Correlation Between Unemployment Rate and Infection Rate in Denver, Colorado

Summer Ai

**Introduction**

Since March 2020, the COVID-19 pandemic has had significant impacts on peoples’ lives in America. The contagious nature of such a virus has fundamentally changed our way to socialize and work with each other. To mitigate the spread of COVID, policies such as mandatory masking and work-from-home have been implemented by governments and companies. However, the influence of those policies is hardly measured. To quantify the effectiveness of such restrictive measures on public health and economy, this project aims to demonstrate the effect of masking on number of confirmed COVID cases and explore the correlation between unemployment rate and infection rate in Denver, Colorado.

**Background**

In the time of COVID, many research has been done on the unemployment crisis, including topics from making predictions to identifying groups at risk. Studies have shown that marginalized groups and young people are particularly vulnerable to poor job conditions (International Labor Organization, 2020) and unemployment rate is estimated to converge to around 5%, similarly to pre-pandemic level, by the end of 2021 (Şahin et al., 2020). While there is research focused on correlation, such as the association between unemployment insurance and food security (Raifman et al. 2020), not enough research studied the correlation between unemployment rate and COVID-related metrics.

For this project, there are two main research questions, both of which focus only on Denver, Colorado. First, I am interested in exploring if and how masking policy changed the progression of confirmed COVID cases. Second, I would like to figure if there is a correlation between unemployment rate and infection rate.

**Method**

There are three datasets used throughout the project. The first one is the COVID-19 data from John Hopkins University, which records cumulatively confirmed cases per country in the U.S. The second dataset is the mask mandates dataset from CDC, which documented masking policy per country in a time series. The third data comes from the website of U.S Bureau of Labor Statistics that contains monthly data about unemployment rate in the US.

To answer the first research question, I decided to use visualizations to show how number of confirmed COVID cases and infection rate change with different status of masking requirement. I started by smoothing the COVID-19 data by substituting daily confirmed cases with 7-day moving average to avoid seeing too many spikes later in the visualization. Then, I defined daily infection rate as number of daily confirmed cases divided by population at risk, which is the difference between total population in Denver and number of cumulative confirmed cases.

As for the second research question, I started by calculating the Pearson correlation coefficient, which ranges from -1 to 1, to measure the global synchrony of the two time series, unemployment rate and daily infection rate (Wright, 1921). Although Pearson r provides a general insight about the data, I furthered my investigation by looking for dynamic signals using Time Lagged Cross Correlation (Podobnik, B., & Stanley, H. E, 2008).

The Time Lagged Cross Correlation identifies how two time series synchronize over time by repetitively calculating the correlation between them, with one time series constantly shifting to create a time offset. In addition, the Cross Correlation method assumes the stationarity of time series, which makes sense in our case since neither unemployment rate nor COVID-19 are heavily affected by seasons.

To summarize what I found from calculating correlations between unemployment rate and infection rate, I compared the two rates and visualized the results over the same timespan to give a more straightforward representation of how they are related.

Throughout the analytical process, human-centered data science principles have been constantly followed to make sure there is no ethical problem for both the datasets and the methods used. All datasets don’t reveal any personal information, rather they only give aggregated data at the county level. Among the three data sources, the data about unemployment rate is the only one that involved surveys to generate final estimates. According to the U.S Bureau of Labor Statistics, the surveys designed strictly protected the confidentiality of the participants to ensure no information, either about individuals or businesses, would be misused (Looking for information about a specific survey?, n.d.). In addition, the response provided will only be used to generate statistics about groups with specific characteristics rather than specific individuals. As for the two statistical measures used in the study, research ethics were strictly followed to make sure results were not manipulated in any way and conclusions were drawn directly from the analysis based on the data. To ensure the reproducibility of the research, the original coding will be provided via a GitHub repository along with detailed descriptions for each step and appropriate copyright for all the data.

