CAFE LOCATION SEARCH :

MANHATTAN

J. Duane

3/13/2021

INTRODUCTION / BUSINESS PROBLEM

For this project we attempted to help a hypothetical client find the best area in Downtown Manhattan to open a new cafe. As Manhattan is a pretty saturated market for cafes and it is an expensive area to operate in, it is crucial for our client to gain a deeper understanding of the market they are entering. Using location data from currently successful restaurant operations, we can regroup our city based on the locations of these high volume restaurants rather than the neighborhoods they fall in, which may give us a different understanding of where people are going out to eat. Our goal is to use data and clustering to find a high volume area that is the least currently saturated with cafes on our list. By positioning their new cafe near other high volume locations, they may be able to capitalize on existing traffic trends. Additionally, If there are less cafes already in existence in an area, it will be easier for them to build a starting customer base.

DATA

To help answer this problem, we will be using a dataset scraped from Restaurant Business Online's Future 50 Restaurants list and made available via Michal Bogacz on Kaggle. This list includes fifty restaurants that were able to successfully grow their operations during a difficult period in the industry, and were chosen for their ability to adapt and likeliness to continue to excel in the future. The data frame includes the following columns:

Rank	Position in ranking
Restaurant	Name of restaurant
Location	Location of origin
Sales	2019 Systemwide Sales($000000)
YOY Sales	Year over year sales
Stores	Number of premises
YOY Store Growth	Year over year premises increase
Unit Volume	2019 Average Unit Volume ($000)
Franchising	Is the restaurant a franchise? (Y/N)

We will focus on the New York Based companies on this list and group their New York City locations based on coordinates. We will then cluster these groups and find the average volumes of each zone. We will see what type of stores appear in which zone, and find which zones could potentially be considered for future cafe locations.

Out[15]:

	Rank	Restaurant	Location	Sales	YOY_Sales	Stores	YOY_Store_Growth	Unit_Volume	Franchising	Franchise?	NYC
0	34	Boqueria	New York, N.Y.	27	22.0	7	16.7	4260	No	0	1
1	25	By Chloe	New York, N.Y.	37	25.6	14	7.7	2800	No	0	1
2	23	The Little Beet	New York, N.Y.	23	26.5	12	33.3	2230	No	0	1
3	29	Dos Toros Taqueria	New York, N.Y.	28	24.0	22	10.0	1375	No	0	1
4	8	Melt Shop	New York, N.Y.	20	39.6	19	35.7	1260	Yes	1	1
5	32	Just Salad	New York, N.Y.	42	22.7	38	26.7	1240	No	0	1
6	14	Bluestone Lane	New York, N.Y.	48	33.0	48	37.1	1175	No	0	1
7	24	Joe & The Juice	New York, N.Y.	47	25.9	69	25.5	760	Yes	1	1

To prepare this data frame for this project, we changed the percent values to floats for our Year over Year values. We also used one hot encoding to add columns for Franchising and whether or not the Restaurant is New York City Based. We used these columns during our exploratory data analysis to check for correlation.

EXPLORATORY DATA ANALYSIS

Through our initial data analysis, we found that 8 of the restaurants in our Data frame originated in New York City, and would therefore have a sizable presence in the city. (figure 1)

We also created a heat map to search for any relevant correlation in our data that could be useful in determining future store locations. From our heatmap, we found a potential negative correlation between Franchising, Unit Volume, and the Amount of Stores (Figure 2). From this we could deduce that more stores overall might lead to lower volume per store, however this information does not directly help us choose a store location so we will save it for another project.

