Affect of Daily commute on COVID cases in Cook County, Illinois

Amrit Bhat

DATA 512 Final Assignment

Dec 12th, 2022

Introduction

The COVID-19 pandemic disrupted human life in many ways. Starting from the sudden rise in use of masks, to only be able to meet people virtually, to not being able to attend long planned festivities with families and friends. Out of these, the one aspect that stood out to me was how COVID-19 disrupted people's travel plans. Being from Bangalore, a city infamous for its prolonged traffic jams throughout the year, the shift to empty roads was upending and shocking to me personally, to say the least. All said and done though, stringent travel bans were possibly the most impactful policies to make a difference in decreasing the spread of COVID-19, from a scientific or practical perspective. With this motivation, I intend to explore the different ways travel bans had an impact on the spread of COVID-19, focusing specifically on the Cook County in Illinois. This is interesting because it would give us more insight into people's perception of risk and openness about travel during the pandemic. This problem statement is especially human centered as it reflects how human travel behavior can influence rise in COVID-19 cases, rise in hospitalized patients, deaths and more importantly, inform public policy during future pandemics or epidemics.

Background/Related Work

Affect of travel on COVID cases

Since the beginning of COVID, there has been quite a lot of research in terms of public mobility in the wake of COVID. As noted in the newspaper article, UIC Study Analyzes How COVID-19 Has Changed Travel Behavior, Lifestyles, researchers from University of Illinois, Chicago, foresaw a shift in travel behavior, with people opting to travel long distance by car rather than plane. Of those surveyed, 43% said they expect to travel by airplane less frequently in the future, with most respondents saying they don't feel safe or comfortable sharing space with others.

Public transit agencies, taxis and ride-hailing services face the same challenge, according to survey results, 93% of respondents viewed public transit, taxis and ride-hailing services as a

potential risk for exposure to the coronavirus. Personal cars, bikes and walking were viewed as the safest modes of transportation, according to the survey results.

In this analysis, I'm interested in getting a better understanding of how the people of Cook County, Illinois responded to travel ban, being the second most populous county after LA County in California.

To be able to answer this question, I will supplement the data from Part 1: Common Analysis with Travel by Distance dataset. The dataset provides us with percent increase (vis -a - vis a baseline) in mobility in public places at the county level. This data is exactly what we need to answer our question.

Research question:

How did inter-regional travel habits between communities impact the growth of COVID cases?

Hypothesis:

Every 10% increase in the number of people staying home decreased overall COVID cases by 6%.

Methodology

Data Gathering and Processing

Two datasets will be combined for this analysis. Daily case data will be gathered from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, representing daily cumulative case counts by county.

We also use the Trips by Distance dataset to understand daily travel behaviors of people in Cook County, Illinois. The dataset summarizes how many people are staying at home during the COVID-19 pandemic and how far people are traveling when they don't stay home.

These two datasets were joined together at the date level, I am anticipating performing additional calculations to create the final dataset. New cases are calculated as the difference in total case count form day to day. We find missing value distribution across the two joined datasets using *missingno*[1] python package.

Analysis

1. **Data Trends:** I plotted the trend of the data to understand rise and drop in COVID cases along with the mask mandate dataset in Part 1 and the Travel by Distance in Part 2.

2. **Correlation analysis**: I started with comparing how each of my covariates vary with the total no. of cases over the period from January 2020 up to the period I have covariate information for.

   Some of the steps I took are:
   - Plot a multi-line time series chart to visually understand how a change in human travel behavior caused a change in number of COVID cases.
   - Utilize automated checkpoint detection methods to identify major inflection points during the period of our analysis.

3. **Regression Analysis**: For my hypothesis to understand how number of COVID cases changed when more people stayed home, I regressed the independent variables describing local travel habits with the dependent variable, new number of COVID cases each day. This step is key, because the magnitude of the coefficients of our model will help us determine how the increase/decrease in the value of the independent variable results in an increase/decrease in the target variable.

