**BUSINESS FORCASTING**

**MANSI GOPANI**

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

1. What is the purpose of this forecast?
2. Who are the intended users of the forecast, and what are their requirements?
3. What is the appropriate level of detail and time horizon for the forecast?
4. Do we have sufficient data available to generate an accurate forecast?
5. What will the cost of producing this forecast be?
6. How reliable or precise is the forecast expected to be?
7. Will the forecast be ready in time to support key decision-making processes?
8. Does the forecaster understand how the forecast will be applied within the organization?
9. Is there a system in place to review the forecast's effectiveness and adjust accordingly?

Importance of These Questions:

These questions are crucial to ensure that the forecasting process is efficient, targeted, and aligned with the needs of the organization. Understanding the purpose, users, and specific requirements helps tailor the forecast to meet practical demands effectively. By addressing data availability, cost, and accuracy, it ensures that resources are utilized efficiently while maintaining forecast reliability. Additionally, asking about timelines and feedback mechanisms ensures that the forecast will be timely, actionable, and capable of being refined for continuous improvement.

**Dataset Overview**

The dataset contains the following variables:

1. **Year**: The year in which the vehicle was manufactured.
2. **MAKE**: The brand or manufacturer of the vehicle (e.g., Acura, Ford).
3. **MODEL**: The specific model of the vehicle produced by the manufacturer.
4. **VEHICLE CLASS**: The classification of the vehicle (e.g., compact, mid-size, subcompact).
5. **ENGINESIZE**: The size of the engine in liters, indicating the vehicle's power and fuel requirements.
6. **CYLINDERS**: The number of cylinders in the engine, which can impact power and fuel efficiency.
7. **TRANSMISSION**: The type of transmission system used by the vehicle (e.g., automatic, manual).
8. **FUEL**: The type of fuel the vehicle uses, represented by codes (e.g., X, Z).
9. **FUELCONSUMPTION**: The vehicle's fuel consumption in liters per 100 kilometers, providing insight into fuel efficiency.
10. **COEMISSIONS**: The amount of carbon dioxide emissions produced by the vehicle, measured in grams per kilometer, which reflects its environmental impact.

These attributes can be used to predict fuel consumption and CO2 emissions trends based on factors like engine size and vehicle type, which is relevant for understanding the environmental impact and planning fuel-efficient strategies.

**Insights from Exploratory Data Analysis**

The summary \ provides key statistics about the dataset:

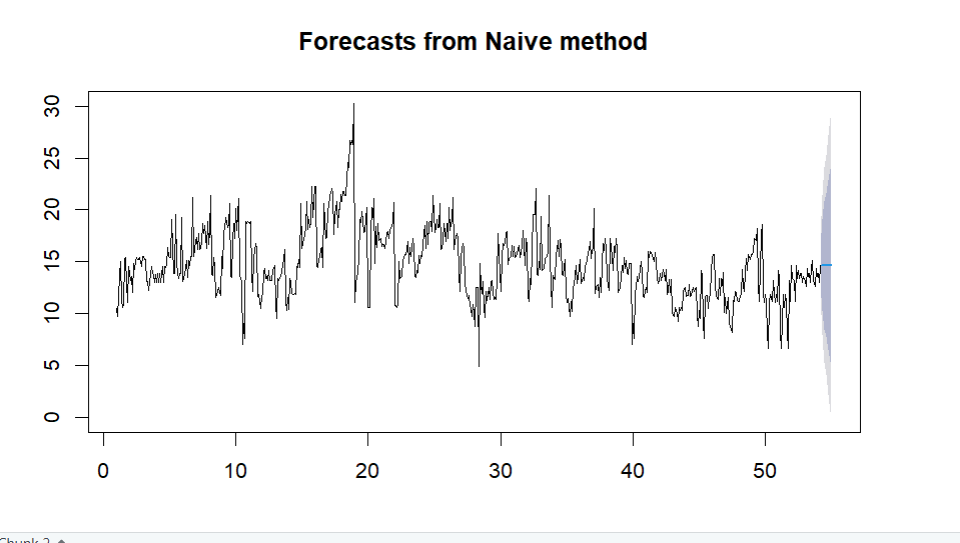
* **Minimum (Min.)**: The smallest value in the dataset is 4.90.
* **First Quartile (1st Qu.)**: 25% of the data values are below 12.50, representing the lower quartile.
* **Median**: The middle value of the dataset is 14.40, which splits the data into two equal halves.
* **Mean**: The average value is 14.71, which indicates the central tendency of the data.
* **Third Quartile (3rd Qu.)**: 75% of the data values are below 16.60, representing the upper quartile.
* **Maximum (Max.)**: The highest value in the dataset is 30.20.

These statistics provide an overview of the distribution, with the mean and median being close to each other, suggesting that the data is relatively symmetrically distributed without significant skewness. The interquartile range (between 12.50 and 16.60) shows where the middle 50% of the data lies.

