Kenneth Hua

Module 1: EDA Project Writeup

Abstract

The goal of this project is to assist the MTA in planning maintenance projects that need to performed on their subway lines.

- The first type of maintenance is a yearly maintenance, which needs to be scheduled at the end of winter to inspect/replace damaged subway tracks. This type of project takes 3 days to complete.
- The second type is a routine maintenance of lubricating the tracks, to prevent friction from wearing out the tracks. This type of maintenance takes 3 hours and needs to be performed weekly.

My goal is to help the MTA find the optimal times to schedule these projects, to minimize the disruption to ridership.

Design and Data

The Data being used is 3 years of MTA Turnstile Data from 2017-2019. This data contains detailed data of MTA turnstile entry and exits for each station collected over periodic intervals. The Data is pulled into a .db file from the web via the get_mta.py program, then queried using SQL Commands.

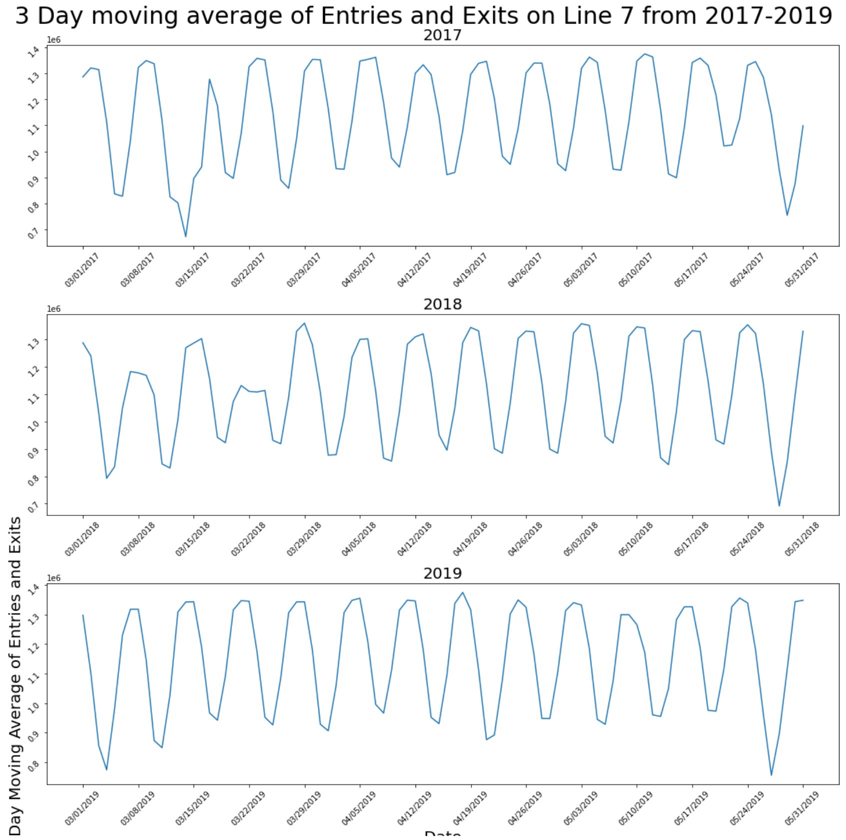
The resulting tables have:
2017: 10258959 Rows, 12 Columns
2018: 10305449 Rows, 12 Columns
2019: 10671895 Rows, 12 Columns

The tables contain information about the Turnstile IDS, the stations, the subways lines accessed by those turnstiles, a cumulative list of entries and exits, and the times and dates that those recordings are logged.

The tables are loaded into pandas dataframes to be further cleaned, processed, and aggregated for the resulting analysis.

