Jillian Sylvester
Professor Ding
DATS 2103
December 11th, 2024

1974–2024:
Fluctuations in American Unemployment and Its Economic Drivers

Question 1: Preparing the Dataset

After downloading the excel files from the Bureau of Labor Statistics, I noticed certain factors about the data sets that helped me decide how to move forward with pairing them together like what year the data started being collected, the unit (monthly, annually, quarterly), and whether it was seasonally adjusted or not.

Datasets I Created:

df_NSA (Not Seasonally Adjusted)

- **All_Items_Urban_Consumers**: Consumer Price Index for all goods and services purchased by urban consumers.
- **Urban_Wage_Earners_Clerical_Workers**: CPI for goods and services purchased by urban wage earners and clerical workers.
- **All_Urban_Consumers_Less_Food_Energy**: Core CPI for urban consumers, excluding food and energy.
- **Urban_Wage_Earners_Clerical_Workers_Less_Food_Energy**: Core CPI for urban wage earners and clerical workers, excluding food and energy.
- **PPI_Final_Demand**: Producer Price Index for final demand, tracking price changes received by producers.
- **PPI_Final_Demand_Less_Food_Energy**: Core PPI for final demand, excluding food and energy.
- **PPI_Final_Demand_Less_Food_Energy_Trade_Services**: Core PPI excluding food, energy, and trade services.
- **PPI_Final_Demand_Finished_Goods**: PPI for finished goods, representing price changes for goods ready for sale.

df_SA_start1974 - Merged seasonally adjusted data from 1974-2024, collected monthly.

- **Civilian_Labor_Force**: Total civilian labor force, including those employed and actively seeking employment (in thousands).
- **Employment_Level**: Number of employed individuals (in thousands).
- **Unemployment_Level**: Number of unemployed individuals actively seeking work (in thousands).
- **Unemployment_Rate**: Unemployment percentage of the civilian labor force.
- **Employees_Nonfarm**: Number of employees in non farm sectors (in thousands).

- **Avg_Weekly_Hours_Production_Nonsupervisory_Employees_Private**: Average weekly hours worked by production and nonsupervisory employees in the private sector.
- **Avg_Hourly_Earnings_Production_Nonsupervisory_Employees_Private**: Average hourly wages of production and nonsupervisory employees in the private sector.

df_combined - Converted df_SA_start1974 to quarterly and merged with other quarterly datasets.

- Includes all columns from df_SA_start1974 plus…
- **Labor_Productivity_Nonfarm**: Output per hour worked in the nonfarm business sector, reflecting efficiency.
- **Labor_Unit_Costs_Nonfarm**: Cost of labor per unit of output in the nonfarm sector.
- **Real_Hourly_Compensation_Nonfarm**: Inflation-adjusted hourly compensation for nonfarm workers.

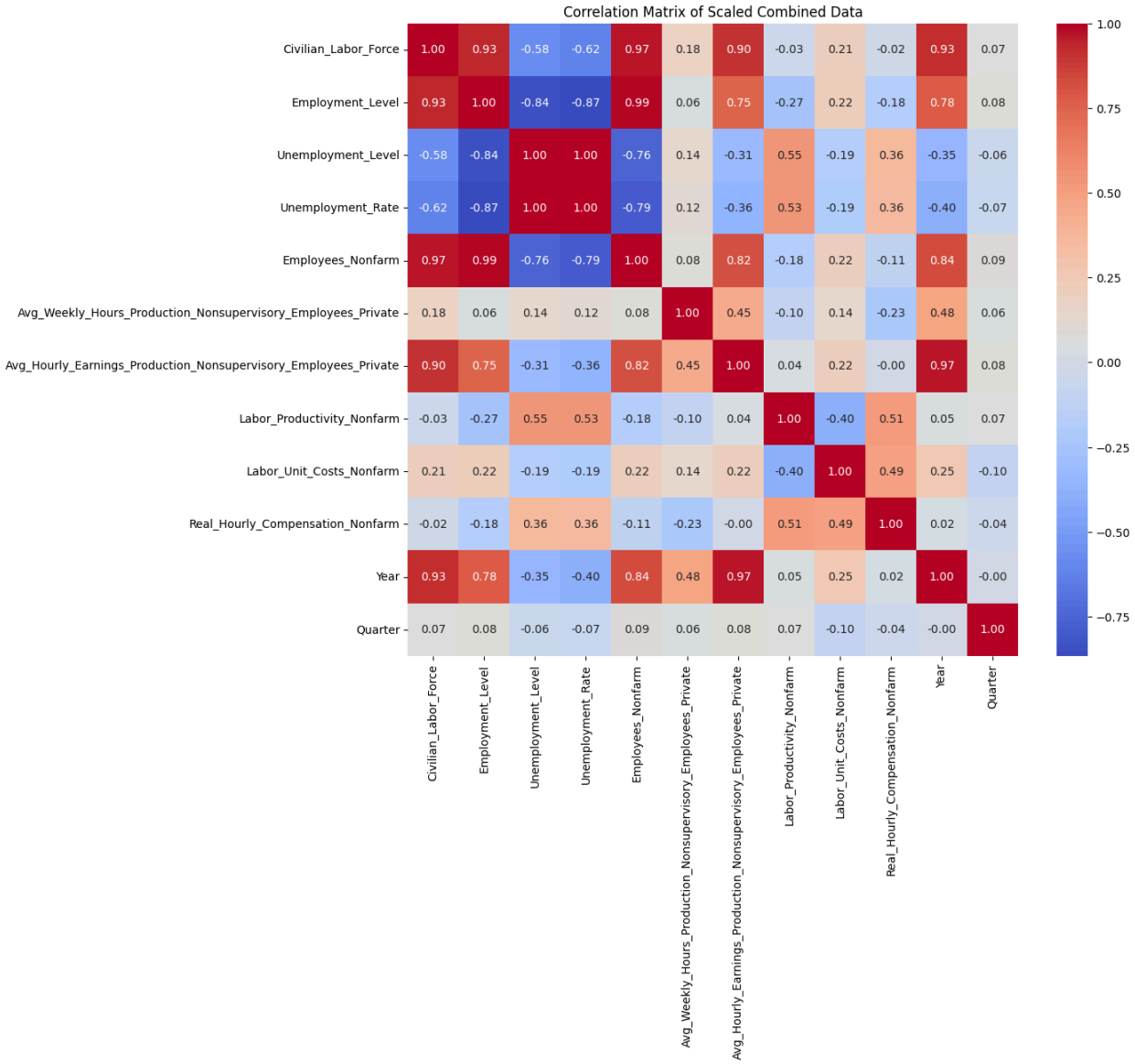
I decided to scale them to help reduce bias, improve my model performance, and make it easier to interpret my analysis.