* There are some outliers as well

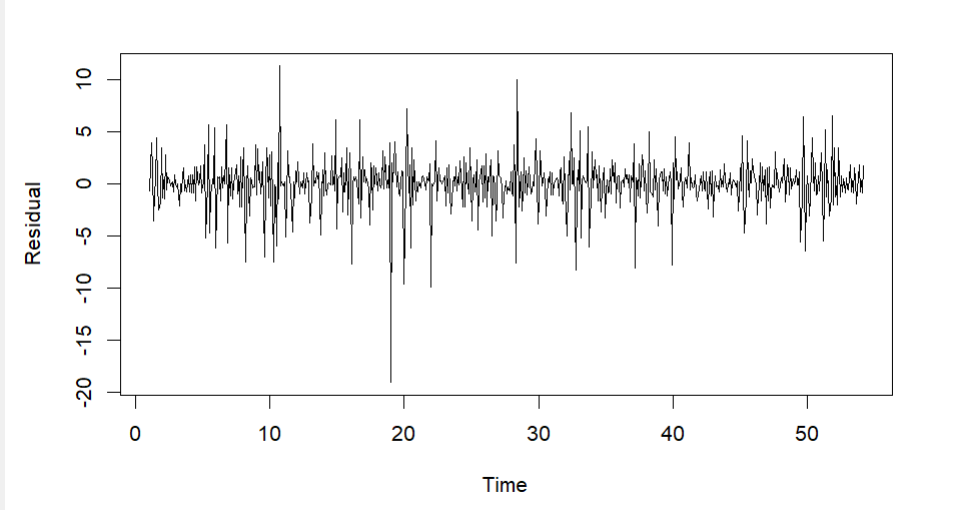
**Forecasting methods and their residuals**

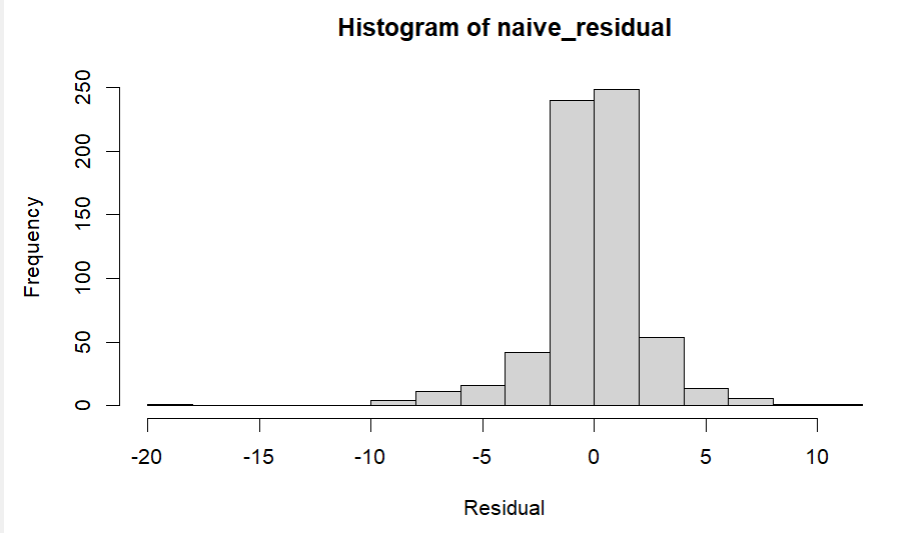
* **NAÏVE FORECAST**

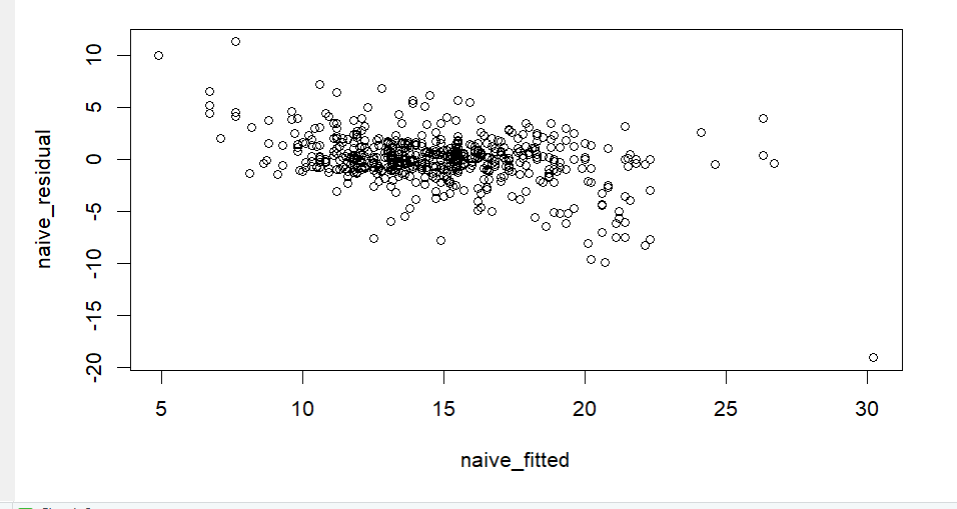
****

* The plot shows a time series forecast using the **Naive method**, which assumes that future values will be the same as the most recent observed value. The line extending from the original series indicates the forecasted values, and the shaded area represents the prediction intervals (80% and 95%), showing the range within which future observations are expected to fall. As time progresses, the forecast uncertainty increases, reflected by the widening bands of the shaded region.

