EDA project description

October 7, 2021

Track Maintenance

Abstract

This project looks into the impact of the summer of hell periods in 2017 and attempts to solve the problem of sudden track repairs by analyzing most recent entries data and finding when the entries are low to schedule regular maintenance.

Design

In 2017, the governor declared a state of emergency and many tracks were out of service due to urgent upgrades. The difficulty is determining when to plan downtime for repairs and maintenance, given that the subway operates 24/7. The MTA asked for help in solving the issue.

Data

Starting with 262 MB of raw data used in the analysis was in two separate years:

- **July 2017** was used to investigate the summer of hell period.
- **Aug and Sep 2021** were used to determine the best time and date to shutdown a line for maintenance

Column used: Station, Line, Name, Date, Time and Entries.

Algorithms

1. Fetch:
    a. Created a UDF to read desired weekly csv files from the MTA website.
2. Clean:
    a. Striped all spaces from strings.
    b. Drop duplicates and unnecessary columns.
    c. Check for nulls and drop them.
    d. Calculate individual entries.
3. Aggregations and statistics:
    a. Get mean of individual entries for each day of the week.
    b. for the day with the lowest mean get the mean of individual entries for each hour.
4. Visualization:
    a. Initial plot of entries to spot the problem.
    b. Plot week days and times of the day over entries.

Tools

The tools that will be used in the EDA:

- IDE:

  VS Code, Python and Jubyter notebook extension by Microsoft.

- Libraries:

  Data manipulation:

    - Pandas
    - Datetime.

  Visualization:

    - Matplotlib
    - Seaborn