BIG DATA PROJECT REPORT

Life in New York City

Group 8 -- Scorpio

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

New York University --Tandon

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

(We don't want open source code for the time being to prevent plagiarism, if you need permission, please send us your GitHub username. Email: js11182@nyu.edu, hz2204@nyu.edu, xj710@nyu.edu.)

05.07.2020

TABLE OF CONTENTS

INTRODUCTION

BACKGROUND

At the end of 2019, the novel coronavirus pandemic (COVID-19) is the defining global health crisis of our time and the greatest challenge we are facing. The New York government asks people to stay home and to stay off public transit, especially subways and buses.[1] Therefore, today, due to the outbreak of COVID-19, the daily lives of people in New York City have changed a lot, especially for people's travel.

PROBLEM

The purpose of the project is to study and analyze what factors affect the COVID-19 confirmed cases in different ZIP code areas in New York City.

Research questions:

1. What is the correlation between the number of people on the subway under each ZIP code in New York City and the spread of COVID-19?
2. Will this relationship or any other factors (e.g. income) cause COVID-19 positive cases in different regions to be different?

APPROACH

Working from home and avoiding public transit play a very important role in the spread of the epidemic. The first goal of the project is to analyze the factors that affect the spread of COVID-19 from the perspective of subway passenger flow.

1. Analyze the daily changes in subway passenger flow in New York City.
2. Compare the subway turnstile data between March 2019 - April 2019 and March 2020 - April 2020. Observe the relationship between the COVID-19 outbreak and passenger flow.
3. According to the policies taken by the New York State government (e.g. work from home), analyze the relationship between specific time points and changes in passenger flow.

Because "The COVID-19 outbreak affects all segments of the population and is

[1] "NYC Asks Commuters to Stay Off Public Transit 'If You Can' to Combat Virus Spread." WNBC New York, https://www.nbcnewyork.com/news/local/nyc-issues-new-commuter-guidelines-to-combat-coronavirus-spread/2317584/.

particularly detrimental to members of those social groups in the most vulnerable situations … including people living in poverty situations…"[2]

1. Look for other factors, such as the average income of NYC, which may affect the number of New York City subway passengers and the spread of COVID-19.

[2] "Everyone Included: Social Impact of COVID-19." United Nations, https://www.un.org/development/desa/dspd/everyone-included-covid-19.html.

DATA

New York City MTA Turnstile Data

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

Description: The data provided by MTA shows the entry/exit register values for each turnstile at control area in each station.

Field	Datatype	Description
C/A	string	Control Area
UNIT	string	Remote Unit for a station
SCP	string	Subunit Channel Position represents an specific address for a device
STATION	string	Represents the station name the device is located at
LINENAME	string	Represents all train lines that can be boarded at this station
DIVISION	string	Represents the Line originally the station belonged to BMT, IRT, or IND
DATE	date	Represents the date
TIME	time	Represents the time (hh:mm:ss) for a scheduled audit event
DESc	string	Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours)
ENTRIES	string	The cumulative entry register value for a device
EXIST	string	The cumulative exit register value for a device

New York City MTA Station by ZIP Code Data

Description: Use Google API to get the ZIP codes of all MTA stations.

Field	Datatype	Description

STATION	string	Station name
ZIP CODE	string	Related station zip code

New York City IRS Income by ZIP Code Data

Link:https://data.world/jonloyens/irs-income-by-zip-code/workspace/file?filename=IRSIncomeByZipCode.csv

Description: The data shows the income of the people in New York City.

Field	Datatype	Description
STATE	string	State
ZIP CODE	string	ZIP code
AVERAGE_INCOME	float	Average income in specific zip code

