Project Report

Introduction

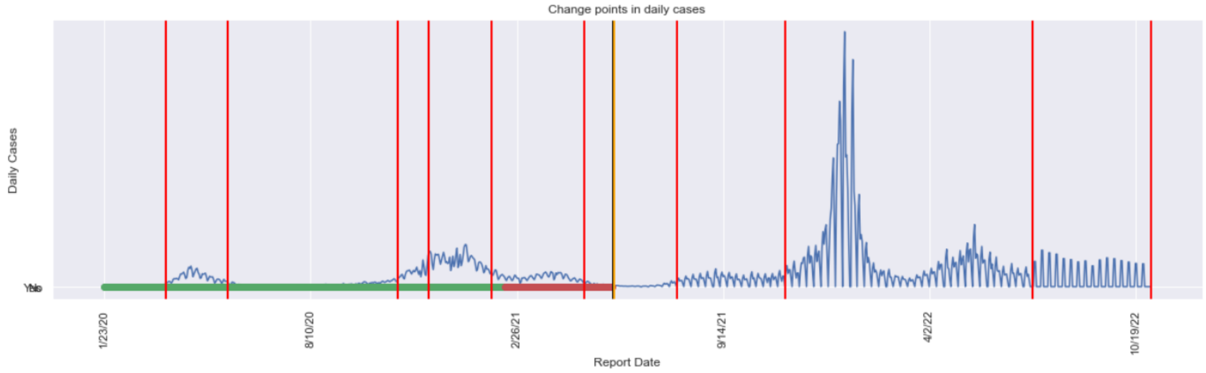
The Coronavirus disease 2019 (COVID-19) poses a severe threat to humans and alters human society. Although prevention and control measures are important in preventing the spread of SARS coronavirus, signs of the existing cases and other factors also affect the further spread of COVID. It is critical to comprehend the migration of covid and how external elements like socioeconomic, climatic, and health aspects are influencing the same, especially because the disease is still spreading and progressing after three years.

Socioeconomic factors are critical in the spread of emerging infectious diseases and a few studies have looked into the role of socioeconomic factors in COVID-19 spread. Hence, the goal of this analysis is to understand how the covid changes over time and how it relates to different socioeconomic and health factors. Understand the contribution of such external factors further so that policymakers and public health officials can use the results to take additional precautions and implement the rules and regulations. Furthermore, data scientists, economists, and public health officials can expand on the results produced to perform causal analysis or any other required analysis to further assist policymakers to make better-informed recommendations, as well as predict the covid cases that may be influenced by the aforementioned factors.

Background/Related Work

Hypothetically, the correlation between X and Y is unity in time series data. However, in reality, we know there are a lot of external factors that show the correlation between x and y is not equal to unity. The same goes for the time series data for the confirmed covid cases. There might be some correlations that are being affected by external factors such as weather, economic, social, health, and many other aspects.

The Mask mandate analysis of this project revealed that 14-day masking and prevention was a significant factor in controlling covid cases. As we can see from the below graph, soon after implementing the masking policy (shown by the orange line) the change in covid cases became stagnant and reduced to a minuscule number as compared to a large number of cases.



However, after a few weeks, the change in the number of covid cases increases and decreases, it kind of fluctuates a lot. So it is pertinent to understand this behavior so that a collective effort can be taken efficiently to curb the disease.

Therefore the research question for this analysis is, "Is there any relationship between the daily covid cases data and the socioeconomic factors in Middlesex county?" It will help in understanding whether the covid cases are affected by any of the socioeconomic and health factors such as income, health issues, housing, and many more. In addition, we also want to understand why Middlesex has the highest covid confirmed cases in the state of Massachusetts.

This analysis is relevant as a similar analysis was performed for major cities of China (where the covid originated from) thus highlighting that this research helped the Chinese officials in helping its citizens. This study helped the officials apprehend that the travelers from Wuhan and rural-to-urban migrants were linked to the COVID-19 outbreak in 39 of China's most developed cities. These findings suggested that in the early stages of the COVID-19 outbreak in well-developed cities, travelers from an epicenter and rural-to-urban migrants should be given special attention. (Lin et al.)

