By Nouf Alotaibi

Email: noufmitla@gmail.com

METRO MOTION

Abstract

This is the first project for the T5 Data Science Bootcamp, which is an exploratory data analysis of the MTA turnstile data set using SQL paired with Python and its libraries *Pandas*, *NumPy*, and *Matplotlib*. This project discusses the work I did to help *Metro Motion* to have a better schedule for repairing and cleaning MTA, and the dataset description, and the tools I will used for the project.

1. Design

The "*Metro Motion*" is a company that provides repairing and cleaning services for metropolitan transportation in New York City. The company is serving a lot of stations that belong to the Metropolitan Transportation Authority.

Metro Motion struggled a lot after the pandemic ended. The company's schedule got messed up because many people are returning to use the transportation services, and their schedule is not meeting the right times for repairing and cleaning. Therefore, the company is wasting a lot of resources. So, I got approached by the company to help them improve their work schedule by estimating the appropriate times for repairing and cleaning. While working on this problem, I answered the questions below.

- What are the most active stations?
- What are the most active/inactive times for stations?
- What are the most active/inactive days of the week for stations?
- Is there a similarity between that dataset prior covid-19 and dataset now?
- What is the best schedule for the company?

The value for the company is that it will save the company's resources and improve their work quality and reputation as a company.

2. Dataset

The Metropolitan Transportation Authority is North America's largest transportation network, serving 15.3 million people across a 5,000-square-mile travel area surrounding New York City through Long Island, southeastern New York State, and Connecticut. The MTA publishes their data in CSV files every week. The table below illustrates the dataset's features/columns and their types.

Column Name	Column type
C/A	Object
UNIT	Object
SCP	Object
STATION	Object
LINENAME	Object
DIVISION	Object
DATE	Object
TIME	Object
DESC	Object
ENTRIES	int64
EXITS	int64

2.1 Scope

- **Sample Size:** I used four datasets, two of them were before covid-19, October and November in 2019, and I concatenate it with two recent datasets from 2021, June and July.
- **Rows:** 3,740,978 rows.
- **Columns:** 11 columns, which are c/a, unit, scp, station, linename, division, date, time, desc, entries, and exits.

3. Feature Engineering

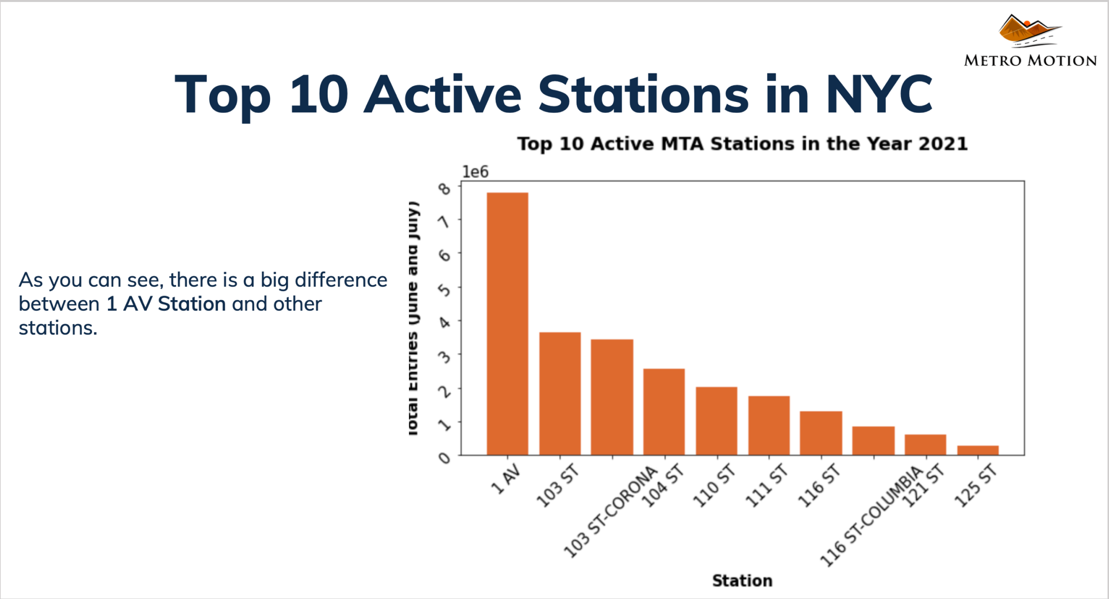
- Concatenating **2019** and **2021** datasets to do a little bit of work on them.
- Stripping columns from whitespace.
- Combining the DATE and TIME columns into one column **DATE_TIME**.
- Extracting the **DAY**, **MONTH** and **YEAR** from DATE column.
- Extracting the **HOUR** from TIME column and split the hours into **TIME_PERIOD** = ['Morning', 'Afternoon', 'Evening', 'Night'].
- Separate the dataset into **turnstiles_19** and **turnstiles_21**.
- Calculating the difference between ENTRIES per a unique turnstile to get **DAILY_ENTRIES**.
- Fixing the DAILY_ENTRIES by removing **outliers** and resolve the **reset** and the **reverse count**.