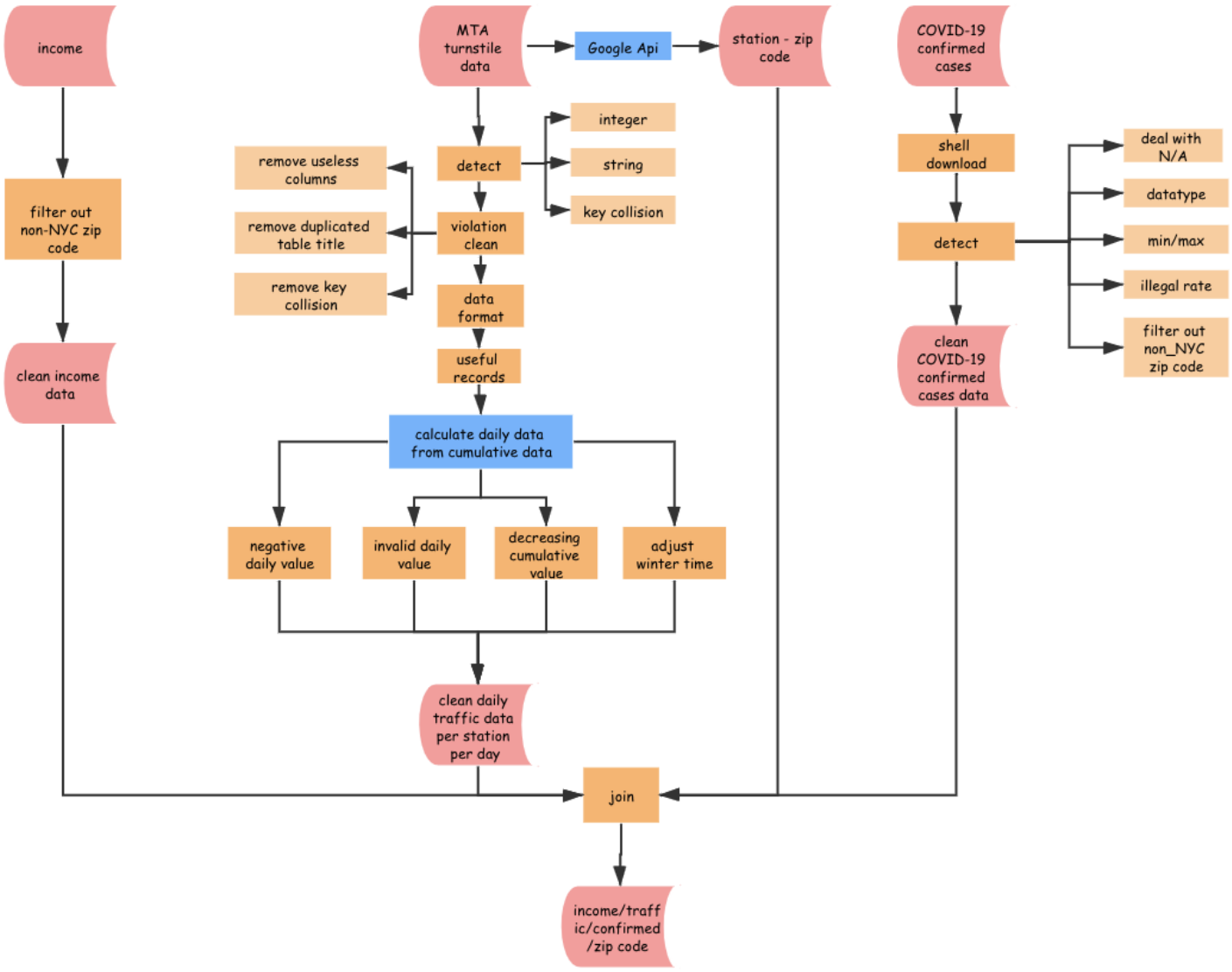
COVID-19 Cases Data

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

Description: The data provided by the New York Times reports the cumulative counts of coronavirus cases in the United States at the ZIP code level.

Field	Datatype	Description
DATE	datetype	Date of test cases
ZIP CODE	string	ZIP code
POSITIVE	int	Positive cases on one day
TOTAL	int	Total test cases on one day

PROCESS OVERVIEW



process flow diagram 1.1

DATA CLEANING

New York City IRS Income by ZIP Code Data

1. Find all ZIP codes in New York City.
2. Filter out the per capita income corresponding to each ZIP code in New York City.

New York City MTA Turnstile Data

There are a total of 70 raw turnstile datasets in txt format and each dataset contains data within 7 days. Each piece of data records the total number of people passing the turnstile during a period of time.