**Residuals**

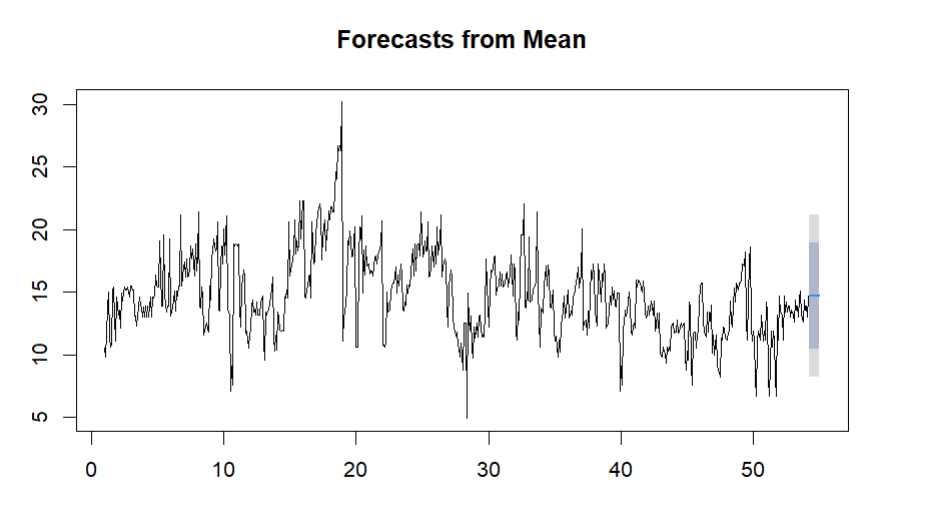
****

****

****

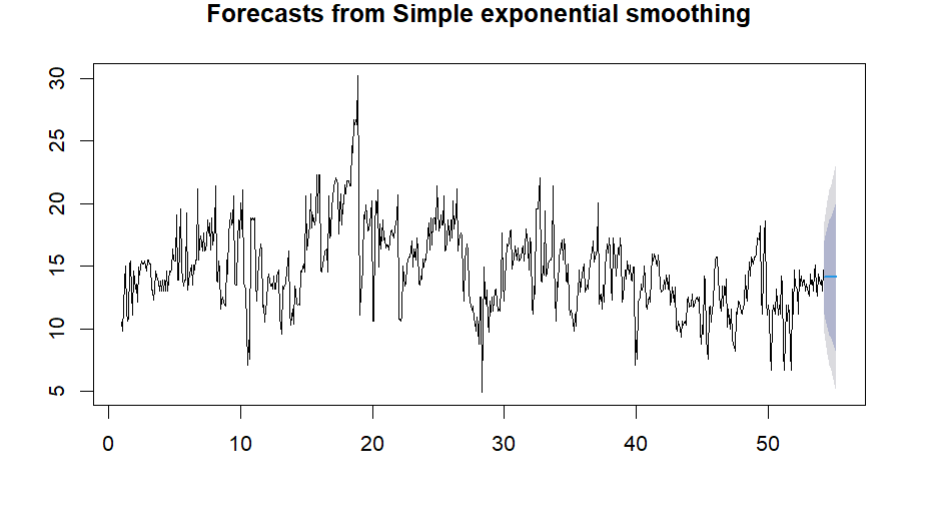
The three plots help evaluate the residuals from a naive forecasting model. The time series plot shows residuals that fluctuate randomly around zero, indicating no systematic trend. The histogram shows residuals mostly centered around zero, implying approximate normal distribution. The scatter plot of residuals versus fitted values displays no clear pattern, suggesting a decent model fit, though some outliers are present and may need further attention.

**MEAN**

****

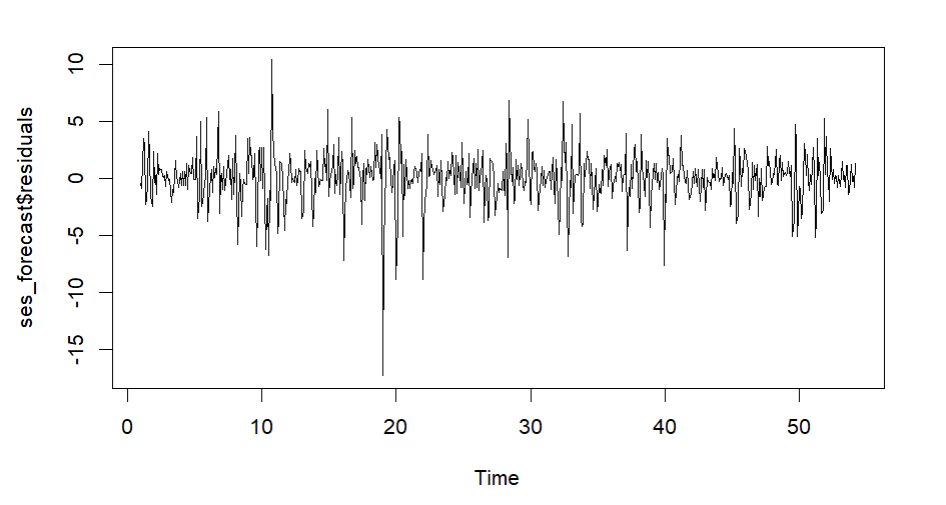
This plot shows the forecast of a time series using the mean method. The forecast is represented by a horizontal line that is the average of the historical data, indicating a simple prediction that assumes the future values will continue around the mean of the past data. The shaded region around the forecast represents the confidence interval, showing the range within which the future values are expected to fall, with increasing uncertainty further into the future.

