**Forecasting U.S Unemployment Trends by Age Groups**

**Group 7 Team Members**

Rasika Vanagutte

Divya Davra

Tejashree Gundeti

Shrey Sheta

**Introduction**

Using a 12-month moving average from the Current Population Survey (CPS), our project analyzes California's monthly unemployment rates by age group (16–19, 20–24, 25–34, 35–44, 45–54, 55–64, and 65+). This Kaggle dataset, which is not seasonally adjusted, provides historical depth and granularity, which makes it appropriate for time series and demographic analysis. Finding patterns, predicting trends, and evaluating the effects of unemployment on various age groups over time are our objectives. Furthermore, we want to investigate relationships with more general socioeconomic metrics like GDP and educational attainment.

**Problem**

One important socioeconomic problem that has a range of effects on people, families, and communities is unemployment. Despite its importance, the emphasis on overall unemployment statistics frequently masks differences between age groups. Unemployment rates by age can differ greatly depending on several factors, including economic policy, industry trends, labor experience, and educational attainment. However, the capacity to create focused treatments that address the particular difficulties faced by each age group is constrained by the paucity of thorough analysis of these differences.

In particular, we are examining unemployment patterns among those in the 20–24 age range, who are generally leaving school to enter labor and begin their careers. The purpose of this analysis is to determine the obstacles that young workers have when they first enter the workforce and determine whether there are enough entry-level opportunities to assist their integration into employment.

**Data**

Dataset Source: <https://www.kaggle.com/datasets/sahirmaharajj/unemployment-by-age-groups-dataset/data>

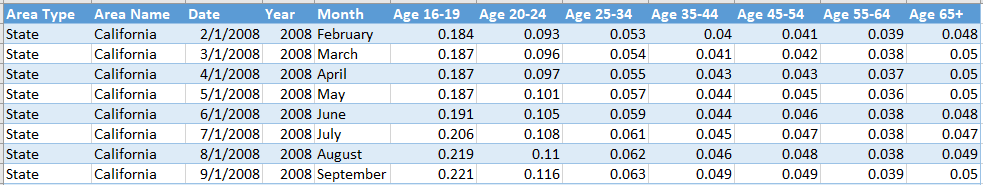


Figure 1

By examining the data source in Figure 1. The dataset utilized for this analysis is sourced from Kaggle, a well-known platform for data enthusiasts and professionals. Specifically, the dataset "Unemployment by Age Groups" was selected due to its relevance and comprehensive coverage of unemployment trends across various age demographics. Kaggle's reputation for hosting high-quality datasets makes this source reliable and suitable for the objectives of this study.

Considering the vast amount of information within the dataset, a focused selection of variables was necessary to align with the study's goals. The primary variables analyzed include:

* **Date**: Capturing the timeline for unemployment trends.
* **Age\_20\_24:** specifically analyzing unemployment trends among individuals aged, a demographic typically transitioning from education to the workforce and starting their careers.

This curated selection enables a targeted exploration of how unemployment rates fluctuate over time, across age groups, providing valuable insights for the analysis.

**Data Cleaning/Validation**

As part of the data cleaning and validation process, we verified whether the dataset contained any missing values. Upon verification, we confirmed that there were no missing values in the dataset. The dataset consists of non-seasonally adjusted California unemployment rates, categorized by age groups, derived from the Current Population Survey (CPS). It is structured based on a 12-month moving average, ensuring smoother trends for analysis.

We partitioned our dataset into an 80/20 split, resulting in 154 records in the training set and 40 records in the validation set. This approach ensures that the training set is sufficiently large to build and train the model, while the validation set provides a reliable basis for evaluating the model's performance on unseen data.

