MTA Project Report

Data Science T5 Boot camp

Name: Faisal Mirza

Gmail: faisal.a.mirza11@gmail.com

GitHub: faisalmirza11

Abstract:

This data science project focuses on finding the NY subway stations with the least amount of people, traffic to enable a film production company to successfully film a movie in the subway without interruption or for them to get a warrant to close off a station in a date and time where not a lot of people need to use the subway, by using the MTA data and by doing some Exploratory Data Analysis and visualizations we attained these low traffic stations successfully.

Framing question: What is the best time and location for having the least amount of people in the subway for filming a movie?

Who benefits: film production companies seeking to film in the subway. Example: Universal Pictures.

Design:

Background: My friend Yasser is a movie writer that works at a film production company called Universal Pictures, he told me that he came up with a horror movie idea that fully takes place in the New York City subway, but in order to film a movie in the subway, the production company needs to know where and when are the best places and times that they can film a movie at that had the least amount of people. I told my friend that the subway transit traffic data is provided online by the MTA, and that me being a data scientist who also happens to like horror movies, I got excited and I offered that I could help them in finding the best stations and times that had the least amount of people so that they could film a movie there. So, my friend Yasser told the production company about me, and they agreed for me to help them.

Data:

The dataset used is provided by the New York Metropolitan Transportation Authority (MTA) and it contains information collected from the turnstile devices in the NY subway stations, some of the columns of this data include:

- Control Area, Unit, SCP: which all represent an individual turnstile.
- Station: represent the station name where the turnstile is located at.
- Entree, Exit values: which shows the number of entries, exits of the station cumulative.
- Date, Time: which represents the date and time of a snapshot of the turnstile info.

Scope:

- The data used, covers a three-month period which is from 3/17/2018 to 6/22/2018 where it is not a summer break to avoid late outgoing tourists, residents, and it's in 2018 to avoid the effects of COVID-19.
- The dataset used had 2.7 million rows, and it had 11 columns.

Algorithms:

I started by viewing the dataset to understand it clearly, then after looking at the data and understanding it, I started to look for issues and problems in the data that required cleaning, and some of the cleaning that I did includes:

- Removing duplicates
- Removing rows where the DESC column equals RECOVER AUD
- Removing rows where the daily entree was equal to zero

Feature engineering:

- Converting DATE, TIME from type object to type date time, and combining them in a new column called DATE_TIME.
- Getting the daily entries for each turnstile each day, and putting it in a new column in a new masked data frame (df_daily) grouped by C/A, UNIT, SCP, STATION by using diff for the entries.
- Handling negative entries and converting them by taking their absolute value, and when a turnstile counter resets we let it return a 0 and putting these fixed values in a new column called FIXED_DAILY_ENTRIES

Then after cleaning the dataset and performing the feature engineering I started to plot and visualize the data using Matplotlib, I used the median to find the least daily entree stations and then I plotted them.

Tools:

Technologies:

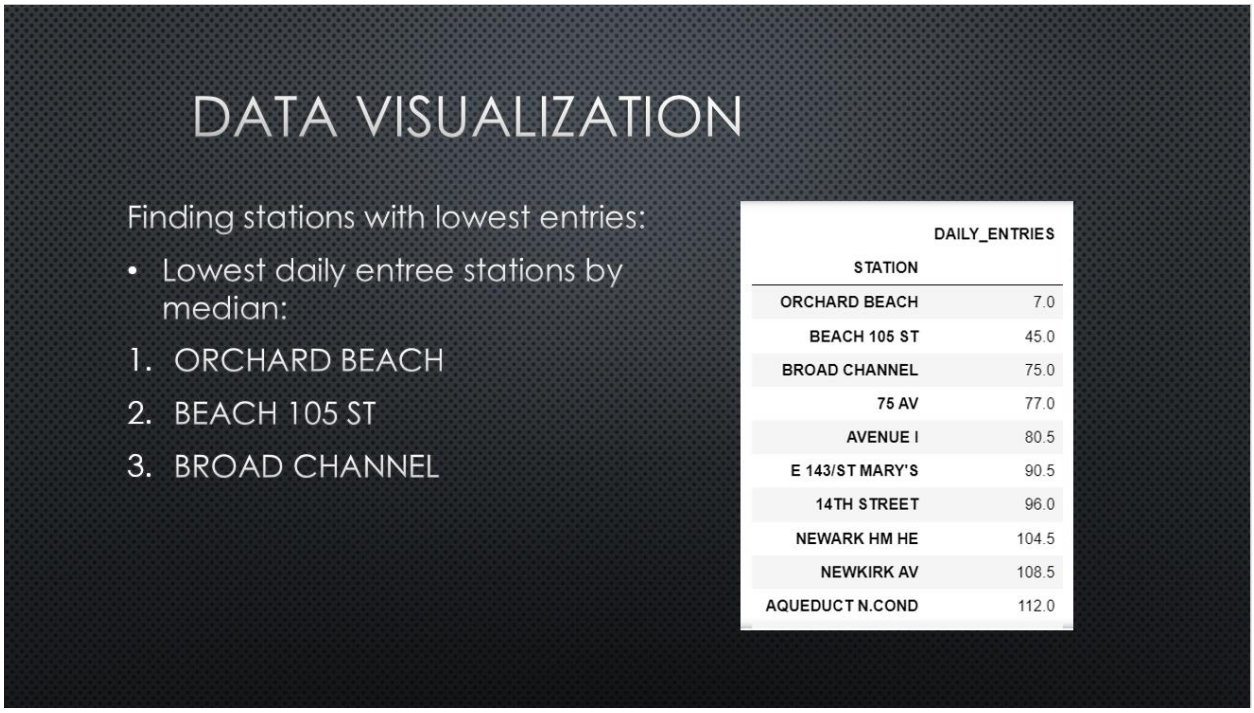
- SQL, sqlite3, to store the data and to manipulate it using queries
- Python, Jupiter Notebook, to perform the EDA

Libraries:

- Numpy, Pandas for data manipulation
- Matplotlib for plotting the data
- Datetime for date manipulation
- Os for importing the data

Communication:

Everything related to this project will available on my GitHub at faisalmirza11 in the repository DS-T5-projects.



STATION	DAILY_ENTRIES
ORCHARD BEACH	7.0
BEACH 105 ST	45.0
BROAD CHANNEL	75.0
75 AV	77.0
AVENUE I	80.5
E 143/ST MARY'S	90.5
14TH STREET	96.0
NEWARK HM HE	104.5
NEWKIRK AV	108.5
AQUEDUCT N.COND	112.0

DATA VISUALIZATION

BEACH 105 ST

- Most daily entries are around 100
- best days are from 3/31/2018 to 5/23/2018
- Overall a low traffic station