1. Merged all 70 raw turnstile data in txt format. Then, converted the merged data from txt format to csv format.
2. Extracted some useful columns in the data for later use.
3. Checked the data type
    a. **Check integer type**. For some columns of the data, find whether have any data types other than integers.
        i. Depending on the integer type, find the maximum and minimum values to determine whether there is invalid data or the data that exceeds normal boundaries.
    b. **Check date type**. Locate the date attribute and convert the data type from string to date.
        i. Original data type for the date attribute is a string, “MM/DD/YYYY”. Thus, when sorting, “2019” will be ranked after “2020”. In this case, the string type should be changed to the date type (“YYYY-MM-DD”) so that it can be sorted accurately.
4. Dealt with key collision
    a. For two different data records, there is a case where the keys of the two records are the same but their values are different. 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).
    b. Then, Analyzed the entries and exits data in Regular and Recover status, and found out that most Recover values are quite Similar to the Regular values.
    c. However, a few of them have much smaller Recover values than the other. Thus, the Recover data is removed if there exists a collision, which needs to write a SQL function to do a minus operation.

New York City MTA Station by ZIP Code Data

There is no matching data for ZIP codes corresponding to each station in New York City online. Thus, the related data was obtained by usingGoogle API.

1. Obtained the dataset of all ZIP codes in NYC.
2. The information of stations are fixed for each week. Thus, we chose to use the station data for one week to reduce the time for data operation.
3. Called Google API to get the ZIP code corresponding to each station.
    a. Extracted all station names and added “New York, NY” as the address.
    b. Searched the transformed address and output a JSON file including all geographic information related to the specific address.
    c. From the JSON file, obtained the related ZIP code.

COVID-19 Cases Data

1. Converted the format of date values to a cleaner format. For example, converted “2020-04-01T12:35:56-04:00” to “2020-04-01”, which only contains the date. The data can be analyzed more conveniently later.
2. Conducted data detection.
    a. Checked top ten maximum and minimum values to find if there are any outliers or illegal values.
    b. Detected and deleted the values of “NA” and “99999” in the data.
    c. Checked illegal rate
        i. Positive rate data that is greater than 100%
            1. On April 26 2020, all values of COVID-19 positive cases and total cases were reversed in the data. That resulted in all positive rates on that day being greater than 100%.
            2. The values of positive cases and the values of total cases were exchanged to recalculate their positive rates.
        ii. Positive rate that is NULL.
            1. Calculated positive rates based on related COVID-19 positive cases and total cases (positive rate = positive cases/total cases).
    d. Calculated the number of distinct dates recorded under each ZIP code for each day.
    e. Calculated the number of different areas recorded under each date by ZIP

code.

f. The ZIP codes between the COVID-19 cases data and NYC MTA station data are compared to filter out different ZIP codes in the data based on the station data.

DATA ANALYSIS & FINDINGS

Calculate daily passenger flow for each turnstile

Calculated the daily passenger flow from the accumulated data for each turnstile. The data was found it has several issues during the calculation. For example, the data of some turnstiles suddenly increased or decreased at some time, which caused the daily passenger flow to be more than 40000 or less than 0. The reason to set the upper threshold to 40000 was that up to 50 people can pass the turnstile in 24 hours in the case of passing one person every 2.16 seconds. Besides, some machines decreased the cumulative number after passengers passed for a long period of time. Those data were cleaned again during the calculation.

Turnstile join zip code data

Calculated daily passenger flow of all turnstiles in each station under each ZIP code. Matched ZIP code with each station.

1. Do the join operation on the zipcode-station relation table later to reduce the size of the data and the running time of the memory.
2. After joining tables, clean the joined table by deleting ZIP codes which are not in NYC and extracting useful columns from the data: zip code, date, inbound traffic, outbound traffic.
3. Set a threshold (the upper bound of the total traffic per turnstile everyday is 40000) to check whether there is a record above the threshold in the data. Finally, all data of the test results are below the threshold, further ensuring the accuracy of the experimental data.

Top 10 decrease

Compared with the number of people taking the subway in March and April in 2019, the maximum and minimum value of the decrease rate of the number of people taking the subway in March and April in 2020 are very different.

