DATA 512: Part 4: Canonical Written Report

1. Introduction

For more than 3 years we have been experiencing a global pandemic as never before. It has taken millions of lives as it spread throughout the whole world. Given that humans have never experienced such an event before, we had no preparation nor understanding on how to manage such a crisis causing multiple building blocks of our society to fail, leaving millions of households without any income while governments struggled to meet the needs of the people.

Given that the pandemic started in 2020, technology and globalization was widely available, allowing for petabytes of data to be collected during these last 3 years. This data availability is a great resource that we now have at our hands to do scientific research, allowing us to learn the multiple ways in which a global pandemic can hit our society and what were the ultimate consequences around this. That data scientists are taking time to analyze and understand this data will have a huge impact in scientific knowledge, governments and experts need to understand what the consequences of their actions were, which policies worked, and which didn't, they also need to understand which populations were most vulnerable and which they should have prioritized more attention to. Insights such as these will allow us to learn from our mistakes, grasp what are the top priorities in the event of a pandemic, and what institutions or policies we need to start building to prepare for a next one. Getting prepared is of utter importance if we want to avoid the tragedy we lived in 2020, optimistically allowing us to save millions of lives and mitigating the pain of the crisis in millions of households that are most vulnerable to these events.

Through this analysis I hope to find insights such as the ones mentioned before, if these insights are useful and novel for Oklahoma County, I will plan to share this in an online article such that this valuable information is shared to the wider community. Sharing has multiple goals, it inspires other data scientists to work on these data sets, it can also drive them to work on top of my work to find additional insights or it can simply inform politicians or advocates about what happened, I would additionally share some suggestions around how this could be prevented in a future pandemic event.

It is important to note that finding insights or contributing in any way to global knowledge around pandemic data is a human-centered problem. We are working with sensible data that represents a tragic story for millions of people, this data for each individual could mean death, pain, suffering, trauma, psychological distress, depression, etc. The pandemic really broke into everyone's lives in a heavily negative way and as such we need to treat this data with respect, being careful with any conclusions that we draw from it, recognizing any biases, caveats, and generalizations that we might find, always finding good explanations as to why this might be true. Nevertheless, as mentioned previously, it is of utmost importance to encourage thousands of data scientists to dedicate their time to work on these kinds of data problems, if a pandemic occurs soon and politicians or advocates take data-driven decisions with the insights that we might find in the data, we will be able to prevent a lot of harm, deaths, and destruction.

As described previously, the pandemic affected and changed almost every aspect of society and human life, hence, there are a vast number of research questions that we could attempt to analyze throughout this project. For part I, I analyze how masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 1, 2021. Given that Oklahoma County did not have any masking mandates imposed, I analyze how does "voluntary masking" looks like. For parts II, III and IV, given that it is a free research project and that I really want to try to find important insights that might help society. I first searched in the worldwide web for the different data sources that Oklahoma County had to offer and found many areas that could potentially be used for analysis such as diversity, economy, civic, education, housing & living, health, etc. Given that I wanted to focus on areas that had greater impact in our society due to the pandemic, I decided to narrow my search to only economical and health data. After quickly checking and analyzing some of the economic data, I grew interest in it as it is an area that is indirect to COVID but at the same time I'm very positive it got truly hit by it. I would like to know in depth how hard did the pandemic affected the economy inside Oklahoma county, how much relationship there is and if there were any actions taken by the county to counter any changes. To do this, I found two interesting indicators of economy, unemployment insurance claims and employment by industry sector that will be the basis of my analysis.

2. Background/Related Work

COVID has been a worldwide pandemic, but it has affected each community in an individual way, it depends largely on the response of the community to the infection and the culture. Hence, work done in one community cannot be assumed to work on any other community, this makes these study cases very focused to each community and the related work that we must find has to be specifically targeted to the same region and community as the one that we are interested in. While there haven't been many studies and related work specifically targeted to Oklahoma County, I have found that there have been several studies in the greater Oklahoma state. Given that Oklahoma County is a very important part of Oklahoma state, I think these studies are relevant as background and related work. I will specifically highlight 2 studies conducted in Oklahoma state, one is a study of the impact of local mask mandates upon COVID-19 case rates in Oklahoma and the other predicts COVID-19 cases using wastewater monitoring across Oklahoma City.

The first research takes advantage of the fact that Oklahoma did not impose a state-wide mask mandate, but multiple municipalities within the state did. The study compares daily case rates between the municipalities that enforced mask and the ones that didn't, at the same time and same state. They perform piecewise linear regression analysis to conclude that "Compared to rates in communities without mask mandates, transmission rates of SARS-CoV-2 slowed notably in those communities that adopted a mask mandate". The second study uses the fact that SARS-CoV-2 can be detected through human feces to create a wastewater surveillance to determine levels of infection and transmission and produce early warnings of outbreaks in local communities, independently from human testing. They can predict the number of cases with an accuracy of 81-92% compared to reported cases, but with the advantage of having these results 4 to 10- days before, allowing for preventive action.

These studies have a big impact on my work, the first study has a huge resemblance to part I, my study is just a subset of theirs, while I study the cases inside Oklahoma County, they study the cases for the whole Oklahoma state. For part II, III and IV, even though I will be focusing on employment and economics data, this research helps me understand that the pandemic can be analyzed in different sectors and fields, that they are all important and that resourcefulness and being able to correlate different important aspects of human life with the pandemic helps us prepare better for the future. It makes it clearer that there are many researchers focusing on COVID-19, meaning that it is a very important field to be working on. Lastly, in the first paper, we observe that they use the fact that Oklahoma state did not enforce mask mandate to study mask importance, this makes me note that Oklahoma County is a very important and special location to study and analyze, especially for the pandemic case where mask mandates were not imposed. Part I and these studies help me be very conscious and extra careful to take any insight from this research with a grain of salt given that it will be for a county that did not enforce any mask mandate throughout the whole pandemic, this is an important piece of information to highlight throughout this analysis where I will analyze the following research question and hypotheses:

Research question: What effect on employment did the pandemic generate inside Oklahoma county?

Hypothesis: I expect to see multiple effects between employment and the pandemic, specifically I consider:

1. The highest peak of unemployment occurred right at the beginning of the pandemic; this initial peak can be up to 30x the average rate of unemployment for the previous year before the pandemic.
2. The industry sectors that got the heaviest unemployment rates were leisure & hospitality as well as other heavy blue collar job industries.
3. Is working in Finance or Government sectors a safer bet when facing pandemic layoff?
4. After 1 year of pandemic, all sectors have been able to recuperate at least 80% of their layoffs.

3. Methodology

3.1 Methodology for Part I

To understand how “voluntary masking” looks like I did the following:

1. Check if there was a similar county in Oklahoma state that did impose a mask mandate, compare between these two counties to understand the difference between voluntary and forced masking, if counties within a state behave similarly

2. Search in the www for dates regarding masking recommendation/mandates and vaccination availability at a national, state and county level. I crossed these dates with our time series analysis to understand if voluntary masking or vaccination has a clear impact on cases, if so, what effect can we see?

I chose these two approaches given that they are what I have available with the scarce resources. Both methods do not use any private or personal information, does not require unethical experimentation and they are widely available data that will give us a better understanding of Oklahoma state and county COVID situation. Checking for these resources in the www will also allow us to encounter news and other useful resources that make our investigation more rich, as we will start to dive deeper and deeper into the lives of people that lived in Oklahoma County during the pandemic, allowing us to understand better what happened, understand better the individual stories of the people, ultimately allowing me to be more intelligent on what research question and hypothesis to focus on, as an example, I noted that Oklahoma County is a big Mining, Logging and Construction sector, hence, I decided to study deeper what happened with employment in different sectors such as this one.

3.2 Methodology for Part I II, III, and IV

I first cleaned the original dataset from part I. This includes transforming the cumulative data into daily confirmed COVID cases. Additionally, given that we know that this data has a weekly seasonality, I increase the decrease the granularity of this data to at least weekly and in some cases monthly. The *Employment by Industry Sector* data is in terms of monthly employees, which makes industry non-comparable, I transformed this to monthly growth (year-over-year), this served two purposes, it allowed me to compare between these different sectors and it removed many of the yearly seasonality that multiple sectors such as Government and Trade, Transportation & Utilities had. Finally, I merged the COVID weekly and monthly summed daily confirmed cases to the unemployment and employment by sector datasets respectively. All these data transformations and feature engineering techniques are performed to allow for a more thorough analysis, aggregating more data allows us to make data less personal/targeted, and merging employment data with COVID cases has no unethical implications to worry about.

