How Does Viral Infection Affect Taxi Service Reliance?

MAST30034 Assignment 1

Xavier Travers
Student ID: 1178369
Github Repository

August 22, 2022

1 Introduction

Viruses are currently on everyone's mind due to the COVID-19 pandemic. With the shift to working from home, lockdowns and fears of infection, it is natural to assume that many people-facing industries are no longer relied on as much. To what extent is such an assumption true? This report investigates possible correlations between a measure 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 a measurement that is more reflective of reliance (or trust).

Specific to taxi services, a key measurement of reliance is average travel radius/distance. This research is performed under the assumption that this measurement will correlate with the level of trust the average person places in a taxi service over alternatives. For example, one may trust a taxi to travel further distances than the local tram. Detailed throughout this report are steps taken to model average weekly trip distances against the case rates of two prominent viruses: COVID-19 and Influenza.

1.1 Timeline

This report focuses on a single 24-month timeline starting in January 2020 and ending in December 2021. Such a large timeline allows for aggregation on a per week basis to yield a large aggregate dataset. It also includes a snippet of time before the COVID-19 pandemic for analysis. Data from 2022 is not included to avoid any risk of using data that is so new that it is incomplete or hasn't been double-checked. Data from before 2020 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 that captures information such as type of taxi, travel distance, general pick-up/drop-off locations/times and other trip data [1]. In this report, the focus is placed on New York's Yellow street hail taxis. The subset of this data denoted by the timeline contains $64'913'648$ trips and 19 features per trip before cleaning. Also included from the same source is a mapping dataset for 265 pick-up/drop-off locations and corresponding boroughs included in the TLC dataset [1].

COVID-19 case rates have been recorded daily by the New York Department of Health and Mental Hygiene [2]. This dataset begins on the last day of February 2020, 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

daily COVID-19 rates by borough (e.g. count of hospitalizations on the day in the Bronx). Of specific interest throughout this report is the daily case count per borough. This dataset contains 900 rows and 67 features per row before cleaning.

Influenza case rates are recorded on a weekly basis by the New York Department of Health [3]. Case rates in this dataset are aggregated by "Morbidity and Mortality Weekly Report" (MMWR) weeks, which are used by the United States Center for Diseas Control and Prevention (CDC) to number the weeks in each year[4]. Each entry in this dataset is associated to an MMWR week, county (within the state of New York), type of Influenza (A, B or unspecified), and case count. This dataset contains 77910 rows and 9 features per row before cleaning.

Since data is aggregated by borough, the population of each borough is 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. This data contains 16 rows and 68 columns per row (one column per county of New York State).

To provide a homogeneous time metric for aggregation, a dataset is generated that defines the MMWR weeks of the data within the selected timeline on a per borough basis. This allows for easier grouping by borough with null values during weeks with no case rates or trips. The MMWR weeks are generated according to the CDC's defined business rules [4]. This generated MMWR week dataset contains 3675 rows and 11 features per row.

All of these datasets provide coverage over the chosen timeline of analysis, allowing for the generation of meaningful models. 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. This dataset contains 5 rows and 5 features per row.

2 Method

2.1 Preprocessing

The datasets require the removal of several entries, and steps taken 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. This potential weakness in the analysis is discussed in this section as well.

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. Imputation is performed on the virus data where no cases were reported for a week or data. The case counts for these non-included rows are assumed to be 0, since the datasets are unlikely to contain missing data due to their importance to the CDC.

Borough vs. County: Each of the 5 boroughs of New York City correspond to a county recognized by New York State [8]. Some datasets contain counties, while others define statistics per borough. The boroughs with corresponding counties of different names are: Manhattan, also called New York County; Staten Island, also called Richmond County; and Brooklyn, also called Kings County [8].

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 pick-up time, trip distance, and pick-up location ID.
3. For each entry, find the associated pick-up borough. Discard all rows where pick-up is not within the 5 boroughs.
4. Discard rows with null values in the above columns or where there is negative distance.
5. 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.

Since only the trip distance is of concern in this report, filtering on other columns is not deemed necessary as long as the entries for trip distance and pick-up borough are consistent and valid.