4. Tools

These are the technologies and libraries that I used for this project:

- **Technologies:** SQL, SQLlite, Python, Jupyter Notebook.
- **Libraries:** Numby, Pandas, Matplotlib, Seaborn.

5. Communications



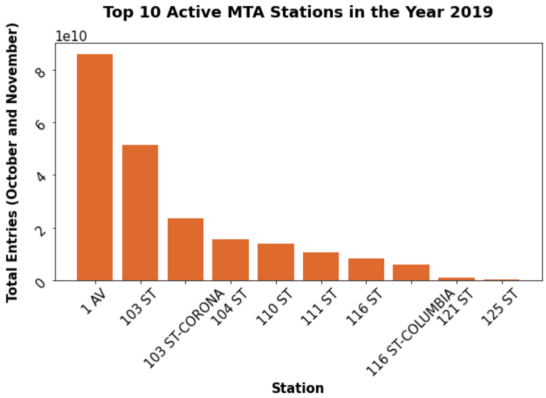
METRO MOTION

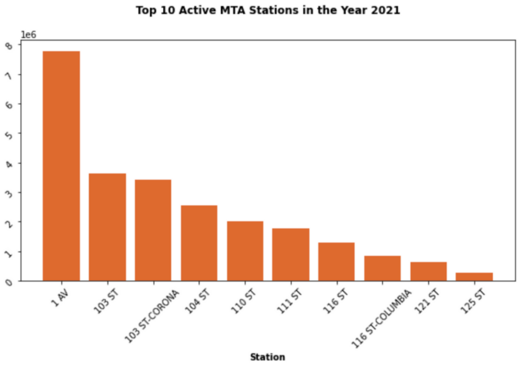
Top 10 Active Stations in NYC

As you can see, there is a big difference between 1 AV Station and other stations.