After cleaning and processing the data, I decided to do multiple statistical analyses and visualizations to answer the research questions:

1. To analyze unemployment peaks, used a timeseries plot, allowing me to compare different peaks, and conclude if the highest occurred right at the start of the pandemic or if there was a lag, the lag was found by plotting this curve alongside the confirmed cases timeseries, I used conditionals to determine if there is a need of a lag or if these was no lag at all, this allowed me to know how much time it took unemployment petitions to peak given COVID. I additionally took the average of the previous year

unemployment, I compared this figure with the 3 highest unemployment peaks, to understand how many times unemployment petitions rose and what was the percentage change compared to pre-covid unemployment rates. Lastly, I created a box plot to compare unemployment growth between pre-pandemic, post-pandemic and new-normal stages.

2. To understand which were the industry sectors that were most heavily hit, I first visualize the difference using timeseries plots. I used the new feature (weekly growth year-over-year) to make it comparable and reduce yearly seasonality found in the data. I couple this with an additional boxplot that better makes sense of the whole post-pandemic period for all sectors, allowing to visually discriminate which were the sectors with greater drops in employment given the pandemic.
3. To check if Finance or Government sectors are safer bet when facing a pandemic layoff, I first did an ANOVA test to check if at least one of the industry sectors has a different growth (was hit more than the rest), then to go deeper I performed multiple mean t-tests corrected using Bonferroni correction, comparing the means of monthly growth between these two sectors and the rest during the pandemic, allowing me to conclude if these were hit less than the rest.
4. To find if after a year of pandemic, all sectors have been able to recuperate at least 80% of their layoffs, I first plot a line graph with the original employment by sector data, allowing me to visualize if there might be some sectors that struggled to recuperate. To be more thorough, I found for each sector which was the number of employees right before the start of the pandemic, in the worst moment (minimum number of employees post-pandemic) and then the last registered number of employees. This allows me to calculate the percentage recovered by dividing the number of employees recovered (last registered number – minimum number) with the number of lay-offs (pre-pandemic number – minimum number). I compare across sectors visually using a bar-graph.

During the whole process, I made multiple human-centered considerations to help me decide what to do. For example, given that I read an article that informed me that Oklahoma is known for their Mining, Logging and Construction sector, I decided to pursue the last hypothesis to check if this sector was able to recuperate at all from the pandemic, this is an example of the importance of thick data, where we are able to get better analysis done by reading and investigating at a more granular and personal level, compared to only analyzing the data directly. I made sure that none of these methods implied ethical concerns or used data that would jeopardize or put anyone at risk. Additionally, I have checked all the data sources I have used, and they have been taken voluntarily by humans in Oklahoma County. Additionally, this is very aggregated data (weekly and monthly), so it is hardly traceable to a single person.

4. Findings

4.1. Findings for part I

4.1.1 Qualitative findings using masking mandate information

I first attempt the first approach mentioned in methodology, we first observe that the whole state of Oklahoma did not enforce any type of masking mandate, hence, we do not have an ideal county to compare with. I move on to the second approach, we first find the following dates for mask mandates at a national level according to the U.S. Department of Defense[1]:

- First Mask Mandate: April 2020
- Mask Mandate is removed: January 2021
- 2nd Mask Mandate: 1 February 2021
- 2nd Mask Mandate is remove June 10, 2021
- Mask Mandate closed spaces is removed: February 25, 2022

I additionally found information from the OK county website regarding masking ordinance, according to their website:

- First Mask Ordinance: August 2020
- Mask ordinance removed: December 2020

To better determine any association, we want to make some change point detection first, I use Facebook open-source tool Prophet[2]. I obtain the following:

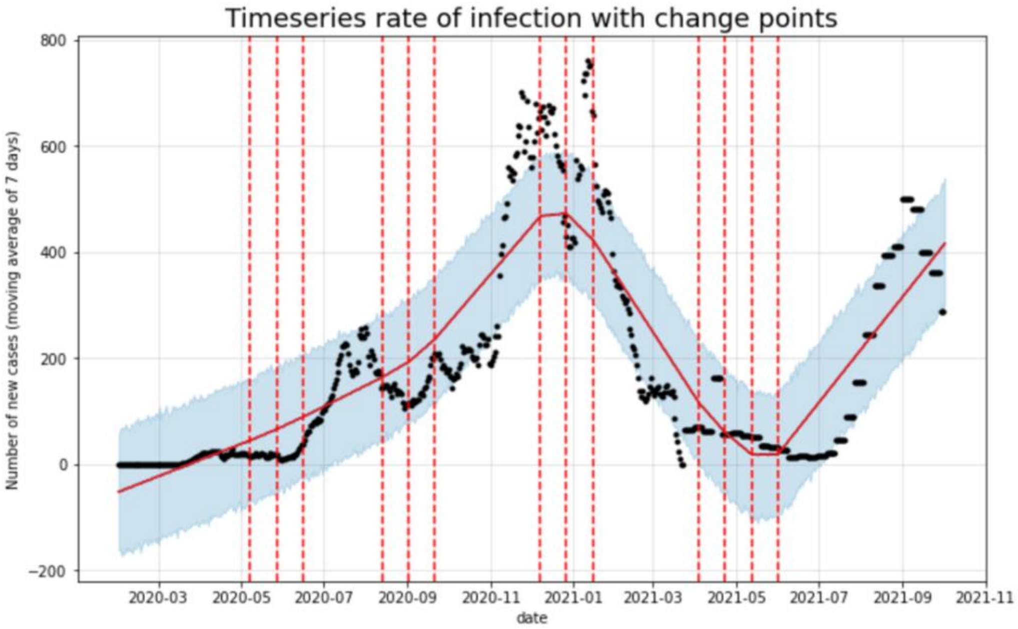


Figure 1: Timeseries rate of infection with change points

[1] The idea to use the CDC masking mandates was shared with me by Urmika

2 The idea to use FB Prophet was shared with me by Charles Reinertson

We can observe clear change points occurring during the following dates: - June 2020 - September, 2020 - January, 2021 - May, 2021

It is very clear to note that when we cross this with the information, we obtained from both the national and county levels, they match very well.

At a national level we have:

National mandate dates for reference: First Mask Mandate: April 2020 Mask Mandate is removed: January 2021 2nd Mask Mandate: 1 February 2021 2nd Mask Mandate is removed: June 10, 2021, Mask Mandate closed spaces is removed: February 25, 2022

1. We can see that when the first mask mandate was enforced (April 2020 to January, 2021), the rate of growth (derivate) of the curve is less than the one period where mask mandates were removed (June 10,2021 to November, 2021), ceteris paribus, we can see that national mask enforcement of face masks appears to have reduced the rate of spread.
2. The 2nd mask mandate (Feb 2021 - Jun, 2021) shows a clear decrease in the rate of new cases aligned with the change points in the graph.
3. When the 2nd mask mandate is removed (June 10,2021 to November 2021) a very clear spike arises.

Caveats: It is normal to observe a spike in cases during December given that: - In holidays people visit family and friends, increasing risk of infection - In holidays people fly to other places that require negative test proof, increasing the number of tests performed at this time of the year. We should ideally be able to normalize this effect with data of # of tests taken.

At the County level we have:

Oklahoma county mandate dates for reference: August 2020 Mask ordinance removed: December 2020

1. As mentioned before in point 1 in the national level, rate of growth of new cases is much slower that that after June 10, 2021.

This all gives us very good cause to think that people in Oklahoma followed the CDC recommendations to mask ("voluntarily masking") given that the dates where these recommendations were enforced, we observe improvements in number of new cases.

Caveat: We are assuming in this case that wearing masks truly diminishes the rate of spread.

Regarding vaccination, I obtain data from Bloomberg COVID vaccine tracker to find the coverage of vaccination in Oklahoma. We find the following:

- Jan 2021: First doses arrive
- Apr 2021: 50 doses per person
- Sept 2021: 100 doses per person
- Nov 2021: 120 doses per person

We can observe that there might be some association between vaccination and change point, from January 2021 onwards there is a very strong decrease in new cases until May 2021. At this point, most of the people that desired to get a vaccine was covered, and vaccination plateaued, so this might also explain why after this date we do not keep seeing a decrease or stabilization of the curve, instead an increase of new cases.

Caveat: This analysis is ceteris paribus, not considering masking or anything else, and if getting people vaccinated progressively would lower the infection rate curve.

