Exploring MTA Turnstile Data

batoul alosaimi

October 2021

Proposal project

1 Introduction:

The purpose of this project is to create a schedule for a stations sterilization company. This schedule illustrates how many sterilization rounds each station needs each day, based on their needs, with the busiest station requiring more than once per day.

1.1 Question/need:

I use MTA turnstile data. To see which stations are the busiest, what are the busiest hours also which days are the busiest. So we can schedule the purges. I'll schedule the ten busiest stops and see how much each stop needs a sanitizing round per day. So I will choose a period from 2021, (6 March to 29 May).

2 Data Description

2.1 MTA Dataset:

The dataset is reported in weekly updates containing roughly 200,000 rows of data. Each row of the dataset represents the recorded by a singular turnstile. Individual turnstiles have unique identifiers built from a combination of their Station, Remote Unit, Control Area (C/A), and Subunit Channel Position. To enter or exit a subway station, one usually needs to pass through a turnstile. The turnstiles record a running tally of the number of entries and exits once every four hours.

3 Tools:

The first tool that I will be using is to import all the datasets into SQL database and join all tables then import these datasets using SQLAlchemy to query the database in Python. Pandas will be mainly used in cleaning datasets such

as handling missing values, column names, station names manipulations and creation of new columns needed for the analysis and visualization. For the visualization part, I will use matplotlib and seaborn libraries also I will try to use more advanced libraries and tools such as Bokeh, Plotly and Tableau.