How Does Viral Infection Affect Taxi Service Reliance?

MAST30034 Assignment 1

Xavier Travers
Student ID: 1178369
TODO: Github Repository

August 14, 2022

1 Introduction

Viral infection is on everyone's mind in the past few years due to the COVID-19 pandemic. With lockdowns and fears of infection, it is a natural assumption that many people-facing industries such as ride-hailing have suffered in demand. To what extent is such an assumption true?

This report aims to contribute to a body of works attempting to quantify the effect that widespread viruses may have on different industries.

Statistical analysis involved in revealing the extent to which case rates of a viral infection correlates with taxi trip rates and passenger counts.

1.1 Timeline

Throughout this report, several timelines are used.

- **Timeline 1:** Starting in March 2018 and ending in February 2021. This 36-month timeline is primarily used in visualizations.
- **Timeline 2:** Starting in March 2018 and ending in February 2019. This 12-month timeline provides a window of time prior to the COVID-19 pandemic. It is used when data analysis focuses only on the effects of the Influenza virus (since the effects of COVID-19 are likely confounding).
- **Timeline 3:** Starting in March 2020 and ending in February 2021. This 12-month timeline provides a window of time during the COVID-19 pandemic. This timeline is used to construct linear models and perform non-visual data analysis.

Data from 2022 is not included, as many of the datasets would be incomplete or unchecked. Data from before 2018 is not included so that Timeline 2 and Timeline 3 are the same duration (12 months), and to reduce code runtime when generating visualizations.

1.2 Datasets

- The New York City Taxi and Limousine Commission (TLC) provides a dataset of taxi service trips which captures information such as type of taxi, travel distance, general pickup/dropoff locations/times and driver-input passenger counts [1]. In this report, the focus is placed on New York's Yellow street hail taxis. The dataset provides coverage over the whole of Timeline 1. Also included from the same source is a mapping dataset for the pickup/dropoff location IDs included in the TLC dataset [1].

- Influenza case rates are recorded on a weekly basis by the New York Department of Health [2]. Case rates in this dataset are dated based on Morbidity and Mortality Weekly Report (MMWR) weeks, which are generated using rules defined by the CDC [3]. Each entry in this dataset contains an MMWR week, county (within the state of New York), type of Influenza (A, B or unspecified), and case count. The dataset provides coverage over the whole of Timeline 1.
- COVID-19 case rates have been recorded daily by the New York Department of Health and Mental Hygiene [4]. This dataset begins on the last day of february, when the first official cases of COVID-19 were recorded in New York City. Each entry in this data set contains a date and several of the daily COVID-19 rates by borough (e.g. count of hospitalizations on the day in the Bronx). Of specific interest is the daily case count per borough.
- Since data is aggregated by borough, the population of each borough needs to be accounted for. For this purpose, the United States Census Bureau's yearly county population totals data is used [5, 6]. This report specifically relies on the population estimates for the counties of New York State.
- To provide a homogeneous time metric for aggregation, a dataset is generated which defined the MMWR weeks of the data within the selected timeline. The MMWR weeks are generated according to the CDC's defined business rules [3].
- For geospatial visualizations, the City of New York's Department of City Planning provides a dataset containing borough outlines [7]. This contains the geometry of each borough as well as their names.

2 Method

2.1 Preprocessing

The datasets require several preprocessing steps to generate aggregate data for proper analysis. The flu dataset contains detail only on a weekly basis, while the other datasets contain daily data. Thus, the most granular time unit by which the data can be analysed is the MMWR week. While data per day allows for more data-points in analysis, Timelines 2 and 3 that ensure around 52 aggregated points of data are available for analysis per per borough.

2.1.1 Cleaning

There are several processes used to remove outliers and unwanted data. Noted are the steps taken to ensure that aggregation by borough and MMWR week is achievable with the TLC, COVID-19 and Influenza datasets. No imputation is necessary, since the scope of the timelines is very wide (approximately 208 million trips' worth), and after removing noise (in the methods described below) from the data, there are still approximately 100 million trips. Neither the COVID-19 nor the Influenza datasets require imputation, since these are maintained to the standards of important medical datasets.

Borough vs. County: Each of the 5 boroughs of New York City correspond to a county recognized by New York State, defined in Table 1 [8]. Some datasets contain counties, while others define statistics per borough.

Borough Name	Bronx	Brooklyn	Manhattan	Queens	Staten Island
County Name	Bronx	Kings	New York	Queens	Richmond

Table 1: Mapping Borough Names to County Names in New York City

Population Dataset: Only data for New York City counties within the years of Timeline 1 is kept.