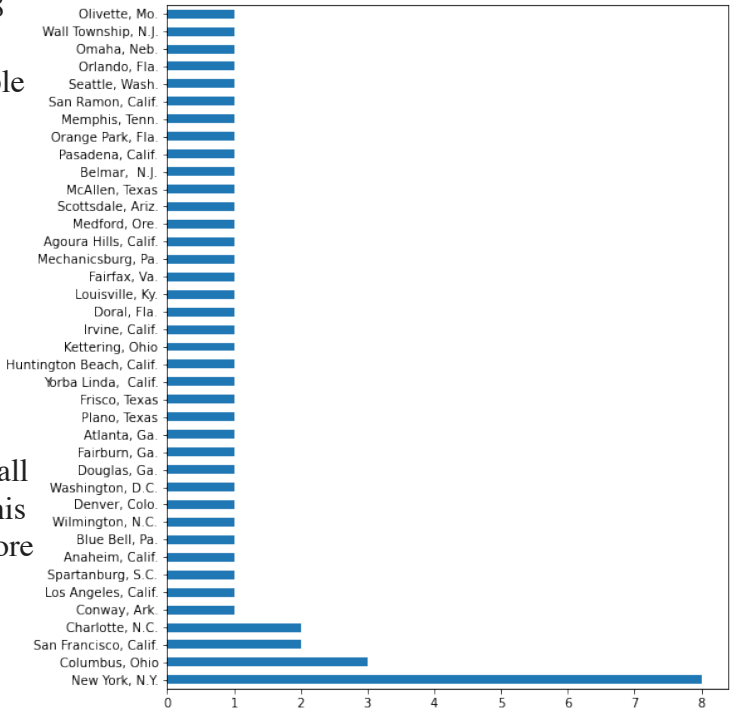
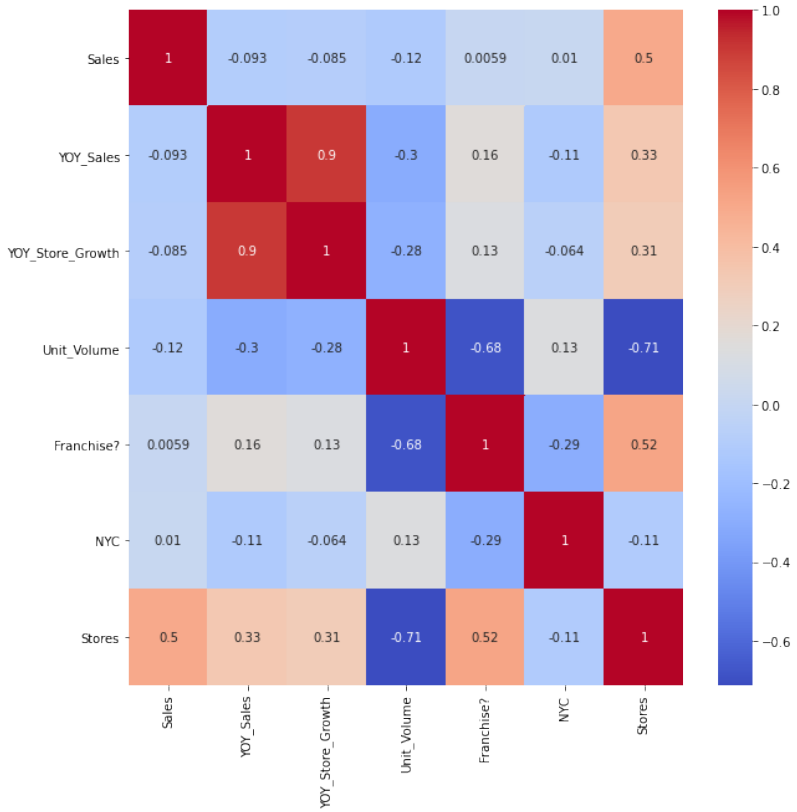


Figure 1

Figure 2



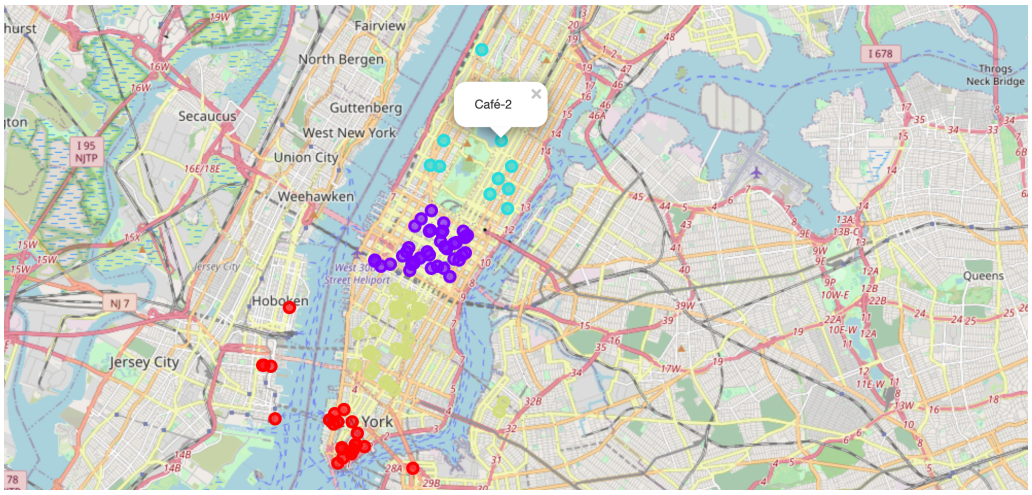
METHODOLOGY

To obtain coordinates for our stores, we ran a search query within a 6 mile radius around Downtown Manhattan for the restaurant names on our list.