Scaled Data Exploration:

Scaled_df_combined correlation matrix notable insights:

- **Nonfarm labor productivity** - negative correlation with **employment level** (-0.275) and **employment nonfarm** (-0.181). This COULD MEAN higher employment levels might not always align with increased labor productivity
- **Nonfarm labor** productivity - positively correlated with **Unemployment Leve**l (0.547) and **Unemployment Rate** (0.531). This COULD MEAN increased labor productivity could be connected with unemployment increases.
- Small positive correlation between **Average Hourly Earnings for Production and Nonsupervisory Employees** and **Labor Productivity** (0.044). This COULD MEAN that increasing wages could increase productivity
- Negative correlation between **Labor Unit Costs** and **Labor Productivity** (-0.404). This COULD MEAN that higher labor costs reduce labor productivity.
- Moderate positive correlation between **Labor Unit Costs** and **Real Hourly Compensation** (0.493). This COULD MEAN that higher labor costs might increase compensation

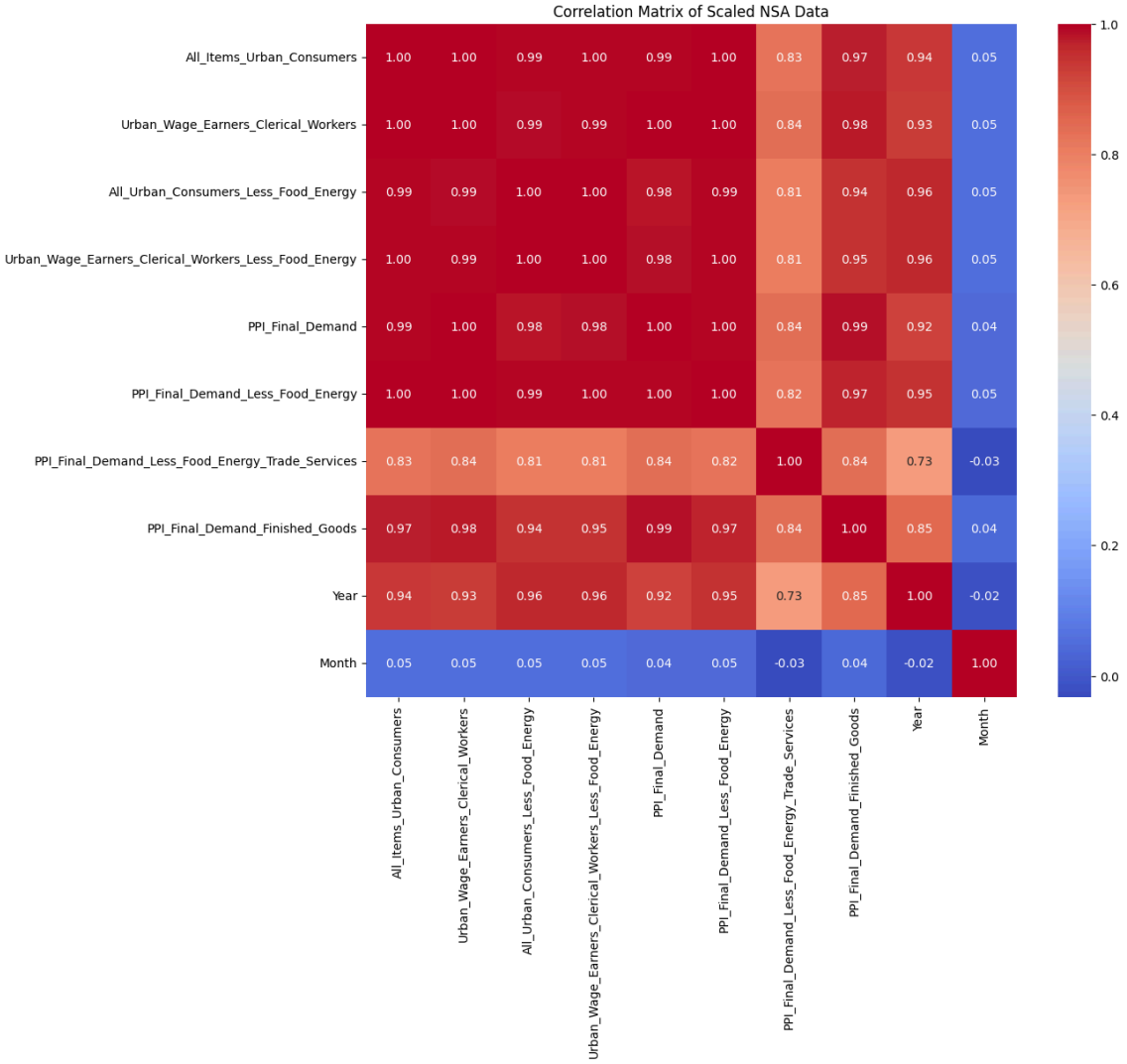
Figure 1: Correlation Matrix of Scaled Combined Data



Scaled_df_NSA correlation matrix notable insights:

- As seen in the heatmap, similar economic variables are highly correlated
- The correlations between the consumer price indices (e.g., **All_Items_Urban_Consumers** and **PPI_Final_Demand**) and the producer price indices (e.g., "PPI_Final_Demand_Less_Food_Energy") are strong. This could mean that changes in producer prices strongly impact consumer prices.
- The strong correlation between **PPI_Final_Demand** and **PPI_Final_Demand_Less_Food_Energy** (0.995) emphasizes how food and energy prices play a significant role in shaping overall producer prices.

Figure 2: Correlation Matrix of Scaled NSA Data

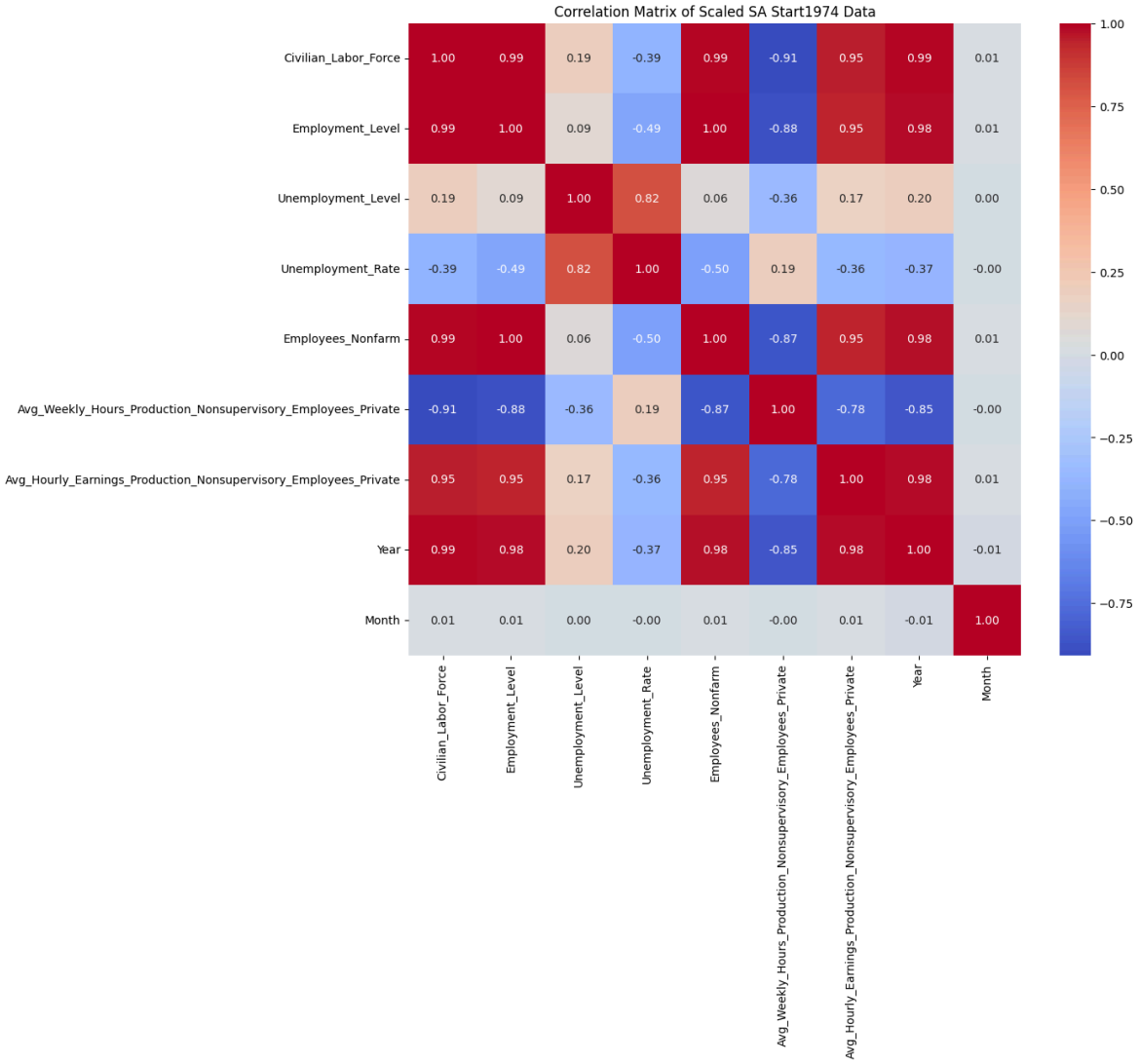


Scaled Seasonally adjusted data that starts at 1974:

