# Exploratory Data Analysis Project Writeup

Advice for NYC Metro Scheduling

Josh Jingtian Wang

1/23/2022

**Abstract:**

The goal of this project was to investigate the trend of the New York City metro ridership and factors that contribute to that trend. I worked with turnstile data from the Metropolitan Transportation Authority (MTA), and looked at how the COVID19 pandemic, rideshare services, weather, time of year and time of week potentially affect the subway traffic.

**Design:**

The COVID-19 pandemic has been devastating to many sectors of the society – public transportation being of them. In this project, the MTA has asked us to investigate when and where they can reduce the number of trains to reduce their budget. Identifying factors that affect the subway ridership will help the MTA maintain their service in a more cost-effective manner.

**Data Description:**

The primary dataset is the MTA turnstile data of New York City (2015-2022): <http://web.mta.info/developers/turnstile.html>

NYC weather data from the National Oceanic and Atmospheric Administration (NOAA) website (2017-2019): <https://www.ncdc.noaa.gov/cdo-web/search>

Rideshare services data will come from the NYC Taxi & Limousine Commission (TLC): <https://www1.nyc.gov/site/tlc/about/aggregated-reports.page>

**Algorithm:**

Turnstile data was grouped by day for the COVID line plot, year/week heatmaps, and the weather scatter plot.

Turnstile data was grouped by month for the ridershare line plot.

**Tools:**

Data retrieval was be done via SQLite. Data cleaning was be done via Python’s pandas package. Data visualization was done with Python’s matplotlib and seaborn packages.

**Communication:**

Please refer to the slides.