Findings

Data Gathering and Processing

Part 1:

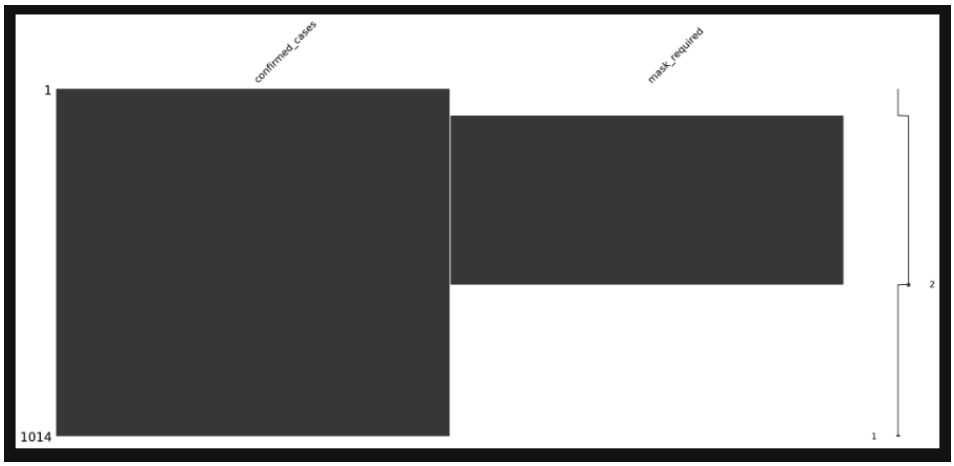

Figure 1. COVID cases dataset combine with mask mandate information

Upon merging the mask mandate information with the COVID-19 cases dataset, we could see that missing values were introduced in the dates where masking information wasn't available. These missing values were replaced with 'No' considering that we don't have any information if a mask mandate was placed during these dates.

Part 2:

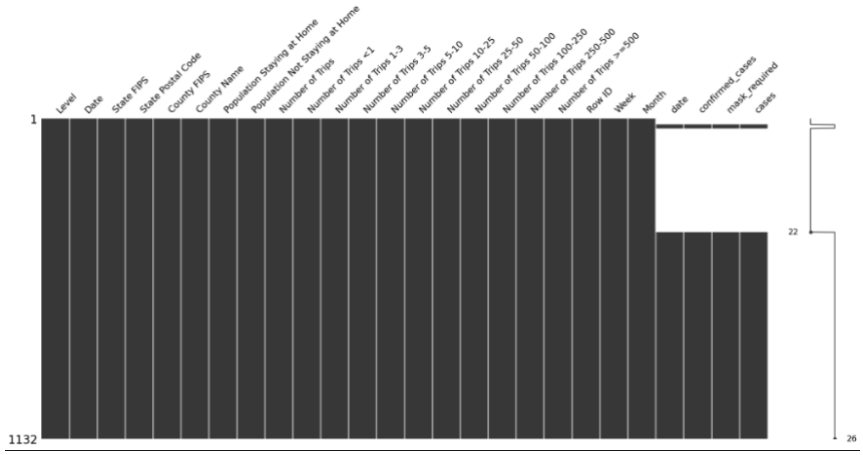

Figure 2.

I merged the cases data with the Travel by County dataset and Daily COVID cases dataset. For some reason, there were some present values in between collection of missing values in this dataset, as can be seen in the top right corner of Figure 2. This is unlikely because covid cases dataset was practically not collected in 2019. A manual inspection of the dataset revealed that for some reason data wasn't sorted by date. Some data (around 14 days from 2021) was placed between 2019 data. This issue was addressed by sorting the data by the “date“ column. We finally had data for 745 days (23 Jan, 2020 - 5 Feb, 2022).