- Strong correlations:
  - Civilian_Labor_Force and Employment_Level have a very high correlation (0.994).
  - Employees_Nonfarm and Employment_Level are also highly correlated (0.998).
  - Avg_Hourly_Earnings_Production_Nonsupervisory_Employees_Private and Employment_Level have a strong positive correlation (0.946808).
  - Similarly, Avg_Weekly_Hours_Production_Nonsupervisory_Employees_Private and Civilian_Labor_Force have a negative correlation (-0.907087).
- Weaker correlations:

- Unemployment_Level and Avg_Weekly_Hours_Production_Nonsupervisory_Employees_Private show a moderate negative correlation (-0.355).
- Month and other variables (except Year) show very low correlations (close to 0), which suggests they do not significantly influence the other variables.

Figure 3: Correlation Matrix of Scaled SA Data From 1974 to 2024



VIF Analysis: exploring multicollinearity

scaled_df_NSA:

	feature	VIF
0	const	3.386638e+07
1	All_Items_Urban_Consumers	1.672085e+05
2	Urban_Wage_Earners_Clerical_Workers	1.122757e+05
3	All_Urban_Consumers_Less_Food_Energy	1.036945e+05
4	Urban_Wage_Earners_Clerical_Workers_Less_Food_...	6.888476e+04
5	PPI_Final_Demand	1.925902e+04
6	PPI_Final_Demand_Less_Food_Energy	1.497881e+04
7	PPI_Final_Demand_Less_Food_Energy_Trade_Services	4.746157e+00
8	PPI_Final_Demand_Finished_Goods	1.566959e+03
9	Year	9.793360e+01
10	Month	1.560447e+00

Scaled_df_combined:

	feature	VIF
0	const	7.117927e+07
1	Civilian_Labor_Force	1.893032e+08
2	Employment_Level	4.383412e+08
3	Unemployment_Level	9.240325e+07
4	Unemployment_Rate	1.283624e+04
5	Employees_Nonfarm	3.497566e+02
6	Avg_Weekly_Hours_Production_Nonsupervisory_Emp...	1.596638e+01
7	Avg_Hourly_Earnings_Production_Nonsupervisory_...	6.361379e+01
8	Labor_Productivity_Nonfarm	1.586026e+01
9	Labor_Unit_Costs_Nonfarm	1.715756e+01
10	Real_Hourly_Compensation_Nonfarm	1.633352e+01
11	Year	2.010206e+02
12	Quarter	2.802447e+00

scaled_df_SA_start1974:

	feature	VIF
0	const	5.461330e+06
1	Civilian_Labor_Force	3.519592e+09
2	Employment_Level	3.412230e+09
3	Unemployment_Level	4.239045e+07
4	Unemployment_Rate	7.072304e+01
5	Employees_Nonfarm	4.071064e+02
6	Avg_Weekly_Hours_Production_Nonsupervisory_Emp...	1.890129e+01
7	Avg_Hourly_Earnings_Production_Nonsupervisory_...	7.880890e+01
8	Year	2.941357e+02
9	Month	1.127503e+00

Insights: This VIF analysis shows there is a lot of multicollinearity across the data frames. In scaled_df_NSA, All_Items_Urban_Consumers and Urban_Wage_Earners_Clerical_Workers had extremely high VIF values, meaning strong correlation with other features. In scaled_df_combined, Civilian_Labor_Force and Employment_Level showed very high VIFs, suggesting a high degree of multicollinearity. In scaled_df_SA_start1974, Civilian_Labor_Force and Employment_Level again had

excessively high VIFs. According to the textbook, when faced with this high multicollinearity, you can either choose to drop some features or can combine them using methods like PCA.

Question 2: Define Your Problem

What factors contribute to the fluctuations in the unemployment rate in the United States, and how do economic indicators such as labor force participation, employment levels, wages, and labor productivity influence unemployment trends?

General Trend Graphs for Reference:

