MTA Turnstile Data Analysis for Cleaning Services

Mohammed Abduldaim

Abstract

this paper is a report of the study and analysis of the MTA turnstile data using tools such as sqlite databases for storing the data, sqlalchemy for importing, pythons pandas library to perform the analysis, and seaborn and matplotlib for visiulizing it. The goal is find the time frame with the least traffic for a station in order to for the cleaning services to operate during them. I looked at a subset of the stations that represent the whole date and found the time frames which has most foot traffic and the least traffic.

Design

BMS is a janitorial and cleaning services provider based in New York. The company has entered a contract with MTA to handle the subway station janitorial and cleaning tasks. BMS is responsible for cleaning the stations however they only perform their services once a day. So, they would like to find the busiest time of day and start operating afterward.

Data

The MTA data set contain information pertaining the turnstile gate for every station. Spanning over a decade, the data records every entry and exist for each gate cumulatively every four hours. It also includes the gate id, where it's station date and time of records and more. The sample I've chosen is 3 month long (June– August) 2021, as it is the newest set of data. The data has 2513079 rows and 11 columns

Algorithms

Data Cleaning

1. Duplicate entries found by grouping each station, turnstile and aggregate by count. I've found 53 total duplicates which I've removed givin their small number.
2. Then using the same grouping method only this time aggregating by the entries difference. Which outputs the entries per time frame.
3. found over 5000 missing values in the column, but that was still a very small percentage of the data so I've removed it as well.
4. Caluclated total foot traffic per time frame using the entries and exits counts.
5. Grouped each time frame under six frames, 12,4,8 AM and 4,8,12 PM
6. Grouped all turnstile by their station and newly assigned time frame and aggregated by the sum of the total entries.

Tools

- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting
- Sqlite and sqlalchemy for importing, exporting and storing data.