**Project Proposal**

**Morooj aldeeb**

# **Introduction:**

Turnstile data is a cumulative time series data that represent transportation traffic of railways in different US stations. The majority of records in the data were collected at fixed time intervals (4 Hours) per device. Each Turnstile device is represented by four different categorical variables. Each record represents an independent record for a turnstile device in a particular station.

# **Initial Data Features:**

Following the original web resource, the data set records have been collected based on 11 fields. Most of the fields are text-based categorical variables, one date and another time field in addition to 2 numerical fields that represent cumulative entries and exits from each station.

Each data record describes a single recording event every 4 hours by a single device deployed at a given identified train station.

# **Data Analysis Approach:**

In this project, we are going to propose a set of questions which tailor or guide the overall data analysis approach that will be taken in order to better understand the data and gain more insights from it to instruct future decisions. We are going to collect 12 Month(s) worth of data from the whole dataset for explanatory data analysis and hypothesis derivation.

## **Research Questions:**

1. Which month(s) are considered busiest? Which week(s) in each month are considered busiest?
2. Calculate transportation efficiencies among different train stations.
3. Is there any significant association between size of the train station and the transportation efficiency of that station?
4. Detect device(s) anomalies in recording data per each train station.
5. Is there any significant relationship between size of the train station and the traffic movement? Not only, we can test if there is any association between size and traffic movement but also, we can try to predict the size of the train station based on knowing the exits and entries in a station which might help authorities in designing better train stations to accommodate in advance to the influx of people at specified geographical location.
6. We can download weather data set from public weather resources for different US states, afterwards, we can merge both data sets by date.

We can investigate the traffic flow in train stations in correspondence with

the weather states, we can deduce if weather conditions had any

effect on the transportation and traffic flow of people in train stations and.

whether this effect is statistically significance and not only that but also, we

can predict the traffic flow given a particular weather condition. this

prediction will be highly effective in traffic control

The target/outcome variable in this case would be the number of entries for a particular station and the explanatory variables will be the weather condition.

**Data Description:**

* in the weather database I will be using year (2018) of the data. We will use date column and link the date column to the weather database.

## for example, I will use the entries and exists numbers of people to know the traffic for a station.

* In the MTA data I will be using the data of year 2018.

## **End User Benefits:**

- The conclusions from this explanatory data analysis project are very important to

different audiences, including but not limited to the government of US, Transportation

ministry in US, Marketing companies and passengers themselves.

- I want to take advantage of crowded stations and stations in poor places to publish

awareness leaflets about weather conditions in crowded and poor neighborhoods.

## **Data Modelling:**

We are going to use multiple linear regression to regress on train station size (Target

variable) using both entries and exits as explanatory variables.

## **Used Tools:**

* Jupyter notebook
* Python

**Used libraries**:

* Pandas.
* NumPy.
* SciPy.
* Matplotlib.
* Seaborn.
* SQLite
* Panda profiling.
* Scikit-learn