Figure 4: Civilian Labor Force Level Over Time

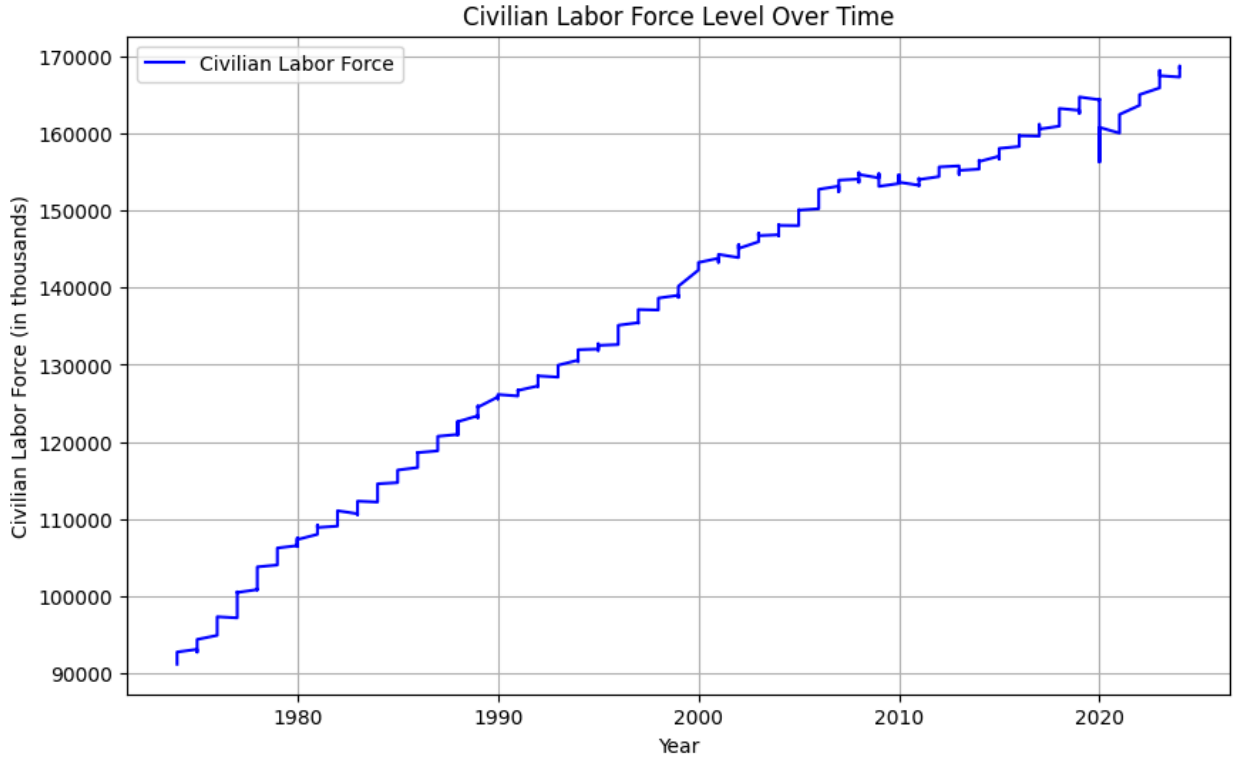
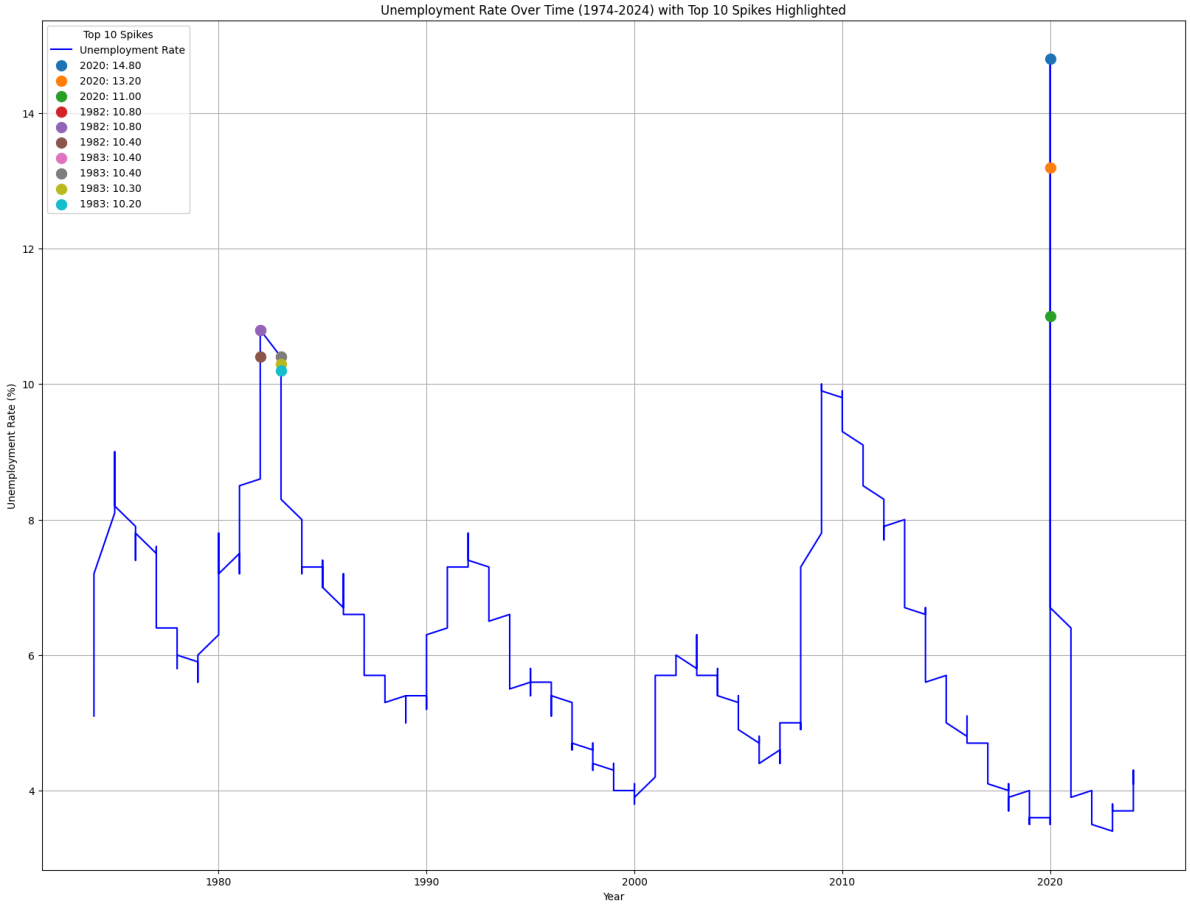


Figure 5: Unemployment Rate Over Time (1974-2024) with Top 10 Spikes



Question 3: Model Selection and Justification

Because there are not a lot of features in the SA data set from 1974 to 2024, I am going to do some analysis across the 50 year span and the shorter dataset with more features.

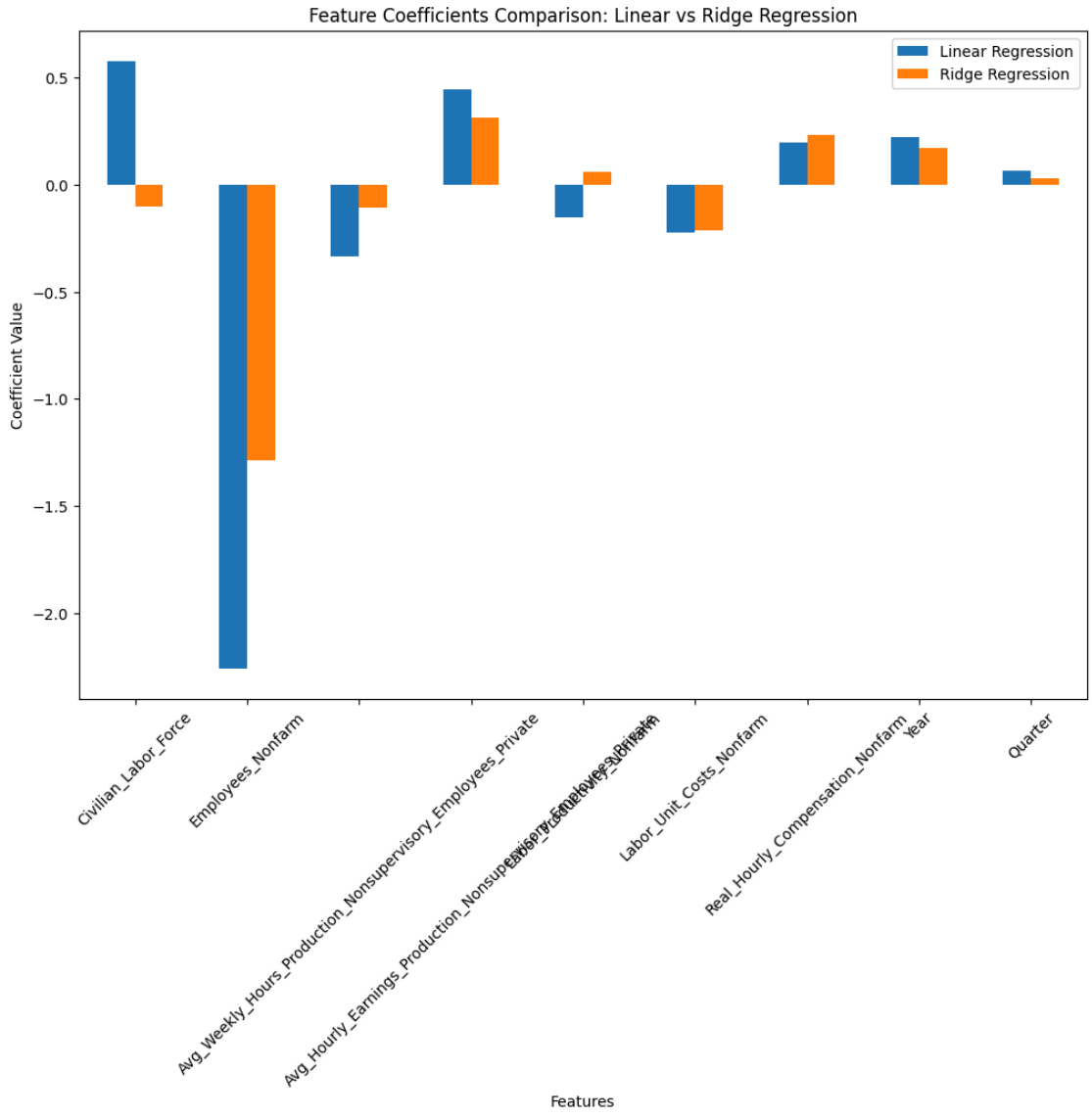
2013-2024: Using the Combined Data Set with more Features

Due to the high multicollinearity, I am going to start with Linear Regressions (as a model to compare to) and Ridge Regression on scaled_df_combined. I am avoiding Best Subset Regression because it does not handle multicollinearity as well. I am also nervous Lasso will drop too many variables from my model. Ridge regression shrinks coefficients without setting them to zero. For the sake of having a better model, I am going to drop the unemployment level and employment level as they are an obvious predictor of the unemployment rate and give me little valuable insight.

