Exploring the History of Unemployment Rates in the United States

Final Report for the MIS-748 Time Series Analysis for Business Forecasting

Anastasia Kurakova
Sabnam Pandit

12/4/24

Table of Contents

1. Introduction

Historically, unemployment rates in the country have been an indicator of economic growth and development. In recent years, the United States have been shaken by events resulting in unprecedented rise in unemployment among the population. COVID-19 pandemic resulted in the average unemployment rates across the country reaching 14.8% in April 2020 [1]. This spike in unemployment was the highest among the leisure industry and hospitality sector (39.8%). After that the population was hit by a wave of layoffs in tech companies triggered by maturing of the sector. An estimated number of 124,000 employees were laid off in 2024 alone adding to overall 428,449 tech workers who lost their jobs in 2022 and 2023 [2]. Inspired by these events that affected thousands of households around the country and concerned about our future, we have chosen to explore this topic.

We analyzed historical unemployment data across the United States to observe patterns, trends, and seasonality. Statistical algorithms were utilized to make short- term and long-term predictions. This analysis is crucial as unemployment rates are key economic indicators that directly impact policymaking, business planning, and personal well-being. Recent research indicates that unemployment rates have an effect on carbon emissions [3]. Accurate forecasts can help the government to stabilize the economy during downturns and plan for workforce needs in periods of growth.

For this project we used R-studio to perform data analysis and implement ARIMA statistical models for forecasting. Our research consists of two parts. First, we looked at the unemployment rates by state with a focus on California. In the second part, we proceeded with analysis of overall and gender specific trends in unemployment rates across America.

2. Data Description

For our research we used two datasets from Kaggle:

https://www.kaggle.com/datasets/axeltorbenson/unemployment-data-19482021/data

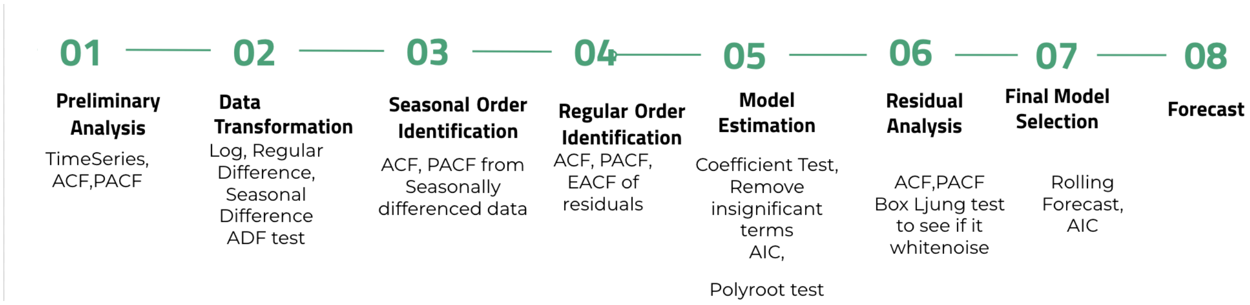
This dataset includes unemployment data from the US from 1947-2021. It contains 11 columns (Date, Unemployment rate for people based on gender and age ranges) and has 887 rows of data records.The data is tracked on a monthly basis.

https://www.kaggle.com/datasets/justin2028/unemployment-in-america-per-us-state/data

This dataset includes unemployment data from the US from 1976-2022. The dataset has 51000 rows and 11 columns. It contains unemployment data for all US states. The data is tracked monthly. It also contains additional statistics including overall population residing in the area along with percentage of employable population and unemployment rates. Analysis of this data could provide us insights into the evolution of unemployment rates for different states across the country over the past 46 years.

3. Analysis and Results

In this report, we do a time series analysis of unemployment rates of the United States by separating them according to states, gender, overall and California. We follow the following roadmap to do analysis and forecasting of the unemployment time series data.



01 **Preliminary Analysis** TimeSeries, ACF,PACF

02 **Data Transformation** Log, Regular Difference, Seasonal Difference ADF test

03 **Seasonal Order Identification** ACF, PACF from Seasonally differenced data

04 **Regular Order Identification** ACF, PACF, EACF of residuals

05 **Model Estimation** Coefficient Test, Remove insignificant terms AIC, Polyroot test

06 **Residual Analysis** ACF,PACF Box Ljung test to see if it whitenoise

07 **Final Model Selection** Rolling Forecast, AIC

08 **Forecast**

3.1. Unemployment Rates Across the States

For this part of the analysis we are using Kaggle dataset that contains the data from 1976-2022. The dataset contained monthly employment information for 50 states as well as special territories: New York City, Los Angeles County, and Washington DC

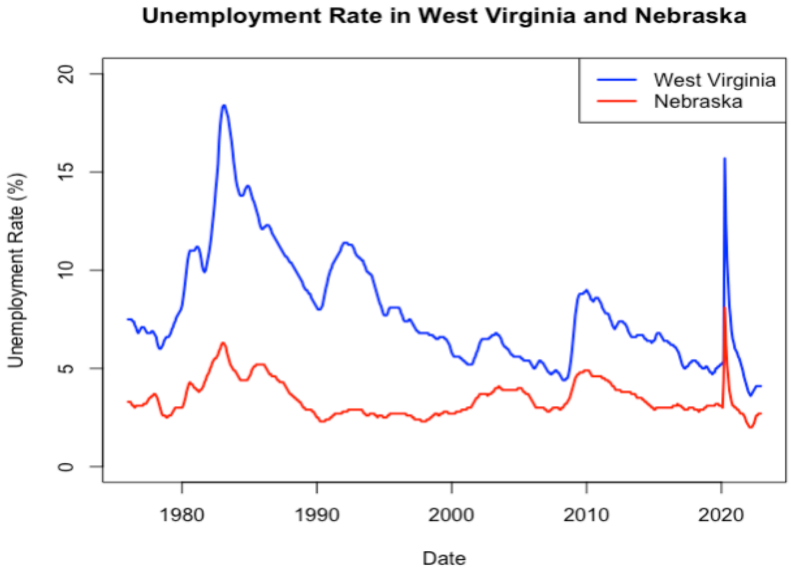
First of all we wanted to see which states have the highest and the lowest unemployment rates in the US.

Highest Rates	
State	**Avg Unemployment Rt**
West Virginia	**7.97**
New York City	**7.76**
Michigan	**7.74**
Alaska	**7.72**
Washington DC	**7.52**

Lowest Rates	
State	**Avg Unemployment Rt**
Vermont	**4.51**
New Hampshire	**4.23**
North Dakota	**3.74**
South Dakota	**3.59**
Nebraska	**3.44**

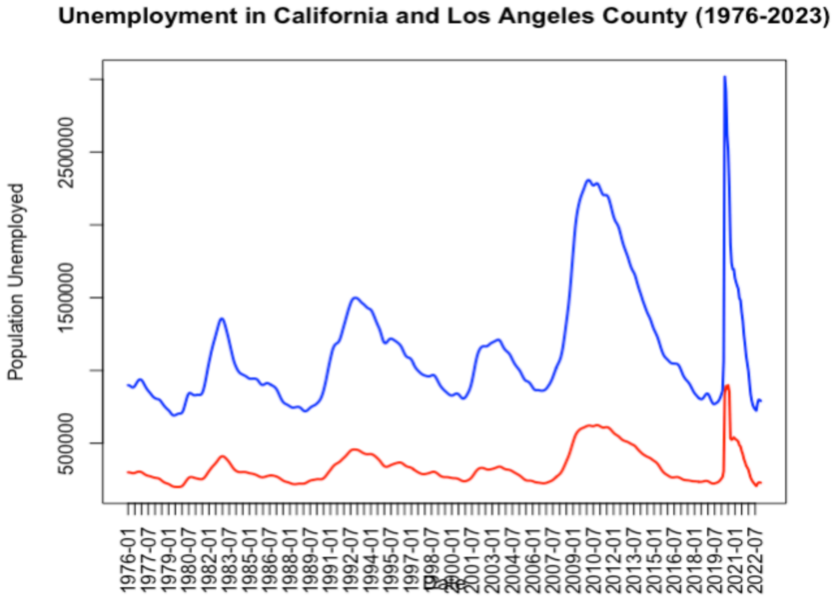
It's an interesting statistics because as we can see it is often not the most developed states and states with the best economy that have the lowest unemployment rates, but often the opposite is true.

We were also able to observe that in 2022 (3.48) the mean unemployment rate in the United States was less than a half of what it was in 1976 (7.17). The country is definitely moving in the right direction. However, there is still a lot to be done. We can also observe how differently states reacted to the global events affecting the economy.
To explore that we compared unemployment rates in the state with the highest mean rate - West Virginia(7.97) and the lowest - Nebraska (3.44).



We can see that rates in West Virginia are more volatile overall and have stronger spikes. That can be noticed by examining the spike caused by COVID-19 pandemic in 2020. But there is also a significant spike in 1992-1993 in West Virginia unemployment that is not present in the Nebraska at all. If we look at the world economy at that time it can be noticed that there is a global decline in demand for coal due to natural gas development and stricter ecological regulations. And West Virginia economy has been historically reliant on coal mining.

In our research, we focus on California State - the 6th highest mean unemployment rates in the country (7.23).

