources that’s been analyzed in the projects are**

**Data1 : NewYork has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.**

**Data Source : newyork\_data.json' https://cocl.us/new\_york\_dataset**

**Data 2 : To Narrow down to one of the boroughs , basis of population /density analysis of the data available in Wikipedia**

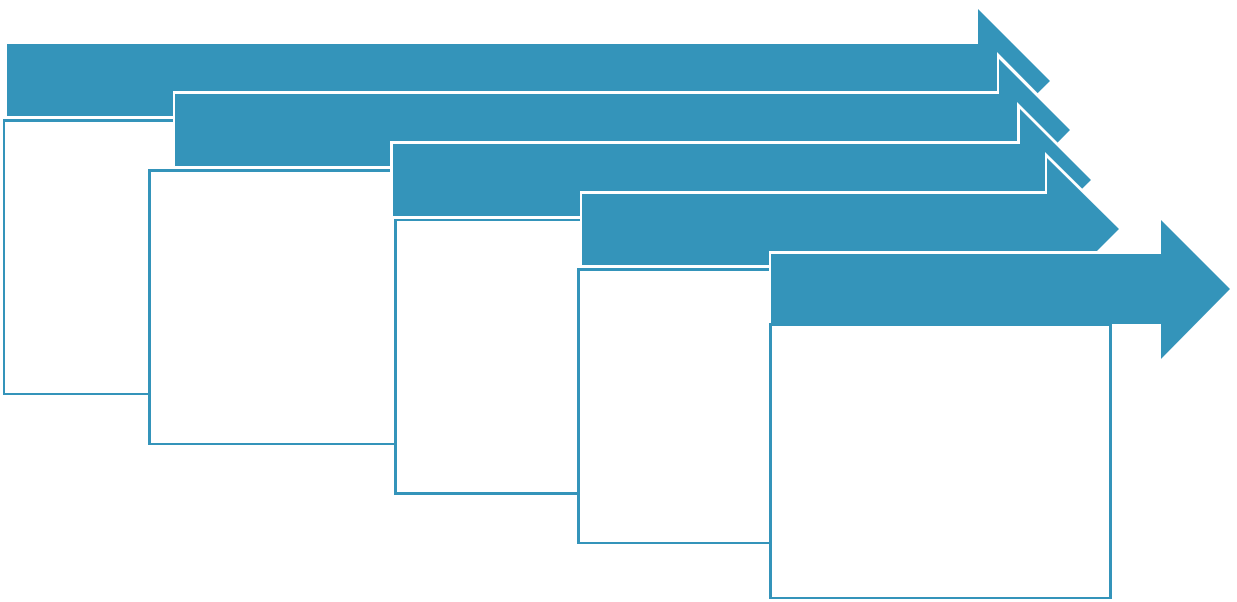
**Data Source : https://en.wikipedia.org/wiki/Demographics\_of\_New\_York\_City**

**Data3 : Exploring the neighborhoods in one of the shortlisted boroughs using FourSquare APIS**

**METHODOLOGY**

**ANALYTIC APPROACH**

**New York city neighborhood has a total of 5 boroughs and 306 neighborhoods. In this project we excluded Manhattan due to high cost and focus only on the rest of the 4 boroughs. From 300 + Neighborhoods across all the boroughs, we have applied the following analytic approach to narrow down to 3 Neighborhood in Brooklyn through multiple data exploratory analysis as explained below.**

