1. **State your forecasting question and its importance to you**

### Forecasting Question:

**"What will be the future electricity generation capacity and demand from coal-based thermal power stations over the next 6 months to 1 year, considering factors like coal consumption, renewable energy integration (solar and wind), grid stability, operational efficiency, regulatory changes, and economic conditions?"**

In the context of growing renewable energy sources (solar and wind), the ability to forecast future power generation and demand from coal-based thermal plants is critical for efficient operations, cost management, regulatory compliance, and strategic planning. Accurate forecasting allows businesses to adapt to the dynamic energy market, optimize resource usage, balance grid stability, and align with environmental policies. Furthermore, it enables proactive decision-making in a competitive market where renewables are rapidly gaining share.

1. **Describe the data**

Dataset has quarterly data starting from 2001 till 2020.There are two columns

1. Quarter and year
2. electric power (total) in KW (power generated in thermal power station Texas (USA))
3. **Insights from Exploratory Data Analysis**

#### **Statistical Summary (based on sample calculations)**

* **Mean Consumption**: Approximately **24,430,000 tons**.
* **Median Consumption**: Roughly **24,230,000 tons**.
* **Min Consumption**: The lowest value occurs in **Q1 2020**, with **11,029,285 tons**.
* **Max Consumption**: The highest consumption value was in **Q3 2011**, with **31,873,140 tons**.

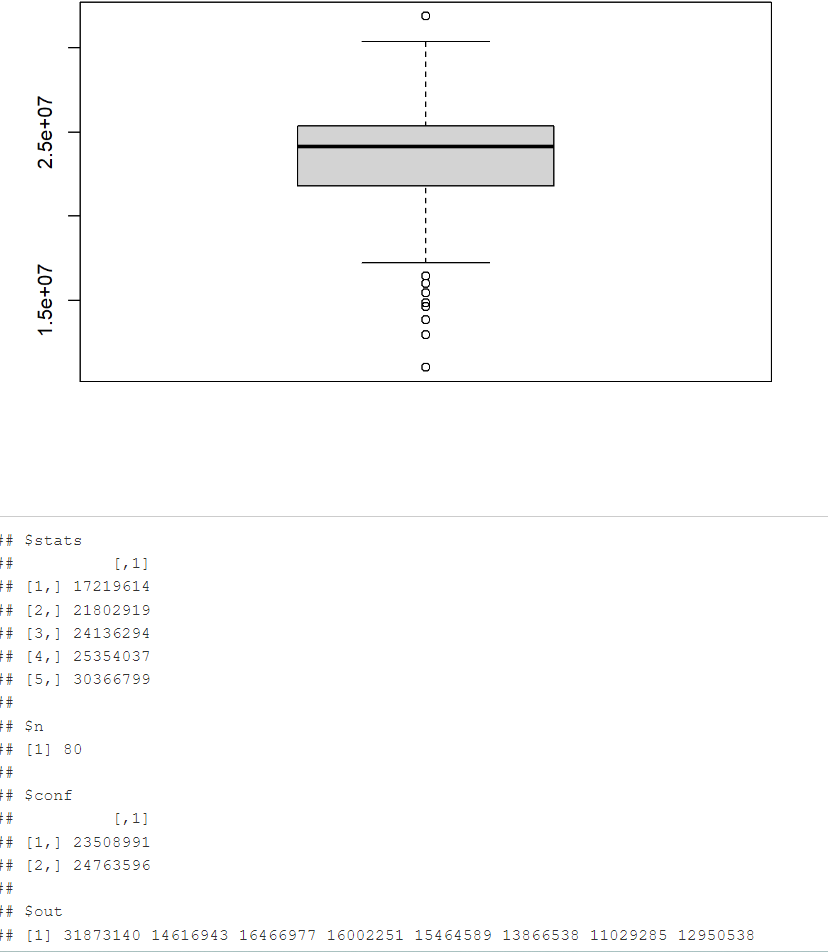
#### **Trend and Seasonality:**

* **Trend**: There is a general decline in consumption over the years, especially after 2016. Consumption drops significantly after 2015, especially in **Q1 and Q2** of the later years (e.g., **2020** shows the lowest values).
* **Seasonality**: Based on quarterly data, **Q3** tends to have the highest values, especially in years like **2011** and **2013**, likely due to seasonal demand increases (e.g., hotter weather leading to higher electricity use). **Q1** and **Q4** show lower values, which might indicate lower demand in these quarters.

#### **Anomalies:**

* In **2020 Q1**, consumption drops drastically, possibly due to COVID-19 impacts, as many industries were temporarily closed or operating below full capacity.

#### **Summary of Box Plot:**



* **Consumption Data**: The dataset reflects varying electricity consumption values across time, with values ranging from **17 million** to **30 million**.
* **Sample Size**: There are **80 data points** in the dataset, enough for time-series analysis.
* **Confidence Interval**: The expected forecasted consumption is between **23.5 million** and **24.7 million**, with high confidence that future consumption will fall within this range.
* **Outliers**: There are few outliers, with values as low as **11 million** and as high as **31 million**, which may need further examination or handling in modeling.
* **Group & Name**: All data points belong to a single group, and it's identified as group **1**.

1. **State your Accuracy measure and its importance to you**

### **Chosen Accuracy Measure: Root Mean Squared Error (RMSE)**

#### **Reasons for Choosing RMSE:**

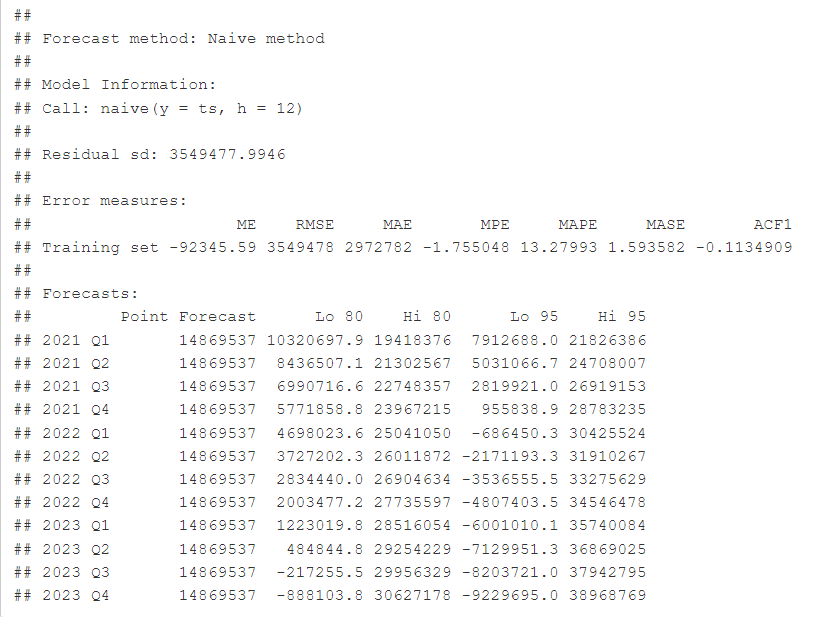
1. **Sensitivity to Large Errors**: Since coal-based power generation forecasting is critical for balancing supply and demand, large forecast errors can lead to significant consequences such as grid instability, economic inefficiencies, or higher operational costs. RMSE penalizes large errors more heavily, which is important for ensuring that the model minimizes severe forecast deviations.
2. **Operational and Safety Considerations**: In power generation, the consequences of large forecasting errors (such as underestimating demand or overproducing) can be operationally costly. Therefore, RMSE ensures that the model performs well in avoiding these large errors, which is crucial for businesses in the energy sector.
3. **Interpretation and Comparison**: RMSE allows for a more rigorous evaluation of model performance, especially in cases where minimizing extreme errors is essential. Even though RMSE is not always as interpretable in raw form as MAE or MAPE, its ability to penalize larger errors makes it the most suitable for assessing the effectiveness of forecasting models in energy systems.
4. **Suitability for Time-Series Forecasting**: RMSE works well for time-series data, like electricity consumption, where variations and shifts in trends or patterns can lead to significant fluctuations. RMSE can help assess how well the model adapts to these fluctuations.