**Exploratory Data Analysis**

As shown in Figure 2, We analyzed and compared unemployment trends across various age groups as part of our forecasting project, identifying distinct patterns and variations in unemployment rates over time for each demographic. Here’s a brief analysis for different age groups:

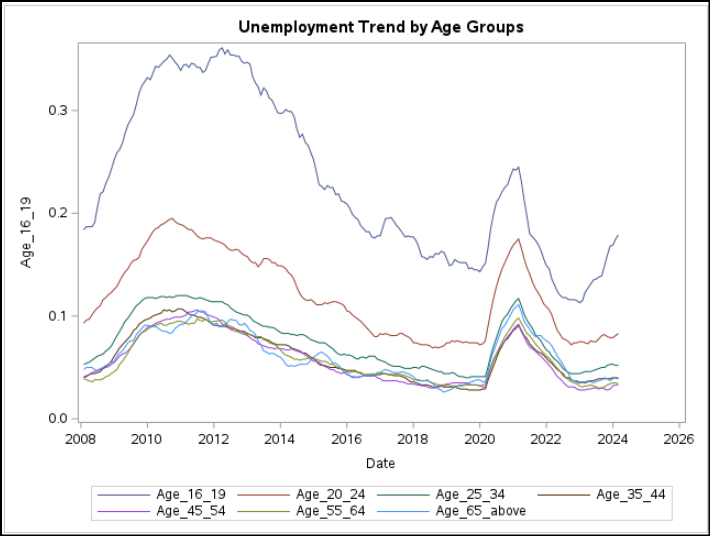


Figure 2

* **Age 16–19:** The youngest group exhibits the highest unemployment rates consistently, peaking sharply around 2010 and 2020. A noticeable spike in 2020 aligns with the economic impact of the COVID-19 pandemic, followed by a decline and a recent increase post-2022.
* **Age 20–24:** This group also has relatively high unemployment rates but consistently lower than those of the 16–19 group. Similar peaks were observed in 2010 and 2020, followed by recovery trends.
* **Ages 25–64 (Prime Working Age):** The unemployment rates for these age groups are significantly lower and more stable compared to the younger groups. While all groups show the 2020 spike, their recovery trends appear steadier post-pandemic.
* **Age 65 and Above:** The unemployment rates are the lowest among all groups, indicating more stability in this demographic, likely due to retirements or fewer active participants in the labor market minimal fluctuation over time except for a slight rise in 2020.

**Key Observations:**

1. The 2020 spike is visible across all groups, reflecting the universal impact of the pandemic on employment.
2. Younger age groups (16–19, 20–24) face more pronounced volatility in unemployment rates, while older groups show stability.
3. The recovery trajectory post-2020 varies, with younger groups recovering slower compared to prime working-age and older demographics.

**Analysis**

Figure 3 identifies the time series plot highlighting a clear downward trend in the unemployment ratio for the 20–24 age group over more than a decade, with two prominent peaks - one in 2010, reflecting the financial crisis, and another in 2020, caused by the economic disruptions of COVID-19. After each peak, there is a steady decline, showing recovery phases as the economy stabilizes. The trend also exhibits non-linearity, as the rate of decline varies over time, reflecting periods of faster and slower recovery based on underlying economic conditions.

The ACF plot further supports this observation, with autocorrelations gradually declining toward zero, suggesting a strong overall trend in the data without any evidence of seasonality. This indicates that the fluctuations in unemployment are driven more by long-term economic factors rather than recurring seasonal patterns.

