أكاديمية سدايا
SDAIA Academy

Data Science Bootcamp

Project Proposal 1

Exploratory Data Analysis (EDA)

For Metropolitan Transportation Authority (MTA)

Presented by: Ashjan Basri

2 September, 2021

Introduction

The Metropolitan Transportation Authority (MTA) conducted a new survey demonstrates that subway riders are more concerned about crime and harassment than the pandemic COVID-19 [1]. The 33.000 people participated in the survey and 87% said the crime the main factor to stop people to use New York subway. Based on the number of crimes and incidents that are increasing on the news on daily basis.

Project Motivation and plan

This project is proposing an analysis study to increase the safety of New York subways and reduces the crime rates. Our solution is to find the optimal subway stations and time to place and distribute police officers who have the task of observing the station areas and facilities. The guideline of this project starts with exploring the MTA dataset, visualising some significant, cleaning the data and apply some aggregation methods to analyse our data.

Data Discerption

Our project uses an opensource dataset MTA turnstile [2], we are aiming to scrape a total of three months of the data. The description of the dataset as follow:

Variables	Discerptions
C/A	Control Area (A002).
UNIT	Remote Unit for a station.
SCP	Subunit Channel Position represents a 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 represents 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 than 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 cumulative entry register value for a device.
EXIST	The cumulative exit register value for a device.

Data Analysis and Tools

Every raw in the data set represents a single turnstile to record the number of entries and exits. Every four hours the turnstile records the number of entries. In our project we will select 3 months of the data starting from April, May and June 2021. The main variables that we will use in our analysis is the Control Area/Unit/Station/DAET/TIME and finds the least stations station are not busy with rider and count the times daily to see the times distribution versus the number of riders.

In our analysis first, we will use some methods to clean the data such as drop duplicate values, missing values, Null values and outlier values. Then we will use a visualisation tool such as pandas, seaborn and matplotlib to see the data for further analysis. Next, we will apply some statistical techniques for our analysis finding the MIN stations that are not busy, finding the MEAN of those stations also, finding the difference in entries on every day on the week.

Conclusion

In conclusion this project use MTA dataset to find insights to improve and reduce the crime rate in NYC public transportation. We will use EDA techniques such as cleaning, visualisation and statical analysis to come up with the final conclusion in the next phase of the project.

References

[1] Eyewitness,abc,MTA survey suggests it's not just COVID keeping riders away from subway accessed on 1, august,2021, http://web.mta.info/developers/turnstile.html.

[2] Metropolitan Transportation Authority, accessed on 1, august,2021 http://web.mta.info/developers/turnstile.html.