Abstract

The metropolitan transportation authority (MTA) has over 400 stations in New York city, and it is for sure that the pandemic has affected them, due to the total and partial lockdown. The goal of this analysis is to know whether the public transportation usage differs before and after the total lockdown that happened in New York City from March 2020 until April 2020, see and visualize how did the pandemic actually affected the public transportation in New York city.

Data structure:

The dataset was provided by Metis of the metropolitan transportation authority in New York city. I considered only six months before and after march 2020 and April 2020, in total 14 months. The dataset consists of 204795 and 217832 rows, and 11 columns in each data set.

C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESc	ENTRIES	EXIT
control area	remote station unit	the address for the device of calculating exits and entries	station name	train lines of this station	line originally the station belonged to BMT, IRT, or IND	Date	Time	scheduled audit event (every 4 hours)	cumulative entry	cumulative exit
Object	Object	Object	Object	Object	Object	Object	Object	Object	Integer	

	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
0	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/24/2020	00:00:00	REGULAR	7477474	2544276
1	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/24/2020	04:00:00	REGULAR	7477477	2544276
2	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/24/2020	08:00:00	REGULAR	7477488	2544294
3	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/24/2020	12:00:00	REGULAR	7477522	2544334
4	A002	R051	02-00-00	59 ST	NQR456W	BMT	10/24/2020	16:00:00	REGULAR	7477644	2544363

Data cleaning:

For the data cleaning and the analysis, I used five of the existing columns and they are: station, date, time, entries, and exits. The cleaning steps are: Striping the columns of all the white spaces and next line space, dropping the C\A, unit, linename, division, and the desc columns, converting the date into datetime, dropping outliers, NA values, the negative values after calculations.

	STATION	DATE	TIME	ENTRIES	EXITS	TOTAL_ENTRIES	TOTAL_EXITS	TOTAL_TRAFFIC
1	59 ST	2020-10-24	04:00:00	7477477	2544276	3.0	0.0	3.0
2	59 ST	2020-10-24	08:00:00	7477488	2544294	11.0	18.0	29.0
3	59 ST	2020-10-24	12:00:00	7477522	2544334	34.0	40.0	74.0
4	59 ST	2020-10-24	16:00:00	7477644	2544363	122.0	29.0	151.0
5	59 ST	2020-10-24	20:00:00	7477785	2544386	141.0	23.0	164.0

Design and algorithm:

First, I merged the days in a specific month, then did the cleaning process on the merge result. I added three new columns named: TOTAL_ENTRIES, TOTAL_EXITS, TOTAL_TRAFFIC, the first two calculate the number of people (not cumulative number), and the last for summing the first two in order to get the total traffic, and the three of them are float. Then I calculated the total traffic in one month, and append it to another data frame, in order to easily display the chart without the need to re-run the whole code.

	MONTH	TOTAL_TRAFFIC
0	OCTOBER 2020	133596429
1	SEPTEMBER 2020	100216541
2	AUGUST 2020	112864786
3	JULY 2020	82855166
4	JUNE 2020	69046048
5	MAY 2020	65611019
6	APRIL 2020	46167307
7	MARCH 2020	162791444
8	FEBRUARY 2020	325014415
9	JANUARY 2020	248903149
10	DECEMBER 2019	261815687
11	NOVEMBER 2019	338969289
12	OCTOBER 2019	276936304
13	SEPTEMBER 2019	272686638

Tools:

- Spyder as developing environment.
- Pandas for data manipulations.
- Matplotlib.pyplot for data visualization.
- Matplotlib.ticker import StrMethodFormatter for the formatting the result.

Conclusion and communication:

In conclusion, there was a change in using public transportation before and after the total lockdown. Moreover, due to the partial lockdown, people were using less public transportation than before march, and after the partial lockdown (after April), people are using public transportation less than before the total lockdown.