Final Project Report

An Analysis of COVID-19's Impact on Unemployment

1. Introduction

2. Background/Related Work

3. Data

4. Method

5. Findings

6. Discussion/Implications

7. Limitations

8. Conclusion

9. References

10. Data Sources

1. Introduction

Unemployment is an important indicator of not only the economic trends of a country, or a part of the country, but also of pertinent world events. The lack of employment has empirically been related to world events, including human and natural. When times are good, the employment rate has been seen to increase. And one would assume that the pandemic induced recession could lead to an increase in unemployment. Through this analysis, we aim to validate this assumption using a scientific approach.

It would be interesting to see the consequences of this on the employment statistics of Harris County, Texas in the United States. The economic vulnerability could potentially be related to the COVID-19 cases. The socio economic impact makes this problem inherently human centered.

The goal of this analysis is to study the economic impacts of covid and also the subsequent unemployment insurance claims. We aim to decipher this relationship between this pandemic caused economic impact in Harris County in Texas. The covid pandemic caused an unprecedented disruption in the economy of the world. The recession that followed the pandemic resulted in steep job losses, and led to an increase in the unemployment rate to a high of 13.0 percent in the USA in the second quarter of 2020, and forced a big chunk of the population to leave the labor force.

2. Background/Related Work

According to an article published in the Texas tribune, Texas lost 1.4 million jobs in the spring of 2020 due to COVID-19 and the unemployment rate peaked at 12.9%in April 2020. The article also stated that the sales tax revenue dropped by almost $2 billion in one year. Another article on the Comprtoller.Texas.Gov states that a record 315,167 Texans filed initial jobless claims during the week ending April 4, 2020. This number is much higher than the number of people who filed the claims during the great recession in 2007-2008. By May 2021, there were strong signs of recovery — 1 million jobs and $1.6 billion in revenue were regained.

The research questions are inspired by these articles and other journals. The two broad questions answered in this project are:

- How was unemployment affected by the covid induced pandemic?
- What was the relationship between the number of covid cases and the unemployment insurance claims in Harris County?

The hypothesis are given below:

- An increase in the number of covid cases in Harris County, Texas led to an increase in the unemployment rate in the area.
- There is a correlation between the growth in the number of covid cases and the unemployment insurance claims in Harris County, Texas

3. Data

1. Unemployment data from FRED

The first dataset that I will use is from FRED, which is short for Federal Reserve Economic Data. It is an online database consisting of hundreds of thousands of economic data time series from scores of national, international, public, and private sources. The Research Department at the Federal Reserve Bank of St. Louis maintains these datasets and they go above and beyond just giving us the data. They aggregate the relevant data, and use a good mix of tools that the users can interact with. These tools are freely available, and help the users tell their story more effectively. This is a trusted source of economic data and has been around for a long time. The data has the unemployment rate from 1990 to 2022.The link for this dataset is https://fred.stlouisfed.org/series/UNRATE

The dataset has 393 rows and 2 columns. It has the unemployment rates from 1990 to 2022. This dataset was last updated on November 4th 2022. The units of unemployment is percentage and has been seasonally adjusted for unemployment based on the economic cycle. The granularity of the data is monthly, which is not the same as the granularity of the other datasets in the analysis.

The dataset gives us the unemployment rates in Houston, and we are considering Houston as a representative of Harris County. Here we use the unemployment data from Houston as a representative of Harris county because of the limited availability of the data from other cities.

Data usage terms and conditions- https://fred.stlouisfed.org/legal/#full-fred-terms

2. Unemployment Insurance Claims data from DOL

The second dataset has been taken from the United States Department Of Labor. The link to the dataset is: https://oui.doleta.gov/unemploy/. The U.S. Department of Labor is providing this information as a public service. These regulations and related materials are maintained on this website to enhance public access to information on Department of

Labor programs. This is a service that is continually under development. It has weekly insurance claims data. The raw data is provided in comma delimited files. The files are updated every morning with data available to the National Office database the previous day by 3:00PM Eastern Time.

This dataset comes under the Freedom of Information Act. Both the sources require us to use data only for non commercial purposes and cite them in our analysis.

4. Method

The research questions for this analysis have been answered using a combination of analytical methods. While choosing the analysis methods, we first needed to identify biases as well as societal and ethical challenges and concerns for effectively harnessing data, and then build a code for reproducible analysis.