Moreover, a similar investigation was conducted by the KFF (Drake and Rudowitz) that helped them understand the public health and economic effects of the pandemic on the well-being of many people living in the United States.

Methodology

For this analysis, the number of covid-19 confirmed cases have been filtered for Middlesex county from the county-wise covid-19 cases. Then this data is merged with the socioeconomic county-wise data for Middlesex.

In non-time series data, the concept of correlation is the same: identify and quantify the relationship between two variables. Because time series data is continuous and chronologically ordered, there is a chance that there will be some degree of correlation between the series observations.

In the context of time series analysis, measuring and analyzing the correlation between two variables can be understood in two ways:

- Analyzing the relationship between a series and its lags, because some of the past lags may contain predictive information that can be used to forecast series events. The autocorrelation function and partial autocorrelation function are two of the most widely used methods for determining the degree of correlation between a series and its lags.
- Analyzing the correlation between two series to identify exogenous factors or predictors that can explain the series' variation over time. In this case, correlation is typically measured using the cross-correlation function. ("How to Perform Correlation Analysis in Time Series data using R? - Luba")

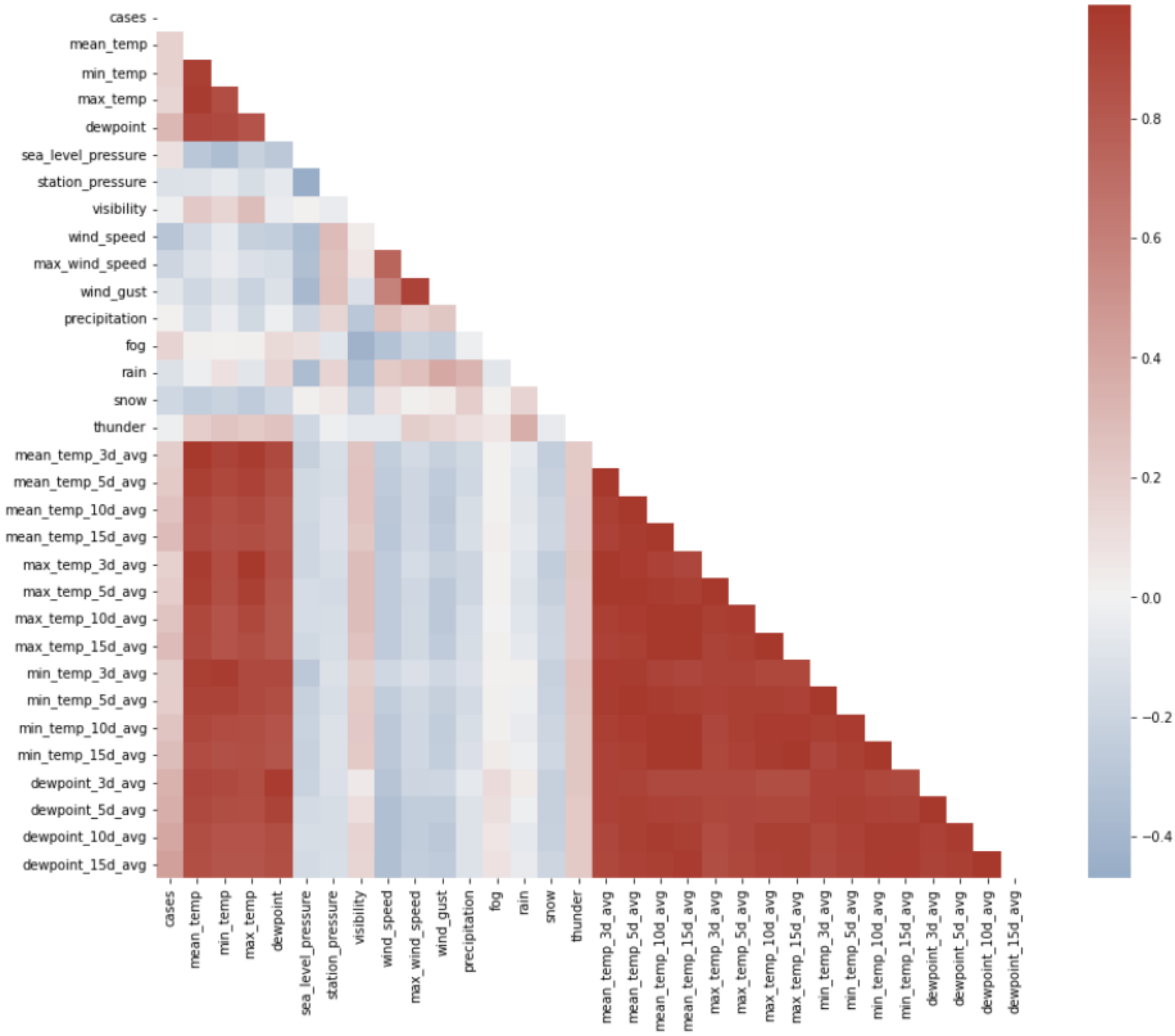
Therefore, the correlation analysis which is lag and causal analysis is performed, to check for any surprises that might arise and to explore the data more. Moreover, check of correlation of cases with any of the factors such as housing, health, income, etc. With the help of this analysis, we will be able to determine what affects the covid cases more.

