A Site Survey of Coffee Shops for NYC Subway Riders

Yongkang Liu

January 07, 2020

Contents

1 Executive Summary

Using the New York subway passenger data and Foursquare venue information, the relationship between the station ridership and the coffee shop number is studied. Main takeaways are summarized as follows

- There exists a relatively strong positive relationship between the station ridership and the number of coffee shops within a walking distance (300 meters) from the stations.
- Geolocation factor plays a primary role to determine the number of coffee shops around subway stations. The stations are clustered with different patterns of the ridership and coffee shop density.
- The data indicates that Starbucks stores exhibit similar site distribution as peers along subway rails, however, they achieve a very effective site selection regarding covering around 90% of all subway riders with a relatively small number of sites.

2 Introduction/Business Problem

New York City (NYC) is well known for its fast-paced city life. As it is well known, NYC has one of the busiest public subway systems in the world. According to the Metropolitan Transportation Authority (MTA), the public transportation administrator in NYC, the annual subway ridership was around 1.757 billion in 2016 [1]. In a weekday it can serve over 5.7 million rides. Besides using the 24-7 continuing subway service, millions of commuters and travelers are also served by numerous coffee shops around the city. It is said that NYC has more coffee shops and cafés than any place else in the U.S. [2] Given these two classes of the most visited venues in NYC, I would like to use data science tools to explore their relationships in this capstone project. This study can serve as part of the site survey for new coffee shop business near subway stations serving their passengers.

In order to better use data to find the business opportunities along subway rails, a question to be asked is whether there is a relationship between the subway ridership and the coffee shop number in the neighborhood.

3 Methodology

In the project, I adopted a methodology to classify the coffee shop market around NYC subway stations to identify specific patterns in clusters of station neighborhoods.

3.1 Overview of Data Sources

To answer the above question, the required data includes the subway ridership information and venue information based on locations. For subway ridership, the features of each station include the identification information of stations, the counts of subway passengers in and out of each station, date and time information regarding the collected ridership in a repeated manner, and geolocation information of each station. For the venue information, each data record should contain the unique id of the venue, categorical labels, and geolocation profiles.

I will mainly use online public data in this project. To meet these requirements, a few data sources have been located including the weekly ridership data and station profile data from MTA, venue database from Foursquare. A brief summary of data sources is shown below.

1) NYC subway station data
   Source: http://web.mta.info/developers/data/nyct/subway/Stations.csv
   Format: csv
   Feature: This file contains all MTA station information including station name, service lines/division, and location (latitude and longitude).
   Usage: This project will use the data to determine the location information (latitude and longitude) of each subway station in the Manhattan island.

2) NYC subway turnstile data
   Source: http://web.mta.info/developers/turnstile.html
   Format: txt/csv
   Feature: Turnstile data contains the passenger counts (enters and exits) at each turnstile of MTA subway stations. The counts are collected every 4 hours from each turnstile unit in a 24x7

schedule. Each record occupies one line in the txt file with column values separated by commas. Every txt file contains all counts in a single week of 7 days.
Usage: The data will be used to label the passenger density in different subway stations at different time.

3) Foursquare location data
Source: https://api.foursquare.com/v2/venues/search
Format: Foursquare API
Feature: Using the location information of the studied subway station in the database queries, the venues around the searched station will be returned for the further analysis.

3.2 Understanding of data

	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
0	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/26/2019	00:00:00	REGULAR	7247322	2455491
1	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/26/2019	04:00:00	REGULAR	7247336	2455499
2	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/26/2019	08:00:00	REGULAR	7247351	2455532
3	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/26/2019	12:00:00	REGULAR	7247463	2455623
4	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/26/2019	16:00:00	REGULAR	7247755	2455679

Figure 1. MTA turnstile data (The first five lines of data in the week of 11/02/2019)

Fig. 1 illustrates the data imported from MTA turnstile data. MTA regularly publishes turnstile data every week. Each file contains information regarding the counts of entries and exits through each turnstile in MTA stations around every 4 hours. Each turnstile is distinguished by UNIT, SCP and STATION. Meanwhile, each station is uniquely identified by the station name, line mark, and division tag. In this project, a few assumptions are made to treat the subway ridership in the turnstile data. First, a person who enters a station and exits another is counted as two riders at different stations. Second, passengers who transfer between different trains/lines are not recorded by the turnstile at the transfer station. Third, a person who enters or exits the station(s) in the same collection period will be counted as multiple riders in the data.

MTA turnstile data includes the ridership information at each station. However, it does not include any real geolocation information of the station or its exit(s). Therefore, such data could not be directly used in the location-based queries to the Foursquare venue database. To solve this problem, a new set of geolocation data is needed.

