Investing in a snack place in New York

Simone Pala

23rd January 2021

1 Introduction and business problem

A restaurant chain is willing to diversify its business and open a new branch to serve the need of commuters in New York, so the question is "*where are the best places to open snack place shops in New York*"?

Marketing department describes the average profile of a commuter person as interested in a fast or "grab & go" meal to avoid spending time waiting at a table and have a meal while travelling.

The business owners are considering opening shops in proximity of New York metro stations, so they'd like to know the presence of competitors in line with the marketing department indications (coffee shops, sandwich, fast food shops...) as well as the potential market size.

2 Data: description and use in the problem solution

The following data are retrieved:

- New York public transportation (subway and bus) system is managed by MTA. Subway stations coordinates are made available on the MTA website https://new.mta.info.
- MTA provides also the average number of people passing for each station (number of transits) of the last 6 years
- Thanks to the Foursquare APIs a list of existing shops in the proximity (radius of 100m) of subway stations

The list of subway stations are used nodes for investigation, for each station:

- it is possible to compare the potential profitability against other ones thanks to the data on average transits
- it is possible to evaluate the number of competitors thanks to information from Foursquare
- The information will be aggregated in a dataframe and shown in a map, furthermore the dataframe will be elaborated with machine learnings to classify the spots into groups and shortlist the subway stations to be considered as location for investment by the restaurant chain.

2.1.1 Details on the data

2.1.1.1 Stations data

Among the fields included in the stations data file (source: http://web.mta.info/developers/data/nyct/subway/Stations.csv) the followings are going to be used:

- Station ID - unique identifier of the Station
- Stop Name - Name of the Subway station
- Borough - Name of the Borough

- GTFS Latitude - Latitude of the station
- GTFS Longitude - Longitude of the station

The Latitude and Longitude are the used information.

2.1.1.2 *Number of transits*

The data are provided in an excel file - source: https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2019

The information provided include the yearly number of transits from 2014 to 2019, last year change and the 2019 rank. The data are organized in three tabs:

- number of transits for the average weekday per subway station
- number of transits for the average weekend per subway station
- total annual number of transits per subway station

An additional data tab lists which and when some of the subway stations were temporarily closed.

For the purpose of the study only the number of transits for the average weekday in 2019 and for the average weekend in 2019 will be considered and imported into a dataframe.

2.1.1.3 *List of existing shops from Foursquare APIs*

Thanks to the Foursquare API is possible to retrieve the following info for each venue in a predefined radius (I'll use 100m) from the subway station:

- id - it is a unique identifier for the venue
- name - name of the venue
- location with address, latitude and longitude - address of the venue and geospatial coordinates
- category - category which the venue belongs to e.g. hotel, bar, restaurant

3 Methodology

section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

3.1 Data import and cleaning

3.1.1 Transits data

Transits data from the above-mentioned sources have been imported and the following operations have been performed to clean data:

1. Drop of column not useful for this study (e.g. % variations)
2. Data casting to the right data type, removal of NaN values and renaming of columns names
3. Grouping of stations with multiple lines as one

This last operation was necessary as the transits data reports the number of passengers per each station divided by metro lines, in other words, there are multiple lines for the same station when multiple subway lines have a stop.

Three data-frames are created to describe the number of transits for the average weekday, for average weekend and total annual number per each year since 2014.

3.1.2 Stations data

Similarly, stations data are imported, cleaned, merged and renamed to have as results a table with *Station name*, *Latitude* and *Longitude*

e.g.

	Station	GTFS Latitude	GTFS Longitude
0	1 Av	40.730953	-73.981628
1	103 St	40.795379	-73.959104
2	103 St-Corona Plaza	40.749865	-73.862700
3	104 St	40.688445	-73.841006
4	110 St	40.795020	-73.944250
...	...	...	...
374	Woodlawn	40.886037	-73.878751
375	Woodside-61 St	40.745630	-73.902984
376	World Trade Center	40.712582	-74.009781
377	York St	40.701397	-73.986751
378	Zerega Av	40.836488	-73.847036

3.1.3 Locations data

A script was created to query Foursquare (via API) and retrieve all the venues in a radius of 100m from each train station, an excerpt of the results is as shown in the table below:

	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	1 Av	40.730953	-73.981628	Hawa Smoothies & Bubble Tea	40.730950	-73.981545	Juice Bar
1	1 Av	40.730953	-73.981628	Trader Joe's	40.730828	-73.980955	Grocery Store
2	1 Av	40.730953	-73.981628	Veeray Da Dhaba	40.730784	-73.982716	Indian Restaurant
3	1 Av	40.730953	-73.981628	Lower East Side Coffee Shop	40.730468	-73.980657	Diner
4	1 Av	40.730953	-73.981628	Domino's Pizza	40.730343	-73.980757	Pizza Place

