Final Project --Data Cleaning and Integration

Group name

Group 8 -- Scorpio

Project participants

Xilin Jiang (xj710), Jia Shi (js11182), Haochen Zhou (hz2204)

Project description

(Note: Due to some drawbacks of our previous datasets. Thus, we slightly changed the theme of our project. What we initially studied was the crime rate affected by COVID-19, and now we change to study that, in different boroughs in New York City, how the number of people taking the subway has changed due to the COVID-19.)

a) Problems

Today, due to the worldwide spread of COVID-19, the lives of people in New York City have changed a lot. In the project, we aim to analyze the relationship between COVID-19 and the number of people who take the subway every day in New York City. Will the flow of people in the subway station be affected by COVID-19? Will social distance policy affect how people travel?

b) Overview

1. The main approach of our project is to accurately analyze the turnstile data to find some interesting points. We will find the changes in people's travel data and COVID-19 cases.
2. As we know, there are five boroughs (The Bronx, Brooklyn, Manhattan, Queens, and Staten Island) in New York City. For each station in the borough, we will look up the swipe records on each device of the station to get the number of people who take the subway every day in different boroughs.

3. Based on the data about the number of people taking the subway, we will find out if there is any relationship between people's travel and COVID-19.

Datasets

a) New York City subway stations information

The data provided by MTA shows the location of each station. Based on the data, we can match each station in a borough.

http://web.mta.info/developers/data/nyct/subway/Stations.csv

b) New York City turnstile information

The data provided by MTA shows the entry/exit register values for each turnstile at the control area in each station. Based on the data, we can find the number of people at each station.

http://web.mta.info/developers/turnstile.html

c) COVID-19

The data provided by The New York Times reports the cumulative counts of coronavirus cases in the United States, at the state and county level.

https://github.com/nytimes/covid-19-data

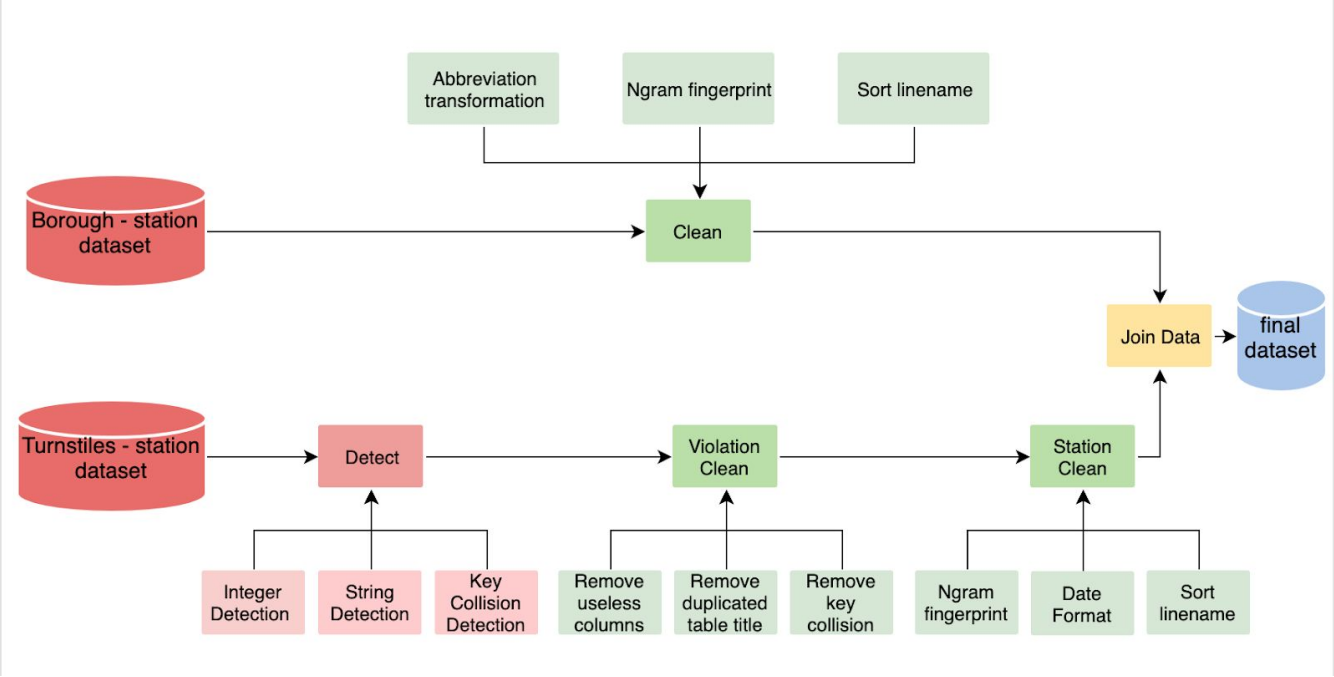
Data Cleaning and Integration

a) Steps

1. We firstly merged datasets (New York City turnstile Information) from different date ranges. Then, we can get one file including all the data about the turnstile information from Dec. 2019 to April 2020 in New York City.
2. We then run another dataset (New York City subway station information) to find what borough does each subway station belong to. Meanwhile, in order to perform an easy operation to join tables in SQL and to avoid the condition that there may exist the same station name in different boroughs, we used the n-gram fingerprint method to transfer the station name into a general format.

3. Besides, we also transferred the borough abbreviation to its full name, such as change “Bk” to “Brooklyn” and etc.
4. Then, we began to perform data clean on the Turnstile dataset. First, we still used the n-gram fingerprint method to turn the station name into a general format and then sort the line name of each station.
5. Look for useless or wrong data about the turnstile. Firstly, check the data type, such as the integer type. According to some integer, we then find the max value and min value to know whether there exists invalid data or the data that exceeds the normal boundary.
6. Find out the date attribute and change the date from a string type to a date type. For example, because the original type of the date is “MM / DD / YYYY” which is a string, "2019" will be ranked after "2020" after sorting. In this case, we have to change it to a date type ("YYYY- MM-DD") so that it can be sorted accurately as we need.
7. We meet the problem of the key collision. It is like, as to two different data records, there may exist their keys are the same but their values are different. More details will be included in the challenge 3.

The Data Clean workflow is as follows:



b) Challenges

1. The original turnstile dataset is complex, and it doesn't have a fixed commit time (the time to submit the data to the terminal) for each turnstile every day. It is difficult for us to find the earliest commit time for each device each day. And It is also difficult to use the 'select'' SQL function to reach the goal. Thus, we finally choose to use "group by" to find the earliest commit time for each turnstile per day.

2. When we deal with the data about borough-station, we didn't consider the corner case that there may exist two stations with the same name in different boroughs. Thus, we cannot just use the "distinct" SQL function to match the borough with the station. Instead, we combine the station name and the "line name", which includes all lines passing through that station, to get the result. In other words, we choose to use two attributes to determine the only station and then combine all lines as the third attributes.

3. We find there exist some collisions in turnstile data. One turnstile may commit its data at the same time but in different statuses (one of the data is committed in "RECOVER AUD" status and the other is committed in "REGULAR" status). We analyzed the entries and exits data in Regular and Recover status. Then we find out that most Recover values are quite similar to the Regular values. However, a few of them have much smaller Recover values than the other. Thus, we decided to remove the Recover data if there exists a collision, which needs to write a SQL function to do a minus operation. At first, we tried to use the "DELETE" function to reach the goal, but that doesn't work. After searching for more information and trying different kinds of methods, we recognized that we can use the "EXCEPT" or "MINUS" function to work it out!

c) Github repo

https://github.com/ShiJia000/COVID-19