COVID-19 and Influenza Datasets: These datasets are very simple, and therefore require very little preprocessing. First, the case counts per day (or per week, for the Influenza dataset) per borough are extracted. Then, the MMWR week associated to each entry, as well as the year and month that the majority of the week participates in are derived (where necessary). Finally, all rows where the MMWR week is not within Timeline 1 are discarded. For the Influenza dataset, counties are converted to their associated borough names for homogeneity of data.

2.1.2 Aggregation

TLC Dataset: This dataset is aggregates by MMWR week and pick-up borough. This allows for a granular look at potential pattern differences between the reliance measures grouped by pick-up locations. The aggregated set is then joined by MMWR year and borough to the corresponding population estimate. For the groupings described above, the number of trips, the average trip distance and passenger count is calculated. This aggregation over Timeline 1 results in approximately 52 rows per borough.

This report only performs grouping by pick-up location, allowing models to reflect a potential taxi customer's perspective on their immediate surroundings before taking a trip. However, exploration of differences between grouping by pick-up or drop-off location on the results of models is a recommended extension to this research.

COVID-19 and Influenza 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.

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), minimizing the variance of the sample means used.

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

Time-Series Analysis: Shown in Figure 1 is the change in average trip distance over time. Interestingly, while most boroughs (pick-up or drop-off) 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.

This uptick in travel distance is likely caused by 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 travelling to and from Staten Island rely on taxis to be safer and ontime over the ferry service. The trip distances between boroughs are generally not very homoegeneous. A likely reason for this heterogeneity is the difference in commutes to work, since many jobs are likely located in Manhattan. The greatest instability in trip distances appears in trips going from Staten Island, while travel from other boroughs appears more stable and similar week-on-week.

This may also come as a result of Staten Island's separation from the other boroughs contributing to many possible factors. Something as simple as culture could play a roles in differentiating the travel patterns between boroughs.

Figure 1 shows the variation in COVID-19 and Influenza cases per 100'000 people over time. One obvious limitation of both datasets is they look very skewed towards the lower values, meaning there will be many datapoints with small case rates, and very few points with larger case rates. This kind of skew suggests that some points of data will have higher leverage when fitting models than others, making the above cleaning and outlier removal processes crucial. Both datasets have spikes of case rates, but COVID-19 experiences a significant jump in January 2022.