Analysis

Data Trends

Part 1

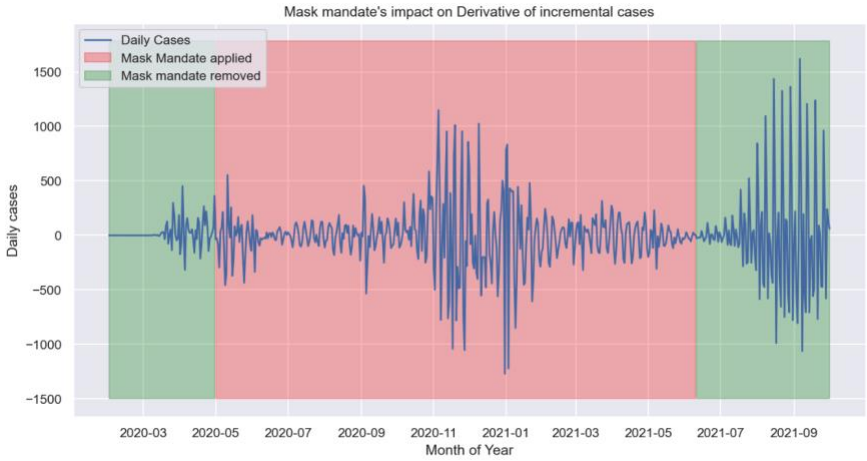

Figure 3.

The above figure shows a time series plot showing the changes in the derivative function of the rate of infection in Cook County of Illinois. We planned to show how the course of the disease was changed by masking policies from around February of 2020 (start of COVID-19) to October of 2022. To achieve this, we plotted a time series plot of the derivative of the rate of

infection and overlayed a plot indicating the days where masking policies were in effect (range of days shaded in RED) vs when they were not (range of days shaded in GREEN).

The X-axes is the “Month of Year” column which represent the specific date we are seeing between February 2020 and October 2021. The Y-axes “Daily cases” represents a derivative of the incremental daily cases. As can be seen, the mask mandate was put in place starting May of 2020. Since that was the beginning of the pandemic, government officials must have gone ahead with this step anticipating higher spread of COVID cases. In the graph, after a slow growth rate until September of 2020, there is significant growth starting mid-September 2020 and the cases are comparatively higher until February 2021, where they start to decline again. Due to the mask mandate in place, we can hypothesize that the cases were significantly lower compared to what they could have been otherwise. Based on the section of the graph after the end of the first mask mandate, we can even prove the validity of the hypothesis since there’s a higher rise in COVID cases following a month of ending the mask mandate in June 2021.

Part 2

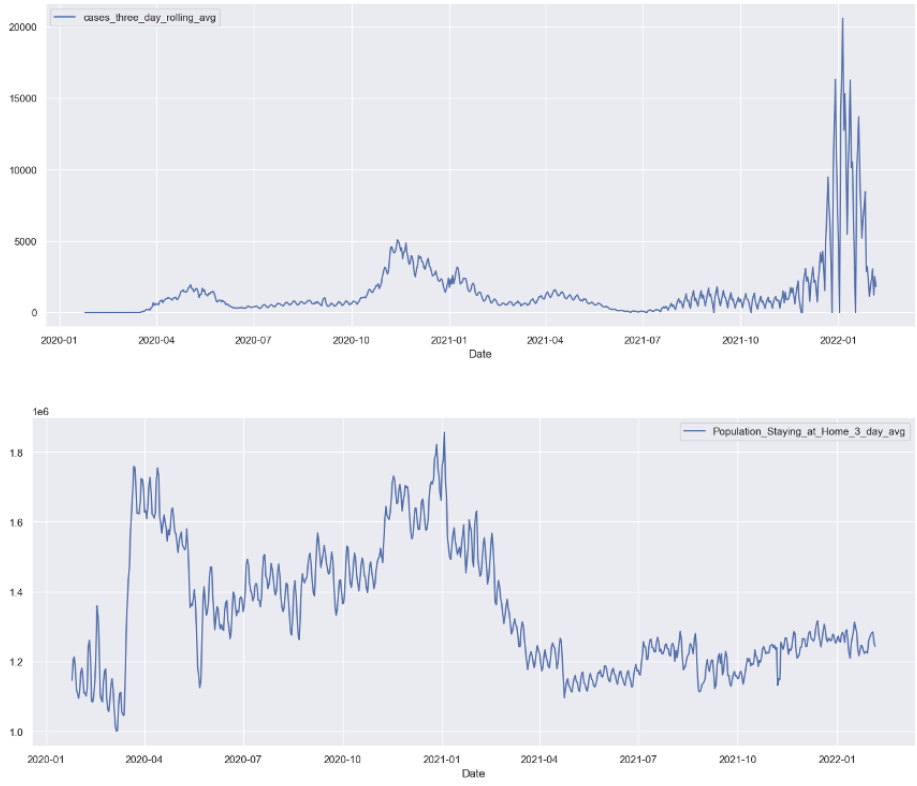

Figure 4.