TLC Dataset:

1. Derive trip duration (in hours) and trip speed (in MPH) columns. Filter out illegal (and likely incorrect) trip entries with a speed greater than 65 MPH, as per New York State law [9].
2. Discard all columns except the pickup time, passenger count, trip distance, pickup location ID, and dropoff location ID.
3. Discard rows with null values in the above columns or where there is negative distance.
4. Derive the MMWR week associated to each trip entry, as well as the year and month that the majority of the week participates in. Discard all rows where the MMWR week is not within Timeline 1.
5. For each entry, find the associated pickup borough and dropoff borough. Discard all rows where either of the location IDs are not within the 5 boroughs.

COVID-19 Dataset:

1. Extract the case count per day per borough. Discard rows with negative case counts.
2. Derive the MMWR week associated to each day entry, as well as the year and month that the majority of the week participates in. Discard all rows where the MMWR week is not within Timeline 1.

Influenza Dataset:

1. Discard rows with negative case counts.
2. Convert counties to their associated borough names for homogeneity of data.
3. Discard all rows where the MMWR week is not within Timeline 1.

2.1.2 Aggregation

TLC Dataset:

From this dataset, two aggregated datasets are generated. One aggregated by MMWR week and pickup borough, and the other aggregates by MMWR week and dropoff borough. This allows for analysis by destination, or starting point of taxi trips. The aggregated sets are then joined by MMWR year and borough to their corresponding population estimated. For each of the groupings described above, the average trip distance and passenger count is calculated, as well as the total number of trips, and the number of trips per capita (derived using the borough's population estimate).

COVID-19 Dataset:

Similarly to the above, the dataset is grouped by MMWR week information and borough. This grouping is joined by MMWR year and borough with the corresponding population estimates. For each grouping, the total number of cases and total per capita are calculated.

Flu Dataset:

Similarly to the above, the dataset is grouped by MMWR week information and borough. This grouping is joined by MMWR year and borough with the corresponding population estimates. For each grouping, the case total per capita is calculated (the case total per MMWR week per borough is already included).

2.2 Analysis and Modelling

2.2.1 Preliminary Analysis

It appears that there very clearly is a dip in passengers, trip rates and trip distances following the start of the COVID-19 pandemic.

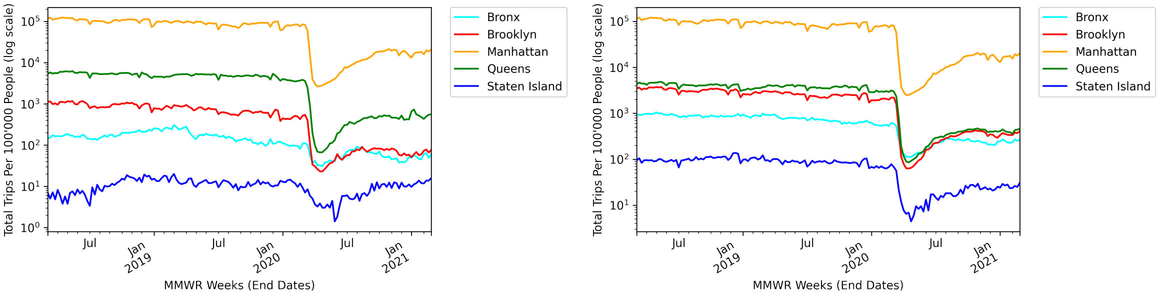

Figure 1: How trip rates per 100k people per pickup (left) and dropoff (right) borough vary over time.

2.3 Geospatial Visualisation

3 Recommendations

4 Conclusions

References

[1] New York City Taxi and Limousine Commission. *TLC Trip Record Data*. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-06.

[2] New York State Department of Health. *Influenza Laboratory-Confirmed Cases By County: Beginning 2009-10 Season*. https://health.data.ny.gov/Health/Influenza-Laboratory-Confirmed-Cases-By-County-Beg/jr8b-6gh6. Accessed: 2022-08-09.

[3] CDC. *MMWR Weeks*. https://ndc.services.cdc.gov/wp-content/uploads/MMWR_Week_overview.pdf. Accessed: 2022-08-09.

[4] Department of Health and Mental Hygiene (DOHMH). *COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths*. https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3. Accessed: 2022-08-09.

[5] United States Census Bureau. *County Population Totals: 2010-2019*. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html. Accessed: 2022-08-14.

[6] United States Census Bureau. *County Population Totals: 2020-2021*. https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html. Accessed: 2022-08-14.

[7] Department of City Planning (DCP). *GIS data: Boundaries of Boroughs (water areas excluded)*. https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm. Accessed: 2022-08-14.

[8] NYC311. *New York City Counties*. https://portal.311.nyc.gov/article/?kanumber=KA-02877. Accessed: 2022-08-09.

[9] New York Safety Council. *Speed Limits in New York*. https://www.newyorksafetycouncil.com/articles/speed-limits-in-new-york/. Accessed: 2022-08-14.