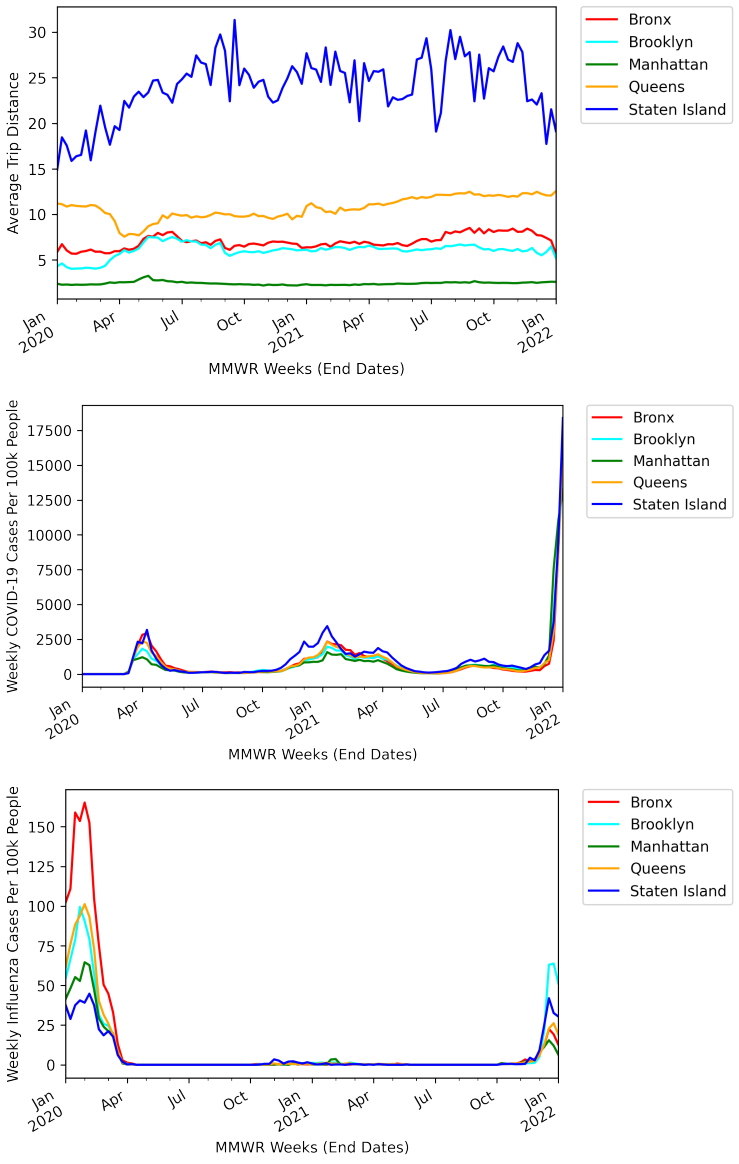

Figure 1: How Average Trip Distances (top), COVID-19 (middle) and Influenza (bottom) case counts per 100k people vary over time.

Geospatial Visualisation:

While time series plots display variation in average trip distance over time, they do not clearly convey the meaning of these differences. Figures 3 and 2 compares the average trip radius overall, the average trip radius for the week with maximum COVID-19 cases per capita, and the week with maximum Influenza cases per capita per pick-up borough.

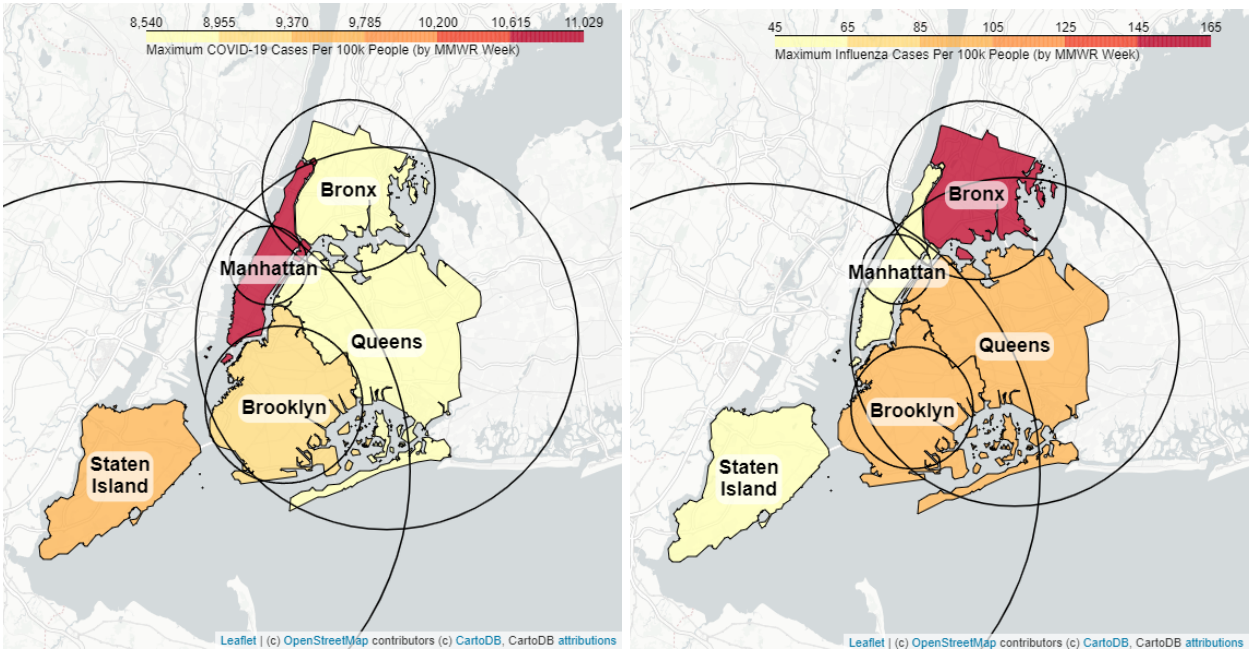

Figure 2: Map of weekly average trip distance following the maximum COVID-19 (left) and Influenze (right) cases rate over Timeline 2.