4.1.2 Visualization to find how the course of the disease changes by masking policies

As stated before, we do not have any masking policies in place according to our masking policy data, hence we show the complete trend of rate of infection:

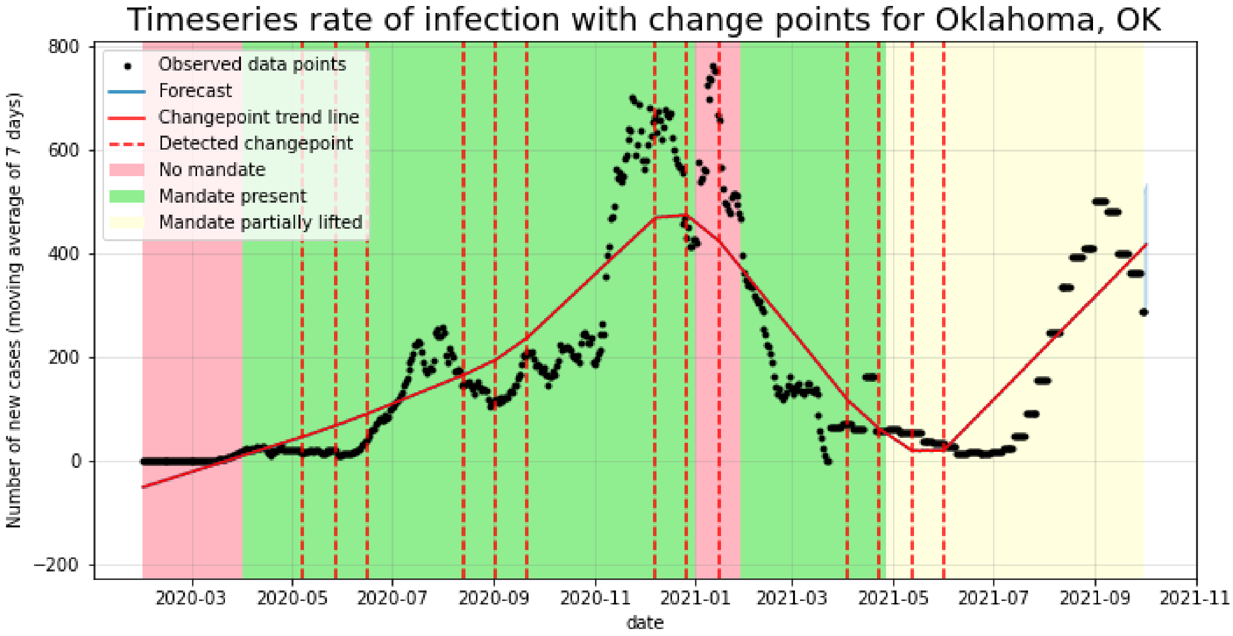


Figure 2: Timeseries rate of infection with change points for Oklahoma, OK.

This plot clearly includes the changes in the derivative function of the rate of infection through the change points and the red lines. The prophet API internally calculates the derivatives for these curves, finds the change points and plots them using the red lines. In the case of Oklahoma, OK, we observe that there are predominantly 4 change points where the rate of infection derivative function changes. If we had data regarding masking enforcement, I would've added this information to the graph to understand and visualize if there is any trend.

Given that we didn't, I used the CDC information I found before and added this to the plot to clearly visualize if there are any relationships between these CDC guidelines and Oklahoma, OK rate of infection.

The visualization was created using the Facebook Prophet library, it shows the timeseries rate of COVID infection from March 2020 to November 2021. This figure uses positive infection cases from the state of Oklahoma, OK that was originally in a cumulative form, this was later transformed to daily new cases feature and furthermore smoothed using a 7-window moving average, that help us remove weekly trends. This daily new cases data is shown in a timeseries to understand if there are any trends and moments where there are clear changes of rate of change. On the x-axis we observe the date in chronological order and in the y-axis, we have # of new cases. The black dots are the observed data points, the red line represents the changepoint trend line, the vertical intermittent red lines represent detected changepoints and then the red, yellow, and green in the background represents the no mandate, mandate partially lifted and mandate present status in each point of time respectively.

We use this plot to understand if there are any clear relationships between the trend changepoints and masking mandates. If there is, we would expect to observe that enforcing masking would affect the trend and we would view an identified changepoint. As an example, we can observe this occur in the plot between Jan 2021 and May 2021, where we observe that the start and end of having the mandate present is associated with the decreasing trend of new cases.

4.2. Findings for part II, III & IV

To answer the research question, I aim to perform the methods described in the previous part for each of the hypotheses I proposed in part 2. I will tackle each hypothesis independently and in the same order as follows:

4.2.1 **Hypothesis: #1:** The highest peak of unemployment occurred right at the beginning of the pandemic; this initial peak can be up to 30x the average rate of unemployment for the previous year before the pandemic.

4.2.1.1 Plot (visual) analysis

I first generate a timeseries visualization of the weekly unemployment insurance claims alongside the confirmed COVID cases:

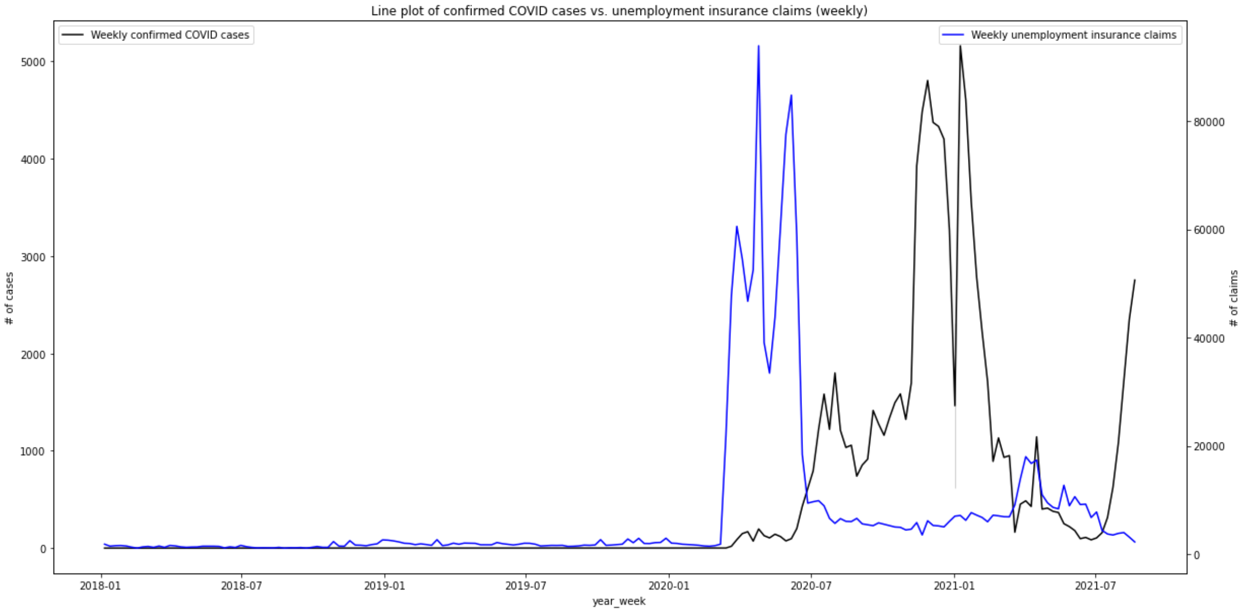
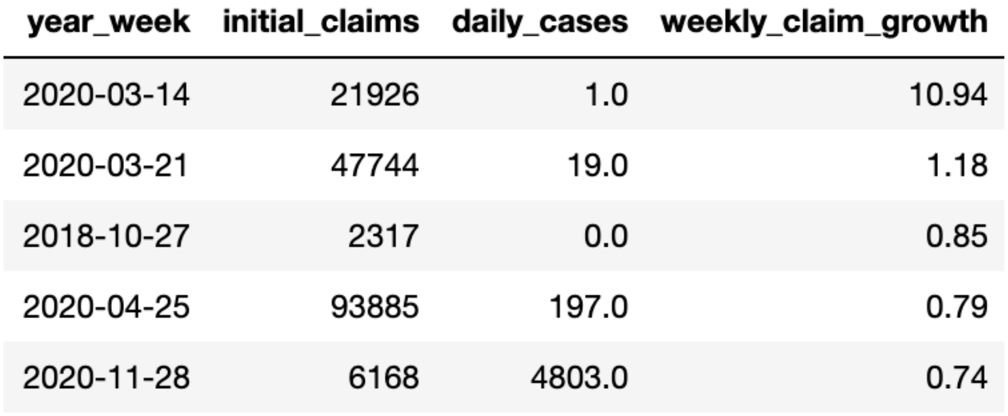


Figure 3: Line plot of confirmed COVID cases vs. unemployment insurance claims (weekly)

In the previous plot we can observe that it is true that the COVID pandemic highly affected the # of unemployment insurance claims. We can observe that in the moment that the pandemic starts, and we observe the first cases arise in the county there is a big jump in unemployment. We would like to know if there was a lag between the start of the pandemic and the start of unemployment claims. If so, by how much?

