Harlem, NYC: Which Cuisine Should You Invest In?

July 2020

I. Introduction

A. Background

New York City is world renowned for its food scene. The availability of different types of cuisine is what draws many consumers and tourists from across the globe. One of the hubs of this food scene is Harlem. Known for both its diversity and its accessibility, you can get affordable food of almost any type of cuisine there. That's what makes it difficult for new restaurant owners to compete – what cuisine should you cater to if there are already so many?

B. Business Problem and Target Audience

One of the most challenging things for restaurant investors and small business owners is to understand what the food scene in the location they are interested in has capacity for. In this analysis, I will uncover which types of cuisines are the most saturated in Harlem, and suggest which cuisines are not represented and could be a good opportunity for the right investor.

The audience for the outcomes of this analysis is restaurant investors. A secondary audience is any small business restaurant owner in NYC and Harlem who wants to understand the market that they are competing in.

C. Personal Interest

Having lived in Harlem for a short time, my favorite memories are centered around the food I tasted. I tried many things there for the very first time. I also watched several new restaurants try to open, some more successfully than others. When many of these restaurants opened, it seemed like only new ideas were succeeding over ones that had many direct competitors. For example, new Italian or Dominican restaurants never seemed to last, and that may have been because there were already so many that were neighborhood favorites.

II. Data

A. Sources

I will be using a few data sources for this analysis. First, I will be using Foursquare's API and restaurant database to collect data on the types of restaurants in Harlem. I will be pulling results from all businesses in the geographic region that Harlem encompasses, with specific focus on the 'Venue Category' feature. Secondly, the geographic locations of each neighborhood can be pulled from this public dataset, containing the latitude and longitude of every zip code in the United States. Third, I will use popularity scores for categories of cuisine pulled from a Google Analytics survey. Finally, the neighborhoods which correspond with each zip code in Manhattan were pulled from this repository published by the NY State Department of Health.

B. Data Cleaning and Feature Selection

The US Zip Code dataset contained the latitude, longitude, zip, and state features needed to run the Foursquare query API. However, it did not indicate which neighborhoods each zip code corresponded to. The repository for the NY State of Health contained the zip and neighborhood features, so I could combine the two data sets to get every feature I need.

My first step was to perform a join operation on the US Zip Codes data with the Manhattan Neighborhoods Zip Code data set. The result was a merged data set of every zip code, latitude, and longitude value for every neighborhood in Manhattan. Then, I narrowed the dataset to just Harlem, which encompasses two sub-neighborhoods: "Central Harlem" and "East Harlem". I did this by filtering the combined data set 'Neighborhood' feature for any name containing 'Harlem'.

Next, using Foursquare's API capabilities, I collected data on businesses near every zip code in Harlem. For this analysis, I needed to know the count of each type of restaurant in each zip code. Therefore, I ran the query to comb a radius of 500 meters from the center point of each zip code, for the top 100 businesses nearby. Because I am only interested in the restaurant businesses, I ran a unique value filter on the 'Venue Category' column and narrowed my data down to only the businesses who are categorized relating to food service.

III. Methodology

A. Exploratory Analysis

First, I was curious to explore the total number of restaurants as compared to other businesses in Harlem. I found that compared to all other types of businesses, restaurants and food services make up over half of all operating venues. Today, restaurants make up 56.3% of all operating businesses, which is depicted in Figure 1.

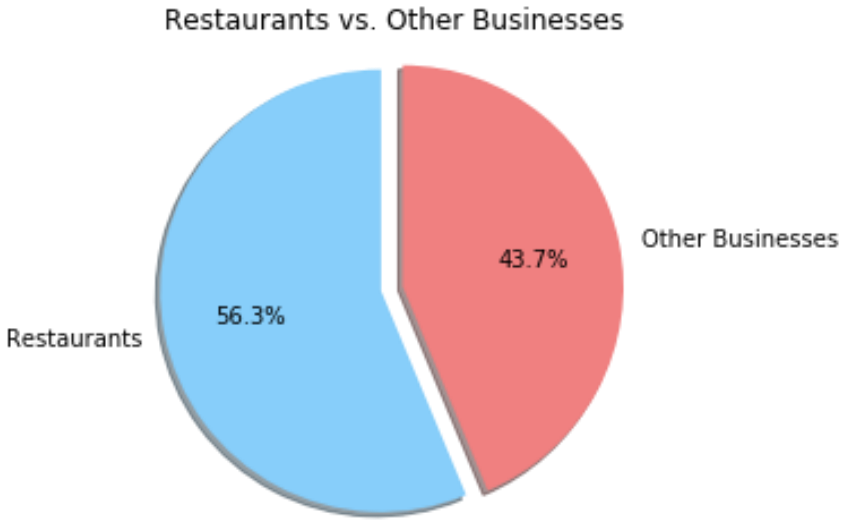

Figure 1. Comparison of restaurant industry stake in local business operations.