****

**NewYork - 5 Boroughs / 306 Neighborhoods**

**~~Manhattan~~**

**1 Borough / 70 Neighborhoods**

**Brooklyn**

**Brooklyn (70 Neigh)**

**7 Neighborhood (100 Venues from FourSquare)**

**Queens**

**The Bronx**

**StatenIsland**

**~~Queens~~**

**~~The Bronx~~**

**~~StatenIsland~~**

**2838 Venues**

**Brooklyn Heights Caroll Gardens CobbleHill Downtown GreenPoint North Side South Side**

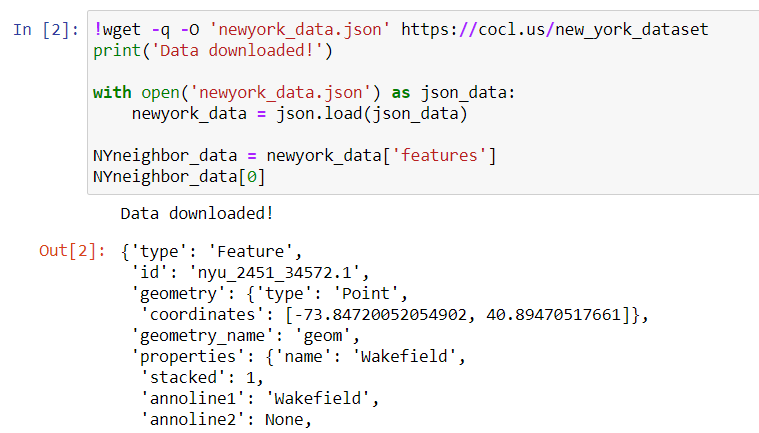
**3 Neighborhood(100Ven & <25 Rest)**

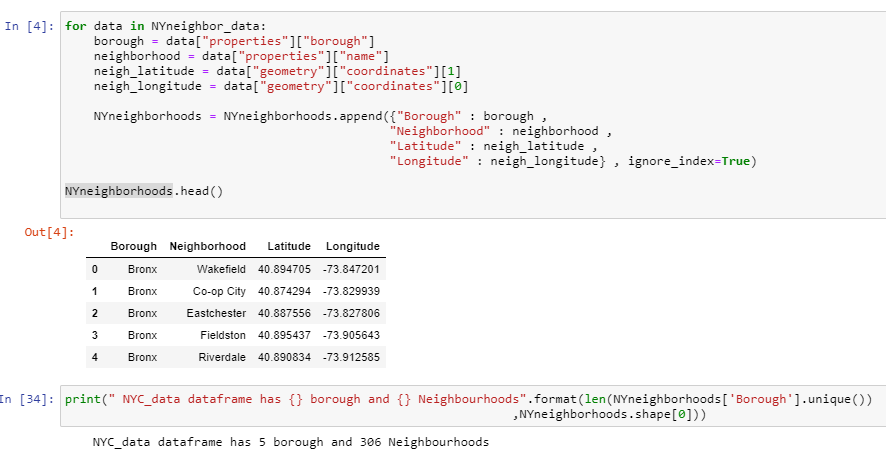
|  |  |
| --- | --- |
| **Brooklyn Heights** | **1Neighborhood & Restaurant Type** |
|  |
| **Caroll Gardens** | **GreenPoint for Italian Restaurant** |
| **GreenPoint** |
| **Carrol Gardens & GreenPoint for** |
|  |
|  | **Indian Restaurant** |

