Investing in a snack place in New York

Simone Pala

Applied Data Science Capstone

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 business problem can be summarized as
  - ***where are the best places to open snack place shops in New York?***

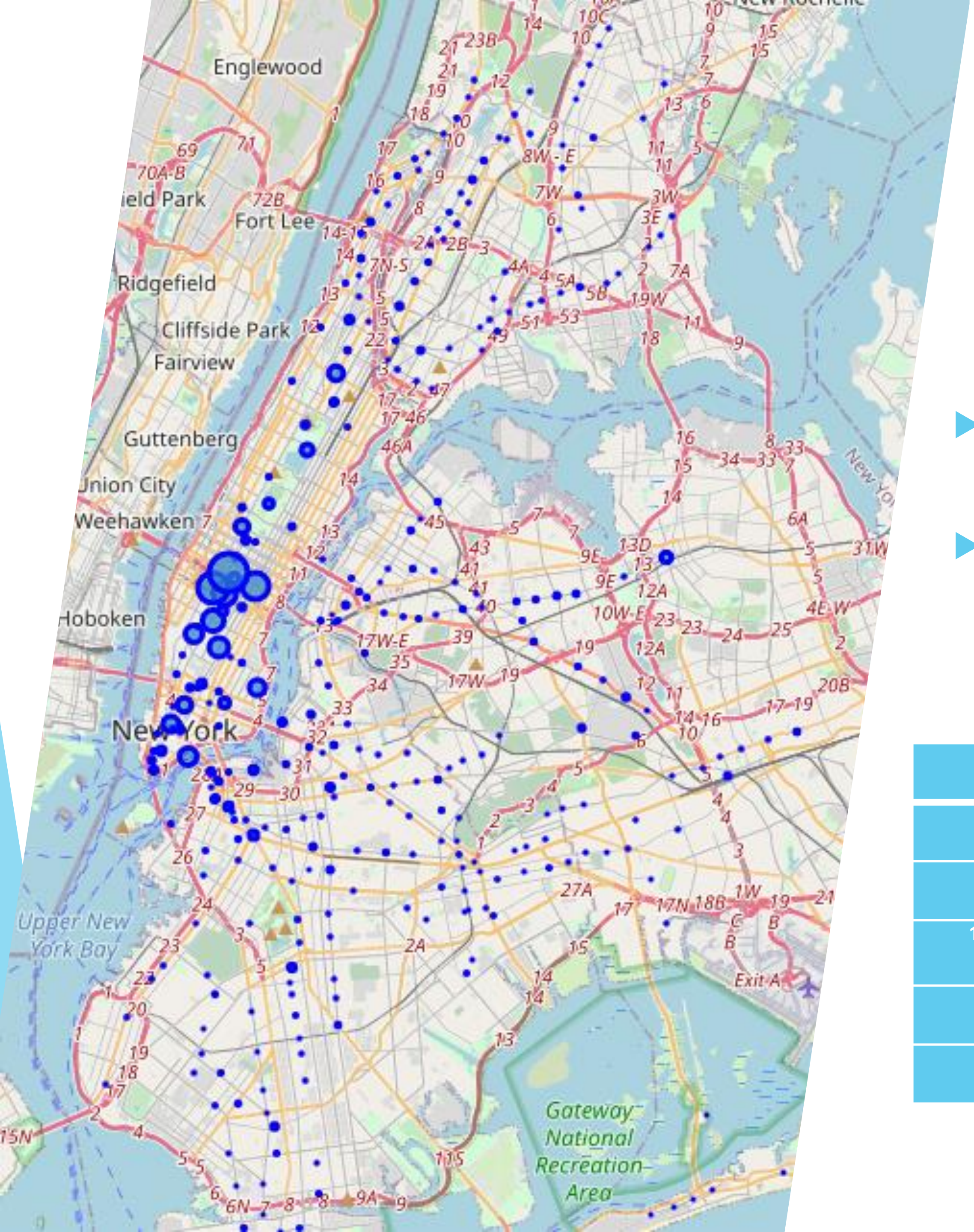
Data: description and use in the problem solution