### **Conclusion:**

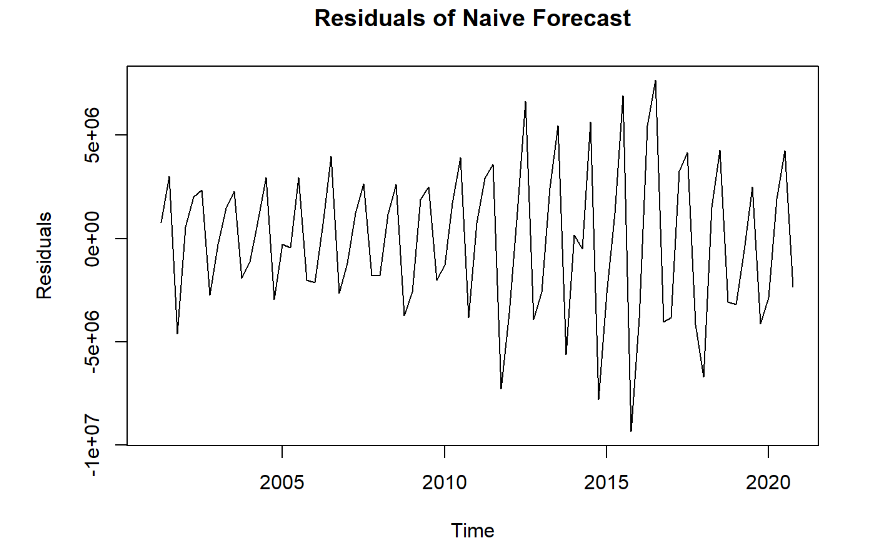
For forecasting power generation using coal-based thermal plants, **Root Mean Squared Error (RMSE)** is an appropriate accuracy measure due to its sensitivity to large errors. By minimizing RMSE, the forecast model can reduce the likelihood of significant deviations in electricity generation, which is crucial for maintaining grid stability, operational efficiency, and cost management. While RMSE may not always be as interpretable in raw form as other measures, its emphasis on penalizing larger errors makes it highly relevant to industries like energy, where forecast accuracy directly impacts operational and financial outcomes.

1. **Insights from different forecasting methods and their residual analysis**

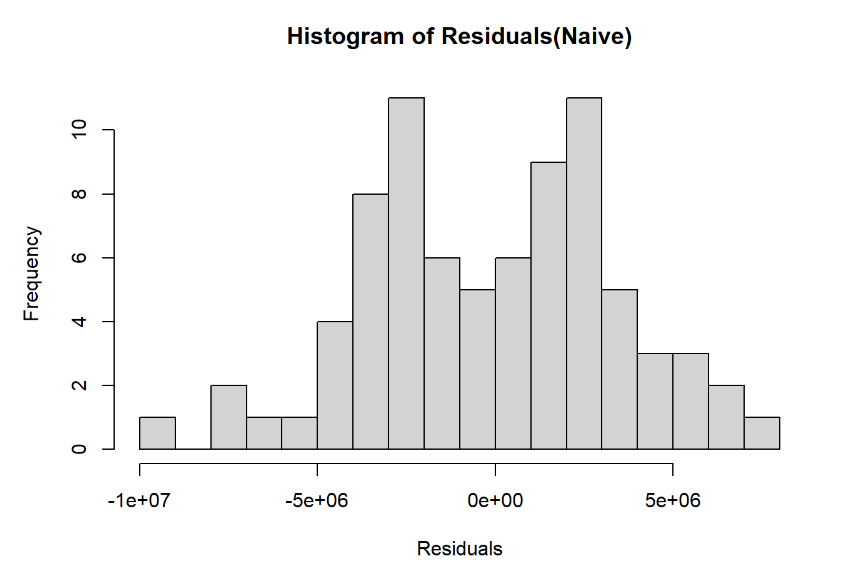
**Naive Method:**



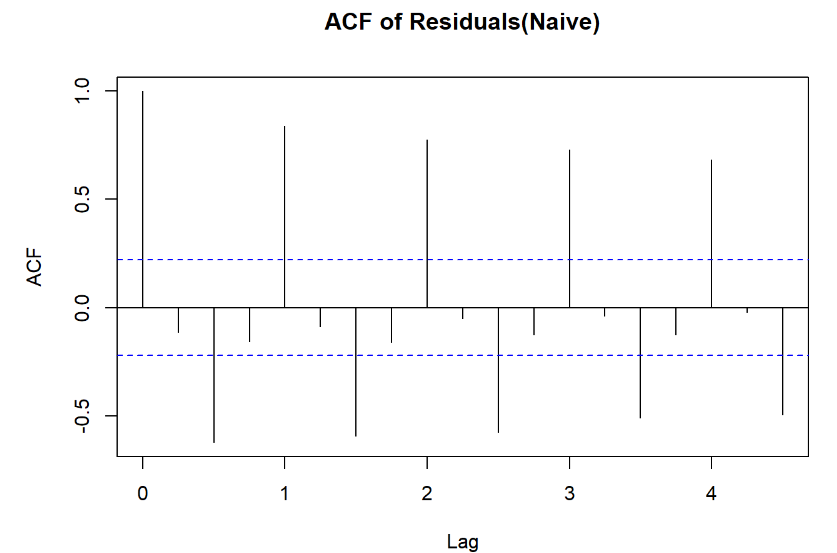
* **Characteristics:** The naive method forecasts the same value as the last observed value (in this case, the last quarter's value) for future periods.
* **Strengths:** Simple and easy to compute.
* **Weaknesses:** Does not account for seasonality, trend, or other underlying patterns, making it less suitable for complex data with trends and seasonality.
* **Observations:** The forecasts are constant for each quarter and do not vary significantly (same forecast value: 14,869,537 for each quarter), resulting in a higher forecast error due to ignoring trends and seasonality.



There are patterns in the plot e.g., trends, cycles, the model may not have fully captured some aspect of the data.