The first step of the analysis is to standardize all the 3 datasets into monthly data by aggregating the base data and the weekly insurance claims. After the data had been standardized and the continuity of the time series was ensured, we calculated the Pearson correlation coefficient between the number of cases and the unemployment rate. And then we found the correlation coefficient between the unemployment rate and the insurance claims. This helped us validate whether our assumption holds true and whether there is actually a correlation relationship between the number of covid cases and the unemployment rate. After ensuring that the assumption is validated, we moved on to build a regression model taking the unemployment rate as the dependent variable and the number of covid cases as the independent variable. Before running the regression, we checked if the assumptions of linear regression hold on the data.

The obvious next step here was to test the null hypothesis that the regression coefficient is not significant. The beta coefficient tells us the unit change in the outcome variable for every 1 unit if change in the independent variable. The t-test assesses whether the beta coefficient is significantly different from zero. When we conduct the regression, we also compute the R squared statistic, which is coefficient of determination. It can be thought of as the fraction of variance in the outcome variable that is explained by the independent variable.

Null Hypothesis : The regression coefficient is not significantly different from zero.
Alternative Hypothesis : The regression coefficient is significantly different from zero.

We considered a confidence interval of 95%, so If the p value is less than 0.05 we can infer that the coefficient is statistically significant and we can use the model for prediction and inference.

5. Findings

We first plot and see what the masking mandates look like over time. And then we try to map the effect of those changes in the change in the number of covid cases.

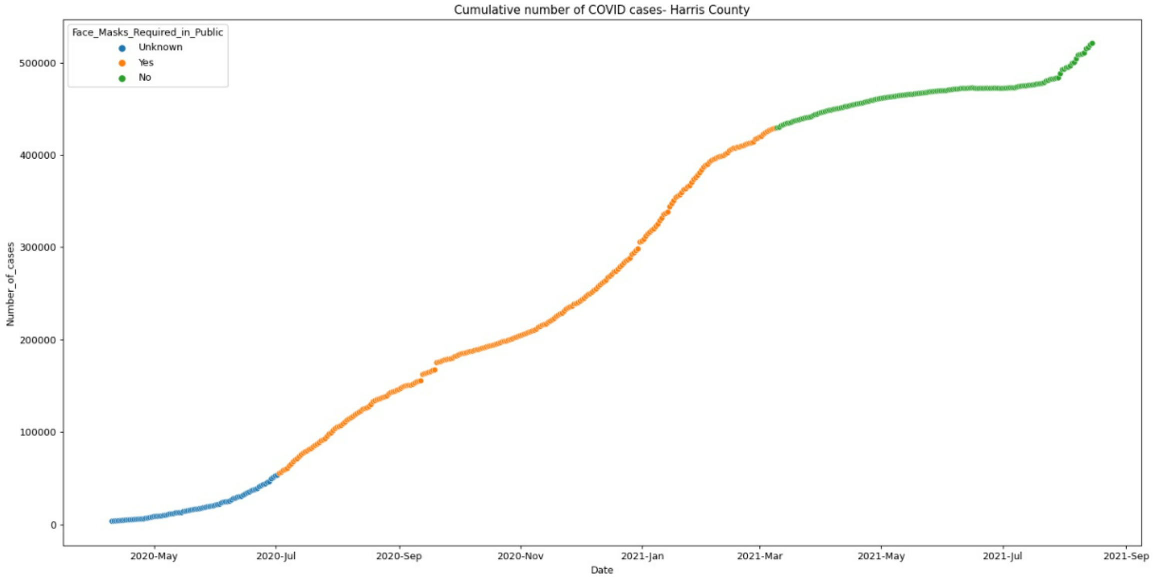


Figure 1a - Cumulative number of covid cases- Harris County

The line graph shows the number of COVID cases daily, and the color of the line depicts the masking mandate at that point in time. For the times when we do not know the mandate, we color the line blue, orange for when there was a masking mandate in place, and green when there was no mandate in place.

