Data 512 - Final Report

Table of Contents

1. Introduction

Covid-19 has impacted many lives over the past 2 years and forced us to live a new lifestyle for a long time. With the goal of finding answers to questions, backed with the available technology, humongous data of all sorts was collected. We were able to gather information like - how many cases are being reported every day, how many people are recovering from covid, what is the number of deaths every day, what is the availability of hospitals and medicines, and so on.

All this data in its raw stage just numbers, which can often be misleading. A number of visualizations were shared across social media displaying the increasing number of cases, irrespective of the measures taken by them. Looking at these charts, people believed that covid is inevitable and also that taking vaccines or wearing masks will not make a difference. I believe that simply presenting raw numbers, cumulative values or increasing time series is unjustified as that provides incomplete information to people.

Hence, with this project, my objective is to understand how changing our lifestyle impacted the spread of covid. This problem is human-centered because it aims to highlight how our actions helped in reducing the spread of covid. In this project, I performed data analysis to bring to light the actual information and hope to show that covid 19 did impact our lives and that actions like the use of masks, or getting vaccinated did make a difference.

2. Background/Related Work

My research focussed on understanding the impact of masking policies and vaccinations on the spread of covid. Thus, the selected research problem is as follows:

Research Question: *How do the actions, such as vaccination, hospitalization, and masking policy, impact the number of cases and deaths during covid in Allegheny County?*

During my research on understanding the impact of masking policies and vaccinations, I came across a large number of articles talking about covid and displaying charts and dashboards based on the available data, but there were limited research papers published on this concept. While I could not

find research specific to my hypothesis, I found a few papers closely related to the same.

In the paper [1] COVID vaccine immunity is waning, Elie Dolgin talks about how the duration of vaccine-based immunity is still evolving. Dolgin mentions that while the vaccines are 90% effective in the first 3 months, it's only 70% effective after 6-7 months. What this means for the impact of vaccination is that we can expect a fluctuation in the number of cases based on the days when mass vaccinations were provided.

Another interesting research paper was Global impact of the first year of COVID-19 vaccination: a mathematical modeling study [3]. This paper performs mathematical and statistical analysis and concludes that COVID-19 vaccination has substantially altered the course of the pandemic, saving tens of millions of lives globally. However, inadequate access to vaccines in low-income countries has limited the impact in these settings. Of all the research papers and articles, these two were the most important ones and closely related to my research topic.

3. **Methodology**

3.1 Data Collection

Selecting the right data source is a crucial step for testing a hypothesis or answering a research question. The data needs to come from a trusted and reliable source. Based on the county I was assigned (Allegheny County), I focussed on looking through official government websites to find any relevant data source that would accompany my research question. The data for this problem comes from the following two sources.

The first data source is by John Hopkins University and is hosted on Kaggle [6]. The files taken from this section are *Raw_us_confirmed_cases.csv* and *Raw_us_death.csv* having data about confirmed cases and deaths during covid. The data is free for public use and has Attribution 4.0 International (CC BY 4.0) License.

The second data source is the department of health and is hosted on the Official Pennsylvania Government Website [7]. There are a number of datasets on this website and for this problem specifically, the data about

vaccination count in Allegheny county is stored under the COVID-19 Vaccinations by Day by County of Residence Current Health section [8]. The website does provide a Public Domain U.S. Government license for data usage.

The data used in this project is generated by humans and has been collected and shared with the public maintaining transparency. I have ensured that all data used provides a license for public use and in no way is biased towards a section of people.