**DATA EXPLORATORY ANALYSIS**

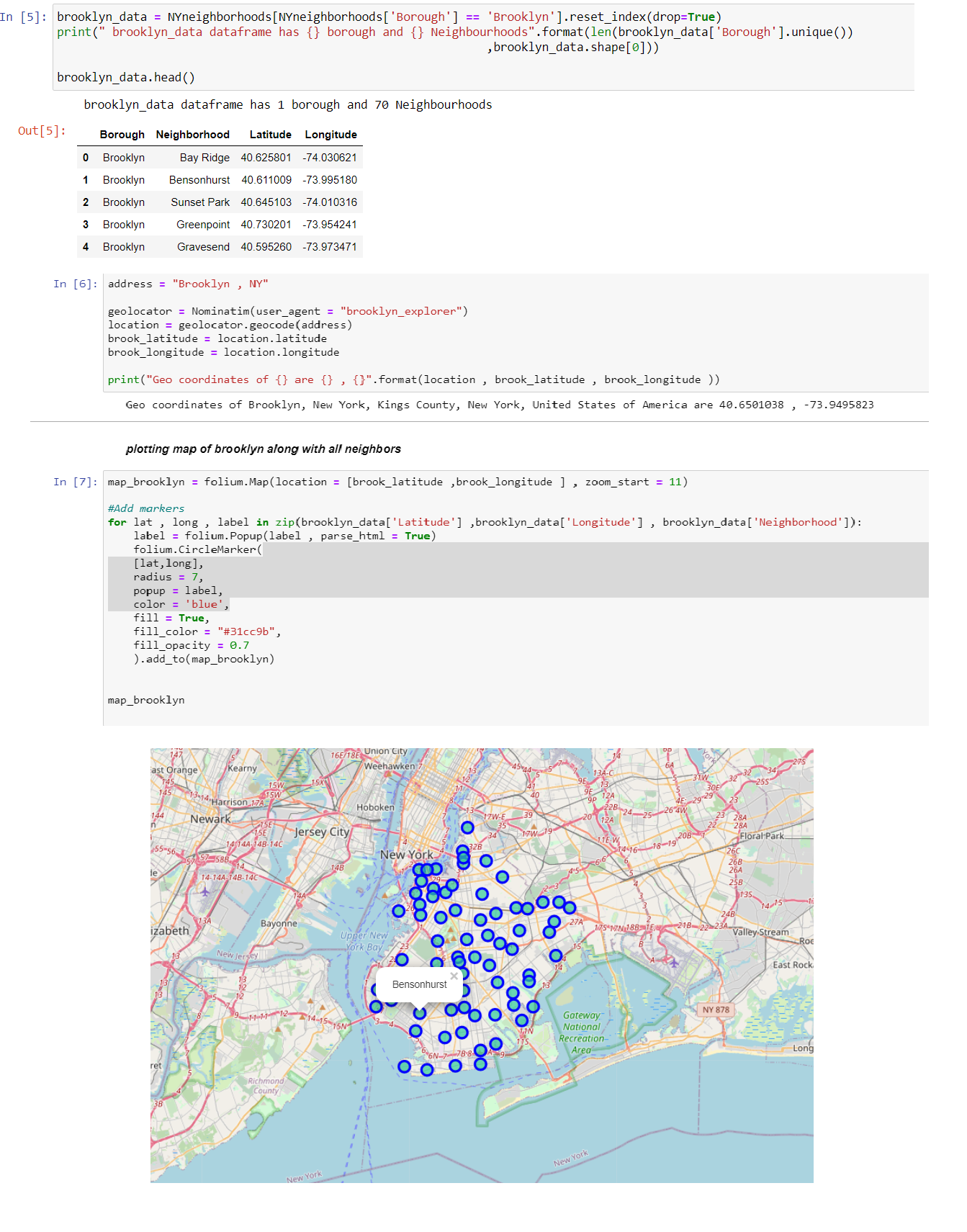
**Solution is approached in seven-step data exploratory analysis as explained below**

**STEP 1: Pull all the boroughs & the respective neighborhood details of the New York data using newyork\_data.json.['newyork\_data.json' - https://cocl.us/new\_york\_dataset]**