	Station ID	Complex ID	GTFS Stop ID	Division	Line	Stop Name	Borough	Daytime Routes	Structure	GTFS Latitude	GTFS Longitude	North Direction Label	South Direction Label
0	1	1	R01	BMT	Astoria	Astoria - Ditmars Blvd	Q	N W	Elevated	40.775036	-73.912034	NaN	Manhattan
1	2	2	R03	BMT	Astoria	Astoria Blvd	Q	N W	Elevated	40.770258	-73.917843	Ditmars Blvd	Manhattan
2	3	3	R04	BMT	Astoria	30 Av	Q	N W	Elevated	40.766779	-73.921479	Astoria - Ditmars Blvd	Manhattan
3	4	4	R05	BMT	Astoria	Broadway	Q	N W	Elevated	40.761820	-73.925508	Astoria - Ditmars Blvd	Manhattan
4	5	5	R06	BMT	Astoria	36 Av	Q	N W	Elevated	40.756804	-73.929575	Astoria - Ditmars Blvd	Manhattan

Figure 2. MTA station profile data

It is found that the station profiles of NYC subway are available in the MTA website. As shown in Fig. 2, this data contains the latitude and longitude information of each station. Therefore, its geolocation feature will be used to complete turnstile data. Joining both datasets will be discussed in the next part.

3.3 Data Preparation

The data sets used in this project come from different sources with their own terms and formats. Before importing them into any analysis algorithm, it is necessary to clean the data first. Details about individual steps can be found in the notebook of Week 4 assignment.

For the facility of future analysis, a unique station ID is first developed. In the turnstile data, there are 377 unique station names, 113 unique line combinations, and 6 division names. Therefore, I use three digits (000-376) to label a station, three digits (000-113) for a line combination, and one digit (0-5) for a division. For example, the station "RIT-ROOSEVELT" in Line "R" of Division "RIT" have the encoded indices, "376", "019", and "5", for the station name, line combination, and division, respectively. Therefore, its turnstile data has the station ID labelled as "t5019376" in the format of "t-Division-Line-Station" where "t" stands for the turnstile data source.

Next, the subway turnstile data is to merge with the station profile data to obtain the geolocation information for each record. Note that both MTA datasets employ a combination of station name, service line, and division name to identify each station uniquely. However, different data sources adopt these names in their own preference, e.g., the abbreviation, nicknames, and upper/lower cases. I use both script-based name matching and manual verification to obtain the one-to-one mapping between the station identification keys in both sets of data. The mapping information is saved as a dictionary where the key is the station ID and the value is the (latitude, longitude) pair of the corresponding station. The processing details can be found in the notebook.

One more feature added is the label for the hour division. Based on the dining period for each daily hour, we can identify the potential needs of coffee consumption in the subway riders in each station. I defined four different periods in a day, which are breakfast time (B, 05:00-11:00), lunch time (L, 11:00-16:00), supper time (S, 16:00-22:00), and night time (N, 22:00-05:00+1Day). Since each turnstile data contains the counts within four hours, if the record happens to pass through a period break, it is necessary to split the number into two periods by their respective shares.

The further step is to group turnstile records by the station and date so that a single station has only one row of data on a specific day. Such operations are performed by the groupby function in the dataframe. Additional columns are also added to provide more features of passenger counts in different segment of a day.

	STATION_IDX	STATION	LINENAME	LOCATION	DAY_WEEK	PASSENGERS	ENTRIES	EXITS	B_Counts	L_Counts	S_Counts	N_Counts
0	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	1	20961.0	12430.0	8531.0	6030.0	5325.0	7767.0	1839.0
1	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	2	22056.0	13185.0	8871.0	6141.0	5729.0	7874.0	2312.0
2	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	3	22499.0	13504.0	8995.0	6430.0	5535.0	7960.0	2574.0
3	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	4	21093.0	12810.0	8283.0	6224.0	5127.0	7085.0	2657.0
4	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	5	21810.0	13062.0	8748.0	5918.0	5711.0	7616.0	2565.0
5	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	6	13457.0	7754.0	5703.0	2939.0	5052.0	5054.0	412.0
6	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	7	11414.0	6603.0	4811.0	1885.0	3202.0	3254.0	3073.0
7	t0001001	5 AV/59 ST	NQRW	(40.764811, -73.973347)	1	27881.0	15751.0	12130.0	7571.0	7995.0	11607.0	703.0

Figure 3. The cleaned data for the daily ridership at each station

Since I am more interested in learning the statistics of the subway ridership. The data is further grouped and averaged by the day of week. Finally, two separate dataframes are obtained, one for the average weekday data and the other for the average weekend data. Both are saved as CSV export data.

