Improving MTA rider experience via analyzing contributing factors of additional waiting time and failures

Problem: While MTA transportation can be convenient for New Yorkers, delays and failures often frustrate riders. Oftentimes, the displayed wait times for subways and buses are incorrect and riders end up waiting longer than they are promised. Moreover, some riders are even sometimes stranded between stations as unprecedented failures arise. Through these issues, MTA subways and buses end up being an unreliable source of transportation. With the limited resources that the MTA has in labor and materials, they have to efficiently choose where to focus their efforts.

Importance: Thus, it is important for decision makers to identify which factors contribute to delays and failures. Factors such as which lines, boroughs, and seasons may shine light on how to recommend which subways and buses to efficiently focus their resources on.

Methodology: To identify these contributing factors, we will run 3 Weibull regressions with covariates for both subways and buses on:

- Mean distance between failures
- Additional platform/bus stop time
- Additional subway/bus time,

as well as a Negative Binomial regression on the number of bus failures. In total, there will be 7 regressions. We choose to use the Weibull distribution for mean distance because it's both continuous and positive. Similarly, we also use Weibull distribution to model both additional wait and travel times. Lastly, we use the Negative Binomial distribution to model the number of bus failures since we are counting a discrete number of occurrences.
In addition to the covariates available to us in the datasets, we will also design covariates for other variables. For example, to incorporate seasonality, we use dummy variables for the four seasons. We will also incorporate our own knowledge of MTA transportation to make logical categories such as east or west side subways or express and local subways.
In our final analysis, we will determine which factors contribute the most to mean distance between failures and additional wait time in order to recommend to the MTA which areas of transportation to focus on.

Datasets: We will be using four datasets:

- MTA Subway Customer Journey-Focused Metrics: Beginning 2015
    - Contains additional platform and train times in minutes and other relevant subway data such as division, line, and number of passengers
- MTA Subway Mean Distance Between Failures: Beginning 2015
    - Contains mean distance between subway failures in miles and other relevant subway data such as subway car type
- MTA Bus Customer Journey-Focused Metrics: Beginning 2017
    - Contains additional bus stop and bus times in minutes and other relevant bus data such as route, borough, and number of passengers
- MTA Bus Mean Distance Between Failures: Beginning 2015
    - Contains mean distance between bus failures in miles and other relevant bus data such as borough and type of bus service

We may omit data from 2020 as transportation behaviors were affected by the COVID-19 pandemic as most individuals worked from home and did not use the transportation to commute to work.