Capstone Paper- Forecasting the Unemployment Rate

Connor Vaughan

Understanding the unemployment rate is important to both policymakers and the federal reserve. One of the FED’s primary goals when conducting monetary policy is to keep the unemployment rate low. It has had various noticeable long-term trends across demographics as well as short run fluctuations. Historically, there has been a large disparity between the unemployment rate of various demographics. For example, the Black unemployment rate has been significantly above the national average but has been getting closer over the past few decades. The goal of this paper is to use data analysis to examine the differences in the unemployment rate across demographics such as gender, race, and age as well as forecast future rates.

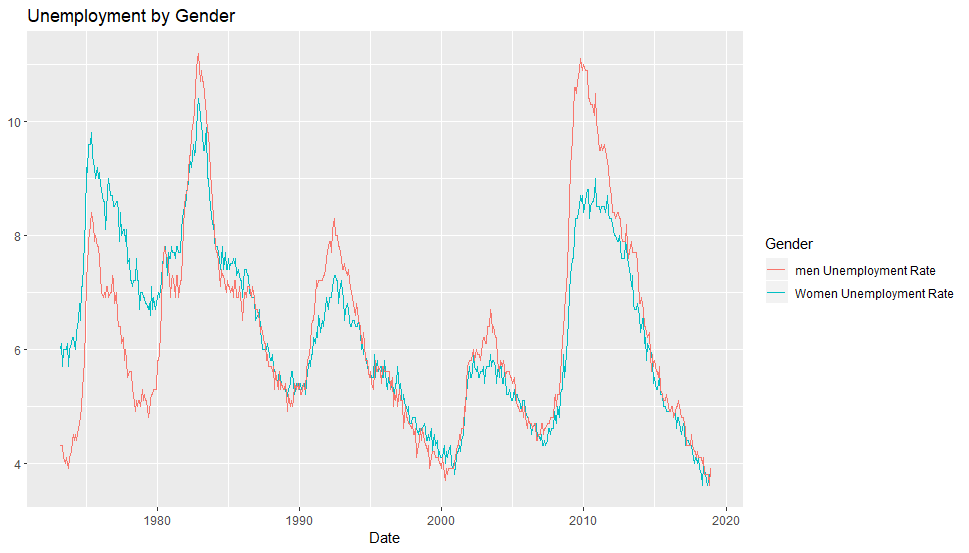
The first step in this project was to gather data on the unemployment rate going back as many years as possible. The Bureau of Labor Statistics has data going back to 1973 for all the demographics needed, except Asian which only has data back to 2003. The data has an unemployment rate for each month, which helps in identifying seasonal trends.

**Data Wrangling**

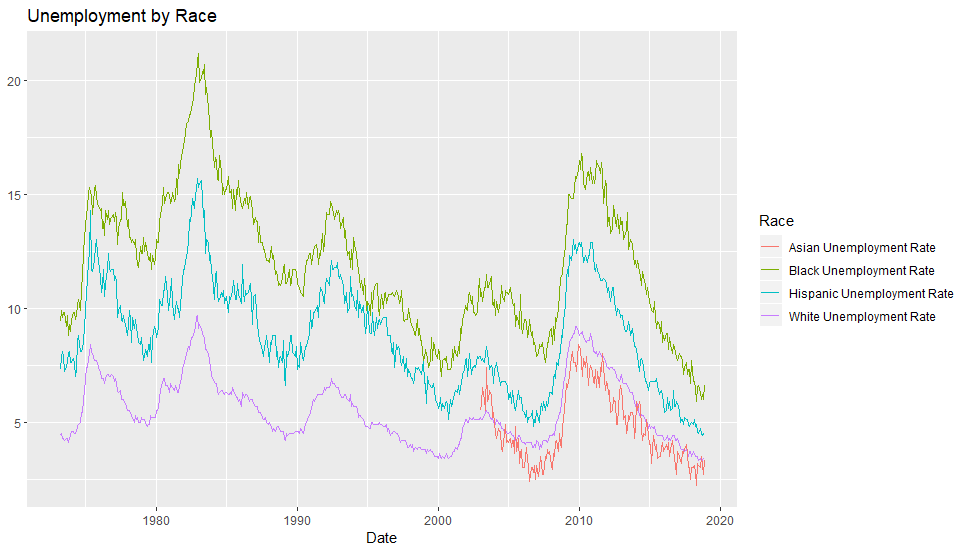
To begin the analysis the data was imported into R using the read.csv function and each dataset was assigned to a data frame by demographic. The data frames had a lot of unnecessary information at the top, so to remove this, each data frame was saved to a new data frame with only the needed rows. At this point the data frames had uninformative names, so each was given a descriptive name based on their demographic using the names function. Then all the data frames were combined into one data frame using the merge function. The Asian data frame couldn’t be combined with the merge function because the data started in 2003, so the left join function was used instead. Now that the data was all in one easy to read data frame, the next step was to look at descriptive statistics to better understand the data as well as create plots to visualize the data. It is interesting to see the trends in unemployment rate by demographic. For example, the female unemployment rate historically has been higher than male, but in the late 80s and early 90s the trend switched, and the male unemployment rate has remained higher. The primary challenge with wrangling the data was getting the data sets to merge properly. When I first tried to merge the data sets R was putting the entire column into 1 cell because I was using the wrong function. After troubleshooting I was able to find the right function, but this didn’t work for the Asian demographic because it only had data back to 2003. This required more troubleshooting and that’s when I found the left join function. One thing I learned from this project is that the majority of data science work is cleaning and wrangling the data. Often times data is not in a ready to use format and it takes a lot of time to get it into that state. Table 1 shows some descriptive statistics and plots 1-3 show unemployment trends over time.