This confirms my hypothesis that investing in restaurants will be a competitive market, when considering the percentage of local businesses dedicated to the restaurant industry.

Once I had the dataset I wanted, I then used the one-hot encoding technique to get the frequency of each type of restaurant. By doing this, I was then able to get a count of each type of restaurant in Harlem, segmented by style of cuisine.

The outcome of this was a data frame (Figure 2) listing every type of cuisine represented in Harlem, including the count of each. This was the starting point for the rest of my analysis.

	Cuisine Type	Count
0	African Restaurant	5
1	American Restaurant	7
2	BBQ Joint	1
3	Bagel Shop	3
4	Bakery	5

Figure 2. Sample head of Harlem Restaurant Frequency data frame.

Next, I wanted to find out which service types of restaurants are the most common. Before determining which style of cuisine will be the most successful, it is useful for investors to know which styles of food service currently exist. To do this, I segmented each cuisine category by an additional feature called 'Service Style.' The options for this category were: 'Full Service Restaurant,' 'Fast Food,' and 'Fast-Casual.' In Figure 3, the distribution of each type of food service business is depicted.

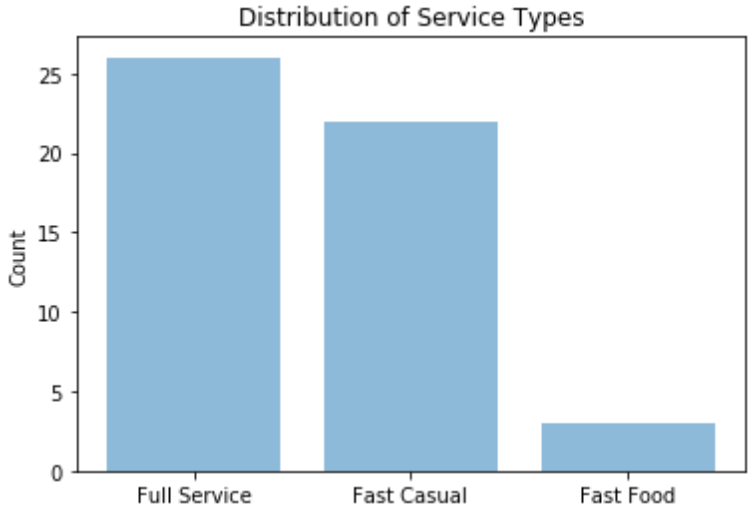

Figure 3. Count of each type of service style for restaurants in Harlem.

From this graph, you can see that the service types of most restaurants in Harlem is either Full Service or Fast Casual. This is an important observation, as it informs any potential investors that the market does not favor Fast Food chains, but does favor more formal and sit-down casual restaurants.

I then appended the popularity score of each type of cuisine to the dataset pulled for restaurants in Harlem. Using aggregate data from a Google Analytics report published, I was able to assign each category of cuisine a score, correlating to how popular that type of food is. The resulting table contained the features: Cuisine Type, Count, Service Type, and Popularity Score (Figure 4).

	Cuisine Type	Count	Service Type	Popularity Score
0	African Restaurant	5	Full Service	40.0
1	American Restaurant	7	Full Service	75.0
2	BBQ Joint	1	Fast Casual	50.0
3	Bagel Shop	3	Fast Casual	70.0
4	Bakery	5	Fast Casual	30.0

Figure 4. Snapshot of dataset and features for machine learning

From this, I can run a K-Means Clustering algorithm to group the types of cuisines and recommend which should be invested in.

B. Modeling

I ran my model off the features acquired in the data exploration phase and through feature engineering. I applied a K-Means algorithm to group similar cuisine types based on their representation in Harlem's neighborhood, their types of service, and their popularity scores. I transformed the Cuisine Types and Service Types into integer values and standardized so that the model would run accurately.