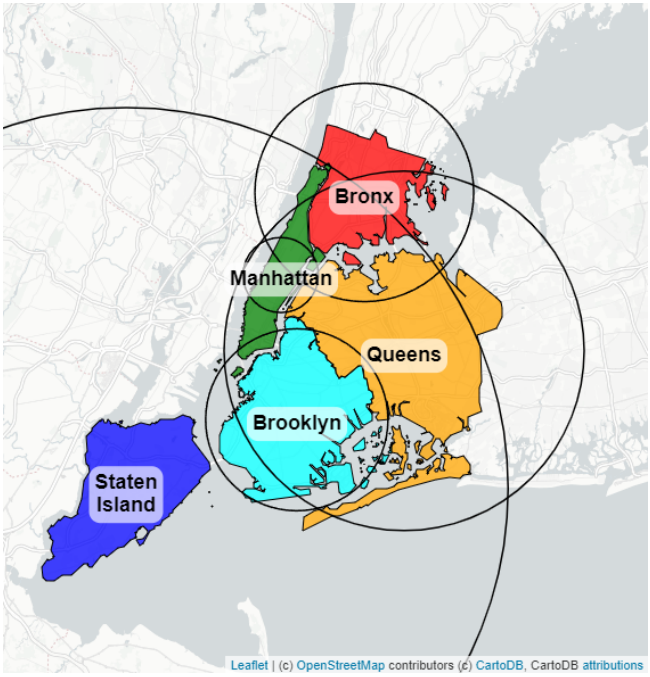

Figure 3: Map of average trip distance overall.

At a first glance, Staten Island experiences the most drastic change in trip distance following instaces of high case rates. This contradicts the findings from Figure 1, where the trip radius for Staten Island increased on average. On the other hand, this supports the initial expectation that COVID-19 and Influenza prominence will decrease trust in taxi services over long distances In general, it appears as though the travel radii per pick-up borough do not change too drastically following an especially high case rate. This may be a reflection of the speed at which case data proliferates among the populations of the boroughs, since not every individual will be checking last week's case rates.

The other boroughs do not vary as significantly in trip distance. This may be due to their proximity to each other. Another contributing factor could be the generally larger populations of these boroughs, meaning that a larger proportion of people are unlikely to change their taxi usage based on outside factors.

Distributions: It is important to note that the average trip distances are random variables with a generally unknown distribution. This means that multiple model candidates may be considered. In order to select two models to compare and contrast, the distribution of the data and its properties are investigated. Trip distances are a continuous metric with a one-sided bound at 0 Miles.

The weekly average trip distances have an overall average of 10.04 miles, with a variance of 58.75. Such a wide variance (where the lower bound of 0 Miles is only within 2 standard deviations of the mean) suggests that models for trip distances may be more prone to variability with irrelevant predictors. According to Figure 4, there appear to be multiple peaks in trip distance. With fewer bins, there are two discernible peaks, each roughly normally distributed. With more bins, the peaks are not as easy to discern, nor determine their distributions. Earlier evidence of positively skewed distribution occurs in Figure 1, where there are several bands of average trip distance when grouped by borough. An argument can be made that Figure 4 shows several normally distributed peaks of trip distances. However, another argument can be made that due to the trip distance being proportional to a measure of duration with a lower limit at 0, it can also be approximated with a gamma distribution.