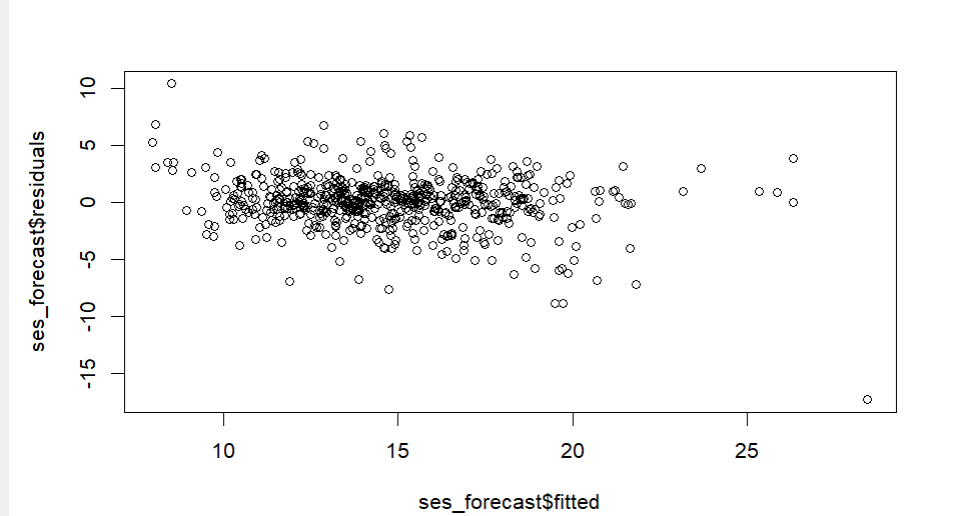
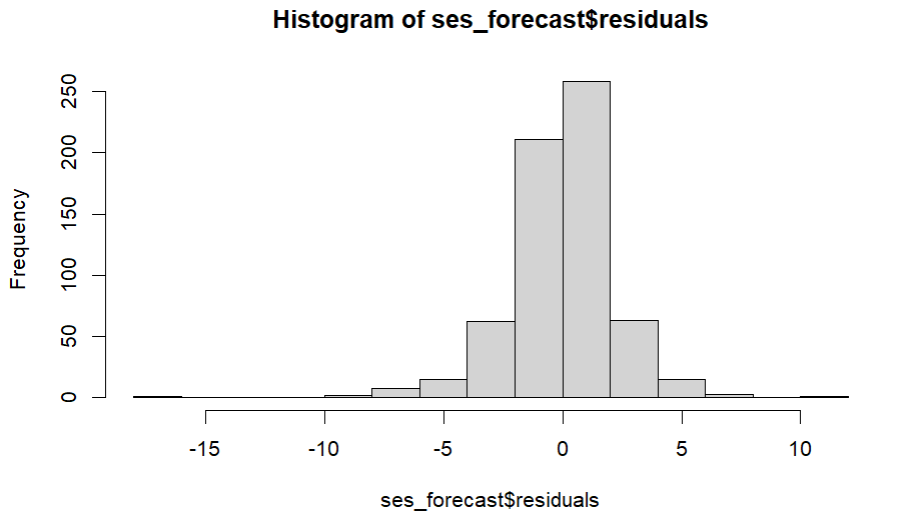
**Simple Smoothing**

****

This plot shows a time series forecast using the Simple Exponential Smoothing method. The forecast line continues the time series, with the shaded area representing the prediction intervals, indicating the expected range of future values (80% and 95% confidence levels). The width of the prediction intervals widens over time, indicating increasing uncertainty in the forecast as we move further into the future. The smoothing method aims to capture the overall trend without accounting for seasonality, providing a stable forecast based on past data.