Linear Regression and Ridge Regression:

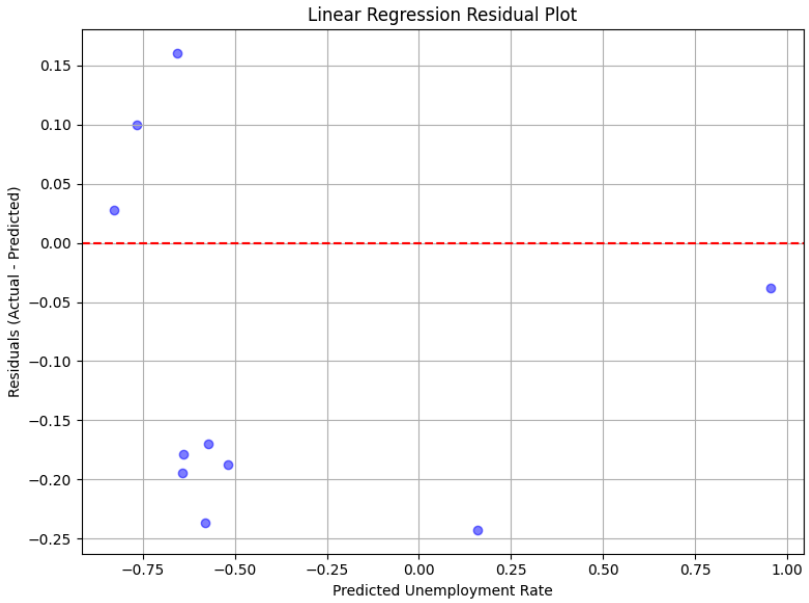
```
Linear Regression Mean Squared Error: 0.02868663596116757
Ridge Regression MSE: 0.06175886983652897
```

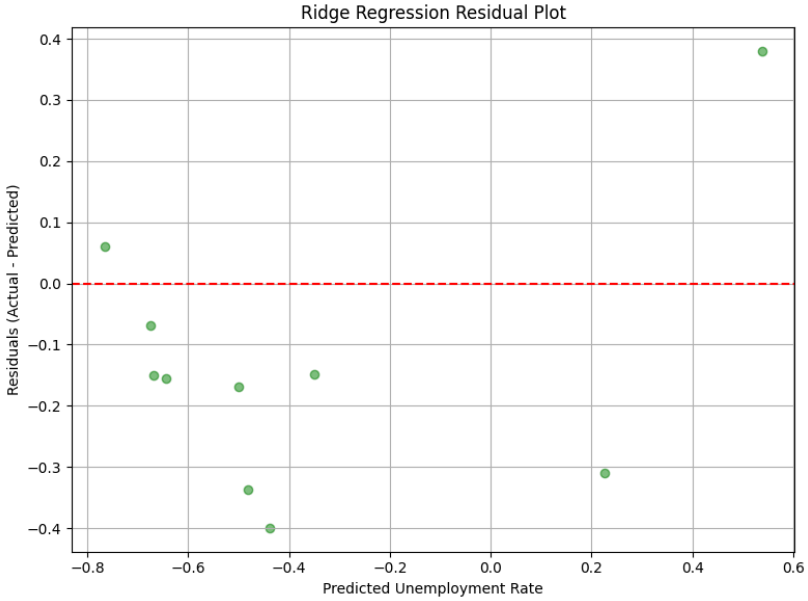
Figure 6: Feature Coefficients Comparison: Linear vs Ridge Regression (2013-2024)



While the linear regression model has a lower MSE, Ridge Regression produced smaller coefficients when it regularized the model, which can help avoid overfitting, especially due to all of the multicollinearity in the data set. I am still concerned about multicollinearity in the model, so I am going to perform PCA because it produces a low-dimensional representation of the dataset, finding a sequence of linear combinations that have maximal variance but are mutually uncorrelated.

Figure 7: 2013-2024 Linear Regression Residual Plot
Figure 8: 2013-2024 Ridge Regression Residual Plot

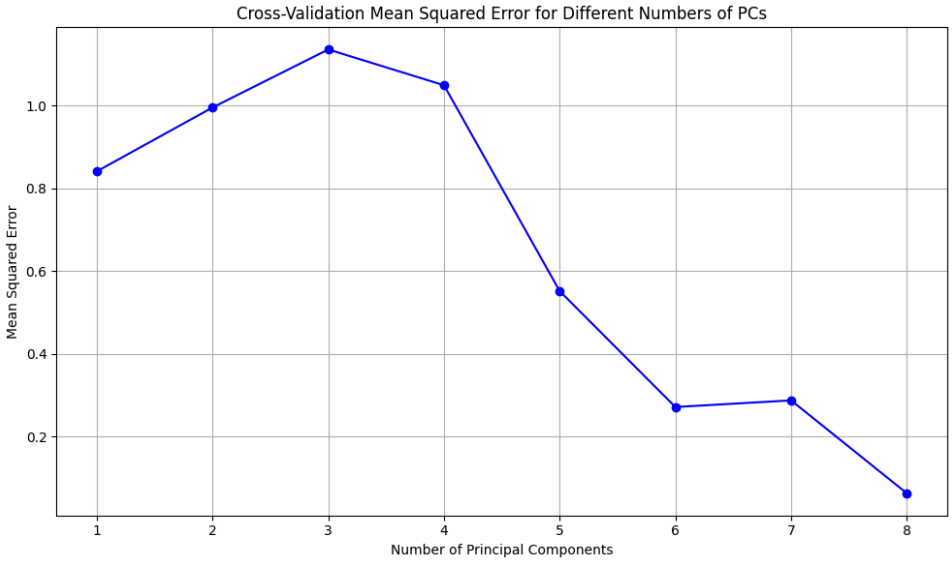




Principal Component Analysis:
I used cross validation to choose the number of components by using skm.GridSearchCV.

Best number of components: 8
Best CV MSE: 0.0637116091619658

Figure 9: 2013 CV MSE for Different Numbers of PCs



I fit the model with 8 components:

Mean Squared Error (MSE) on Test Set: 0.04706214976971712
R-squared on Test Set: 0.826798265542752