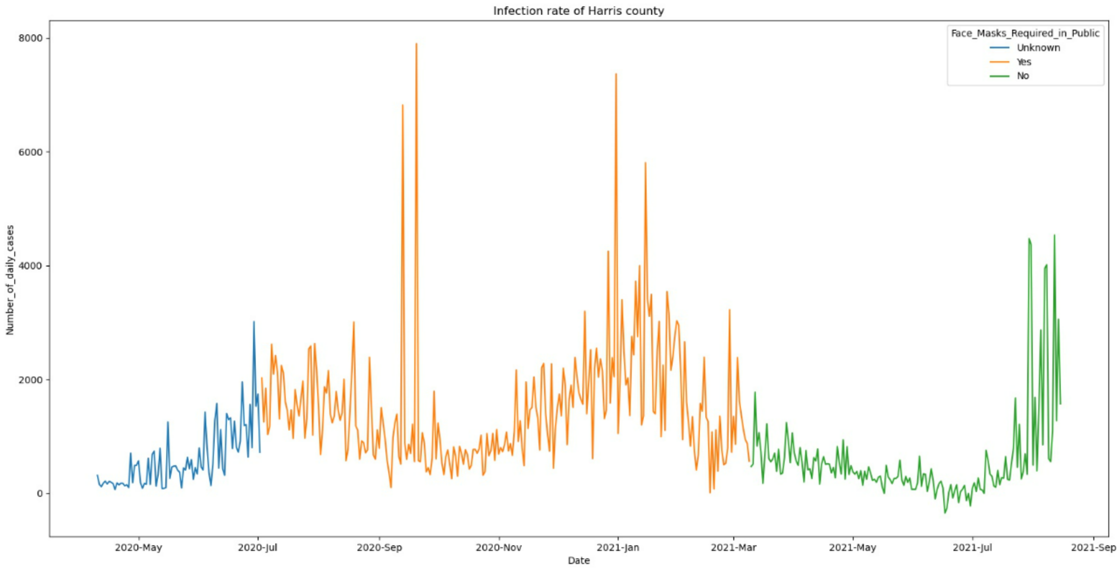


Figure 1b - Daily covid cases- Harris County

When the mandate was put in place, there was a dip in the number of cases, and we also see a rupture point at the place near the dip in the graph. This validates our assumption that there are decreases in the number of covid cases when people start wearing masks.In addition to this one, there are more rupture points, in the graph showing changes in the infection rate. The population of the county is assumed to be static over the time period considered for analysis, so the trend in the visualization is not affected whether we take into account the infection rate or the number of COVID cases.

We then see the time series trend in the number of active covid cases, unemployment rate and the Unemployment insurance claims on a single graph to do a visual inspection of the trends. The primary axis is the number of covid cases and the average insurance claims, and the secondary axis is the unemployment rate.

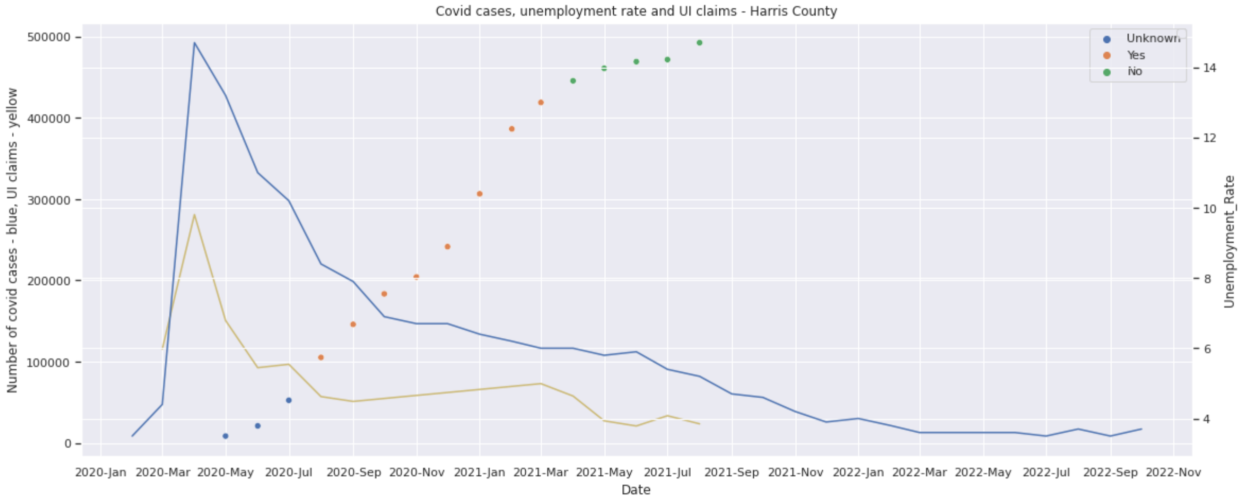


Figure 2 - Covid cases, unemployment rate and UI claims - Harris County

We see that as the cumulative active cases rise, the unemployment rate peaks, and then gradually decreases. This can be attributed to the vaccination becoming available in Harris County, Texas in August 2021. We also see that the unemployment rate also varies significantly within different sectors. The leisure and hospitality industry saw the biggest hit in employment.