The resulting model grouped the neighborhoods into 5 clusters (Figure 5).

	Cluster Labels	Cuisine Type	Count	Popularity Score	Type	Service Type
0	0	African Restaurant	5	40.0	1	Full Service
1	4	American Restaurant	7	75.0	1	Full Service
2	0	BBQ Joint	1	50.0	2	Fast Casual
3	4	Bagel Shop	3	70.0	2	Fast Casual
4	1	Bakery	5	30.0	2	Fast Casual

Figure 5: Data including assigned clusters

These clusters can be summarized by looking at the most common features in each cluster. In cluster '0', the average popularity score was 42 across 12 restaurants. In cluster '2', the popularity score was on average 61 across 9 restaurants.

IV. Results

When taking a closer look at these two clusters, you notice that they represent the Full Service, Ethnic cuisines. This confirms the hypothesis that Harlem restaurant-goers tend to like to try new types of cuisines.

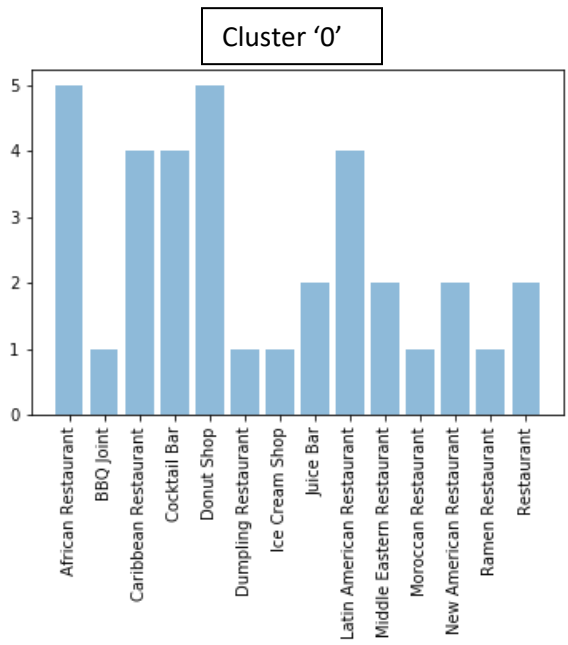

Figure 6: Cluster Zero count of popular restaurant types

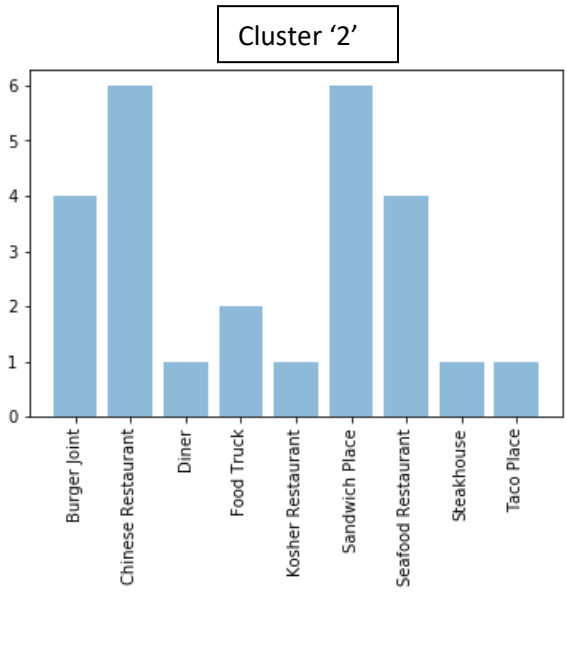

Figure 7: Cluster Two count of popular restaurants

The model helped to narrow down the type of cuisines that are most common and popular in Harlem, and categorized them so that they could be considered for investments. This is useful, but also requires further analysis to drill down on what our results are really telling us.

The distribution of each cluster's popularity score varies. Figure 8. Depicts the range of popularity scores for each cluster. We can see that in clusters two and zero the range and median is high.

