Final Project: Modeling the Impact of COVID-19 on Unemployment and Financial Markets

Priyanka Gunduboina, Iman Ismail, Jaya Johnson, Robi Rahman

Project Summary and Research Motivation

In this project, we investigate the economic effects of the COVID-19 pandemic. Historically, most harmful economic events have caused simultaneous downturns in employment and stock markets, as people lose their jobs and businesses lose revenues. However, although the spread of the coronavirus has cost millions of people their jobs and caused an initial hit to the stock market, financial indices quickly recovered while employment is slow to return. Therefore, we are interested to find out whether these different effects can be explained or predicted using statistical modeling.

Our research is broken down into four sections, where each of us investigated a different aspect of COVID-19's effects on employment, financial markets, or the global economy. We have modeled unemployment rate as a multilinear function of different contributing factors across different American demographic groups, analysed the impact of COVID cases and fatalities on state-level unemployment rates, analyzed whether or not COVID cases have had an impact on national stock indices in different countries, and analyzed the disparate effects of COVID on the unemployment in different industries, as well as whether virtual operations have allowed them to escape the harmful effects of the pandemic.

Data for this project was obtained from the Bureau of Labor Statistics, Centers for Disease Control COVID tracker, 2020 United States Census, and other sources. Data cleaning, analysis, and visualization were done using R.

Section 1: COVID, Demographics, and Unemployment

Jaya Johnson

Research Question and Motivation

In this section we analyze the impact of various factors that may or may not contribute to the overall unemployment rate during the pandemic. As a result of the pandemic, there were many job losses and operations across different industries like entertainment, transportation were adversely affected. Our motivation is to see if COVID cases have a direct effect on the unemployment rate.

We consider the overall impact of variables like unemployment rate within different demographics, number of hours worked across different industries, job loss percentage and the number COVID cases in the US. Our dataset spans 30 months from October 2018 to March 2021. The datasets were obtained from the Bureau of Labor Statistics[1].

Hypothesis

Our hypothesis is that different factors like population demographics (e.g. race, gender), average number of hours worked across different industries, job loss percentage and the number of COVID cases in the US are correlated to unemployment, and can be used to model the unemployment rate. We then would like to predict the unemployment rate in the US for the month of April based on the regression model.

The following variables were taken into consideration to build the model.

Unemployment Rate White Americans	Average number of hrs - Transportation and Warehousing
Unemployment Rate Asian Americans	Average number of hrs - Financial Sector
Unemployment Rate Black Americans	Average number of hrs - Health and Education Sector
Unemployment Rate Hispanic Americans	Average number of hrs - Leisure and Hospitality Sector
Unemployment Rate Men	Average number of hrs - Private Sector
Unemployment Rate Women	Married over 16 (with spouse)
COVID Cases US	Job Losers (Percent)

[1] https://www.bls.gov/cps/effects-of-the-coronavirus-covid-19-pandemic.htm#concepts

Data Exploration and Visualization

Below we have some plots that show information about the distribution of data and the trends.

In *FIGURE 1* below, we see that most of the data variables are skewed left, indicating that the spike in unemployment rates in the different demographics was a result of the initial impact of the pandemic. Weekly hours in the hospitality, leisure and transportation industries also show left-skewed distributions; this could be an artifact of the numbers prior to the pandemic.

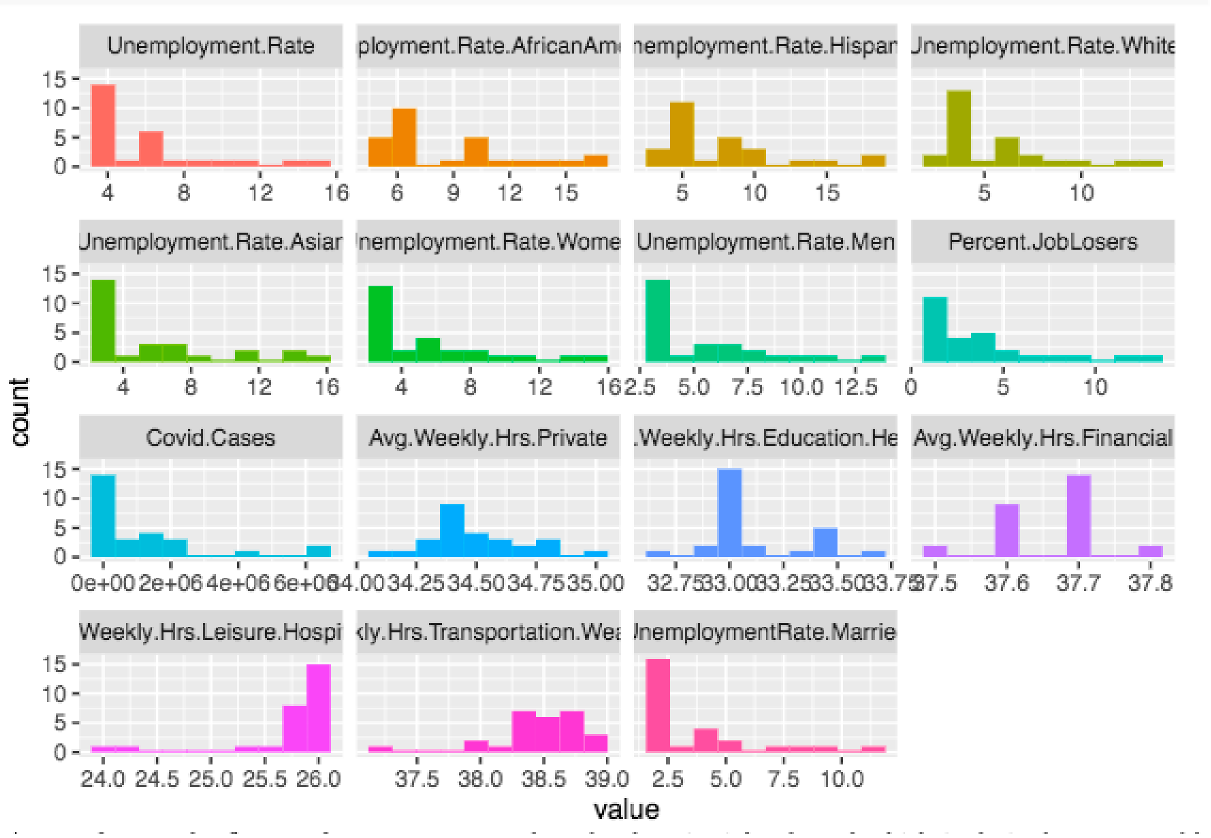


FIGURE 1. Histograms of different predictor variables.