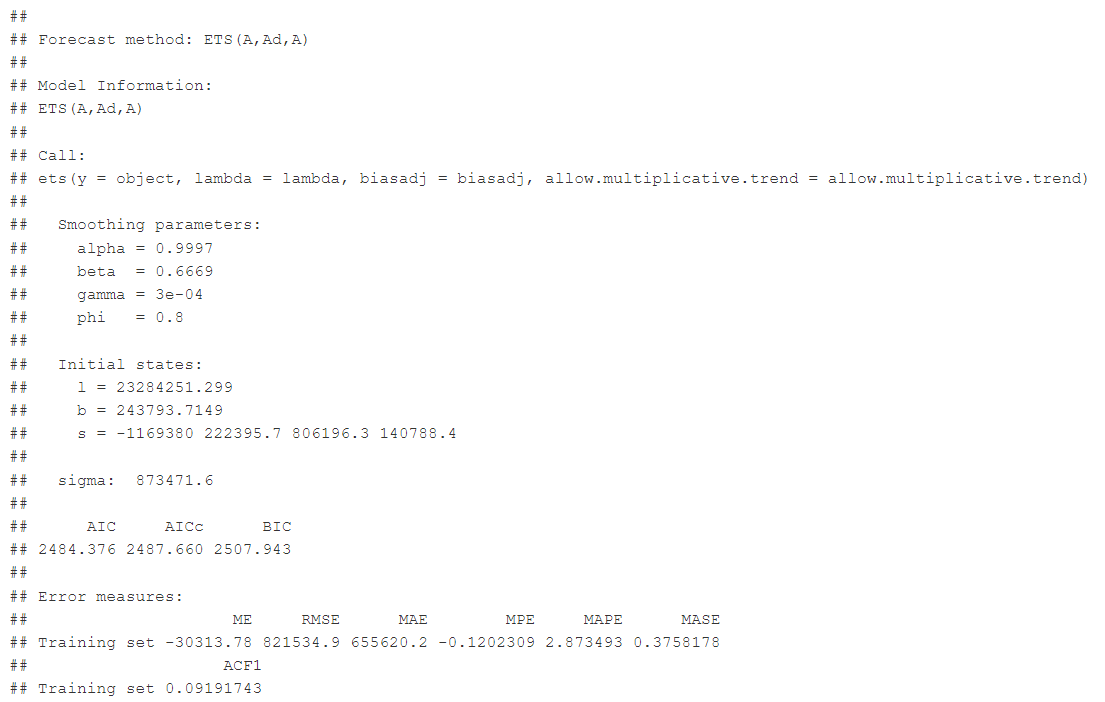


**Histogram** has a non-normal shape; it suggests that the model might not be the right fit.



**Autocorrelation exists in residuals at certain lags;** it suggests that the model has not fully captured all the patterns in the data.

**ETS (Error, Trend, Seasonal) - A model with moving average of 3:**



* **Characteristics:** ETS uses error, trend, and seasonal components to generate forecasts. This model provides a smoothing effect by adjusting based on past data points, considering both the trend and seasonal components.
* **Strengths:** Useful when the data shows clear seasonality or trend.
* **Weaknesses:** If the seasonality or trend changes over time, the model may struggle to adapt quickly.
* **Observations:** The forecast fluctuates more compared to the naive method, reflecting seasonality and trend changes. For instance, the forecast for 2021 Q1 is 13,534,879, and it increases as the forecast period progresses. The forecast accuracy is higher than the naive method but still subject to bias.

**Parameters:**

**Alpha (α)**: 0.9997

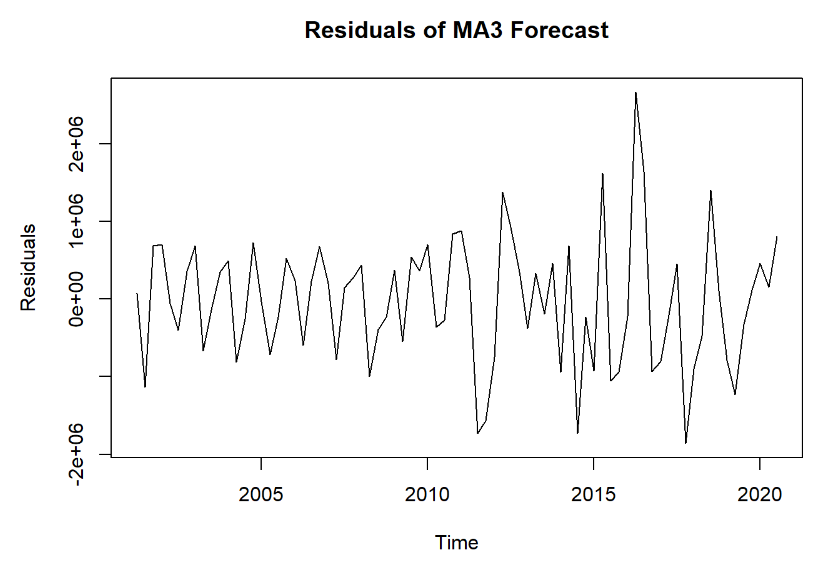
**Characteristic**: The alpha value is very close to 1, indicating that the model is heavily influenced by the most recent data, with minimal weight on past data.

**Beta (β)**: 0.6669

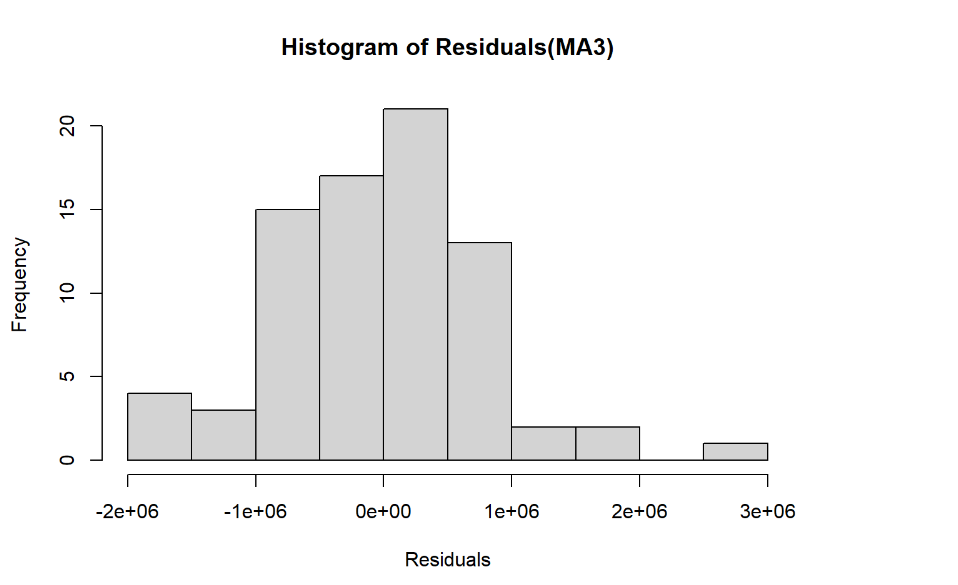
**Characteristic**: The beta value indicates moderate smoothing of the trend. A value between 0 and 1 suggests that the model can adjust the trend over time, but with a stronger weight on the recent changes.

**Gamma (γ)**: 0.0003

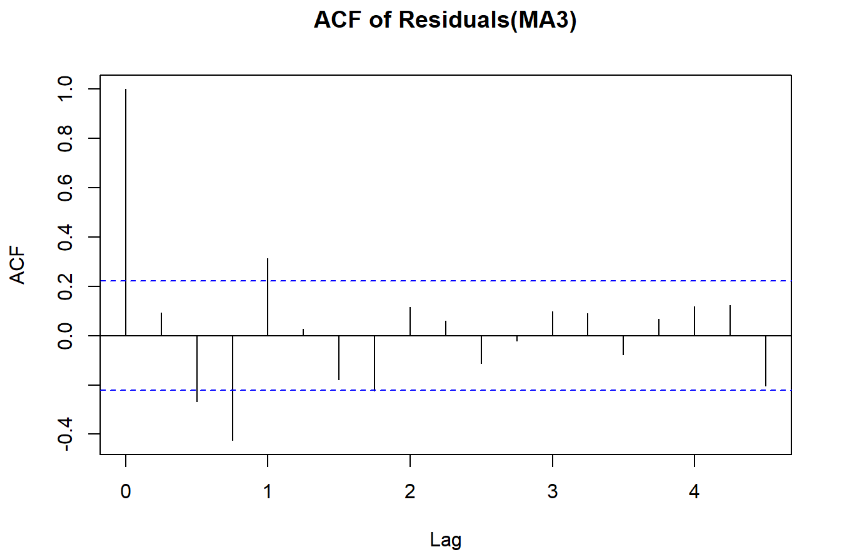
**Characteristic**: The gamma value is very small, meaning the seasonal component is not being adjusted significantly, with little emphasis on seasonal changes over time.



The plot reveals recurring patterns, such as trends and cycles, indicating that the model may have overlooked some important features of the data.