|  |  |  |  |
| --- | --- | --- | --- |
| Demographic | Mean | Min | Max |
| Women | 6.35 | 3.6 | 10.4 |
| Men | 6.30 | 3.6 | 11.2 |
| Black | 11.97 | 5.9 | 21.2 |
| Hispanic | 8.82 | 4.4 | 15.7 |
| White | 5.54 | 3.3 | 9.7 |
| Asian | 4.78 | 2.2 | 8.4 |
| 16-19 | 17.86 | 12 | 27.2 |
| 20-24 | 10.30 | 6.5 | 17.2 |
| 25-34 | 6.22 | 3.5 | 11 |
| 35-44 | 4.67 | 2.4 | 9 |
| 45-54 | 4.14 | 2.1 | 8.1 |
| 55 & Over | 3.93 | 2.3 | 7.4 |

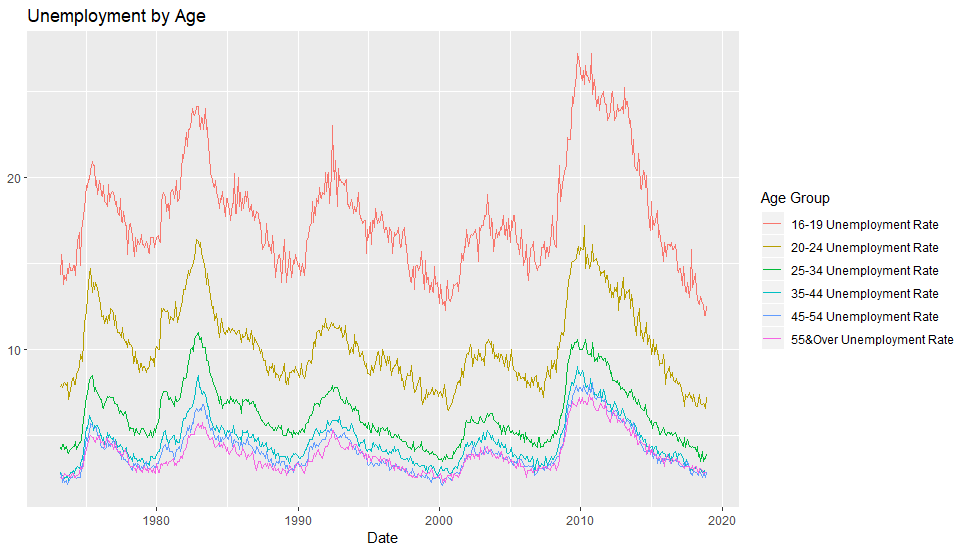
Table 1



Plot 1



Plot 2



Plot 3

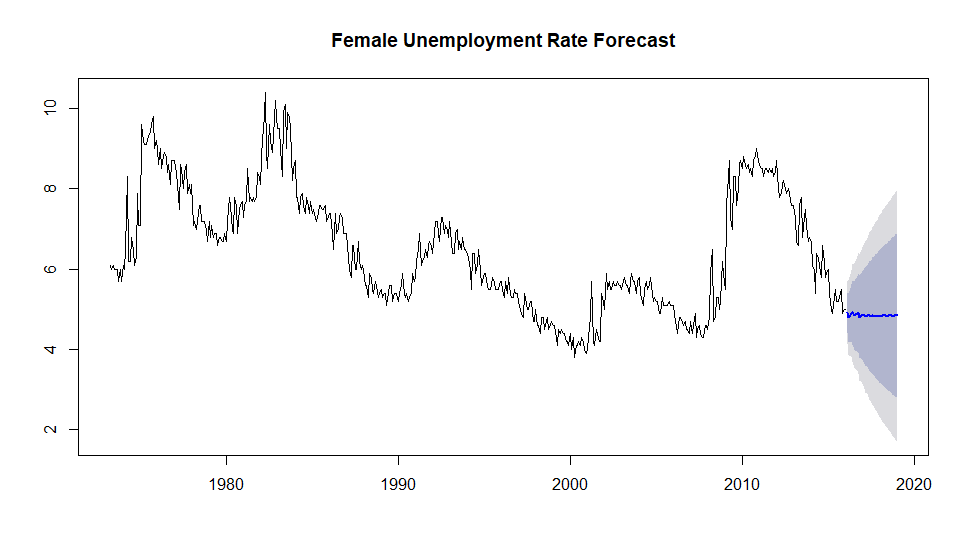
**Forecasting**

**Training**

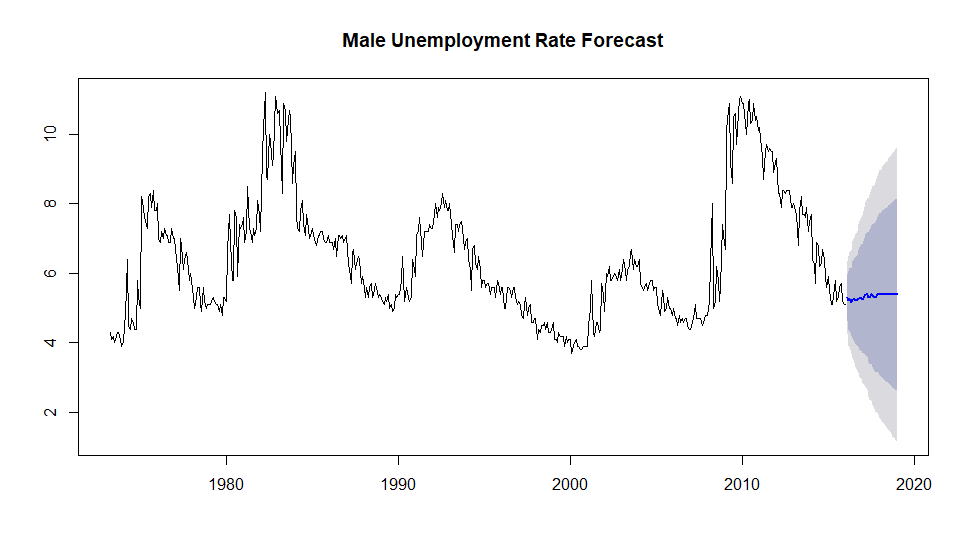
The previous section provided a good understanding of the trends the unemployment rate has displayed historically, but it’s important to predict where the unemployment rate is going. The primary tool used to do this was the ARIMA forecasting package. This package was chosen because the data is seasonal and has linear trends which allows ARIMA to create forecast models. To use this model the data was converted to time series then was used to create a training and test model. The data from 1973 to 2017 was used to train the model, which was then used to forecast future rates for 36 months.