**Results**

As shown in Fig 1, the two time-series graphs indicate the trend of number of confirmed cases and infection rate for Denver, Colorado from February 1st, 2020, to October 15th, 2021. Both graphs are encoded with three colors to represent the mask policy for a specific range of dates, with black being unknown, red being required, and green being not required. Viewers can first look at the top graph to get a sense of how the number of cases changes daily, and then look at the bottom one to observe the trend of infection rate.

The x-axis for both graphs indicate the month and year of the data. For the top graph, the y-axis is the number of cumulatively confirmed cases ranging from 0 to 81000. For the bottom graph, the y-axis is the daily infection rate, which is calculated as defined in the previous section and can be interpreted as the derivatives of the number of cumulatively confirmed cases.

The effect of masking policy on confirmed cases is obvious from the two graphs. The most rapid increase happened around late November 2020 as the slope is very steep on the first graph and daily infection rate is very high in the second plot as well. Among 9 months of masking policy, the most rapid increase happened about halfway through, yet the derivative of confirmed cases dropped dramatically since January 2021. The slope of the first graph is a lot smoother, and the infection rate is significantly smaller in the second graph. Such a trend even continues when masks are no longer required in April 2021, which means a relatively safe environment is guaranteed with a low infection rate in Denver.

Graphical user interface, chart

Description automatically generated

Fig 1. Number of cumulatively confirmed cases and daily infection rate in Denver, Colorado

The trend of unemployment rate in Denver between January 2019 and October 2021 is shown in Fig 2 below. There was an unprecedented rapid increase around April 2021, when multiple regions in the U.S started to see a rapid infection of COVID-19. However, the unemployment rate didn’t stay at the peak for very long. It rapidly decreased from 13% to around 9% three months later, and it has been steadily decreasing until October 2021. Although continuous declining trend is reassuring, it worth noting that the unemployment rate has not returned to pre-pandemic level yet. To examine if there is any correlation between unemployment rate and infection rate in Denver, I only focused on the unemployment rate from February 2020 to make sure the two time series share the same length of time.

As indicated in Fig 3, when putting the two rates in the same graph, I noticed there is no obvious correlation overall. However, to find evidence supporting my claim, I computed the Pearson coefficient correlation score for the two time series as planned. The Pearson r for the two rates is around -0.035, which indicates that there is almost no correlation between them. However, as addressed in the last section, Pearson r only provides insights about the global synchrony without any details about dynamic signals. Because it is possible to see the trend of one time series happen a few days before that of the other, I continued the analysis by calculating the cross correlation between the two rates.

Chart, line chart

Description automatically generated

Fig 2. Unemployment rate between Jan. 2019 and Oct. 2021 in Denver, Colorado

Chart

Description automatically generated

Fig 3. Comparison between infection rate and unemployment rate in Denver, Colorado

For cross correlation, I repetitively calculated Pearson r score as the unemployment rate time series was shifted to the right or to the left, at an increment of one day once at a time, which yields a list of Pearson r score as a function of time offset in days. The graph of the cross-correlation analysis is shown in Fig 4. While the dashed black line shows the correlation when there is zero offset, the red dashed line indicates the time offsets when the greatest correlation was observed, which is 203 days. Referring to Fig 3, it was not surprising to notice that the time difference between the peak of the unemployment rate and the infection rate was about seven to eight months, which means the strongest correlation indicated by cross correlation is for the two peaks.

Chart, line chart

Description automatically generated

Fig 4. Pearson r score between unemployment rate and infection rate for different offsets

However, if the one time series were to trigger or happen before another, 203 days is too large for an ideal synchrony to happen. For example, if one believed the rapid increase in unemployment rate followed by an increase in infection rate in Denver, then the ideal synchrony between the two rates should be observed when the offset is between approximately negative 7 to negative 30 days, not 203 days. Therefore, both a single Pearson r and the cross-correlation analysis yield the same result. That is, there is no evidence to suggest a strong correlation between the unemployment rate and infection rate in Denver, Colorado.

