Applied Data Science Capstone
Final Assignment

FINDING A LOCATION TO OPEN A RESTAURANT IN MANHATTAN

Leobardo Gómez

Introduction

As an entrepeneur in the food business, it is always important to find a good location to open a restaurant.

Given that New York City is one of the biggest cities in the world, it is very difficult to find the best place to start a new business. Manhattan is one of the most important boroughs in New York City, because it is very densely populated and for the fact that is the economic, administrative and cultural center of the city. This makes Manhattan and ideal spot for starting a restaurant.

Data Science is the use of data to gain insights about a given problem. Described briefly, it is the process to answer questions about different areas with the use of information.

Business Problem

Is it possible to find a good location in Manhattan to start a resturant?

The goal of this paper is to use Data Science to discover the answer to the previous question. So, the target audience for this project are entrepeneurs with the idea of starting a restaurant in Manhattan.

Data

To find a good location to start a restaurant is a complex task that requires some analysis. Based on the web article:

https://fitsmallbusiness.com/choose-a-restaurant-location/

It is possible to identify some important features for choosing the right place:

- Restaurant concept
- Type of ideal customer

As the article describes it, the Restaurant concept is based on the combination of *cuisine type* and *restaurant style.* These features are going to be essential in the process of clustering the information.

For the type of ideal customer, some demographic information is needed.

The following page holds information about population per neighborhood:

https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp

The file _New_York_City_Population_By_Neighborhood_Tabulation_Areas_ has some data to filter neighborhoods:

	Borough	Year	FIPS County Code	NTA Code	NTA Name	Population
0	Bronx	2000	5	BX01	Claremont-Bathgate	28149
1	Bronx	2000	5	BX03	Eastchester-Edenwald-Baychester	35422
2	Bronx	2000	5	BX05	Bedford Park-Fordham North	55329
3	Bronx	2000	5	BX06	Belmont	25967
4	Bronx	2000	5	BX07	Bronxdale	34309

The following URL contains information about groups of age in New York, divided by Borough:

https://www1.nyc.gov/site/planning/planning-level/nyc-population/census-2010.page

From this site, it's possible to download information about age.

The file *totpop_5yrgrps_nta* contains the information in CSV format:

	Borough	FIPS County Code	Code	Name	Total Population	Under 5 Years	5-9 Years	10-14 Years	15-19 Years	20-24 Years	25-29 Years	30-34 Years
0	Bronx	5	BX01	Claremont-Bathgate	31078	2890	2816	2769	3113	2658	2325	2012
1	Bronx	5	BX03	Eastchester-Edenwald-Baychester	34517	2225	2436	2568	3194	2684	2344	2064
2	Bronx	5	BX05	Bedford Park-Fordham North	54415	4517	4183	4058	4623	4693	4663	4262
3	Bronx	5	BX06	Belmont	27378	2076	2073	1969	3458	3937	2157	1846

Also, it is important to take in account the type of business around the area.
Foursquare datasets will be needed to get insight about the nearby venues.

Methodology

Based on the web article:

https://fitsmallbusiness.com/choose-a-restaurant-location/

It is possible to determine the features for the analysis. The article suggest that the required data to determine a good location could be:

- Demographic information, specifically about age and economics of the population
- Types of venues in the location
- Restaurant concept (meaning the type of cuisine and the style of the business)

The *demographic information* from the web resources (downloaded in CSV format) could be segmented in three categories or age groups of population; the *types of venues* were classified using the information available from FourSquare.

It is important to notice that the age groups and the types of venues were classified based on some simplified criteria:

- The *age groups* were segmented by
  - Young – 15 – 29 years
  - Middle Age – 30 – 49 years
  - Old – 50 + years
- The *types of venues* were segmented by words contained in the category information (downloaded in JSON format). The categories were defined as: Bar, Entertainment, Fast Food, Health/Sports, Restaurant, Spots of Interest and a category for anything else not considered (other)

Merging both datasets, will create the structure needed for the classification:

	Name	Code	Young	Middle Age	Old	Bar	Entertainment	Fast Food	Healt/Sports	Other	Restaurant	S
0	Marble Hill	MN01	10936	14038	13650	0	0	2	3	4	4	
1	Hamilton Heights	MN04	13227	14568	13146	7	0	3	2	20	19	
2	Manhattanville	MN06	6298	6356	6063	2	1	0	1	13	17	
3	Morningside Heights	MN09	22290	13326	14520	2	0	4	2	14	8	
4	Upper West Side	MN12	22909	41259	50714	9	0	2	4	16	29	

The **K-Means** algorithm was selected because of its simplicity to cluster information into groups with similar features. The number of clusters selected were 6.

