

Question/need:

- What is the framing question of your analysis, or the purpose of the model/system you plan to build?

Categorize the subway stations based upon the drop in turnstile traffic seen during the Covid pandemic.

- Who benefits from exploring this question or building this model/system?

Target commercial realtor companies that purchase, sell, or lease commercial properties in and around the subway stations. The realtor companies will be able to use this data to adjust commercial cap rates to help their buyers and sellers better price real estate opportunities in and around the subway stations. For example, subway stations that had a small traffic drop from pre-pandemic times should have a lower adjusted cap rate than those with high a drop in turnstile traffic.

Data Description:

- What dataset(s) do you plan to use, and how will you obtain the data?

MTA turnstile data

<http://web.mta.info/developers/turnstile.html>

- What is an individual sample/unit of analysis in this project? What characteristics/features do you expect to work with?

Identify Covid Pandemic traffic drop as a percentage by comparing two six week periods of turnstile data at individual station level. The two time periods are Feb-mid March 2019 and Feb – mid March 2021.

Each row of this data includes the following information:

- **C/A** = Control Area (e.g., A002)
- **unit** = Remote Unit for a station (e.g., R051)
- **SCP** = Subunit Channel Position represents a specific address for a device (e.g., 02-00-00)
- **station_code** = C/A + unit, locating a station
- **turnstile** = C/A + unit + SCP, locating a turnstile

- **Station** = Represents the station name the device is located at
- **date** = Represents the date (MM-DD-YY)
- **time** = Represents the time (hh:mm:ss) for a scheduled audit event
- **datetime** = date + time (MM-DD-YY hh:mm:ss)
- **desc** = Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours)
- **entries** = The cumulative entry register value for a turnstile device
- **exits** = The cumulative exit register value for a turnstile device

Since we only need traffic data at station and date/time level we can drop the following data indexes: C/A, unit, SCP, station_code, and turnstile.

We will aggregate entry and exit values at station and date level.

- If modeling, what will you predict as your target?

No modelling being done, this project is only doing exploratory analysis.

Tools:

- How do you intend to meet the tools requirement of the project?

SQL Database and Querying tools, Python analytic and graphing libraries.

- Are you planning in advance to need or use additional tools beyond those required?

No

MVP Goal:

- What would a [minimum viable product \(MVP\)](#) look like for this project?

A table that lists categorizes all the subway stations into:

Change in NYC Turnstile Traffic by Station during COVID-19

- 0 – 25% Percentile
- 25 – 50% Percentile
- 51 – 75% Percentile

- 76 – 100% Percentile
- Outliers