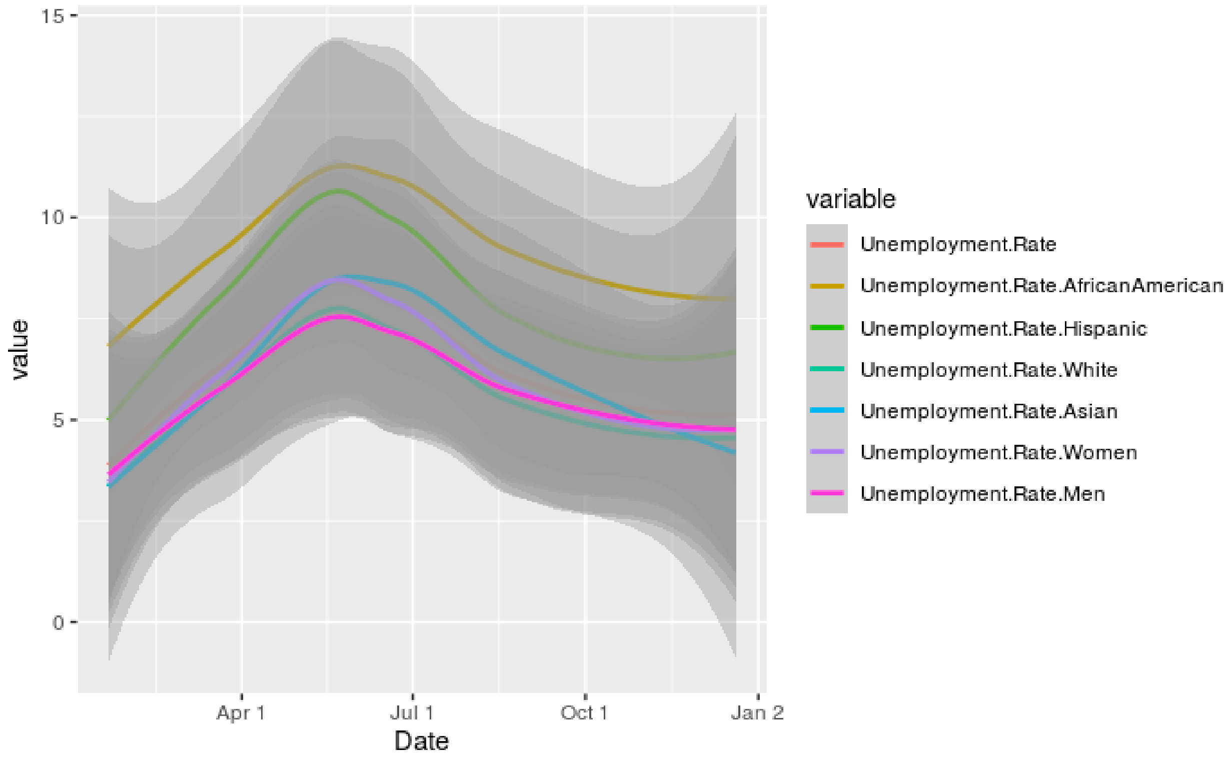


FIGURE 2. Time Series Plot of Predictor Variables.

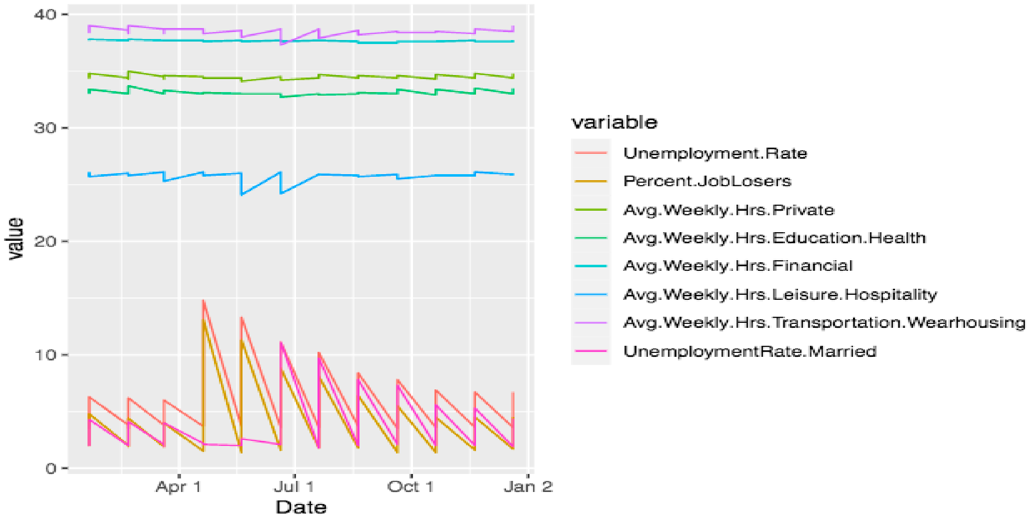


FIGURE 3 Time Series of Predictor Variables.

The figures above show the different trends over time. We observe that the unemployment rate for individual indicators by race and gender moves similar to the overall unemployment rate. In the second figure, we see that the average working hours remain consistent except for a drop in leisure and hospitality, as expected, and an increase in transportation and warehousing.

Model Summary

Methodology: We analysed the scatter plots between the unemployment rate and other predictor variables.

Most of the plots showed a linear relationship between the predictor variables and unemployment rate. However COVID cases and average hours worked in the private sector did not have a linear relationship.

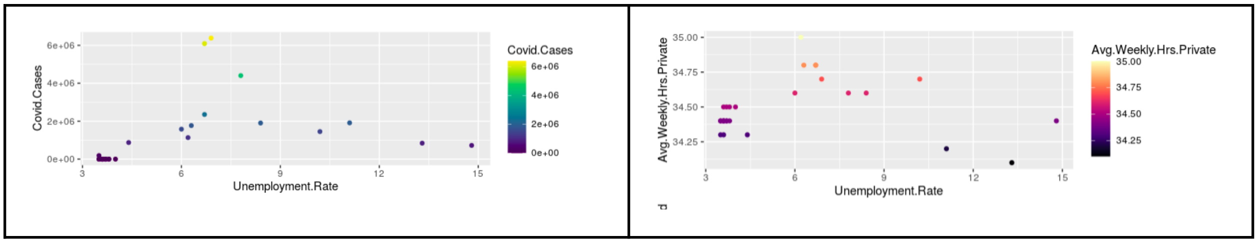


FIGURE 5. Non linear relationships in the data set. (COVID cases, Average Weekly Hrs Private Sector)

Multicollinearity: After running multicollinearity diagnostics on the full model and looking at the cor matrix and partial cor matrix, we used the Kendall method because of the small number of data points we had. The following elements were dropped from the model:

- Unemployment Rate for Men. (Overlap with the unemployment rate by race that took men and women into account)
- Unemployment Rate for Women. (Overlap with the unemployment rate by race that took men and women into account)
- Average Number of Hours for the Private Sector. (Overlap with breakdown of different private sector categories.

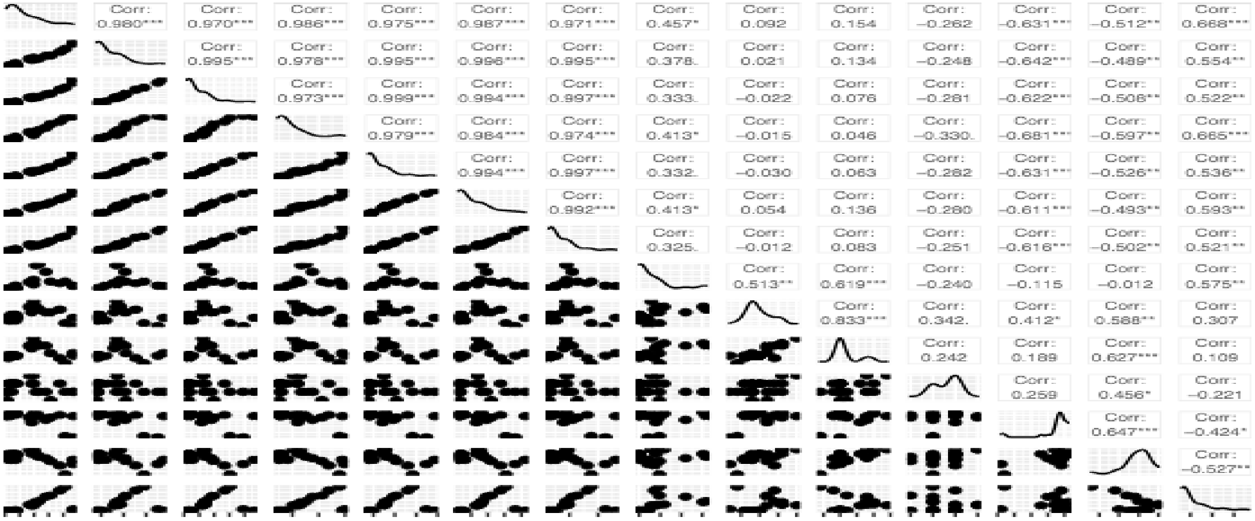
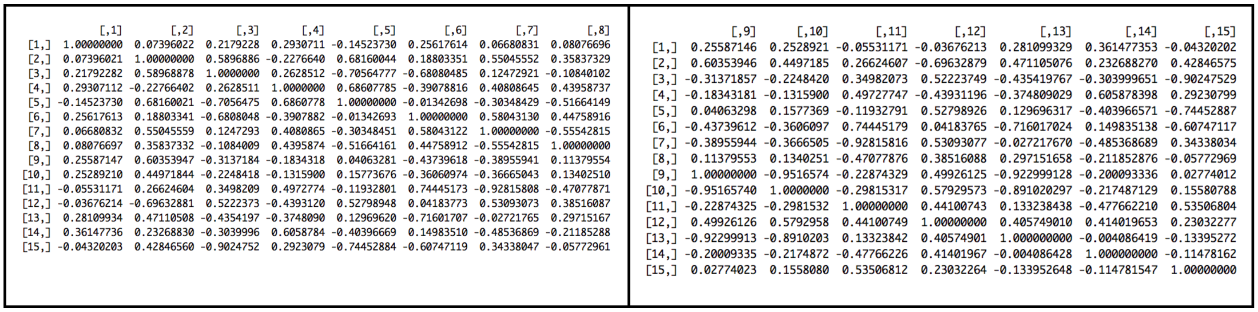


