# Collecting the most about of signatures for the WTWY GALA with most efficient predictive schedule

Amir Khoeilar

## Abstract

The Women Tech Women Yes gala that is going to be held early summer ( June ) 2022, needs to attract attention and have as much attendance as possible. I have approached the data from the Metropolitan Transportation Authority (MTA) for New York City in order collect information regarding the high traffic station in order to collect as many emails (by street teams) for sending out tickets from the 5.5 million riders of NYC. In order to optimum situation for the WTWY gala invitations, I have used many graphs and tables in order to provide a specific timetable for the street teams to approach the high traffic stations at certain days and times of the week.

## Design

The WTWY organization holds a gala at the beginning of summer each year due to attract attention in regards of women working in the technology field and fundraise in order to pave an easier path each year in that aspect. Having access to street teams, the organization is going to use this privilege to approach people at NYC subway stations to gather signature and send out tickets. Knowing the gala being held in June 2022, we need to gather data about subway stations which is one of the mostly used public transportations of the city where street teams are mostly going to be active and successful with our goal. By determining the highly traffic stations while knowing the traffic of the stations during times of the day and days of the week, the street teams can perfect their task of invitations. ( also not repetitive people )

## Data

Knowing the Galla being held in in June 2022, I have approached the data set of MTA turnstile data for the subway stations of New York City for the months of March, May and June of 2021.

The dataset contains 2717757 of data showing entries and exits for very turnstile in every station in the entire NYC. Each data is defined with 11 attributes. Besides date, time, entries and exits and stations which are the obvious characteristics of a turnstile, couple of other important attributes are: C/A =  Control Area,Unit = Remote Unit for station,SCP = Subunit Channel Position representing an specific address for the device,DESC = ‘REGULAR’ scheduled audit event (every 4 hours).

By analyzing the data for the turnstiles and eventually the eateries and exits for each station on each of these 3 months, we can collect a great understanding of traffic ( eateries + exits ) for different stations during days and hours of these months and

ultimately project that for our prediction of traffic for the year of 2022, three months before the WTWY gala.

## Algorithms

*Feature Engineering*

1. Acquiring data from the MTA website, reading and studying it using SQL and DS Browser.
2. Going through data in DB browser using simple code to find any NULL in the data set.
3. Using SQLalchemy library in Jupiter notebook in other to implement same kind of observations inside Jupiter notebook.
4. Eventually using pandas and numpy library in order to manipulate data given and finding the outliers and extraordinary data.

*Models*

Pandas was the main help in order to come up with specific meaningful tables in order to plot graphs supporting our cause. Our cause being coming up for a solution for the WTWY organization to collect as many possible signatures. With the help of Matplotlib and seaborn, I reached the desired graphs that described traffic in stations properly.

*Model Evaluation and Selection*

In the conclusion, with the help of the libraries mentioned above and the graphs, I achieved the highly populated (ridership traffic) for all station. Besides that found out that over the 3 month of my data a cumulative of 17 mill people will be passing by the stations.

I have provided a simple chart for the WTWY organization to send out their street teams in order to collect as many signatures as possible. There has been some pessimistic assumption made that in the long run would help WTWY organization reach its certain attendance for the gala in June 2022.

## Tools

- Numpy and Pandas and SQLalchemy for data manipulation
- DB Browser and SQLite Querying from that database and observation
- Matplotlib and Seaborn for plotting
- Mac Keynote and Numbers for the presentation

## Communication

There is going to be the code for data manipulation and also PDF slides of powerpoint presentation available on my GitHub account.

https://github.com/rezxkoi/WomenTechWomenYes/tree/layer