Data was furtherly elaborated so to count the number of the venues classified by *Venue category*. The venue categories were classified into two macro classes: "*Snack places*" and "*Restaurants*", these represents the possible competitors for the business. The result is the following:

Station	Snack place	Restaurant
1 Av	1	3
103 St	0	0
103 St-Corona Plaza	5	3
104 St	0	0
110 St	4	1
...	...	...
Woodhaven Blvd	1	0
Woodlawn	2	0
Woodside-61 St	4	5
York St	0	0
Zerega Av	1	2

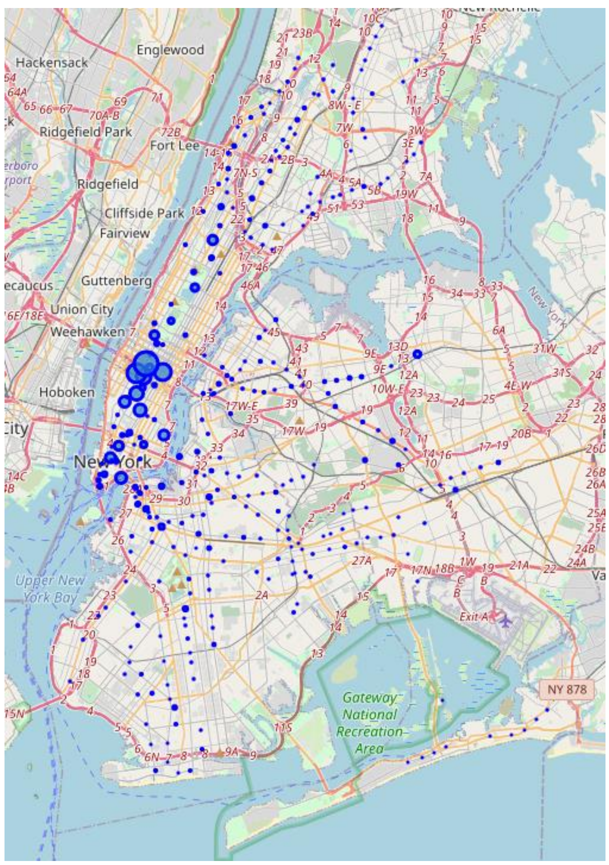
3.2 Merging transits and locations data

Data-frames with stations geographical coordinates and transits data have been merged to obtain a single data-frame containing the station names, their coordinates and transits data. E.g.:

	Station	GTFS Latitude	GTFS Longitude	average_weekday	average_weekend	total_annual
0	1 Av	40.730953	-73.981628	18393.11020	12273.23080	5345371.0
1	103 St	40.795379	-73.959104	9982.05510	10230.49360	9303988.0
2	103 St-Corona Plaza	40.749865	-73.862700	19943.12990	24170.55770	6399657.0
3	104 St	40.688445	-73.841006	2275.30115	1408.16345	1311812.0
4	110 St	40.795020	-73.944250	10579.86610	11426.80770	3316061.0

A preliminary visualization on a map to show the location of each station has been created. Each station is represented by a circle, the largest the circle the higher the annual transits of passengers. See image below.

The map below shows that most transits are happening in New York downtown, more precisely in Manhattan.



3.2.1 Exploratory data analysis and statistical testing

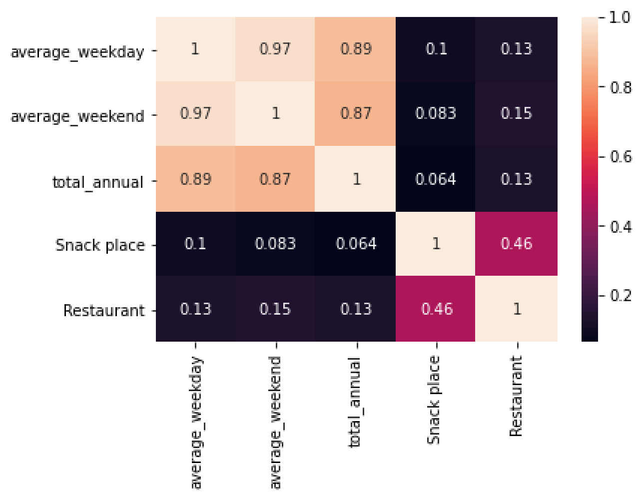
The data-frames prepared as explained in the previous sections were merged, the result is a single data-frame containing stations data with coordinates, traffic and number of nearby food-serving venues. E.g.

