Recommendation on commute time and station in NYC

Abstract

The goal of the project is to understand the subway traffic trend in 2021 as people come back to work. I first used MTA and NYC positive covid case data to determine if the pandemic is a legitimate concern, then compared 2020 and 2021 daily ridership trends to determine the focused year. After plotting hourly ridership and compared ridership per station, I finalized the recommendation regarding commute hour and station.

Design

The hypothetical client would be a small company, the company wants to advise their employees on how to minimize exposure to the virus while taking public transportation. Following the CDC guidance, one should avoid close contact (within about 6 feet) with other people, which translates to avoiding crowds in a closed small space. What we need to find out is the lower trafficked times and stations.

Data

I have two data sources:
1. Daily positive covid test cases: each row represents at a specific date, at a NY county how many new positive test cases are confirmed.
2. MTA turnstile data: this data contains the number of people entered or exited through a specific NYC subway turnstile. This file is at a four hours data grain.
For both data sources, I limited three months of data (Feb~April) in 2020 and 2021. As I deep-dive into the data, the data scope narrows down to 13 busiest stations.

Algorithms

1. Identify turnstile reset by comparing cumulative entries to the previously recorded entries, fix the negative entries issue.
2. Identify outliers where the cumulative entries are more than two standard deviations over the median. Exclude outliers from the analysis.
3. Calculate the volume of null values, after evaluating the data size (0.52%), decide the drop the nulls.
4. Evaluate if a data is true zero or a data entry error. One turnstile at 59 street station missed two days of data, this is likely a true zero, especially since the two days are weekends. While the RIT-ROOSEVELT station missed two months of data, this station is likely closed or the turnstile is broken. In another word, not true zero.

Tools

- Data Cleaning: SQL, Python
- Analysis: Pandas
- Visualization: Tableau, Seaborn

Communication

Please find the slide and code in my Github account.