	STATION_IDX	STATION	LINENAME	LOCATION	PASSENGERS	ENTRIES	EXITS	B_Counts	L_Counts	S_Counts	N_Counts	LATITUDE	LONGITUDE
0	t0000000	59 ST	NQR456W	(40.762526, -73.967967)	21683.8	12998.2	8685.6	6148.6	5485.4	7660.4	2389.4	40.762526	-73.967967
1	t0001001	5 AV/59 ST	NQRW	(40.764811, -73.973347)	28250.4	16195.6	12054.8	7665.2	7832.8	11732.4	1018.8	40.764811	-73.973347
2	t0001002	57 ST-7 AV	NQRW	(40.764664, -73.980658)	62275.4	36188.6	26086.8	16510.4	16266.8	22055.6	7442.6	40.764664	-73.980658
3	t0001003	49 ST	NQRW	(40.759901, -73.984139)	43546.8	24051.8	19495.0	10989.6	10251.8	15957.6	6347.8	40.759901	-73.984139
4	t0002004	TIMES SQ-42 ST	ACENQRS1237W	(40.755983, -73.986229)	40079.4	16989.6	23089.8	12394.2	9851.2	14409.4	3424.6	40.755983	-73.986229

Figure 4. The cleaned data for the average weekday ridership at each station

4 Results

[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.]

4.1 Visualizing the Subway Station Data

First, let's visualize the data obtained to have the first impression. Using the subway station data, the program identifies 145 stations in the turnstile data. Fig. 5 shows the locations of these subway stations in the Manhattan map using the Folium module. When clicking each station in the map, a label will pop up showing the station's name and service line(s). In the week of November 02, 2019, the average subway ridership in Manhattan in a weekday is 5887125.0. And the average subway ridership in Manhattan in a weekend is 2728774.5.

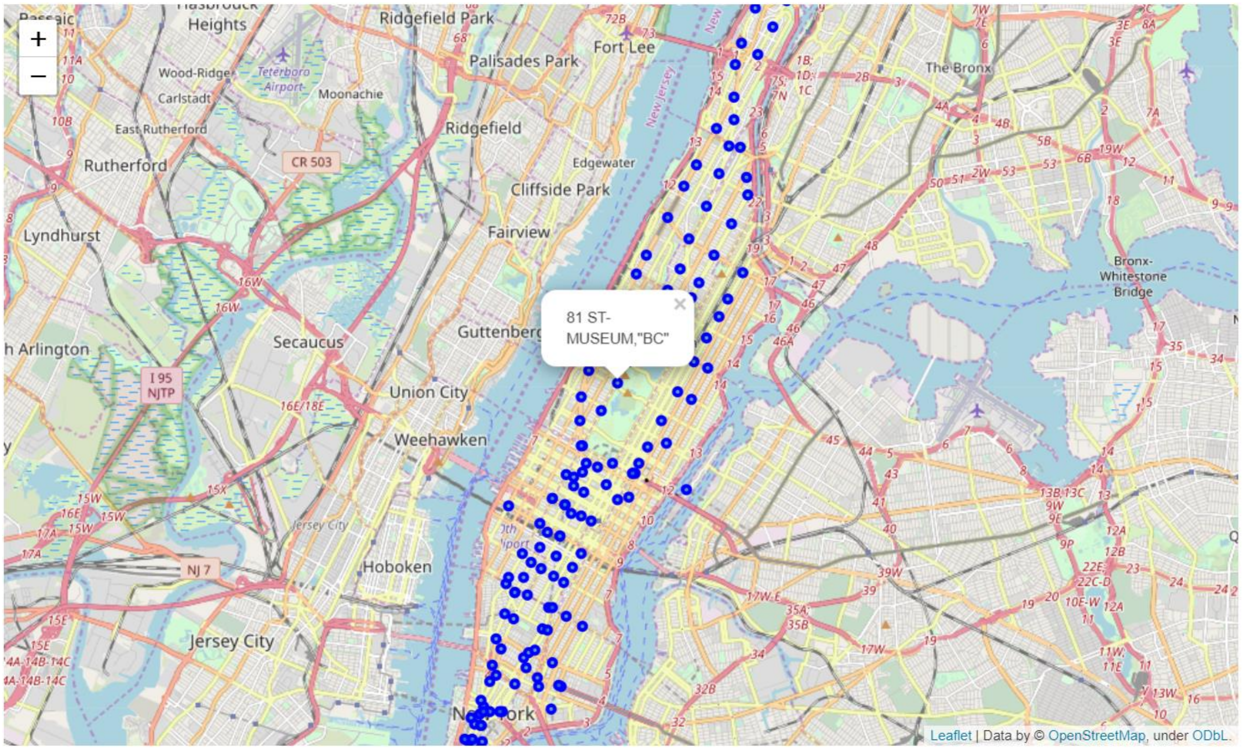

Figure 5. Subway stations in NYC (Manhattan)

