MTA Turnstile Data: Exploratory Data Analysis

A case study using the results of the analysis to know the times of overcrowding in order to give the largest number of people a vaccine for the Covid 19 virus.

A Proposal by:
"Tahani Aldosari"

Beneficiary of MTA Turnstile Data: Exploratory Data Analysis

The beneficiary of the data analysis in this project is the Ministry of Health, which in turn tries to reach the largest number of recipients of the Covid 19 vaccine, including people at metro stations, which is a large number throughout the day. In particular, the beneficiaries of this project are the corona vaccine work team, including doctors and nurses, who are entrusted with the task of giving the vaccine to all people in the country.

Problem Statement

These days, we are witnessing a wide campaign by governments and countries to prevent the spread of the Covid 19 virus and try to reduce human losses by various means. Governments, especially ministries of health, form committees to give the largest number of people the necessary vaccine to overcome the spread of infection among them. What the committees need is for a large number of people to be present at the same time to speed up the process of giving the vaccine and save time. Among the means that lead to the gathering of the largest number of people are the train or metro stations, but it is necessary to know the peak times when these stations are crowded with people. To do this, we will go in this project to study and analyze historical data belonging to the "MTA Turnstile Data" and extract useful indications about the days most crowded with people and what exactly are the most filled stations.

Project's Questions:

From the above, we conclude two main questions that lead us to produce the current project using data analysis:

1. What are the peak times when these stations are crowded with people.
2. what exactly are the most filled stations.

Dataset Description

The dataset used in this project is a CSV file published weekly by MTA (Metropolitan Transportation Authority), and adds the most recent download link to the top of a very long list on their website. There are no fancy open data portal frills here, just CSVs and some supporting files to explain what each field is, and how to connect a row of data to a specific subway station. The dataset is the latest update of the data (Saturday, September 25, 2021) which contains 210401 rows.

The following is the description of each columns in the datatset:

```
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
```

Tools and technologies

Technologies:

1. panads.
2. Seaborn.

3. Sqlite3.
4. matplotlib.

Tools:

1. python.
2. Sql.
3. Jupiter notebook.