**Discussion**

The analysis for the two research questions were very interesting and it is important to discuss why the results might be counterintuitive at first. Although Fig 1 has demonstrated the effect of masking policy in Denver, one may still question why the peak of infection rate appeared in the middle of the period when masks were mandatory, and even higher than the time when masks were not required.

In this case, there are two factors that need to be considered. First, the highest infection rate in Denver was in late November and early December 2020, which is around Thanksgiving when people travel around to see their families. More number of visitors would test for COVID before boarding, so it was likely to identify many positive but asymptotic cases than usual. Thanksgiving was also a chance for people to meet with family members they don’t often see, which could also increase the rate of transmission to cause the peak. Second, though a rapid decrease in infection rate was not observed right after the implementation of mandatory masking policy, the exponential growth of rate of infection, as indicated in Fig 1, could otherwise be non-stoppable without it. Further research could investigate potential reasons of seeing the highest infection rate in the middle of the period when masks were required. If holiday season were one of the main reasons for the rapid increase, it would further support the effectiveness of masks because the choice of people unmasking themselves in gatherings could associate with a higher infection rate.

As for the research about unemployment rate, whenever studying the unprecedented spike in April 2020, it was rare to not connect it with the pandemic. However, my finding revealed that, at least for Denver, there is no strong correlation between unemployment rate and infection rate, which is one of COVID-19 metrics. Although statistical significance was not detected, there are still many factors that can connect COVID-19 with high unemployment rate. For example, at the beginning of the pandemic, the public could be extremely panic over things they had never experienced before. The fear for an unforeseeable future may lead many companies to adapt more conservative approaches, such as stop recruiting new employers or even laying off the current ones to minimize the probability of going through a financial crisis. Normally, such factors are harder to quantify and therefore rare to see relevant research. To study unemployment rate during the pandemic, future research could focus on different COVID-19 metrics, such as consumers’ panic index, to explore their relationships. Overall, it is likely that a combination of many factors led to the high unemployment rate observed in April 2020, which explains why it is hard for researchers to detect significant result from a one-to-one metric comparison. However, reporting insignificant result is also very important for the literature so that other researchers do not repetitively conduct similar studies and can pick up the analysis from where they have already known.

**Limitation**

In general, there are three critical limitations of the current study, ranging from flaws in data to assumptions in statistical methods, that could have impacted the result in an important way.

First, while daily number of confirmed cases is easy to obtain, the unemployment rate data is only monthly. To identity the correlation between the time series, unemployment rate was assumed to be stable over each month, which is not necessarily true in the real life. However, daily unemployment rate is hard to estimated or obtained, so the effect of data with two types of frequencies is hardly changeable, unless other methods are applied to gain daily unemployment rate, such as smoothing techniques.

In addition, the definition of our key metric, infection rate, was based on personal assumptions. As mentioned before, daily infection rate was calculated as the ratio of daily smoothed confirmed cases to population at risk, where population at risk was defined as the difference between total population and number of cumulative cases. Such a definition not only excludes number of death due to COVID but also assumes recovered people won’t contract it again. If there were a serious understatement or overstatement about the number of death due to COVID in Denver or disproportionally high reinfection rate, then the estimated daily infection rate would be a lot different than what was used in the study. Therefore, careful examination about the assumptions made for such a definition are needed should there are relevant research papers.

The last major limitation of this study relates to the assumption of using Pearson r. The correlation coefficient score has been repetitively used for the detection of both the overall and the local synchrony. One assumption held by Pearson r is the homoscedasticity of the data, which might not be true given the fluctuations in both two time series. In the case of cross correlation, an addition assumption is the stationarity of the time series, which was mostly satisfied because neither unemployment rate nor infection rate were heavily affected by seasonality. Because of the limitation for the assumptions of analytical method, the power of relevant statistics was inevitably less than if all conditions were met. Therefore, it is important to point this out for readers so that they can interpret the results with an appropriate context.

