**Section 1: COVID, Demographics, and Unemployment**

**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. https://www.bls.gov/

**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.

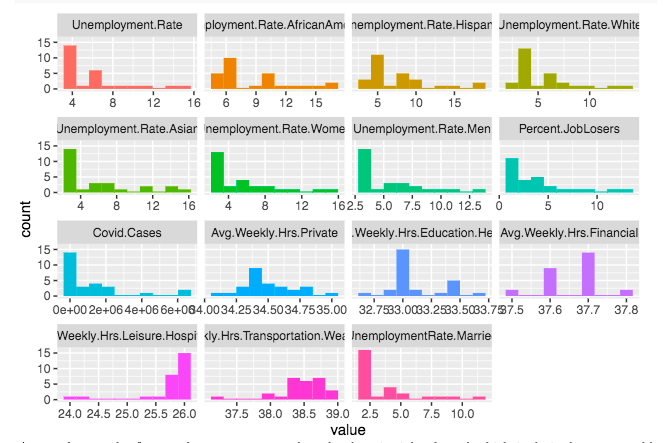
**Predictor Variables:** The following variables were taken into consideration to build the model.

|  |  |
| --- | --- |
| 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) |

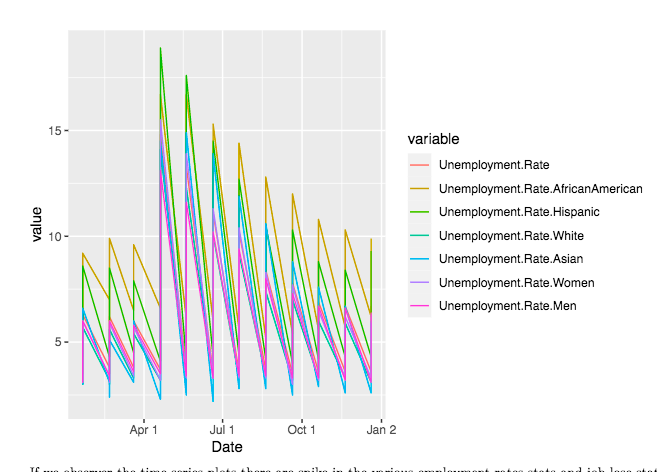
**Data Exploration and Visualization:**

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

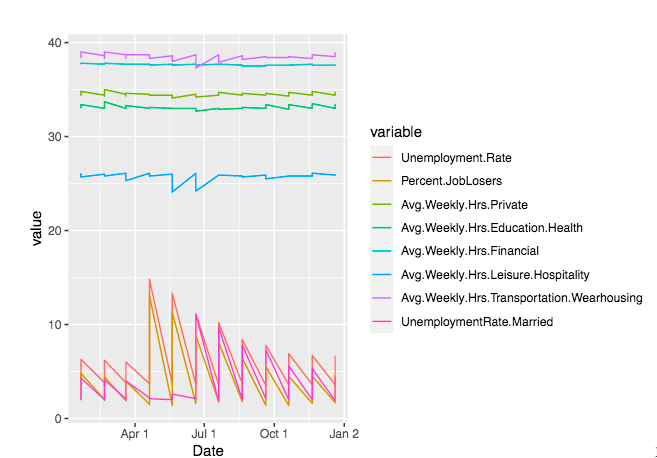
In *FIGURE 1*below, we see that most of the data variables are skewed left, indicating that the spike in unemployment rates in the different demographics was a result of the initial impact of the pandemic. Weekly hours in the hospitality, leisure and transportation industries also show left-skewed distributions; 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 unemployment rate for individual indicators by race , 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.

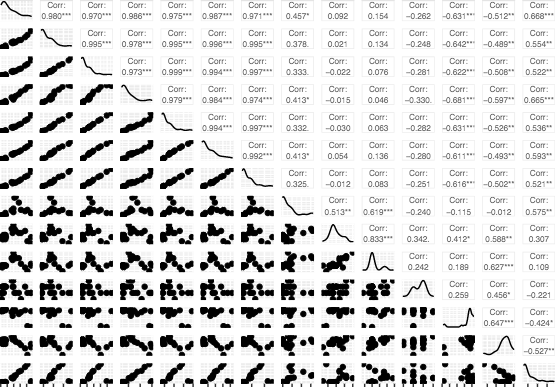
Most of the plots showed a linear relationship between the predictor variables and unemployment rate. However COVID cases and average hours worked in the private sector did not have a linear relationship.

|  |  |
| --- | --- |
|  |  |

**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.

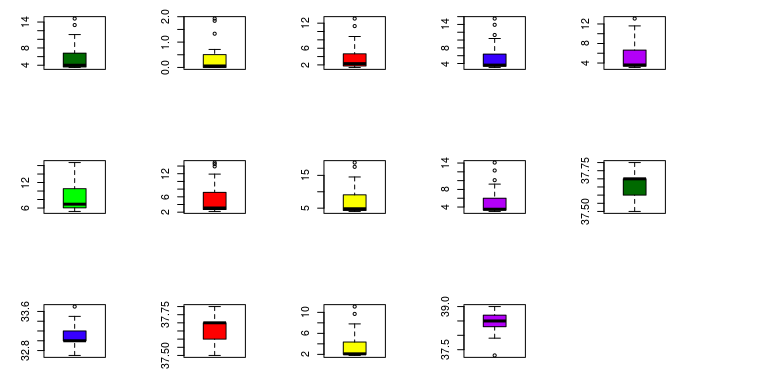


**FIGURE 6**. Correlation Matrix

|  |  |
| --- | --- |
|  |  |

**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*:

* Normalizing counts to percentages was the first step. 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 families and link.
* We would like to try the box-cox transformation to the lm model to see if additional predictors become significant.