4.2.1.2 Lag between start of pandemic and unemployment claims

By using conditionals in Python, we prove that the week of the first aggressive peak change in number of unemployment claims is the same as the first week with conformed COVID cases in Oklahoma, OK. To make sure it was the week with highest growth change, we obtain the following table:



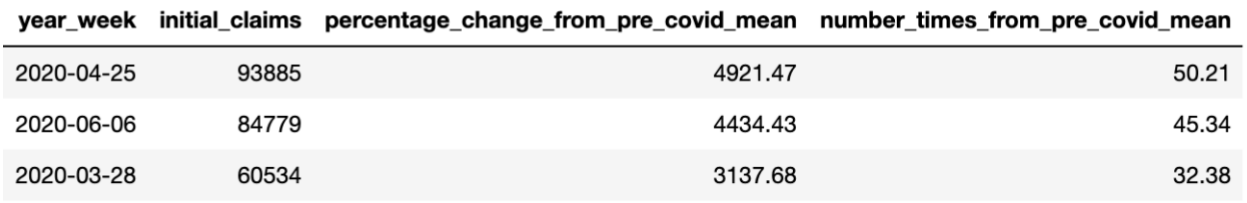
year_week	initial_claims	daily_cases	weekly_claim_growth
2020-03-14	21926	1.0	10.94
2020-03-21	47744	19.0	1.18
2018-10-27	2317	0.0	0.85
2020-04-25	93885	197.0	0.79
2020-11-28	6168	4803.0	0.74

Figure 4: Table with the top 5 days with highest growth change for initial unemployment claims.

We know that the first week with a confirmed COVID case in Oklahoma county was the 14th of March, 2020, hence we have discovered that the first week with confirmed cases of COVID is the same week with the highest change in unemployment claims, this answers the first question of the hypothesis. Now let's answer the second.

4.2.1.3 Percentage change between highest 3 unemployment peaks and the average unemployment claims

To now answer the second part of the hypothesis, we want to calculate the percentage change in unemployment for each of the 3 highest peaks with respect to pre-COVID 1-year average unemployment. I first calculate the average unemployment claim of 1 year (52 weeks) before COVID start week (2020-03-14) and find that it is 1869.67 claims. I calculate the number and dates of the 3 highest peaks of unemployment claims (2020-03-28, 2020-04-25 and 2020-06-06) and then calculate the percentage change for each of the peaks, obtaining the following table:



year_week	initial_claims	percentage_change_from_pre_covid_mean	number_times_from_pre_covid_mean
2020-04-25	93885	4921.47	50.21
2020-06-06	84779	4434.43	45.34
2020-03-28	60534	3137.68	32.38

Figure 5: Table with the percentage change in unemployment for each of the 3 highest peaks with respect to pre-COVID

We surprisingly find that our hypothesis is conservative, as we can observe, the highest peak was 50x the average rate of unemployment for the previous year before the pandemic. The next two peaks are at least 45 and 32 times the average rate.

4.2.1.4 Box plot of the YoY weekly growth of pre vs. post pandemic unemployment claims

We would like to go further and visualize using boxplots how different were the unemployment claims pre and post pandemic. I use the weekly growth (year-over-year) transformation in this case to reduce any seasonality effects as explained in the methodology sector and remove outliers to obtain the following plot:

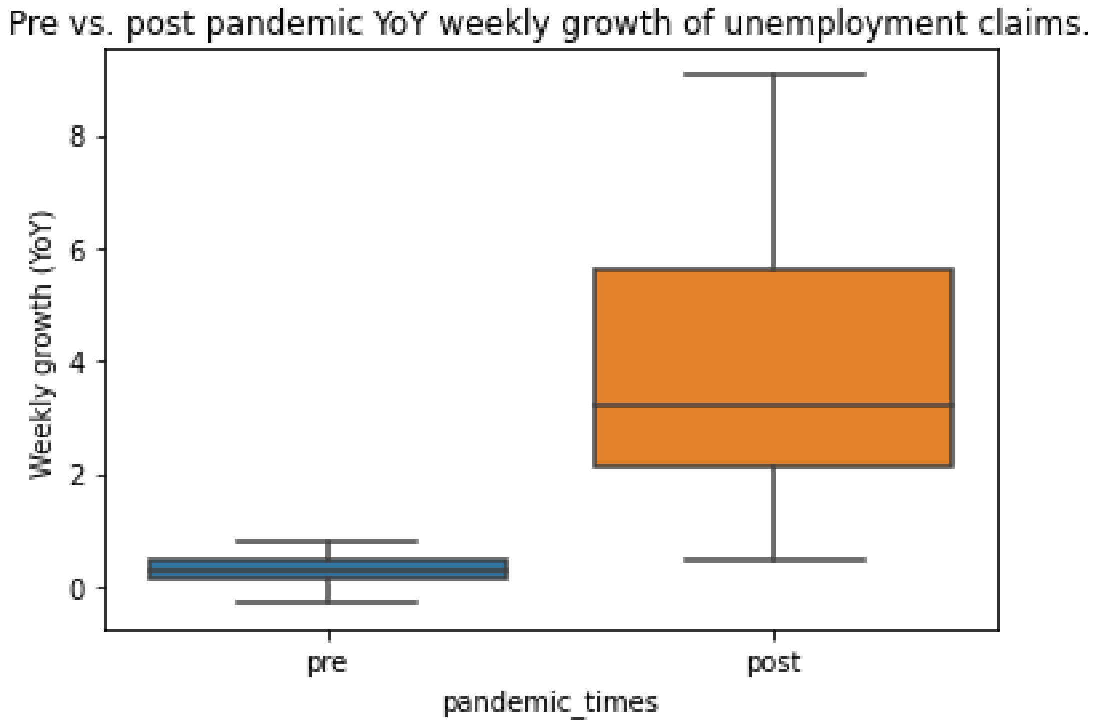


Figure 6: Box plot with pre vs. post pandemic YoY weekly growth of unemployment claims

We know that between the 14 of March and the 27 of June, there were big peaks with a lot of volatility, we want to observe the post-normality as well. Hence, I create a new box plot separating this and generating a new segment we name *new normal*:

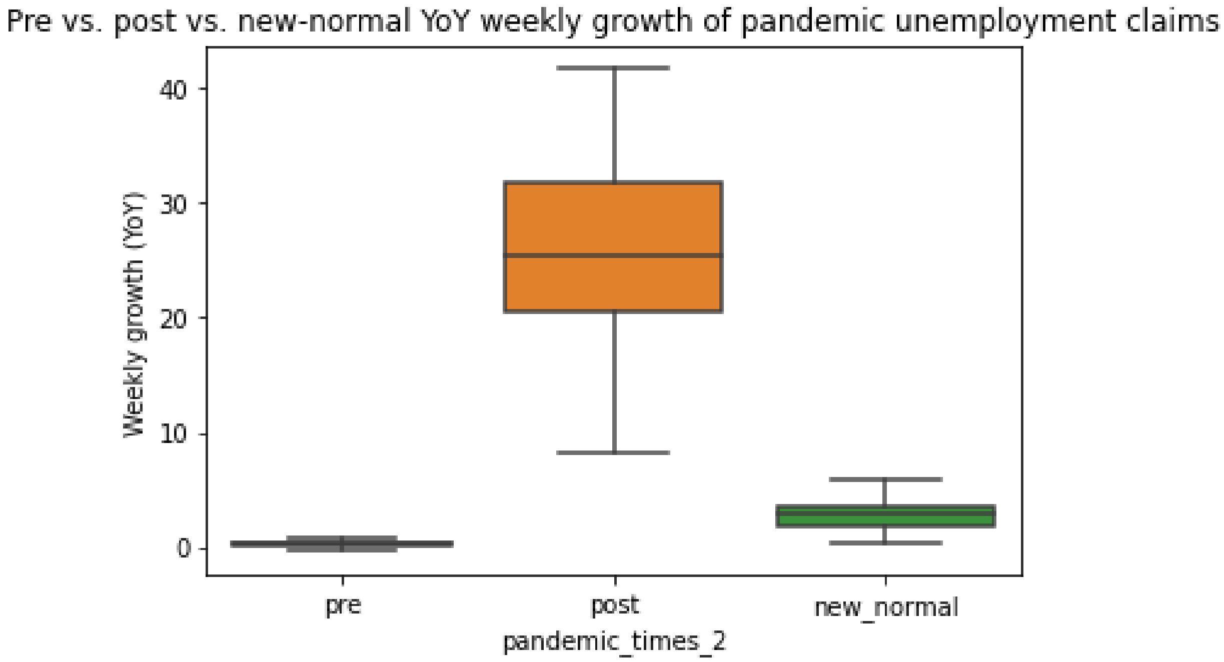


Figure 7: Pre vs. post vs. new-normal YoY weekly growth of pandemic unemployment claims

4.2.2 **Hypothesis #2:** The industry sectors that got the heaviest unemployment rates were leisure & hospitality as well as other heavy blue collar job industries.