The information of maximum and minimum rate is as follows:

	ZIP code	Decrease rate
Maximum	10040	89%

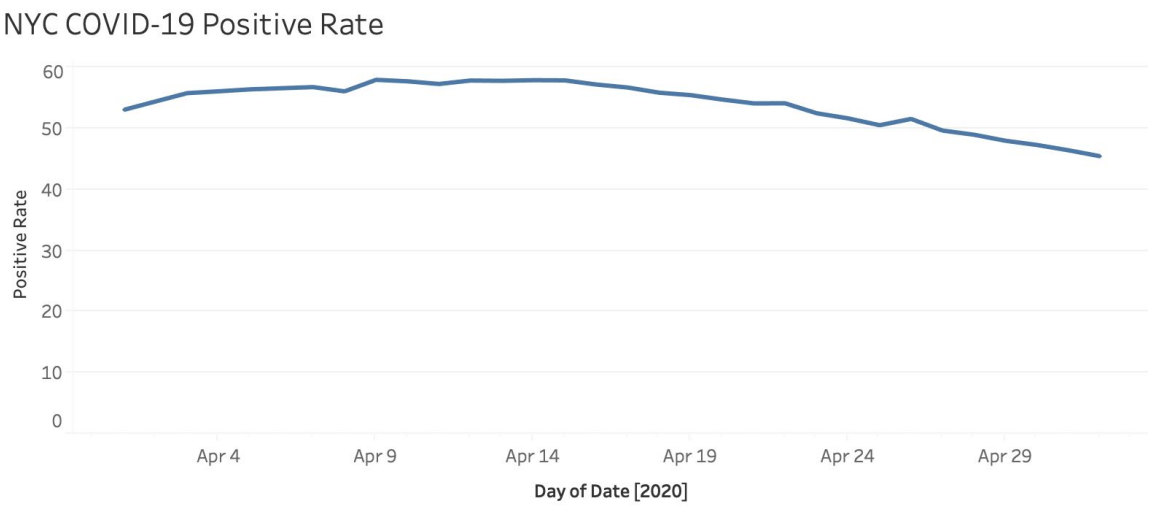
Minimum	11207	49%

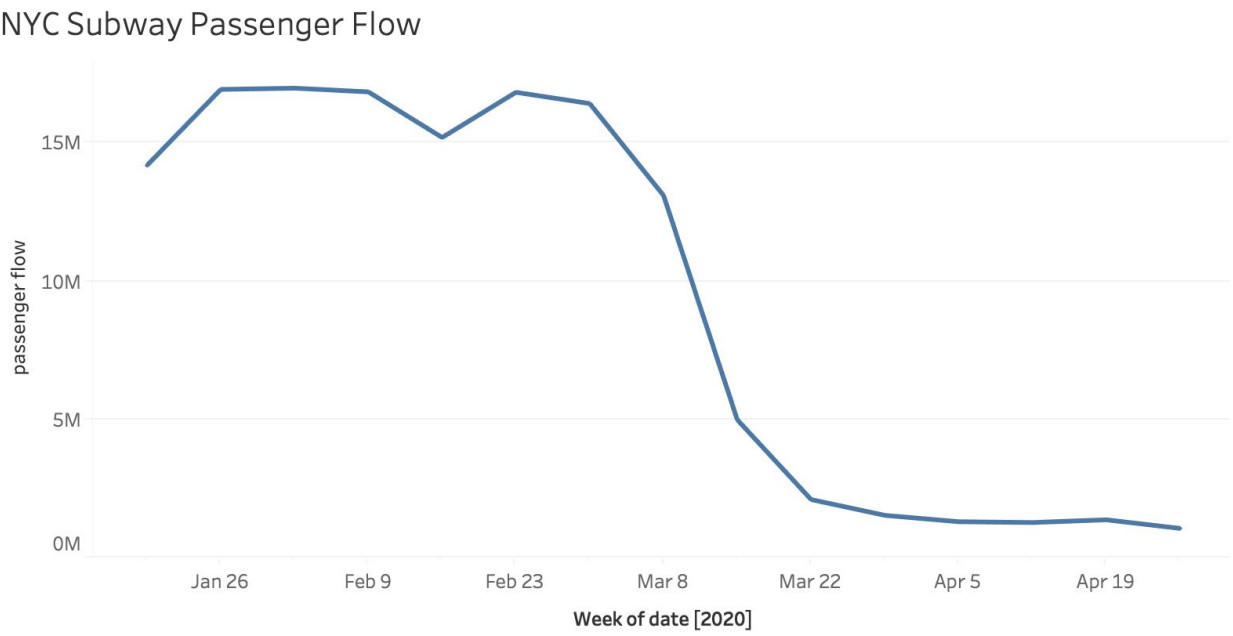
The decrease rate is high, which means that due to the COVID-19, the number of people taking the subway has changed greatly compared with the same period last year. There are less people going out.