	Station	GTFS Latitude	GTFS Longitude	average_weekday	average_weekend	total_annual	Snack place	Restaurant
0	1 Av	40.730953	-73.981628	18393.11020	12273.23080	5345371.0	1	3
1	103 St	40.795379	-73.959104	9982.05510	10230.49360	9303988.0	0	0
2	103 St-Corona Plaza	40.749865	-73.862700	19943.12990	24170.55770	6399657.0	5	3
3	104 St	40.688445	-73.841006	2275.30115	1408.16345	1311812.0	0	0
4	110 St	40.795020	-73.944250	10579.86610	11426.80770	3316061.0	4	1
...	...	...	...	...	...	...	...	...
293	Woodhaven Blvd	40.713493	-73.860402	12473.82680	12498.74040	7718919.0	1	0
294	Woodlawn	40.886037	-73.878751	6679.57870	7255.34610	2094285.0	2	0
295	Woodside-61 St	40.745630	-73.902984	16683.89370	20097.03850	5345369.0	4	5
296	York St	40.701397	-73.986751	12638.32680	13023.86540	3927129.0	0	0
297	Zerega Av	40.836488	-73.847036	2676.38190	2099.75000	795756.0	1	2

Thanks to the correlation matrix, it's possible to understand that:

- There is a good correlation (close to 0.9) among *total annual passenger*, *average weekend* and *weekday* transits.
- There is a small correlation (0.46) between *Snack place* and *Restaurant*.

This can indicate that train stations in proximity of a high number of snack places are normally not the same close to restaurants.



An analysis of statistical indicators (after minimum-maximum scaling) reveals that:

- Stations have a mean of 0.072 total annual passenger and 0.18 snack places
- The 50th percentile of stations have 0.034 total annual passenger and 0.1 snack places

	average_weekday	average_weekend	total_annual	Snack place	Restaurant
count	298.000000	298.000000	298.000000	298.000000	298.000000
mean	0.058079	0.051137	0.072873	0.181208	0.172260
std	0.095782	0.086046	0.117517	0.176715	0.207093
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.019745	0.016651	0.019437	0.000000	0.000000
50%	0.032620	0.027627	0.033972	0.100000	0.111111
75%	0.060720	0.056660	0.071776	0.300000	0.222222
max	1.000000	1.000000	1.000000	1.000000	1.000000

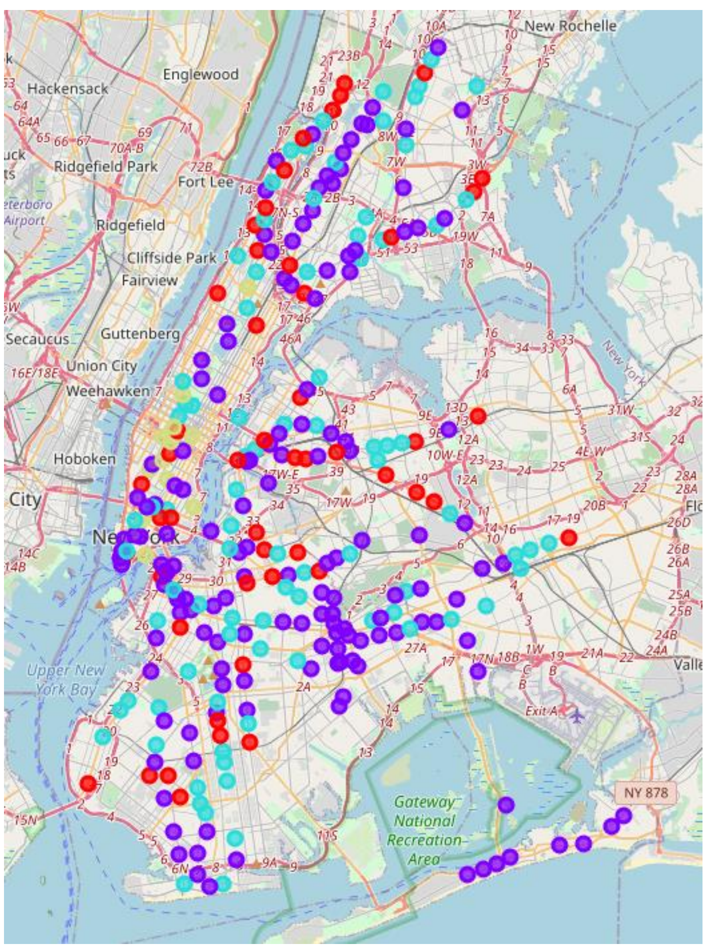
3.3 Clustering stations

A *k-mean* algorithm to cluster stations into 4 groups was used. The variable used for the clustering are:

- average_weekday
- average_weekend
- total_annual
- Snack place

4 Results

The algorithm has assigned each station to a cluster, the results can be visualized on the map below:



- Cluster 0 is with red circles
- Cluster 1 is with purple circles
- Cluster 2 is with light blue circles
- Cluster 3 is with yellow circles

The following table summarizes the mean and 50th percentile for each of the 4 clusters, the table includes also the reference mean and 50th percentile of all the stations as a reference and facilitate the comparison:

Cluster no.	**count**	***Reference Mean total_annual***	**Mean total_annual**	***Reference Mean Snack place***	**Mean Snack place**	***Reference 50% percentile Mean total_annual***	**50% percentile Mean total_annual**	***Reference 50% percentile Snack place***	**50% percentile Snack place**	**Comments**
0	50	***0.073***	0.068	***0.181***	0.488	***0.034***	0.049	***0.100***	0.450	Medium transits, high snack places
1	149	***0.073***	0.046	***0.181***	0.046	***0.034***	0.021	***0.100***	0.000	Low transits, low snack places
2	86	***0.073***	0.052	***0.181***	0.237	***0.034***	0.031	***0.100***	0.200	Low transits, high snack places
3	13	***0.073***	0.537	***0.181***	0.177	***0.034***	0.453	***0.100***	0.100	High transits, medium snack places

5 Discussion

The data presented and summarized in the previous section can be read as in the following comments:

- Cluster 0 (*red*) has a total annual transit (0.068) in line with the reference value (0.073), nearby snack places are in average higher (0.488) than the reference (0.181). 50th percentile of snack places is higher than its reference value.
- Cluster 1 (*purple*) has a total annual transit (0.046) lower than the reference value (0.073), nearby snack places are in average lower (0.046) than the reference (0.181)
- Cluster 2 (*light blue*) has a total annual transit (0.054) lower than the reference value (0.073), nearby snack places are in average higher (0.237) than the reference (0.181). 50th percentile of snack places is higher than its reference value.
- Cluster 3 (*yellow*) has a total annual transit (0.537) higher than the reference value (0.073), nearby snack places are in average higher (0.177) than the reference (0.181).

Therefore, it is possible to conclude that stations in:

- *cluster 0 have a medium number of transits and a high number of snack places*
- *cluster 1 have a low number of transits and a low number of snack places*
- *cluster 2 have a low number of transits and a high number of snack places*
- *cluster 3 have a high number of transits and a medium number of snack places*

the following table allows an easy-to-read representation:

Cluster	Transits vs reference	Snack place vs reference
0	Medium	High
1	Low	Low
2	Low	High
3	High	Medium

5.1 Business suggestions

It is possible to conclude that the suggestion is to invest in stations where transit of people is higher and the competitors are less than the norm, this indicates that **investment in businesses close to stations in cluster 3 (yellow) should be preferred**.

Locations in cluster 0 (red) can also be considered, but it is to be considered that a higher competition from other businesses need to be won.

Perhaps, given the centrality of the location in cluster 3, real estate rent/buying prices are higher than other locations and the availability reduced.

5.2 Analysis limitation and future directions

The presented analysis has the following limitations, future improvements can address them:

- the time of the day when people commutes is not taken into account, indeed people may be more willing to buy food in certain time of the day (e.g. breakfast and lunch time)
- in some stations the flux of people may be biased (e.g. all people tend to walk in a specific direction) while the analysis is collecting data of venues on a radius of a station
- transits data of some stations were merged because they have multiple exits. One exit can be more used than other ones
- this study is considering all the snack place with the same importance, but perhaps a further filtering, and a subsequent qualitative study may produce more precise business recommendations.

6 Conclusion

In this study it is presented the analysis of possible locations in New York where to invest in a snack place business.

The number of total transits per each subway station are retrieved from New York's subway operator, along with stations coordinates. These data are used to determine the potential customer demand of each location.

Foursquare API is used to retrieve the number of venues in the proximity of each station, the venues are then filtered and classified to find the potential competitors in each area.

Finally, a k-means classification was used to cluster stations into 4 groups based on their volume of passengers (customer base) and the number of existing snack places (competitors). This allowed to produce business suggestions.