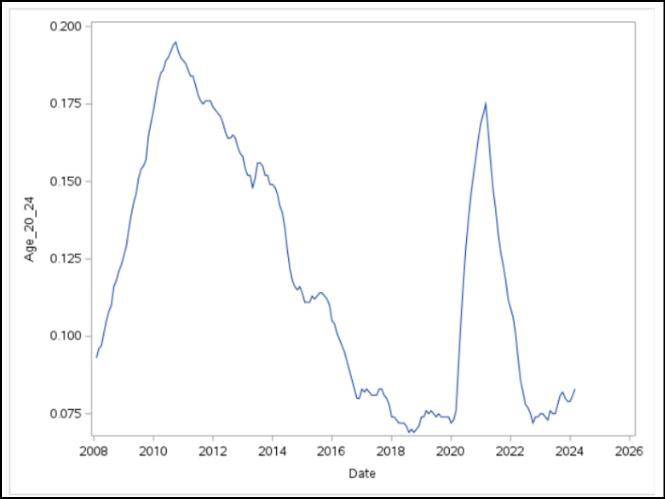
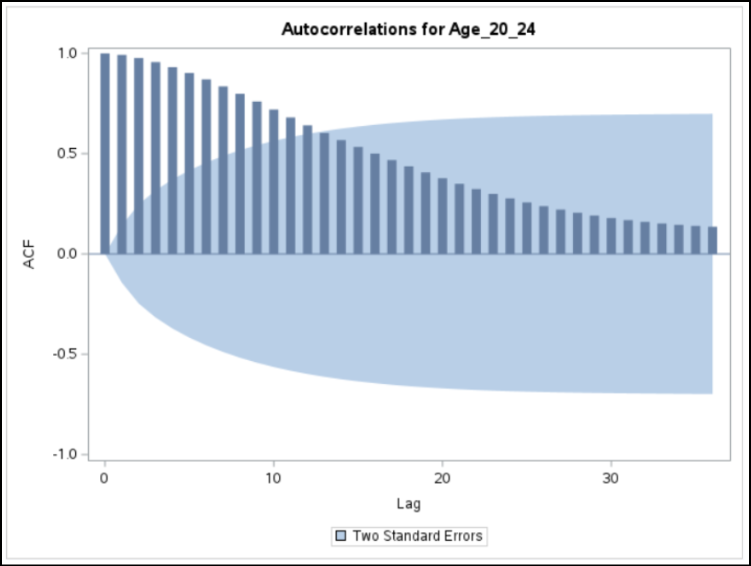
** **

Figure 3

**Model Selection**

Considering the outcome of the time series plot showing the strong trend in the time series data without any seasonal pattern, we have decided to use below-forecasted models:

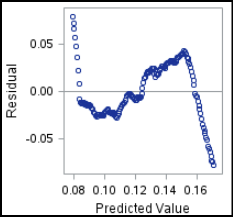
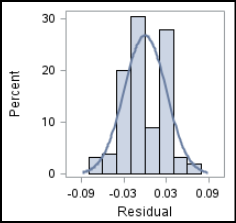
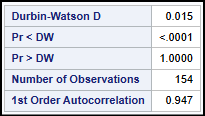
* Simple Linear Regression
* Non-Linear Regression Model
* Damped-Trend Exponential Smoothing
* Holt's Exponential Smoothing
* ARIMA

**Modeling Process:**

1. **Simple Linear Regression**

We began our analysis with a Simple Linear Regression model to predict unemployment trends using time as the independent variable. The model was statistically significant, as its p-value was lower than the significance level (𝛂). The equation derived from the model was Ŷ = 0.1715 - 0.0006t, with R² value of 45%. However, this R² value indicates that the model explains only a minimal portion of the variance in unemployment trends.

As shown in Figure 4, the regression model failed to meet two of three critical assumptions of linear regression. The residuals displayed non-constant variance, violating the assumption of homoscedasticity. The histogram of residuals (middle) shows a near-normal distribution, supporting the normality assumption for residuals. The Durbin-Watson test result (D = 0.015) indicates strong positive autocorrelation in the residuals, as confirmed by the 1st order autocorrelation value of 0.947 and the significant p-value (< 0.0001). These issues highlight the limitations of the linear regression approach in accurately modeling the data.

  Figure 4

While the model effectively captured the overall downward trend in unemployment, it struggled to represent the non-linear patterns evident in the time series data as seen from the plot between actual vs predicted values. These results suggest the need for a more advanced modeling approach, such as polynomial regression or a time series model, to better capture the complexities of the unemployment trends.

Figure 5 summarizes model performance, with a Mean Absolute Error (MAE) of 0.025 for the training set 0.036 for the validation set, and a Mean Squared Error (MSE) of 0.001 for training and 0.002 for validation. The error measures are very close, indicating no significant sign of overfitting.

A graph with a line and a red line

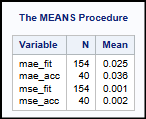
Description automatically generated 

Figure 5

1. **Non-Linear Regression Model**

We selected a non-linear regression model for our analysis since the time series data exhibited a non-linear trend. The model was statistically significant, as its p-value was below the specified 𝛼 level, and it showed a slight improvement in the R² value, increasing to 49%. The parameter estimates revealed that while the linear time component was not statistically significant, the quadratic time component was statistically significant within the model.