3.2 Feature Engineering

The initial data extracted required a significant amount of data cleaning and pre-processing to ensure the extracted data is specifically for Allegheny County and is in a required structured format. Post that I extracted a few features to make the analysis -

Variable Name	Variable Definition
current_population	Updating daily population based on death count and people who got covid a day before
rate_of_infection	Defined using confirmed cases and the current population of the county

The rate of infection is was calculated using the following equation:

$$\text{Infection Rate} = \frac{Number\ of\ Infections}{Population\ Size}$$

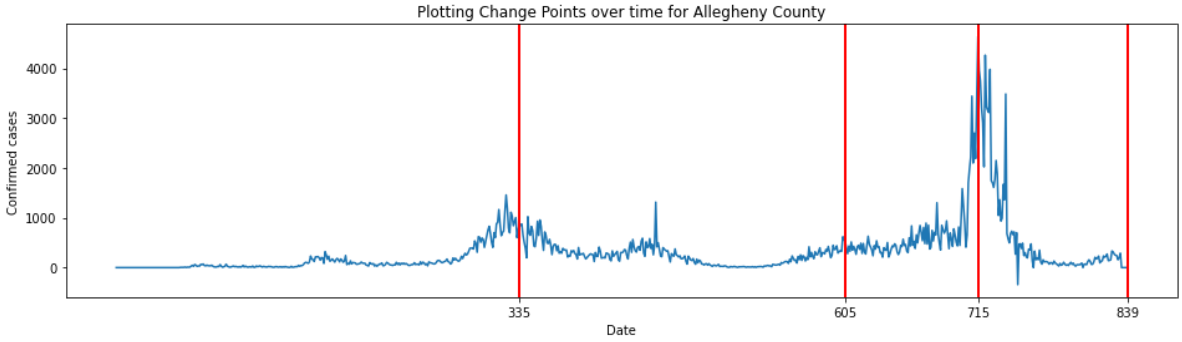
Source: New York Times

These variables were selected to ensure that the information communicated to the viewers uses updated data and instead of showing an upwards cumulative trend, depicts the actual change on a day-to-day basis.

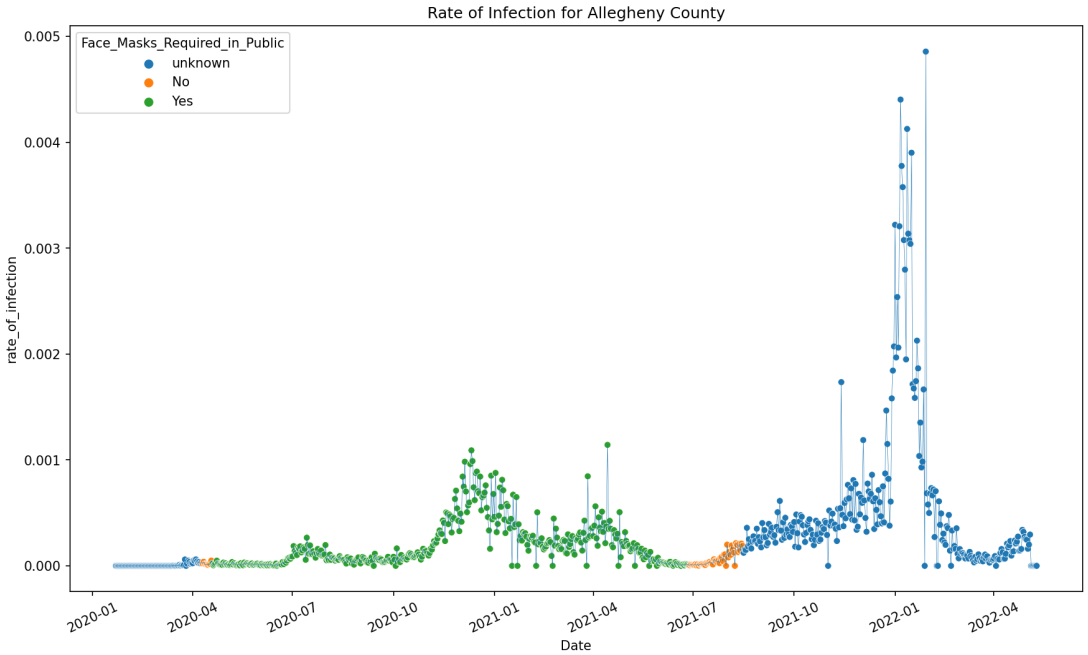
3.3 Preliminary Analysis

The initial analysis involved data from only John Hopkins University for the confirmed covid cases and masking policies. This analysis focussed on two parts - changepoint detection and graphical analysis.