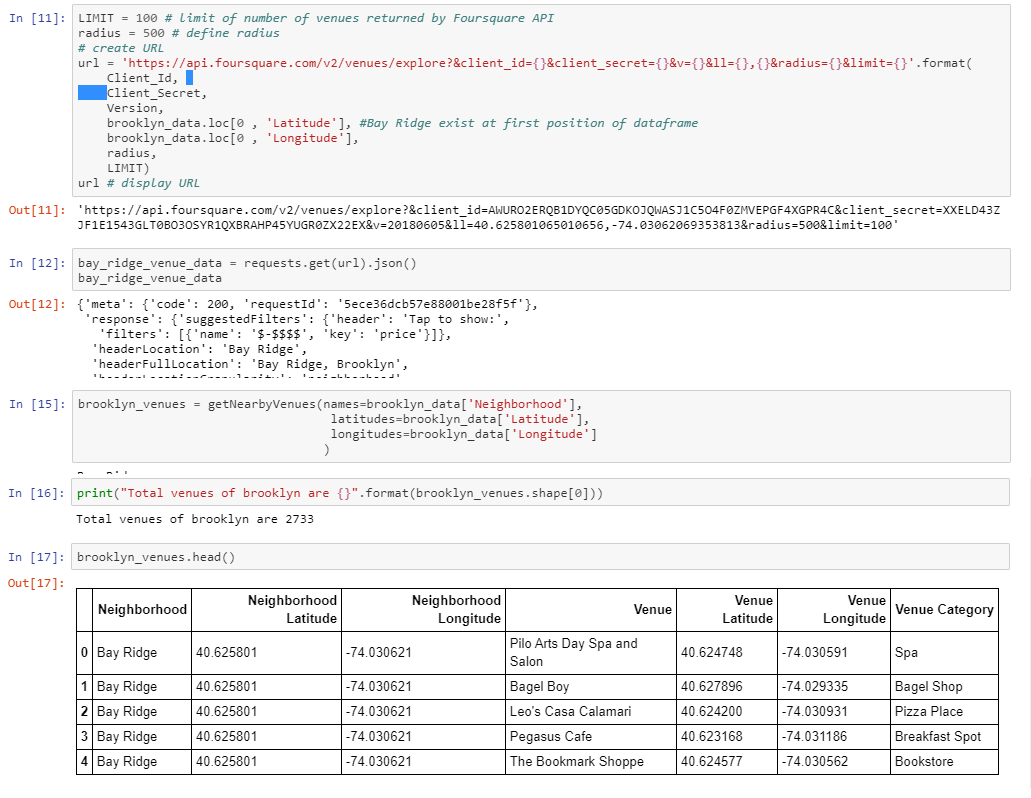
****



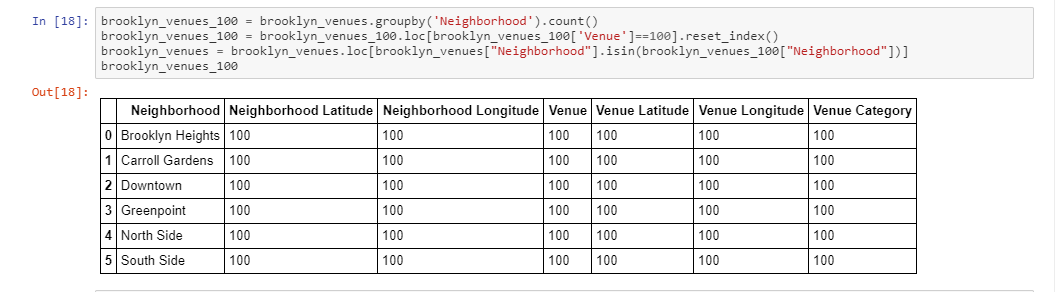
**STEP 2: Deep Dive into the shortlisted Borough from Step 1 Using FourSquare APIs**

****

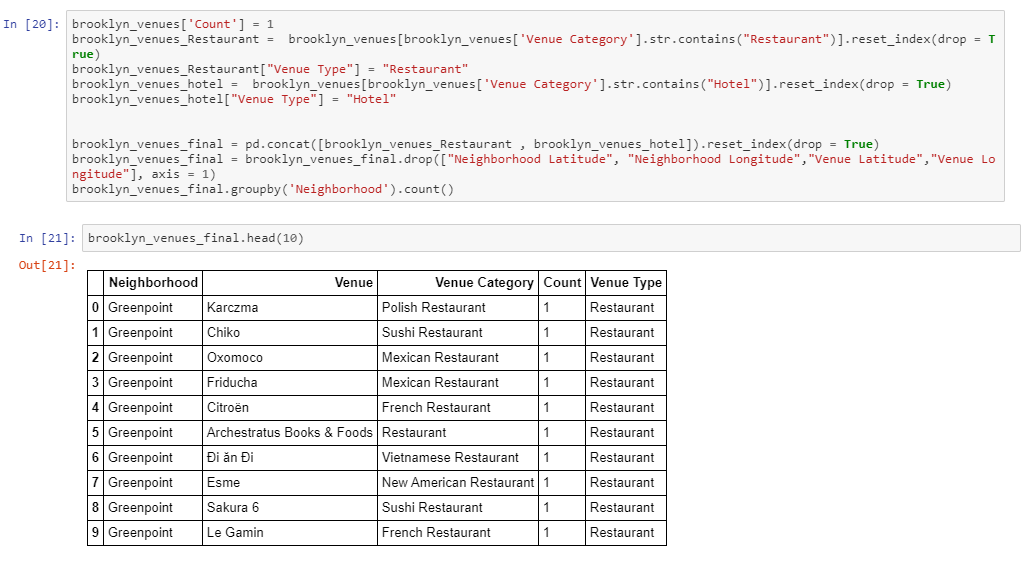
**STEP 3: Explore Venues across the neighborhoods in that Borough & Narrow down to handful of it based on larger number of Venues Vs less number of Restaurants +Hotels**

