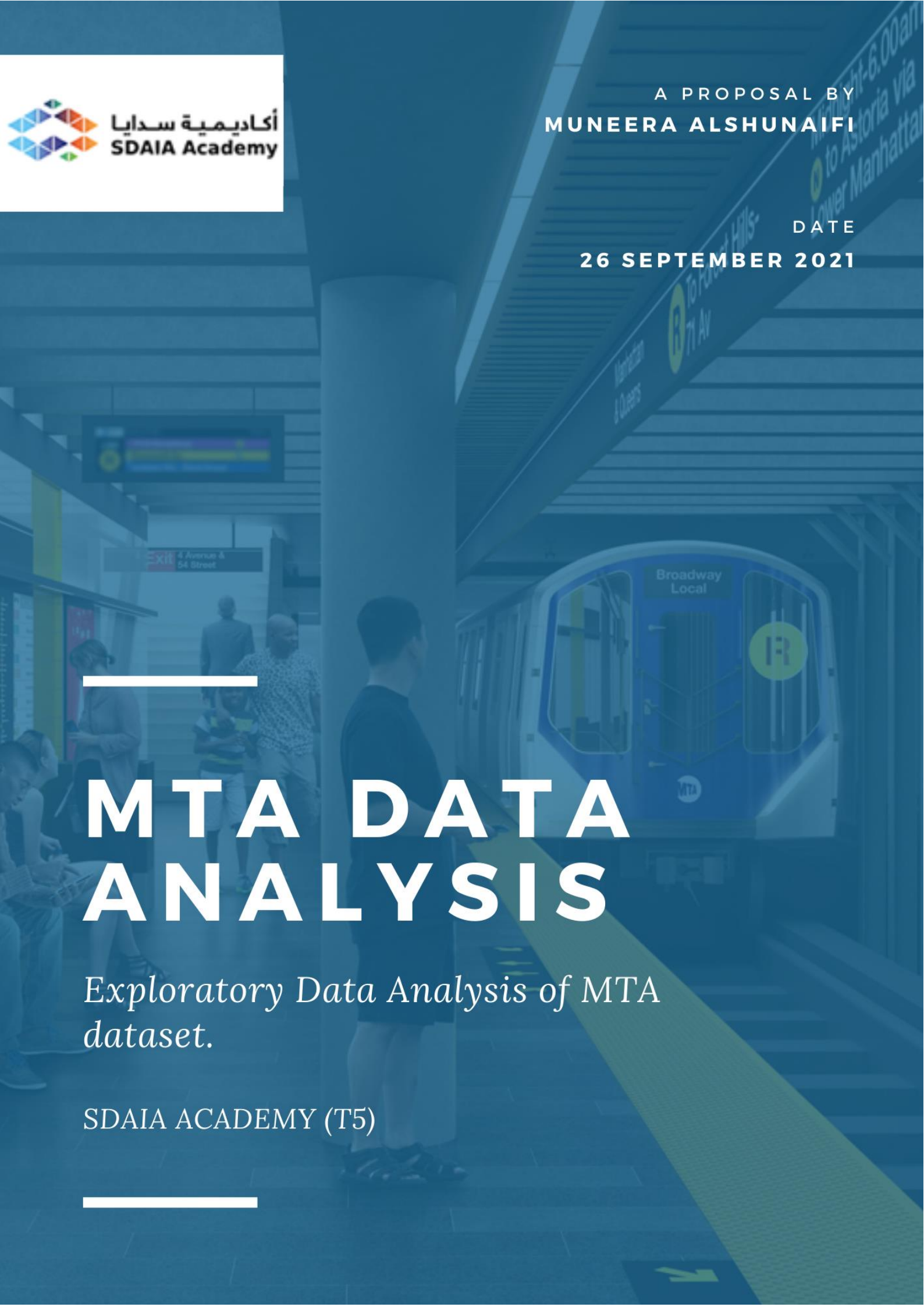
أكاديمية سدايا
SDAIA Academy

A PROPOSAL BY
MUNEERA ALSHUNAIFI

DATE
26 SEPTEMBER 2021



MTA DATA ANALYSIS

Exploratory Data Analysis of MTA dataset.