- Subway stations coordinates are made available on the MTA website, MTA is New York’s public transportation managing company - https://new.mta.info
- MTA provides transits data as the average number of people passing for each station (number of transits) of the last 6 years
- Foursquare APIs are used to retrieve the list of existing shops in the proximity (radius of 100m) of subway stations

- **Station's data and transits -> Potential customer base of the snack places**
- **Existing snack places in a radius of 100m from stations -> Potential competitors**

Data: merging and cleaning

- Transit data were imported, the following operations were 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 (necessary as the transits data reports the number of passengers per each station divided by metro lines)
- Stations data have been imported and:
  1. Useless columns dropped and removal of NaN
  2. Grouping of stations with multiple lines as one (necessary as the transits data reports the number of passengers per each station divided by metro lines)



Data: merging and cleaning

- Excerpt of cleaned and merged data
- Visualization on map (the larger the dot the higher the **total annual transit**)

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

Venues data

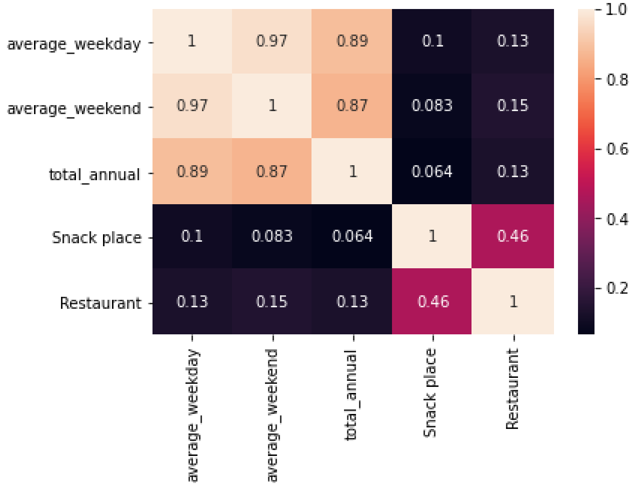
- A list of all the venues in a radius of 100m from the station were retrieved via Foursquare API, then:
  1. Data was furtherly elaborated so to count the number of the venues classified by *Venue category*.
  2. The venue categories were classified into two macro classes: “*Snack places*” and “*Restaurants*”, these represents the possible competitors for the business.
  3. For each station the number of “Snack place” and “Restaurant” is obtained