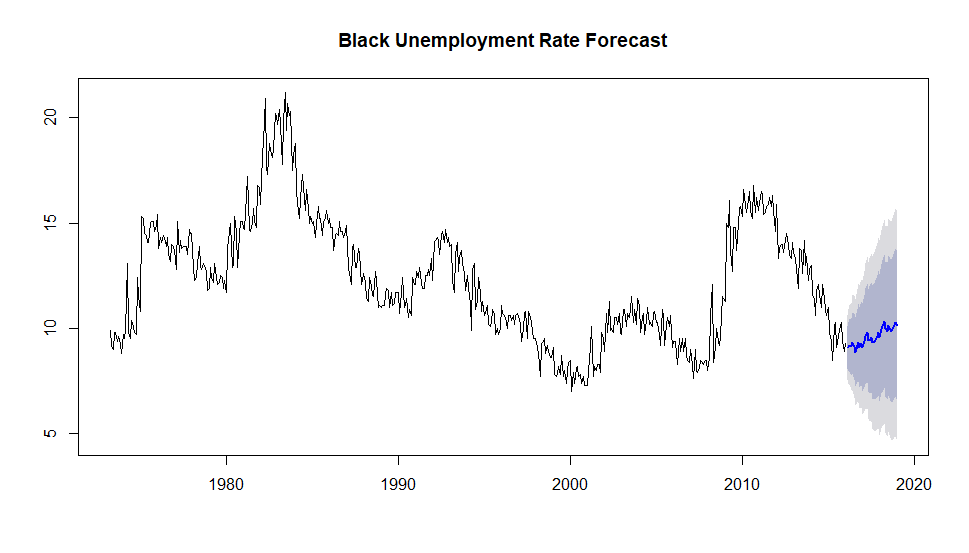
**Evaluating Accuracy**

To determine the accuracy of the model, it was used to predict values of the unemployment rate for each month in 2018, then these values were compared to the actual unemployment values for each month in 2018 to see how close the model got. The closer the model’s predictions are to the actual values, the more accurate the model is at forecasting. There are also statistical measures used to determine the accuracy of forecasting models, particularly MAPE and RMSE. MAPE is an acronym for mean absolute percent error, and it is found by taking the difference of the actual and the forecast then dividing by the actual and multiplying by 100, then take the mean of all those values. RMSE is an acronym for root mean square error and is found by taking the standard deviation of all the errors between the forecast and the actual values. Both these tools give an empirical measure on the quality of the machine learning model. Table 2 shows the MAPE values for the training and test set as well as the RMSE values. (Dalinina) MAPE is an acronym for mean absolute percent error, and it is found by taking the difference of the actual and the forecast then dividing by the actual and multiplying by 100, then take the mean of all those values. RMSE is an acronym for root mean square error and is found by taking the standard deviation of all the errors between the forecast and the actual values. While the MAPE values are higher than preferable, they do still predict the future unemployment rates with about 70-80% accuracy. Part of the reason they are so high is because this model doesn’t account for the recessionary periods in the training set that led to large increases in the unemployment rate across all demographics. This model also suffers from overfitting where it follows the training data set too closely which makes it difficult for it to create accurate predictions. Plots 4-16 show visually what the forecasting model predicts the unemployment rate will do for each demographic. The model predicts the unemployment rate will rise for every demographic in the following months, which intuitively makes sense because the unemployment rate is lower than it has historically been.