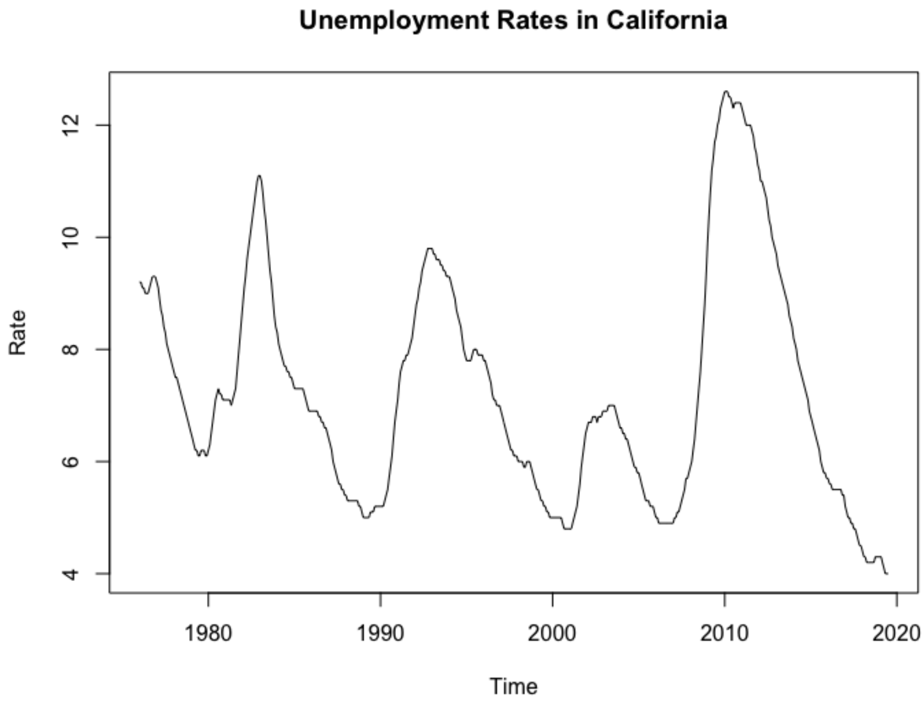


If we compare the Unemployment population in Los Angeles County (red) and California overall (blue) we can see how closely they are related meaning that LA county has a dramatic impact on the state economy overall. The lowest unemployment rates in California were observed in July 2022 - 3.8 and the highest were recorded during COVID-19 pandemic in April 2020 - 16.1.

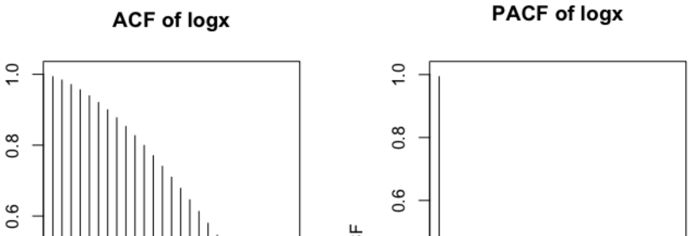
3.2. Unemployment Rates in California

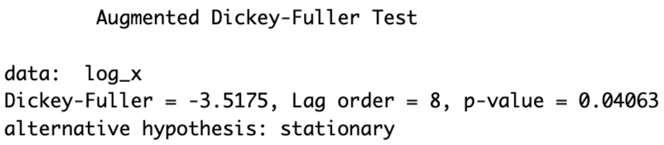
According to the documentation provided by Bureau of labor statistics, the dataset we are working with was adjusted for seasonality meaning we won't detect strong seasonal spikes due to farming expected in summer months [4]. We are exploring unemployment rates in California from 1976-2022

We have used data till the February 28,2020 to avoid COVID-19 event that had a huge impact on the unemployment rates across the country and particularly in California with the highest rate of 16.1% of state population. 7 months of data were left for testing.



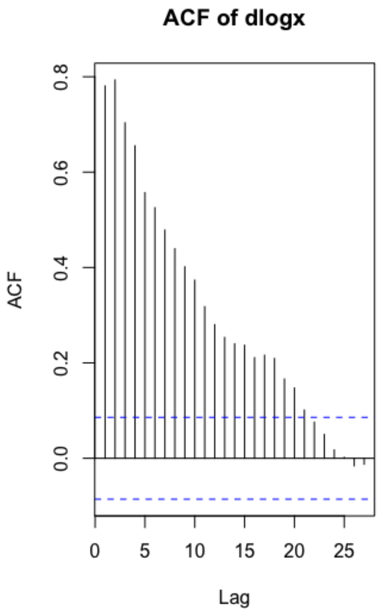
Our data is not stationary, so the first thing was to take a log of the data.

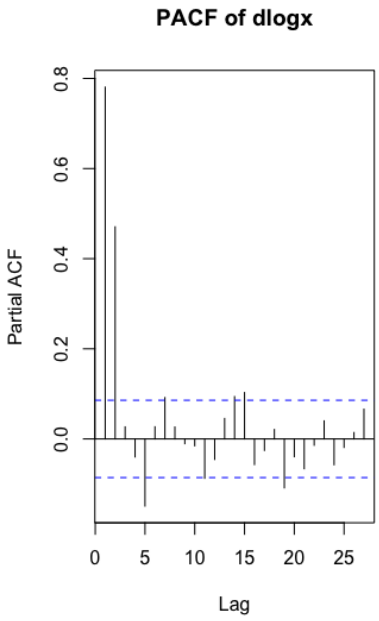


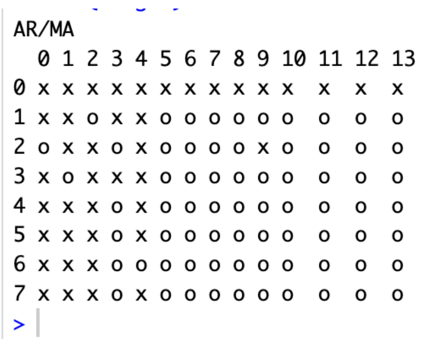


```
        Augmented Dickey-Fuller Test

data:  log_x
Dickey-Fuller = -3.5175, Lag order = 8, p-value = 0.04063
alternative hypothesis: stationary
```

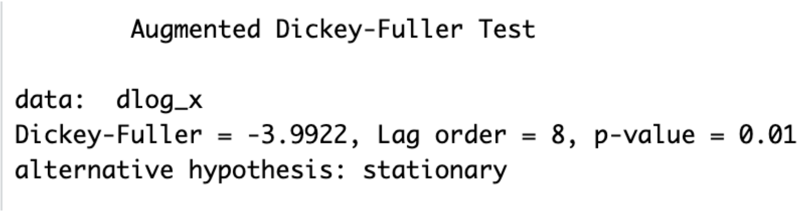
Data still doesn't look stationary with ACF decaying slowly and PACF spikes at lags 1,2, and 3. Although, adf test had a significant p-value<0.05, it was decided to proceed with differencing the data. ACF and PACF of dlog_x look better, and adf has a p-value of 0.01 which is significant meaning the data is stationary.


ACF of dlogx


PACF of dlogx



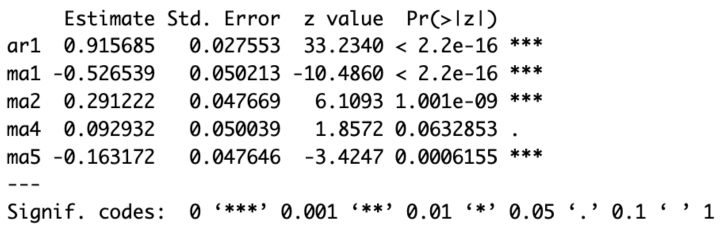
```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x 
1 x x o x x o o o o o o  o  o  o 
2 o x x o x o o o o x o  o  o  o 
3 x o x x x o o o o o o  o  o  o 
4 x x x o x o o o o o o  o  o  o 
5 x x x o x o o o o o o  o  o  o 
6 x x x o o o o o o o o  o  o  o 
7 x x x o x o o o o o o  o  o  o 
> 
```



```
        Augmented Dickey-Fuller Test

data:  dlog_x
Dickey-Fuller = -3.9922, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary
```

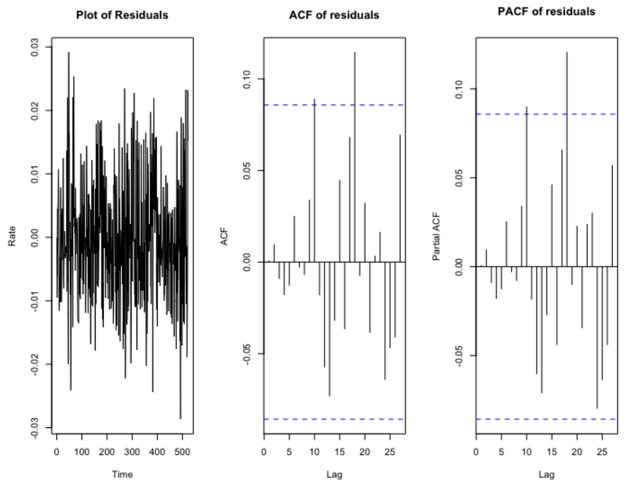
We proceed with eacf to pick a best fitted model. We have chosen to include ARMA(1,5), ARMA(1,9) and ARMA(3,5) in the evaluation.