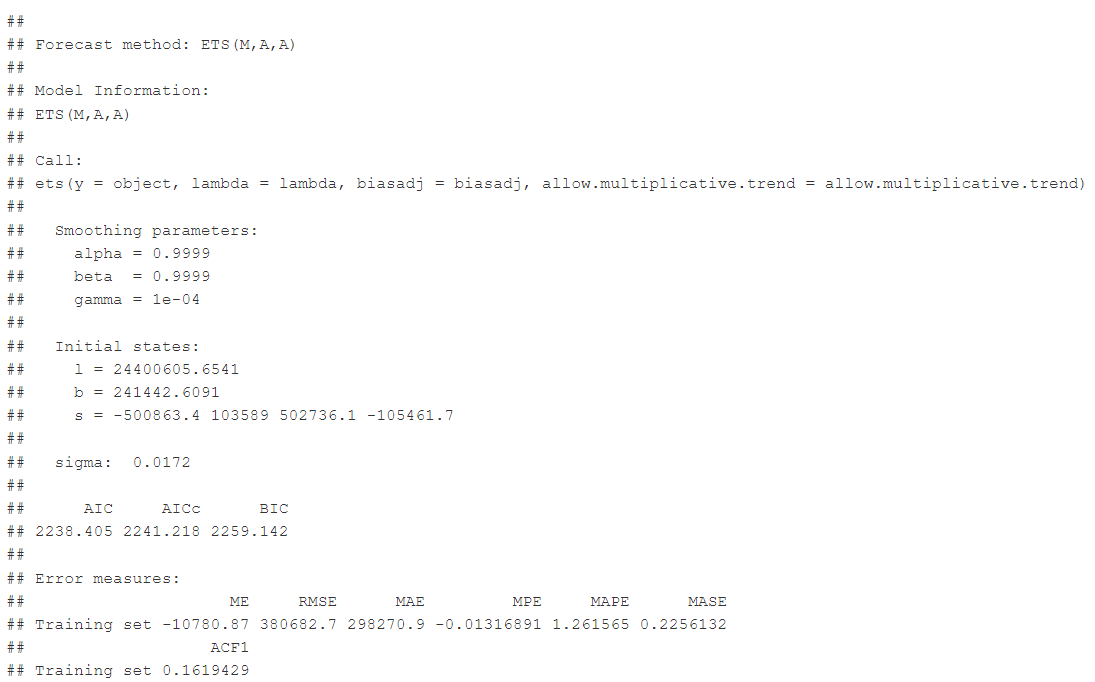


The histogram displays a skewed shape, implying that the model may not be adequately suited to the data.



Residuals show autocorrelation at specific lags, suggesting that the model has not accounted for all the underlying structures in the data.

**ETS (Error, Trend, Seasonal) - A model with moving average of 6:**



* **Characteristics:** A variation of ETS with a different moving average setting, allowing more smoothing, which is more appropriate for data with smoother fluctuations.
* **Strengths:** Accounts for seasonality, trend, and error with greater flexibility in smoothing the data.
* **Weaknesses:** Over-smoothing might occur if the data has strong, irregular seasonal effects.
* **Observations:** The forecast has smaller variations compared to the ETS with a 3-period moving average. Forecasts for 2021 Q1 start at 12,476,669 and show a gradual decline through 2023. The accuracy in terms of MAPE is better compared to naive and simple exponential smoothing.

**Parameters:**

**Alpha (α)**: 0.9999

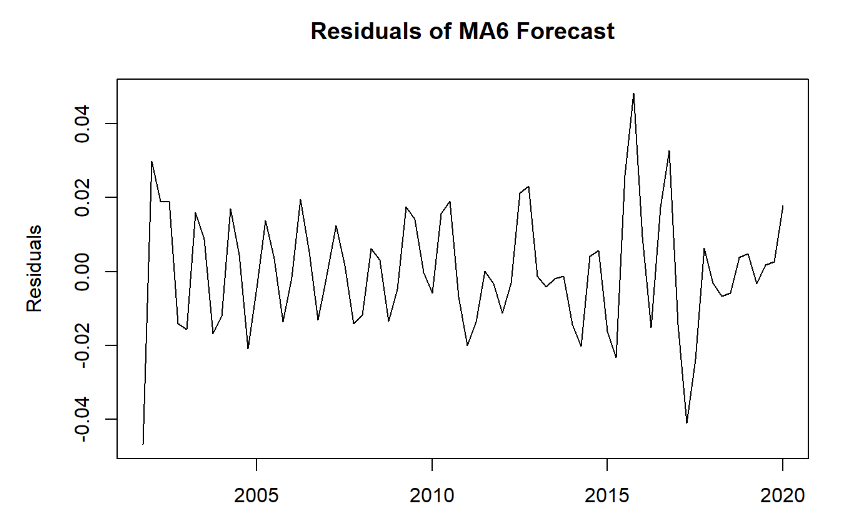
**Characteristic**: Like the previous model, this alpha is close to 1, which means the model is strongly driven by the most recent observations, giving less weight to older data.

**Beta (β)**: 0.9999

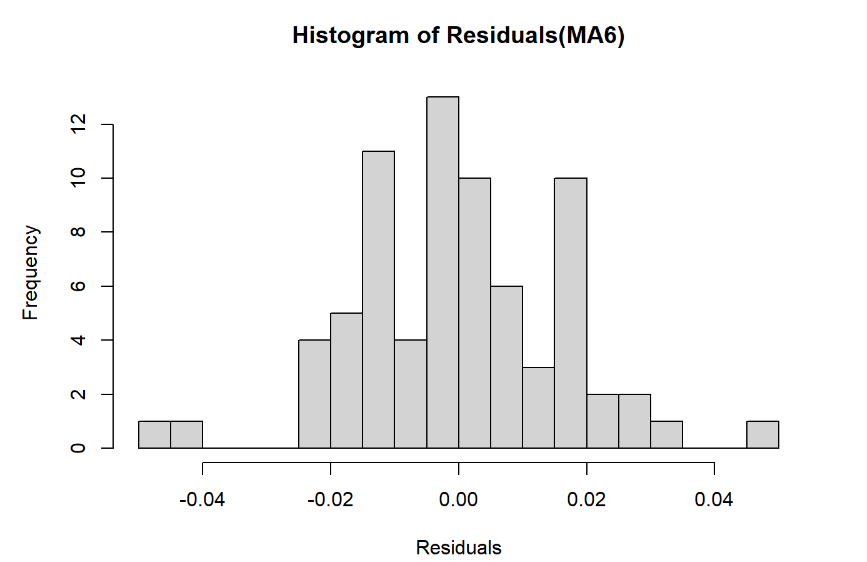
**Characteristic**: The high value for beta indicates that the model reacts very quickly to changes in the trend, making it highly responsive to recent shifts in the underlying data.