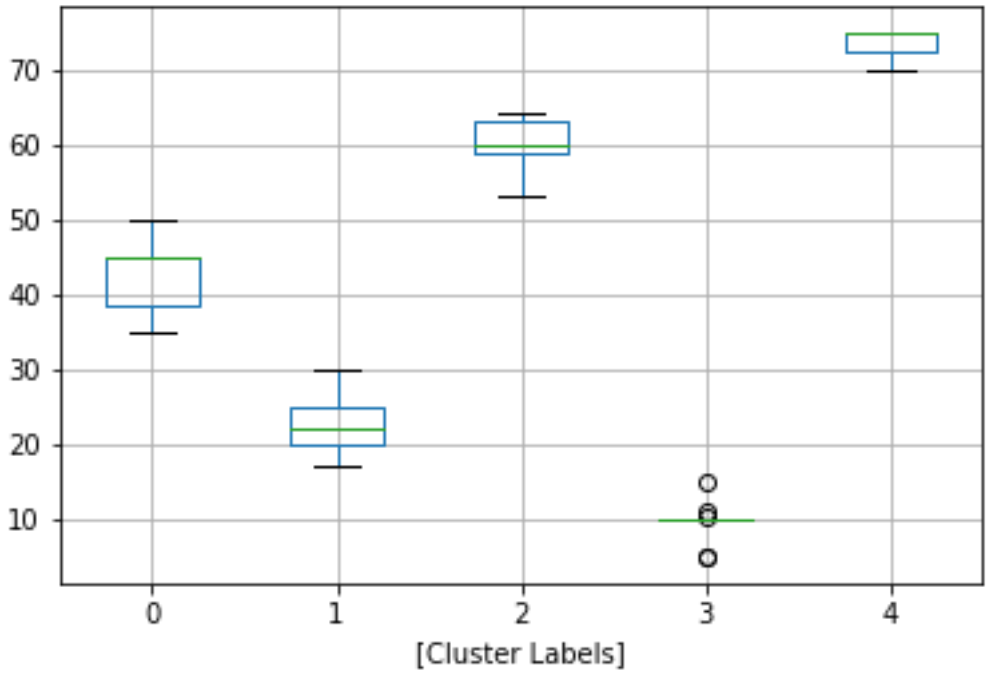

Figure 8. Popularity Score distribution by cluster.

It is of note that the cluster ‘4’ has a significantly high popularity score. When taking a closer look at this cluster, there are only 3 restaurants represented. For this reason, this cluster appears to be an outlier group of very popular restaurants, but not the most saturated cuisine types.

In Figure 9, the top 10 types of food service businesses are depicted. By taking a closer look at these, we can see which are the most highly saturated in the Harlem restaurant industry. In Figure 10, the popularity scores of the most highly saturated businesses are represented.

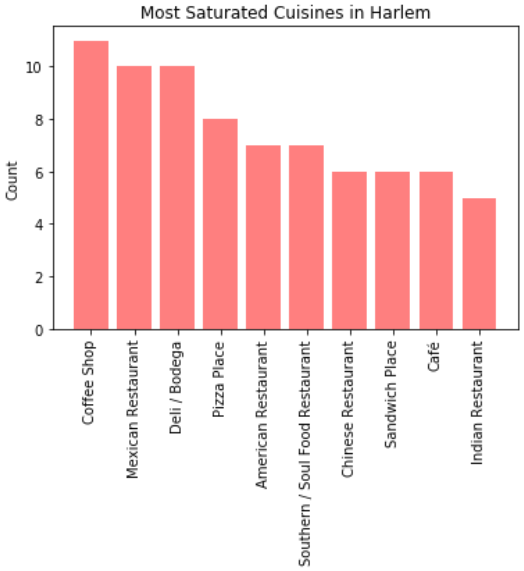

Figure 9. Count of most saturated cuisine types.

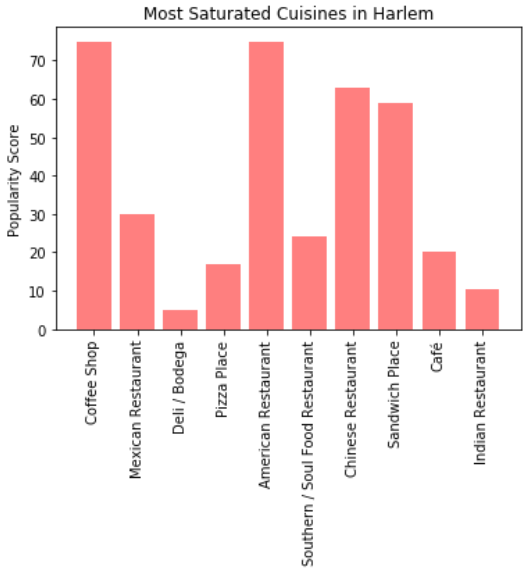

Figure 10. Popularity scores of cuisine types.