Figure 10: 2013-2024 PCA and Linear Regression Plot

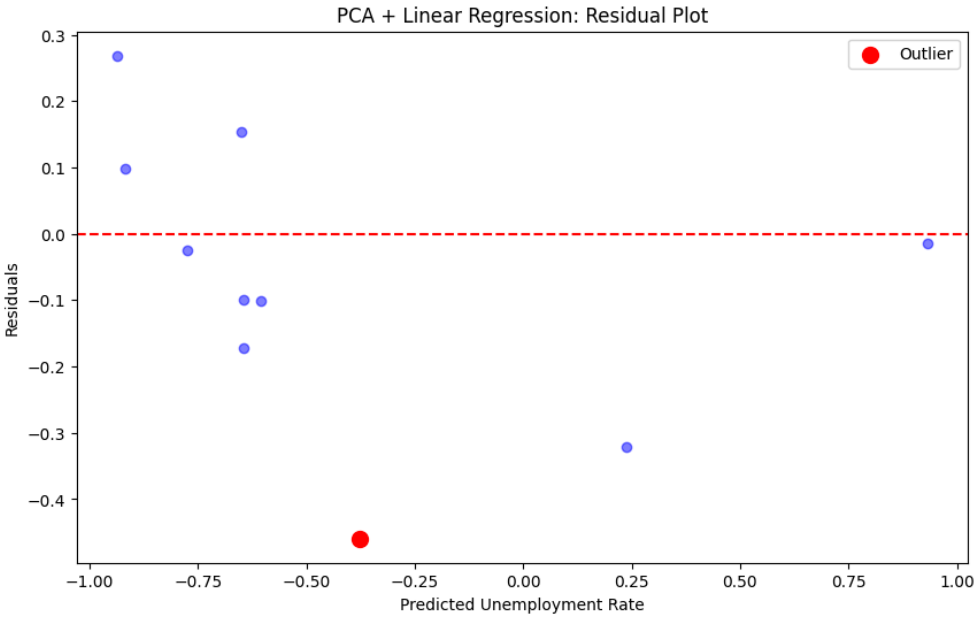
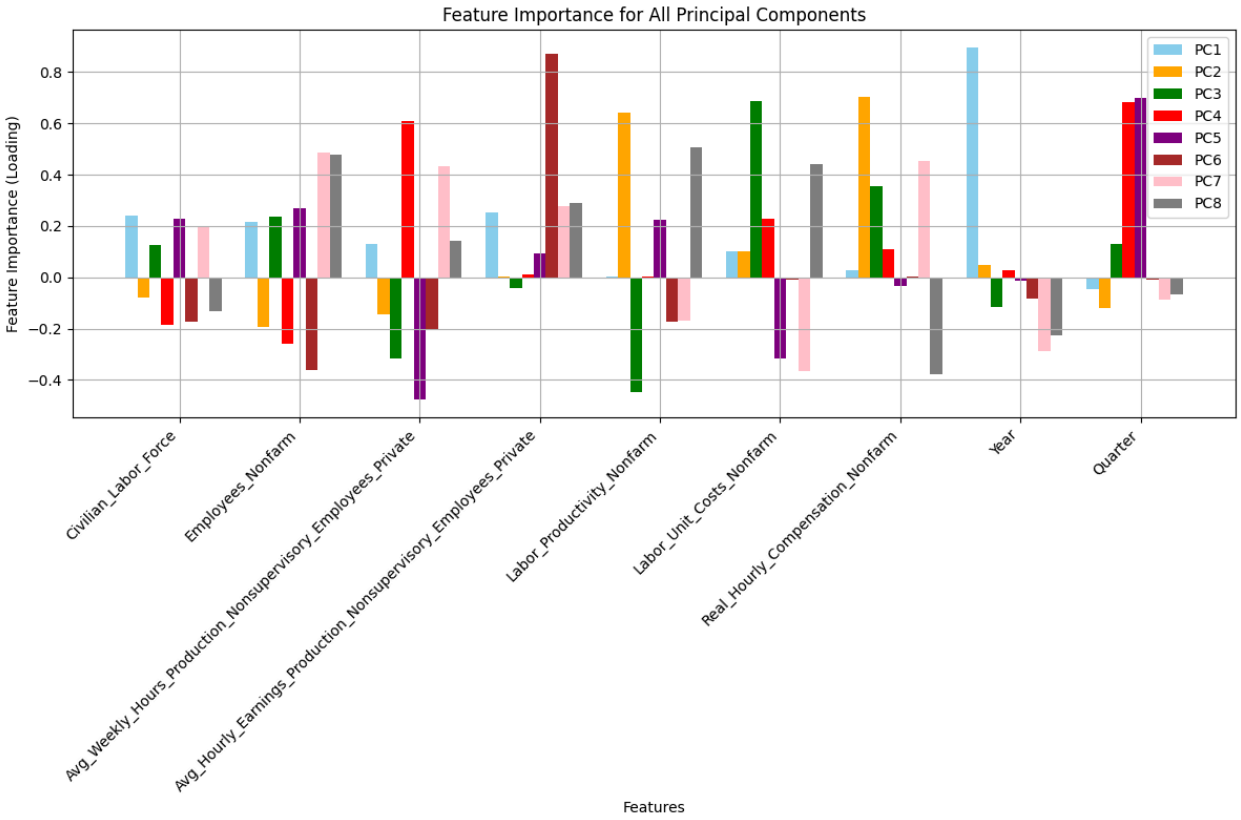


Figure 11: 2013-2024 PCA Feature Importance



PC1: represents general economic expansion over time as Year large influence.

PC2: represents the relationship between productivity and compensation. Are workers who are productive getting appropriately paid for their contributions?

PC3: represents productivity versus labor costs and how changes in productivity relate to changes in costs and compensation

PC4: represents fluctuations in weekly hours worked versus quarterly patterns. For example, the negative relationship with Nonfarm employees COULD MEAN that weekly hours vary when there are more or less people working.

PC5: represents the inverse relationship between weekly hours and productivity, maybe saying that working longer does not equal higher productivity

PC6: This component is very influenced by wages, as hourly wages are linked to changes in costs/employment.

PC7: represents the dynamics between employment levels and labor costs. Higher employment means changes in hours worked and unit costs.

PC8: shows the relationship between labor productivity, unit costs, and compensation is pretty balanced, suggesting trade-offs between increased productivity and wage pressures.

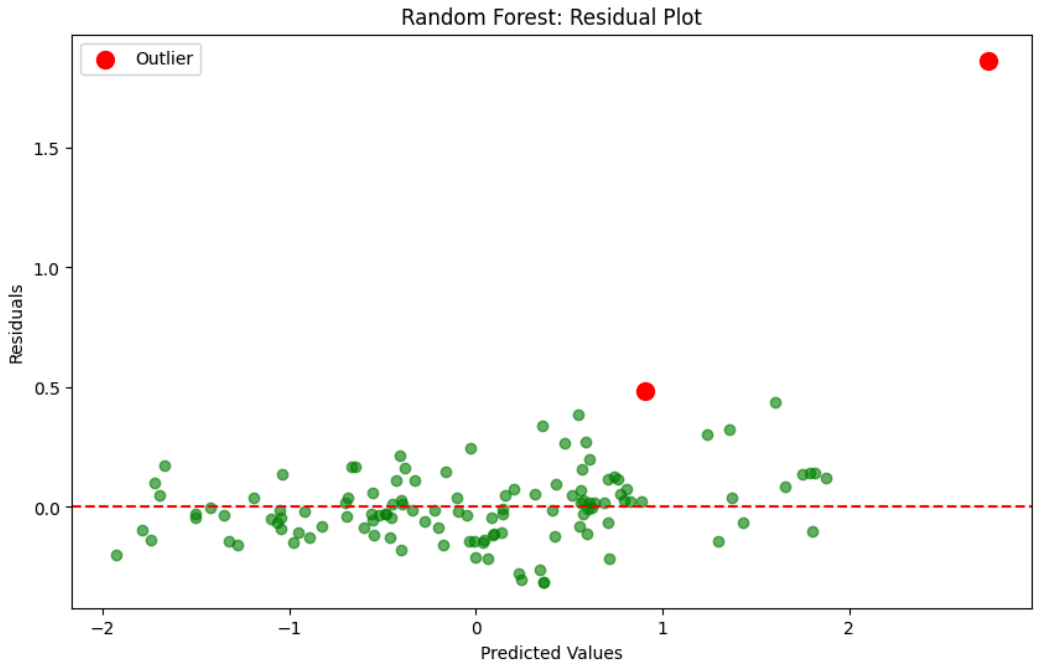
PCA outperformed Ridge.

Random Forest:
From here, I then decided to do Random Forest Regression because it decorrelates the individual trees it makes. The features are highly correlated, so the fact that Random Forest uses a random subset of features at each tree split helps stop the most dominant features from over influencing my model. I ran grid_search.best_estimator_ to present the best achieving parameters.

