By Nouf Alotaibi
Email: noufmitla@gmail.com

METRO MOTION

1. Introduction

This is the first project for the T5 Data Science Bootcamp, which is an exploratory data analysis of the MTA turnstile data set using SQL paired with Python and its libraries *Pandas*, *NumPy*, and *Matplotlib*. Below is a detail about the company I am collaborating with to help them with their problem, the dataset description, and the tools I will be using for the project.

2. Background

2.1 Company Information

The "*Metro Motion*" is a company that provides repairing and cleaning services for metropolitan transportation in New York City. The company is serving a lot of stations that belong to the Metropolitan Transportation Authority.

2.2 Problem Statement

Metro Motion struggled a lot after the pandemic ended. The company's schedule got messed up because many people are returning to use the transportation services, and their schedule is not meeting the right times for repairing and cleaning. Therefore, the company is wasting a lot of resources. So, I got approached by the company to help them improve their work schedule by estimating the appropriate times for repairing and cleaning. While working on this problem, I will be answering the questions below.

- What are the most active/inactive times for each station?
- What are the most active/inactive days of the week?
- When a station gets busy, does the satiations that are near getting busy too? Is there a relation?
- Is there a similarity between that dataset prior covid-19 and dataset now?
- What is the best schedule for the company?

2.3 Value for the Company

The value for the company is that it will save the company's resources and improve their work quality and reputation as a company.

3. Dataset

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 publishes their data in CSV files every week. The table below illustrates the dataset's features/columns and their types.

Column Name	Column type
C/A	Object
UNIT	Object
SCP	Object
STATION	Object
LINENAME	Object
DIVISION	Object
DATE	Object
TIME	Object
DESC	Object
ENTRIES	int64
EXITS	int64

3.1 Scope

- **Sample Size:** I will be using a dataset before covid-19, so I chose October and November in 2019, and I will concatenate it with a new dataset for June and July from 2021.
- **Rows:** 3,740,978 rows.
- **Columns:** 11 columns, which are c/a, unit, scp, station, linename, division, date, time, desc, entries, and exits.

4. Tools

These are the technologies and libraries that I will be using for this project:

- **Technologies:** SQL, SQLlite, Python, Jupyter Notebook.
- **Libraries:** Numby, Pandas, Matplotlib, Seaborn.