4.2.2.1 Plot (visual) analysis

I first generate a timeseries visualization of the monthly employment YoY growth by sector alongside the confirmed COVID cases:

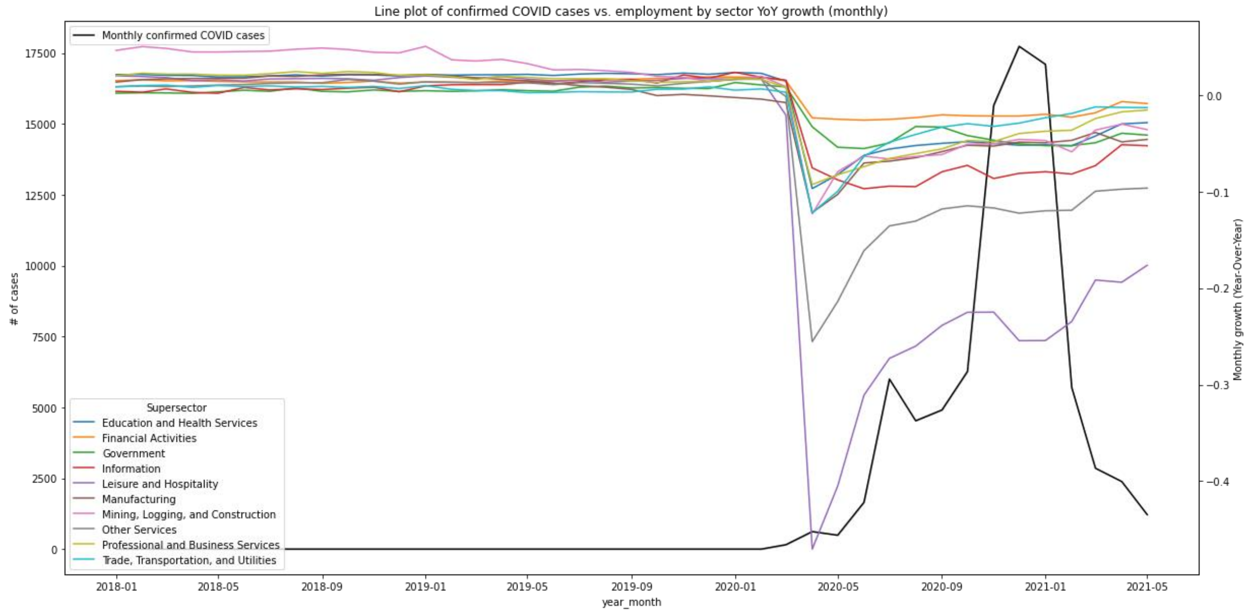


Figure 8: Line plot of confirmed COVID cases vs. employment by sector YoY growth (monthly)

We can observe that the industry sectors that got the heaviest unemployment rates were leisure & hospitality as well as other heavy blue collar job industries. We want to dive deeper so we create box-plots for post-pandemic Monthly growth (YoY)

4.2.2.2 Box-plot across industry sectors to have a better separation

A box plot of the post-pandemic portion of Figure 8 will allow us to have a 1-dimensional comparison across sectors:

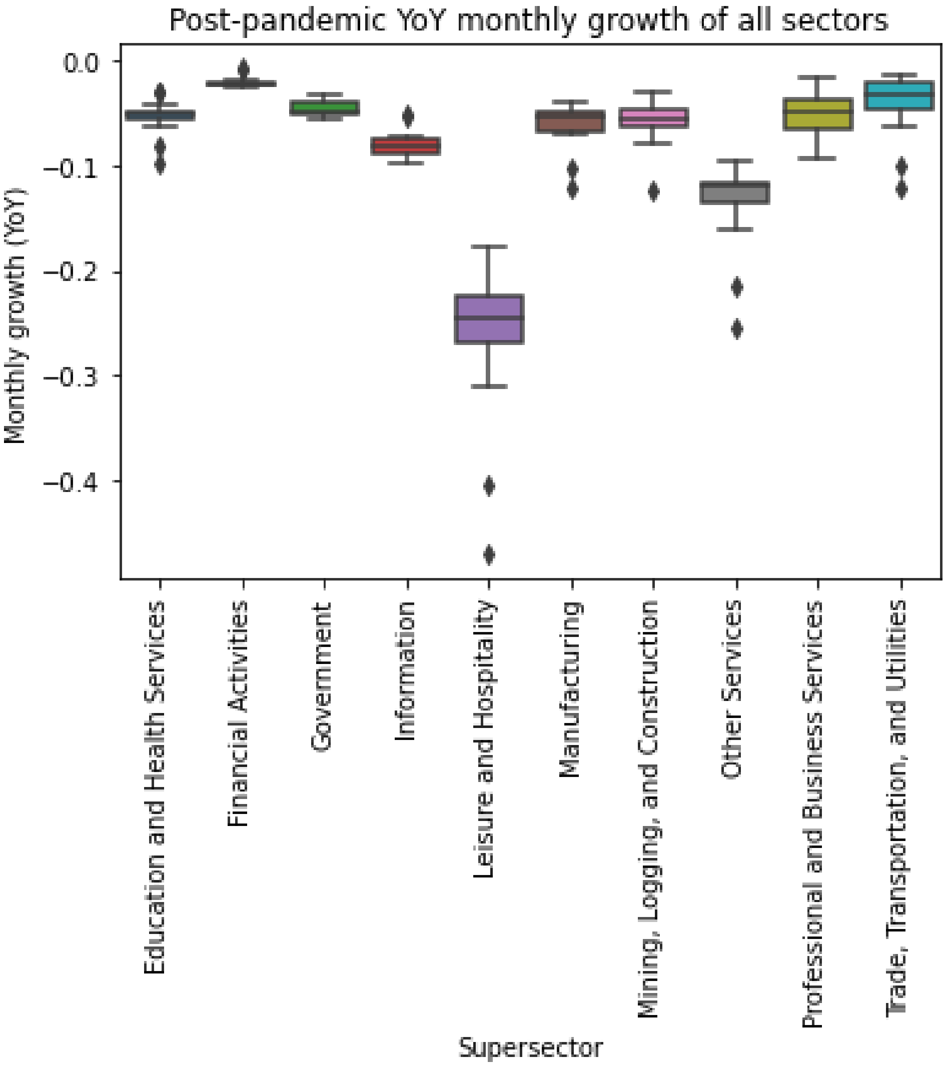


Figure 9: Box plot: Post-pandemic YoY monthly growth of all sectors

In the above plot, when COVID cases start to appear in Oklahoma, OK, all industry sectors suffer a very strong loss of employment. We can also visually observe that some sectors were more hardly hit than others. Sectors such as Leisure and Hospitality, Other Services, were specially hit while other sectors such as Financial Activity and Government were the ones that suffered less. We want to observe this better using box plots by sector.

We can conclude that the industry sectors with heaviest unemployment rates were Leisure & Hospitality and other blue-collar jobs such as "Other Services" and "Manufacturing". But what about the other side of the coin? Which are the sectors that were less hit and could be considered "safer"?

4.2.3 **Hypothesis #3:** Is working in Finance or Government sectors a safer bet when facing pandemic layoff?