**Gamma (γ)**: 0.0001

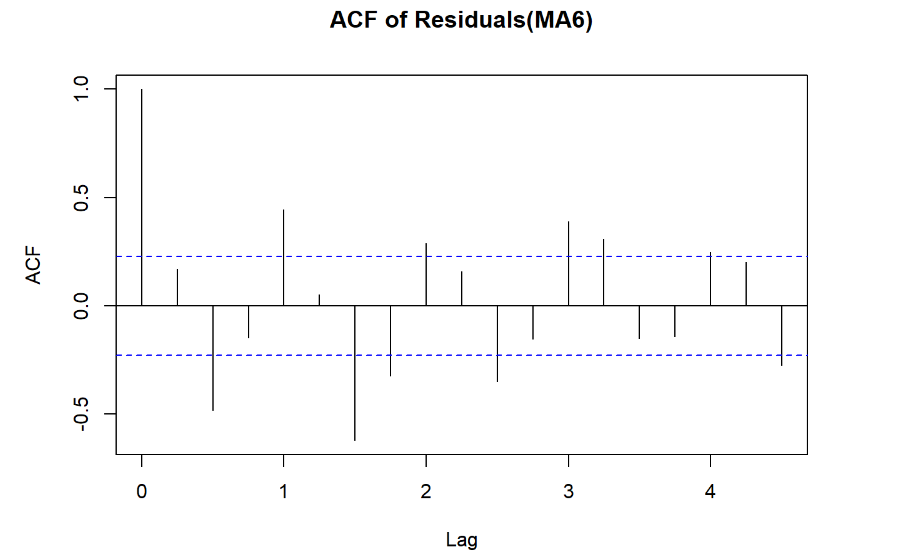
**Characteristic**: Similar to the previous model, the gamma value is very small, indicating minimal emphasis on seasonal changes, suggesting that the model does not adjust much for seasonality.



The plot shows clear patterns, including trends and cycles, which suggests the model may have missed some key aspects of the data.

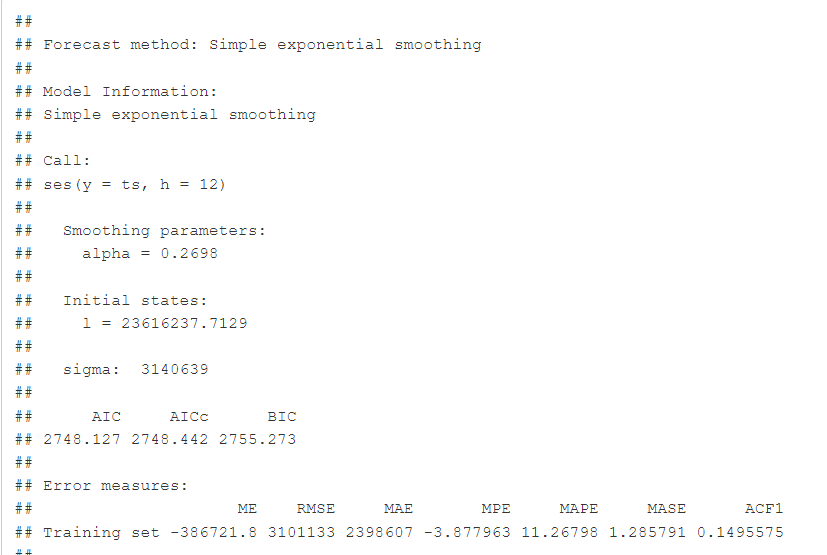


The histogram exhibits a non-normal distribution, pointing to the possibility that the model is not the best fit for the data.



There is autocorrelation present in the residuals at certain lags, indicating that the model has not captured all the dynamics within the data.

**Simple Exponential Smoothing:**



* **Characteristics:** A model that averages past data and applies a smoothing factor (alpha). This method is simpler and does not explicitly model trend or seasonality.
* **Strengths:** Simple and good for data without strong trends or seasonality.
* **Weaknesses:** Does not capture seasonality or trend, which limits its application on seasonal or trending data.
* **Observations:** The forecast is more stable and does not capture large fluctuations, with the forecast for 2021 Q1 remaining at 15,268,514. The method smooths out the data but does not incorporate any significant seasonality, making it less flexible for complex data.

**Parameters:**

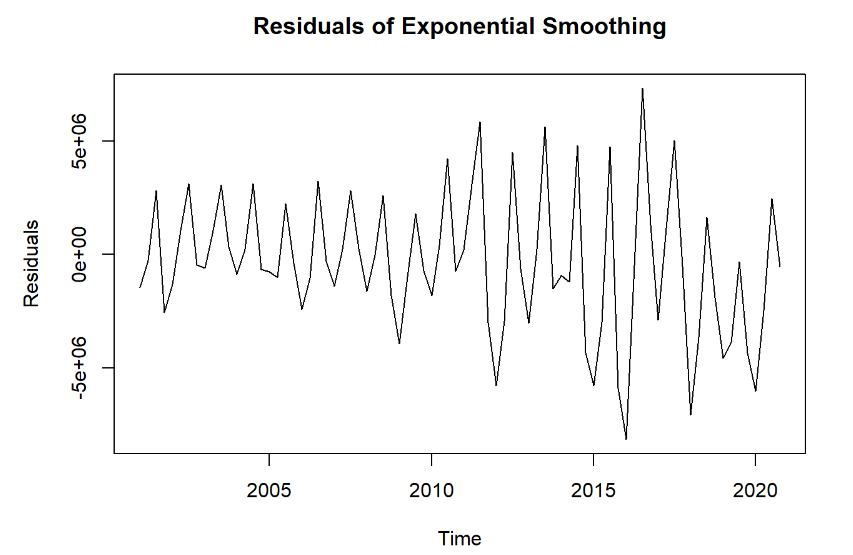
**Alpha (α)**: 0.2698

**Characteristic**: The alpha value is relatively low, indicating that the model places more importance on the long-term historical data and less on recent data, leading to a smoother forecast.

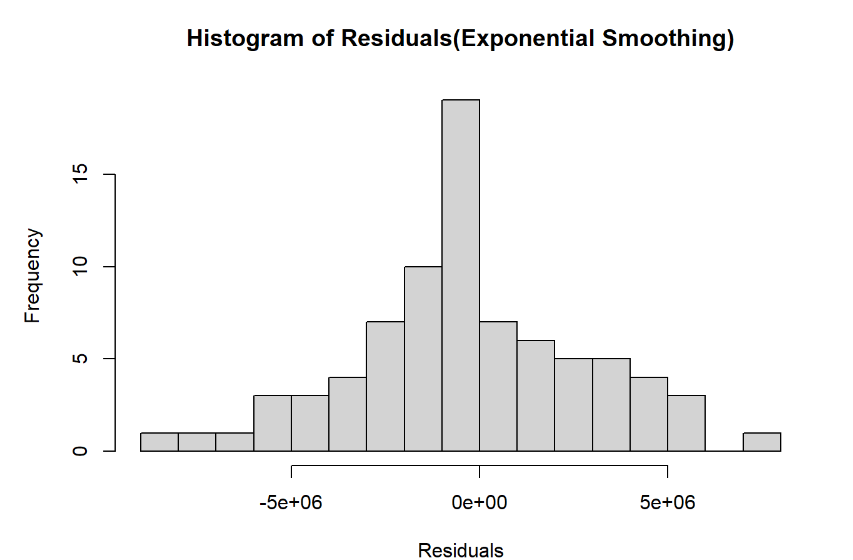
**Beta (β)**: Not used (since no trend component is involved).

**Gamma (γ)**: Not used (since no seasonal component is involved).

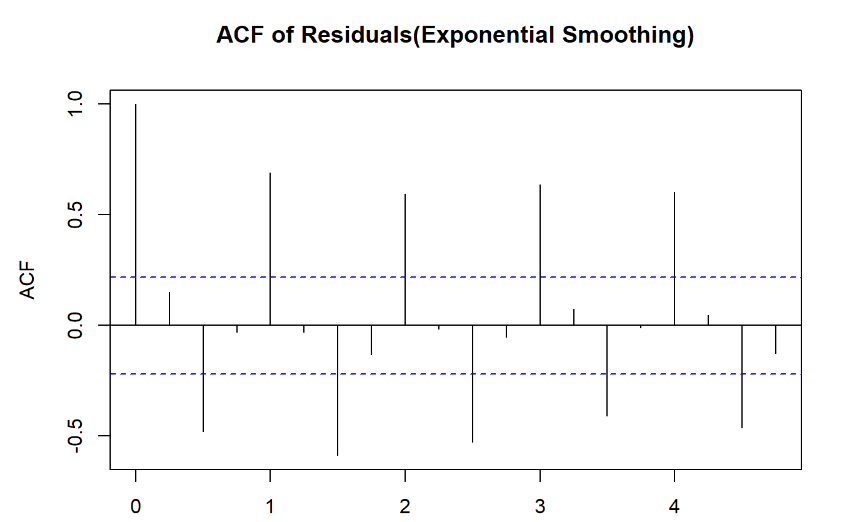
**Note**: SES models are primarily used for data without clear trend or seasonality, thus only alpha is significant here.