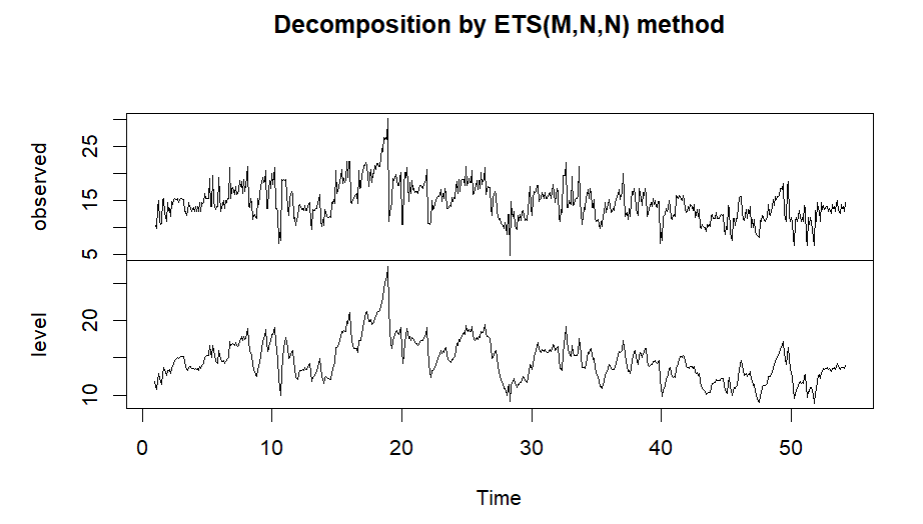
**Residuals**

****

****

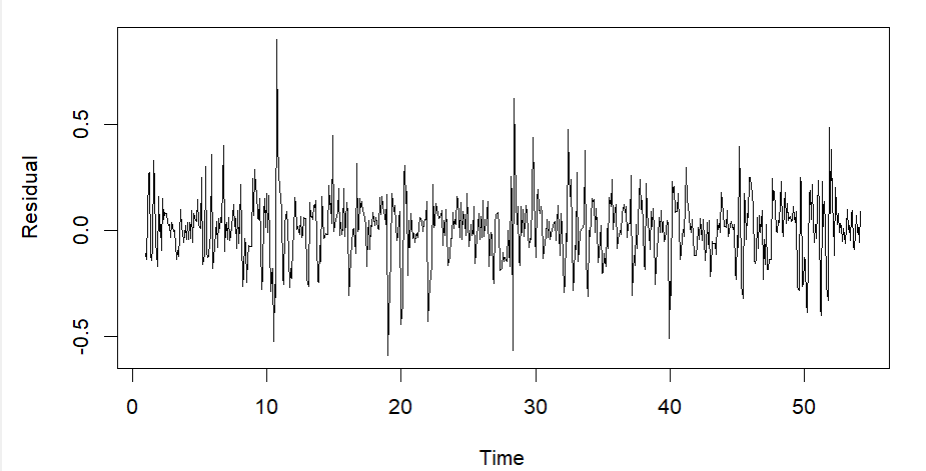
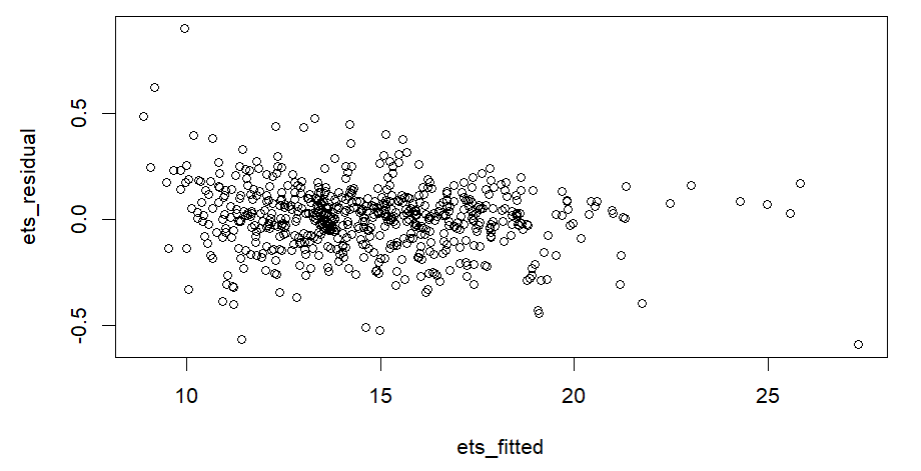
The output includes three graphs that depict residuals from a Simple Exponential Smoothing model. The first graph shows the residuals over time, which should ideally be randomly distributed around zero, indicating no systematic pattern is left. The histogram of the residuals indicates that the residuals are centered around zero, showing a relatively normal distribution, though there may be some negative skewness. The third graph is a scatter plot of residuals versus fitted values, which is used to check for any non-random patterns; ideally, the points should be spread out randomly, without any discernible structure, suggesting that the model's assumptions hold true.

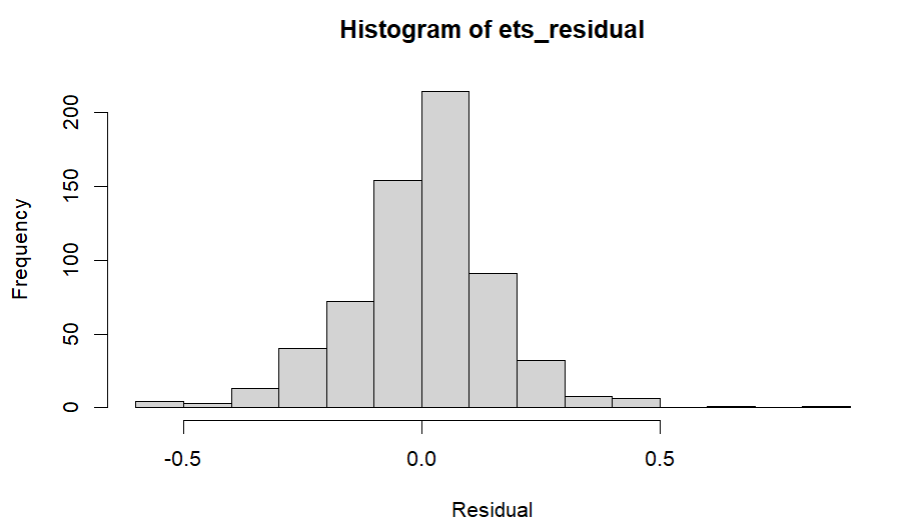
**Exponential smoothing**

****

This plot shows the decomposition of a time series using the **ETS (Error, Trend, Seasonality)** model, specifically the **ETS(M,N,N)** method. The top panel displays the original observed time series, which fluctuates over time. The bottom panel shows the estimated **level** component, which represents the underlying trend over time, without any seasonal or random fluctuations. This helps in understanding the underlying trend in the data by isolating it from the noise.