FIGURE 6. Correlation Matrix



```
              [,1]        [,2]       [,3]       [,4]        [,5]        [,6]        [,7]        [,8]
 [1,]  1.00000000  0.07396022  0.2179228  0.2930711 -0.14523730  0.25617614  0.06680831  0.08076696
 [2,]  0.07396021  1.00000000  0.5896886 -0.2276640  0.68160044  0.18803351  0.55045552  0.35837329
 [3,]  0.21792282  0.58968878  1.0000000  0.2628512 -0.70564777 -0.68080485  0.12472921 -0.10840102
 [4,]  0.29307112 -0.22766402  0.2628511  1.0000000  0.68607785 -0.39078816  0.40808645  0.43958737
 [5,] -0.14523730  0.68160021 -0.7056475  0.6860778  1.00000000 -0.01342698 -0.30348429 -0.51664149
 [6,]  0.25617613  0.18803341 -0.6808048 -0.3907882 -0.01342693  1.00000000  0.58043130  0.44758916
 [7,]  0.06680832  0.55045559  0.1247293  0.4080865 -0.30348451  0.58043122  1.00000000 -0.55542815
 [8,]  0.08076697  0.35837332 -0.1084009  0.4395874 -0.51664161  0.44758912 -0.55542815  1.00000000
 [9,]  0.25587147  0.60353947 -0.3137184 -0.1834318  0.04063281 -0.43739618 -0.38955941  0.11379554
[10,]  0.25289210  0.44971844 -0.2248418 -0.1315900  0.15773676 -0.36060974 -0.36665043  0.13402510
[11,] -0.05531171  0.26624604  0.3498209  0.4972774 -0.11932801  0.74445173 -0.92815808 -0.47077871
[12,] -0.03676214 -0.69632881  0.5222373 -0.4393120  0.52798948  0.04183773  0.53093073  0.38516087
[13,]  0.28109934  0.47110508 -0.4354197 -0.3748090  0.12969620 -0.71601707 -0.02721765  0.29715167
[14,]  0.36147736  0.23268830 -0.3039996  0.6058784 -0.40396669  0.14983510 -0.48536869 -0.21185288
[15,] -0.04320203  0.42846560 -0.9024752  0.2923079 -0.74452884 -0.60747119  0.34338047 -0.05772961
```

```
              [,9]       [,10]       [,11]       [,12]        [,13]        [,14]       [,15]
 [1,]  0.25587146  0.2528921 -0.05531171 -0.03676213  0.281099329  0.361477353 -0.04320202
 [2,]  0.60353946  0.4497185  0.26624607 -0.69632879  0.471105076  0.232688270  0.42846575
 [3,] -0.31371857 -0.2248420  0.34982073  0.52223749 -0.435419767 -0.303999651 -0.90247529
 [4,] -0.18343181 -0.1315900  0.49727747 -0.43931196 -0.374809029  0.605878398  0.29230799
 [5,]  0.04063298  0.1577369 -0.11932791  0.52798926  0.129696317 -0.403966571 -0.74452887
 [6,] -0.43739612 -0.3606097  0.74445179  0.04183765 -0.716017024  0.149835138 -0.60747117
 [7,] -0.38955944 -0.3666505 -0.92815816  0.53093077 -0.027217670 -0.485368689  0.34338034
 [8,]  0.11379553  0.1340251 -0.47077876  0.38516088  0.297151658 -0.211852876 -0.05772969
 [9,]  1.00000000 -0.9516574 -0.22874329  0.49926125 -0.922999128 -0.200093336  0.02774012
[10,] -0.95165740  1.0000000 -0.29815317  0.57929573 -0.891020297 -0.217487129  0.15580788
[11,] -0.22874325 -0.2981532  1.00000000  0.44100743  0.133238438 -0.477662210  0.53506804
[12,]  0.49926126  0.5792958  0.44100749  1.00000000  0.405749010  0.414019653  0.23032277
[13,] -0.92299913 -0.8910203  0.13323842  0.40574901  1.000000000 -0.004086419 -0.13395272
[14,] -0.20009335 -0.2174872 -0.47766226  0.41401967 -0.004086428  1.000000000 -0.11478162
[15,]  0.02774023  0.1558080  0.53506812  0.23032264 -0.133952648 -0.114781547  1.00000000
```

FIGURE 7. Partial Correlation Matrix

Outliers: The box plots do show us extreme values in almost every category; however, these outliers truly represent the spike in the unemployment numbers at the beginning of the pandemic. Therefore, we did not remove these points from the dataset.

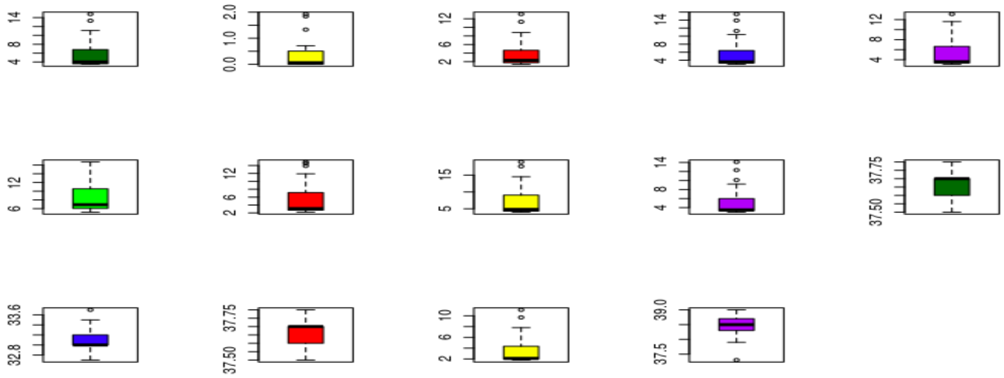


FIGURE 8. Box plots for all variables.

Transformations: The simple linear model had a p-value lower than .05; however, only two of the predictor variables were significant. We looked at the VIF and AIC after running diagnostics for the model and then performed different transformations, repeated the process to tune the variables and optimize selection.