First Model we looked at was ARMA(1,5). ma3 was insignificant and after removal we got a model out15.1: AIC(out15.1) = -3416.3, BIC(out15.1) = -3388.75

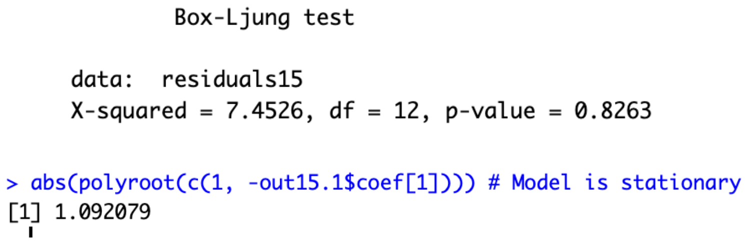


```
     Estimate Std. Error  z value  Pr(>|z|)    
ar1  0.915685   0.027553  33.2340 < 2.2e-16 ***
ma1 -0.526539   0.050213 -10.4860 < 2.2e-16 ***
ma2  0.291222   0.047669   6.1093 1.001e-09 ***
ma4  0.092932   0.050039   1.8572 0.0632853 .  
ma5 -0.163172   0.047646  -3.4247 0.0006155 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Next, we checked the residuals to confirm that they behave as white noise.

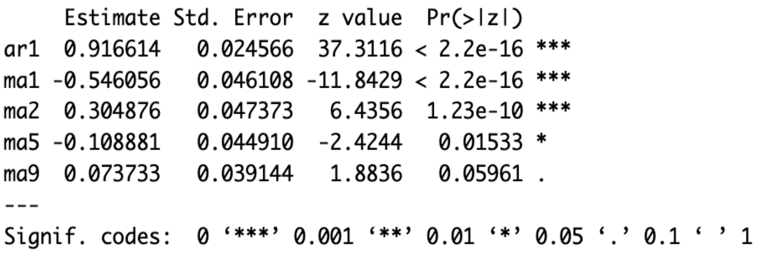


We also used Box-Ljung test to confirm that there is no autocorrelation

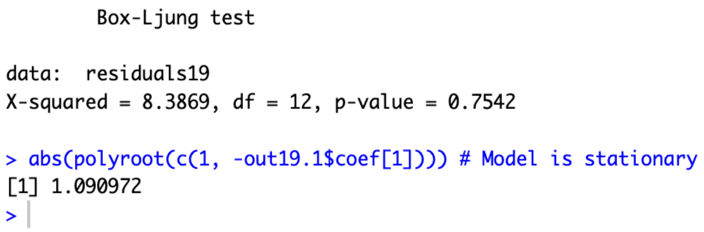


```
	Box-Ljung test

data:  residuals15
X-squared = 7.4526, df = 12, p-value = 0.8263

> abs(polyroot(c(1, -out15.1$coef[1]))) # Model is stationary
[1] 1.092079
```

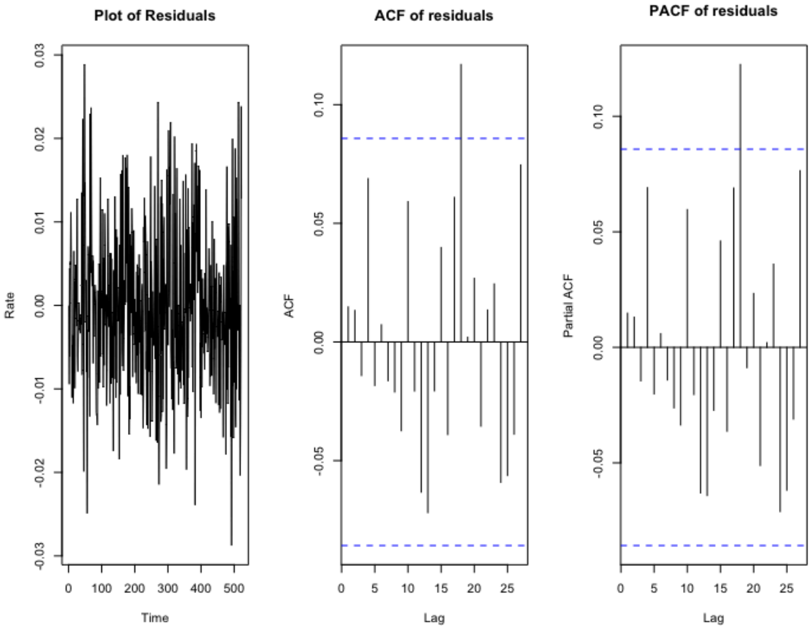
Second model we explored is ARMA(1,9). Again there were multiple insignificant terms that had to be removed. The final model is out19.1: AIC = -3416.27, BIC = -3388.28



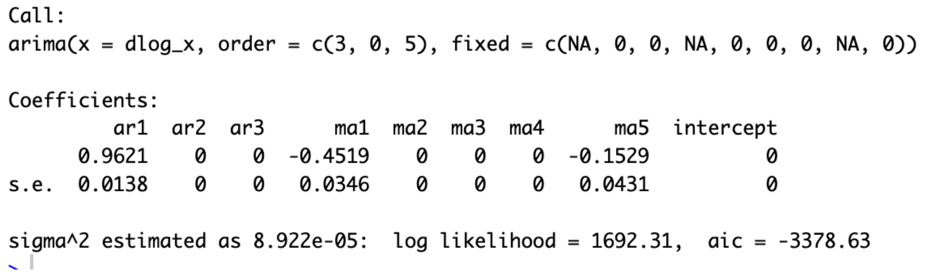
```
     Estimate Std. Error  z value  Pr(>|z|)    
ar1  0.916614   0.024566  37.3116 < 2.2e-16 ***
ma1 -0.546056   0.046108 -11.8429 < 2.2e-16 ***
ma2  0.304876   0.047373   6.4356  1.23e-10 ***
ma5 -0.108881   0.044910  -2.4244   0.01533 *  
ma9  0.073733   0.039144   1.8836   0.05961 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
	Box-Ljung test

data:  residuals19
X-squared = 8.3869, df = 12, p-value = 0.7542

> abs(polyroot(c(1, -out19.1$coef[1]))) # Model is stationary
[1] 1.090972
> 
```

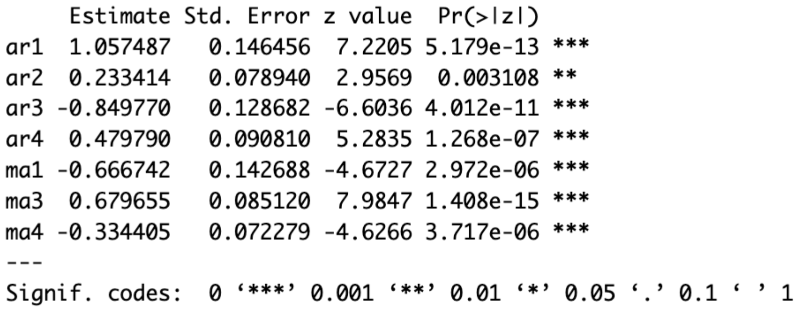


Then we looked at the ARMA(3,5) model. After inspecting the coefficients and excluding insignificant terms, we were left with the model that has an AIC significantly worse than the other 2 model. We chose not to proceed with it.



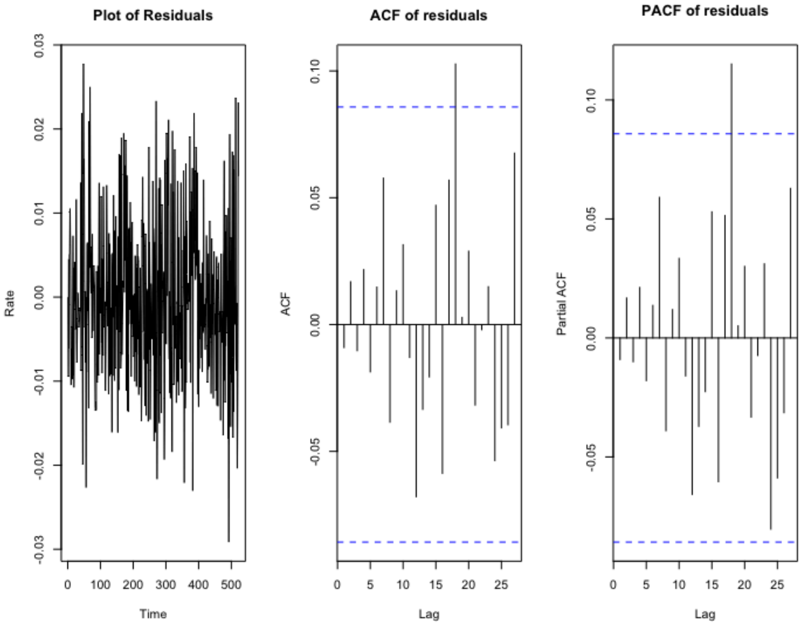
```
Call:
arima(x = dlog_x, order = c(3, 0, 5), fixed = c(NA, 0, 0, NA, 0, 0, 0, NA, 0))