The non-linear quadratic regression model failed to meet most of the critical assumptions of the regression model as seen in Figure 6. The residuals displayed non-constant variance, violating the assumption of homoscedasticity. The residuals appear to have a normal distribution. The residuals exhibited positive autocorrelation, which breached the assumption of independence. The Durbin-Watson test result (D = 0.016) indicates strong positive autocorrelation in the residuals, as confirmed by the 1st-order autocorrelation value of 0.94 and the significant p-value (< 0.0001).

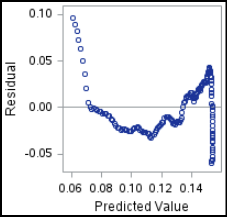
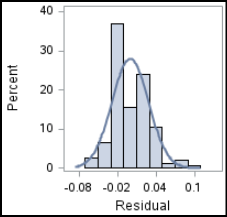
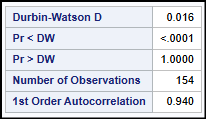
**  **

Figure 6

The plot of Actual vs. Predicted values indicates that the non-linear model captures the overall curvature of the trend more effectively than the linear model. However, it struggles to account for the significant cyclical peaks observed in the time series, suggesting that the model cannot fully capture periodic fluctuations or short-term variations in the data.

The error measures calculated for model fit and model accuracy indicate potential signs of overfitting as the MAE of the validation set is slightly higher than the MAE of the training set, and the MSE value of the validation set also showed a significant increase compared to the training set.

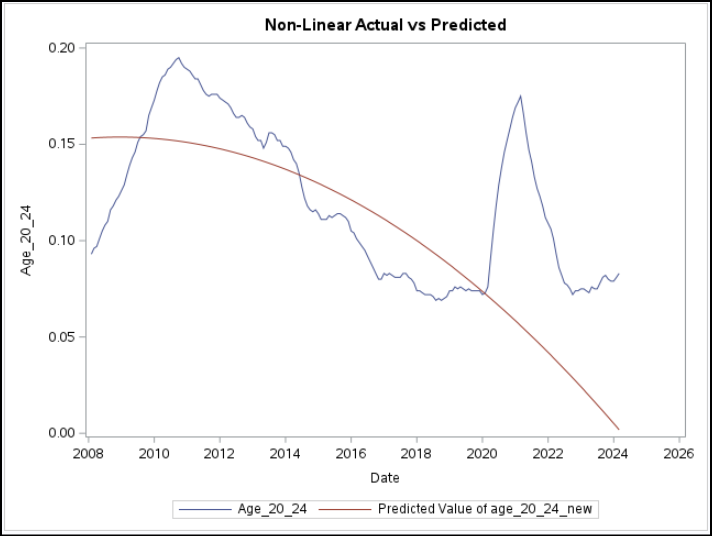
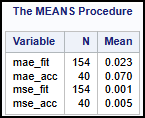
** **

Figure 7

1. **Exponential Smoothing Models**

Two types of exponential smoothing models were implemented: Linear Exponential Smoothing and Damped-Trend Exponential Smoothing. Their performance was evaluated based on parameter estimates, predictive accuracy, and visual trends in their forecasts.

As per Figure 8, The Linear Exponential Smoothing model employs two key parameters, Level Weight and Trend Weight, with estimated values of 0.9990 and 0.68915, respectively. These parameters are statistically significant, as indicated by p-values < 0.0001. In comparison, the Damped-Trend Exponential Smoothing model incorporates an additional parameter, the Damping Weight. Its parameter estimates Level Weight (0.97659), Trend Weight (0.80480), and Damping Weight (0.87263) demonstrated strong statistical significance as well.