Results

The clustering with **K-Means** grouped the information into 6 different clusters:

Cluster	Age	Type of venue
Cluster 1	OLD & MIDDLE AGE groups	1. **Restaurants** are the most common 2. **Other** & **Shopping** venues are popular
Cluster 2	OLD group	1. **Restaurants** are the most common 2. **Other** & **Shopping** venues are popular
Cluster 3	Balanced in all groups, with more OLD	1. **Other** venues are the most common 2. **Restaurant** & **Shopping** venues are popular
Cluster 4	YOUNG group	1. **Restaurant** venues are the most common 2. **Other** & **Shopping** venues are popular
Cluster 5	YOUNG & MIDDLE AGE groups	1. **Shopping** venues are the most common 2. **Spots of interest** & **Other** venues are popular
Cluster 6	MIDDLE AGE group	1. **Other** venues are the most common 2. **Restaurant** & **Shopping** venues are popular

Discussion

The web article for finding a good spot for a restaurant also indicates some evaluation in the economic data. Despite it was possible to obtain information about age (census of NYC) and types of venues (FourSquare), it is not so simple to get all the data about economics.

Some web datasets have the information about income per household or mean income per district, but the sources are not condensed or are not in the required level (some information is grouped by borough).

I think that the segmentation by the **K-Means** could be easier or more refined if the economic data is available by neighborhood.

Conclusion

The web article:

https://fitsmallbusiness.com/choose-a-restaurant-location/

Gives some guidelines to choose the correct spot for open a restaurant. These guidelines could be related to the CLUSTERS that were obtained with the **K-Means** algorithm.

There are some considerations to take in account depending on the type of restaurant.

Type of Restaurant	Description	Related cluster
Fast Food	15 – 35 age range. Regularly is connected to another activity like shopping or activities with friends	CLUSTER 4 – The age group is similar, and there are not much fast food places in this cluster. Also, there are shopping venues around.
Bar & Bistro	25-45 age range. Takes places after work, relaxed social ambiance. The customer could drink alcohol	CLUSTER 5 – The age group is similar and there are not much competitors around. CLUSTER 6 – The age group is somewhat similar
Casual Dining	Families with children, often connected to another activity like shopping or going to movies	CLUSTER 3 – This cluster has a balanced population in all groups of age. It could be possible to have customers (families). There are shopping and other venues nearby

		CLUSTER 5 – The majority of the population is young and middle age people. There are shopping venues around
Fine Dining	35 years or more, couples and executives. High price point	CLUSTER 1 – The population is primarily in 35 + range of age. This could mean higher income as well CLUSTER 2 – The population is in the 50 + range of age.

Finally, based on the previous table, each cluster could be mapped to a neighborhood in Manhattan.

The initial question posed at the beginning of this paper:

Is it possible to find a good location in Manhattan to start a resturant?

Could be answered if the type of restaurant is given:

Type of Restaurant	**Related cluster**	**Location/Manhattan neighborhood**
Fast Food	CLUSTER 4	Murray Hill, Gramercy, East Village
Bar & Bistro	CLUSTER 5	Battery Park City
	CLUSTER 6	Hudson Yards, Lincoln Square, Clinton
Casual Dining	CLUSTER 3	Marble Hill, Manhattanville, Stuyvesant Town
	CLUSTER 5	Battery Park City
Fine Dining	CLUSTER 1	Morningside Heights, Upper West Side, West Village, Lower East Side, Lenox Hill, Yorkville
	CLUSTER 2	Hamilton Heights, Midtown, Turtle Bay, Chinatown, Upper East Side