METRO MOTION

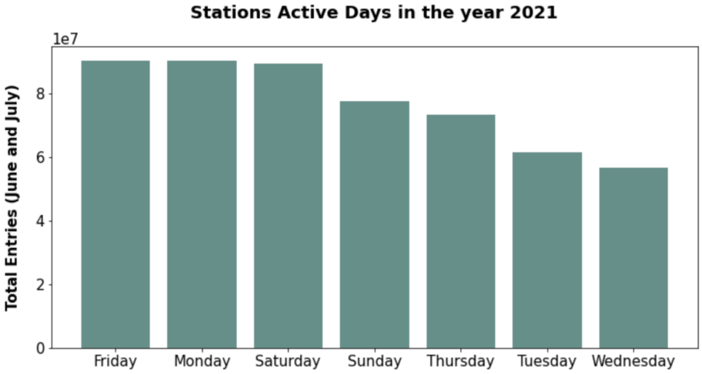
2019 VS 2021





METRO MOTION

Stations Active Days in NYC

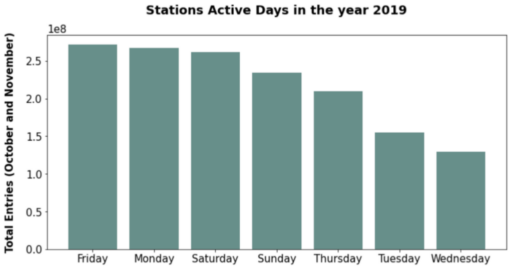


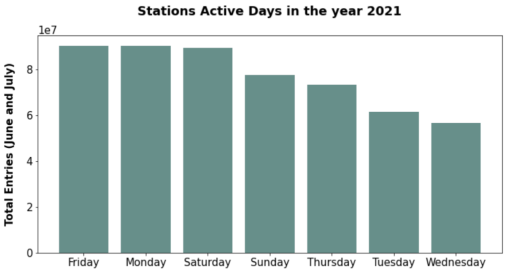
Most active days are weekends, and the start of the week.

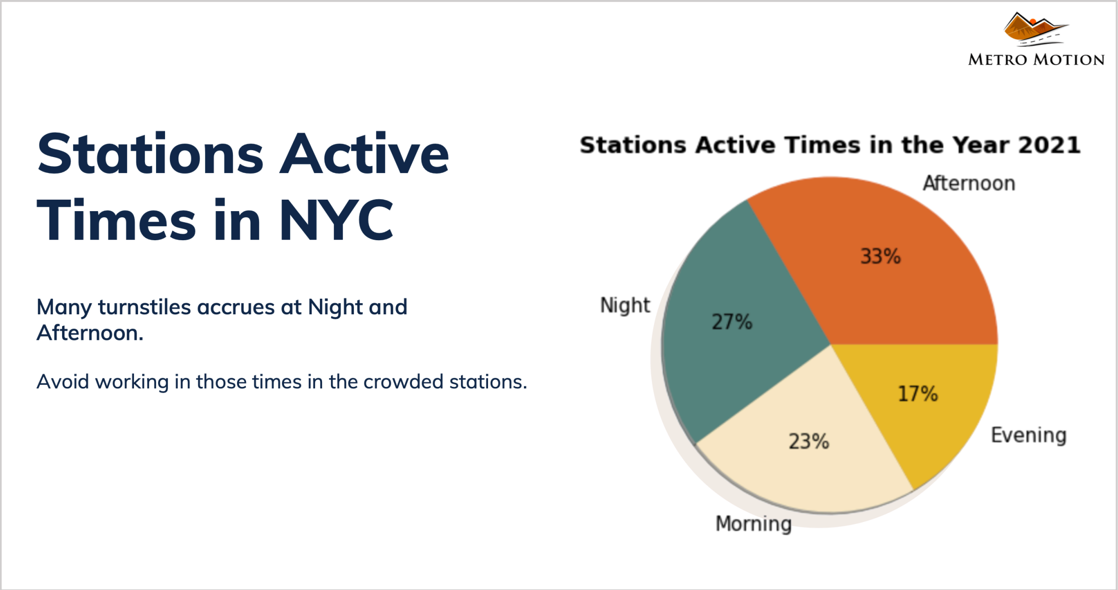
It gets less by the end of the week.