Results

Maintenance Project 1: 7 Line Track Replacement

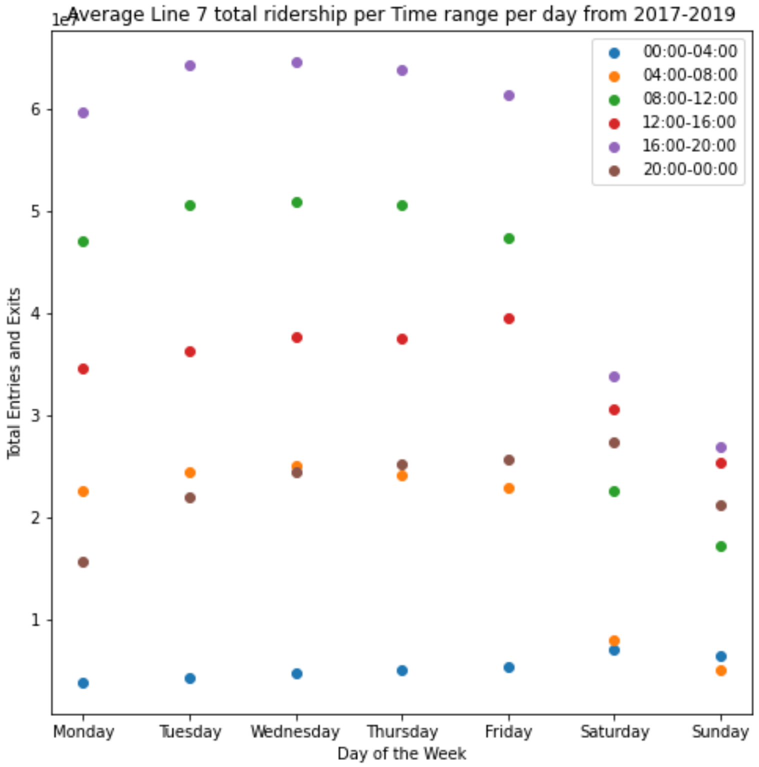


From this table, we can see the 3 day moving average of combined entries and exits across the spring season, for 2017, 2018, and 2019. From this figure, we can see a few patterns:

- The figure is clearly periodic, with the periodic intervals occurring weekly.
- Weekdays are the higher ridership (peaks), while weekends (valleys) are lower ridership
- There is a low valley on the right side of each year, around the end of May. This dip likely represents the result of a 3-day weekend, memorial day.
- On average, the beginning of March has lower weekend dips than late March/April.

Based on these observations, I would recommend a weekend in the beginning of March to perform the maintenance. While Memorial Day Weekend has low ridership, it may not be ideal to perform the maintenance on a holiday weekend, as this may upset riders. **Therefore, the first weekend of March is the most ideal time to perform this 3-day maintenance project.**

<u>Maintenance Project 2:</u>
<u>7 Line Track Lubrication:</u>

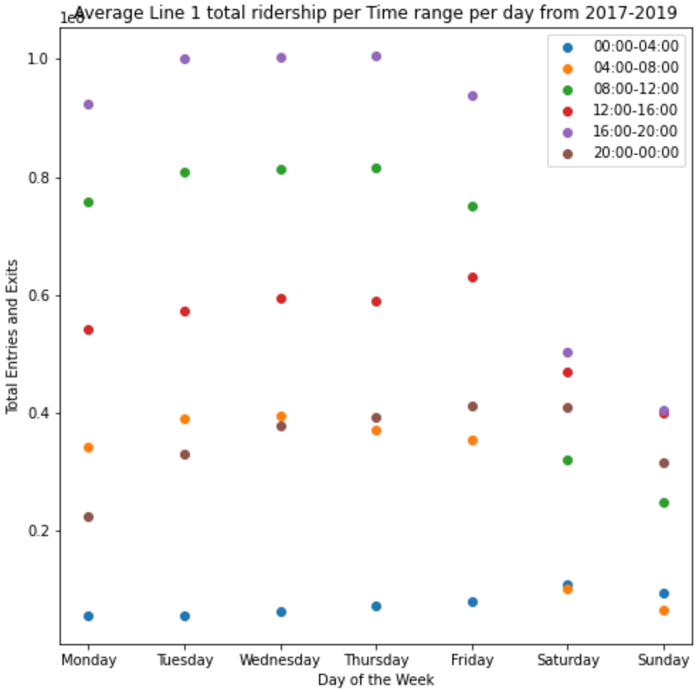


From the plot above, we can see the total entries and exits for different time intervals, plotted along an axis of each day of the week. We can observe a few general trends.

- On weekdays, average ridership is very high, with the busiest times the rush hour intervals of 16:00-20:00 (Afternoon rush hour) and 08:00-12:00 (Morning rush hour)
- The lowest ridership on weekdays occurs at 00:00-04:00
- Monday ridership is typically lower than other weekdays
- On weekends, the behavior is a bit different, with peak times being 16:00-20:00 and 12:00-16:00 (daytime).
- On weekends, the morning ridership is very low on
- Saturday ridership is consistently higher than Sunday

Based on these observations, the lowest ridership is on a weekday morning on the 00:00-4:00 time slot. **There aren't many people out late, due to riders having to work the next day. In particular Monday from 0:00-4:00 seems to have the least ridership, so that is the best time to perform this maintenance.**

1 Line Track Lubrication



For Line 1, when we compare to Line 7, for the most part the same observations hold. There are small shifts in the relative position of each data point, but the general trend is for the most part consistent.

When looking at the lowest ridership time interval of the week however, for Line 1, there is almost a tie between Monday 00:00-4:00, Tuesday 0:00-4:00, and Sunday 4:00-8:00. This can likely be attributed to the fact that Line 1 is not only a commuter line, but also a line that riders use to get around to functions, nightlife, and other activities around Manhattan. Therefore, people tend to stay out a bit later each night, and the ridership from 00:00-04:00 on weekdays is a bit higher than on Line 7.

Since the construction team will be performing maintenance on Line 7 at Monday 00:00-4:00, it is best to perform maintenance on Line 1 on Sunday 04:00-08:00.

Tools

The python tools used for this project are:

SQL for data loading
Pandas for data cleaning, aggregation, manipulation
Matplotlib for visualization

Communication

For additional information, please contact kenhua15@gmail.com. The project will also be posted on my github found here: https://github.com/kenhua15/MTA-Exploratory-Data-Project