**1. Introduction**

**1.1 Background**

The concept of grinding coffee beans and steeping the grounds in water has been around for centuries. The aromatic drink has drawn the attention of nearly every type of person. It is a product that is as ubiquitous as water itself, served around the world on every continent, and retains a high level of demand in every corner of Earth. In fact, interest has not waned but seems to be ever increasing as of late. So, with economic demand appearing nearly inelastic, and companies like Starbucks selling cups for $5 in some places, it would make sense for someone to want to enter the coffee bean distribution business. To achieve maximum cashflow, one might want to operate in the one of the largest markets available, New York City.

**1.2 Problem**

Opening a coffee bean distribution service in New York City might seem like a winning idea, but there are factors that need to go into the planning. For one, research should be done in order to determine an opportunistic location for the business. New York City is large, and there are countless possibilities for the placement of a bean supplier. One should consider, perhaps, the populations of New York City's boroughs and neighborhoods, prevalence of nearby cafes and coffee shops, and other retail stores that may influence the success of the business.

**1.3 Interest**

Those with interest in an analysis of coffee-centric neighborhoods would include an audience of the aforementioned coffee bean suppliers, plus a potential group of other stakeholders like cafe owners, local real estate investors, and even coffee-enthused tourists or residents. City officials may want to read such a study to help guide community planning as well.

**2. Data Extraction and Cleaning**

**2.1 Data Sources**

A comprehensive list of New York City's boroughs and associated neighborhoods can be found at the following URL: <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json>. From the .json file, we can additionally locate coordinates which we will use to visualize our research. Retail data will be extracted using a Foursquare API. Foursquare possesses information on a range of venues with details including reviews, photos, and product or service categories among others, but we will be primarily concerned with location and category data for the sake of the analysis.

**2.2 Data Cleaning**

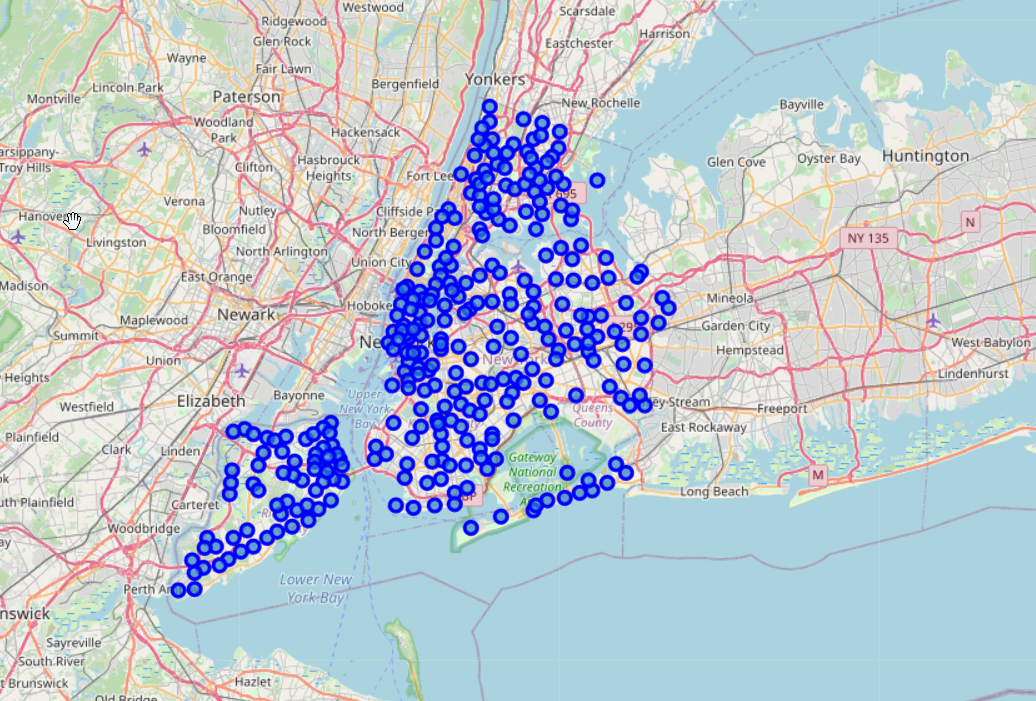
Data from sources will be transformed before loading to include the appropriate features, including but possibly not limited to, borough and neighborhood names, latitude and longitude or locations, venue names, and venue categories. Cleaning feature records/rows may be required to purge vacant data from cells as to not hinder any visualizations or computations. Also, lists should be scrubbed to omit erroneous data that could inadvertently affect the results of the analysis.