Next, we see the correlation between the number of cases and the unemployment rate. There is a weak but negative correlation. This can be attributed to the fact that as the number of cases increases, there are remedial policies brought about to tackle that problem. Correlation Coefficient between the number of covid cases and unemployment rate is -0.221.

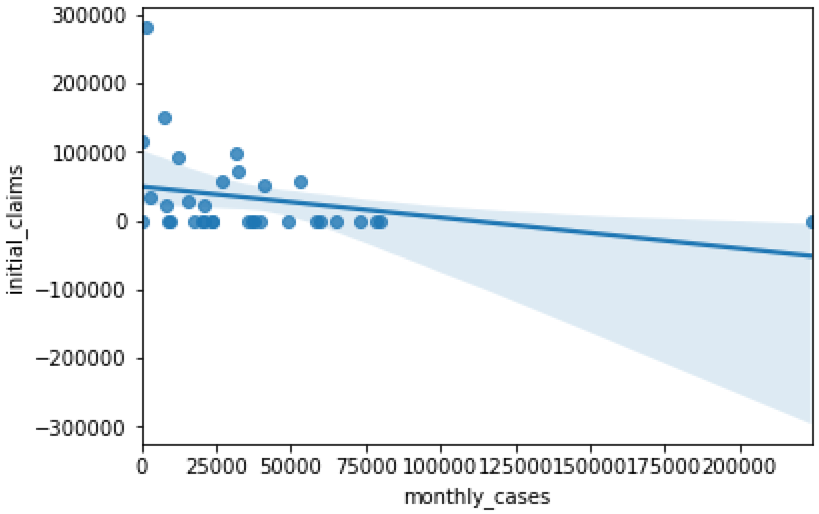


Figure 3 - Correlation between number of cases and UI claims

Similarly, there is a weak negative correlation between the number of covid cases and the unemployment rate. Correlation Coefficient between the number of covid cases and unemployment rate is -0.306

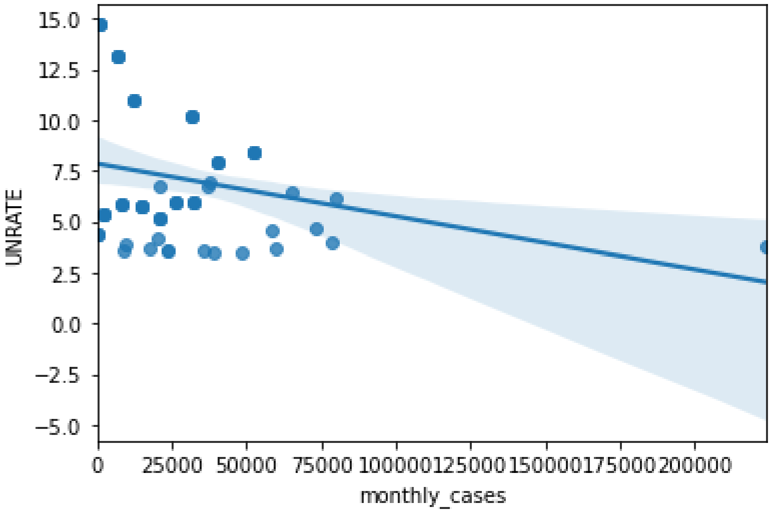
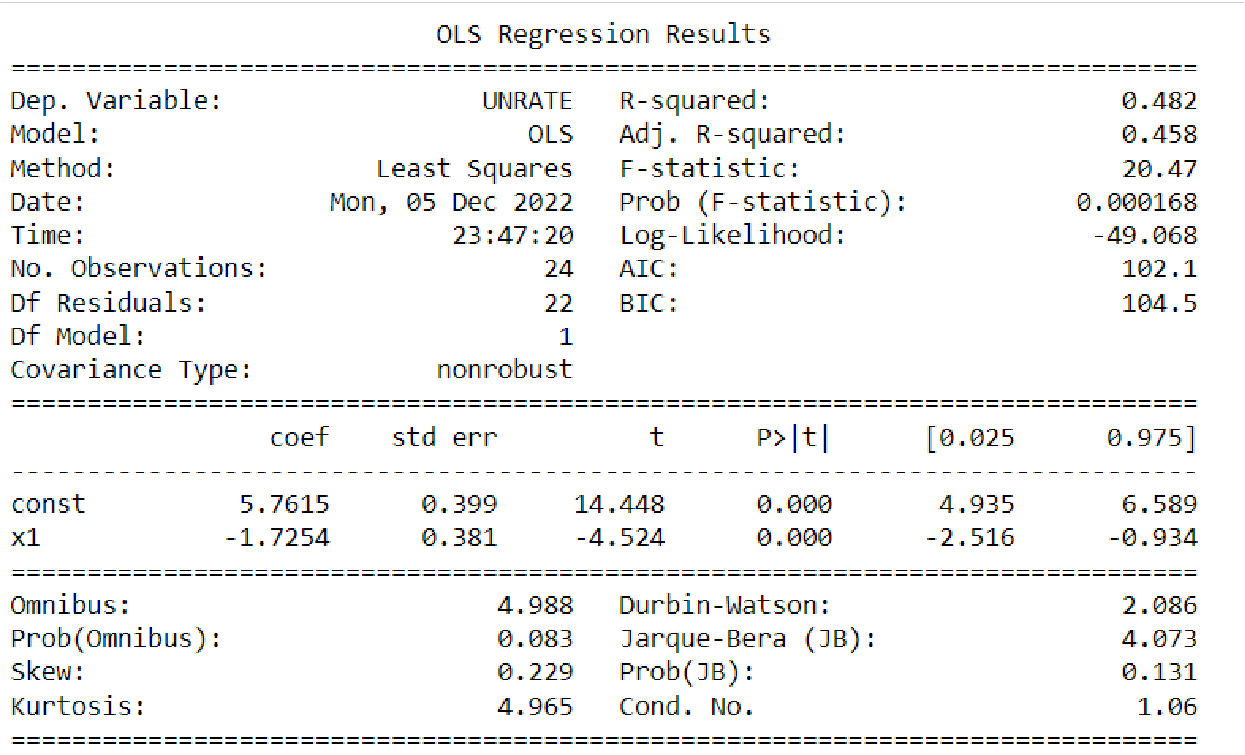