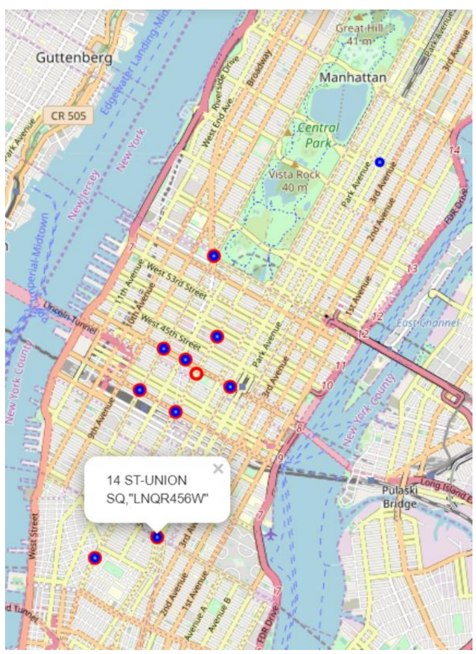

Figure 5. 10 busiest stations in a weekday (RED) and weekend (BLUE)

Fig. 6 illustrates the top 10 busiest stations in a weekday and weekend. The two sets are almost the same which mark the transportation hubs and places of interests for daily commuters and travelers. Most of these locations are in midtown.

Figure 7. Top 10 busiest stations in a weekday in different periods (breakfast in BLUE, lunch in RED, supper in GREEN, and night in BLACK)

In Fig. 7, a further study is visualized for the busiest stations in different time segments of a weekday, which further confirms that busy stations always go busy. The peaks of ridership are mainly spotted in midtown although there is slight difference varying with the time in a day.

More observations can be made to fully understand the pattern of the ridership in Manhattan subway stations. However, it is not the main goal of this project. The readers are encouraged to explore the harvested data for their own interest.

4.2 Spatial Analysis of Foursquare Venue Data

In the following, I will integrate the subway data into the venue information obtained through the Foursquare database. Using the geolocation information of each station, we can search the nearby venues through the Foursquare API. Be 300 meters the walkable distance from a station to its neighborhood venues, the following discussions will focus on the returned coffee shop information within this range from each station.

The query is organized to include "coffee" as the keyword to search for all venues within 300 meters near each subway station. The records for each station are saved into one row summarizing the number of returned venues. Meanwhile, all records of "Starbucks" are also saved in a separate column for the later discussion.

	STATION_IDX	STATION	LINENAME	LATITUDE	LONGITUDE	Starbucks	StarbucksAll	100M	200M	300M
0	t0000000	59 ST	NQR456W	40.762526	-73.967967	4	6	1	9	20
1	t0001001	5 AV/59 ST	NQRW	40.764811	-73.973347	1	2	1	3	7
2	t0001002	57 ST-7 AV	NQRW	40.764664	-73.980658	9	12	6	18	31
3	t0001003	49 ST	NQRW	40.759901	-73.984139	7	11	8	21	34
4	t0002004	TIMES SQ-42 ST	ACENQRS1237W	40.755983	-73.986229	6	10	7	19	50

Figure 8. The counts of nearby coffee shops around subway stations in NYC

Fig. 8 shows the results where "100M", "200M", and "300M" store the numbers of coffee shops from each station in a distance of 100 meters, 200 meters, and 300 meters, respectively. "Starbucks" stores the number of Starbucks venues within a radius of 300 meters from each station. Note that the returned results have shown a maximum limit of 50 results per Foursquare query [4]. A separate round of venue queries were performed to obtain the total number of Starbucks within 300 meters from each station using the keyword "Starbucks" the results of which are stored in the column "StarbucksAll".