****

**FILTERING NEIGHBORHOODS HAVING 100 VENUES**

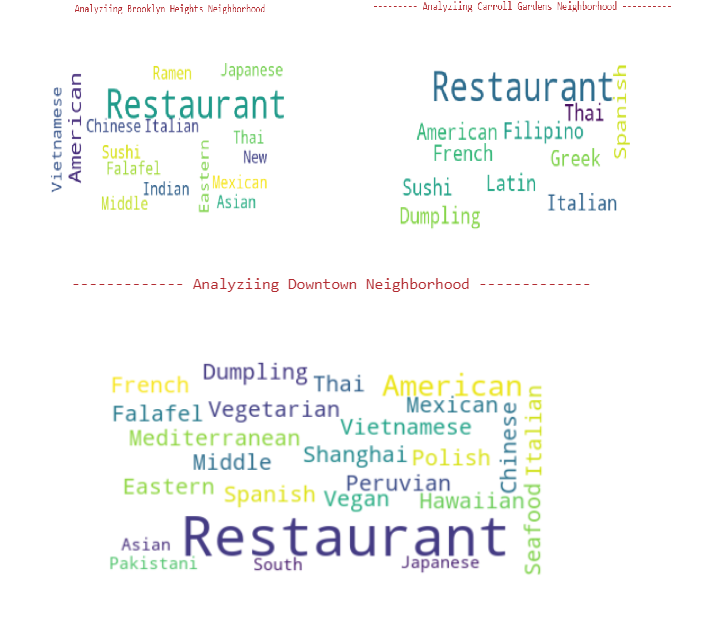
****

**FOCUSSING ON THE “RESTAURANTS & HOTELS” IN THE VENUE CATEGORY**

****

**STEP 4: Deep Dive into the shortlisted neighborhoods using, Word Cloud, Means of frequency of each category of Restaurants & identifying the Top5 Common Restaurants/Hotels**

**a) WORD CLOUD to look at the Restaurant Types among the Seven Neighborhoods**

****

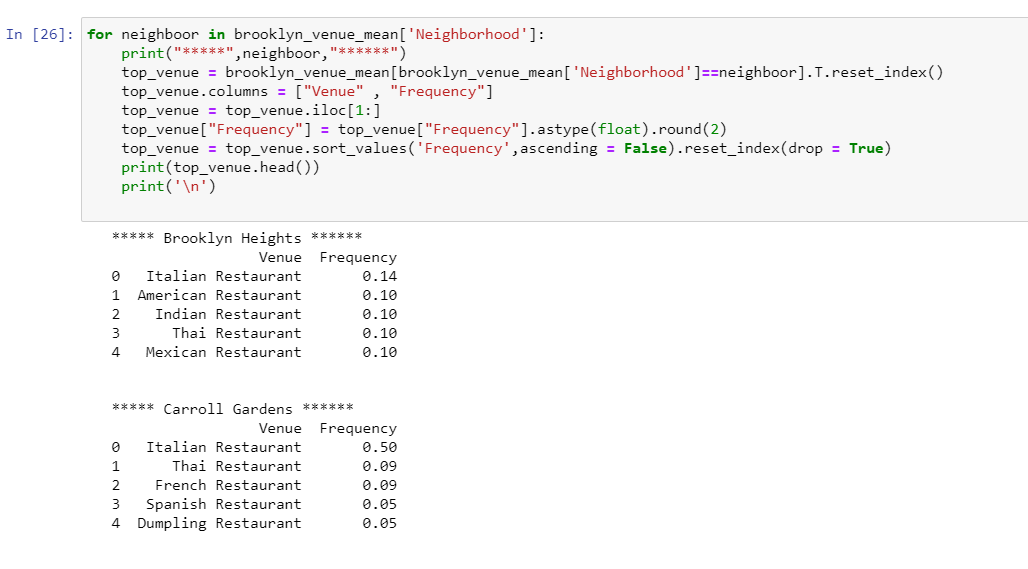
1. **PIVOT to Look at the Less Restaurants/Hotels Venues with in the shortlisted 7 Neighborhoods**