SDAIA ACADEMY (T5)

Table of Contents

Introduction

The Metropolitan Transportation Authority (MTA) is a public benefit corporation responsible for public transportation in the New York City metropolitan area of the U.S. state of New York. The MTA is the largest public transit authority in the United States, serving 12 counties in Downstate New York, along with two counties in southwestern Connecticut under contract to the Connecticut Department of Transportation, carrying over 11 million passengers on an average weekday systemwide, and over 850,000 vehicles on its seven toll bridges and two tunnels per weekday. [1]

In this project, I will be using Exploratory Data Analysis approach as known as EDA, which aims to analyze and investigate data sets and summarize their main characteristics mostly by employing data visualization methods. [2] However, I am planning to open a coffee shop at one of the Metropolitan Transportation stations, to do so, the EDA will facilitate the process of measuring the decision of opening a coffee shop at one of the Metropolitan Transportation stations whether is good and profitable idea or not; That is, by posting ADS and notice the people interaction.

The challenge is to know which stations are the most crowded because it will be our target to post the ADS in. Additionally, knowing the rush times (e.g., days, hours) will help me enhance the ways of posting ADS (e.g., adding digital ads). Also, I want to measure COVID-19 effect on the Metropolitan Transportation stations in order to know how to deal with related phenomena while opening the coffee shop in the future.

Data Description

The data will help us understand the nature of metro stations' congestions days, hours, and stations. Moreover, MTA is a website that provides a series of data files containing numbers of cumulative entries and exits by stations, turnstile with their dates and time specified. The metro data records are weekly produced and mainly collected every 4 hours.

To carry out the insights, I will use the data of six months starting from October 2020 until March 2021. The Features provided in the dataset are:

- C/A = Control Area (e.g., A002) which is a string
- UNIT= Remote Unit for a station (e.g., R051) which is a string
- SCP = Subunit Channel Position represents an specific address for a device (e.g., 02-00-00) which is a string
- STATION = Represents the station name the device is located at which is a string
- 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. which is a string
- DIVISION = Represents the Line originally the station belonged to BMT, IRT, or IND which is a string
- DATE = Represents the date (MM-DD-YY) which is a data type
- TIME= Represents the time (hh:mm:ss) for a scheduled audit event which is a time type
- DESc= Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours) Audits may occur more than 4 hours due to planning, or troubleshooting activities.

Additionally, there may be a "RECOVR AUD" entry: This refers to a missed audit that was recovered. Which is a string

- ENTRIES = The cumulative entry register value for a device which is integer
- EXIST = The cumulative exit register value for a device which is integer I will add more features which are the following:
- Turnstile_location = which is a combination of C/A + unit + SCP can be used to locate the near by places around the turnstile on google map
- entries_num = which is the number of entries for the station timestamp observed by taking the difference of the cumulative entries and the previous one.
- exits_num = which is the number of entries for the station timestamp observed by taking the difference of the cumulative exits and the previous one.
- Weekday = the weekday name to distinguish between weekdays and weekends.
- Congestion = which is the number of entries and exists added up to know how busy the station is.

Tools

To explore and analyze the data, I will be using Jupyter notebook to use python language and Python libraries, such as:

- Matplotlib and Seaborn for data visualization.
- Numpy and Panda for data read and write operations

Conclusion

MTA Turnstile data analysis and exploring will give us insights about the congestion areas which can help in opening a coffee shop by posting some ads, this can be carried by using python data visualization and manipulation libraries. The expected results is to have crowded stations on the weekends and morning work hours therefor more ads can be posted either digitally or the traditional way, also stations near to work and entertainment areas can have more turnstile because it's expected to have more entrances than others. Moreover, I am expecting to get fewer usage of MTA stations during last months of 2020 due to COVID-19 pandemic and lack of vaccine.

References

[1] https://en.wikipedia.org/wiki/Metropolitan_Transportation_Authority

[2] http://web.mta.info/developers/resources/nyct/turnstile/ts_Field_Description.txt