4.2.3.1 One-way ANOVA multiple mean test (F-test) to check if there is statistical difference between sectors

We would like to first check if there is any difference between all the sectors, to do this I perform a simple one-way ANOVA multiple mean test (F-test) as follows:

$$H_0: \ \mu_{Education\ and\ Health\ Services} = \mu_{Financial\ Activities} = \mu_{Mining,Logging,and\ Construction}$$

$$= \ \mu_{Government} = \mu_{Information} = \ \mu_{Leisure\ and\ Hospitality} = \mu_{Manufacturing}$$

$$= \ \mu_{Other\ Services} = \mu_{Professional\ and\ Business\ Services} = \ \mu_{Trade,Transportation,and\ Utilities}$$

$$H_1: At\ least\ one\ is\ different$$

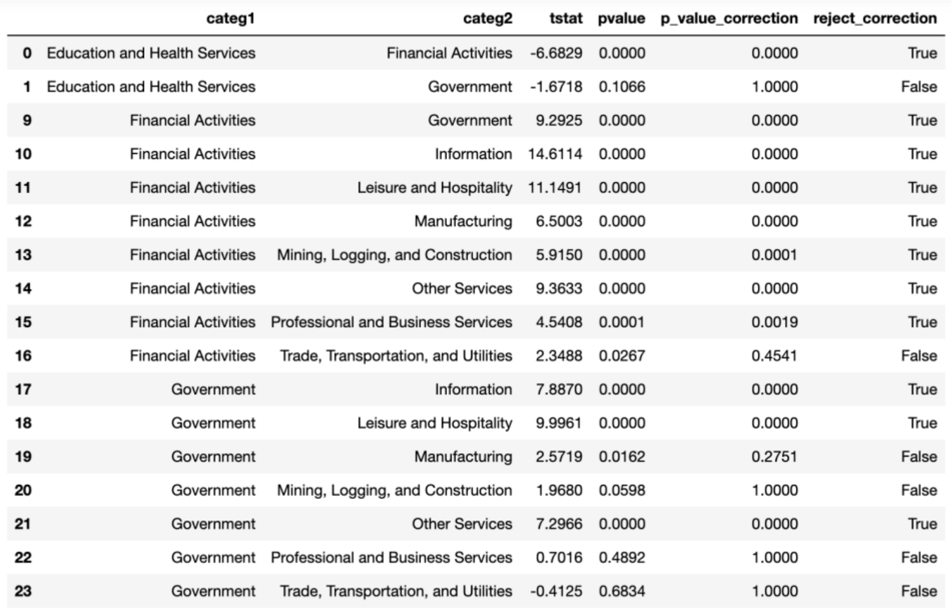
We obtain the following result:

```
F_onewayResult(statistic=58.95877089886192, pvalue=1.2964906766364846e-41)
```

With 99% confidence we reject the null hypothesis (all means are equal), hence, at least one of these groups have a different population mean.

4.2.3.2 Multiple t-test between Finance or Government sectors and the rest of sectors

We want to run a multiple t-test between Finance or Government sectors and the rest to understand if there are any pair differences, these will allow us to answer the hypothesis question. We obtain the following:



	categ1	categ2	tstat	pvalue	p_value_correction	reject_correction
0	Education and Health Services	Financial Activities	-6.6829	0.0000	0.0000	True
1	Education and Health Services	Government	-1.6718	0.1066	1.0000	False
9	Financial Activities	Government	9.2925	0.0000	0.0000	True
10	Financial Activities	Information	14.6114	0.0000	0.0000	True
11	Financial Activities	Leisure and Hospitality	11.1491	0.0000	0.0000	True
12	Financial Activities	Manufacturing	6.5003	0.0000	0.0000	True
13	Financial Activities	Mining, Logging, and Construction	5.9150	0.0000	0.0001	True
14	Financial Activities	Other Services	9.3633	0.0000	0.0000	True
15	Financial Activities	Professional and Business Services	4.5408	0.0001	0.0019	True
16	Financial Activities	Trade, Transportation, and Utilities	2.3488	0.0267	0.4541	False
17	Government	Information	7.8870	0.0000	0.0000	True
18	Government	Leisure and Hospitality	9.9961	0.0000	0.0000	True
19	Government	Manufacturing	2.5719	0.0162	0.2751	False
20	Government	Mining, Logging, and Construction	1.9680	0.0598	1.0000	False
21	Government	Other Services	7.2966	0.0000	0.0000	True
22	Government	Professional and Business Services	0.7016	0.4892	1.0000	False
23	Government	Trade, Transportation, and Utilities	-0.4125	0.6834	1.0000	False

Figure 10: Table with results of multiple paired t-tests for Financial Activities or Government vs. all other sectors.

As we can observe, we correct the p-value using Bonferroni correction to err on the safe side when concluding. We can observe that when working in Financial Activities, with 95% confidence we can conclude that the monthly employment growth YoY is different and better than the rest of sectors when facing the COVID-19 pandemic, except for Trade, Transportation and Utilities. On the other hand, Government is in the middle with 4 sectors that are statistically significantly different but 5 sectors where we cannot reject the null hypothesis. We can then conclude that Financial Activities shows a safer bet when facing pandemic layoff.

4.2.4 Hypothesis #4: After 1 year of pandemic, all sectors have been able to recuperate at least 80% of their layoffs

4.2.4.1 Plot (visual) analysis

We would first want to visually understand how the # of employees by sector behaves throughout the pandemic, the initial and the end of each line, to have an approximation of which could be the most and least resilient.

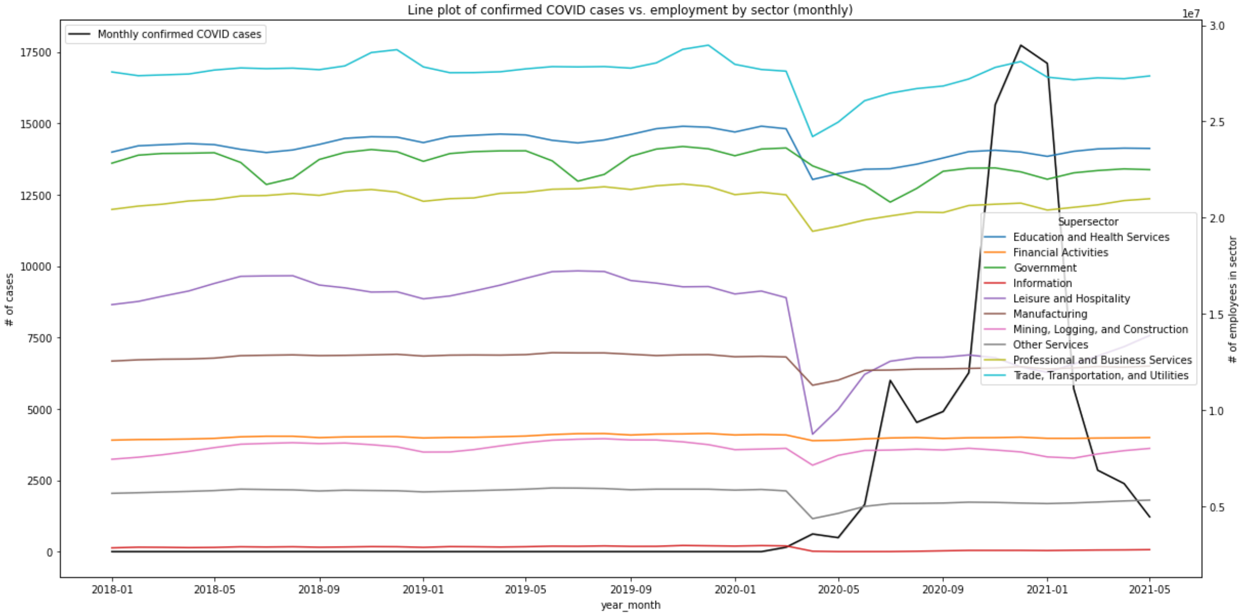


Figure 11: Line plot of confirmed COVID cases vs. employment by sector (monthly)

We can observe that many of the sectors seem to fail to recuperate from the layoff, especially Information seems to not recuperate at all. We want to get a better view of this, so we plot a bar-graph

4.2.4.2 post-pandemic percentage of recovered layoffs by sector with bar-graph

We first find for each sector the minimum number of employees after the pandemic; we then find for each sector the last number of employees we have in records lastly we find the number of employees that each sector had the month right before the pandemic hits (February). We calculate the layoff for each sector and the number of employees recovered after the layoff using the 3 previous numbers. Then we calculate the percentage recovered per sector and plot it in the following bar-graph:

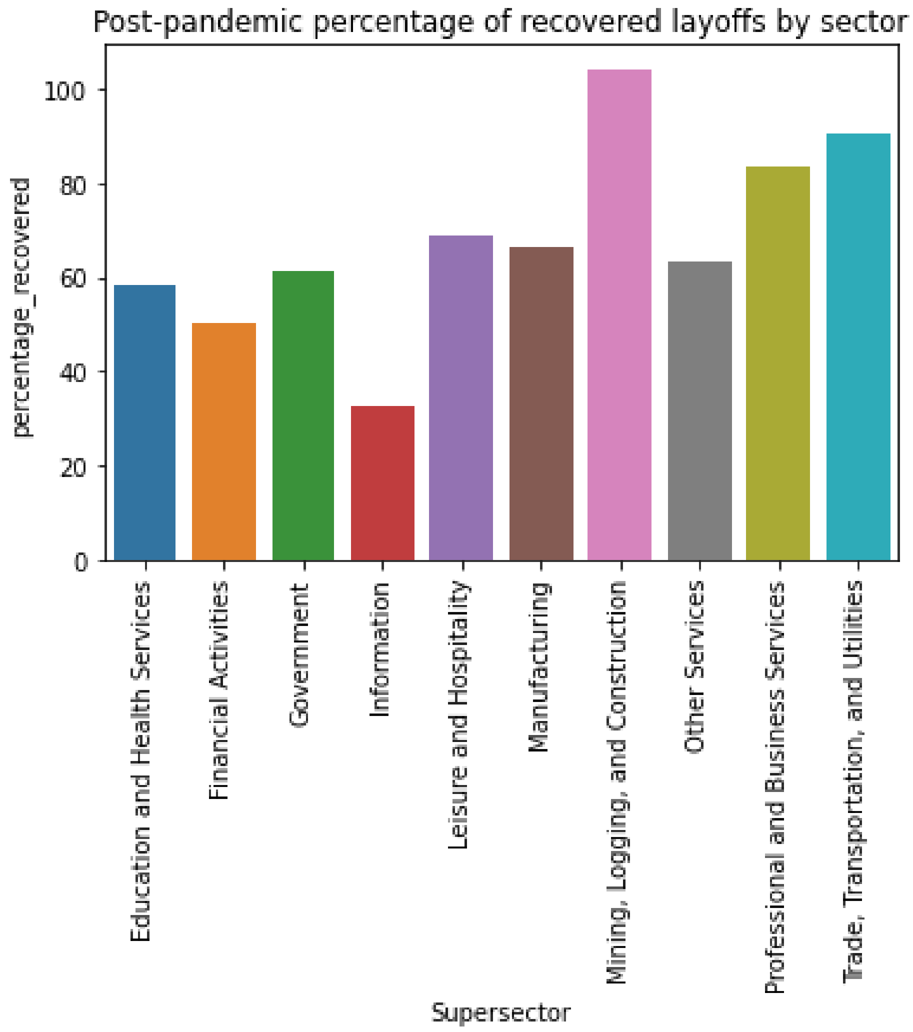


Figure 12: Bar plot of post-pandemic percentage of recovered layoffs by sector

We can notice that not all sectors have been able to recuperate equally. Sectors such as Mining, Logging and Construction were able to have even more jobs than before, while others such as Information haven't been able to recuperate more than 1/3 of what they lost. As we mentioned in the previous section, we got to know through local news that Mining, Logging and Construction is the most important sector in Oklahoma county, hence it is good news that it was able to recuperate.

5. Discussion/Implications

In the previous section, we have been able to respond to all the different hypotheses that we initially had using statistical and graphical methods. Our research question is highly important as it has to do with jobs, which is a very important part of human development and survival in modern society. Jobs are very important as they bring regular paychecks to families, it provides food and shelter and any other necessary goods. It also gives sense of identity, meaning and purpose, it rises intellectual challenge, access to community and socializing, it helps humans understand the world better and it brings health benefits.

Given all the previous characteristics that jobs give to humanity, it is essential to have a healthy economy that can provide enough job opportunities to anyone who needs one according to their area of expertise. When faced with a pandemic, we have proved through our findings that the job market plummets, it is harshly affected negatively for all sectors with big drops of opportunities, layoffs, and unavailability. This is a critical moment, as millions of people become job-less and will start to struggle to meet their financial debts and will lack many of the important qualities mentioned before.

Our research and findings allow us to be much better prepared for when a future pandemic strikes in Oklahoma county. We now know that Oklahoma county must be prepared to face a high peak of unemployment at the beginning of the next pandemic, that these peaks can be 50 times as high as normal unemployment rates. We have also discovered that Leisure & Hospitality and blue-collar jobs are the most affected sectors at the beginning, so Oklahoma county can begin to plan what mechanisms it can install to try to make these sectors stronger such that there is layoff prevention. We also have found that Finance is the only sector that doesn't get harshly hit like the rest, this means that it is a sector that Oklahoma county can be less worried about and won't require as much attention as the rest in the event of a pandemic. They can also do a deeper investigation as to why this sector is so strong and attempt to learn from this to apply to the other sectors. Lastly, we have found that after more than a year through the pandemic, not all sectors recuperate equally in Oklahoma county, we learn that the most important sector, namely, Mining, Logging and Construction is very resilient and can recuperate fully, while others such as Information will struggle much more. This allows Oklahoma county to prioritize which sectors to pay much more attention to after a pandemic happens to try as fast to bring their numbers back to normal.

All the hypothesis that I decided to explore in this research were thoughtfully built to give Oklahoma a much better insight around what is most important and critical around employment in the eventuality of a pandemic. It gives insight on which moments and sectors to focus on, given that there are more critical moments and sectors in a pandemic crisis. In the end, jobs are one of the most essential parts of a human being's life and being able to provide a county the opportunity to act more wisely and learn from experience where to focus will save and change lives, preventing harm, death, and destruction. Given this huge human-centered impact, comes with the same amount of responsibility, hence, we need to be very careful with our process, methodology and especially when generating actionable insights and conclusions that could be used in the next pandemic by politicians and advocates. We need to recognize biases and caveats in our analysis and be sure to comply with all the limitations and assumptions that our models have before we share with the world our findings, as much as they could make a lot of good, they could also cause a lot of harm if not done correctly, therefore we will further discuss this research's limitations in the next section.

6. Limitations

- During the analysis after section 4.2.1.3, I wanted to create a confidence interval of the YoY weekly growth of pre vs. post pandemic unemployment claims. I found that assuming normality, we have that the mean YoY weekly growth of unemployment claims has a confidence interval between 374 and 905 percent. But then I noticed that I should check for normality in the feature, so I ran a histogram and obtained the following plot:

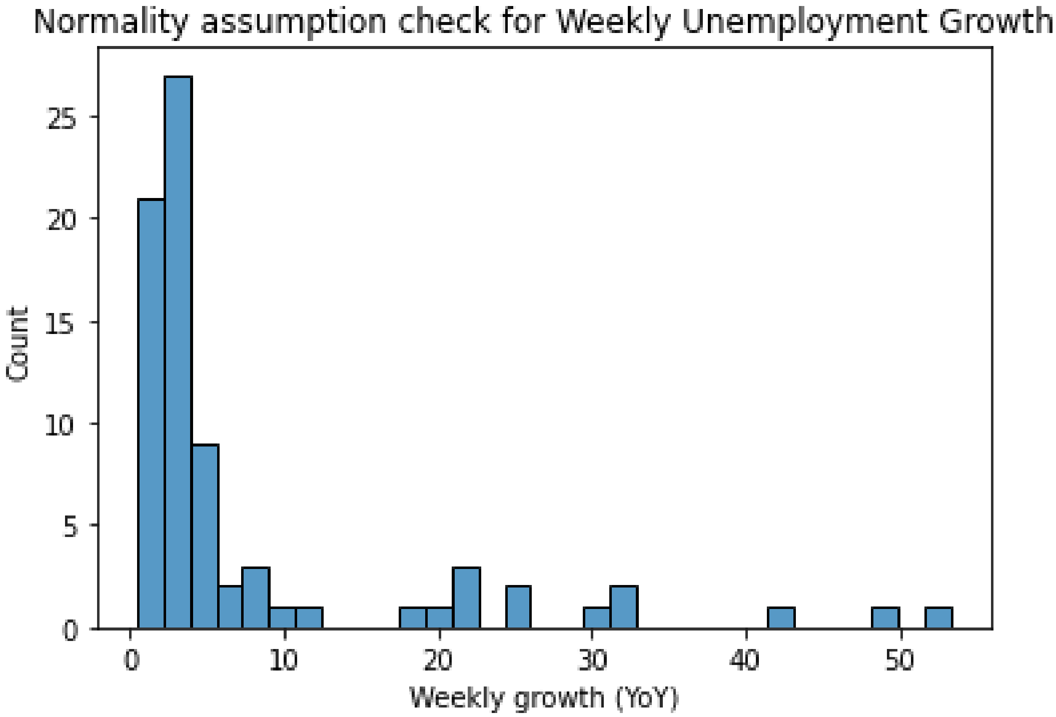


Figure 13: Normality assumption check for Weekly Unemployment Growth

- We can clearly observe that the feature does not comply the normality assumption, hence, I could not use the confidence interval. This whole process made me very conscious of the importance of checking assumptions in the data, if I didn't I could've very wrongly concluded insights that are untrue and that could be later been used for some decision making.
- Since the beginning we used unemployment insurance claims as a proxy for unemployment. While this makes sense, it is an assumption that our whole analysis relies on. There could be biases in this assumption given that not necessarily all people file claims when they are unemployed, many might be uninterested in doing so for a particular reason. Also, we are not contemplating unformal jobs in our analysis, hence, our analysis is heavily biased to the data that we were able to collect, but there is definitively a larger panorama and story to be told from many sectors that are informal and that we cannot view in unemployment insurance claims.
- COVID confirmed cases dataset has a big bias that has been known, we know that there is possibility of false positives and false negatives in this data.