Process followed:

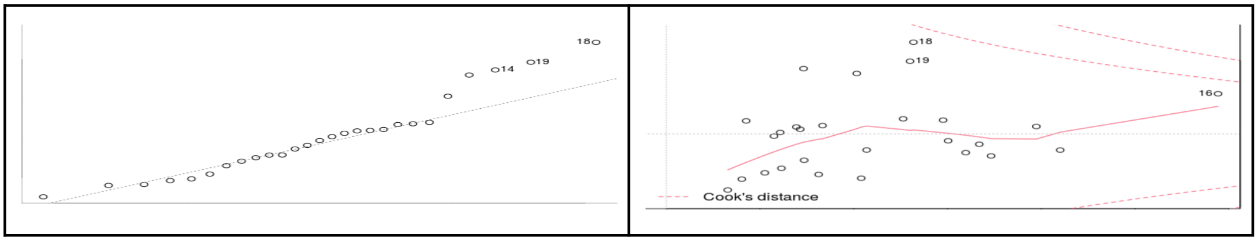
- Counts were normalized so that all the data points were in percentages.
- We tried the following transformations: log, sqrt, polynomial. We noted a drop in chi-square and VIF values when running the multicollinearity diagnostics in the sqrt, log transformations. We picked log and sqrt transformed equations to further our model development.
- Then we ran the ols_step_both_p function to determine the best possible variables to include in our model.
- This helped us further eliminate variables, leaving us with only *unemployment rates for African Americans, unemployment rates for White Americans, average weekly hours financial, Job losers percent, and Unemployment among married couple*s.
- These attributes resulted in statistically significant predictor variables as well.

Regression Model

```
Call:
lm(formula = log(reduced_df$Unemployment.Rate) ~ log(Unemployment.Rate.AfricanAmerican) +
   log(Unemployment.Rate.White) + log(Avg.Weekly.Hrs.Financial) +
   log(UnemploymentRate.Married) + log(Percent.JobLosers), data = reduced_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.018700 -0.006783  0.003299  0.006695  0.011409

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                           10.535792   3.966330   2.656  0.01477 *
log(Unemployment.Rate.AfricanAmerican)  0.261380   0.030805   8.485 3.16e-08 ***
log(Unemployment.Rate.White)            0.689164   0.025515  27.010  < 2e-16 ***
log(Avg.Weekly.Hrs.Financial)          -2.905807   1.090888  -2.664  0.01453 *
log(UnemploymentRate.Married)          -0.008684   0.005389  -1.612  0.12198
log(Percent.JobLosers)                  0.055826   0.017624   3.168  0.00464 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009731 on 21 degrees of freedom
Multiple R-squared:  0.9996,	Adjusted R-squared:  0.9996
F-statistic: 1.156e+04 on 5 and 21 DF,  p-value: < 2.2e-16

Overall Multicollinearity Diagnostics


                   MC Results detection
Determinant |X'X|:       0.0004         1
```

```
Farrar Chi-Square:      183.4034        1

Red Indicator:          0.6728          1

Sum of Lambda Inverse:   122.6031       1

Theil's Method:         -0.2363         0

Condition Number:       7139.9157       1


1 --> COLLINEARITY is detected by the test

0 --> COLLINEARITY is not detected by the test



Call:

imcdiag(mod = model)
```

Plots:



Conclusion

The null hypothesis, that the predictor variables are not correlated to the unemployment rate, was rejected due to the results from the initial and follow-up models. Our hypothesis that the unemployment rate can be modeled by these predictor variables was consistent with the model results; the most significant predictors were unemployment rate for African Americans, unemployment rate for White Americans, financial sector average weekly labor hours, percentage of job losers, and unemployment among married couples. However, the number of COVID cases is not a statistically significant contributor per our analysis, even though at the beginning of the pandemic we did note spikes.

Data Set Limitations and Future Considerations

- Since the coronavirus pandemic began spreading worldwide in January 2020, and the datasets were obtained in April 2021, there were only 15 months of data available for analysis. If this line of research were continued, the modeling and diagnostics could be run again with more data points, which could yield more robust conclusions.

- Working with percentages was a limitation to the analysis, as we did not have the raw data to get the unemployment counts. Transforming to counts may help alleviate the issue of high variance inflation factors for some of the variables.
- Additional variables such as telecommuting, remote schooling, vaccinations, level of education, and workforce age demographics may have significant effects, and including them could improve the predictive capability of the model.
- Once we have larger data sets in place, we would like to use methods like random forest and extreme gradient boosting to enhance our models.
- We would try the glm package with different function families.
- We would like to try the box-cox transformation to the lm model to see if additional predictors become significant.

Section 2: COVID and Unemployment Across US States

Robi Rahman

Research question and motivation

Due to the risk of coronavirus transmission within crowded workplaces and public venues, national and state governments have issued shutdown orders and other restrictions on businesses, with the goal of preventing widespread illness and hospital overcrowding. In recent political debates, a common topic was the unemployment caused by these restrictions and whether the resulting economic and social costs were justified by the health risks of the pandemic. This section investigates whether there is a relationship between COVID-19 cases and deaths within states, and the unemployment rate of those states.

Hypothesis

The hypothesis for this analysis is that states with higher per capita case and death rates will experience increases in unemployment, as states implement restrictions and people living in badly affected areas are more likely to avoid going to businesses and stores.

Data visualization

Unemployment was modeled as a response variable, using COVID cases and deaths as predictor variables, for the country overall as well as 52 individual states and areas. Graphs are available in the .Rmd output in the project repository, and some representative results are shown graphically below.

Methodology and assumptions

Unemployment data was obtained from the US Bureau of Labor Statistics for the time period of the pandemic. Daily COVID-19 case and death counts were extracted from the CDC dataset, then summed using R to get monthly totals for each state. State population data for 2020 were scraped from Wikipedia. [cite all three sources] Per capita case counts were determined by dividing statewide cases by each state's population. The time series data of monthly unemployment rates by state were used to produce a dataframe of monthly changes in unemployment by subtracting entries across consecutive months. The change in monthly unemployment rates were then analyzed against state COVID case and death rates using univariate and multivariate linear regressions; this analysis was conducted for the national composite dataset, and repeated 51 times using a for loop to iterate through the time series data for each individual state.

Limitations and diagnostics

Diagnostics show that the model has moderately good performance, with not much of a pattern to the residuals and no points having excessive leverage. However, the Q-Q plot indicates that quantiles in the residuals are not linear with the theoretical quantiles predicted by the model outside of quantiles in the range -1 to 1. This suggests it may be possible to produce a better model by running a Box-Cox transformation on the data, so this is a promising avenue for follow-up investigation.