Coefficients:
         ar1  ar2  ar3      ma1  ma2  ma3  ma4      ma5  intercept
      0.9621    0    0  -0.4519    0    0    0  -0.1529          0
s.e.  0.0138    0    0   0.0346    0    0    0   0.0431          0

sigma^2 estimated as 8.922e-05:  log likelihood = 1692.31,  aic = -3378.63
```

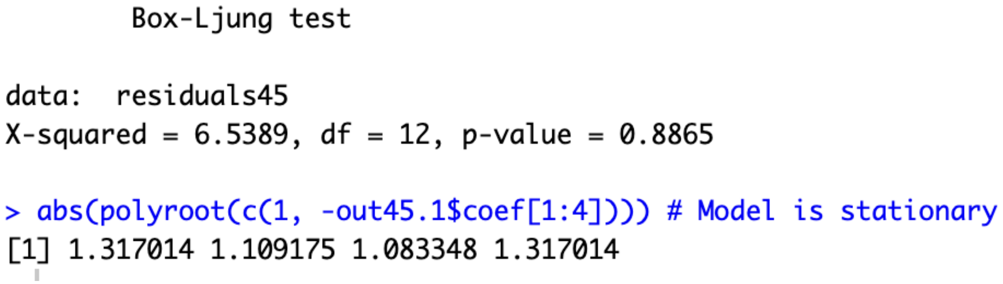
ARMA(4,5) after removing insignificant terms:
AIC=-3415, BIC= -3379



```
     Estimate Std. Error z value  Pr(>|z|)    
ar1  1.057487   0.146456  7.2205 5.179e-13 ***
ar2  0.233414   0.078940  2.9569  0.003108 ** 
ar3 -0.849770   0.128682 -6.6036 4.012e-11 ***
ar4  0.479790   0.090810  5.2835 1.268e-07 ***
ma1 -0.666742   0.142688 -4.6727 2.972e-06 ***
ma3  0.679655   0.085120  7.9847 1.408e-15 ***
ma4 -0.334405   0.072279 -4.6266 3.717e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

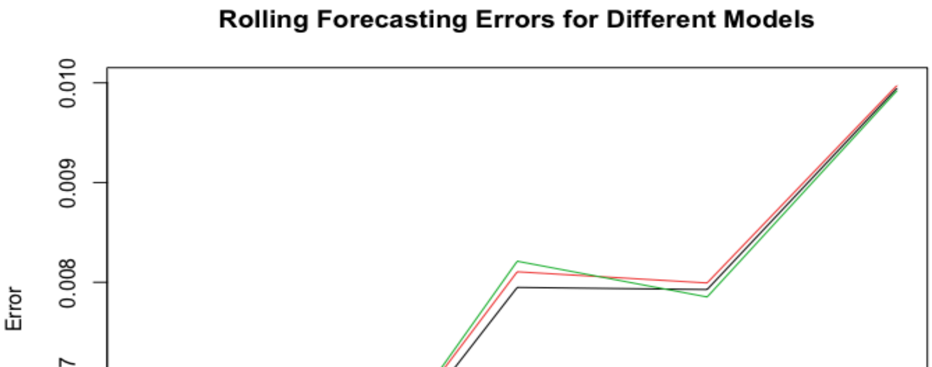
Residual analysis:



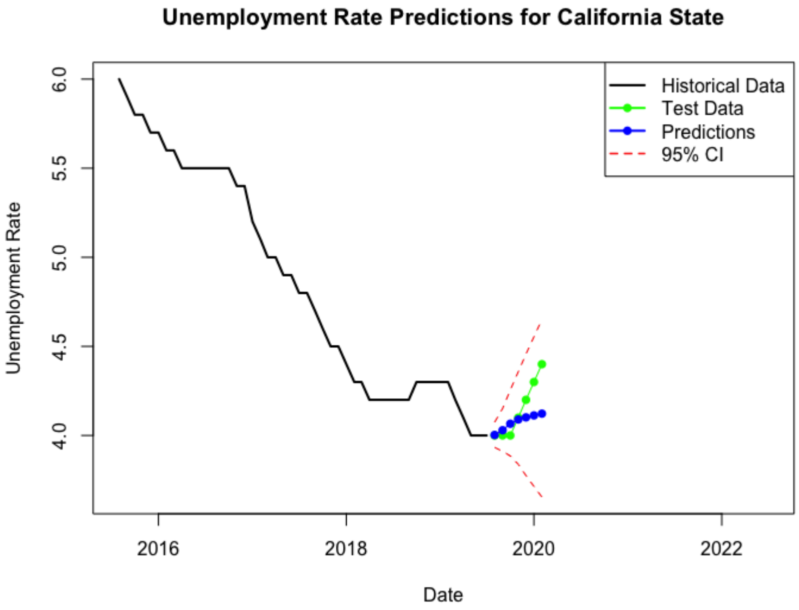


```
        Box-Ljung test

data:  residuals45
X-squared = 6.5389, df = 12, p-value = 0.8865

> abs(polyroot(c(1, -out45.1$coef[1:4]))) # Model is stationary
[1] 1.317014 1.109175 1.083348 1.317014
```

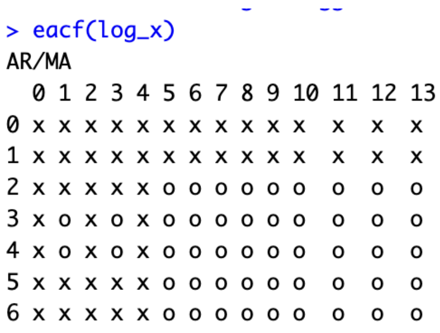
We performed rolling forecast analysis to pick the best model and it looks like ARMA(1,0,5) showed the best results and the lowest error for the rolling forecast.



However, when we proceeded to create an actual prediction, the results produced by ARMA(1,5) model were not very accurate.

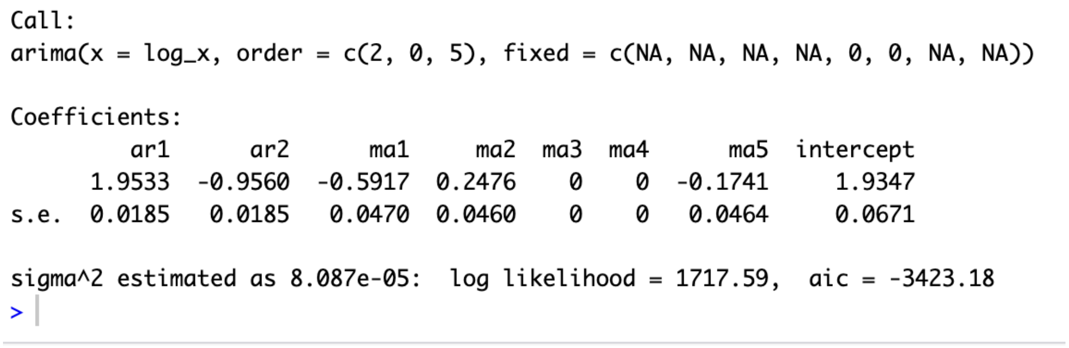


Due to this discrepancy, we have chosen to return to log(x) data and try the model using that data since adf showed stationarity.

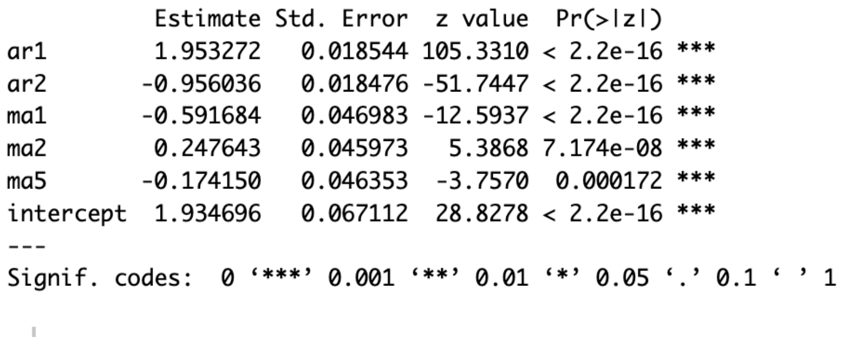


```
> eacf(log_x)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x x x x x x x x x x  x  x  x
2 x x x x x o o o o o o  o  o  o
3 x o x o x o o o o o o  o  o  o
4 x o x o x o o o o o o  o  o  o
5 x x x x x o o o o o o  o  o  o
6 x x x x x o o o o o o  o  o  o
```

Looking at the eacf plot of log(x),
We have chosen ARMA(2,5) model.