METRO MOTION

2019 VS 2021







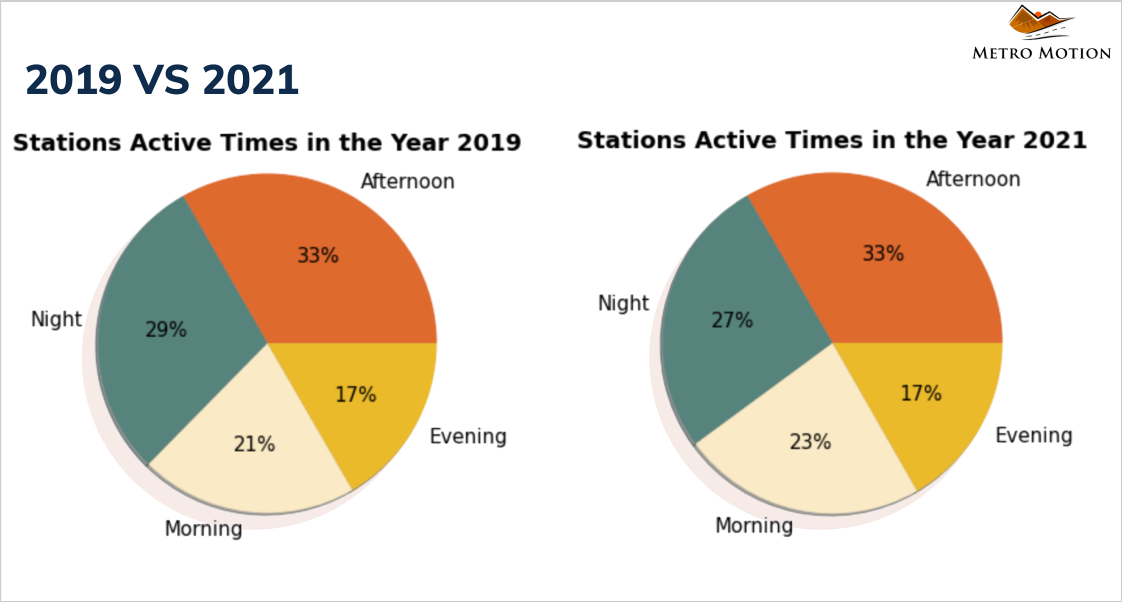
METRO MOTION

Stations Active Times in NYC

Many turnstiles accrues at Night and Afternoon.

Avoid working in those times in the crowded stations.

Stations Active Times in the Year 2021

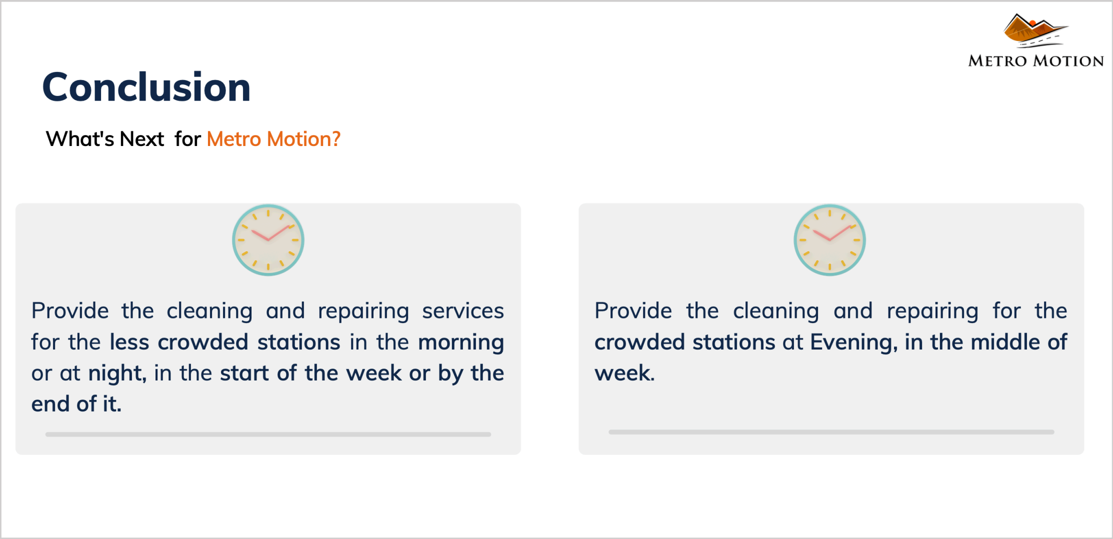


METRO MOTION

2019 VS 2021

Stations Active Times in the Year 2019

Stations Active Times in the Year 2021



METRO MOTION

Conclusion

What's Next for Metro Motion?

Provide the cleaning and repairing services for the **less crowded stations** in the **morning** or at **night**, in the **start of the week or by the end of it.**

Provide the cleaning and repairing for the **crowded stations** at **Evening, in the middle of week.**