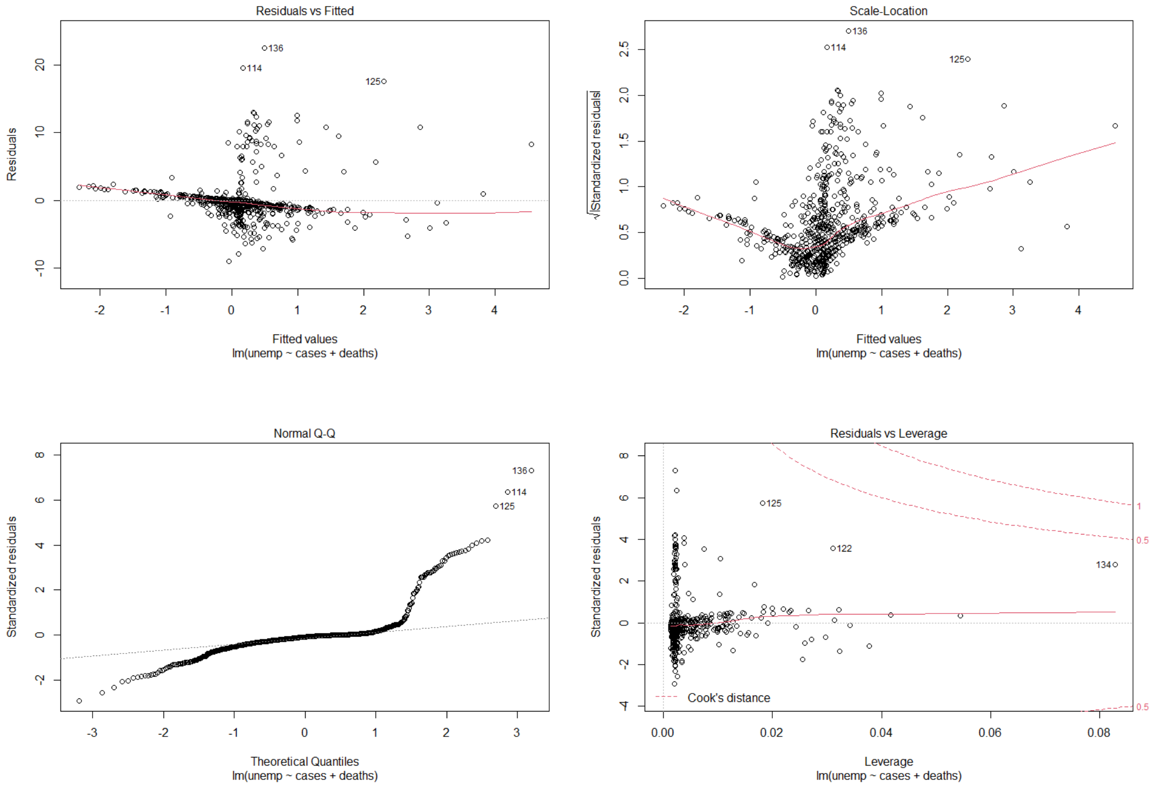
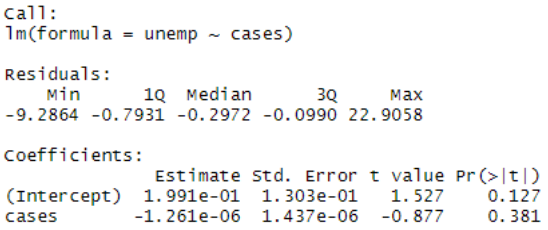


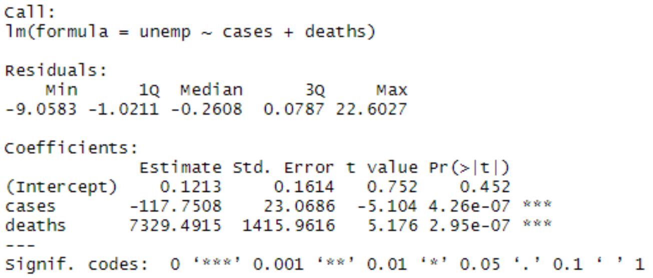
Figure 1: Diagnostics for the national model of monthly employment change vs cases and deaths