**3. Methodology**

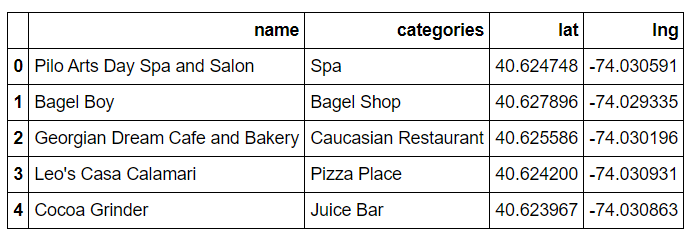
**3.1 Exploratory Analysis**

The first requirement of extracting data was locating a file that would contain a complete list of New York City’s boroughs and neighborhoods, along with each neighborhood’s corresponding latitude and longitudinal coordinates. For this objective, we turn to a familiar source used in a previous course lab in the form a .json file. Upon looking at the elements of the data, we see that our relevant fields are in the “Features” key. After isolating only the necessary information from the file, we convert the data into a *pandas* dataframe to better explore the data.

The cleaned dataframe shows us that we have details on five borough and 306 neighborhoods which should tell us that the .json list is fairly comprehensive, if not complete. So, we initiate an instance of the geocoder and map the coordinates using *folium* to see if the visualization displays realistic and accurate data.



Another main source of data we will utilize is Foursquare, for its extensive collection of retail and business information. Using the API, and a subset of the New York City neighborhood data, we look at the top 100 venues within a 500 meter radius of the first indexed neighborhood in Brooklyn (Bay Ridge). We can see that the code returns 79 venues and view a sample of the results.