Best Parameters: {'max_depth': 10, 'max_features': 0.5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
Best Random Forest Model Mean Squared Error (MSE) on Test Set: 0.06379693864441274
Best Random Forest Model R-squared on Test Set: 0.765209611538293

- The most impactful features are **Employees_Nonfarm, Labor_Productivity_Nonfarm, and Civilian_Labor_Force**. These suggest that the model is heavily influenced by employment and productivity levels in the nonfarm sector.
- Features like **Quarter** and **Year** have very low importance, meaning time features have little influence on prediction. This makes sense because it is seasonally adjusted data.
- Residuals look reasonable - big outlier year is 2020

Figure 12: 2013-2024 Random Forest Residual Plot



1974-2024: Dataframe Over Longer Period with Less Features:

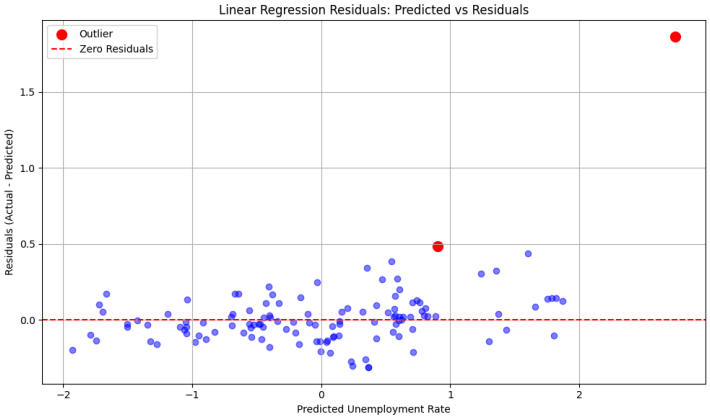
Linear Regression Versus Ridge Regression:

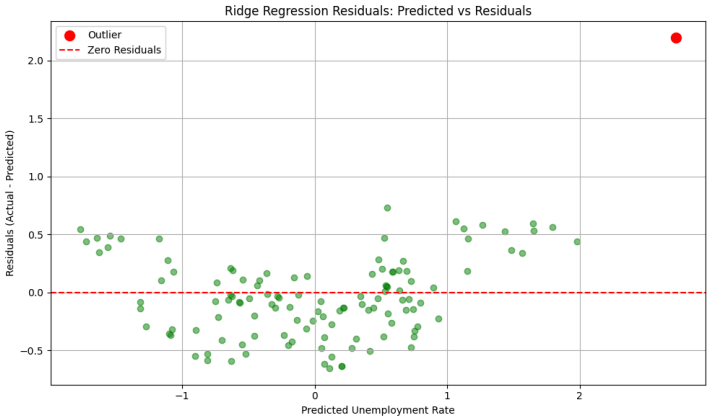
```
Linear Regression Mean Squared Error: 0.132502427058951
Ridge Regression MSE: 0.1520708539208546
```

Ridge regression's MSE is higher, indicating that the regularization (penalization of large coefficients) model has slightly worse performance in terms of MSE. However, Ridge tends to prevent overfitting and generalization. According to the coefficients, the linear regression model is more sensitive to changes in variables like **civilian labor force** and **nonfarm employees**, leading to higher coefficients. The Ridge model shows smaller coefficients for many variables due to regularization.

```
Linear Regression Outlier year: 2020
Ridge Regression Outlier year: 2020
```

Figure 13 (LEFT): 1974-2024 Linear Regression Residuals
Figure 14 (RIGHT): 1974-2024 Ridge Regression Residuals





Boosting:

In gradient boosting, each new tree is built to correct the residuals of the previous trees. It is fairly good at capturing complex relationships between variables.

```
Gradient Boosting MSE: 0.05014073108856459
Outlier year: 2020
```

This model has a much lower MSE than Linear regression or Ridge Regression.

Figure 15: 1974-2024 Feature Importance from Gradient Booting

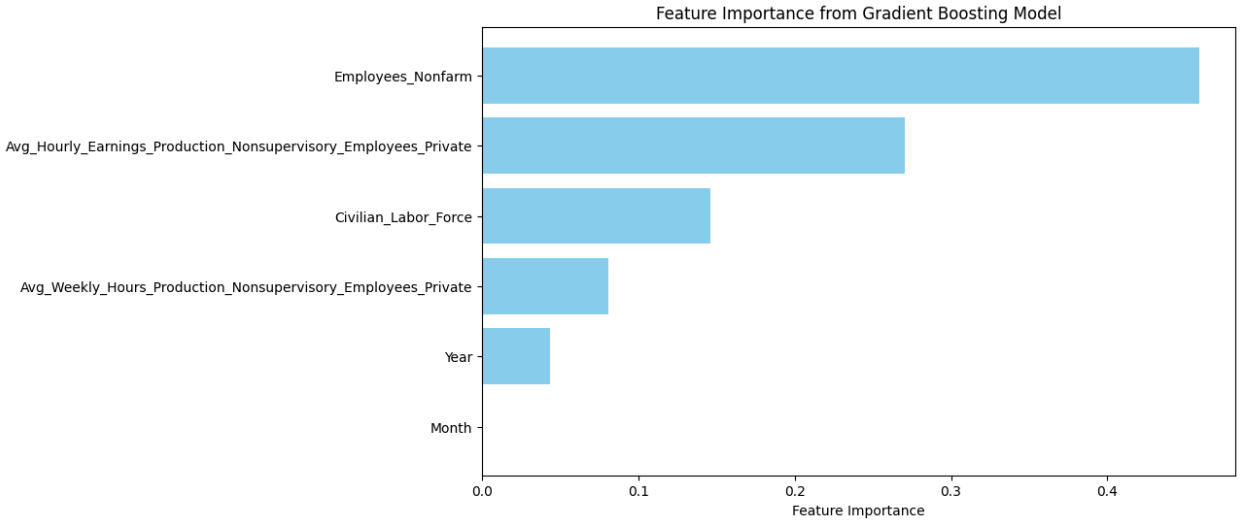
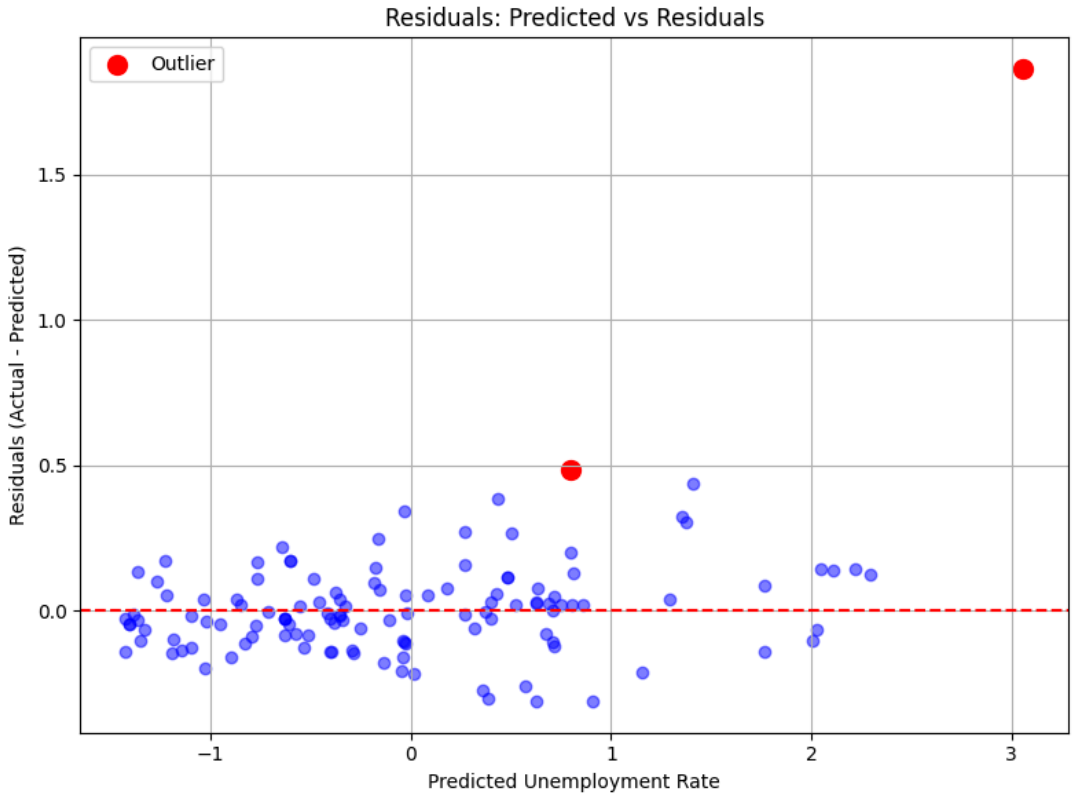