The change point detection focussed on finding abrupt changes in data when a property of the time series changes. The notebook consists of reproducible codes to implement this on any dataset using the ruptures library in python. Below are the results from the change point detection -



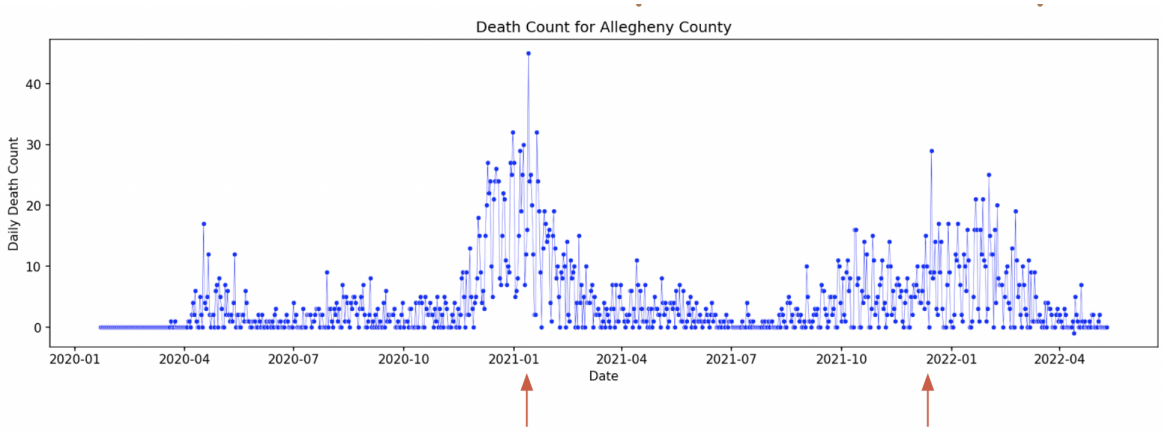
It was useful to understand how the change point detection was able to identify 4 points of change in the data. To better visualize where these abrupt changes happened, below is a plot depicting the rate of infection over time for Allegheny County along with the masking policies.



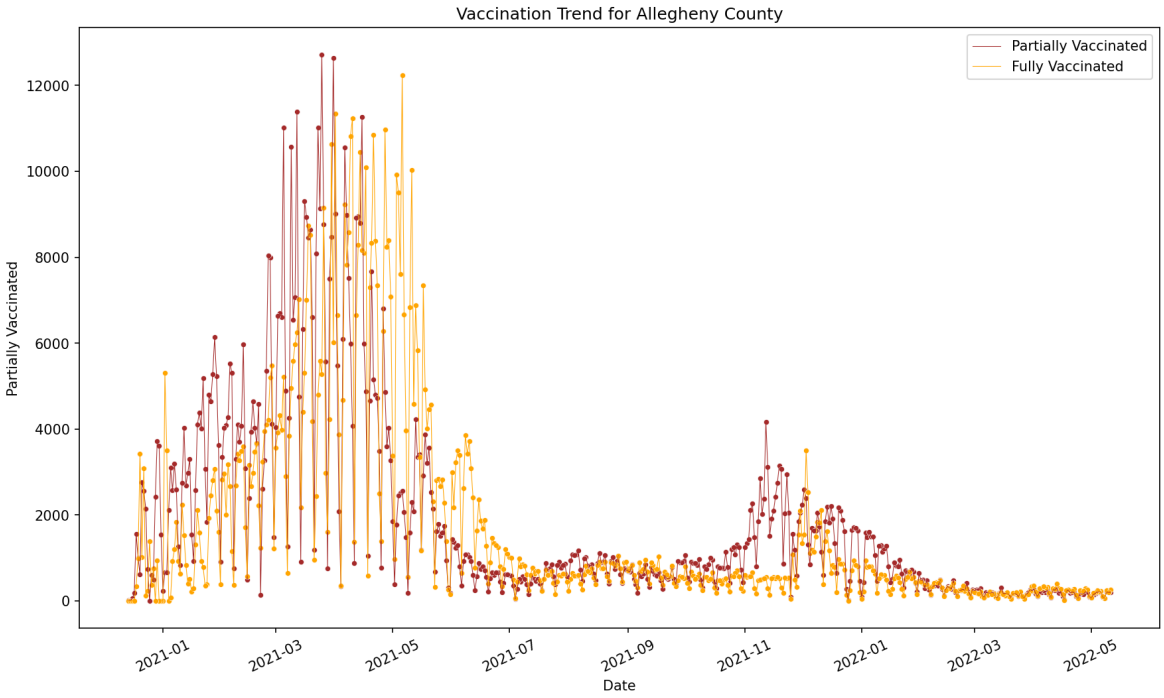
From the plot, we can see that we observed two waves in the span of 2 years. The first wave appears close to December 2020 and January 2021 while the second wave is seen in January 2022. We see that the mask mandate was removed in July and August 2021 and we can clearly see the spread of infection exponentially increased soon afterward. During the second wave, the spread of infection and the number of confirmed cases were much higher than the wave one.