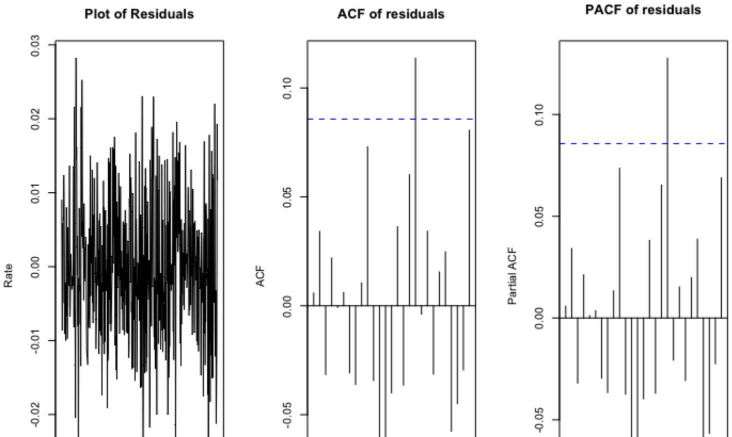
```
Call:
arima(x = log_x, order = c(2, 0, 5), fixed = c(NA, NA, NA, NA, 0, 0, NA, NA))

Coefficients:
         ar1      ar2      ma1     ma2  ma3  ma4      ma5  intercept
      1.9533  -0.9560  -0.5917  0.2476    0    0  -0.1741     1.9347
s.e.  0.0185   0.0185   0.0470  0.0460    0    0   0.0464     0.0671

sigma^2 estimated as 8.087e-05:  log likelihood = 1717.59,  aic = -3423.18
>
```

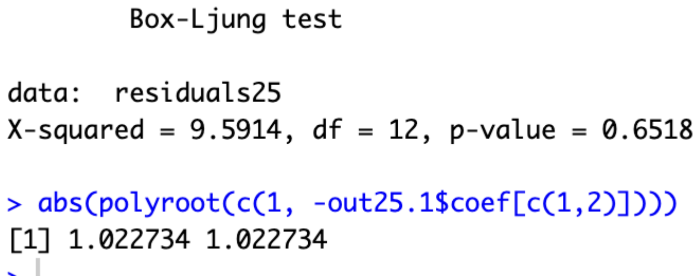


```
           Estimate Std. Error  z value  Pr(>|z|)    
ar1        1.953272   0.018544 105.3310 < 2.2e-16 ***
ar2       -0.956036   0.018476 -51.7447 < 2.2e-16 ***
ma1       -0.591684   0.046983 -12.5937 < 2.2e-16 ***
ma2        0.247643   0.045973   5.3868 7.174e-08 ***
ma5       -0.174150   0.046353  -3.7570  0.000172 ***
intercept  1.934696   0.067112  28.8278 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can observe that final AIC (-3423) of this model is better than the ones we were able to obtain previously. BIC = -3391

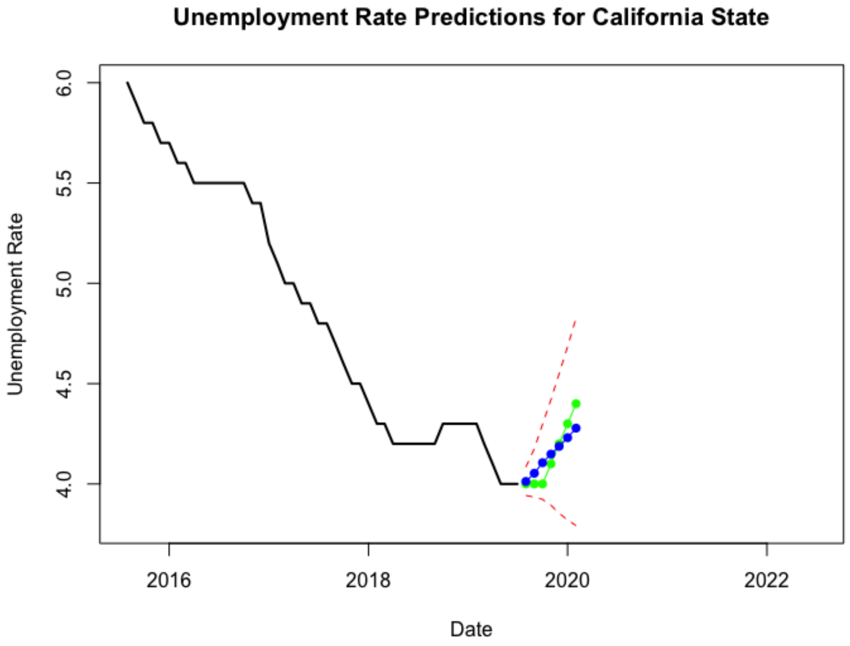
Looking at the residuals we can see that they behave as white noise.