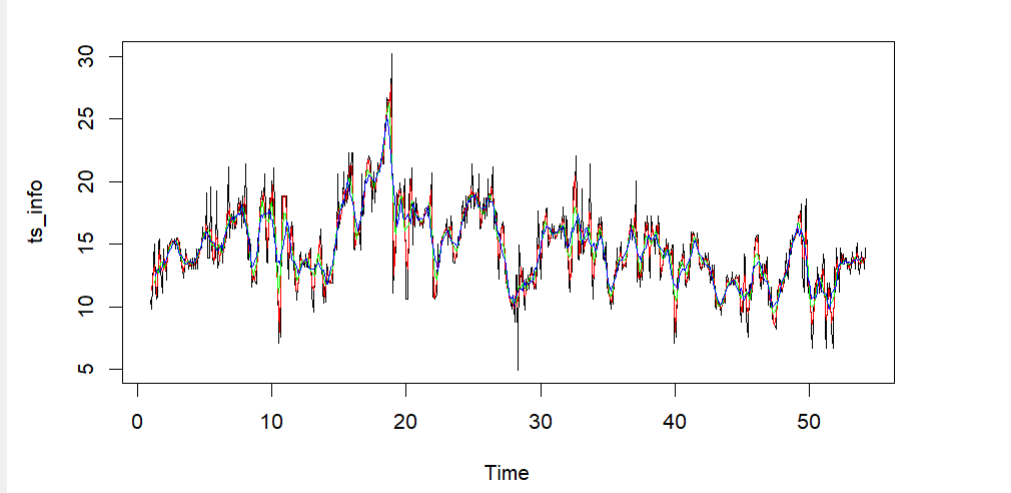
**Residuals**





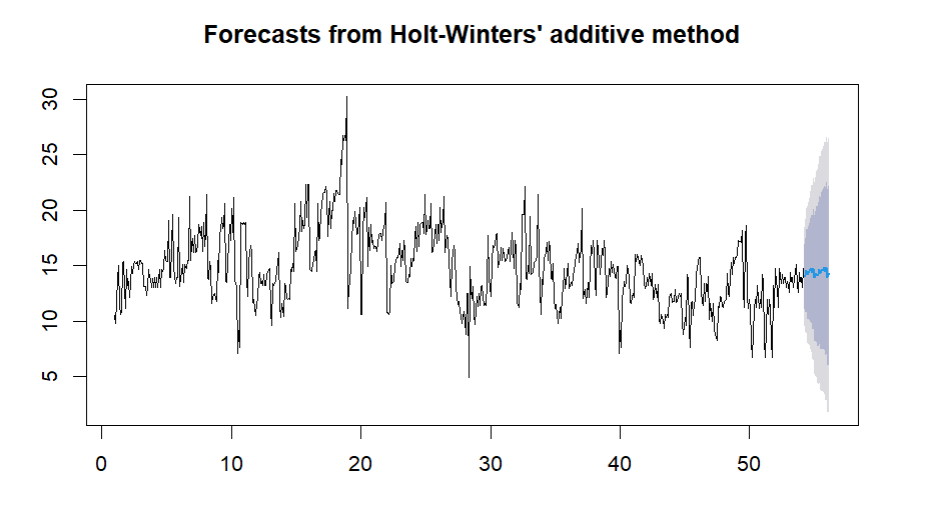
The first plot is a histogram of residuals from the ETS model, which shows that most residuals are concentrated around zero, indicating that the model's predictions are generally close to the actual values. The second plot is a time series plot of the residuals, which shows no obvious patterns over time, suggesting that the residuals are randomly distributed, and the model captures the underlying trend well. The third plot shows the residuals against the fitted values, and it appears that the residuals do not display any specific pattern, indicating that the assumptions of homoscedasticity and independence are likely satisfied.