However, the decrease rate is low means that the number of people taking the subway has not changed much compared with the same period last year. There are still lots of people who will choose to go out to take the subway.

Relationship between COVID-19 and Passenger flow

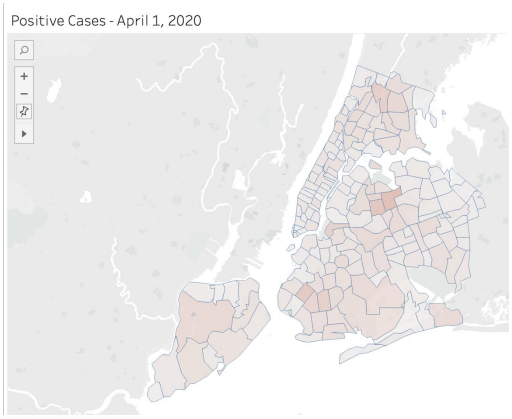
1. Calculated the daily positive rate from April to May 2020. (daily positive rate = sum of positive cases / sum of total cases in NYC)
2. According to the daily positive rate, draw a correlation chart of the positive rate and the date to analyze how the positive rate changes over time.
3. Draw a correlation chart of passenger flow and the date to see how passengers flow changes over time at all subway stations in NYC.
4. Compared the COVID-19 positive rate graph with NYC subway passenger flow graph.
5. **Findings:**
    a. Due to the spread of COVID-19, the number of people taking the subway in NYC has decreased.

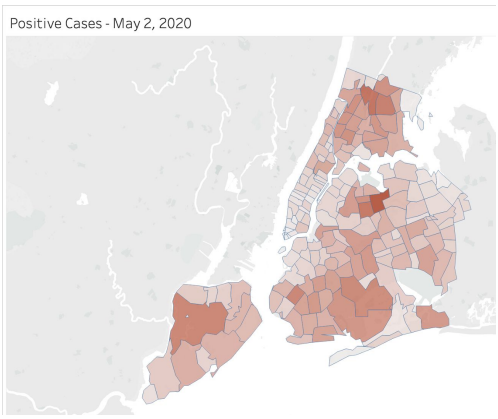




Find relationship between COVID-19 positive cases and other factors

1. Analyzed the number of positive cases in each ZIP code area of New York City. Performed the data by drawing a geographic map using Tableau. The darker the location on the map, the more COVID-19 positive cases in the area. Then, it was found that over time, some areas are much darker than others, which means that there were more cases in that area.