Figure 4 - Correlation between number of cases and unemployment rate

The next step in the analysis is regression. The predictor variable is the number of active covid cases and the target variable is unemployment rate.

Our model explains only 50% of the variance of the target with the predictor. And this is expected. Because in the real world, our independence assumption does not hold and this is a very hard prediction to make.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 UNRATE   R-squared:                       0.482
Model:                            OLS   Adj. R-squared:                  0.458
Method:                 Least Squares   F-statistic:                     20.47
Date:                Mon, 05 Dec 2022   Prob (F-statistic):           0.000168
Time:                        23:47:20   Log-Likelihood:                -49.068
No. Observations:                  24   AIC:                             102.1
Df Residuals:                      22   BIC:                             104.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.7615      0.399     14.448      0.000       4.935       6.589
x1            -1.7254      0.381     -4.524      0.000      -2.516      -0.934
==============================================================================
Omnibus:                        4.988   Durbin-Watson:                   2.086
Prob(Omnibus):                  0.083   Jarque-Bera (JB):                4.073
Skew:                           0.229   Prob(JB):                        0.131
Kurtosis:                       4.965   Cond. No.                         1.06
==============================================================================
```

Figure 5 - OLS Regression Results

The regression coefficient is statistically significant. However, the model performance is not great. The R squared on the training data 0.49 and on test data 0.52. We cannot predict unemployment accurately using just the number of covid cases in Harris county. This is because in the real world, the independence assumption does not hold, and the other confounding factors play a role!

6. Discussion/Implications

We see that the number of active cases in Texas follows the same trend visually as the total number of active cases in the USA. So this analysis is interesting and we can possibly extend it to the county's statistics. We have used both descriptive analysis and inferential statistics to make sense of the data, while keeping in mind that there could have been biases that could have crept in while data collection. So we actively eliminate methodological biases.

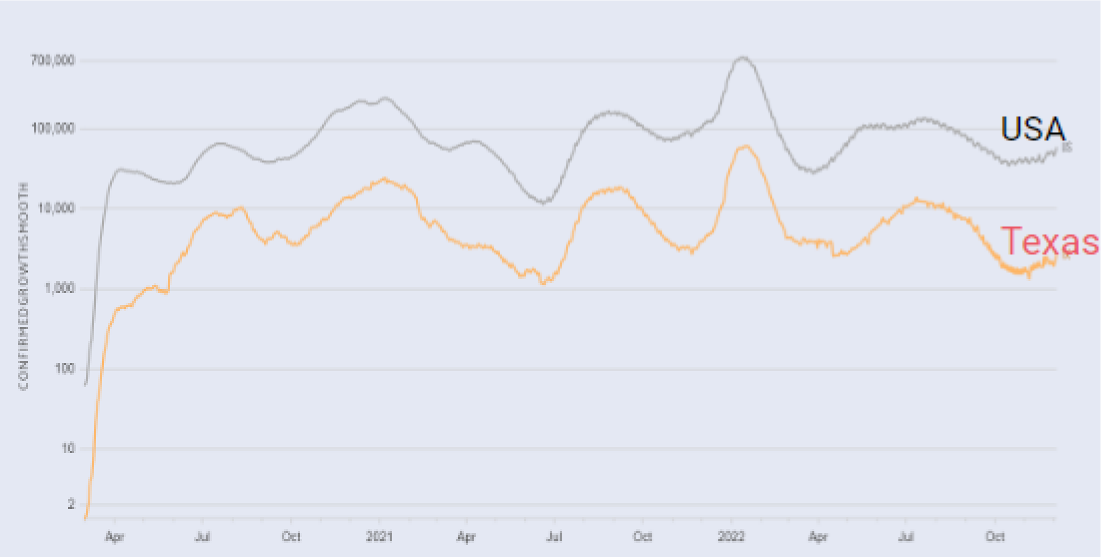


Figure 5 - Number of confirmed covid cases in USA vs Texas

We are all familiar with the challenges that the pandemic brought about – unemployment being one of the major ones.There were many factors that affect unemployment – masking policies, vaccination requirements and availability.

The covid induced pandemic also caused a multitude of socio economic consequences, and it completely changed life for a section of the population. The two statistics analyzed here, unemployment and UI claims, are indicative of changes in the quality of people because of the pandemic, making this problem inherently human centric.

7. Limitations

The datasets chosen for the analysis are not available for the County and this limited data availability makes us use the information of the metropolitan city in the county or information about the state and consider that as a representative of the information about the county. The metropolitan city in Harris County for which data is available is Houston, so we take the unemployment data of Houston, and the insurance claims data of Texas.