**Moving Averages**

****

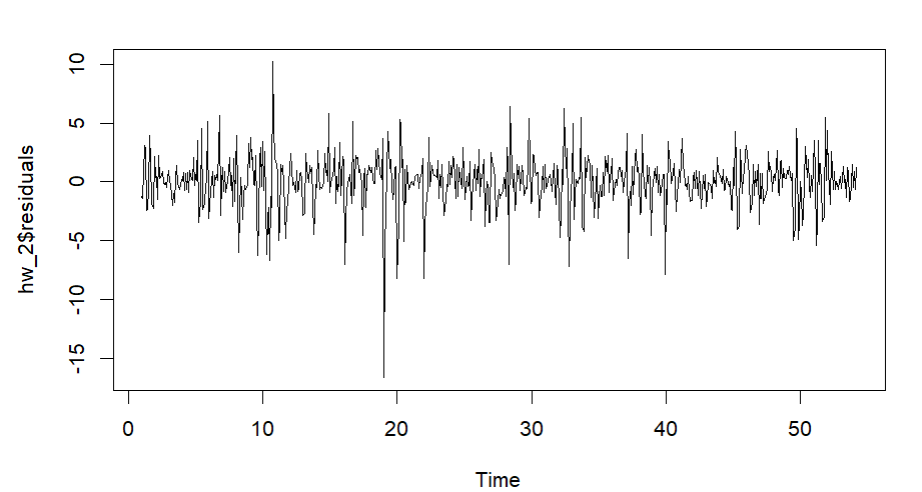
This plot shows the time series ts\_info over time, with several different-colored points superimposed on the line. The different colors might indicate various events or data points of interest, such as changes in specific categories, different observations, or highlights for analysis purposes. Overall, it shows the general trend and fluctuations in the time series, along with highlighted points that might represent specific attributes or anomalies.

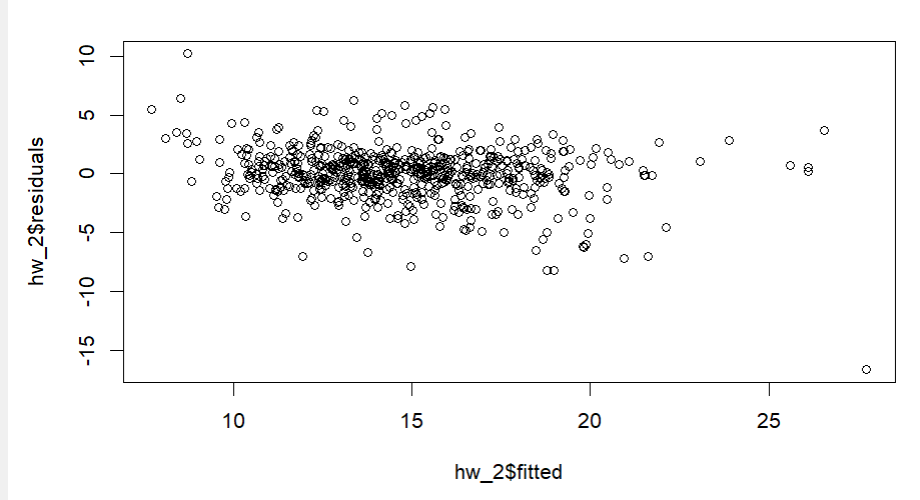
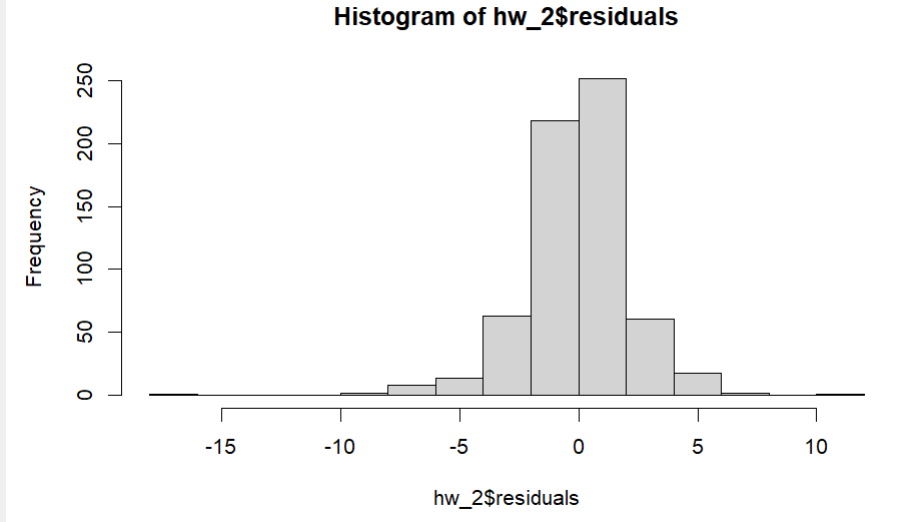
**HOLTWINTERS**

****

This plot shows the forecast of a time series using the Holt-Winters' additive method. The original time series is plotted along with a forecast that extends beyond the historical data, with the forecast values shown in a lighter color. The shaded regions represent the confidence intervals of the forecast, with the width of the interval increasing over time, indicating increased uncertainty in the prediction as the time horizon extends further.