```
        Box-Ljung test

data:  residuals25
X-squared = 9.5914, df = 12, p-value = 0.6518

> abs(polyroot(c(1, -out25.1$coef[c(1,2)])))
[1] 1.022734 1.022734
```

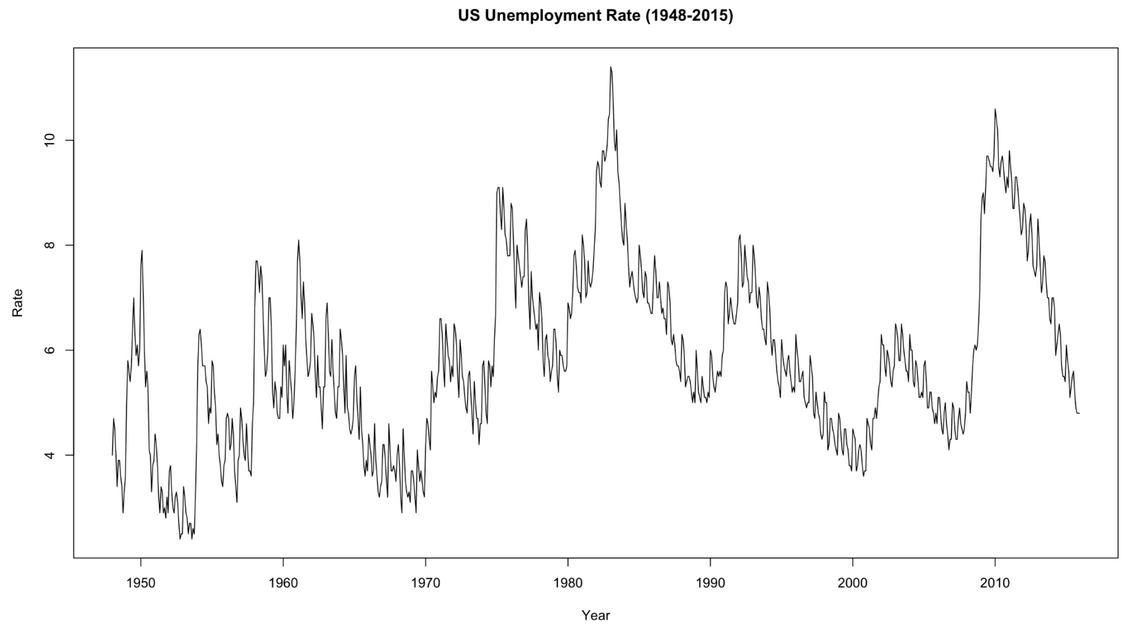
The forecasting results for this model were more reliable and closer to the testing data compared to the previous predictions.

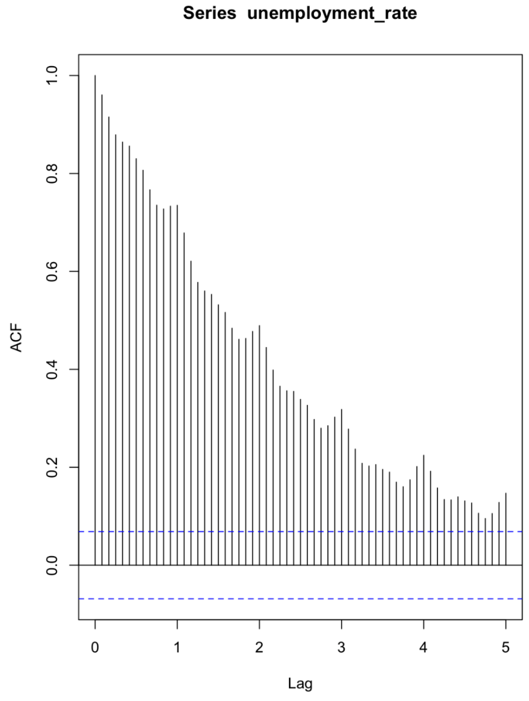


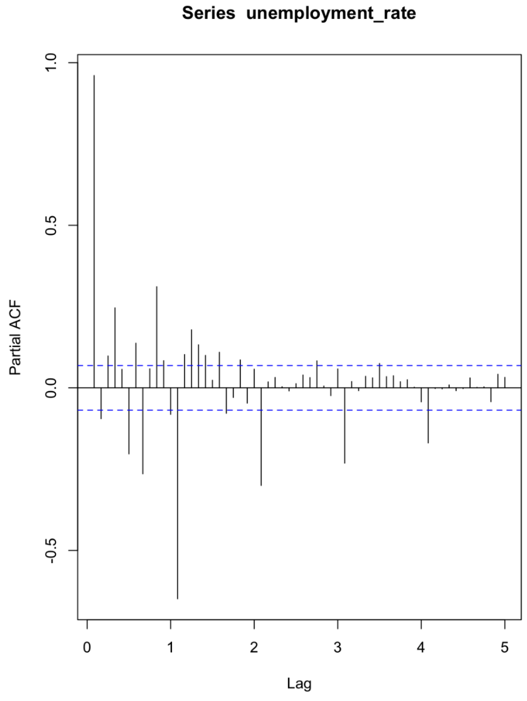
3.3 Total Overall Unemployment Rate of US

We are working on the Unemployment Data from Dataset 1. The dataset contains the data from 1948 to 2021. Our aim is to successfully forecast the unemployment rate in the US. The approach that is carried out in this part of the project is to first divide the datasets into training and test sets. We are taking the data from 1948 to 2015 as training data and the rest of the data for testing purposes.

Since we are working with the overall unemployment data at this part of the project ,we just take that column. The first step is to visualize the nature of the data for which we plot the time series data and ACF, PACF for the data.


US Unemployment Rate (1948-2015)

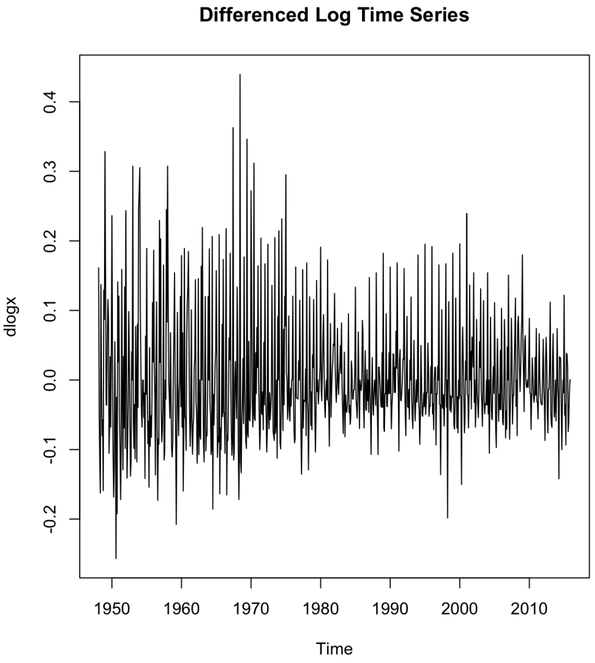

Series unemployment_rate


Series unemployment_rate

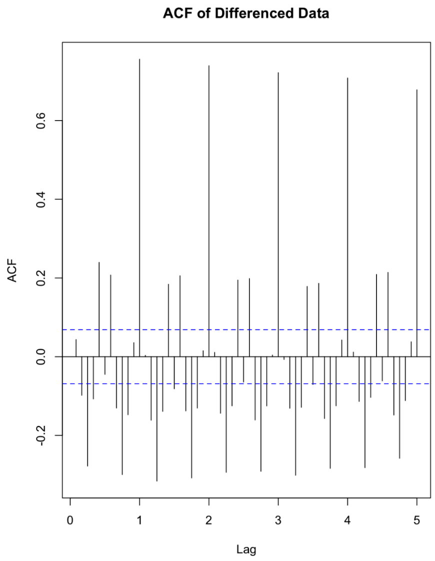
We observe that there is a periodic trend in the time series plot . The ACF and PACF also suggest that the data is non-stationary.

Since the data is non-stationary and has non constant mean and non constant variance we take the log of the data first and the difference of the logged data.

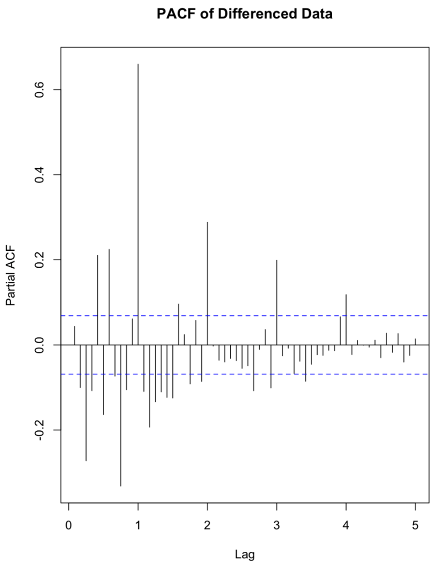
We plot the logged and diff data and its ACF and PACF to see patterns and stationarity.



Differenced Log Time Series

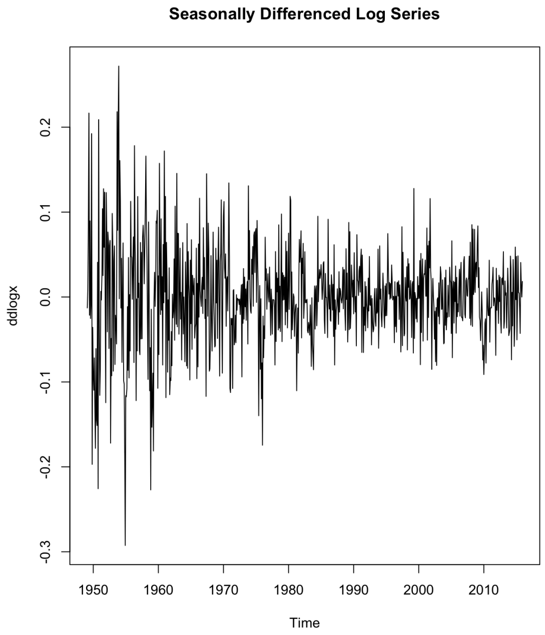


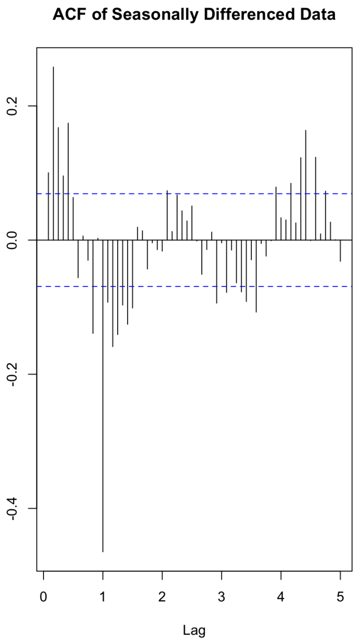
ACF of Differenced Data

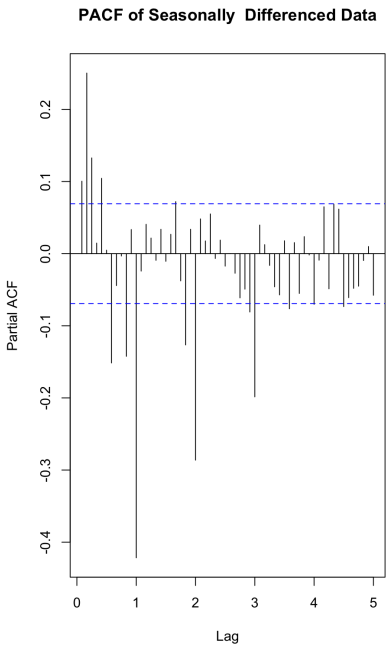


PACF of Differenced Data

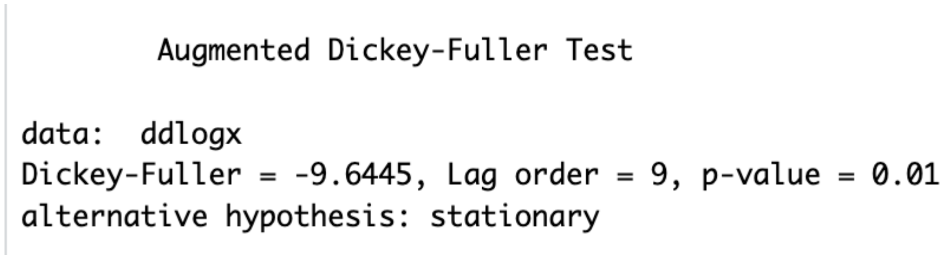
The ACF and PACF plot suggest that there is a seasonality. As we can observe the spikes at s=12. To handle the seasonality, we take the seasonal difference on the (log and diff) transformed data. After transforming the data with the seasonal difference (s=12) ,we get the following plot and ACF, PACF.







The adf test verifies the stationarity of the data.



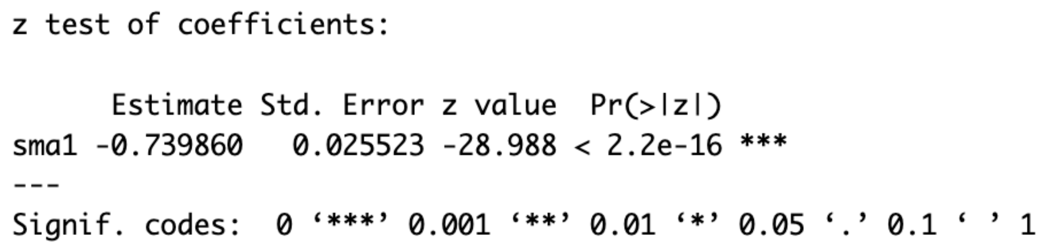
```
        Augmented Dickey-Fuller Test

data:  ddlogx
Dickey-Fuller = -9.6445, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

In order to get the seasonal order of the seasonal model we observe the ACF and PACF of seasonally differenced data. From the ACF , PACF plot we have two possible orders of P, Q for the Seasonal model.