3.4 Detailed Analysis

After looking into the masking policies and infection trends, I started digging deeper in order to understand how the vaccination policy had an impact on the number covid cases and deaths. The objective here was to understand the trend in the number of deaths (at a daily level), the trend in vaccination (at a daily level), and how these correlate with themselves and with the previous analysis. Here is the initial trend for the number of deaths due to covid in the Allegheny county -



The above plot shows the number of new death cases for Allegheny county and the highlighted dates shows the change points from the changepoint detection. These can also be seen as drastic changes in the number of cases when we observed the highest number of cases or peaks. A major difference between these two peaks is the introduction of vaccines. In Allegheny county, the vaccination process started in early 2021. Furthermore, when we look through the vaccination policies, we see the following trend -



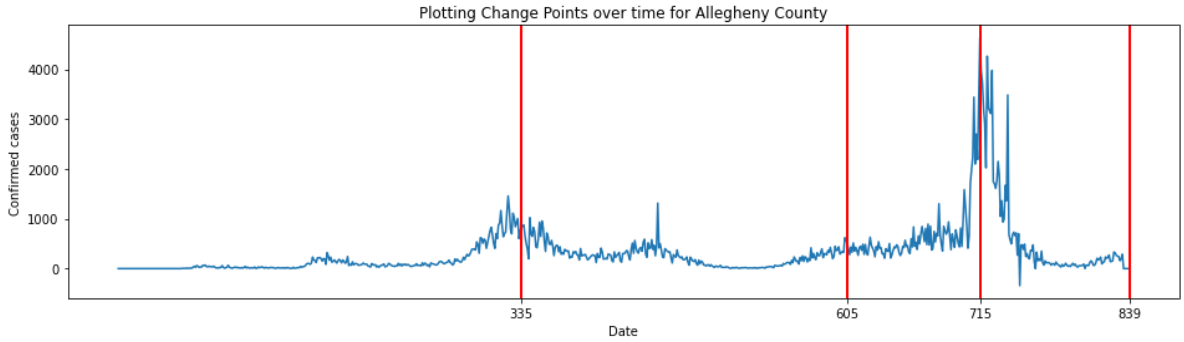
Further crunching the numbers shows that about 50 percent of the population was vaccinated within the first 6 months. We observed that as people started getting vaccinated, even though the cases increased rapidly, the number of deaths did not increase at the same rate.