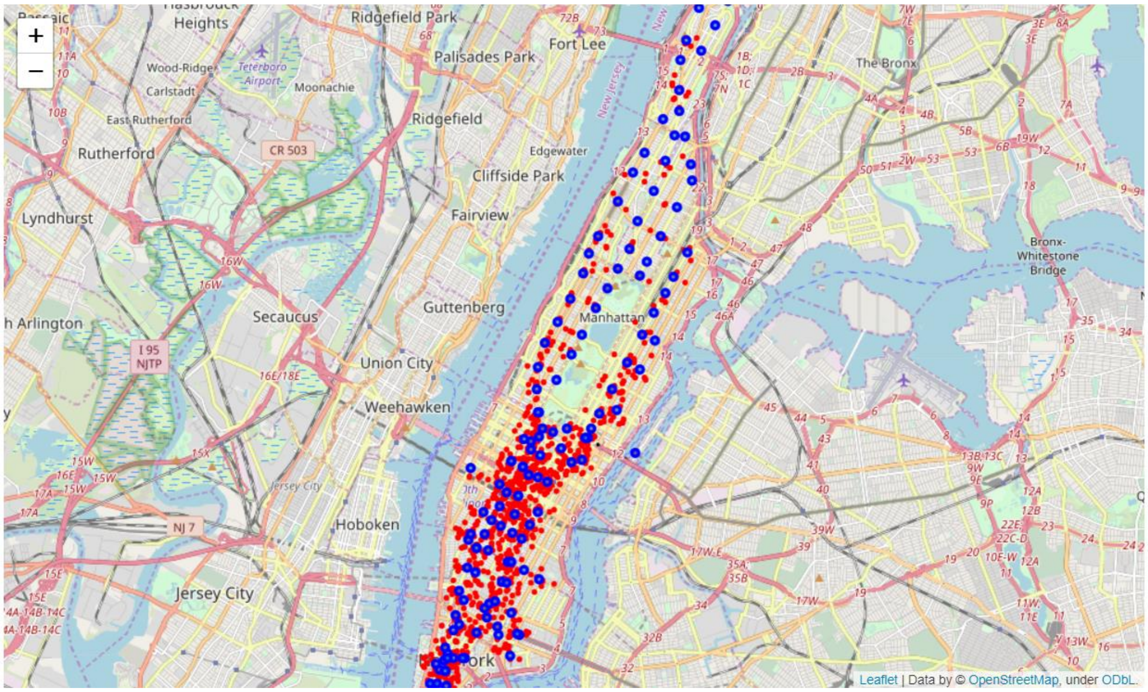

Figure 9. Coffee shops near subway stations in NYC (Coffee shops in RED, and subway stations in BLUE)

As shown in Fig. 9, like the subway rider distribution, most of such coffee shops are located in downtown and midtown of NYC, which agrees with the observation that these two areas are the most crowded and busiest in the city.

Figure 10. The top 10 busiest subway stations (BLUE) V.S. The top 10 subway stations that have the highest number of coffee shops within a radius of 300 meters (RED)

I plot 10 stations with the highest ridership and 10 stations having the most coffee shops in the vicinity in Fig. 10. As the two sets are not exactly the same, it indicates that access to subway is only one of the many factors that determine the sites of coffee shops.

4.3 Statistical Analysis of the Data

Next, using the obtained data, we will check the relationship between the average weekday ridership of each station and the number of nearby coffee shop(s). As shown in Fig. 11, the relationships show a few patterns at different stations.

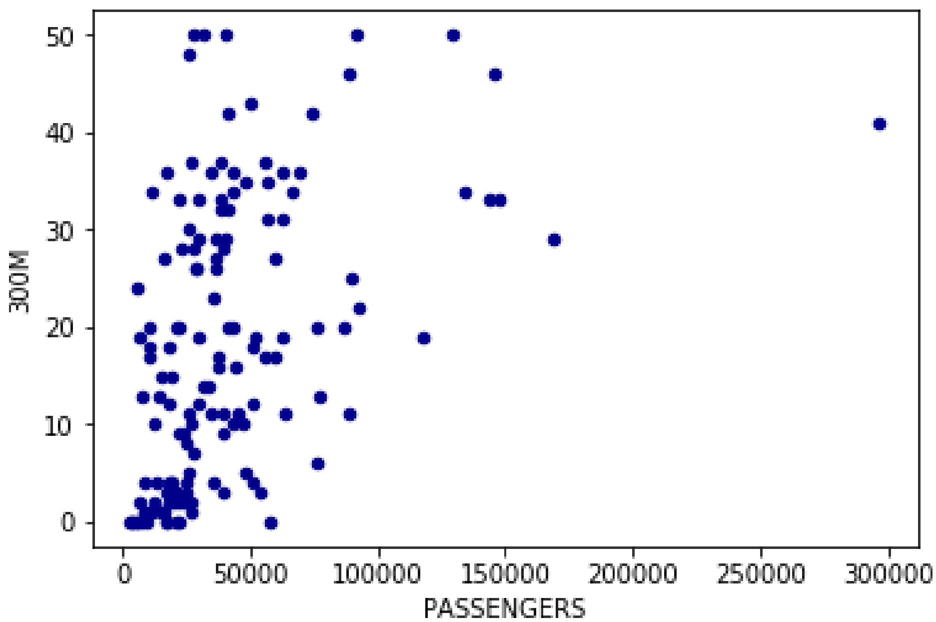

Figure 11. The scatter plot of the weekday subway passengers and the number of coffee shops within 300 meters around subway stations

Running the test of the correlation between the ridership and coffee shops, it returns the value of 0.461, which indicates a relatively strong positive correlation between these two features.