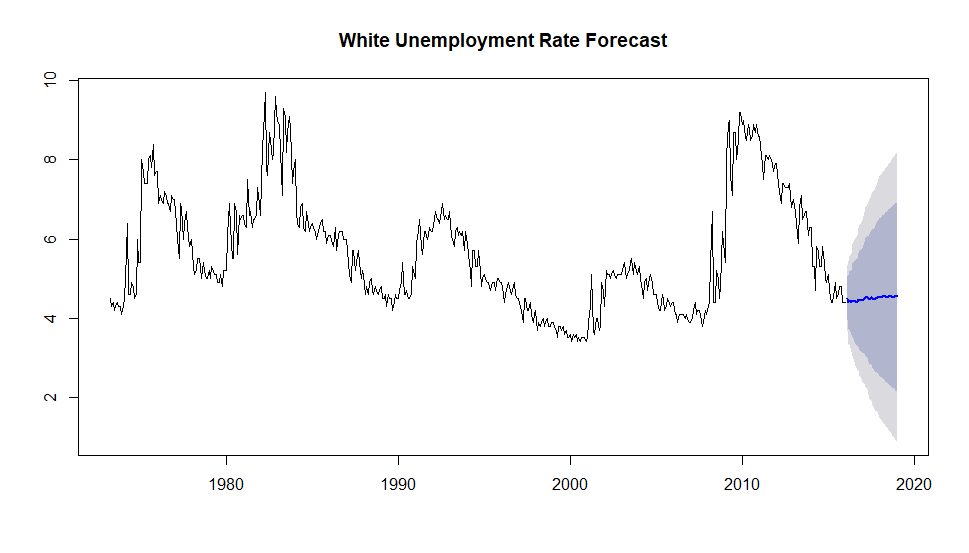
Plot 4

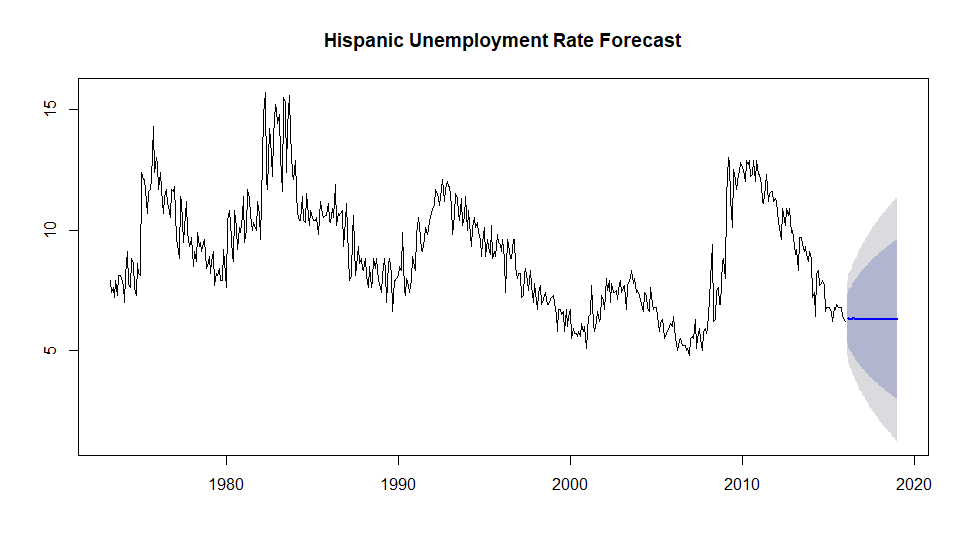




Plot 5

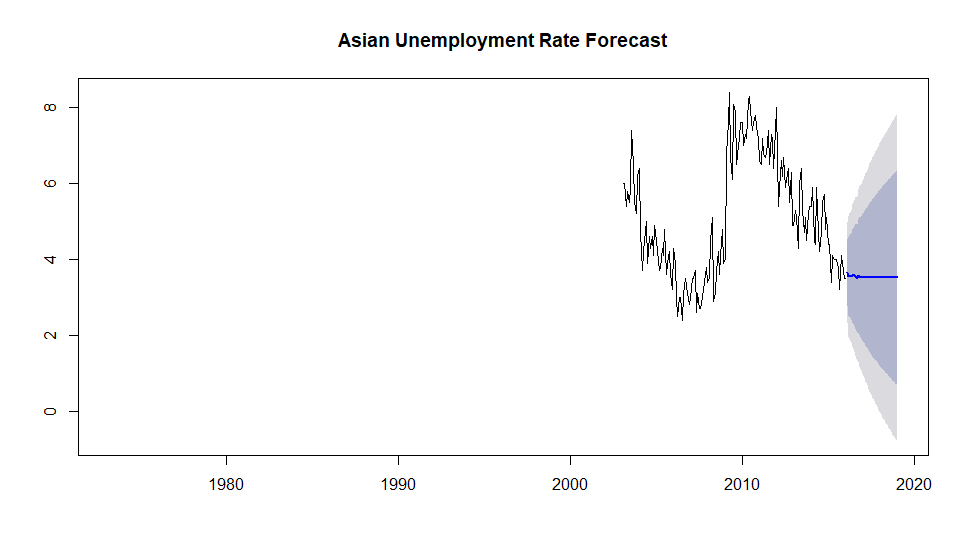
Plot 6

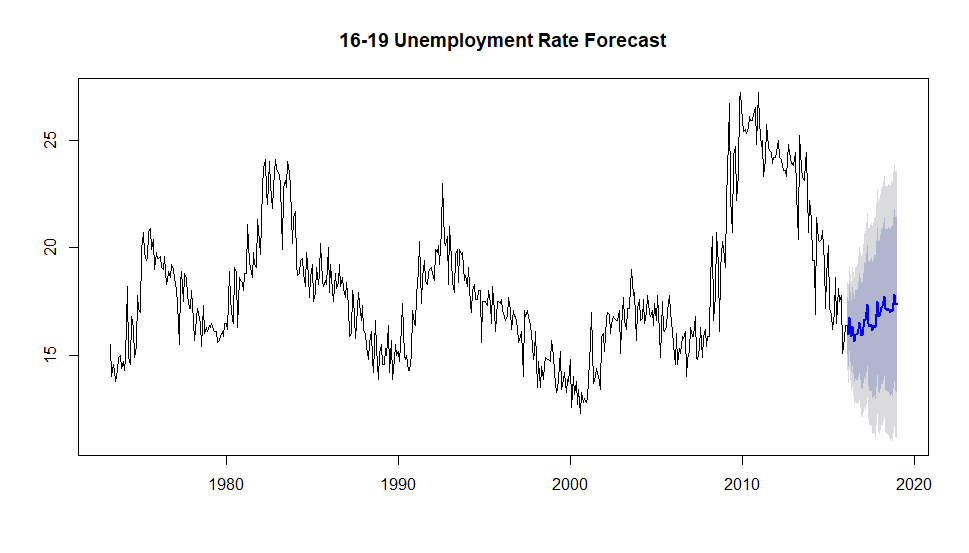




Plot 8

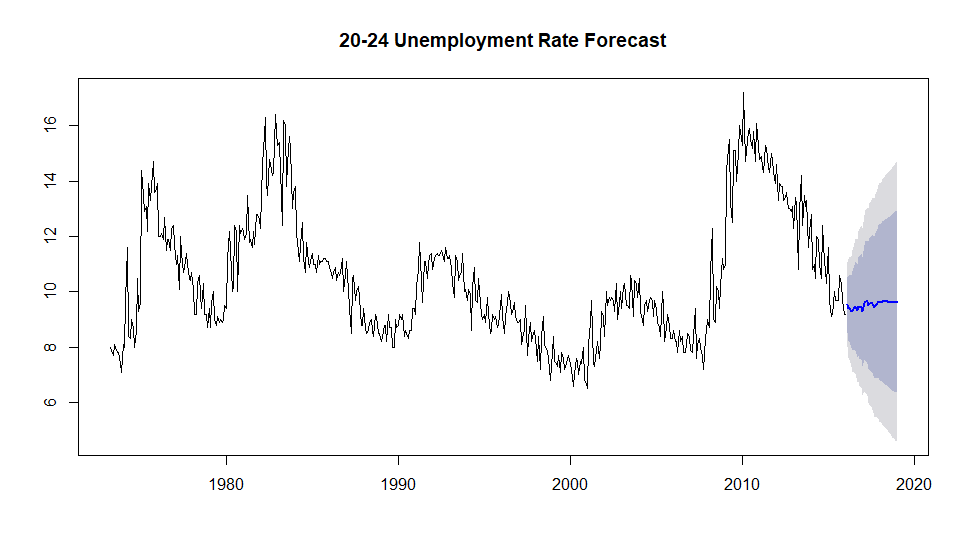
Plot 7



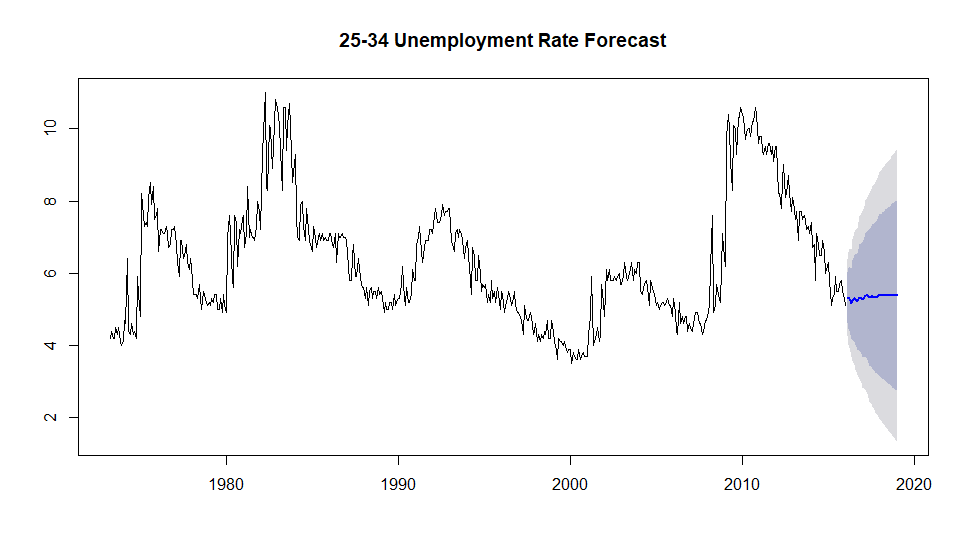


Plot 10

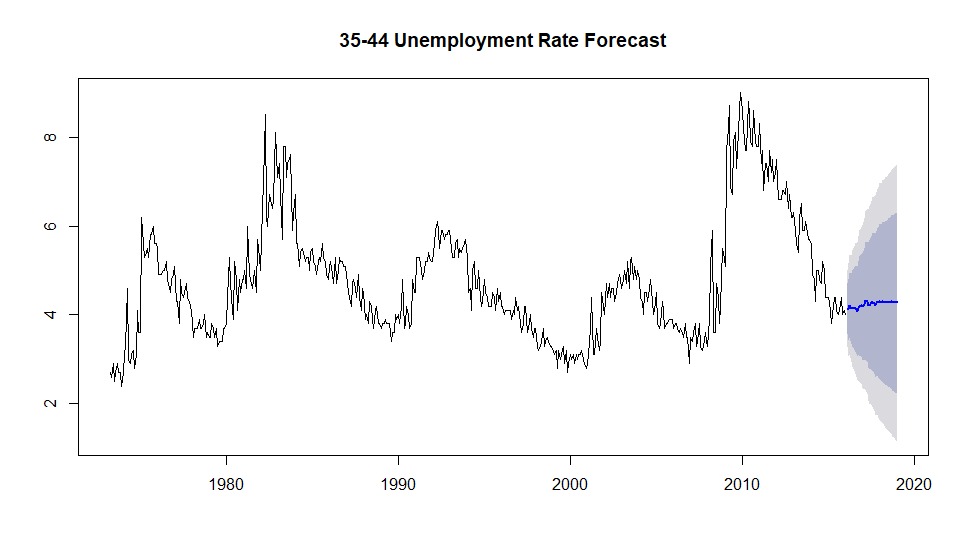
Plot 9



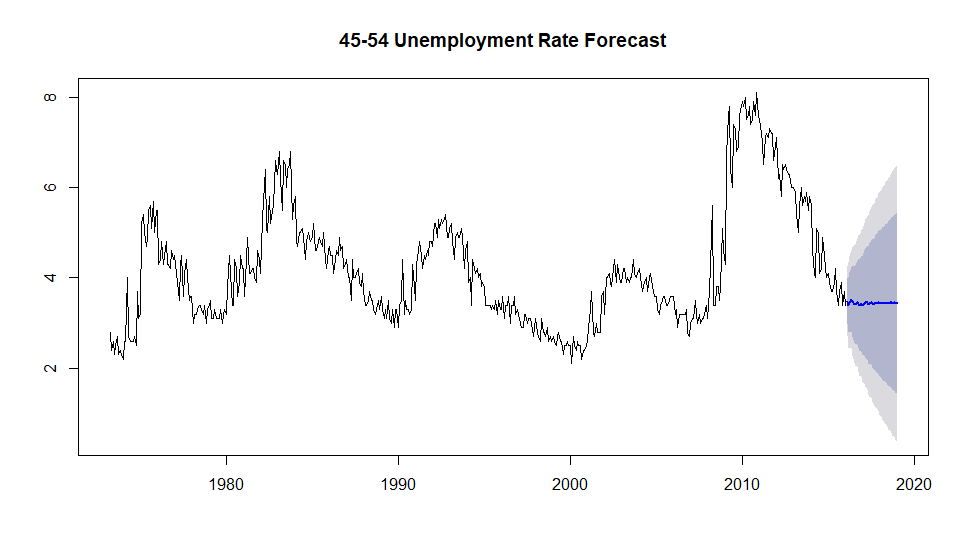
Plot 11



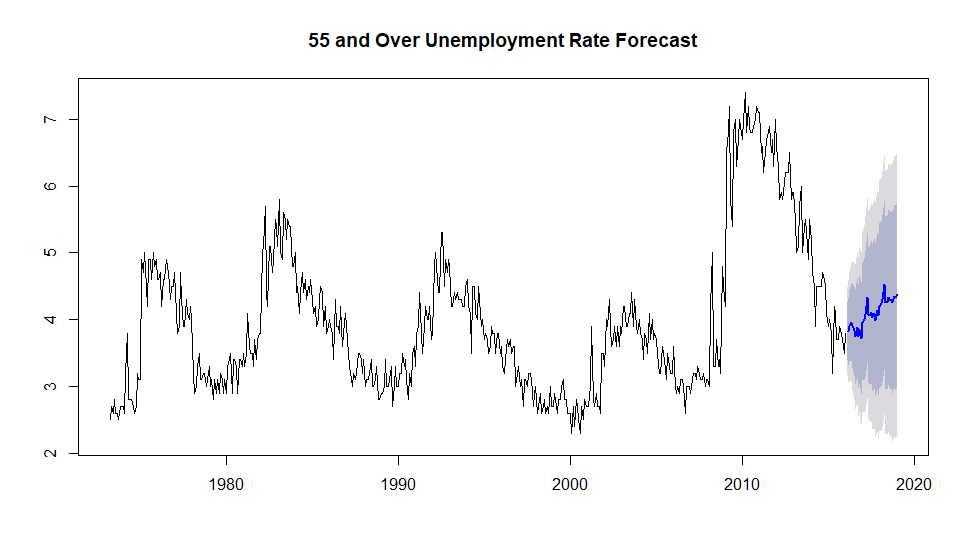
Plot 13



Plot 14



Plot 15



Plot 16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Demographic | Training MAPE | Test MAPE | Training RMSE | Test RMSE |
| Women | 4.06 | 13.72 | 0.38 | 0.67 |
| Men | 4.95 | 22.07 | 0.52 | 1.04 |
| Black | 4.85 | 30.83 | 0.77 | 2.46 |
| Hispanic | 6.21 | 22.72 | 0.78 | 1.23 |
| White | 4.53 | 17.95 | 0.419 | 0.77 |
| Asian | NA | 13.09 | 0.66 | 0.49 |