The plot displays identifiable patterns such as trends and cycles, which suggests the model may not have fully accounted for all aspects of the data.

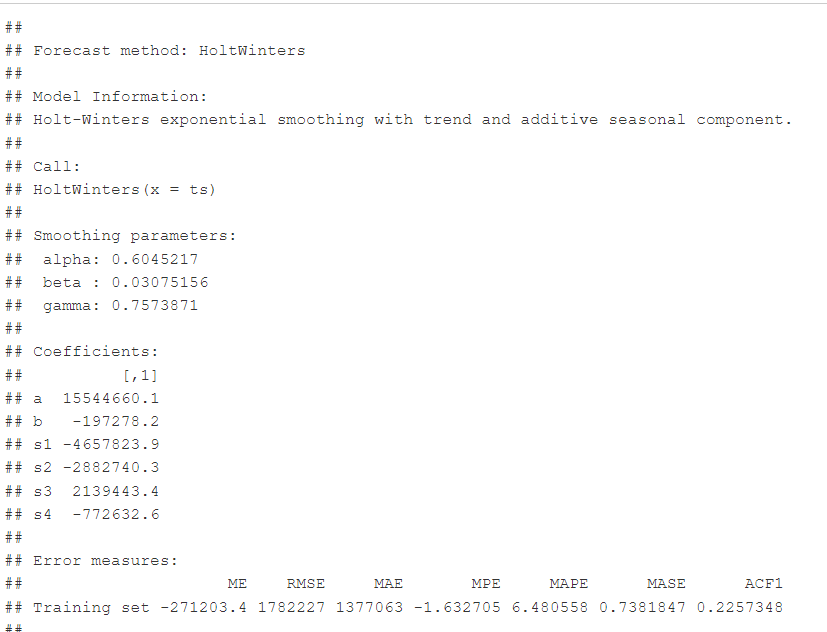


The histogram shows a non-normal shape, implying that the model might not be an appropriate fit for the data.



Autocorrelation is present in the residuals at specific lags, indicating that the model has failed to capture all the relevant patterns in the data.

**Holt-Winters Exponential Smoothing:**



* **Characteristics:** Accounts for both trend and seasonality by smoothing past values with parameters (alpha, beta, gamma) for level, trend, and seasonality.
* **Strengths:** Excellent for capturing data with both trends and seasonality.
* **Weaknesses:** Computationally more complex and requires fine-tuning of parameters.
* **Observations:** Forecasts fluctuate considerably, reflecting both trend and seasonality. For example, the forecast for 2021 Q1 is 10,689,558 and shows periodic changes, capturing seasonal effects better than other methods. The forecast is generally more aligned with real-world fluctuations.

**Parameters:**

**Alpha (α)**: 0.6045

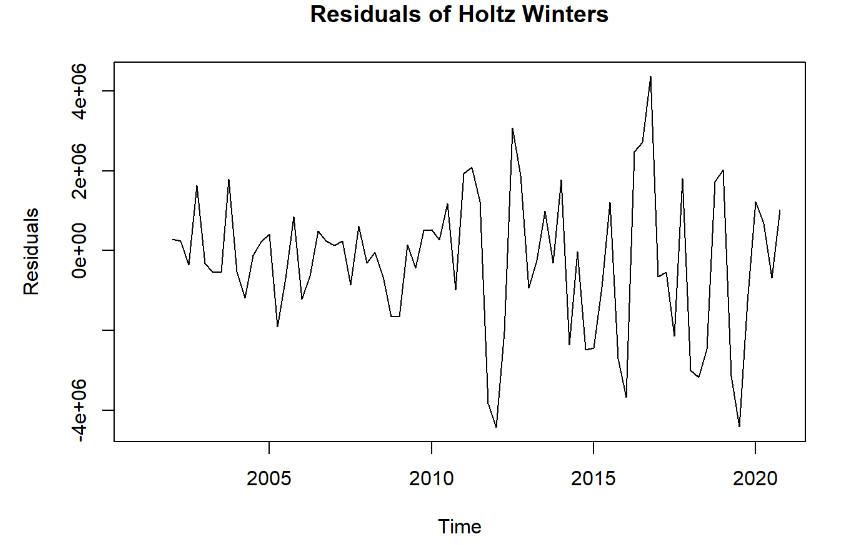
**Characteristic**: This alpha value indicates a moderate smoothing of the level, giving weight to recent observations but still accounting for past data. It strikes a balance between responsiveness and smoothing.

**Beta (β)**: 0.0308

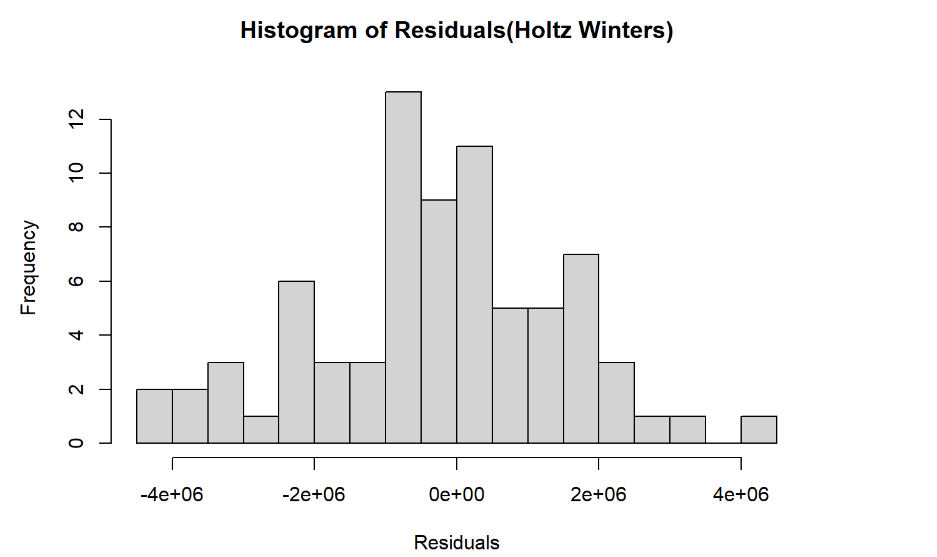
**Characteristic**: The small beta value suggests that the trend is not heavily weighted and changes slowly over time, meaning the model does not quickly adjust to sharp changes in trend.