	PASSENGERS	300M
PASSENGERS	1.000000	0.461635
300M	0.461635	1.000000

The reasons why they are not strongly correlated with each other, i.e., the value approaches to 1, can be explained in the following ways. First, coffee shops are popular places for refreshment. Therefore, many coffee shops also serve local residents and workplaces in the neighborhood although they are next to subway stations whose passengers are part of their customers. Second, there are other venue categories serving food and drinks other than coffee shops which are not included here, such as local restaurants, fast food chains, and convenience stores. For example, McDonalds and Dunkin Donuts are not labelled as “coffee” venues in Foursquare’s database.

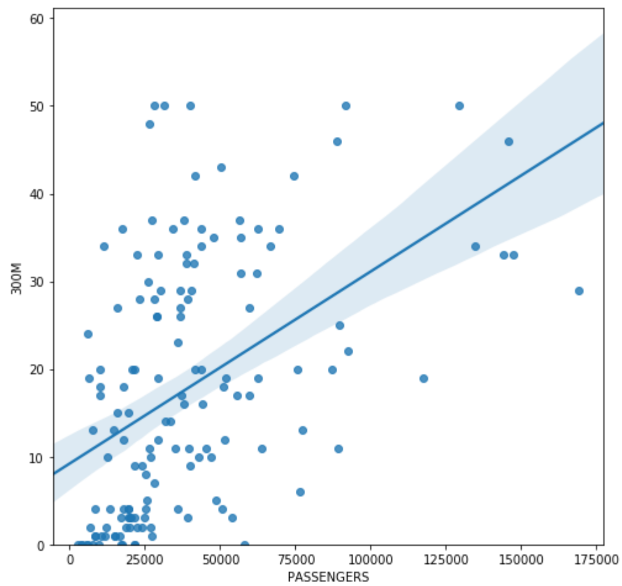

Figure 12. The scatter plot with the linear regression prediction of the number of coffee shops within 300 meters around subway stations under the weekday subway passengers

As shown in Fig. 12, a linear regression method is used to obtain the projected linear relationship between the ridership and the number of coffee shops. Here, the linear regression is obtained based on the dataset after removing the outlier, i.e., the station "GRD CNTRL-42 ST" which has a ridership of nearly 300,000 on an average weekday. In Fig. 12, each dot represents one station's metrics. Using this figure, it can show a few interesting results regarding different clusters of data.

First, let's use the categorical feature to label different stations. Here, "Cat_1" indicates the stations with a low ridership, i.e., less than 30,000 passegers per weekday; "Cat_2" indicates the stations with a moderate ridership, i.e., whose value between 30,000 – 60,000; "Cat_3" indicates the stations in heavy use, i.e., between 60,000 – 120,000; For "Cat_4" stations, they are usually the hubs with a ridership figure beyond 120,000.

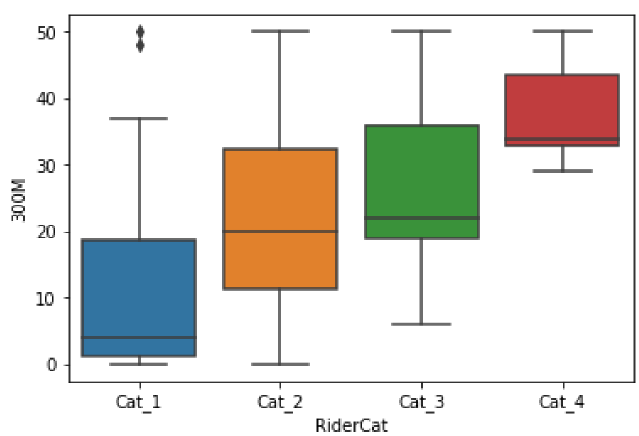

Figure 13. The box plot of the number of coffee shops within 300 meters around subway stations of different categories

As shown in the box plot of Fig. 13, the coffee shop number varies with the categories which agrees with the earlier observations that a positive correlation exists between the ridership and the number of coffee shops.

5 Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

Let's continue the above discussion on some additional patterns shown in the relationship between the ridership and the coffee store number.

5.1 Coffee Shop Distribution Patterns around NYC Subway Stations

Figure 14. The subway stations with a low ridership (<30,000 per weekday) and few nearby coffee shops within 300 meters (no more than 5)

The first cluster of stations has very few coffee shops (less than or equal to 5) nearby and a low ridership (30,000 or less per weekday). The locations of such stations are shown in Fig. 14. It is found that such stations are mainly located in the north of Manhattan island which are far away from the conventional business hubs, i.e., downtown and midtown.

The second cluster to watch includes stations with a low ridership (30,000 or less per weekday) but having affluent coffee shop choices for their riders in the walking distance (no less than 35). As shown in Fig. 15, most of them reside in the neighborhoods of downtown and midtown.

