Report

Abstract

There is an electric company that wants to work with New York MTA subway stations to do lighting maintenance and you want to know the peak times to avoid and non-peak times to send the maintenance team to the station

Data Sourcing

Download MTA turnstile data files for the following three months (July , August and September in 2021)

Data Cleaning

-display columns names

-The strip() method removes any leading (spaces at the beginning) and trailing (spaces at the end)

-display data information

-The value_count() for some columns

-describe dataframe

-Display null values

-create datetime and day column

-drop duplicates value

-delete the unnecessary columns

-"group by column "STATION", "UNIT","C/A" and "SCP

-create column TIME_INTERVAL

-calculate DAILY_ENTRIES and DAILY_EXITS

-merge data frame entries and data frame exits

-calculate the traffic(DAILY_ENTRIES + DAILY_EXITS)

-calculate the traffic at each station descending and ascending poltted

-plotting and calculating peak times for some stations

TOOLS

Data analysis with the following

Panda , Numpy , SQLalchemy , Matplotlib and Seaborn