The accuracy and authenticity of the data is beyond our control, and so is the granularity of the data. The frequency of unemployment rate is weekly and the frequency of the claims information is monthly. The frequency of the base dataset from part 1 is daily. So there is a need to standardize these datasets.

Another important point to keep in mind is that unemployment could be impacted by other external phenomena too. However, for the purpose of this analysis we consider that in the time period in consideration here, other factors remain constant and the relationship we observe is solely due to the pandemic.

While these insurance claims are on the important factors of the secondary impacts of covid, there could be multiple other transitive impacts that could make this analysis more interesting and complete. Another interesting social impact of the pandemic induced recession could be on the mental health of the people who got unemployed as a result of covid.

In addition to these unknowns and dependencies, we need to ensure that the assumptions of the statistical methods used are satisfied by the data. Here, we check the assumptions of linear regression - linear relationship, homoscedasticity, normality of residuals,

There are no explicit ethical considerations in using the datasets being used here. The insurance claims data may not be a good representative of the socio-economic impacts of the pandemic induced recession because the whole population is not aware of their rights and the fact that they can claim unemployment benefits. So this analysis can potentially give us interesting insights into the gap between the two phenomena. The datasets are relevant to the research questions and helped in effectively building on Common analysis and expanding the inferences from the Common analysis.

8. Conclusion

As the cumulative active cases rise, the unemployment rate peaks, and then gradually decreases. This can be attributed to the vaccination becoming available in Harris County, Texas in August 2021. It also varies significantly within different sectors. The leisure and hospitality industry saw the biggest hit in employment.

The regression model explained in Figure 5 reinstates the fact that it is very hard to predict unemployment from the number of covid cases in isolation. It is so because the socio economic happenings in the world are not independent of each other, and we cannot model them quantitatively without considering the qualitative aspect of it. This problem that we are trying to address here is at the core of human centered data science, and we need to actively consider biases and other challenges while solving it quantitatively.

The results of the analysis are relevant because they help us with the following:

- Understanding the secondary impact of the pandemic
- Better equipped to minimize the impact in future.
- Uncover the economic vulnerability related to a pandemic.
- Understand better the socio economic impact of the problem while keeping humans at the center of the analysis.

9. References

Below is a list of resources referenced during the course of this project

https://apps.texastribune.org/features/2020/texas-unemployment/
https://comptroller.texas.gov/economy/fiscal-notes/2020/may/unemployment.php

10. Data Sources

- Unemployment data from FRED : https://fred.stlouisfed.org/series/UNRATE
- UI Insurance Claims data from DOL : https://oui.doleta.gov/unemploy/