4. **Results**

4.1 Findings

Changepoint Detection:

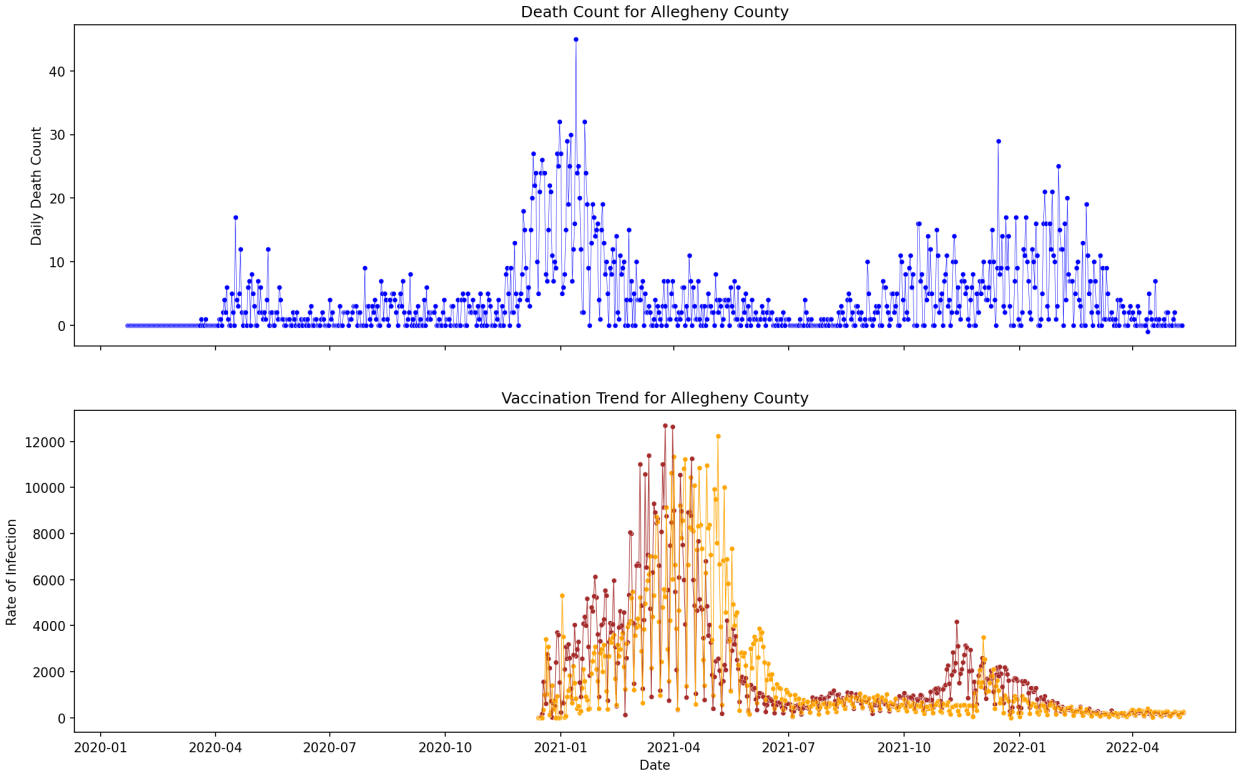
The project showed very promising results. The changepoint detection accurately predicted the rise and fall of both waves (hence 4 extreme changepoints). This was a good starting point and led to further analysis around this duration.



Data Analysis:

Both data analysis sections focussed on finding answers to the one question - how did masks and vaccinations help with the number of covid cases and deaths?

After a detailed graphical analysis we observed that for months with no mask mandate, the rate of infection was really high but with a mask mandate, the rate of infection was under control. On the other hand, the number of deaths was lower post-vaccination (even though the rate of infection and confirmed cases were 3 times than before).



4.2 Discussion/Implications

Why is this project important?

The goal of this project was to understand how our actions made an impact on the spread of covid. As we are being greatly impacted by covid, financially, physically, and emotionally, it is important to know

what steps can help us better fight this. Being able to show, through numbers and graphs, that us changing our lifestyle - wearing masks and getting vaccinated - has in turn made us stronger and more immune to the virus is indeed helpful.