****

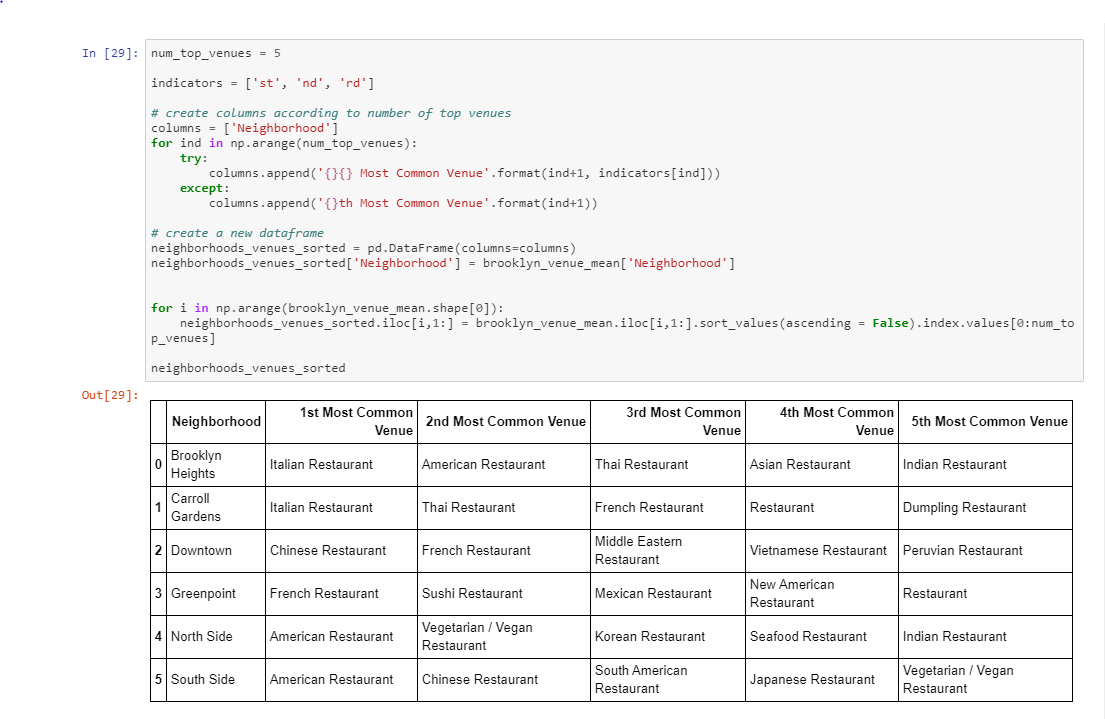
**c) Grouping the Neighborhood Using Means of Frequency of each Category**

