DATA 608 - Final Project Proposal

Michael D'Acampora

10/30/2019

Background

There are approximately 472 subway stations in NYC each with numerous turnstiles that count passenger entries and exits. This data is made public by the Metropolitan Transportation Authority.

Proposal

The goal is to analyze how the data is given in raw form, aggregate and summarize turnstile data for each subway station and create a searchable dashboard in a shiny app.

It would be interesting to know which stations and subway lines are the busiest, as well which subway divisions. Divisions are essentially split between the numbered lines and the lettered lines, which historically goes back to when there were three private companies operating the subway in New York. You may have noticed the numbered trains (1/2/3, 4/5/6) are smaller than the lettered lines (A/C/E, N/R/W). It is because there were design differences between the divisions.

Finding anomolies in the data may be able to help the MTA find better ways to locate turnstile jumping or develop better controls for counting entry and exit of passengers. Other subway enhancements like improving bottlenecks or altering scheduling could also be considered.

Data

It will take some time to understand the data set, since each turnstile has its own SCP number and each of those must be grouped for each station. There may be other turnstiles in this data set that aren't necessarily subway turnstiles that must be taken into considerations also.

Additionally the entry and exit counts are cumulative, like an odometer. Since the data set is a week's worth of turntile information, the reading at the end of the week must be subtracted from the beginning of the week for each turnstile in order to obtain the weekly count.

The data is provided in CSV format with 11 features:
C/A = Control Area (A002)
UNIT = Remote Unit for a station (R051)
SCP = Subunit Channel Position represents an specific address for a device (02-00-00)
STATION = Represents the station name the device is located at
LINENAME = Represents all train lines that can be boarded at this station Normally lines are represented by one character. LINENAME 456NQR repersents train server for 4, 5, 6, N, Q, and R trains.
DIVISION = Represents the Line originally the station belonged to BMT, IRT, or IND
DATE = Represents the date (MM-DD-YY)
TIME = Represents the time (hh:mm:ss) for a scheduled audit event
DESC = Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours)
>>1. Audits may occur more that 4 hours due to planning, or troubleshooting activities.
>>2. Additionally, there may be a "RECOVR AUD" entry: This refers to a missed audit that was recovered.
ENTRIES = The comulative entry register value for a device
EXIST = The cumulative exit register value for a device

The data can be found here. http://web.mta.info/developers/turnstile.html