- SAR(3)
- SMA(1)

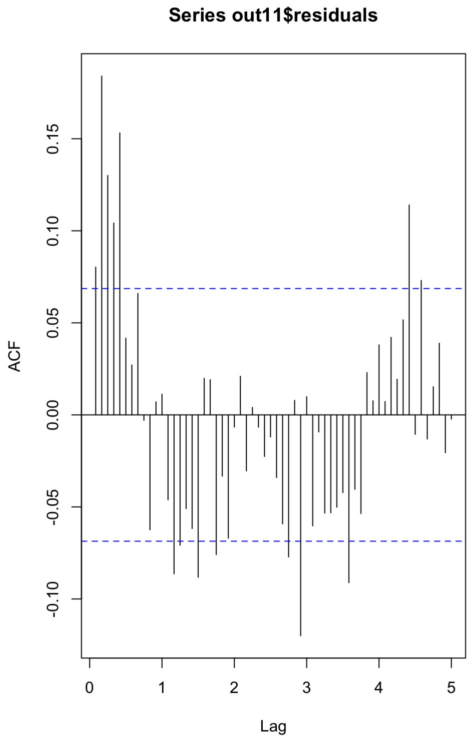
Model	AIC
SMA(1)$_{12}$	-2600.898
SAR(3)$_{12}$	-2576.051

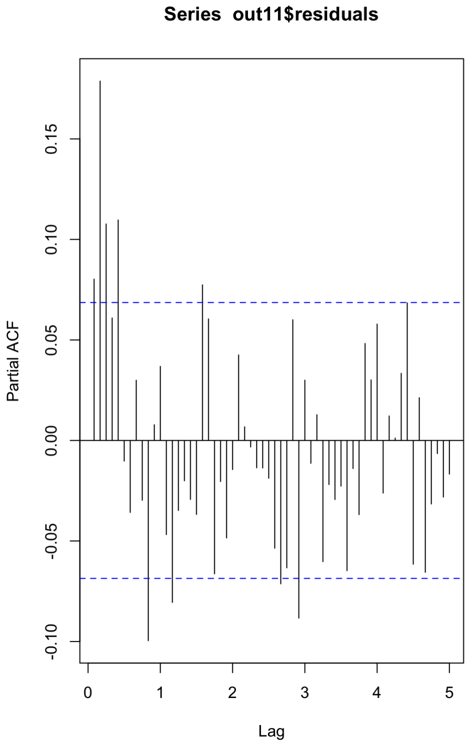
We move forward with Seasonal MA(1) as AIC is less than Seasonal AR(3).

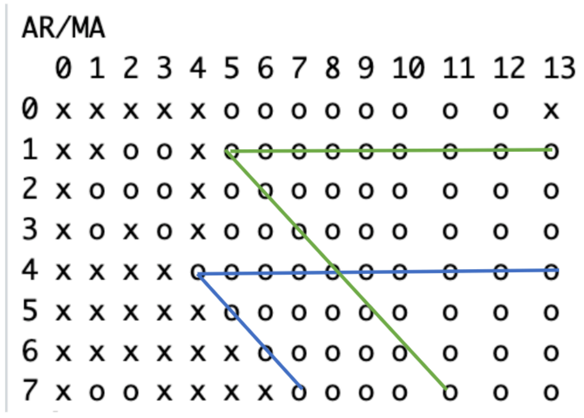


```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)
sma1 -0.739860   0.025523 -28.988 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For the SMA(1) ,intercept is not significant so we remove it.





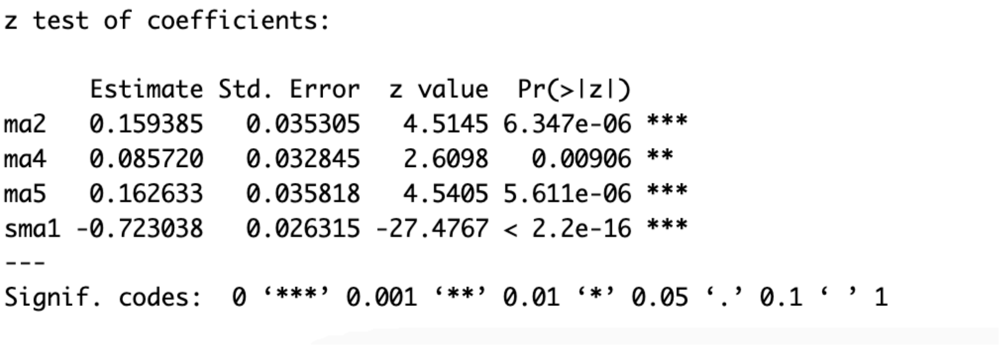


```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x o o o o o o  o  o  x
1 x x o o x o o o o o o  o  o  o
2 x o o o x o o o o o o  o  o  o
3 x o x o x o o o o o o  o  o  o
4 x x x x o o o o o o o  o  o  o
5 x x x x x o o o o o o  o  o  o
6 x x x x x x o o o o o  o  o  o
7 x o o x x x x o o o o  o  o  o
```

We look at the EACF of the residuals of the model to identify the regular order p,q. The plot suggests the two possible models:

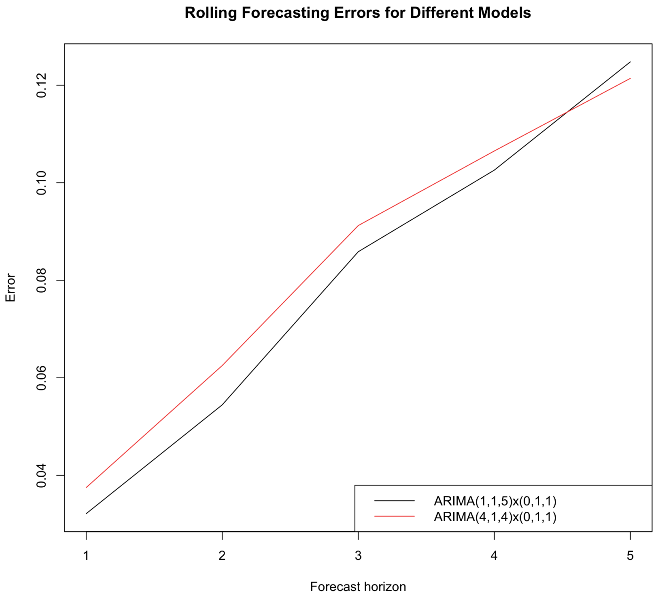
- ARIMA(1,0,5)
- ARIMA(4,0,4)

	ARIMA(1,1,5) x SARIMA$_{12}$(0,1,1)	ARIMA(4,1,4) x SARIMA$_{12}$(0,1,1)
AIC	**-2647.143**	-2646.773
Abs Polyroot	**1.383744 1.386822 1.386822 1.383744 1.669706**	1.106218 1.001000 1.001000 1.106218
Box L-Jung	0.1957	**0.2414**

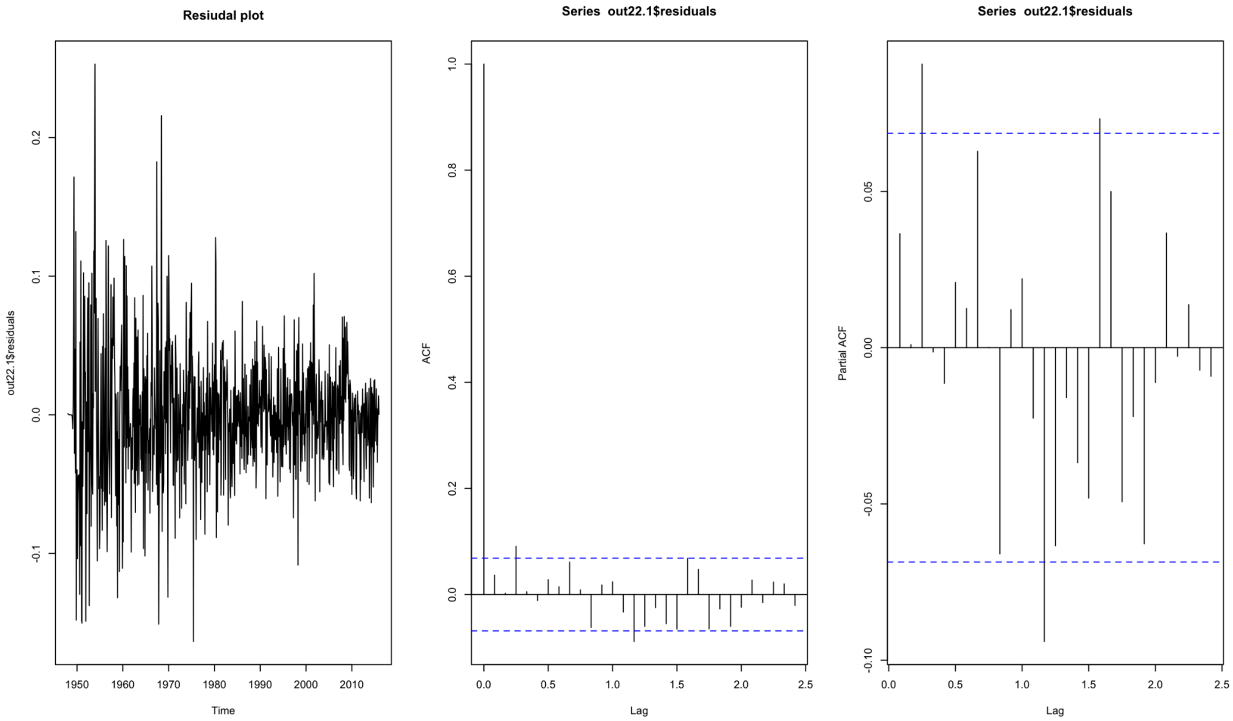


```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)    
ma2   0.159385   0.035305   4.5145 6.347e-06 ***
ma4   0.085720   0.032845   2.6098   0.00906 ** 
ma5   0.162633   0.035818   4.5405 5.611e-06 ***
sma1 -0.723038   0.026315 -27.4767 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