For the second part of the analysis, to understand why Middlesex has the highest number of recorded covid cases, the number of COVID-19 cases in the counties in Massachusetts was aggregated. Then plotted the regression analysis plot to understand how they correlate to the socioeconomic factors.

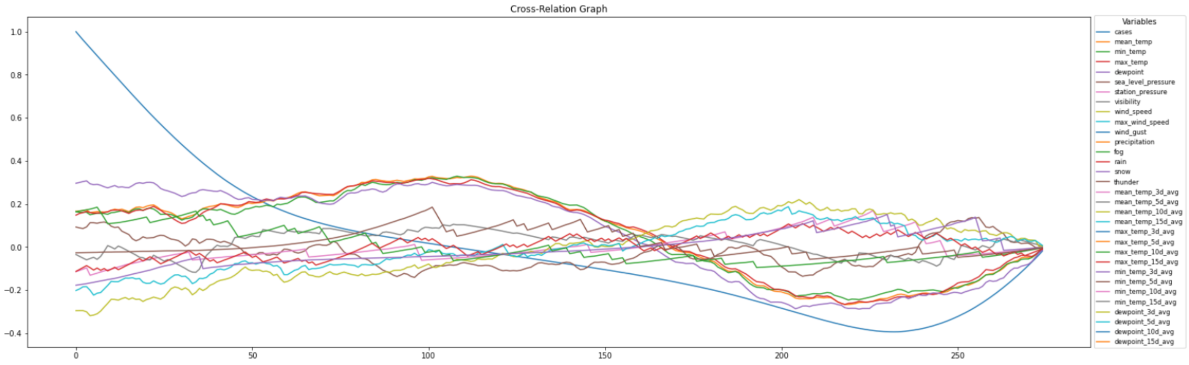
After performing a manual investigation on the dataset, I found that a few attributes are empty to get rid of them before starting the analysis. Moreover, I found that most of the attributes were constant throughout. So made a new data frame for those variables for the analysis. And converted the categorical to the numerical values, since these will affect the correlation analysis.

Furthermore, visualized the correlation using heatmaps to determine the correlation coefficient among various factors. Then I will set up a threshold that is not zero to comprehend the affecting factors for increasing covid cases.