| 16-19 | 4.35 | 18.93 | 1.04 | 3.05 |
| 20-24 | 5.20 | 27.21 | 0.77 | 2.12 |
| 25-34 | 5.44 | 19.94 | 0.53 | 0.99 |
| 35-44 | 6.17 | 26.01 | 0.44 | 0.94 |
| 45-54 | 6.50 | 11.82 | 0.39 | 0.43 |
| 55 & Over | 5.94 | 25.94 | 0.33 | 0.93 |

Table 2

**Conclusion**

Throughout this project I have developed skills that will enable me to better analyze data in the future. The tools that I learned to utilize in R enable me to do tasks that other statistical programs are unable to do. If I were to continue to work on this project, I would try to incorporate the recessions into the training model which would allow it to create more accurate predictions. The insights from this project are particularly useful to policymakers who make laws and regulations that could potentially impact the labor force. Understanding where the unemployment rate is likely going for each demographic is an important consideration. This model predicts that for every demographic the unemployment rate will rise over the following months, which seems reasonable considering how low it currently is. This is also useful to the FED who conduct their policy based on current and expected future unemployment rates. These results could potentially also be used by businesses and investors because the unemployment rate is often considered one of the best proxies for macroeconomic conditions. One of the primary skills I developed through this project is problem solving issues with R. Rarely did I try I function or line of code and have it work first try, which required me to use other resources to solve the issue. One of the advantages of working with computer software is that many people post questions and solutions to problems online. I found that when I ran into a problem while working on this project, there were many other people who had similar problems and there would usually be multiple forums that helped me resolve it. In conclusion, I feel that this project has taught me skills that will be incredibly valuable throughout my career.

References

Dalinina, Ruslana. “Introduction to Forecasting with ARIMA in R.” *Oracle + DataScience.com*, www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials.

“Databases, Tables & Calculators by Subject.” *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, [www.bls.gov/data/#unemployment](http://www.bls.gov/data/#unemployment).

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Parker, Jeffery. *Models of Unemployment*, www.reed.edu/economics/parker/s11/314/book/Ch14.pdf.