‘Analysis of NYC’s people concentration using MTA subway stations data’

Project Proposal

i. Background

Company Info:

The Government of New York City, headquartered at New York City Hall in Lower Manhattan. The city government is responsible for public education, correctional institutions, public safety, recreational facilities, sanitation, water supply, and welfare services. NYC has a plan to make the city healthy and minimize the negative impact of any future health pandemic.

Problem:

New York City Government has a plan to balance the existence of the services of city around the whole city as one of its defense mechanisms towards any health pandemic. Services of the city include banks, universities, schools, companies, public institutions, factories, and any place that the New Yorkers may need to go to regularly. Currently, the government needs to know specific peaks of crowds at specific times of the day and week.

The value:

To help the government understand what areas of the city are more crowded at what time of the day and the week to prevent having more concentrations of people in specific areas in specific times.

ii. Data Description

Two datasets will be used in this project. One is the MTA dataset collected from The New York Subway MT, and the other is the New York subway stations data.

A) MTA Dataset

- The New York subway MTA turnstile data is a series of data files containing cumulative number of entries and exits by station, turnstile, date and time. Data files are produced weekly, data records are collected typically every 4 hours with some exceptions.

- Features

Field Name	Description
C/A	Control Area (A002)
UNIT	Remote Unit for a station (R051)
SCP	Subunit Channel Position represents an 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
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)
ENTRIES	The comulative entry register value for a device
EXITS	The cumulative exit register value for a devic

- Time Interval (9-2019, 10-2019, 11-2019)
- Number of rows: 2676241

B) The New York subway stations data.

- The dataset has the New York City subway stations names and location.
- Features

Field Name	Description
URL	These text maps give you information about station stops, service, and transfer information for each subway line in New York City.
OBJECTID	Index number for the station
NAME	Name of the station
the_geom	Geometric location of the station
LINE	Line of trains in the stations
NOTES	Notes about operation times

- Number of rows: 473

iii. Tools

A) Technologies: SQL, SQLite, Python, Jupyter notebook, Teblue
B) Libraries: Numpy, Pandas, Matplotlib