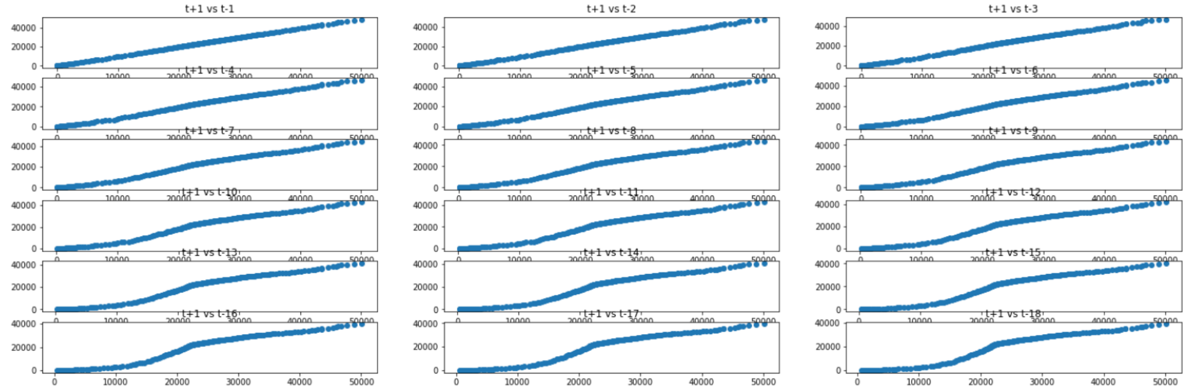
Findings



From this heatmap, it can be easily said that the temperature and dewpoint are highly correlated to the spread of covid cases. Moreover, not just the current values, the 3-day average, 5-day average, 10-day average, and 15-day average are also contributing significantly.

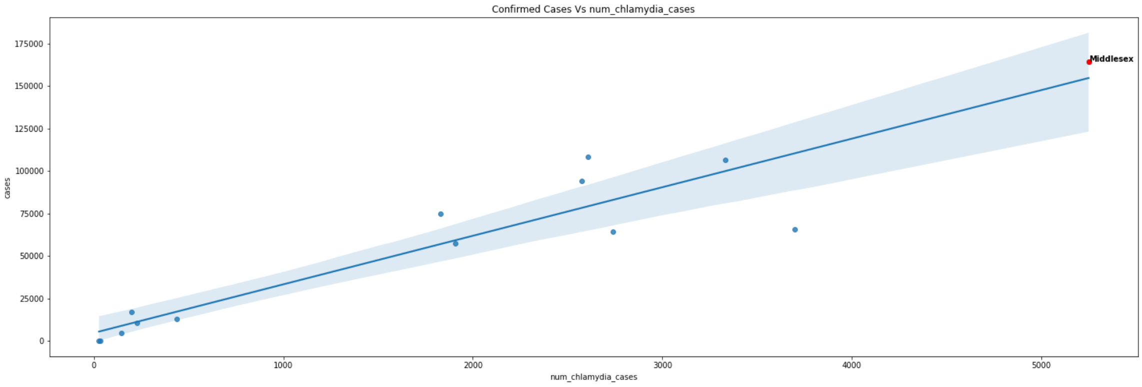


This cross-correlation plot, further corroborates our heatmap analysis, that the covid cases are indeed affected by lagged temperature and dewpoint weather data.

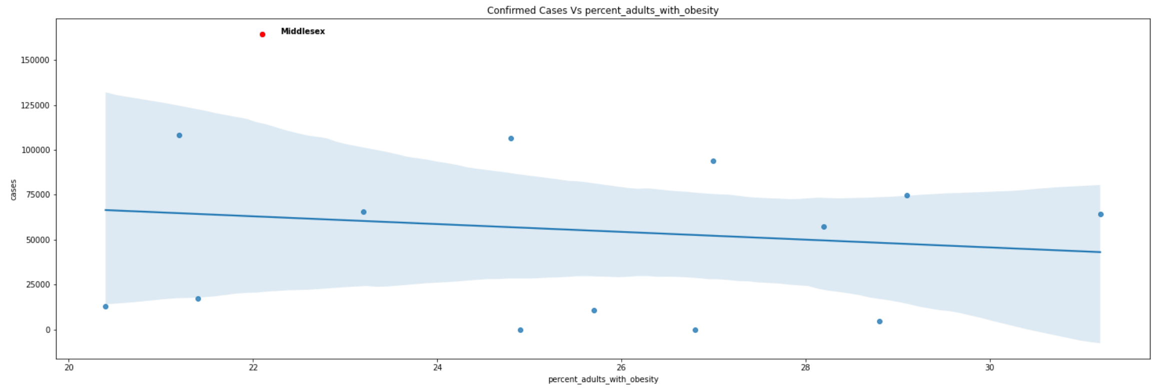


The covid cases are not only affected by the weather conditions but also by their historical value. In the above plot, the plots show a linear pattern between confirmed cases and their 14-day lag, it suggests autocorrelation is present. A positive linear trend (i.e. going upwards from left to right) is suggestive of positive autocorrelation. Moreover, The tighter the data is clustered around the diagonal, the more autocorrelation is present. This plot confirms that the current cases are highly correlated with the last 14 days of the plot.

