# MTA Turnstile EDA

## Abstract:

The goal of this project was to leverage turnstile usage data in order to guide the Starbucks franchise in selecting a location to open a new store at, thereby expanding their customer reach. I worked with data provided by the Metropolitan Transit Authority (MTA), along with United States Census data, which provided average household income and population information, specified for various zip codes across New York City. After gathering this data, I imported it to pandas for exploratory data analysis, I used matplotlib and seaborn to visualize my results.

## Design:

The design or approach of this project was achieved in 5 steps:

1) Acquiring data from the MTA website and selecting a pertinent time frame for the objective. Since I am recommending potential locations for a new Starbucks store opening, I wanted to select a time frame that would allow the client to understand what changes take place, if any, at the turn of a season. For that reason, I select data between Feb 1, 2019 to April 30, 2019; to show how turnstile usage changes as we drift from cooler to warmer temperatures. (Unfortunately, because I was crunched for time, I wasn't able to sift through the data to analyze these seasonal changes in my project.
2) Familiarizing myself with the data to understand what each column meant, what entries/ exits were tracking
3) Cleaning the Data to remove nulls, outliers and create a datetime series.
4) Exploring the data in conjunction with demographic data to evaluate busiest stations which were located in densely populated zip codes with high average household incomes
5) Created visualizations of busiest stations

## Data:

1. MTA Data: http://web.mta.info/developers/turnstile.html
2. United States Census: https://data.census.gov/
3. CSV file with stations and zip: https://raw.githubusercontent.com/wnobles/Project1/main/stations_zip.csv

## Algorithms:

1. Ingested MTA turnstile data and extracted into python
2. Function to import data from mta link
3. Creation of datetime column using date and time columns
4. Calculating foot traffic by subtracting incremental exits from incremental entries (date vs the previous date)
5. Removing nulls and outliers
6. Grouping dataframe by station and sorting mean foot traffic of each station.
7. Importing stations.zip file and joining it with turnstile data
8. Importing demographic file from US census and joining with dataframe from set 6.
9. Sorting mean foot traffic having grouped by stations, discovering top 10 busiest stations, comparing these stations to those that also have an average household income above 60K

## Tools:

- SQLAlchemy
- Jupyter
- MATPLOTLIB
- Seaborn
- Pandas

## Communication:

This project aimed to recommend a subway station with the highest foot traffic to Starbucks so that they could consider it when searching for locations to expand their franchise to. Because Starbucks is considered a premium coffee brand, its target customers have average household incomes of at least 60K. Using the MTA data in conjunction with the demographic data, I suggested two potential locations to Starbucks: 14th St-Union Square and 72 st. Both were in zip codes where the income bracket was above 60K, and both had high levels of foot traffic, indicating that it was a highly frequented station. In order to improve on this project, I would've

looked further into how foot traffic changes throughout the day, week, and season, as it would allow Starbucks to customize their service according to the implications of those discoveries.

1. Acquiring Data - Weekly subway turnstile data since May 2010 is available on the MTA's website. Acquiring the data was a simple matter of pulling several week's worth of data from the site. We decided to use May 2016 data to approximate traffic during the anticipated deployment time period of May 2017.
2. Understanding the Data - Turnstile counts were tracked as cumulative entries and exits, for several measures of time during the day, for an individual turnstile. To see patterns in the data by station, it became apparent that we would have to aggregate the data across individual time intervals for each station. To make things a little more complicated, the time measures were not uniform and there were occasions where turnstile counts were reset or irregularly audited.
3. Cleaning and Prepping the Data - This was primarily executed using Pandas and included:
    - Removing trailing whitespace
    - Parsing date time values
    - Calculating incremental entries/exits
    - Removing outliers
4. Exploring the Data - To investigate our project objectives, we performed the following aggregations to understand which stations had the highest volume and how turnstile traffic varied by time:
    - Time of day
    - Day of week
    - Station
5. Creating Meaningful Visualizations - We wanted to share our findings with others, and as the old adage goes, a picture speaks a thousand words. Each graph we plotted illustrated a key point.