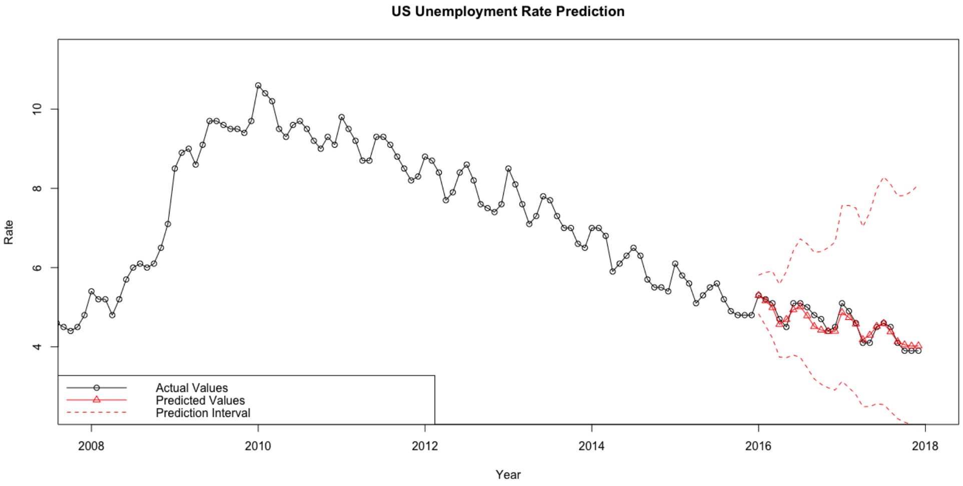
ARIMA(1,1,5) x SARIMA$_{12}$(0,1,1) is a better model since it has lower AIC and its roots are way greater than 1 signifying stationarity. The Box L-Jung method justifies that it behaves like white noise. In contrast, the other model has roots very close to 1 and worse AIC.



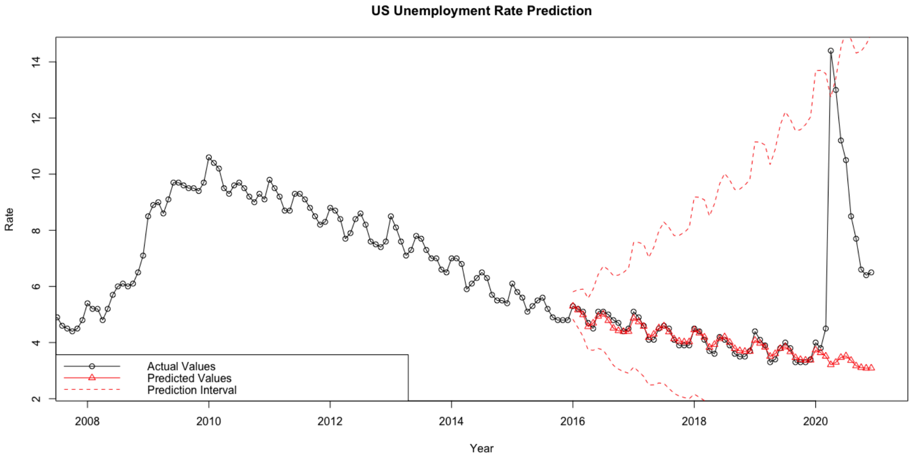
Since ARIMA(1,1,5) x SARIMA12(0,1,1) consistently has lower forecast errors than ARIMA(4,1,4) x SARIMA12(0,1,1) across the horizons, it is the better model based on the rolling forecast performance.



The residuals behave like white noise which is also verified through the Box Ljung test. After we have chosen our preferred final model we will use the model for the prediction.



We have predicted the data from 2016 to 2018. The performance of the model looks good and is a close match with actual data from test data.



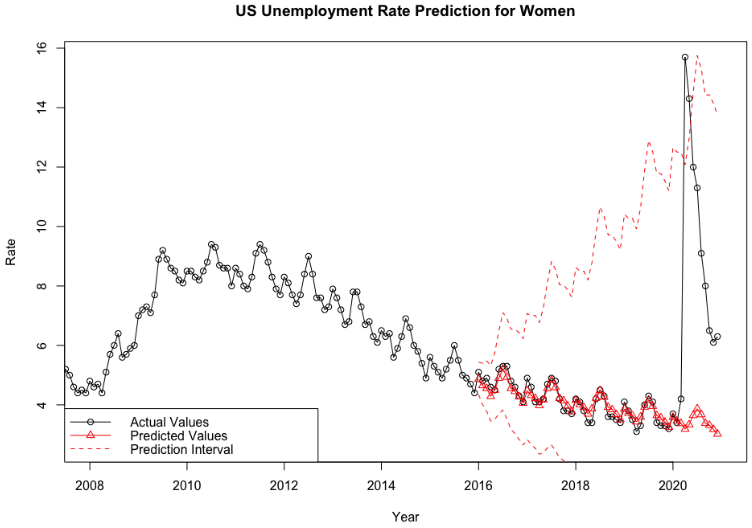
Our model does not predict well if the test data distribution itself contains outliers because of extreme situations like Covid-19. However, it looks almost within the upper boundary.

3.3. Unemployment Rate of Females

We carried out a female unemployment rate analysis similar to the analysis that we performed for the overall unemployment rate. The approach and the steps taken is same as above.The best chosen model for the female unemployment rate is

ARIMA(1,1,6) x SARIMA12(0,1,1)

The prediction given by the model looks like this:

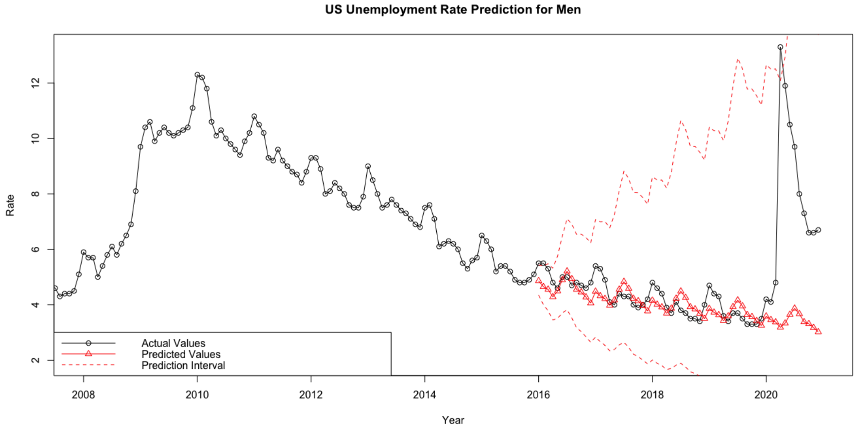


We observed that the order of the ARIMA model is slightly higher (order=6) compared to that of the overall unemployment rate. This is likely due to half year patterns in the female unemployment rate which was absent in the overall unemployment rate.

3.4. Unemployment Rate of Males

We carried out a male unemployment rate analysis similar to the analysis that we performed for the overall unemployment rate. The approach and the steps taken is same as above.The best chosen model for the male unemployment rate is ARIMA(6,1,6) x SARIMA12(0,1,1)

The prediction given by the model looks like this:



4. Conclusion

In this report, we did a detailed analysis of historical unemployment rates in the United States, focusing on state-specific, gender based and overall trends. Using time series techniques, like ARIMA and SARIMA models, we examined data spanning several years to identify patterns, seasonality, and factors influencing unemployment. In addition, we were able to demonstrate that the created model were able to predict future unemployment rates within confidence intervals.

The key findings highlighted disparities across states, with states like West Virginia exhibiting consistently high unemployment rates due to economic reliance on volatile industries, while states like Nebraska demonstrated stability with significantly lower rates. California, as a focal point, showcased the impact of regional economic hubs like Los Angeles on statewide unemployment trends.

The gender based unemployment rate and overall unemployment rate analyses revealed distinct patterns in unemployment rates between men and women, emphasizing the need for tailored economic policies to address unique challenges faced by different groups. The models developed in this study demonstrated robustness in forecasting under normal conditions but struggled to account for outliers caused by extraordinary events, such as the COVID-19 pandemic.

The report underlines the critical role of time series analysis in understanding and forecasting unemployment trends, providing valuable insights for policymakers, businesses, and researchers to address workforce challenges and plan for economic stability.

References

1. Congressional Research Service. (2020). *The coal industry: Historical trends and current status* (Report No. R46554). Retrieved from https://crsreports.congress.gov/product/pdf/R/R46554/9

2. Sayegh, E. (2024, August 19). *The great tech reset: Unpacking the layoff surge of 2024.* Forbes. Retrieved from https://www.forbes.com/sites/emilsayegh/2024/08/19/the-great-tech-reset-unpacking-the-layoff-surge-of-2024/

3. Wang, Q., & Li, L. (2021). The effects of population aging, life expectancy, unemployment rate, population density, per capita GDP, urbanization on per capita carbon emissions. *Sustainable Production and Consumption*, *28*, 760-774.

4. U.S. Bureau of Labor Statistics. (n.d.). *Local area unemployment statistics: Table note.* Retrieved December 3, 2024, from https://www.bls.gov/news.release/laus.tn.htm

Dataset 1

Torbenson, A. (n.d.). *Unemployment data (1948–2021) [Dataset]*. Kaggle. Retrieved from https://www.kaggle.com/datasets/axeltorbenson/unemployment-data-19482021/data

Dataset 2

Justin2028. (n.d.). *Unemployment in America per US state [Dataset]*. Kaggle. Retrieved from https://www.kaggle.com/datasets/justin2028/unemployment-in-america-per-us-state/data