	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

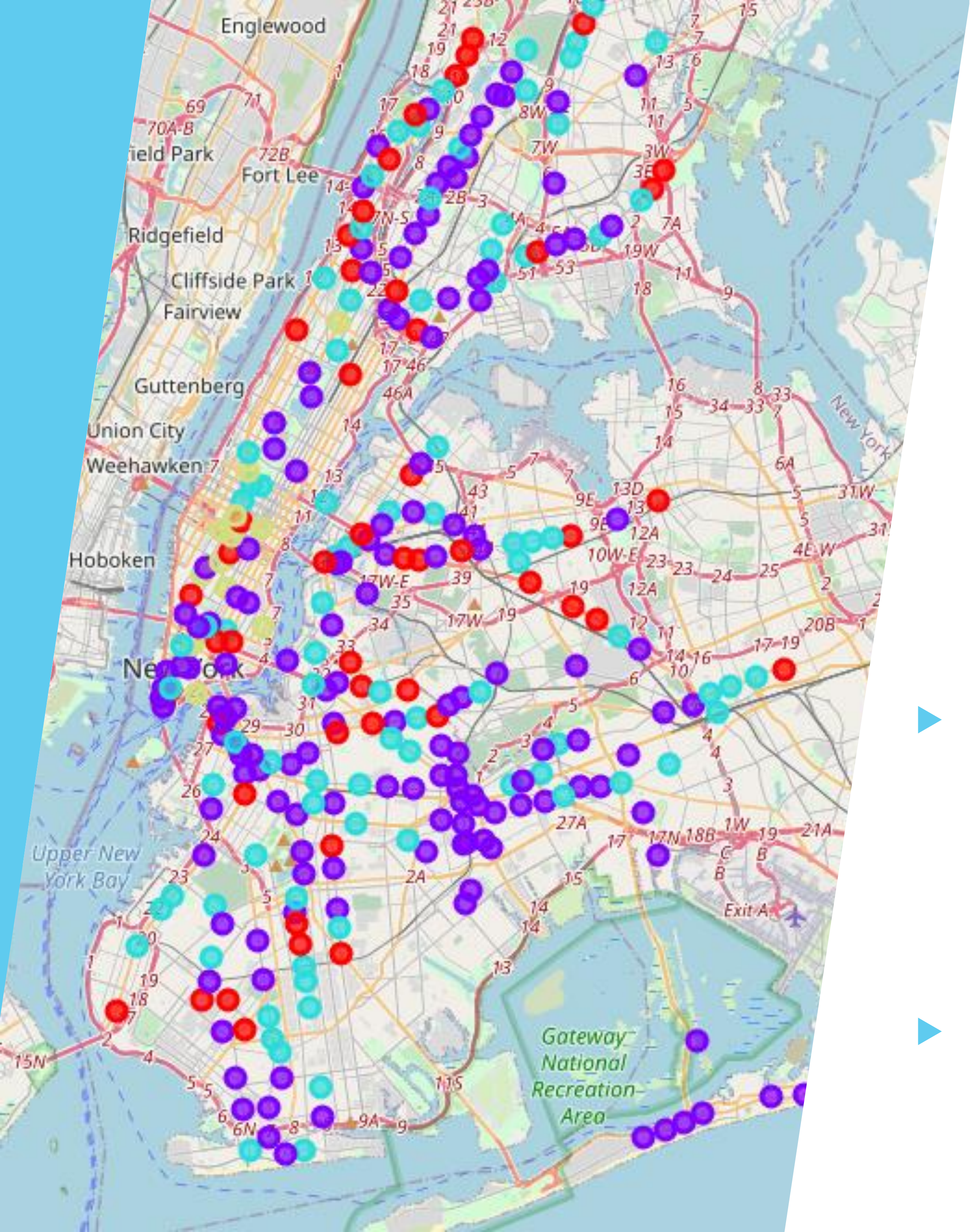
Station	Snack place	Restaurant
1 Av	1	3
103 St	0	0
103 St-Corona Plaza	5	3
104 St	0	0
110 St	4	1
...	...	...
Woodhaven Blvd	1	0
Woodlawn	2	0
Woodside-61 St	4	5
York St	0	0
Zerega Av	1	2

A single dataset to analyze: merging the data

- The venues and stations data were merged, the result is a single data-frame containing stations data with coordinates, traffic and number of nearby food-serving venues
- 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*



	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
…	…	…	…	…	…	…	…	…
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



Clustering with k-mean algorithm

Cluster	Marker	Transits volume vs reference	Snack place vs reference
0	Red	Medium	High
1	Purple	Low	Low
2	Light Blue	Low	High
3	Yellow	High	Medium

- Suggestion is to invest in stations where transit of people is higher and the competitors are less than the norm
  - **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.

Conclusion and future directions

- Preprocessed, cleaned and merged data
- Built a classification to suggest where to invest in snack places targeting for commuters
- Limitations and potential future improvements:
  - the time of the day when people commutes is not taken into account, transits may be biased in certain time of the day
  - 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 of multiple exits, one exit can be more used than other ones
  - snack places from Foursquare have been considered with equal importance, a further filtering and qualitative study may produce more precise business recommendations