Results and conclusions

Based on the state-by-state case data for January 2020 through March 2021, it has turned out that statewide changes in unemployment are not significantly correlated with state case counts when these variables were examined one-on-one (see Figure 2). It appears that rather than shutting down businesses when case counts become high, states preemptively impose restrictions to stop cases from rising. Unemployment spiked before cases peaked, and across the 50 states there was not an overall trend of high case counts coinciding with negative changes in employment (see Figure 3). When unemployment change is modeled as a function of cases and deaths, both predictor variables are significant, with a negative coefficient on cases and a positive coefficient on deaths (Figure 2). This means that high death rates are associated with loss of employment, but high case counts without fatalities are not. This is historically reasonable, as states and countries have tightened restrictions when hospitals are overcrowded, death rates are high, and elderly and at-risk populations are most affected, as seen in the outbreaks in New York's dense apartment complexes and nursing homes. Conversely, they have taken a more lax approach when infection rates are high but deaths are low and cases are concentrated among young, healthy people with milder symptoms, such as Florida's outbreaks among beachgoing college students. Thus, the results of these models confirm the news reports and anecdotal experiences about varying degrees of outbreak severity and state lockdown policies.



```
Call:
lm(formula = unemp ~ cases)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.2864 -0.7931 -0.2972 -0.0990 22.9058 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.991e-01  1.303e-01   1.527    0.127
cases       -1.261e-06  1.437e-06  -0.877    0.381
```



```
Call:
lm(formula = unemp ~ cases + deaths)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.0583 -1.0211 -0.2608  0.0787 22.6027 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.1213     0.1614   0.752    0.452    
cases       -117.7508    23.0686  -5.104 4.26e-07 ***
deaths      7329.4915  1415.9616   5.176 2.95e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: results of the model of unemployment vs cases, compared to the model of unemployment vs cases and deaths. There is no relationship between cases and unemployment when they are considered one-on-one, but when deaths are introduced, the relationship is revealed: COVID fatalities are highly correlated to increased unemployment.

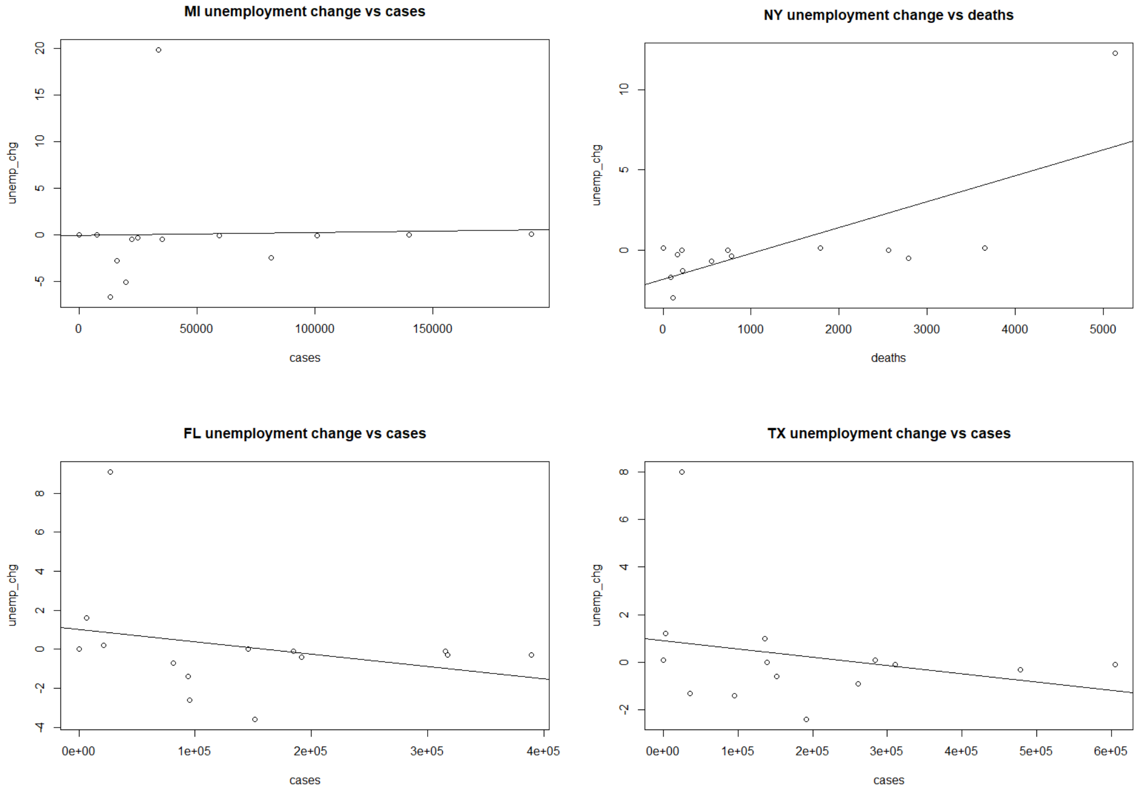


Figure 3: four representative scatterplots were selected from the 106 graphs produced by iterating through the datasets. Michigan has a typical pattern, showing little to no correlation between its monthly case count and the unemployment change in those months. New York exhibits a steep slope between deaths and unemployment - severely overcrowded hospitals led the state government to impose strict lockdowns, causing widespread unemployment soon after. Florida and Texas exhibit another trend, where lockdowns were relaxed when high case counts were accompanied by mild infections and low death rates.

Further discussion and research

These results suggest several avenues for investigation to follow up on this analysis. Although there is no overall national trend between cases and unemployment, this is not necessarily true if the data could be clustered or binned into appropriate categories. Simpson's Paradox demonstrates that even if an aggregated dataset shows a certain trend or no trend at all, subcategories within the data may exhibit a different or totally opposite trend. These could occur among clusters of states that exhibit common characteristics: for example, COVID may affect employment rates differently in urban vs rural states, or warm vs cold states.

It is theorized that the level of viral load during exposure affects severity of the subsequent coronavirus disease, with higher initial exposure leading to more severe infection, while fewer viruses during the initial exposure are more often associated with asymptomatic cases[2]. Thus, states where transmission occurs indoors or in crowded places (urban or cold areas) may need to enact different restrictions than places where transmission occurs outdoors (rural or warm areas), leading to different effects on employment. There may also be different trends in states governed by different political parties, since opposing politicians also tend to support different policies regarding pandemic lockdowns and economic stimulus.

Adding state-by-state data on other factors could improve this model, either by serving as predictor variables and improving the coefficient of determination, or by helping to control for confounding factors that may obscure the relationship between COVID and unemployment. Therefore, this analysis could benefit from the addition of more categorical and quantitative data for each state, such as the political party of its governor, the annual mean temperature, or the percentage of residents living in urban vs rural municipalities.

[2] https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30232-2/fulltext

Section 3: Effects of COVID-19 on Financial Markets

Iman Ismail

Research question

As the novel coronavirus spread across the world and triggered a global pandemic, economies were hit and went into sudden recessions, causing widespread unemployment and market downturns. However, unlike in past recessions where stock prices and employment rates were both hurt for long periods, global stock markets recovered relatively quickly from the initial shock of the pandemic, possibly due to innovative remote work arrangements that allowed companies to resume business, or because of government stimulus policies aiming to revive their economies.

In this section we analyze how various financial markets around the world were impacted by the global spread of the novel coronavirus. Stock indices such as the Dow Jones Industrial Average in the United States, Bovespa Index in Brazil, and MERVAL in Argentina were modeled using cumulative and new COVID cases as predictors.

Hypothesis

The hypothesis for this analysis is that COVID cases did not significantly affect stock market indices, as these quickly recovered from the initial crash and were revived by government stimulus programs.

Analysis and methodology

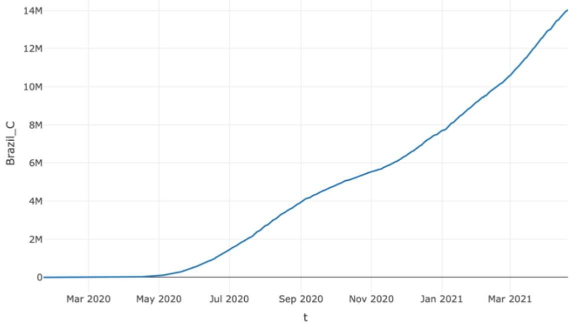
Daily time series data were gathered for case counts and stock index values in the United States, Argentina, Brazil, and Mexico for the time period from the beginning of the pandemic to the start date of the project. Each country's stock index was modeled as a function of its case counts; it was found that applying a logarithmic transformation to the cases improved the significance, robustness, and diagnostics of the models.

Results and conclusions

Plotting the time series data of cases and stock prices over time reveals that the COVID stock crashes of early 2020 happened before the majority of cases occurred, while the counts were still low. Stock markets then recovered and increased despite rising case counts. This may have been because all of the expected losses and business damage from the pandemic was priced into stock values up front, and then companies bounced back in spite of the hindrance to business caused by illnesses and lockdowns. These trends can be seen in Figures 1a and 1b, time series graphs of the Brazilian case counts and stock index.

Figure 1a, COVID cases in Brazil over time

Figure 1b, Bovespa Index over time



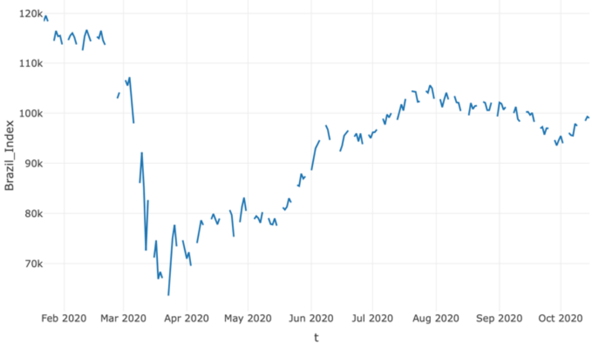
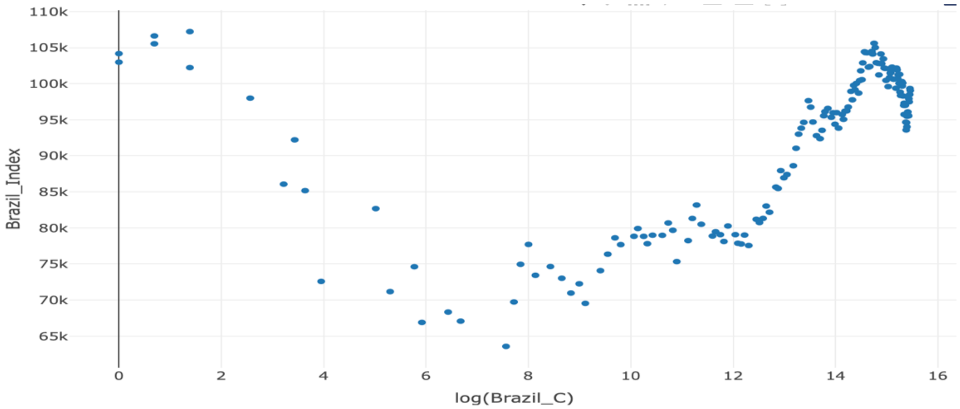


Figure 2, Bovespa index value vs logarithm of case count



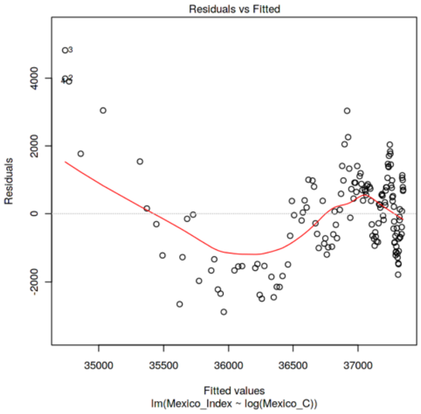
0.449395114889693

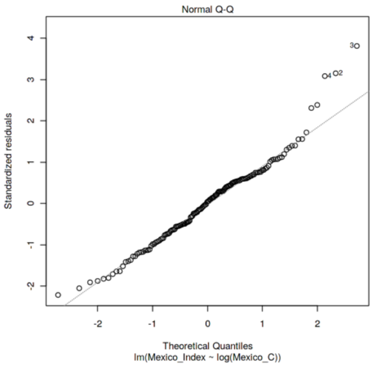
Figure 2 shows the correlation between the Brazilian stock market index and the logarithm of the number of COVID cases reported up to that date. After COVID cases exceeded 3000 (where log(cases)=8 on the x-axis), the trend was for cases and stock value to increase in tandem, but the opposite relationship was seen in the earlier days of the pandemic, while cases had not yet reached that level.

This analysis was repeated for Argentina, Mexico, and the United States, with similar results. The model outputs showed highly significant relationships between COVID cases and stock index values for all four countries, indicating that there was in fact a relationship between the two variables, and so the hypothesis was incorrect. However, diagnostic plots for the models showed that in each case, some of the data points had high leverage, and some extreme points at the tails of the Q-Q plots did not fall in line with the others, indicating that linearity may not be a safe assumption for all points in the dataset. A typical example is the output of the model for the Mexican data, below, showing that COVID cases are a highly significant factor of the

Mexican stock market value, but as in most real-world datasets, much of the variation remains unexplained.