Figure 4. shows the change in a 3-day rolling average of COVID cases (upper time series plot) comparing it with a 3-day rolling average of population staying at home (lower time series). We can infer that on a high level, cases came down when more people stayed at home. For example, there was a sudden uptick in the number of people staying home (1.1 million -> 1.75

million) in the period between April 2020 and June 2020. There is some growth in COVID cases during April 2020 and Mid-May 2020 but after which the cases decrease. This can majorly be attributed to the various international and domestic travel bans imposed, however there can always be other confounding factors involved. Since it was the start of the pandemic, there was an initial uptick in cases but more people staying home would have reduced the transmissibility (R0 value [2]).

Another example is of the period between December 2020 and March 2021. There was also an uptick in cases mid-December 2020 but that also subsequently decreased by Jan 2021 as the number of people decreased.

Figure 5.

From above graph, there seems to be a straight-line relationship with number of COVID cases showing slight increase with higher Population Staying at Home. Based on the histogram for each axis, we infer that the 'Population staying at home' variable has an approximate normal distribution, however the 'cases' variable has a skewed distribution.

Correlation Analysis

From the previous section we understand that since our independent and dependent variables are ordinal data with some features having skewed data distributions, we use the spearman correlation metric to measure correlated variables.

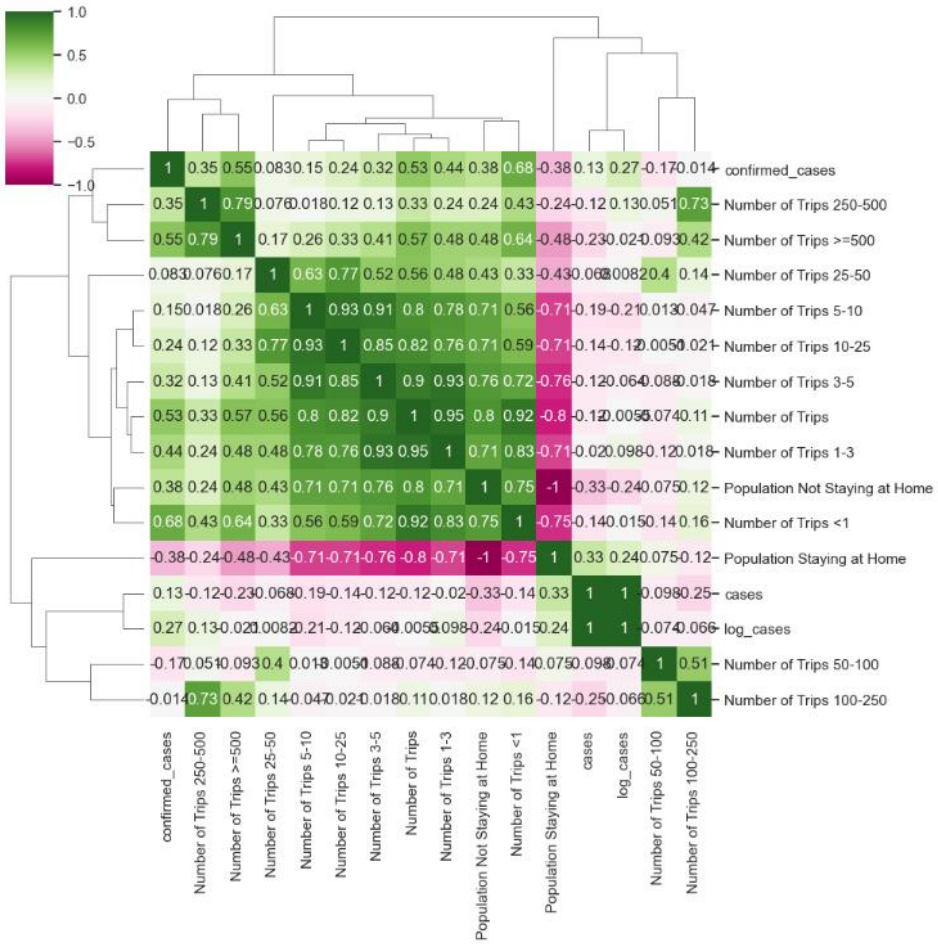

Figure 6. Spearman correlation matrix

Using a hierarchical correlation plot, we can see that the "Number of Trips" variables are generally highly correlated. However, for picking a single variable for our regression analysis, we focus on the correlations for the 'cases' dependent variable, where we see that the Population Staying at Home variable has the highest correlation of 0.33 (shown in Figure 7).