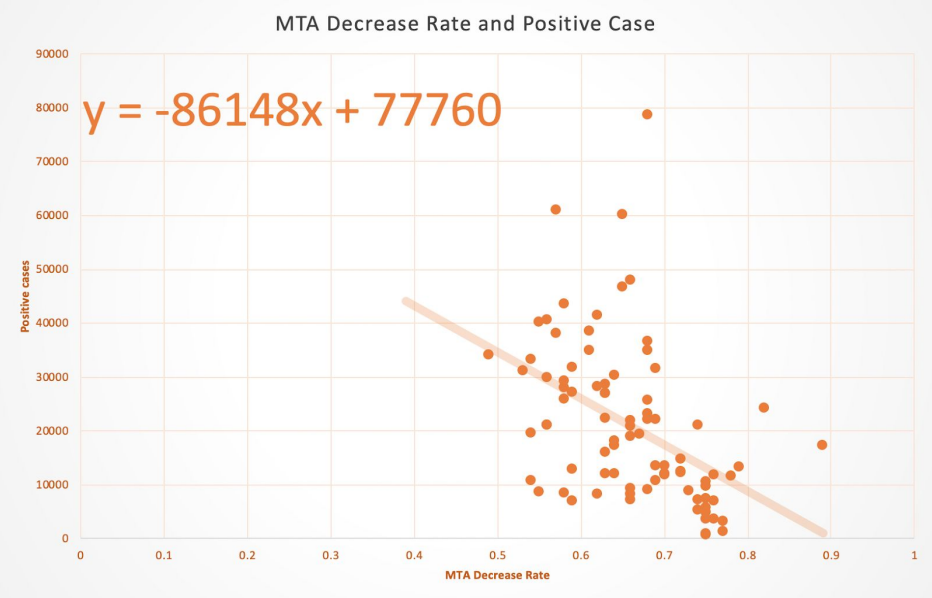




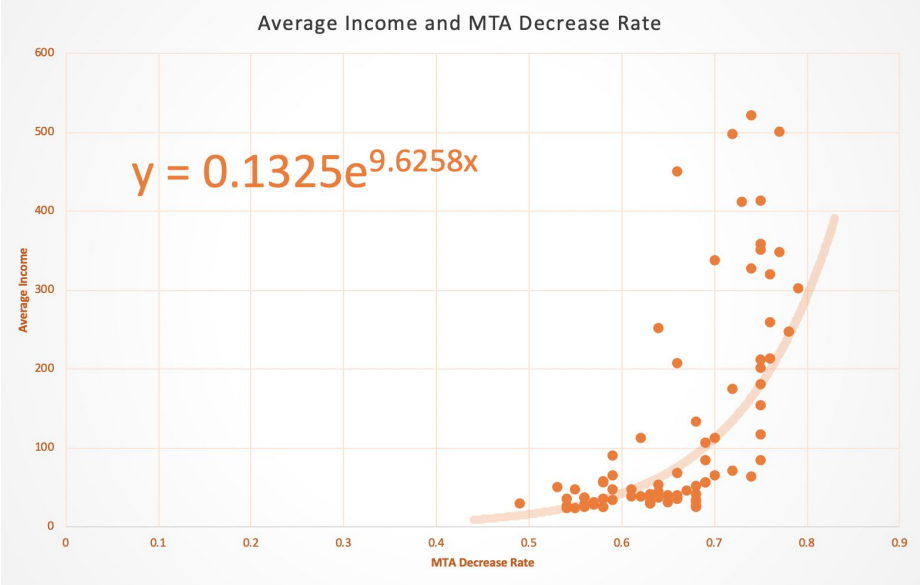
2. Based on the finding and the article “Everyone Included: Social Impact of COVID-19”[3], we speculated that per capita income may be related to the passenger flow of the New York City subway and the spread of COVID-19.

[3] “Everyone Included: Social Impact of COVID-19.” United Nations, https://www.un.org/development/desa/dspd/everyone-included-covid-19.html.