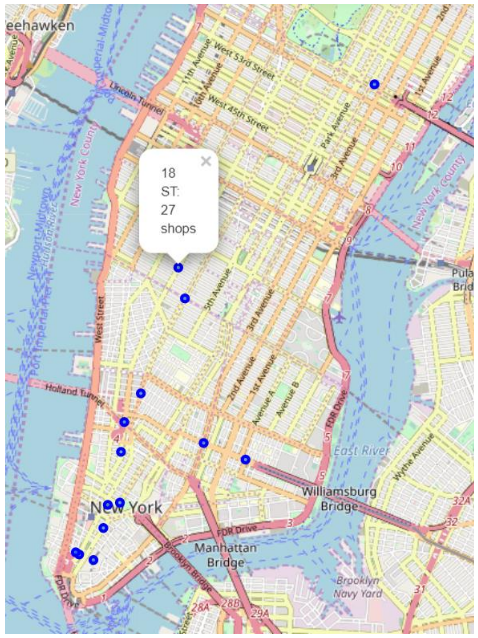

Figure 15. The subway stations with a low ridership (<30,000 per weekday) and abundant nearby coffee shops within 300 meters (no less than 35)

Next, let’s turn to the hub stations which usually serve over 80,000 passengers in a single weekday. Thanks to the huge number of riders, the neighborhoods also enjoy many coffee stores as shown in Fig. 16. Most of such sites are found in the midtown.

Figure 16. The subway hub stations with a high ridership (>80,000 per weekday) in RED and the top 10 subway stations having the most nearby coffee shops within 300 meters in BLUE

Last, let's look at opportunities for opening a new coffee shop near the subway station. From above discussions, a subway station with a busy schedule, e.g., 60,000 passengers per weekday, is supposed to be able to support a few local venues including coffee shops. I use the criteria of more than 60,000 riders per weekday and less than 5 coffee shops to search for the stations that may be the candidate locations for new coffee shop owners.

Figure 17. Subway stations with a ridership larger than 60,000 per weekday but having few coffee shops (less than 5)

Use such criteria, it returns three of such sites. Compared to the stations of cluster one in the same area, these three stations serve more passengers. A good sign is that two of them have shown a few coffee shops. However, the suggested locations should be reviewed with more environmental factors, such as alternative service vendors (e.g., fast food chains), competition restaurants, and other environmental factors.

	STATION	LINENAME	PASSENGERS	ENTRIES	EXITS	RiderCat	300M	Starbucks
0	145 ST	ABCD	50805.2	25418.0	25387.2	Cat_2	4	1
1	125 ST	ACBD	58153.0	31024.6	27128.4	Cat_2	0	0
2	125 ST	456	53925.8	29241.2	24684.6	Cat_2	3	0

5.2 Starbucks around NYC Subway Stations

Starbucks, as a popular refreshment venue, has its own footprints in NYC. Using the available data in this project, let's examine its store locations around subway stations and compare them to the other coffee shops.

Note: Foursquare API with the query of "coffee" only contains a partial view of the all available "Starbucks" venues in the results. Therefore, a separate round of queries with "Starbucks" as the keyword was performed.

The result of the ridership and the total number of found Starbucks stores is shown in Fig. 18.

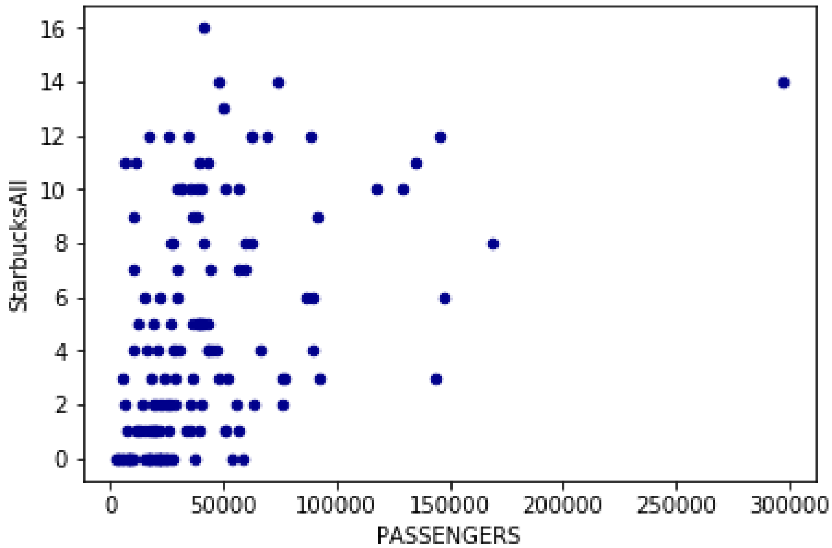

Figure 18. The scatter plot of the weekday subway passengers and the number of Starbucks within 300 meters around subway stations

	PASSENGERS	StarbucksAll
PASSENGERS	1.000000	0.434211
StarbucksAll	0.434211	1.000000