****

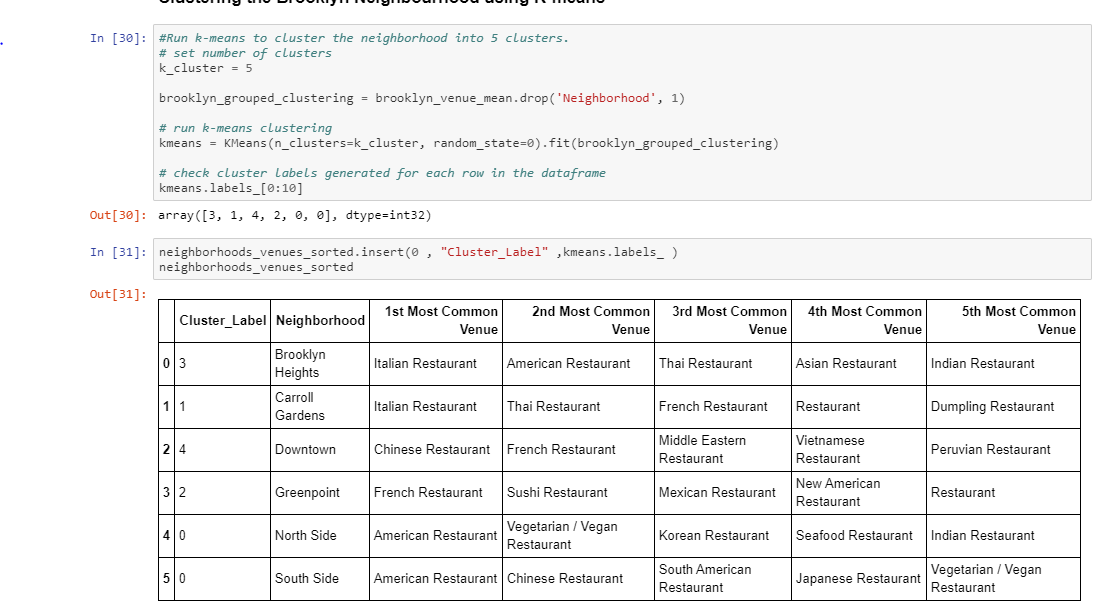
**d) Exploring Each Neighborhood along with top 5 Common Restaurants/Hotels**

****

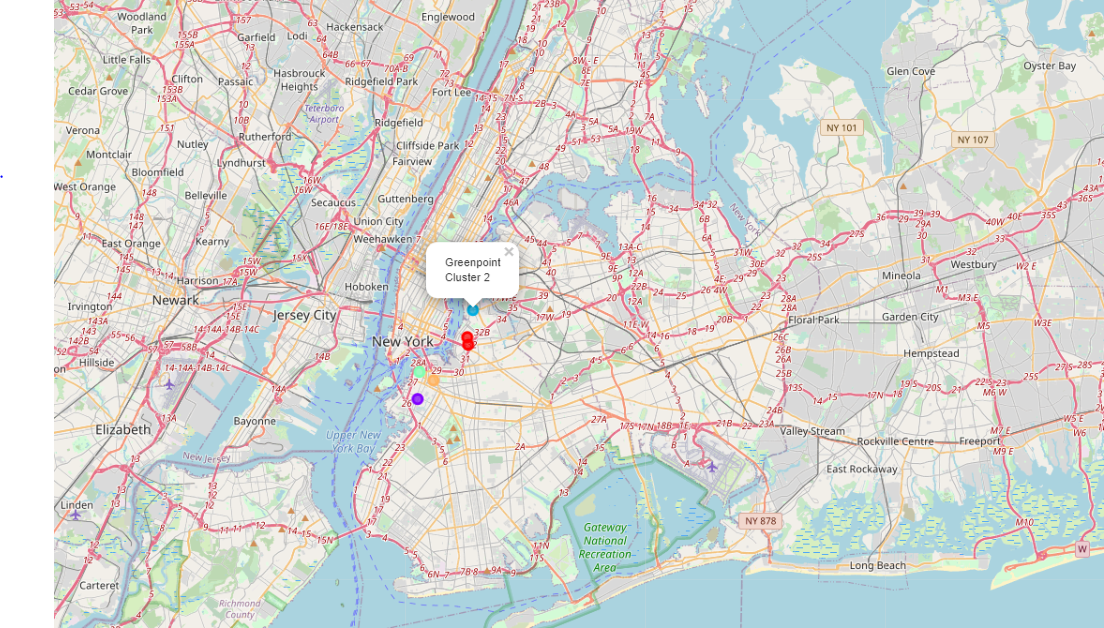
**e) Sorting the Venues in the Descending Order**