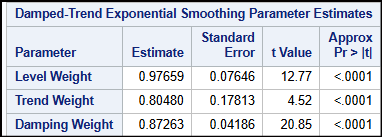
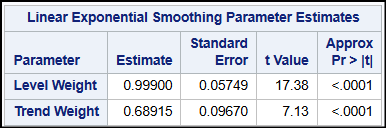


Figure 8

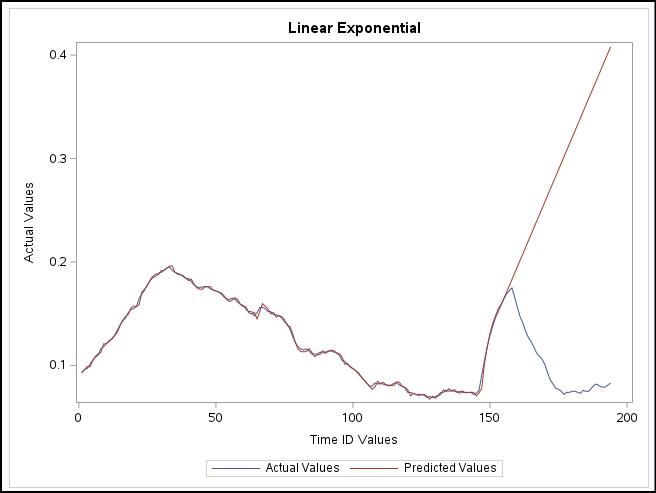
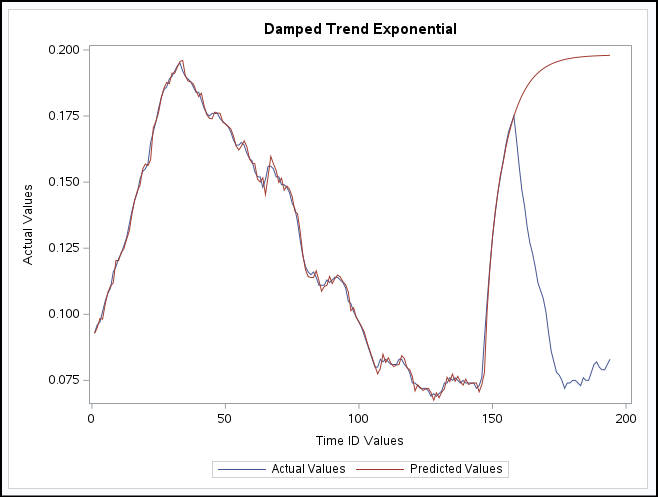
 ****

Figure 9

Table 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Smoothing Models Model Fit** | **Model Fit Model Fit** | | | **Model Fit Model Accuracy** | | |
| **MSE** | **RMSE** | **MAE** | **MSE** | **RMSE** | **MAE** |
| **Linear Exponential** | 0 | 0.002 | 0.001 | 0.044 | 0.21 | 0.18 |
| **Damped Trend** | 0 | 0.002 | 0.001 | 0.009 | 0.098 | 0.089 |

Table 1 and Figure 9 provide a detailed evaluation of two smoothing models, Linear Exponential and Damped Trend, for both model fit and accuracy. The p-value (<0.0001) indicates the models are statistically significant. Additionally, the R-square value of 0.99 demonstrates that 99% of the variance in the unemployment rate is effectively explained by the model, showcasing its strong predictive capability.

For model fit, both smoothing models achieved excellent results with near-zero Mean Squared Error (MSE), Root Mean Squared Error (RMSE) of 0.002, and Mean Absolute Error (MAE) of 0.001. This highlights the models' ability to closely align predicted values with actual values within the training dataset.

Regarding model accuracy, the Damped Trend model outperforms the Linear Exponential model in terms of lower MSE (0.009 vs. 0.044), RMSE (0.098 vs. 0.21), and MAE (0.089 vs. 0.18) when applied to unseen data. The charts further support these observations, as the Damped Trend model better captures patterns in the data, while the Linear Exponential model diverges significantly in certain regions, particularly towards the end.

Overall, while both models perform well, the Damped Trend model demonstrates superior accuracy and generalization, making it the preferred choice for this dataset. The high R-square value and low error metrics reflect the robustness of this model in explaining and predicting unemployment rates.

1. **ARIMA Model**

We applied the ARIMA model, as it is generally considered a better option for long-term forecasting due to its ability to capture both trend and seasonality in time series data. Since our data exhibits a strong trend component and is non-stationary, we applied first differencing twice to transform the data into a stationary series, making it suitable for the ARIMA model as shown in figures 10, 11, and 12.