log_cases	1.00
Population Staying at Home	0.33
confirmed_cases	0.13
Number of Trips 1-3	-0.02
Number of Trips 25-50	-0.07
Number of Trips 50-100	-0.10
Number of Trips 250-500	-0.12
Number of Trips 3-5	-0.12
Number of Trips	-0.12
Number of Trips 10-25	-0.14
Number of Trips <1	-0.14
Number of Trips 5-10	-0.19
Number of Trips >=500	-0.23
Number of Trips 100-250	-0.25
Population Not Staying at Home	-0.33

Figure 7. Feature correlations with 'Cases'

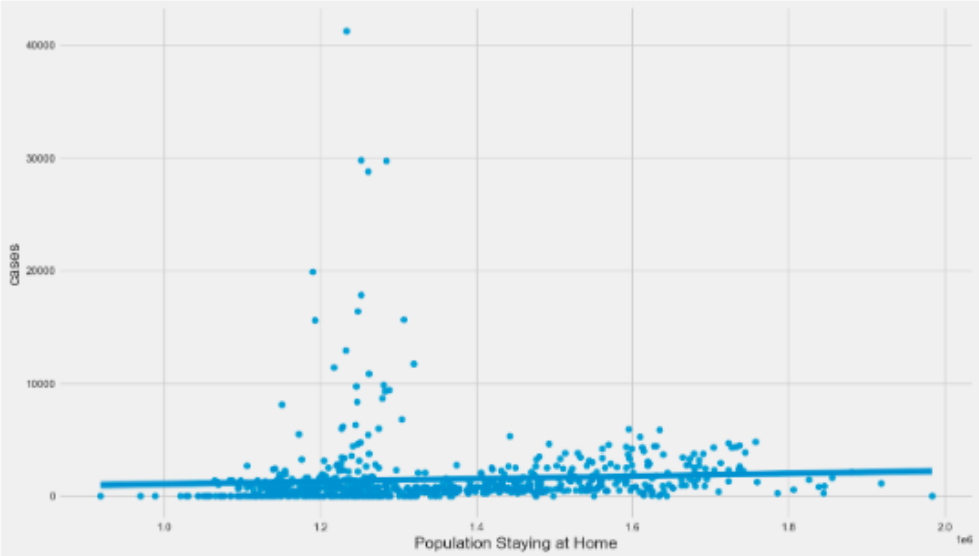

Figure 8. Linear Regression model

	0
Feature Name	Population Staying at Home
R Squared	0.004816
Intercept	-96.107163
Slope	0.001163

Figure 9. Model parameters

We regress the Population Staying at home variable as the independent variable along with the cases variable as the dependent variable. Here, we infer that COVID cases decreased when people went out more, albeit a very low R squared value (0.004816), signaling low confidence on the obtained results. The model suggests that every 10% increase in number of people results in a 0.116% increase in new daily COVID cases.

Discussions/Implications

This analysis was structured to facilitate insight into human travel behavior and attempt to arrive at verifiable and actionable results. With the goal of creating a product that users could evaluate against their own background knowledge, a regression model was created to infer correlational results that could be explained in simple language. The focus of this analysis was towards enabling policy makers to make data driven decisions from previous human travel behavior about regulating travel during future pandemic event.

The underpinning of the analysis was that as people travelled out more, there was an increase in the transmissibility of the virus. However, according to this linear analysis, we saw an anti-result, where the COVID cases increased when more people went out. It is noteworthy though that the R-Squared of this analysis is very low (0.004816) and so we should try modelling this relationship with non-linear models such as compartmentalized models [3] and time series neural networks.