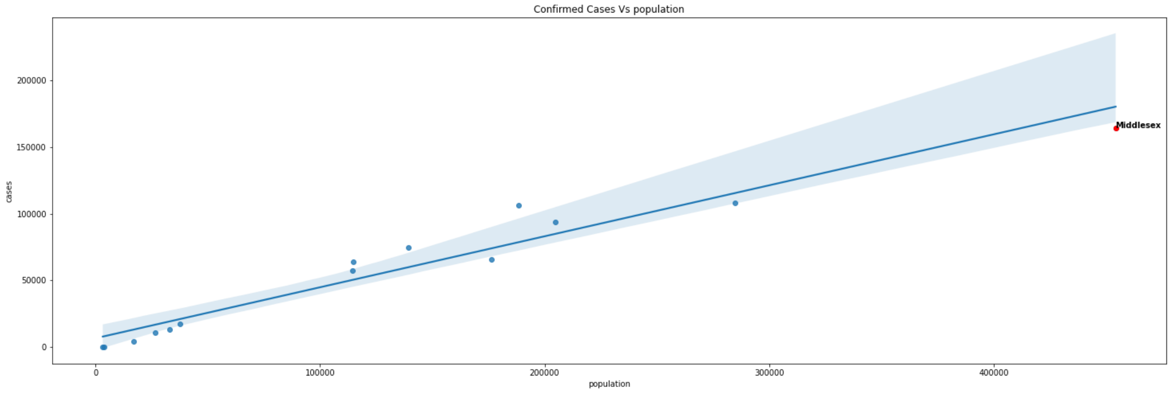
The regression plots tell us that there are a lot of factors that are correlated with the covid and there are many factors that are not related at all and do not impact on covid at all. A few of the regression plots are shown below, you can find the rest of the plots in the GitHub repository in the folder "*./pic/relational_pic/*".



This graph shows that the chlamydia rate and covid cases are correlated together. Moreover, Middlesex has the highest chlamydia and covid cases recorded.



The plot conveys that covid and obesity are not related at all. Moreover, obesity is not contributing to Middlesex's covid cases. And additionally, the population is also one of the major contributing factors in the increase of covid confirmed cases.



After investing all the plots we find that covid cases in Middlesex are highly correlated with the number of uninsured, % with access to exercise, population, num of mental health care providers and physicians, income, food environment index, Chlamydia cases, deaths, dentist, etc. On the other hand, factors like obesity, children in poverty, excessive drinking, physical inactivity, single-parent households, % smokers, population density, crime rate, etc do not have ramifications on covid cases.

Discussion/Implications

Over 4.3 million confirmed cases and over 290,000 deaths have resulted from the COVID-19 pandemic worldwide. It has also raised concerns about the impending economic crisis and recession. Social isolation, self-isolation, and travel restrictions have resulted in a reduced workforce across all economic sectors, resulting in the loss of many jobs. Schools have closed, and the demand for commodities and manufactured goods has dwindled. In contrast, the demand for medical supplies has skyrocketed. The food industry is also experiencing increased demand as a result of panic buying and stockpiling of food products. Thus socioeconomic effects of COVID-19 impact the individual level as well as the global level, it is indispensable that we understand these impacts and implications to their fullest (Nicola et al.).

Now we have found the important socioeconomic factors or highly correlated aspects. Data scientists, and data analysts, can use these features to create an appropriate machine learning model that takes into account the external factors and not just the historical values to help in predicting future covid cases and their movement.

