

Exploratory Data Analysis (EDA) for The NYC Blood Center on the MTA Dataset



Haneen AlHomoud

Abstract

This project was conducted for the T5 Data Science Boot Camp, where we performed Exploratory Data Analysis (EDA) on the MTA turnstiles data. The project aims to give an excellent geographical plan to a blood donation center located in NYC to allocate the centers, busses, and booths properly, take advantage of people planning to visit the stations and reach the highest number of donators visiting a specific area station.

Design(Background company)

- **Company info:** Blood donation center provides busses, booths and centers designated for blood donations.
- **Problem/opportunity:** The lack of precise geographical informatics to allocate the centers busses and booths properly in order to reach the highest number of donators.
- **Value for the company:** the ability to receive more donations, and reduce the operational fees.

Data

About MTA:

The Metropolitan Transportation Authority is North America's largest transportation network, serving 15.3 million people across a 5,000-square-mile travel area surrounding New York City through Long Island, southeastern New York State, and Connecticut. The MTA network comprises the nation's largest bus fleet and more subway and commuter rail cars than all other U.S. transit systems combined. The MTA's operating agencies are MTA New York City Transit, MTA Bus, Long Island Rail Road, Metro-North Railroad, and MTA Bridges and Tunnels.

Data Description:

Data obtained from <http://web.mta.info/developers/turnstile.html>.

Columns description:

Field Name	Description
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)
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)
ENTRIES	The comulative entry register value for a device
EXITS	The cumulative exit register value for a device

Algorithms

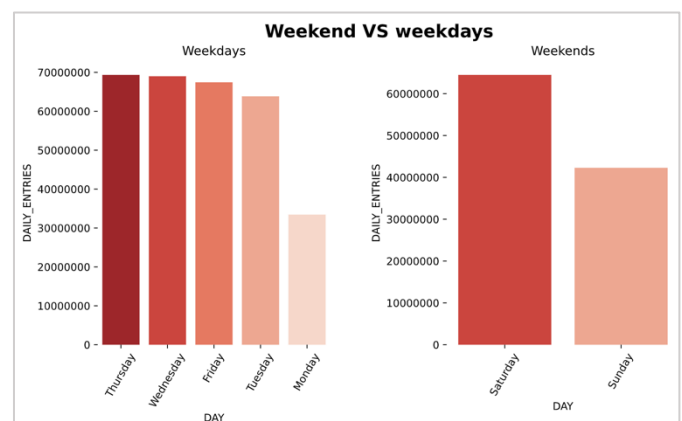
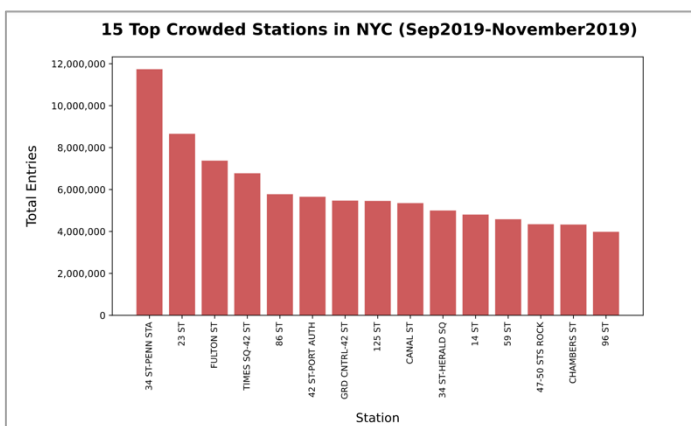
1. Problem Understanding
2. Dataset Exploration (Data Cleansing)
 - Null Values
 - structural errors
 - Outliers
 - Duplicated rows
3. Exploratory Data Analysis (EDA)
4. Insights
5. Conclusion

Tools:

- Technologies: SQL, SQLite, Python, Jupyter Notebook
- Libraries: Numby, Pandas, Matplotlib, Seaborn, sqlalchemy, sklearn

Communication

- Charts:




- Presentation snips:

TS DATA SCIENCE BOOTCAMP

Exploratory Data Analysis (EDA) for The NYC Blood Center on the MTA Dataset

Prepared by: Haneen Alhomoud




New York City Blood Center

Problem Description



- Nowadays, not enough people donate blood
 - Only 4 people out of 100 who are able to donate blood do so
- The blood supply needs to be replenished constantly
 - Blood products have a short shelf life, from 5 to 42 days

 The New York Blood Center stated that it's difficult nowadays to find blood donors

Thursday, 9 September 2021 https://www.hofstra.edu/studentaffairs/blood/blood_why.html

1

Metropolitan Transit Authority (MTA) Dataset

3 Files:

January, 2019
February, 2019
March, 2019

Rows
2,676,241
Columns
11

Thursday, 9 September 2021

2

Dataset Exploration (Data Cleansing)



Null Values
None



Duplicated rows
None



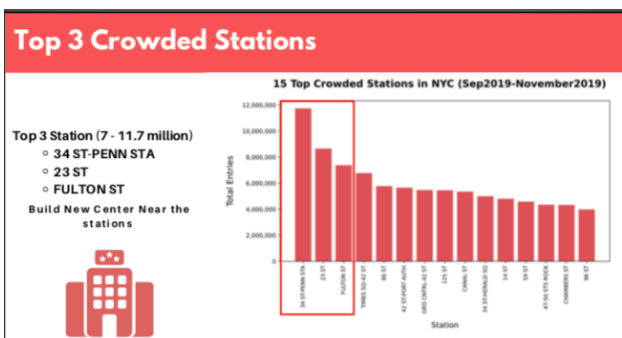
structural errors
None



Outliers
Entries-Exits

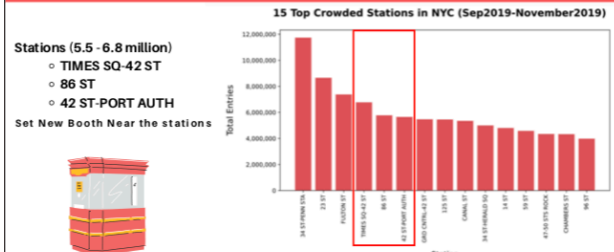
Thursday, 9 September 2021

3



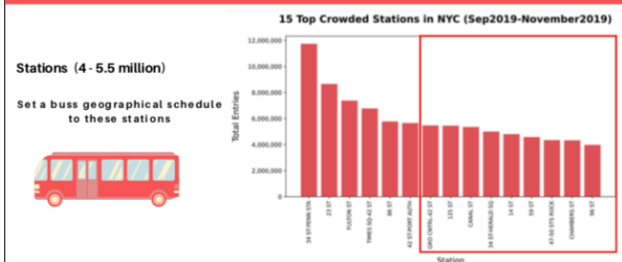
5

Stations With Above 1 Million Visitors



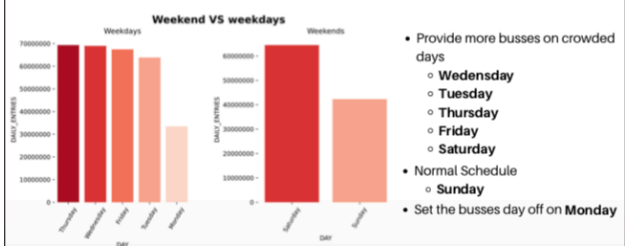
6

Crowded Stations



7

Weekends VS Weekdays Station Visitors



8