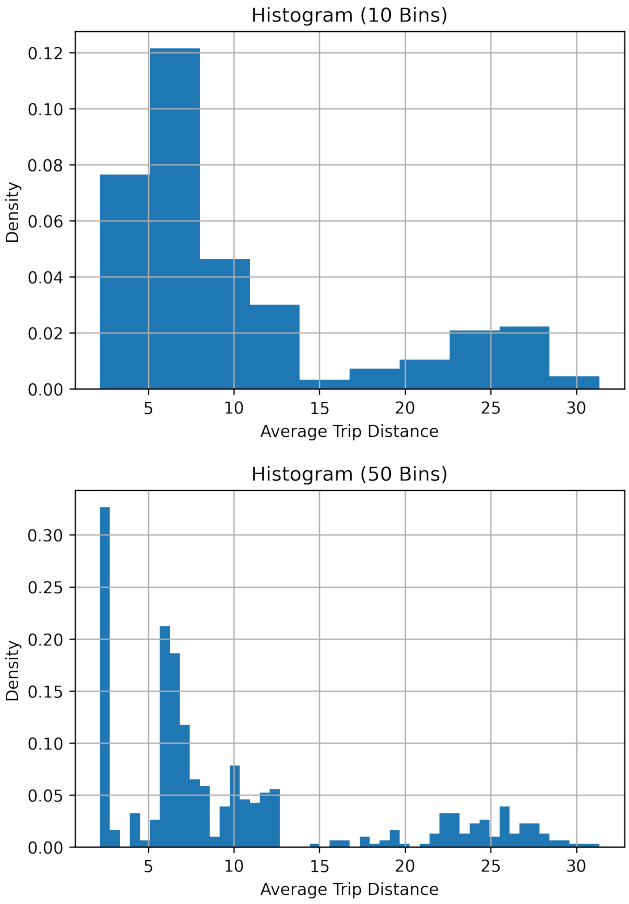

Figure 4: Probability density histograms of average trip distance with 10 bins (top) and 50 bins (bottom).

2.2.2 Modelling

Modelling continuous data is best done with linear models. Both the pick-up borough and time appear to affect the selected measures of reliance, meaning that they both need to be considered in the generated models. This means that for each week's average reliance measure, a linear model is generated with the predictors: borough, the preceding week's index in the timeline, the preceding week's COVID-19 case rate per 100 thousand, and the preceding week's Influenza case rate per 100 thousand, along with interaction between the borough (a non-ordinal categorical) and each of the viral case rates.

Gaussian Linear Model: The weekly average trip distance is first modelled using an ordinary least squares (OLS) or Gaussian linear regression, due to the potential presence of normal distributions per borough. The generated model displays a negative relationship between the viral case rates and trip distances, due to the parameters being negative ($\beta_{\text{COVID-19}_1} \approx -3 \times 10^{-4}$, and $\beta_{\text{Influenza}_1} \approx -4 \times 10^{-3}$). This confirms the theory that increased case rates disincentivise the use of taxis over longer distances. Other specific OLS parameter values are not mentioned, since this is a less than full rank linear model (where there are infinitely many parameter solutions due to the inclusion of a categorical variable). Instead, the model is analyzed using ANOVA testing and a comparison of fitted and observed data.

Table 1: ANOVA of chosen features in predicting average weekly trip distance

	SS	DF	$\mathcal{F}$	$\mathbb{P}(>\mathcal{F})$
Borough	2.901×10^4	4	4.438×10^3	Negligible
Preceding week index	1.136×10^2	1	6.952×10^1	7.099×10^{-16}
COVID-19 cases	7.514×10^0	1	4.598×10^0	3.249×10^{-2}
Borough Interaction	4.567×10^0	4	6.986×10^{-1}	5.931×10^{-1}
Influenza cases	2.067×10^1	1	1.265×10^1	4.115×10^{-4}
Borough Interaction	4.505×10^2	4	6.892×10^1	1.313×10^{-46}
Residuals	8.318×10^2	509		

According to Table 1, the most significant predictor in the linear model is the borough interaction term with Influenza case rates, while the least significant is the interaction between COVID-19 cases per 100 thousand and borough. At a 95% confidence level, only the COVID-19 interaction terms are considered irrelevant to the model. According to Figure 5, the residuals of this model will be heteroskedastic for the data. Unfortunately this is a strong indicator that the data does not follow a linear relationship. This suggests the need for further investigation into the type of relationship which is present. The Gaussian model yields an adjusted R^2 of 0.972, which is generally quite high. A measure to consider for comparison to the next is the log-likelihood of the linear regression, which is -865.9.

Gamma Regression: Continuous data with bounds that measures duration can often be accurately modelled using a generalized linear model in the gamma family. Since distance is approximately proportional (if not simply a multiple of) the time duration of each taxi trip, it may be the case that a gamma regression with the default inverse link is more applicable. The parameters from this model also supports the theory that trip distance decreases with increased case rates ($\beta_{\text{COVID-19}_2} \approx 6 \times 10^{-6}$, and $\beta_{\text{Influenza}_2} \approx 2 \times 10^{-4}$). Due to the use of the inverse link function, positive parameter values have an inverse effect on the response. Again, other specific parameters for the model are not mentioned. Instead, the relevance of the model can be determined through deviance and the Cox and Snell pseudo R^2. For this data, the model yields a deviance of 4.185 and a pseudo R^2 of 1.000. The deviance is very low, which suggests little variance in the model results. The pseudo R^2 is perfect, which indicates that this model is probably overfitting the data. This model yields a log-likelihood of -542.2.

