## ABSTRACT

The goal of this project was to provide the WomenTechWomenYes (WTWY) organization with assistance on how to effectively increase attendance to their yearly gala as well as increasing awareness of the lack of women in the tech industry. In order to achieve their goals, WTWY deploys a “street team” throughout the city’s subway stations to acquire email signatures from passengers who are commuting. By utilizing the MTA’s public turnstiles database, I was able to pinpoint what were the busiest subway stations in the city over the 2021 Spring to Summer time period to see which months were busiest and at what points of the week traffic was heaviest in order to effectively create a marketing strategy for next year’s gala.

## DESIGN

Web Scraping the MTA Turnstile data allowed for an estimate of passenger entrances and exits throughout each subway station in NYC. Given that the WTWY gala takes place in summer, the specific date range used for this analysis was from March 2021 - August 2021. I separated the analysis into a higher level overview of the busiest stations in New York City overall, followed by a further in depth analysis of the top 5 busiest stations. The deeper dive into these stations showed fluctuations in entries throughout the season, as well as a trend of heavier traffic mid-week when compared to weekends or the beginning of the week. Using these observations, WTWY could use these points to deploy a larger number of members to certain stations throughout the season.

## DATA

The database provided by the MTA as a whole has roughly 7.5 million rows in its entirety. Provided for us in their database, we found several unique turnstiles identifiers to help us identify which turnstiles we were looking at and in which stations. Each turnstile also provided a rolling sum of entries and exits that were calculated in 4 hour intervals. Although identifying each unique turnstile was not a crucial output to assist WTWY, through the data cleanup process, we were able to identify several issues in the MTA database, such as duplicative data, inaccurate entry counts, and incorrect names. Given these discrepancies, the cleanup process led to removing the inaccurate data entries, and using averages for each overall station rather than an actual sum of the data provided. This allowed me to get a more accurate estimate of the daily amount of foot traffic through each station.

## ALGORITHMS

- By grouping each data column that ties in with each unique turnstile (i.e C/A, SCP, Unit Numbers, and Stations), we were able to ensure that each unique turnstile only had one given entry for each 4 hour interval and remove any duplicate entries given.
- To reduce the possibilities of errors in the entry counts themselves, we used a function within the data frame to revert any counts given by a turnstile that counted over 1 million back to zero. We found that any counts over this threshold were highly likely to be errors as it would mean that roughly 42,000 people would be entering the station every hour. This equates to roughly 700 people per minute which was highly improbable.
- Graphing Data was aggregated by extracting date data from the database and grouped by months, or by day of the week and were set to average entrance values for each station.

## TOOLS

- Seaborn and MatPlotLib were used within python to plot and group data points and visualize the foot traffic from each station.
- SQLite was used for the initial database analysis, to give an overall representation of the data provided by the MTA.
- Pandas within Python was used for the more in-depth querying of the database.

## FINDINGS AND RESULTS

The MTA database showed rather inconsistent data, however, was able to give a high level picture on the overall patterns of foot traffic throughout the city. Trends that are likely centered around tourism season were notable - as the more tourist heavy stations had larger fluctuations than stations that serve as transportations hubs throughout most of the year. The five stations that required the most focus for a marketing strategy were Penn Station, Herald Square, 86th Street, PATH New WTC, and Fulton St. Out of these top stations, the New WTC station, Herald Square, and Fulton St stations were found to be tourist and season dependent as the largest upticks in volume were seen during the summer when compared to regularly trafficked stations such as Penn Station.