**Non-Stationary Data:**

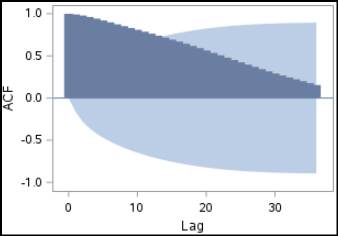
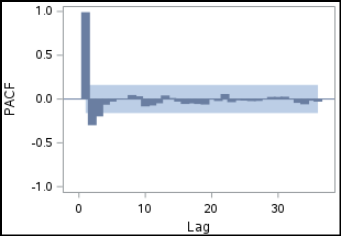
** **

Figure 10

**First Order Differencing:**

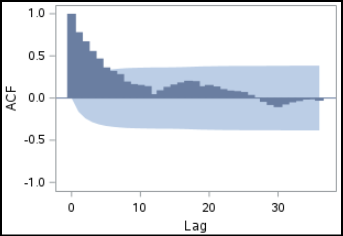
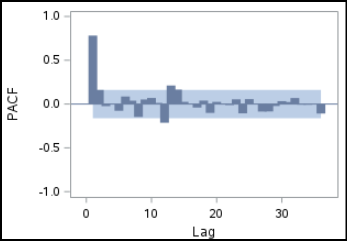
** **

Figure 11

**Second Differencing:**

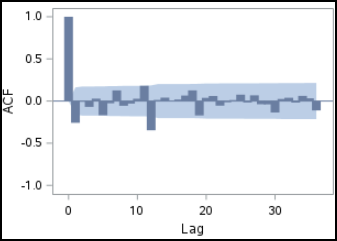
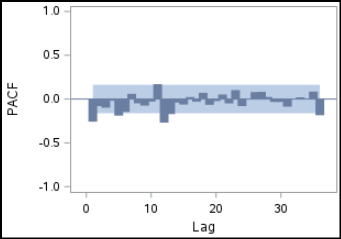
** **

Figure 12

The few ARIMA models we tried before finalizing the order of ARIMA are:

* ARMA(4,2,0)
* ARMA(6,2,0)
* ARMA(8,2,0)
* ARMA(4,2,3)
* ARMA(6,2,3)
* ARMA(8,2,2)
* ARMA(8,2,3)

In all the models we tested, the residuals failed to exhibit a white noise pattern based on the number of significant peaks, leading us to discard those models. After exploring various combinations of AR and MA parameters, we finalized the ARIMA(12,2,3) model. This decision was based on several key factors: a significant peak at lag 12 in both the ACF and PACF plots, which indicated the need for AR(12); three significant peaks in the ACF plot, supporting the inclusion of MA(3); and the application of second differencing to achieve stationarity in the data.

**ARIMA(12,2,3)**

The autocorrelation check for the residuals shows p-values greater than the value of 𝛼, confirming that the residuals follow a white noise pattern. Additionally, the standard error estimate, AIC, and BIC values are quite low, indicating a good model fit. This suggests that the ARIMA(12,2,3) model effectively captures the underlying patterns in the data and provides a reliable forecast.

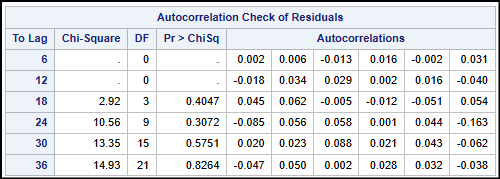
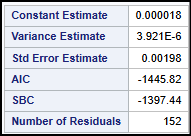
** **

Table 2

The plot of Actual vs. Predicted values shown in Figure 13 indicates that the predicted values successfully capture the initial underlying trend. Since predicting the cyclical component can be challenging, the forecasted line does not reflect the dip observed in the actual values. However, the model does a better job of capturing the overall trend compared to the previous models, demonstrating an improved fit for the long-term forecast.

The error measures for model fit and model accuracy are quite close enough indicating a very low possibility of model overfitting. This suggests that the model generalizes well to the data and is not overly complex, making it a reliable choice for forecasting.