Meanwhile, the correlation test returns a value of 0.4342 which indicates a relatively strong relationship between these two factors.

```
In [151]: print("In a weekday")
          print(f"Starbucks shop number: {df_dict_starbucks.shape[0]}, covering the total ridership: {sum(result[result['Starbucks']!=0]['P/
          print(f"Starbucks shop number: {df_dict_starbucks_all.shape[0]}, covering the total ridership: {sum(result[result['StarbucksAll']
          print(f"Coffee shop number: {df_dict_shops.shape[0]}, covering the total ridership: {sum(result['PASSENGERS'].tolist())}")

In a weekday
Starbucks shop number: 136, covering the total ridership: 5206800.000000001 (0.8844384992674691)
Starbucks shop number: 246, covering the total ridership: 5325643.2 (0.9046254665902285)
Coffee shop number: 1020, covering the total ridership: 5887125.000000002
```

```
In [150]: df_tt_slim_weekend = pd.read_csv("turnstile_weekend.csv")
          result_weekend = pd.concat([df_tt_slim_weekend, df_station_coffee[['300M', 'Starbucks', 'StarbucksAll']]], axis=1, join='inner')
          result_weekend.head()
          print("In a weekend")
          print(f"Starbucks shop number: {df_dict_starbucks.shape[0]}, covering the total ridership: {sum(result_weekend[result_weekend['St
          print(f"Starbucks shop number: {df_dict_starbucks_all.shape[0]}, covering the total ridership: {sum(result_weekend[result_weekend
          print(f"Coffee shop number: {df_dict_shops.shape[0]}, covering the total ridership: {sum(result_weekend['PASSENGERS'].tolist())}"

In a weekend
Starbucks shop number: 136, covering the total ridership: 2349247.0 (0.8609165029942929)
Starbucks shop number: 246, covering the total ridership: 2413639.0 (0.8845139090826303)
Coffee shop number: 1020, covering the total ridership: 2728774.5
```

The performance of the market coverage is managed at the same level of or slightly better than the average performance for all coffee shops in this study. Starbucks cover most of stations and subway passengers, i.e., it opens 100+ stores that cover almost 90% NYC subway ridership. Therefore, in terms of location effectiveness, Starbucks plays very well in selecting the sites to cover as many stations and subway passengers. Among the total 1020 coffee shops around 145 subway stations in NYC, Starbucks occupy 136 locations which cover 88.44% of the total daily subway riders. I further use Foursquare API to get the complete set of Starbucks around these stations, i.e., “StarbucksAll, it returns 246 Starbucks venues which increase the coverage to 90.46%. In a weekend, these Starbucks locations can also serve around 88.45% of the 2.7 million subway riders from the nearby subway stations.

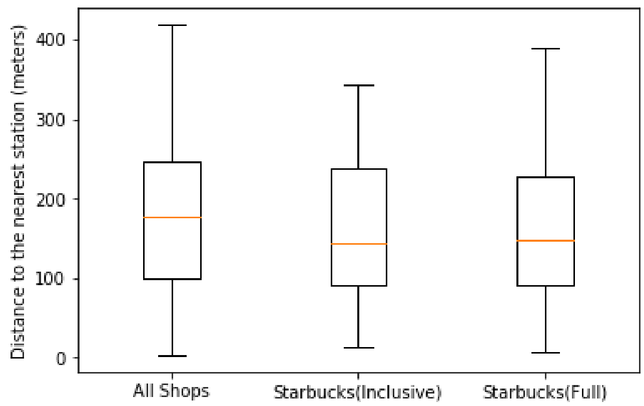

Figure 19. The box plot of the distances from Starbucks stores to the nearest subway stations

Let’s consider another metric. Using all the returned Starbucks venues around subway stations, the distances from them to the nearest subway stations are calculated and shown in Fig. 19. If the distance metric represents the easiness for subway riders to visit the coffee shop, Starbucks has adopted a similar strategy of choosing its locations compared with the other competitors in serving subway passengers.

6 Conclusion

Subway stations and coffee shops as two common venues in the daily life of NYC provide a way of sniffing its business activities. Through this report, the time and spatial variations have been illustrated based on the obtaining, processing, and analysis of available public data. There are a lot of potentials for further analysis to discover new data-oriented business opportunities, such as find location-based services, and evaluate the health of stock market company’s portfolio and investment strategies.

Reference:

[1] http://web.mta.info/nyct/facts/ffsubway.htm

[2] https://patch.com/new-york/new-york-city/nyc-americas-coffee-shop-capital-study-says

[3] https://coffeemachinegrinder.com/2019/09/10/how-many-starbucks-in-nyc/

[4] https://stackoverflow.com/questions/14211120/maximum-number-of-results-in-foursquare-api