****

**STEP 5: Clustering the neighborhood using K-means & identifying the locations on the Map.**

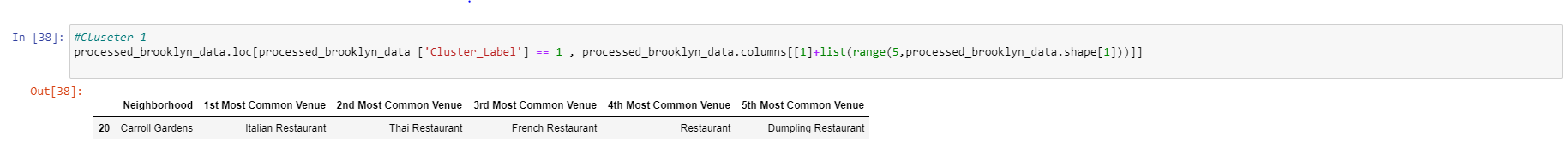
****

**CLUSTER MAP**

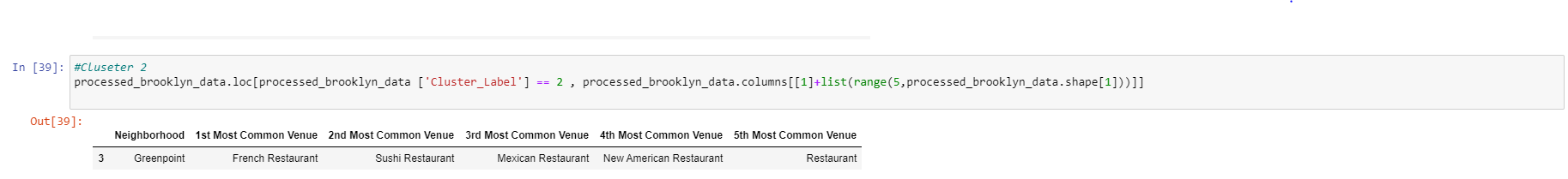
****

**STEP 6: Concluding the Choices of Restaurants & Locations basis of the data analysis in Step**

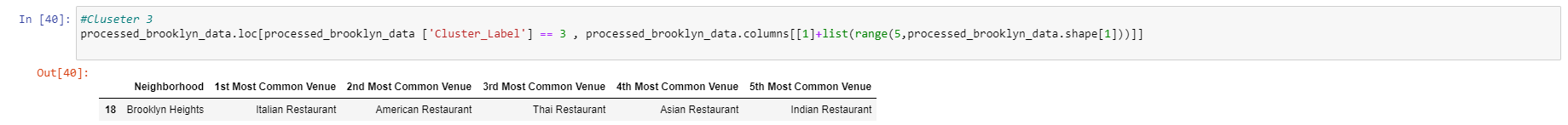
1. **Examining the Cluster -1 – Carrol Gardens**

****

1. **Examining the Cluster -2 - Green Point**

****

1. **Examining the Cluster -3 – Brooklyn Heights**

****

**RESULTS**

**Out of those shortlisted three Neighborhoods, Asian & Indian Restaurants are not that common in Cluster 1 or in Cluster 2, whereas it’s quite common in Brooklyn Heights. So Indian Restaurant would be preferred in Carrol Gardens or GreenPoint. If It’s Italian Restaurant, best bet would be @ GreenPoint.**

**DISCUSSION**

* **When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.**
* **For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.**
* **Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.**

**It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.**

**On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.**

**CONCLUSION**

**It’s an attempt to explore the different possible analysis we could do in the available data and rationalize the decision. Although all of the goals of this project were met there is definitely room for further improvement by analyzing few more supplementary data points like demographic information, Average Spent of the population, Proximity of other crowd pulling venues like Malls, shopping complex, Cinema halls etc. However, this project could definitely be handy to narrow down a Neighborhood and a**