The results of the study demonstrated that the use of masks helped reduce the spread of covid infection. As the masking policies were removed, we observed a surge in the number of covid cases and the rate of infection. Moreover, the number of deaths on a daily basis reduced soon after the vaccination policy was in place.

How could future research build on this study?

While we are able to see that the mask mandate and vaccination policy have helped keep the covid in control, there can still be a number of questions that would help gain a deeper understanding. Future work on this would include understanding how the different vaccines (first dose, second dose, booster dose) help individuals. Understanding how people with one dose versus people with multiple doses react to the covid spread would create more awareness among the individuals and allow them to make the right decisions.

Why is this project Human-Centered?

The fact that coronavirus affects humans does not make this problem human-centered. In addition to that, throughout the project, I worked on ensuring that this project is aimed at helping people understand and see the real picture. Moreover, the data collected was from reliable sources with a public access license.

In addition to that, the complete analysis is reproducible and highly interpretable. Anyone with no technical background can still observe the visualizations and see how the trends have changed over time.

4.3 Limitations

Limitations for model building:

While the initial idea was to implement a time series model, it did not work well with the given scenario. Firstly the existing data did not satisfy the stationarity conditions and secondly, the implementation

became difficult to infer after a point. In order to keep the results reproducible and interpretable, changepoint detection and graphical analysis were selected as two major techniques in this project.

Assumptions taken analysis:

During the feature engineering phase, I created a feature called “rate_of_infection”. While I was able to update the current population at a daily level, it only took into account the deaths due to covid. There could be deaths for other reasons in the county which are not considered while updating the population.

Limitation for reproducibility:

Another limitation to keep in mind is that the data used and analysis done are specific to the selected county (Allegheny County). Although all the techniques used in this project can be used for any other county, the interpretations and results might vary depending on the selected county.

5. Conclusion

Covid hit the globe in unexpected ways and changed our lives forever. Wearing masks and making sure that we get vaccinations and booster shots in a timely manner has become a usual thing for us today. While there was a lot of data collected during this pandemic, just numbers make little sense and do not give out the correct information. Keeping in mind all these things, I wanted to show how the changes in our lifestyle impacted the spread of covid. With this motivation, I worked on answering the question: How do the actions, such as vaccination, hospitalization, and masking policy, impact the number of cases and deaths during covid in Allegheny County?

Through this research, I was able to share the trends of covid infection spread over time. Through the plots generated, I was able to conclude that having a mask mandate actually helped keep the infection under control during wave 1. And secondly, even though the rate of infection kept increasing in wave 2, the number of deaths did not increase at the same rate! I hope that these inferences encourage us to keep up with and move forward in life with the required precautions.

6. **References**

6.1 List of Publications

[1] Dolgin, E. (2021). COVID vaccine immunity is waning - how much does that matter? Nature, 597(7878), https://doi.org/10.1038/d41586-021-02532-4
[2] Padma, T. V. (2021). India's COVID-vaccine woes - by the numbers. Nature (London), 592(7855), 500–501. https://doi.org/10.1038/d41586-021-00996-y
[3] Watson, O.J. *et al.* (2022) *Global impact of the first year of COVID-19 vaccination: A mathematical modeling study*, *The Lancet Infectious Diseases*. Elsevier. Available at: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(22)00320-6/fulltext
[4] *Wesa Covid-19 Dashboard 90.5 WESA*. Available at: https://www.wesa.fm/health-science-tech/wesa-covid-19-dashboard

6.2 List of Data Sources

[5] New York Times - Mask wearing survey data https://github.com/nytimes/covid-19-data/tree/master/mask-use
[6] Covid19 Data from John Hopkins University https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university
[7] Commonwealth *of* Pennsylvania http://www.pa.gov/
[8] COVID-19 Vaccinations by Day by County of Residence Current Health https://data.pa.gov/Covid-19/COVID-19-Vaccinations-by-Day-by-County-of-Residenc/bicw-3gwi