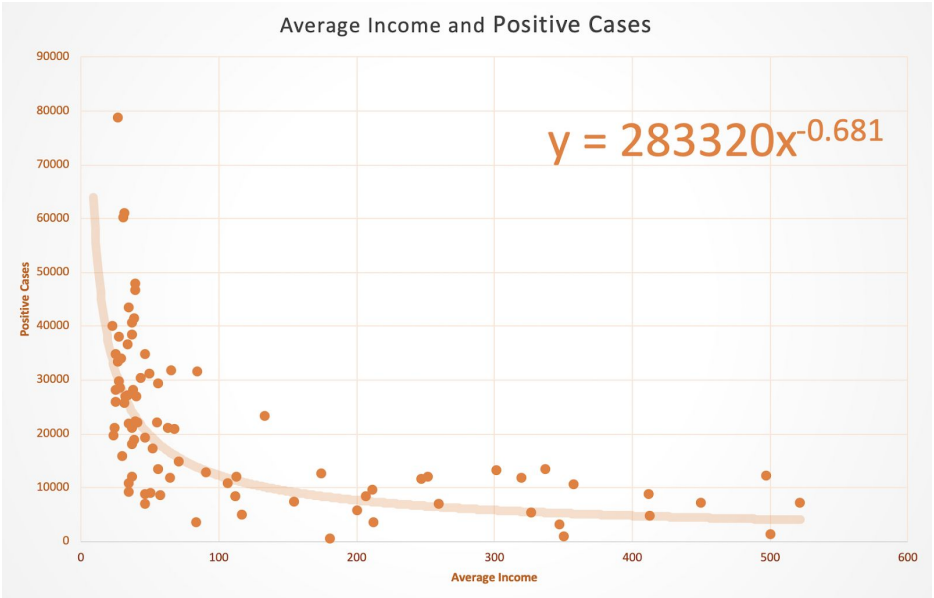
3. Join the data of average income, the number of COVID-19 positive cases and the number of passengers in NYC by the same ZIP code.
4. According to the joined data, draw scatter plots to analyze the relationship between each pair of the three data. Established regression models based on scatter plots to find detailed correlations.
5. **Findings:**
    i. Relationship between passenger decrease rate and COVID-19 positive cases based on the ZIP code in NYC: In the following graph, passenger decrease rate and the positive cases shows a negative correlation. The line between the points which is the trend line indicates that the district with higher passenger decrease rate has fewer COVID-19 positive cases.



ii. Relationship between average income and MTA decrease rate based on the ZIP code in NYC: In the following graph, average income and MTA decrease rate shows a positive correlation. The trend line indicates that the area with higher passenger decrease rate has higher average income for local people.



iii. Relationship between average income and COVID-19 positive cases based on the ZIP code in NYC: In the following graph, people's average income and the positive cases show the negative correlation. The line between the points which is the trend line indicates that the area with higher average income has fewer COVID-19 positive cases.



CHALLENGES

1. Due to the winter time, the time on March 9th was reduced by one hour, resulting in data cleaning errors. The graph caused by the data error has a great fluctuation, which is abnormal. Then, we started to search what happened that day. We searched the news, Twitter, etc. of the day. It is found that whenever there are large fluctuations in the data, Trump or the New York governor will make the relevant remarks the day before yesterday. We guessed at the beginning that they were connected, but we still have some doubts. After debugging, we found that there is little data on March 9th, which is caused by the problem of winter time.
2. Since there is no zip code dataset that can completely match the station in the turnstile data, we take the method of calling google api to find the corresponding zip code by obtaining all station names.
3. For the original data of turnstile, when converting the accumulated traffic data into daily traffic data, there are three challenges:
    a) When calculating the daily value through the formula, there exists some calculated daily data that have negative values, meaning there is a sudden change. For example, the daily number suddenly changes from 2000 to 200 within a time interval, which leads to the existence of negative numbers. In this case, we take the data from the previous day to ensure the stability of the overall trend of the data.
    b) The accumulated data shows a decreasing state. The abnormal decline rate and the increase rate of normal data were calculated, and it was found that the values were close, so the negative daily data was taken as the opposite number as the daily data for this day.
    c) In the calculated daily data, there are people who exceed the limit of one day, that is, data that does not conform to common sense, such as hundreds of millions of people walking a turnstile a day. In this case, we assume that the maximum threshold for a day is 40,000. So for these abnormal data, we take the data of the previous day, in order to ensure that the overall trend will not change significantly.
4. It's hard to accurately describe the correlation between the degree of income and the total number of confirmed cases. So we consulted the students who studied statistics. After asking them, we drew the scatter plot in Excel to generate the regression equation, which is an accurate way to express the correlation.

CONCLUSION

The project aims to find out what factors affect the different confirmed cases rate in different ZIP code areas. We analyzed and studied the relationship between two possible factors, which are average income of NYC and the NYC subway passenger flow, and the spread of COVID-19. Based on our study above, we found that the greater the number of people taking the subway in one area, the greater the total confirmed cases in the area. Besides, the higher personal income in one area, the less the total confirmed cases in the area, showing an inverse relationship.

low-income people need to go out to work, which may cause the decrease rate of the number of people taking the subway does not change too much. These people may have a higher risk of infection than others. Therefore, the government should pay more attention to protect the lower income people in the future to make our lives in NYC better.