MTA Turnstile Exploratory Data Analysis

Abstract

This exploratory data analysis is done based on the MTA turnstile data, which is a collection of the cumulative traffic data in the New York City subway stations. The analysis is done to find out which stations, at which dates and times are the busiest using the entries and excites record in the stations. This data can be used by a coffee company that has multiple shops near some of the stations. The results will help the coffee company to come up with a plan on when to open their shops after the covid-19 closings and how to allocate their resources and also plan new openings.

Design

This data analysis is designed to:

- Collect the required data from the MTA website and put it into the local database
- Clean the data
- Perform the data analysis using python libraries; Pandas, Matplotlib, Seaborn

Data

The data used for this project is a 14 weeks period Turnstile records for Dec 19, 2020 to March 20, 2021. The data shows cumulative entry and exit counts for each station at a four hour interval each day.

Algorithm - Example

```
import urllib.request

url =
"http://web.mta.info/developers/data/nyct/turnstile/turnstile_{}.txt"
week_nums = [210320,210313,210306,210227, 210220, 210213, 210206,
210130,210123, 210116,210109,210102,201226,201219]


for week_num in week_nums:

urllib.request.urlretrieve(f"http://web.mta.info/developers/data/nyct/turn
stile/turnstile_{week_num}.txt",
                               f"data/turnstile_{week_num}.txt")

from sqlalchemy import create_engine
import pandas as pd

engine = create_engine("sqlite:///mta.db")
all_data = pd.read_sql('SELECT * FROM mta_data;', engine)
all_data = all_data.drop(index=[0,1])
print(all_data.head())
```