Economists can use these research findings to come up with economic ideas and tools to better understand the crisis and come up with good ideas for policy. Additionally, economists are very creative in finding new data or new ways to display them and thus can contribute in important ways to better policy options on top of this research (Bütler).

Based on the analysis and suggestions provided by the data scientists, economists, and public health officials the policymakers can make informed decisions and laws to mitigate the pandemic. Moreover, using this analysis the government, policymakers, and politicians can be better prepared for any future pandemic.

Limitations

This analysis also has certain limitations like data cleaning techniques, statistical techniques, specific assumptions, and preconditions which are -

1. This study is limited to Middlesex county. So this study can be extended to any other by performing a similar analysis and changing that data as required for any particular city/county/ or country in question and then concluding results based on that analysis. In other words, this analysis is reproducible but the results cannot be projected for any other city, county, state, or country.
2. There were a few attributes such as precipitation, average grade performance, etc. that had null values in them. So I dropped these attributes, so this analysis cannot comprehend these attributes.
3. This analysis does not consider how strictly the mask mandate laws were being followed, which was the initial motivation of this analysis.
4. Different datasets were available for different data ranges. So after merging them the dataset and analysis are now performed on Jan 21, 2020, to Dec 04, 2020.

5. The time series analysis is often done on stationary time series data. So for this analysis, we have assumed that the time series in question is stationary in nature.

Conclusion

This analysis attempted to examine the effect of Covid-19 and the mask mandates policy that was put into effect at the beginning of covid. Specifically, I tried to apprehend the sudden increase in covid cases even though the mask mandate was implemented. The hypothesis was that the external factors were contributing to the same and the finding also concluded that. Many socioeconomic and weather factors were actually correlated with the increase in covid-19 cases in Middlesex county making it the county with the maximum number of confirmed cases in the state of Massachusetts.

References

- Bütler, Monika. "Economics and economists during the COVID-19 pandemic: a personal view - Swiss Journal of Economics and Statistics." *Swiss Journal of Economics and Statistics*, 8 November 2022, https://sjes.springeropen.com/articles/10.1186/s41937-022-00097-1 . Accessed 12 December 2022.
- Drake, Patrick, and Robin Rudowitz. "Tracking Social Determinants of Health During the COVID-19 Pandemic." *KFF*, 21 April 2022, https://www.kff.org/coronavirus-covid-19/issue-brief/tracking-social-determinants-of-health-during-the-covid-19-pandemic/ . Accessed 12 December 2022.
- "How to Perform Correlation Analysis in Time Series data using R? - Luba." *LOB.DATA*, 15 September 2020, https://www.lobdata.com.br/2020/09/15/how-to-perform-correlation-analysis-in-time-series-data-using-r/ . Accessed 12 December 2022.
- Lin, Yiting, et al. "Association Between Socioeconomic Factors and the COVID-19 Outbreak in the 39 Well-Developed Cities of China." *Frontiers*, Frontiers in Public Health, 25 September 2020, https://www.frontiersin.org/articles/10.3389/fpubh.2020.546637/full . Accessed 12 December 2022.
- Nicola, Maria, et al. "The socio-economic implications of the coronavirus pandemic (COVID-19): A review." *NCBI*, National Library of Medicine, 17 April 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7162753/ . Accessed 12 December 2022.

Data Sources

This analysis research question will require several different datasets. The datasets are as follows:

1. The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data. This data is updated daily.

2. The CDC dataset of masking mandates by county.

Note that the CDC stopped collecting this policy information in September 2021.

3. The New York Times mask compliance survey data.

The majority of this data is by US County by Day. The mask compliance is a single-shot estimator that gives you a compliance estimate for every County in the US.

4. US Counties: COVID-19 + Weather + Socio/Health data Socioeconomic Time series dataset I

5. US Counties: COVID-19 + Weather + Socio/Health data Socioeconomic Dataset II

The dataset contains county-level data on health, socioeconomics, and weather can help us address to identify which populations are at risk for COVID-19 and help prepare high-risk communities.

For this analysis, the focus is on ***Middlesex county*** in the state of ***Massachusetts***.