Future research can build on this study by setting up experiments to understand how increase in the number of times people went out or number of trips they took (one trip is an individual visiting a place and staying there for more than 10 mins) increased the number of COVID cases.

Limitations

Linear Relationship

In this analysis, we use Linear Regression for examining the relationship between COVID cases as the dependent variable and Population staying at home as the independent variable. By its nature, linear regression only looks at linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them. Sometimes this is incorrect. One can tell if this is a problem by looking at graphical representations of the relationships. In this analysis, we clearly see that there might be a non-linear pattern in the dataset, which can be attributed to several different counter-factual considerations such as vaccine adaption, mask mandates, herd immunity, virus mutations and more.

Sensitivity to outliers

It is also sensitive to outliers. Outliers can be univariate (based on one variable) or multivariate. In our analysis we could clearly see some multivariate outliers, therefore, we will have to be very careful while relying on the results of this analysis.

Data Independence

Linear Regression assumes that the data are independent. That means the number of COVID cases on one day has nothing to do with cases on another day. This is probably the biggest assumption that we can make in this analysis because COVID cases on one day would have a higher probability to be affected by the number of COVID cases on the previous day.

Spearman correlation (non-normal distribution)

I have used spearman correlation primarily because of two reasons. One is that the data is ordinal and that it follows a non-normal distribution. The use of this correlation measure will not be valid however, if these assumptions are not met. The invalidity of this assumption would also mean that our final inference from the model will be incorrect.

Estimated values

One ethical consideration we should account for is that these data are experimental, which means they are created using new data sources or methodologies that benefit data users in the absence of other relevant products. Since we can't be sure about the accuracy of the numbers in the data, we should be careful about the fidelity of the interpretations from any analytical studies on this data, in the pursuit to inform travel policy considerations during future.

Conclusion

This study informs the readers on how inter-regional travel habits of people impact the growth of COVID cases using human centered data science principles. It primarily relies on assumptions that circle around human behavior and therefore inform the reader through the lens of human centered data science. Data was gathered from the Bureau of Transportation Statistics and John Hopkins University to help follow this plan. The project focused on Cook County in Illinois state.

The hypothesis was that every 10% increase in the number of people staying home decreased overall COVID cases by 6%. However, using a linear regression model, we infer that COVID cases decreased when people went out more, albeit a very low R squared value, signaling low confidence on the obtained results. The model suggests that every 10% increase in number of people results in a 0.116% increase in new daily COVID cases. In the context of COVID, this result isn't dependable because it is bereft of several different counter-factual considerations such as vaccine adaption, mask mandates, herd immunity, virus mutations and more.

References

1. Using Missingno to Diagnose Data Sparsity, https://www.kaggle.com/code/residentmario/using-missingno-to-diagnose-data-sparsity
2. What Is "R-naught"? Gauging Contagious Infections, https://www.healthline.com/health/r-naught-reproduction-number
3. Compartmental models in epidemiology, https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology
4. Spearman's Rank Correlation Coefficient Using Ordinal Data, https://towardsdatascience.com/discover-the-strength-of-monotonic-relation-850d11f72046#:~:text=Spearman's%20rank%20correlation%20requires%20ordinal,%2C%20Neutral%2C%20Disagree%2C%20Strongly%20Disagree
5. Correlation Types and When to Use Them, https://ademos.people.uic.edu/Chapter22.html

Data Sources

1. The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data. This data is updated daily. There are no ethical considerations I

feel we have to consider with this dataset, as these are the actual cases that were reported publicly. Licensed under Attribution 4.0 International (CC BY 4.0). Overall, this dataset will contain the daily COVID case count, which will act as our dependent variable in the analysis.

2. The Trips by Distance dataset to understand daily travel behaviors of people in Cook County, Illinois. The dataset summarizes how many people are staying at home during the COVID-19 pandemic and how far people are traveling when they don't stay home. It consists of 22 columns out of which I felt 16 relevant for our analysis. This dataset is licensed under U.S. Government Works. I couldn't find any information on Terms of Use, however.