Project Proposal

Background :

Here we are the marketing team at the stations. We want to do data analysis to help us lead rental

space, whether it's advertising screens, food, and drink machines, stores, or pharmacies, etc...

We want to know the busiest stations, the times of the crowd, and where there is this crowding

when entering or exiting.

So that we can distribute the rental spaces appropriately.

Question :

What are the busiest stations?

What is the time of day with the highest traffic per station?

Where is the most crowded when entering or leaving?

Data Description :

MTA Dataset :

Field Description

C/A: Control Area (A002)

UNIT: Remote Unit for a station (R051)

SCP: Subunit Channel Position represents an specific address for a device (02-00-00)

STATION: Represents the station name the device is located at

LINENAME: Represents all train lines that can be boarded at this station. Normally lines are

represented by one character. LINENAME 456NQR repersents train server for 4, 5, 6, N, Q, and

R trains.

DIVISION: Represents the Line originally the station belonged to BMT, IRT, or IND.

DATE: Represents the date (MM-DD-YY).

TIME: Represents the time (hh:mm:ss) for a scheduled audit event.

DESc: Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours)

      1. Audits may occur more that 4 hours due to planning, or troubleshooting activities.

      2. Additionally, there may be a "RECOVR AUD" entry: This refers to a missed audit that was recovered.

ENTRIES:The comulative entry register value for a device.

EXIST: The cumulative exit register value for a device.

Tools :

- Python

- Python Libraries: Pandas, matplotlib

- SQL

- SQLITE

- Jupyter notebook

https://github.com/Nadah16/EDA_Projects