**Gamma (γ)**: 0.7574

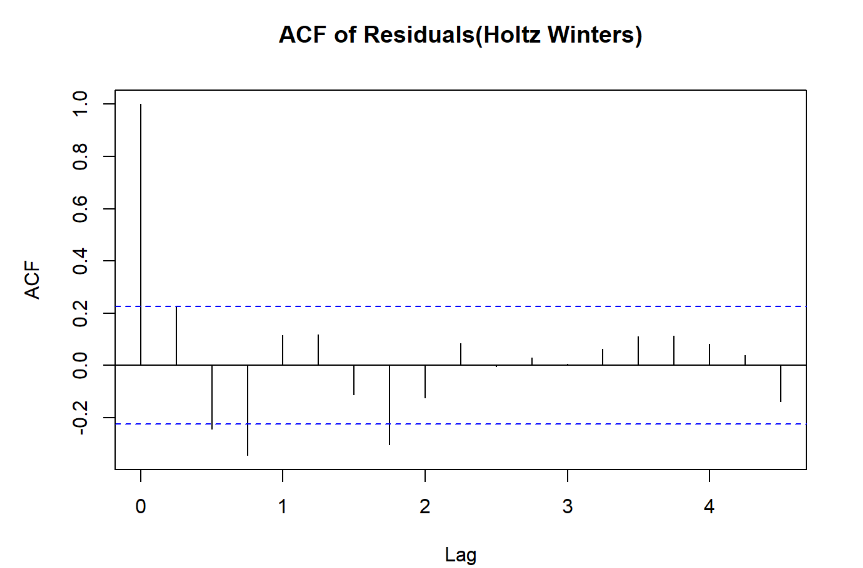
**Characteristic**: The gamma value is large, reflecting a strong emphasis on seasonal components, allowing the model to adjust quickly to seasonal fluctuations and more accurately forecast data with seasonal patterns.



The plot shows some patterns, such as trends and cycles, suggesting that the model captures part of the data's structure, but not all of it.

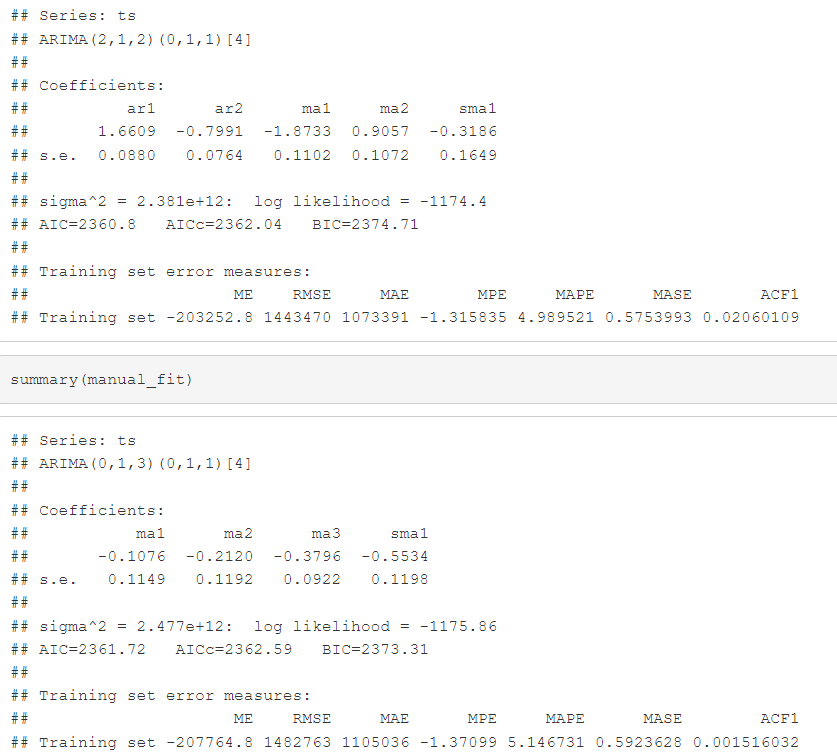


The histogram is somewhat close to normal, indicating that the model fits the data to some extent, though there may still be room for improvement.

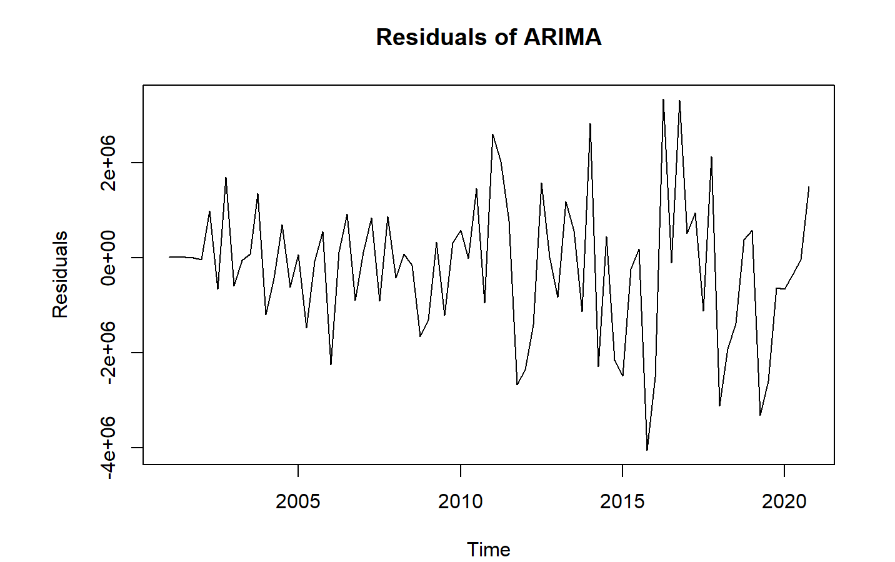


There is slight autocorrelation in the residuals at certain lags, implying that the model captures most patterns in the data, but might have missed a few details.