Additionally, we should ideally be able to normalize this data by the number of tests taken in a day. In this whole research, we are assuming that testing is constant, and that people take tests at the same rate week-by-week, which isn't true, especially in holidays when we know people are keener to get tested to be able to fly. This means that some peaks in cases could also be associated with more tests being done in that week rather than more cases arising in the area, this is a big bias that we are not accounting for in our research.

- We are analyzing the pandemic through a very narrow lens by only using our economic data to describe and understand the number of confirmed COVID cases, assuming that this is the only important factor during the pandemic. There are many other that can affect the pandemic that we should consider such as vaccination rate, recovery rate, death rate, incubation period, among others. If we would like to make a much deeper analysis, we would have to begin to include these factors in, as they play a very important role and are confounder variables in our analysis.
- Some of the statistical tests performed in this analysis such as the F-test and t-tests are highly dependent on having a normal distribution. I used a Bonferroni correction that normally is highly conservative to make even more sure that any conclusions we draw are highly conservative, but it would be better to check if the features used for these statistical tests comply with the test's assumptions.
- In the last conclusion for hypothesis #4, we are assuming that the same people stay in Oklahoma county, that there are no transfers or people moving out or in, which is not necessarily true and a big assumption. As an example, we are saying that not even 1/3 of the layoff Information workers have been able to rejoin their sector, but this assumes that the lay-off people stayed in Oklahoma county and would only be willing to come back to Information sector.

7. Conclusion

In this research we have been able to answer the following research question and hypotheses:

Research question: What effect on employment did the pandemic generate inside Oklahoma county?

Hypothesis: I expect to see multiple effects between employment and the pandemic, specifically I consider:

1. The highest peak of unemployment occurred right at the beginning of the pandemic; this initial peak can be up to 30x the average rate of unemployment for the previous year before the pandemic.
2. The industry sectors that got the heaviest unemployment rates were leisure & hospitality as well as other heavy blue collar job industries.
3. Is working in Finance or Government sectors a safer bet when facing pandemic layoff?

4. After 1 year of pandemic, all sectors have been able to recuperate at least 80% of their layoffs.

We have found throughout the research that the COVID-19 pandemic did have a strongly negative effect on employment in Oklahoma country. We have noticed that the highest peak of unemployment did occur at the beginning of the pandemic, that the peaks can rise as high as 50 times the normal rate. Regarding sectors, we discovered that the heaviest hit was Leisure & Hospitality and other blue-collar jobs while Finance and Government were the least hit. We concluded that Finance is the least hit of them all, being a safe bet when being employed in this sector in Oklahoma county during a pandemic. Lastly, I was able to find that that after more than a year after the pandemic began, sectors such as Information haven't been able to recover as much as 1/3 of their original amount, while others such as Mining, Logging and Construction are fully recovered.

All these insights are very useful to face a new pandemic that might eventually happen anytime in the future. Oklahoma county government might use this information learnt from COVID-19 and apply it to a new pandemic, this can imply saving or changing thousands of lives in a positive way as learning from the past and taking defensive action and preparation normally yield stronger and more resilient communities, sectors, and people. That data scientists are taking time to analyze and understand this data will have a huge impact in scientific knowledge, governments and experts need to understand what the consequences of their actions were, which policies worked, and which didn't, they also need to understand which populations were most vulnerable and which they should have prioritized more attention to. Insights such as these will allow us to learn from our mistakes, grasp what are the top priorities in the event of a pandemic, and what institutions or policies we need to start building to prepare for a next one. Making this research comes with additional perks, it inspires other data scientists to work on these data sets, it can also drive them to work on top of my work to find additional insights.

8. References

Centers for Disease Control and Prevention. (n.d.). *U.S. state and territorial public mask mandates from April 10, 2020 through August 15, 2021 by County by day*. Centers for Disease Control and Prevention. Retrieved December 8, 2022, from https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i

Goldbloom, A. (2022, December 8). *Covid-19 data from John Hopkins University*. Kaggle. Retrieved December 8, 2022, from https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university

Kuhn, K. G., Jarshaw, J., Jeffries, E., Adesigbin, K., Maytubby, P., Dundas, N., Miller, A. C., Rhodes, E., Stevenson, B., Vogel, J., & Reeves, H. (2022). Predicting covid-19 cases in diverse population groups using SARS-COV-2 wastewater monitoring across Oklahoma City. *Science of The Total Environment*, *812*, 151431. https://doi.org/10.1016/j.scitotenv.2021.151431

The New York Times. (2020, April 1). *Oklahoma coronavirus map and case count*. The New York Times. Retrieved December 8, 2022, from https://www.nytimes.com/interactive/2021/us/oklahoma-covid-cases.html

Nytimes. (n.d.). *Covid-19-data/mask-use at master · Nytimes/covid-19-DATA*. GitHub. Retrieved December 8, 2022, from https://github.com/nytimes/covid-19-data/tree/master/mask-use

Oklahoma. The COVID Tracking Project. (n.d.). Retrieved December 8, 2022, from https://covidtracking.com/data/state/oklahoma

Staff, K. O. C. O. (2022, December 1). *Oklahoma reports more than 5,200 new COVID-19 cases, 17 additional deaths*. KOCO. Retrieved December 8, 2022, from https://www.koco.com/article/oklahoma-covid-19-numbers-december-1/42124598

Taylor, J. D., McCann, M. H., Richter, S. J., Matson, D., & Robert, J. (2022). Impact of local mask mandates upon covid-19 case rates in Oklahoma. *PLOS ONE*, *17*(6). https://doi.org/10.1371/journal.pone.0269339

9. Data Sources

For part I, not all data sources have full data, some contain nulls in multiple days.

The raw US confirmed cases file is sourced from a Kaggle repository of John Hopkins University COVID-19 data. This data is updated daily. You can use any revision of this dataset posted after October 1, 2022, has a Attribution 4.0 International (CC BY 4.0)

The CDC dataset of masking mandates by county. Note that the CDC stopped collecting this policy information in September 2021. Data Provided by Mara Howard-Williams, Public Health Law Program, Center for State, Tribal, Local, and Territorial Support, Centers for Disease Control and Prevention

The New York Times mask compliance survey data asks for the following attribution:

> This data is licensed under the same terms as our Coronavirus Data in the United States data. In general, we are making this data publicly available for broad, noncommercial public use including by medical and public health researchers, policymakers, analysts and local news media.
> If you use this data, you must attribute it to “The New York Times and Dynata” in any publication. If you would like a more expanded description of the data, you could say “Estimates from The New York Times, based on roughly 250,000 interviews conducted by Dynata from July 2 to July 14.”

> If you use it in an online presentation, we would appreciate it if you would link to our graphic discussing these results https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html.
> If you use this data, please let us know at covid-data@nytimes.com.
> See our LICENSE for the full terms of use for this data.

For part II,III and IV, I used data from the datausa.io web page, they offer extensive US public data that is conveniently at county granularity. Specifically for Oklahoma County, we will use two additional timeseries data:

- **Employment by Industry Sector:** Time series containing monthly employees per sector for Oklahoma county (not-seasonally adjusted) between December 2017 and April 2021.
- **Unemployment Insurance Claims:** Time series containing weekly unemployment insurance claims in Oklahoma county (not-seasonally adjusted) between January 2018 and September 2021.

The links to these data sets are the following:

- **Employment by Industry Sector:** https://api-ts-vibranium.datausa.io/tesseract/data.jsonrecords?cube=BLS Employment - Supersector Only&drilldowns=Month of Year,Supersector&measures=NSA Employees&State=04000US40
- **Unemployment Insurance Claims:** https://datausa.io/api/covid19/employment/latest/

In the following link: https://datausa.io/about/usage/, datausa.io specifies all licence/terms of use required to use this data, they allow access to this data for educational purposes, it is presented under a GNU Affero General Public Licence v3.0 (GPLv3). The content can be copied, downloaded for own use provided that suitable acknowledgment of Data USA as source is given.

These two time-series data sources summarize unemployment, given that it is at a county granularity, there is no personal information from any of the citizens. Given this, it is not sensible data that we would require to handle with greater care. We still need to be conscious that this data represents the suffering and pain of thousands of citizens that passed through harsh moments during the pandemic, as such, this data must be treated fairly and conclusions must be thoroughly inspected and proved as these could be eventually used as data-driven decisions in a next pandemic, affecting thousands to millions of lives.