Comparing Models:

According to Figure 5, both the Gaussian and gamma regressions perform well with the given dataset. While the models are similarly accurate with lower values of trip distance, at higher fitted values of trip distance, the ordinary least squares model tends towards a certain trip distance where there is actually a lot of variation. The gamma model appears to fit results with more homoskedasticity in the data. This suggests that average trip distances are better modelled using a gamma distribution.

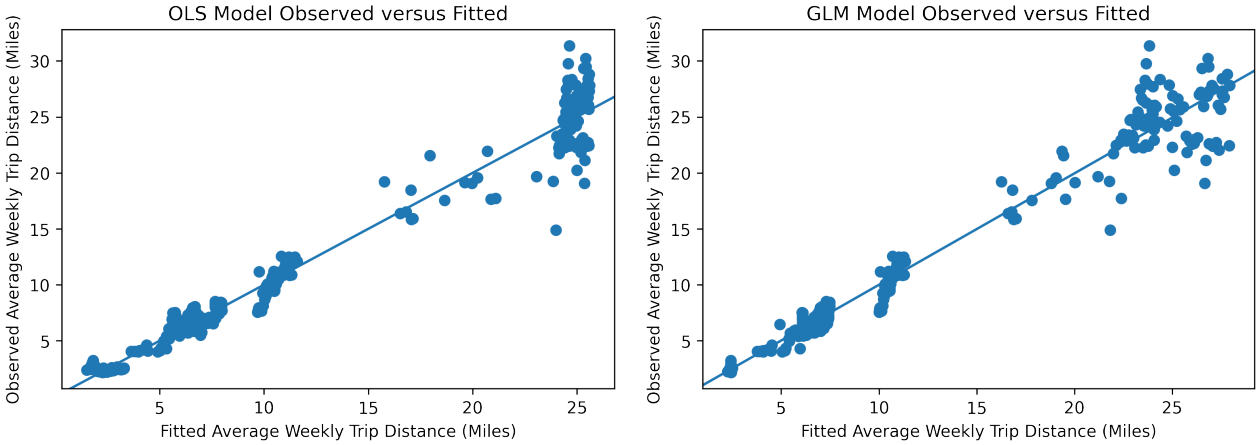

Figure 5: How observed and fitted values compare for the trip distance Gaussian Linear Regression (left) and gamma Regression (right).

The models can very easily be compared on their log-likelihoods, since they're both used on the same predictor. The gamma regression has a log-likelihood of -542.2, whereas the Gaussian (OLS) regression has a log-likelihood of -865.8. Since the gamma regression has a higher log-likelihood, the probability of its parameters being more correct is higher. This suggests that the trip distance data is indeed better represented with a gamma regression model.

3 Conclusions

In conclusion, there is merit to the thinking that increased case rates of viruses have an effect on the reliance on taxis at different distances. According to the linear models generated, both COVID-19 and Influenza have a small, but relevant relationship with the average taxi trip's distance. These relationships can be shown either with an ordinary least squares linear regression, or with a gamma regresion, with the latter yielding a higher R^2. However, there is the risk that the latter is overfitting the data, which may be the case due to the relatively small (< 1000 points) sample size. It is important to note that no causal relationships were proven or investigated in this report, and thus the generated models cannot be relied on as anything more than indicators of underlying correlations. In the same vein, while the models can be used to predict future reliance measures of yellow taxis, they should by no means be the sole models used. Instead they are best suited for use in conjunction with several others.

4 Further Research

There are several paths for further research based on this report. First, consider other datasets which merit inclusion in this analysis, such as other virus case rates, or other forms of transport. With more data to analyze, a clearer picture of the true relationships between infectious disease and preceived need for transport services. As mentioned in the paper, the TLC data is aggregated by pick-up location. The same analysis should also be performed on aggregation by drop-off location. There may be value in comparing and contrasting the resulting models with this different grouping. Finally, specifically 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.

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] 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.

[3] 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.

[4] CDC. *MMWR Weeks*. https://ndc.services.cdc.gov/wp-content/uploads/MMWR_Week_overview.pdf. 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.