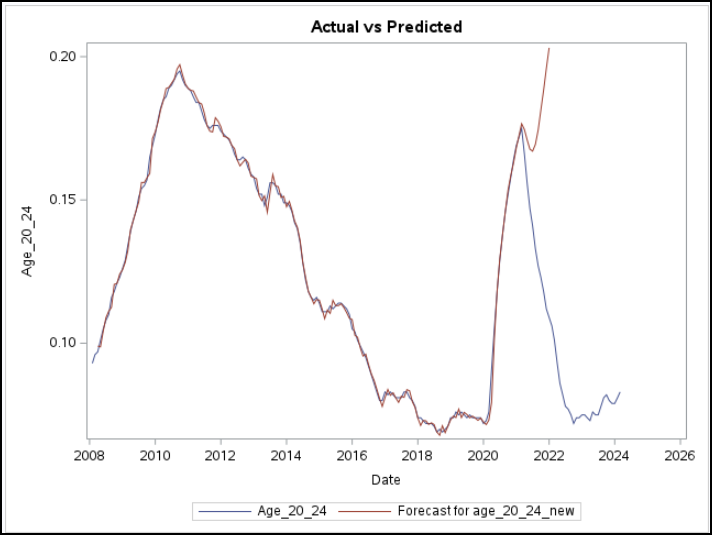
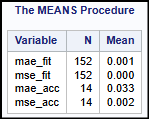
** **

Figure 13

**Model Comparison/Evaluation**

Table 3

| **Data Models** | **Model Fit** | | **Model Accuracy** | |
| --- | --- | --- | --- | --- |
| **MSE** | **MAE** | **MSE** | **MAE** |
| **Linear Regression** | 0.001 | 0.025 | 0.002 | 0.036 |
| **Non-Linear Regression** | 0.001 | 0.023 | 0.005 | 0.070 |
| **Linear Exponential** | 0 | 0.001 | 0.044 | 0.18 |
| **Damped Trend Smoothing** | 0 | 0.001 | 0.009 | 0.089 |
| **ARIMA** | 0 | 0.001 | 0.002 | 0.0033 |

**Final Model Validation and Accuracy**

On comparing the error measures across all the chosen data modeling techniques, we observe that Holt’s Linear Exponential model, Damped-Trend Exponential Smoothing model, and the ARIMA model all show low error measures on the training set, indicating a good model fit for each of them. This suggests that all three models are effective in capturing the underlying patterns in the unemployment trend. However, when we compare the model accuracy, the ARIMA model stands out with the lowest error measures, indicating that it generalizes better on unseen data compared to the exponential smoothing models.

**Generalization**

Ultimately, the ARIMA model demonstrated strong performance with the data. Since it accurately predicts the unemployment rate for the 20-24 age group, there is potential for this model to be applied to forecast unemployment rates for other age groups as well. This indicates that the ARIMA model could be a reliable tool for predicting unemployment rates across different demographic groups.

**Suggestions for Future Studies**

Future studies on unemployment trends in the 20-24 age group could expand by analyzing other demographic segments, such as gender or education level, to provide a broader understanding of employment patterns. Incorporating external economic factors like GDP growth or inflation could enhance model accuracy by capturing the wider economic influences on unemployment. Exploring models that address seasonality and cyclical trends, such as SARIMA or machine learning approaches, may improve forecasting. Additionally, comparing ARIMA with advanced models like LSTM networks, studying the impact of policy interventions, and examining regional variations could provide valuable insights. Integrating real-time data would also help track short-term fluctuations and improve model responsiveness.

**Conclusion**

The final conclusion of our data analysis on unemployment trends in the 20-24 age group reveals that the ARIMA model is highly effective in capturing the underlying trend and making accurate predictions. The model successfully accounts for long-term fluctuations in unemployment, with a good fit to the data and low error measures on both the training and validation sets. While the ARIMA model does not fully capture the cyclical peaks, it offers a strong foundation for forecasting the unemployment rate in this age group. This analysis suggests that ARIMA could potentially be extended to other age groups or demographic segments, making it a valuable tool for broader unemployment trend forecasting.

**Appendix** 