```
Call:
lm(formula = Mexico_Index ~ log(Mexico_C), data = ALL_data[(ALL_data$Mexico_C) >
    6, ], na.action = na.omit)

Coefficients:
                 Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)     34308.46     446.33  76.867  < 2e-16 ***
log(Mexico_C)   222.83       38.64   5.767 4.37e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared:  0.1795, Adjusted R-squared:  0.1741
```





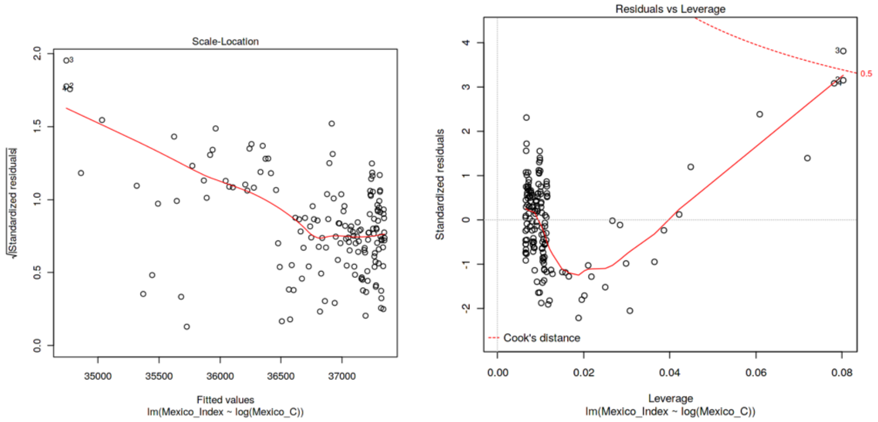


Figure 3, diagnostic plots for the model of Mexican stock index vs COVID cases. Note that the residuals exhibit a U-shaped trend vs number of cases, so there may be more links in the relationship between the variables than this linear model can capture. Most quantiles among residuals were linear with their theoretical quantiles, except for some points far from the rest of the data. Although there were no outliers in the dataset, these points warrant further investigation.

Section 4: Unemployment Rates Across Various Industries Within the United States and Their Significance

Priyanka Gunduboina

Research question and motivation

Many industries rely on employees operating in-person and on-site. Due to the COVID 19 pandemic, operating as such has become increasingly difficult. Pandemic precautions include wearing masks, disinfecting surfaces, and maintaining 6' social distancing between persons. Such precautions are more difficult to operate under in certain industries compared to others. Therefore, in this section, the unemployment rates per industry were investigated within the United States with the aim of understanding which of these exhibited the highest rates of unemployment during the pandemic. The significance of an industry's ability to operate virtually was also measured.

Hypothesis

The hypothesis for this analysis is that industries that cannot transition to operating through virtual means will exhibit the highest rates of unemployment, including sectors such as construction, manufacturing, and wholesale trade.

Assumptions and Methodology

The employment data was obtained by the U.S. Bureau of Labor Statistics on a monthly basis through the distribution of the Current Population Survey (CPS)[3]. The survey was given to 60,000 households comprising about 110,000 individuals sampled per survey. The Census bureau selected a sample of 800 geographic areas to represent each state and the District of Columbia for which the survey was distributed. The sample was designed by each state to select geographic regions that best reflect both its urban and rural areas but also industrial versus agricultural and major geographic divisions. Therefore, the employment data are representative of the United States employment numbers among various industries across the country.

The employment data are partitioned into 17 major industries: mining and logging, construction, manufacturing, wholesale trade, retail trade, transportation and warehousing, utilities, information, financial activities, professional and business service, education and

[3] https://www.bls.gov/cps/cps_htgm.htm

health services, leisure and hospitality, other services, government, federal government, state government, and local government. It is assumed that the mining and logging, construction, manufacturing, wholesale trade, transportation and warehousing, utilities and leisure and hospitality industries cannot switch their primary operations to virtual means, and therefore will exhibit the highest rates of unemployment per month. Of the four different government industries, only the 'government' category is considered as it is a consolidation of the later three.

Per the Centers for Disease Control and Prevention, the COVID 19 pandemic was believed to have become widespread within the United States March 2020, and is still ongoing. Therefore, data were considered for the period from March 2020 to March 2021.