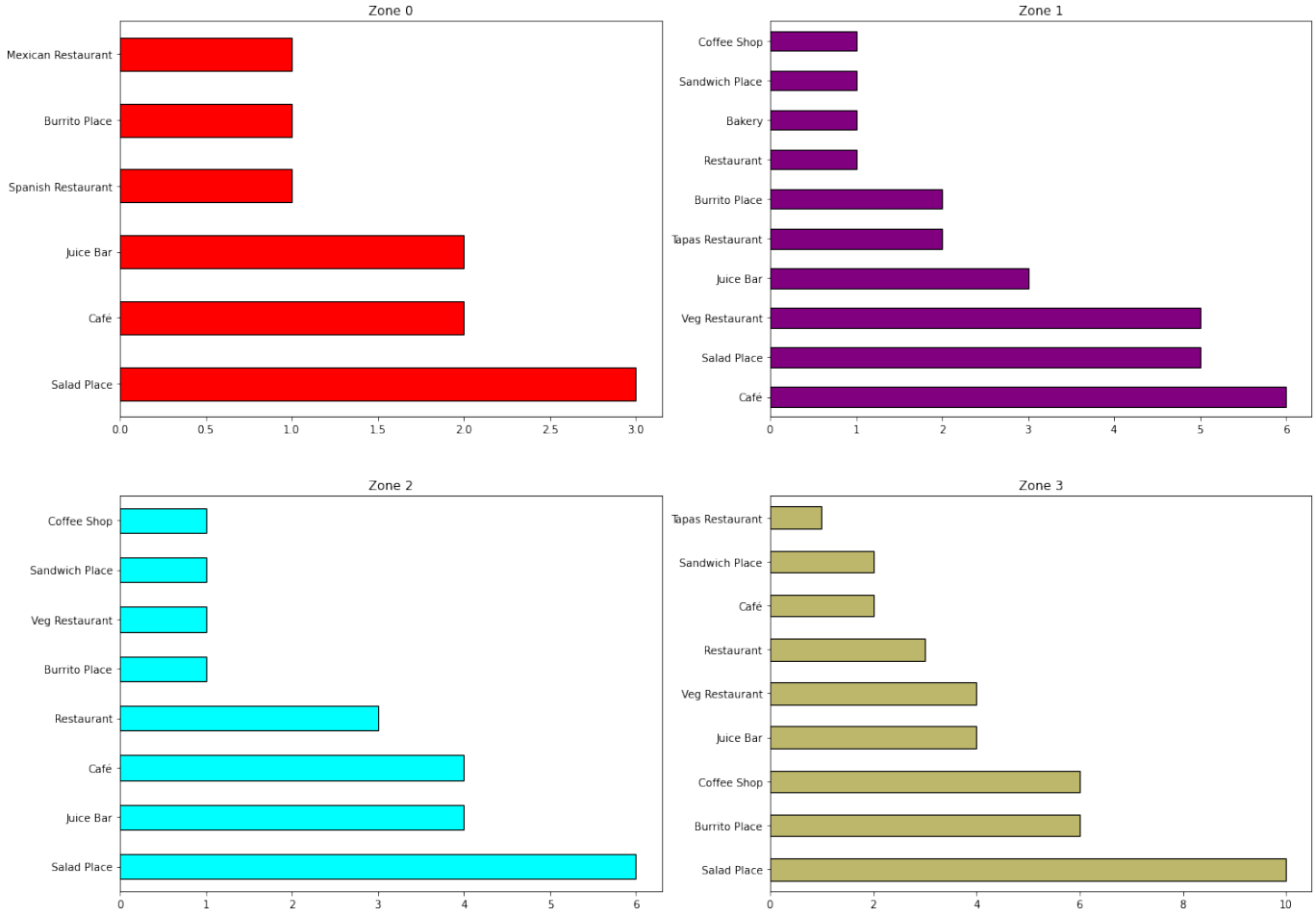
Our dataset now contained coordinates for over 80 successful, high volume restaurant and cafe locations in downtown Manhattan.

Our next step would be to cluster these stores and find which areas have the highest volume per zone, with secondary consideration focusing on the composition of store types within each zone. To achieve this we would cluster our dataset into zones based on their coordinates.
We used the elbow curve method to find the appropriate number of clusters for our data based on restaurant location. The elbow curve method allows us to find the point where returns become diminished and are no longer worth adding additional clusters. In this case our result was 4 (figure 4).

After k means, we had split our data into four clusters with an average size of 24 stores per cluster. We visualized our points over a map of Manhattan using Folium, and labelled each point with its store type and cluster zone number(figure 5) Exploring this map would allow us to find any relevant landmarks that would affect volume, such as Grand Central Station's location in zone 1.



The store 'type' breakdown was as follows:



We also calculated the mean store volume for each cluster zone:

cluster zone	average store volume
zone 0	1460.0
zone 1	1766.48
zone 2	1310.0
zone 3	1505.0

RESULTS

We could consider looking more closely into zone 1 which has the highest average volume, however it holds the second most cafes and coffee shops. This seems to be an extremely high traffic area containing two of the major transportation hubs for the City, and this clustering process can be repeated within this zone to gain more accurate results.
From this data we would also likely consider Zone 0, which has significant volume and less cafes and coffee shops when compared to the other zones.
We can also use this method more accurately by incorporating different store types outside of the original dataset, in order to gain a better fit for a specific restaurant type. While this example was focused around finding a high volume location for a cafe, we could add in different search keywords and find underserved areas for different concepts.