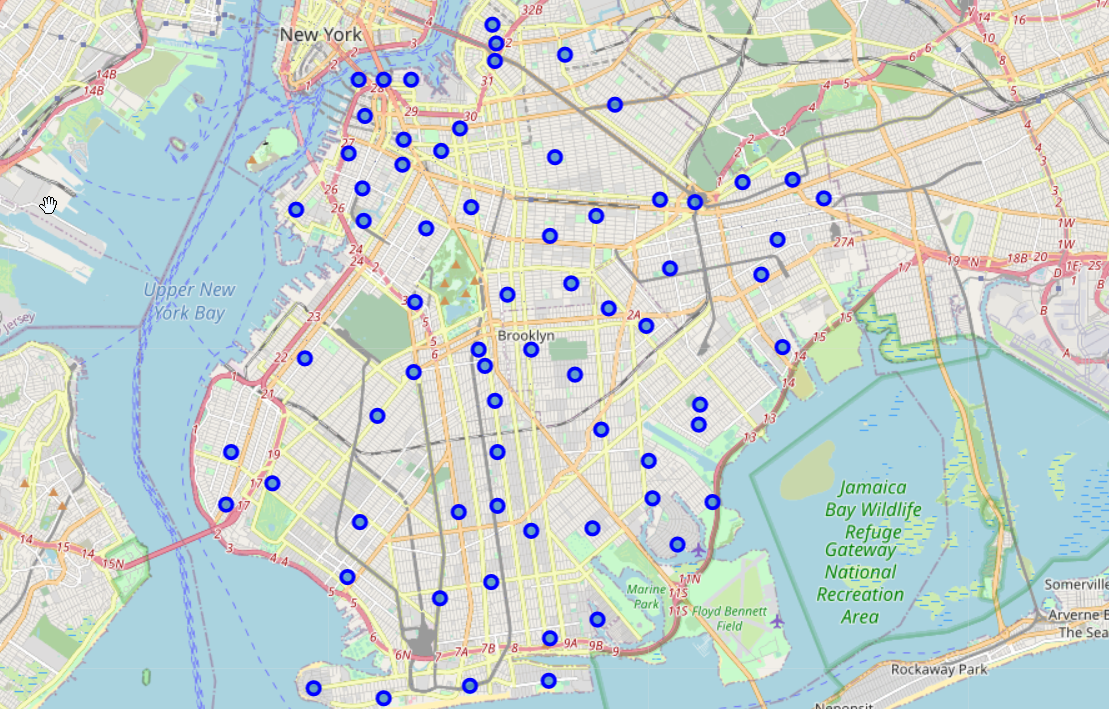


The code appears to work correctly, and we proceed with our knowledge of the data source nuances. From here, we will continue to shape and then model the data.

**3.2 Modeling**

Because of our data exploration into the neighborhoods and venues of New York City, we can understand that the magnitude of this metropolitan area is very large. Only looking at subsets of whole NYC communities and the venues residing in each tells us that we should first identify a promising borough upon which we can perform more thorough research. With each of the five boroughs having a large enough population to hypothetically constitute is own city, it would not be unreasonable to pick the one with the largest population, one that formerly did stand alone as an independent city at one point. Brooklyn is the largest, in terms of population, of New York’s boroughs, and it is dense with shops and residential neighborhoods. Someone looking to open a business of any type would not be faulted for wanting to look at Brooklyn. For the purpose of this analysis, we will also create a new dataframe focusing solely on the neighborhoods of NYC’s Brooklyn borough.

After filtering the geographic dataframe, we can take another look, using *folium*, at the layout of Brooklyn’s neighborhoods.

