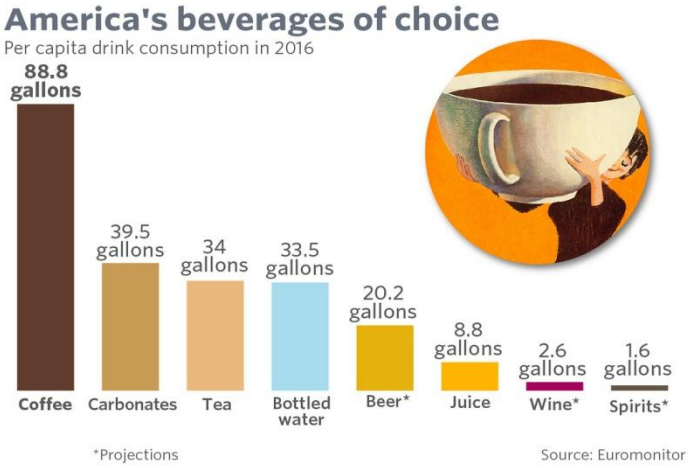
Data Science Capstone Project Report

The Battle of the Neighborhoods – New York City

New York City (NYC), often called New York (NY), is the most populous city in the United States. It's the city that never sleeps – and the caffeine helps. **New York City revolves around coffee**. In the early-1800s, New York was one of the largest coffee roasting centers in the United States. More recently, mobile startup Massive Health found that New Yorkers drink 6.7 times the amount of coffee consumed by the average denizen of any other US city. NYC has a varied coffee landscape.



Business Problem:

General Overview of the Industry

Coffee is one of the world's largest commodities. The top green coffee producing countries are Brazil, Colombia, and Vietnam. Many grower countries are small, poor developing nations that depend on coffee to sustain local economies. The U.S. is the world's largest importer of green coffee beans and the largest consumer of coffee.

Coffee consumption is highest in the Northeast, where over 60 percent of the population consumed coffee daily in 2005, according to the National Coffee Association (NCA). Per capita consumption is highest in the Central U.S., where coffee drinkers average 3.7 cups per day.

As consumers, deciding where to buy coffee is relatively easy. For the most part, our decisions come down to figuring out how far will you travel and how much will you pay to get coffee from your favorite coffee shop?

As coffee shop owners, deciding where to locate is more complex. The following questions are faced by coffee shop owners:

- Should you set up shop near Battery Park city (Manhattan) despite an oversaturated competitive landscape?
- Should you locate in a less competitive area like Tottenville, (Staten Island) despite lower customer density?
- Should high-end coffee shops go to Gramercy (Manhattan) since apartment rents are higher than other parts of the city?

Competitive Landscape

Consumer taste and personal income drive demand. The profitability of individual companies depends on the ability to secure prime locations, drive store traffic, and deliver high-quality products. Large coffee chain companies like Starbucks, have advantages in purchasing, finance, and marketing. Small companies can compete effectively by offering specialized products, serving a local market, or providing superior customer service.

Coffee shops compete with businesses such as convenience stores, gas stations, quick-service and fast-food restaurants, gourmet food shops, and donut shops. The US coffee industry is concentrated: the eight largest companies account for about 70% of revenue. Third wave coffee shop like Blue Bottle Coffee, Stumptown coffee roasters are pushing the competitive edge even further. This makes it difficult for local owners to setup their coffee shops easily. The main concern of the local coffee shop owner is which location to choose in order to open a coffee shop. Location can make or break a coffee shop. The next is the competition faced from mainstream coffee chains like Starbucks, Dunkin' etc. which have strong financing, marketing budget plans. After choosing a location, the next concern is the rent of the place, availability of hiring staff, expense on labor etc.

Problem Statement

Keeping this into view, the focus of this capstone is to help local coffee shop owners who want to open a coffee shop in one of the neighborhoods of New York city. Therefore, the analysis and results of this project would interest stakeholders who are interested in opening an independent coffee shop in New York city. I use Foursquare Data to get the coffee venues in neighborhoods of NYC. I also use the American Community Survey Data to get the details of the city's detailed demographic, socioeconomic, and housing characteristics. I group the neighborhoods into clusters based on the number of venues, demographic characteristic like population, socio-economic characteristics like median gross rent and housing characteristics like median household income. I use the Folium library to visualize the neighborhoods in New York City and their emerging clusters. On basis of the assignment of neighborhoods to individual clusters, ideal location of the next coffee shop venues are predicted.

Data Acquisition

Extraction from data sources and cleaning:

1. Foursquare data:

Foursquare's data, with over 3 billion visits/month around the globe, 105 million global venues, and 25 million people globally (both in and outside of the apps, via the Pilgrim SDK) who have

opted in to always-on location sharing, is incredibly valuable to advertisers, businesses and developers. The Places API offers real-time access to Foursquare's global database of rich venue data and user content to power your location-based experiences in your app or website.

The API endpoints like search, explore finds the nearest venue in a given radius and returns different venue characteristics. The API call returns a JSON file and I have turned that into a data-frame. Here I've chosen coffee venues(coffee shops, cafes, tea rooms) for each neighborhood within a radius of 500 meters.

Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bronx	Kingsbridge	40.881687	-73.902818	Mon Amour Coffee & Wine	40.885009	-73.900332	Coffee Shop
Bronx	Kingsbridge	40.881687	-73.902818	Gold Mine Cafe	40.878916	-73.904698	Café
Bronx	Kingsbridge	40.881687	-73.902818	Tony's Cafe	40.879280	-73.905228	Café
Manhattan	Marble Hill	40.876551	-73.910660	Starbucks	40.877531	-73.905582	Coffee Shop
Manhattan	Marble Hill	40.876551	-73.910660	Starbucks	40.873755	-73.908613	Coffee Shop

Fig : A snapshot of the data returned by Foursquare API.

2. New York City data

NYC 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. Luckily, this dataset exists for free on the web in the form a geojson file which is further cleaned by extracting key and values and prepared in the following format.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Fig : A snapshot of the data of NYC neighborhoods

3. American Community survey(2018)

The American Community Survey (ACS) is the most extensive nationwide survey currently available. From its annual releases we are able to examine the city's detailed demographic, socioeconomic, and housing characteristics. Each month, questionnaires are sent to a sample of

approximately 295,000 addresses across the country, so households are continuously receiving and responding to the ACS. I have selected 5-year estimates from the ACS. The primary advantage of using multiyear estimates is the increased statistical reliability of the data for less populated areas and small population subgroups.

Cleaning of data sources:

Characteristics like population, median household income, median gross rent are selected across individual datasets compiled from the survey. I have cleaned and collected all the data from different datasets into a single dataframe. One thing to note is that the data is collected for a bunch of neighborhoods called neighborhood tabulation areas(NTA). NTAs are the geographic areas for use with both the 2010 Census and the American Community Survey (ACS) combined together for the purpose of better estimates. Each NTA consists of one or more neighborhoods with a unique code(GEOId) called the NTA code.

A snapshot of the compiled data from different datasets along with the NTA codes

	GeogName	GeoID	Borough	Pop_1E	MdHHIncE	MdGRE
0	Allerton-Pelham Gardens	BX31	Bronx	31993	71414	1417
1	Bedford Park-Fordham North	BX05	Bronx	57685	37282	1259
2	Belmont	BX06	Bronx	29115	28484	1172
3	Bronxdale	BX07	Bronx	39423	38587	1207
4	Claremont-Bathgate	BX01	Bronx	35560	25061	860
5	Co-op City	BX13	Bronx	47400	51613	1032
6	Crotona Park East	BX75	Bronx	22103	22123	933
7	East Concourse-Concourse Village	BX14	Bronx	64850	32012	1140
8	East Tremont	BX17	Bronx	44057	23863	997
9	Eastchester-Edenwald-Baychester	BX03	Bronx	37887	54568	1247
10	Fordham South	BX40	Bronx	28164	26122	1171

Pop_1E – Population Estimate

MdHHIncE – Median Household Income

MdGRE – Median Gross Rent

Feature Selection

Competition

Competition is a very important factor while setting up a coffee shop. Proximity to other competing businesses plays a key role in deciding the success of the coffee shop. Establishing which competitors are in your area and their offering could help guarantee you choose the right location for your business. If there is too much competition then it is a warning sign to find a new location. While setting up a new coffee shop, it is crucial to know the number of similar businesses in a neighborhood. By leveraging the power of Foursquare API, we can determine the density of coffee venues in every neighborhood. This will give the coffee shop an idea of the competition in a given area.

Population

Needless to say, the best place to locate a café /coffee shop is in an area with a high population. Finding the population base of an area can tell you the income range of your potential coffee shop location. Population tells you whether there will be enough people nearby with sufficient expendable income to support your establishment, and it can provide guidance as to what they're most likely to spend their disposable money on.

Median Household Income:

Median household income is a measure of affluence. Median income is the amount that divides the income distribution into two equal groups, half having income above that amount, and half having income below that amount. Median household income is one measure, among many, that gauges the economic well-being of a region. Median household income provides information about the financial resources available to households. Higher household incomes are commonly associated with a greater means of acquiring goods and services.

The median household income in a given city and its neighborhood is closely tied to employment levels, educational attainment, and regional economic opportunities. The wealthiest metropolitan areas are home to large college educated populations. They also have a healthy job market and tend to pay higher wages. The median household income of the residing population can be useful while deciding the price of the items sold buy a coffeehouse. High incomes are also reflected in metro areas property values and rent. While deciding to open a coffee shop in neighborhoods having a low to moderate income a comparatively larger population, it is reasonable to keep affordable prices. Thus, Median household income is a good indicator as high incomes suggests that people have extra money to dine out.

Median Gross Rent

Gross rent is the contract rent plus the estimated average monthly cost of utilities (electricity, gas, and water and sewer) and fuels (oil, coal, kerosene, wood, etc.). A coffee shop owner needs

to choose a location that has a reasonable rent. The total cost of rent and rates should be less than 10% of the sales. A neighborhood with a lower rent can be highly favorable but at the same time it should not be a place with low customer density. High rents can affect owners who are tight on budget. Thus, rent of a neighborhood plays a decisive role while opening a coffee shop.