**Conclusion**

For this research, there were two main questions I would like to address. First, I was curious to figure out how the masking policy changed the progression of confirmed cases in Denver, Colorado. The visualization greatly helped me gain an understanding about the trend of confirmed cases as well as infection rate, and how they were impacted by mandatory masks. As for the peak observed around December 2020, I proposed different hypotheses that would be in the interest of future researchers. Despite the highest infection rate happened when masks were required, it did not counter the observation regarding the positive influence brought by mandatory masks in Denver.

As for the second question, I would like to see if there was any correlation between the unemployment rate and infection rate in Denver. After computing Pearson r for overall and local synchrony, it turned out there was no evidence supporting a strong correlation between the two time series. However, other relevant factors, as discussed in detail before, can be investigated by future research to investigate the relationship between unemployment rate and COVID-related metrics.

Doing this project has greatly enhanced my understanding of human centered data science by allowing me to go through every step that data scientists have to follow for most of the analysis projects. From collecting to cleaning data, I was constantly reminded to think about potential ethical issues, such as privacy, and validity of assumptions, such as the term ‘population at risk’. When doing the analysis, I was more aware of algorithmic bias brought by partial violation of statistical assumption than before, and I recorded them in the paper to make the whole pipeline more transparent and accountable. Legal considerations, such as copyright of data, along with other documentary steps will be provided as well to ensure the reproducibility of the project and help future researchers to extend relevant studies.

**References**

International Labor Organization. (2020). World Employment and Social Outlook – Trends 2020.

Retrieved December 10, 2021, from https://www.ilo.org/global/research/global-

reports/weso/2020/lang--en/index.htm.

Jin Hyun Cheong, P. D. (2021, March 13). *Four ways to quantify synchrony between time series*

*data*. Medium. Retrieved November 10, 2021, from

https://towardsdatascience.com/four-ways-to-quantify-synchrony

-between-time-series-data-b99136c4a9c9.

Podobnik, B., & Stanley, H. E. (2008). Detrended cross-correlation analysis: a new method for

analyzing two nonstationary time series. *Physical review letters*, *100*(8), 084102.

Raifman, J., Bor, J., & Venkataramani, A. (2021). Association between receipt of Unemployment

Insurance and food insecurity among people who lost employment during the COVID-19

pandemic in the United States. *JAMA Network Open*, *4*(1).

https://doi.org/10.1001/jamanetworkopen.2020.35884

Şahin, A., Tasci, M., & Yan, J. (2020). The unemployment cost of COVID-19: How high and how

long? *Economic Commentary (Federal Reserve Bank of Cleveland)*, 1–7.

https://doi.org/10.26509/frbc-ec-202009

U.S. Bureau of Labor Statistics. (2016, August 5). *Linking and copyright information*. U.S.

Bureau of Labor Statistics. Retrieved November 10, 2021, from

https://www.bls.gov/bls/linksite.htm.

U.S. Bureau of Labor Statistics. (n.d.). *Looking for information about a specific survey?*

U.S. Bureau of Labor Statistics. Retrieved November 10, 2021, from

https://www.bls.gov/respondents/.

Wright, S., 1921. Correlation and causation. *Journal of agricultural research,* 20(7), pp. 557–585

**Data Sources**

There are three data sources used in this project.

The first one is the [COVID-19 data](https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv) from John Hopkins University.

Licensed under [Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/legalcode), and I am allowed to share and adapt the data.

The second one is the [mask mandates](https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i) data from CDC.

As per [Q&A section](https://www.cdc.gov/Other/policies.html) on CDC, this dataset is in the public domain and may be freely used or reproduced without obtaining copyright permission.

The third one is the [unemployment rate](https://beta.bls.gov/dataViewer/view/timeseries/LAUCN080310000000003) data from U.S. Bureau of Labor Statistics.

According to the [linking and copyright information](https://www.bls.gov/bls/linksite.htm), this dataset is in the public domain and is free to be used without specific permission, but required to cite the U.S. Bureau of Labor Statistics as a source.