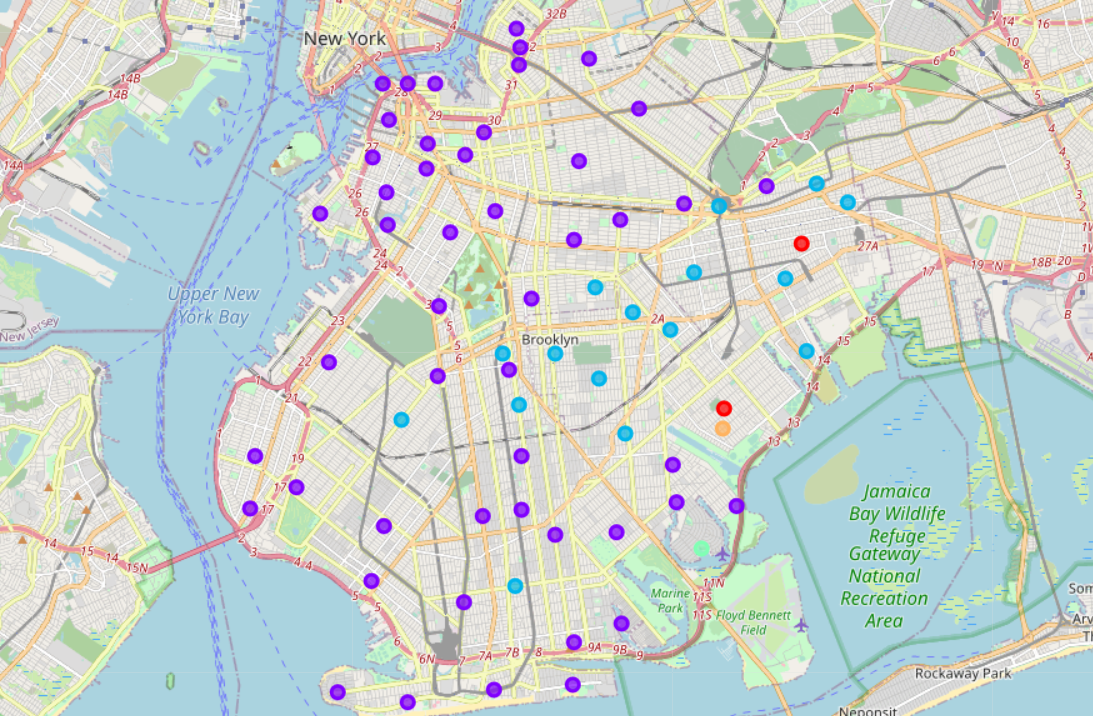


And just as we had located venues in the proximity of the Bay Ridge neighborhood, we expand our code to include all the neighborhoods in Brooklyn. A new dataframe named “brooklyn\_venues” is also created. The size of the new dataframe is (2767, 7), and there are 291 unique venue categories we can use in our analysis.

So, we begin to analyze each neighborhood with one hot encoding and create yet another dataframe that will allow us to view the mean frequency of occurrence of each venue category, grouped by neighborhood name. The size of the dataframe looks much different after grouping (40, 332). If we re-sort the dataframe to list each neighborhood’s top 10 venue categories by name, we get a more intuitive chart.



Subsequently, we will run *k-means* to cluster the neighborhoods into five distinct clusters. And once again, we turn to *folium* to visualize the clusters on a map of Brooklyn. We also display tables of the clusters to interpret details of the five.



**4. Results**

The cluster details show us that clusters 0, 3, and 4 are represented by very few neighborhoods. Specifically, just four of Brooklyn’s neighborhoods are clustered into three of the five clusters. And looking at each of these clusters, it is clear that they are niche neighborhoods on the outskirts or bays of New York City with uncommon businesses (i.e. bus line, pool, harbor/marina). The other two clusters appear to border one another with many more neighborhoods comprising each.

A closer analysis reveals that clusters 1 and 2 identify more consumer-centric retail shops like restaurants, banks, and clothing stores. But for the purposes of this report, we need to figure out where the shops that affect coffee sales reside. The most common venues in neighborhoods contained in cluster 2 look like they might be very food heavy. We have several neighborhoods with “Friend Chicken Joint” representing the first most common venue. Caribbean and Chinese restaurants seem to be the most common venue in few additional neighborhoods. This would lead me to believe that there are a lot of dinner-focused restaurants in cluster 2. If we look for signs of coffee along the columns of this cluster’s chart, we do see a few instances of coffee shops and cafes, but they do not seem as prevalent as other restaurant types or grocery stores. Having many grocery stores might play favorably into the decision to place a coffee supply business in proximity, however, we also need to consider the characteristics of our final cluster.

Cluster 1 is the largest of the clusters and appears situated primarily on the north, west, and south sides of Brooklyn, where the east side seems to be occupied by neighborhoods of clusters 0, 2, 3, and 4. A closer look at the types of venues contained in cluster 1 shows neighborhoods targeting a more diverse market of consumers. We can see everything from spas to bodegas in the table. But among the most common venues found in each neighborhood, we also see several coffee shops and cafes. To see coffee shops along with sushi restaurants, wine shops, and spas says to me that these neighborhoods are likely striving for a more sophisticated appearance. Whereas cluster 2 had an ethnic and comfort food feel, cluster 1 is covering a different demand overall.

**5. Discussion & Conclusion**

It would seem to me that anyone wanting to open a coffee bean supply business would want to reconsider doing so on the east side of Brooklyn, where there looks to be a different kind of commercial feel. This is assuming that whomever opening the business desires to be within the proximity of many coffee shops, cafes, bodegas, or other businesses where coffee might be sold. A different market research might be conducted on the feasibility and economics of opening a physical store in these neighborhoods, and whether it may be less expensive to operate on the east side. How marketing might be a factor in working close to your clientele, and how much that is worth, would also be a consideration.

Models in this study relied upon quantitative data and did not take into consideration qualitative factors possibly including the sentiment of the residents that live in Brooklyn’s neighborhoods and how they feel which trends might be affecting them. This project is also open to additional models and analysis for future expansion of insight.