The most highly saturated type of food service business in Harlem is a Coffee Shop. That is followed by Mexican Restaurants, and Delis. These are also very high on the popularity score. These restaurants correlate to cluster '4'. Although they are very popular, there are already too many of these types of restaurants to compete with. It would not be the best choice for a restaurant investor or new business owner to open a restaurant falling into these categories. For this reason, cluster 4 will not be considered for the best place to invest.

Examining the clusters for popularity score, the top popularity scores are depicted in figure 11. We can see that most of these cuisine types also appear in the cluster we selected for: Cluster 2.

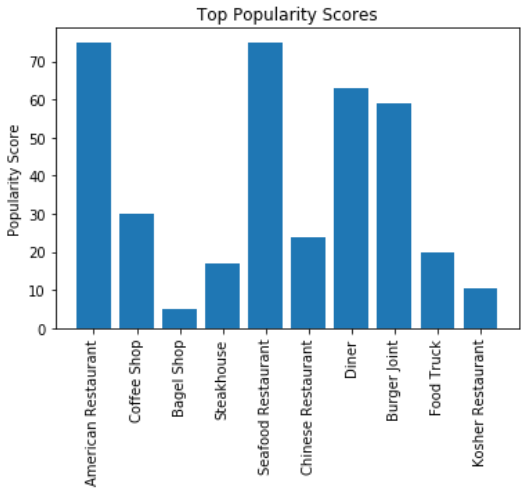

Figure 11. Cuisines by Popularity Score

Having assessed both the saturation of each cuisine type, and the popularity score of each cuisine type, we can conclude that the restaurant types in Cluster 2 (figure 12) are the best options for new investors. These have high popularity scores, have a presence in Harlem, and are not over saturated already.

	Cluster Labels	Cuisine Type	Count	Popularity Score	Type
7	2	Burger Joint	4	60.0	2
10	2	Chinese Restaurant	6	63.0	1
16	2	Diner	1	63.0	2
23	2	Food Truck	2	60.0	3
31	2	Kosher Restaurant	1	59.0	1
40	2	Sandwich Place	6	59.0	2
41	2	Seafood Restaurant	4	64.0	1
45	2	Steakhouse	1	64.0	1
46	2	Taco Place	1	53.0	2

Figure 12. Cluster Two Data Frame

V. Discussion

As stated previously, Harlem is a unique hub of great food in New York City. It is extremely diverse, and highly populated with restaurants. In order to recommend to potential restaurant investors which types of restaurant cuisines will have a likelihood of succeeding, I ran an analysis on the current restaurant market.

I began by considering the total density of restaurants as it compared to other businesses in Harlem. Restaurants account for over half of businesses, so are clearly very popular. I used the Kmeans algorithm to cluster restaurants based on their service type, cuisine, and number of existing businesses. If this analysis was to be repeated, I would add additional features to better inform the machine learning algorithm. I believe with additional features, this would increase the accuracy of the model and could narrow the clusters, leading to better decision making.

I also performed exploratory analysis on the top types of restaurants, by each of these features. By cross referencing the clusters produced in my algorithm with the data on which restaurant types are very popular or overly saturated, I was able to narrow my focus to two main clusters: 0 and 2. I then compared these two, focusing on saturation of cuisine and popularity. I concluded that although both are popular, cluster 0 is too highly saturated in the existing market and will likely be too difficult to compete with.

VI. Conclusions

In this study, I examined the Harlem restaurant industry to determine which types of restaurants, by cuisine, a potential investor should consider. I identified cuisine, popularity score, and service type as the main features that can group restaurants. I ran an unsupervised K-Means clustering algorithm to determine which types of restaurants were most similar to each other based on these features. From there, I examined each cluster to determine its likelihood to be over saturated or low on the popularity score index. I was able to reduce the 5 clusters down to just one, cluster 2, which is the best option for new investors. The restaurant-types in cluster 2 have high popularity scores, but are not already over saturated. They exist already so they have been proven to be successful in the market. Those restaurant types are: Burger Joints, Chinese Restaurants, Diners, Food Trucks, Kosher Restaurants, Sandwich Shops, Seafood restaurants, Steakhouses, and Taco Places.