How Does Viral Infection Affect Taxi Service Reliance?
MAST30034 Assignment 1

Xavier Travers
Student ID: 1178369
TODO: Github Repository

August 15, 2022

1 Introduction

Viruses are currently on everyone's due to the COVID-19 pandemic. With the shift to working from home, lockdowns and fears of infection, it is natural to assume that people-facing industries are no longer relied upon to the same degree. To what extent is such an assumption true? This report investigates possible correlations between measures of reliance on taxi services, and virus case rates. While most research is satisfied with measuring correlations with demand in the form of usage frequency, the focus for this report is placed on measurements which reflect reliance more than demand.

Specific to taxi services, two key measurements of reliance are average travel radius/distance and average passenger count. Both of these measurements will correlate to the level of trust the average person places in a taxi service over its alternatives. For example, one may trust their own cars for interstate travel, but may be happy to pay for a street-hail on the way home from work (and similar logic may apply to trust in taxis with numbers of passengers).

Detailed throughout this report are steps taken to model average weekly trip distances and passenger counts against the case rates of two key viruses of recent importance: COVID-19 and Influenza.

1.1 Timeline

Throughout this report, two timelines are used.

- **Timeline 1:** Starting in July 2018 and ending in June 2021. This 36-month timeline is used exclusively for time-series plots.

- **Timeline 2:** Starting in July 2020 and ending in June 2021. This 12-month timeline provides a window of time during the COVID-19 pandemic. This timeline is used to construct linear models and generate other all other visualizations.

Data from 2022 is not included, as many of the datasets would be incomplete or unchecked. Data from before 2018 is not included 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 2. 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: This dataset is very simple, and therefore requires very little preprocessing.

1. Extract the case count per day per borough.

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: Much like the COVID-19 dataset, this one is very simple, and therefore requires very little preprocessing.

1. Convert counties to their associated borough names for homogeneity of data.

2. 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 aggregated by MMWR week and dropoff borough. This allows for a granular look at potential pattern differences between reliance measures grouped by dropoff or grouped by pickup locations. 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. This aggregation over Timeline 1 results in approximately 3 years × 12 months × 4 weeks = 144 rows per borough.

COVID-19 and Flu Datasets:

Similarly to the above, the datasets are grouped by MMWR week information and borough where necessary. This grouping is joined by MMWR year and borough with the corresponding borough's population estimates in a given year. For each grouping, the weekly case rate per 100'000 people in the borough is calculated.

2.2 Analysis and Modelling

This section of the report highlights the analysis performed on the aggregated data and describes the models generated for passenger counts and trip distances.

2.2.1 Preliminary Analysis

Aggregation Averages:

Since aggregation is performed on a mass scale, there is the risk of losing a lot of information present in the granularity of per-trip or per-day datasets. However, aggregating on a mass scale is beneficial in that the averages calculated (for distance and passenger count) a less prone to variability. As described by the Central Limit Theorem, as sample size n increases for a sample mean $\bar{X}$,

$$\text{var}(\bar{X}) \propto \frac{1}{n}$$

Therefore, since there over 150 million trips contained within Timeline 1, the sample size per week per borough is in the hundreds of thousands trips (assuming an even distribution of trips).

Time-Series Analysis:

It is important to note a feature which appears in nearly any time-series that includes the months of March to May 2020. The following time-series plots will all contain a similar "COVID dip".

Shown in Figure 1 is the change in average trip distance over time. Interestingly, while most boroughs (pickup or dropoff) experience a slump in average trip distance following the "COVID dip", Staten Island appears to recover from the effects, and even increase in average trip distance during the COVID-19 pandemic.
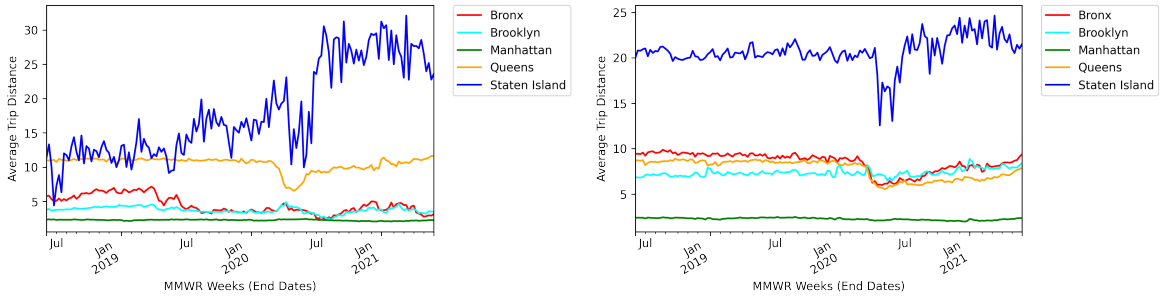

Figure 1: How trip rates per 100k people per pickup (left) and dropoff (right) borough vary over time.

This uptick in travel distance is likely caused to do the reduced usage of the Staten Island ferry and the reduced schedule during the pandemic [10]. If this is the case, then Figure 1 suggests that people from Staten Island rely on taxis to be safer and ontime over the ferry service.

2.3 Geospatial Visualisation

3 Conclusions

This report provides a brief, but thorough look at correlations to simple reliance statistics.

4   Recommendations

This has been a thorough, but brief look into correlations between viruses and reliance metrics. The following are several paths on which further research in this area could begin:

- There is the obvious path of modelling demand in the form of trip rates with respect to viral case rates.

- It may be beneficial to begin generating models of repeated trips between the same locations, which could suggest patterns of use.

- Another way in which the research could be deepened would be to select a subset of specific locations where trips going to/from are likely work-related. This would provide a window into measuring the shift to working from home caused by viruses.

- Finally, the results generated in this report are far from conclusive or exhaustive. As data is collected over the years, more accurate models may be constructed. While the focus may remain on modelling against COVID-19 and Influenza rates, other viral statistics such as the prevalence of HIV/AIDS should also be considered for potential correlations.

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.

[10] "NYC DOT Announces Reduced Staten Island Ferry Service". In: *DOT Press Releases* (Mar. 2020). URL: `https://www1.nyc.gov/html/dot/html/pr2020/pr20-014.shtml`.