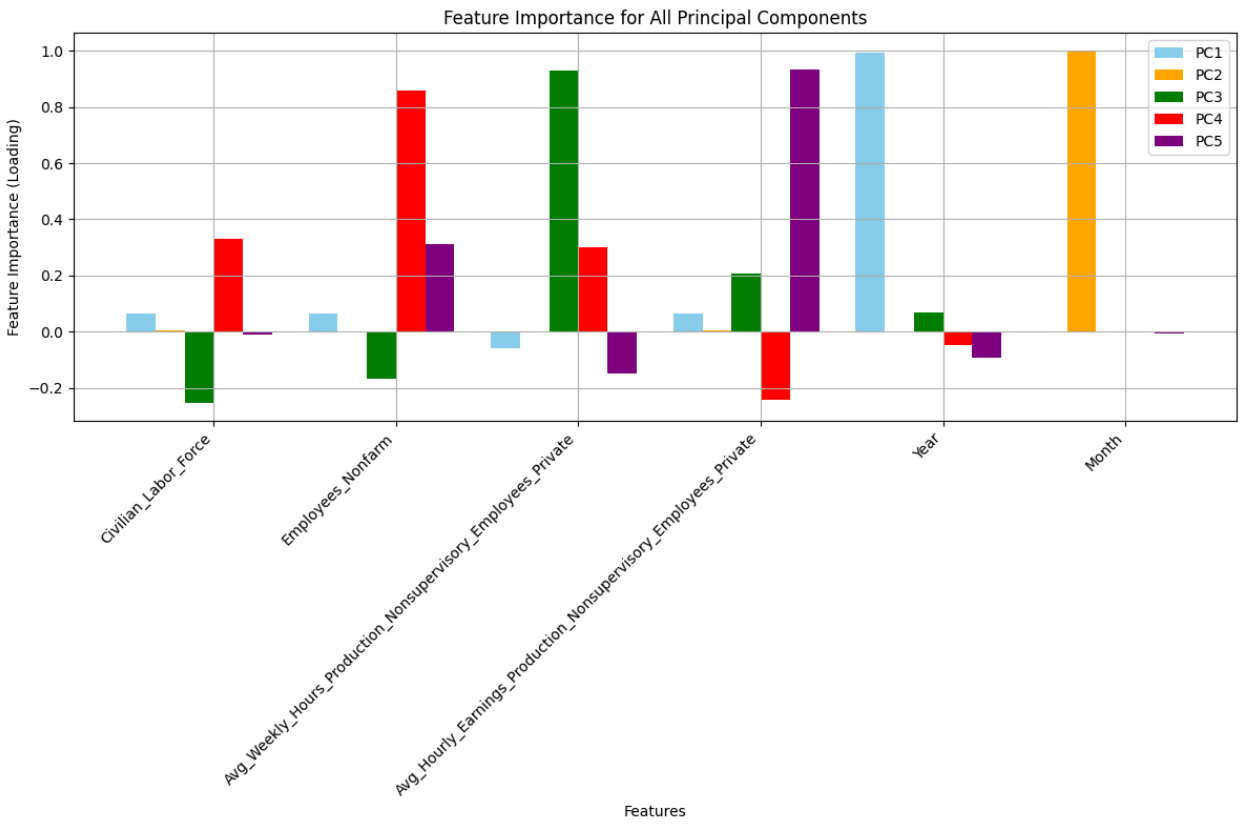
Figure 16: 1974-2024 Boosting Residuals



I wanted to make a PCA model here for comparison purposes:

```
Mean Squared Error (MSE) on Test Set: 0.19469264304712144
R-squared on Test Set: 0.820842957837627
```

Figure 17: 1974-2024 PCA Feature Importance



Question 4: Drawing a Data-Driven Conclusion

As we view the past fifty years of unemployment in the United States in figure 5, it is not hard to notice that the highest spikes of unemployment happen at some of the hardest points of economic hardship in recent history, like the 2009 recession and COVID-19. However, one graph is not enough to explain the complexity each available feature has in its relationship with unemployment.

While my models that relied on data from 2013 to 2024 performed better, it is important to recognize the need for a larger time span to gain more valuable insight into trends. Looking at my fairly successful Boosting model and figure 15, we can see key features, including nonfarm employees (unsurprisingly as less employees = more unemployment), average hourly earnings for private nonsupervisory and production employees, and civilian labor force. Seemingly, a larger civilian labor force may lead to higher unemployment, as more people actively seek employment, potentially outpacing available jobs. For average hourly earnings for private nonsupervisory and production employees, looking at figure 3, we can see that there is a negative correlation between itself and the unemployment rate, which suggests that as hourly

earnings increase, the unemployment rate tends to decrease. This however does not happen every time hourly earnings increase, but its feature importance is notable.

For comparison of feature importance, consider the principal component analysis done on the data from 2013 to 2024 that had more features available than the data starting in 1974. This was the most successful model for this data set. For the sake of comparison, I performed PCA on the data from 1974 to 2024 as well. This did not perform as well as my boosting model. Here are some notable differences:

- The strong influence of the "Year" variable in both models suggests that long-term economic trends, such as business cycles and technological advancements, significantly impact unemployment rates.
- Features that are related to labor force and wages are also important in both models. Changes in these variables can affect labor demand and supply, leading to fluctuations in unemployment rates.

However, my 1974-2024 PCA model was not as good as my Boosting model was which highlighted some of these features as important to understanding unemployment patterns:

- Employees_Nonfarm: This suggests that changes in non-farm employment have a significant impact on unemployment rates.
- Avg_Hourly_Earnings_Production_Nonsupervisory_Employees_Private indicates changes in hourly earnings can influence unemployment rates, potentially through their impact on consumer spending and business investment.
- Civilian labor force's importance implies that changes in the size of the labor force impacts unemployment, as mentioned earlier.

As this was my first experience performing data mining analysis on real-world datasets, choosing the right model was the most challenging aspect. Here is why I believe my choices were correct, as is reflected in their low MSE values and high R-squared values. For the data over a shorter period of time but with more features (scaled_df_combined), PCA was the correct choice for its ability to reduce dimensionality but still captures the important patterns in the data. PCA helped simplify the complex relationships shown in the correlation matrix. For my longer data set (scaled_df_SA_start1974), Gradient Boosting was the right choice because it captures non-linear relationships and interactions between features while creating trees that correct the errors of the tree made before it. Gradient Boosting handled the complexity of this longer dataset best out of Linear regressions, Ridge, and PCA. Nevertheless, to ensure the continued evaluative success of these models, the data frames should be updated with new data as provided by the BLS and success metrics should be monitored as it is re-run with new data. This will ensure the models are not stagnant and unrelated to more modern times as data continues to be collected over the years.