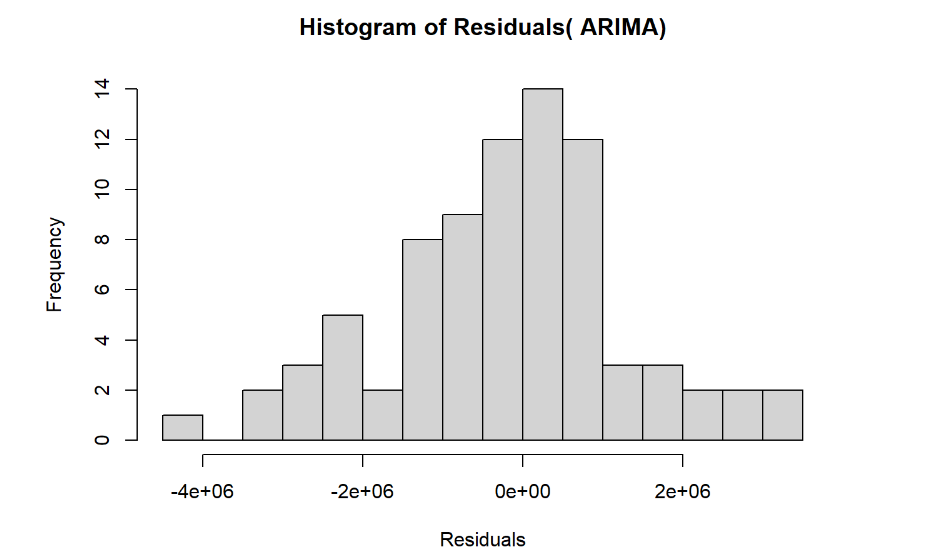
**ARIMA:**

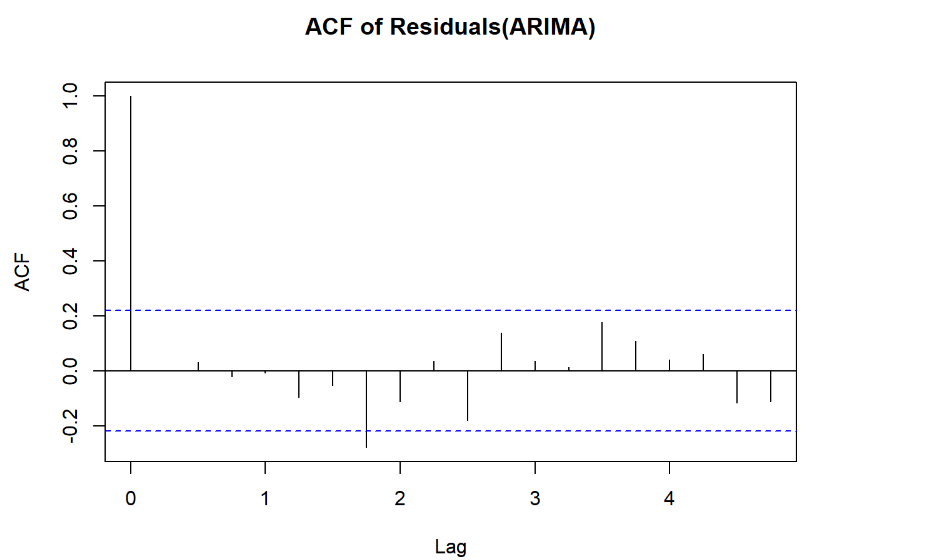


* **Characteristics:** Autoregressive Integrated Moving Average (ARIMA) uses past values and their linear relationships (AR terms), moving averages (MA terms), and differencing (I term) to forecast future values.
* **Strengths:** Powerful and flexible model, especially for time series with a strong autocorrelation structure.
* **Weaknesses:** Assumes stationarity and requires careful model selection.
* **Observations:** The ARIMA model offers a dynamic forecast that adjusts based on past data, with forecast values showing a reasonable balance between trend and seasonality. The model shows forecasts starting at 13,534,879 for 2021 Q1, with later forecasts reflecting the integrated seasonal and trend adjustments.



The plot clearly shows that the residuals are distributed randomly without any significant pattern indicating an excellent fit to the data.

  
The histogram closely follows a normal distribution, suggesting that the model is a very good fit for the data.



There is no significant autocorrelation in the residuals, indicating that the model has mostly captured the underlying structure of the data.

1. **Prediction and Accuracy summary from different forecasting methods**

**Naive Method:**

* **Predictions:** The forecast remains constant at 14,869,537 for all future quarters, showing no trend or seasonality adjustments.
* **Accuracy:** The RMSE and MAPE values are quite high (RMSE = 3,549,478 and MAPE = 13.28%), indicating a poor fit for this data, especially considering the presence of trends and seasonality.

**ETS (A,Ad,A) Moving Average 3:**

* **Predictions:** The forecast varies each quarter, reflecting the seasonal and trend components. For instance, the 2021 Q1 forecast is 13,534,879, and it increases over time.
* **Accuracy:** The RMSE is 821,534 and MAPE is 2.87%, showing much better accuracy than the naive method. This method accounts for trends and seasonality, improving prediction quality.

**ETS (M,A,A) Moving Average 6:**

* **Predictions:** Similar to the ETS method with 3-period moving average but smoother due to the 6-period smoothing factor. The 2021 Q1 forecast is 12,476,669.
* **Accuracy:** The RMSE is 380,682 and MAPE is 1.26%, indicating high accuracy, especially compared to both naive and simple exponential smoothing methods.

**Simple Exponential Smoothing:**

* **Predictions:** The forecast is constant at 15,268,514 for each quarter, showing limited sensitivity to seasonal trends.
* **Accuracy:** The RMSE is 3,101,133 and MAPE is 11.27%, which is better than the naive method but worse than the ETS methods, reflecting its lack of adaptation to seasonality and trends.

**Holt-Winters Exponential Smoothing:**

* **Predictions:** The forecast fluctuates significantly, reflecting both seasonal and trend components. The forecast for 2021 Q1 is 10,689,558, showing periodic variation.
* **Accuracy:** The RMSE is 1,782,227 and MAPE is 6.48%, indicating reasonable accuracy with good seasonal adaptation.

**ARIMA:**

* **Predictions:** The ARIMA model adjusts its forecast based on past data with a significant amount of variation for each period, capturing both trend and seasonal effects. The forecast for 2021 Q1 is 13,534,879.
* **Accuracy:** The RMSE is 1,443,470 and MAPE is 4.99%, showing good predictive performance, though slightly less accurate than ETS models due to the complex nature of ARIMA and possible overfitting.

1. **State your decision based on the analysis**

Here is the ranking of the models based on their forecasting accuracy and error measures, from best to least effective:

1. **ARIMA(2,1,2)(0,1,1)[4]**

Strong performance with low RMSE value. The model captures trend and seasonality effectively, with a relatively low AIC and BIC, indicating a well-fitting model for the data.

1. **ARIMA(0,1,3)(0,1,1)[4]**

Like the first ARIMA model but with a slightly less optimal fit as indicated by a higher AIC and BIC. Still, the error measures are relatively low, and it handles seasonality well.

1. **Holt-Winters**

Performs well with trend and seasonal components. While its RMSE is higher than ARIMA, it still provides valuable forecasts. The smoothing parameters (alpha, beta, gamma) help the model adjust to seasonal fluctuations, although its accuracy could be slightly better.

1. **ETS(A,Ad,A) Moving Average 3**

This model also shows good performance, but its RMSE is slightly worse compared to Holt-Winters and ARIMA models. However, it is still reliable for medium-term forecasting with moderate error rates.

1. **ETS(M,A,A) Moving Average 6**

Slightly less accurate in terms of RMSE and MAPE compared to the above models. This model may not perform as well for capturing sudden changes in trends or seasonality but could still be useful for data with moderate seasonal components.

1. **Simple Exponential Smoothing**

The least accurate of all models with the highest RMSE. Simple exponential smoothing is generally less effective for data with strong seasonality or trend compared to models like ARIMA and ETS, which account for these components better.

1. **Naive Method**

Provides the least accurate forecasts with the highest RMSE, reflecting its inability to capture underlying trends or seasonality effectively. It is the simplest model and can be used for baseline comparisons but is not suitable for accurate long-term forecasting.

**Conclusion** :

The **ARIMA(2,1,2)(0,1,1)[4]** model ranks the highest in accuracy, followed by **ARIMA(0,1,3)(0,1,1)[4]** and **Holt-Winters**, while the **Naive method** provides the least accurate forecasts.

1. **Provide some ideas to improve your forecasts**

To improve forecasting accuracy, here are a few ideas:

1. **Feature Engineering**:
   1. Incorporate additional variables like temperature, holidays, or economic indicators to capture seasonal and cyclical patterns that may impact electricity consumption.
   2. Explore lag features, such as previous quarters’ consumption, to capture time-based dependencies better.
2. **Advanced Models**:
   1. Experiment with more complex models like **LSTM (Long Short-Term Memory)** neural networks, which are better suited for capturing long-term dependencies in time series data.
   2. **SARIMA** (Seasonal ARIMA) could also be considered if the seasonality pattern is more complex than what ARIMA models capture.
3. **Hyperparameter Tuning**:
   1. Fine-tune hyperparameters for models like ARIMA or ETS to avoid overfitting or underfitting. Use techniques like grid search or random search to find the optimal parameters.
4. **Model Ensembling**:
   1. Combine multiple models (e.g., ARIMA, ETS, Holt-Winters, and machine learning models) to form an ensemble forecast. Weighted averaging or stacking can improve robustness and accuracy.
5. **Cross-Validation**:
   1. Use **time-series cross-validation** instead of simple training/test splits to ensure the model generalizes well to unseen data. This method respects the time dependencies between observations.
6. **Data Preprocessing**:
   1. Apply data smoothing or transformation techniques like **Box-Cox** to stabilize variance and remove trends in the data for more accurate forecasting.
   2. Address outliers or missing values by imputation or robust techniques to avoid model distortions.
7. **Error Analysis and Model Refinement**:
   1. Analyze residuals (error terms) to ensure there are no patterns left unexplained by the model. If residuals show trends or cycles, consider incorporating them into the model or trying a different approach.