Unemployment rates by industry and month were calculated by measuring the percent change between employment numbers of a particular month and employment numbers of the previous month. The average unemployment rate during the pandemic was then calculated for each industry from the monthly unemployment figures. The original approach was to create regression models for each industry and record slopes as the rate of unemployment per month, but due to the lack of data points (since there are only 13 months of pandemic unemployment figures to date), the models proved to be insignificant and unreliable.

The significance of industry type on the unemployment rate was measured by consolidating data into two columns: percent change of employment per month and industry. A factor variable indicating each industry’s ability or inability to operate virtually was also added in a third column. These variables were then used to create a multilinear regression model and significance of industry type and virtual operating ability were measured as a function of percent change of employment.

Data Visualization

Figure 4.1 Histograms of unemployment rates per month per industry

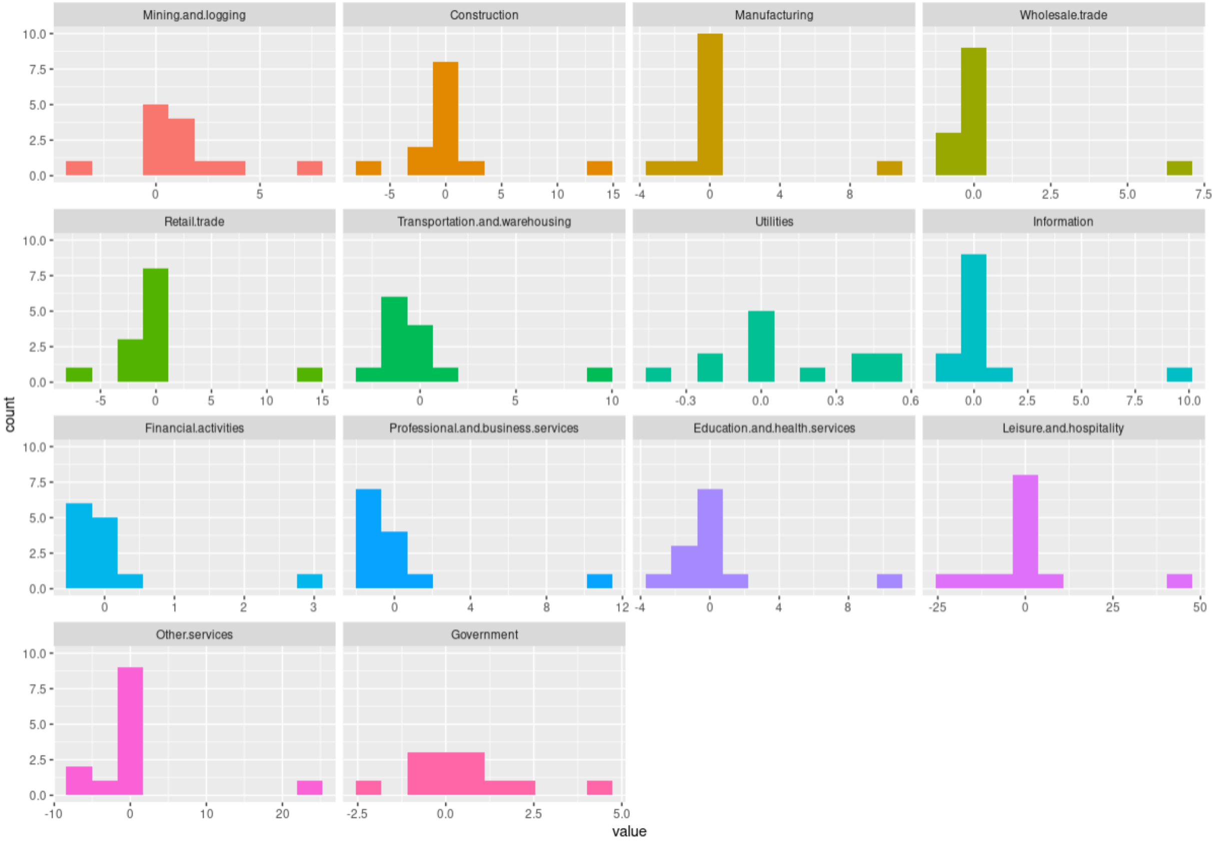


Figure 4.2: Average unemployment rate per industry during the COVID 19 Pandemic

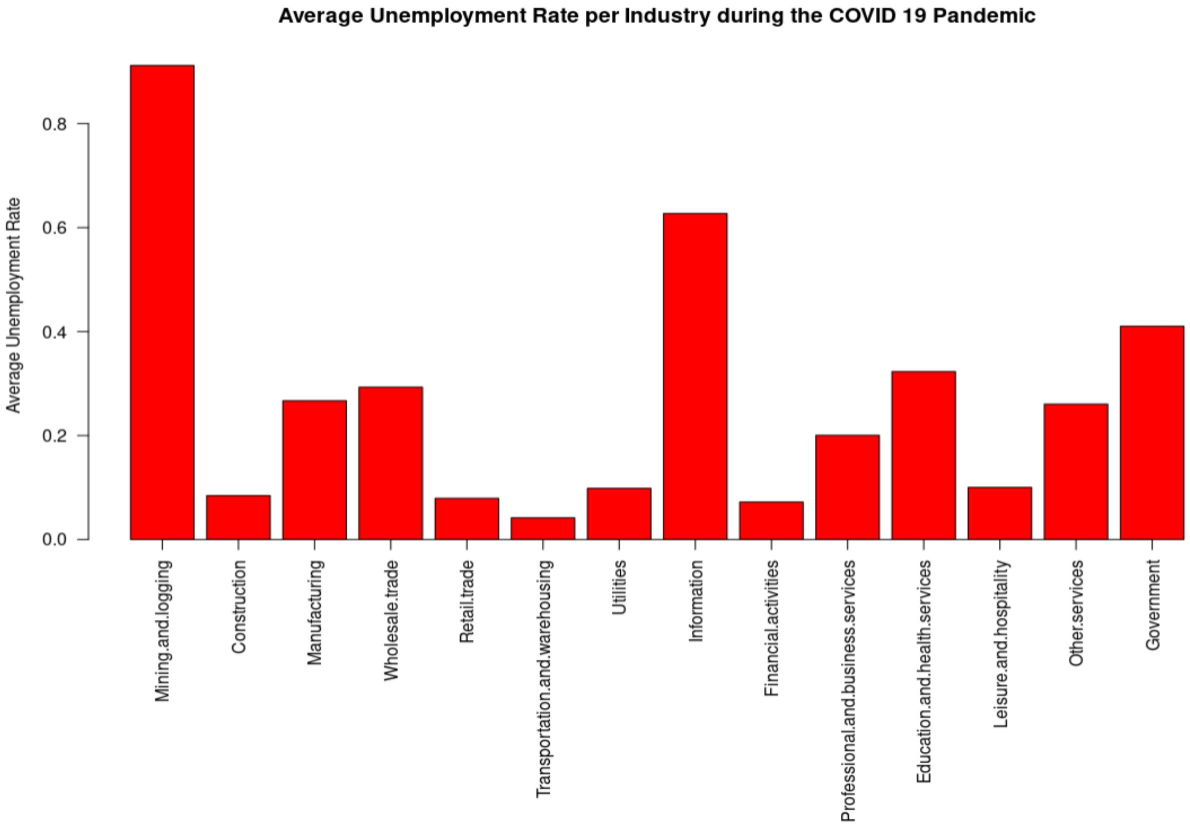
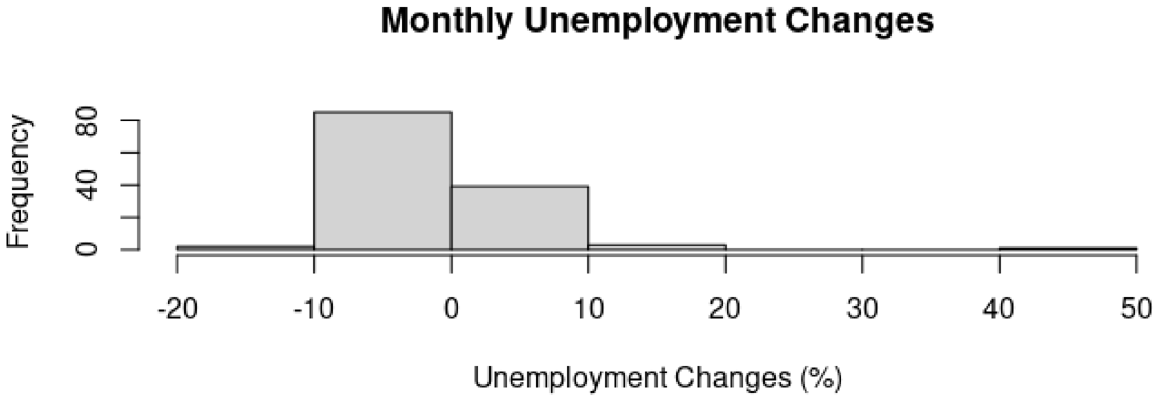


Figure 4.3: Histogram of Monthly Unemployment Changes



Limitations and Diagnostics

There were many limitations associated with this analysis. Only thirteen data points were available for unemployment data per industry as there have only been thirteen months during the pandemic. Therefore, it was difficult to run analyses on unemployment rates per industry during the pandemic. However, by averaging the percent changes in employment per month per industry, I was able to compare unemployment rates across industries during the pandemic as a whole.On average, the logging and mining industry exhibited the highest unemployment rate, followed by the information and government industries. The industry to exhibit the lowest unemployment rate was transportation and warehousing followed by financial activities and construction (Figure 4.2).

However, when observing the histograms (Figure 4.1) of the unemployment rates per industry, all industries except utilities were right-skewed. This indicates that there are a few small data points contributing to the skewness, and the mean may not best represent the data.Therefore, a linear regression model was also created to determine the significance of each industry and unemployment rate.

Results and Conclusions

All data were consolidated into one data frame, which gave 130 data points to work with. The histogram of the overall unemployment rates was also left skewed. Industries 'Professional and Business Services', 'Other Services' and 'Government' were removed from this analysis as it was not easily identifiable whether such industries would be able to operate virtually. After analysing the data, it was very apparent that due to the tested models, the variables: Industry and Virtually Operable do not explain the variance in the unemployment rates within the United States.

Overall, this analysis indicates that industry type and the industry's ability to function virtually does not have a significant impact on unemployment rates, and the hypothesis was incorrect.

Next Steps

As the pandemic progresses, there will be more data regarding unemployment per industry. Once more data are available, regression models can be built to better forecast monthly unemployment rates within each. This data can also be used to calculate the significance the type of industry has on unemployment rates. Also, there are many jobs available within each industry that can be fully remote. Therefore, future models can include the distribution of job titles that tend to be more correlated to remote work among various industries and model which industries can primarily function virtually.