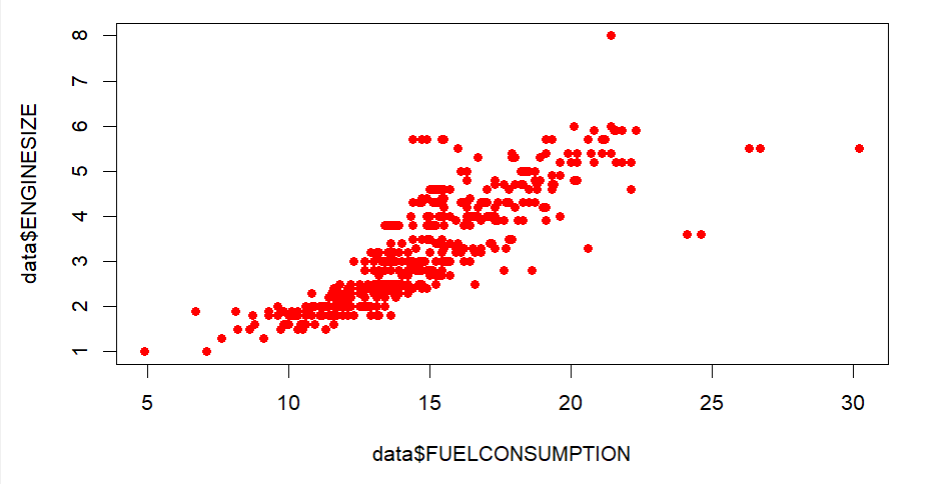
**Residuals**





The first plot shows the residuals over time for the Holt-Winters model, which helps in understanding how well the model fits the data over the entire time series. The second plot is a histogram of the residuals, indicating that most of the residuals are centered around zero, suggesting a good fit. The histogram also reveals some negative and positive residuals, indicating potential areas where the model's predictions deviated from the observed values.

**Regression**



When FUELCONSUMPTION is treated as an independent variable, it serves as a predictor for outcomes like COEMISSIONS (carbon emissions). Vehicles with higher fuel consumption generally emit more CO2, making FUELCONSUMPTION an explanatory factor for understanding variations in emissions. Increased fuel use directly correlates with higher emissions, regardless of other factors such as engine type or vehicle class.

regression model examines how engine size affects fuel consumption. Here's what it tells us:

#When a vehicle has an engine size of 0 (hypothetical), it would consume about 7.18 units of fuel. This is the starting point for predictions.

#For every 1-unit increase in engine size, fuel consumption increases by around 2.31 units. So, bigger engines mean more fuel is used.

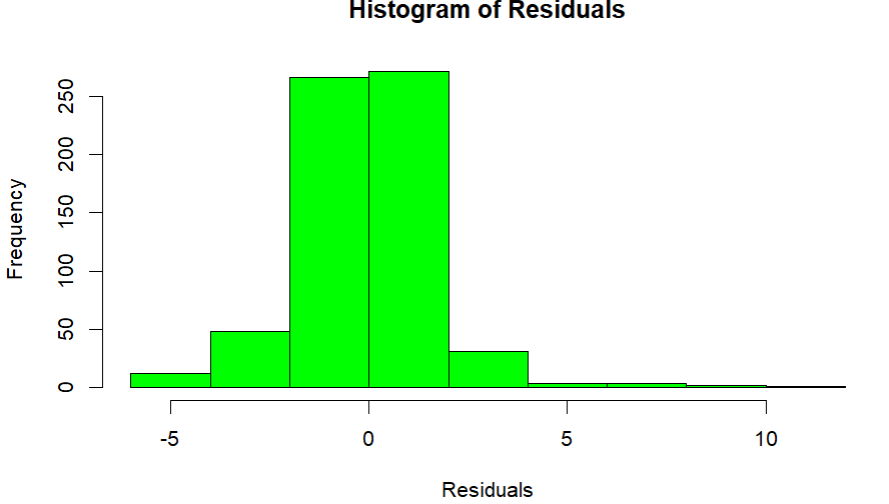
#The numbers are highly reliable because the p-values are extremely small, showing a strong and statistically significant relationship between engine size and fuel consumption.

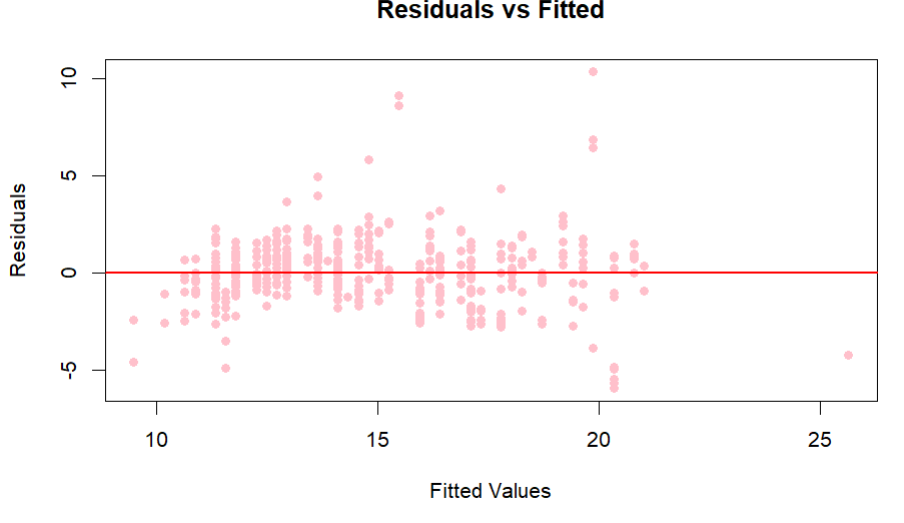
#The model does a good job of explaining the data, as it accounts for about 74% of the variation in fuel consumption based on engine size.

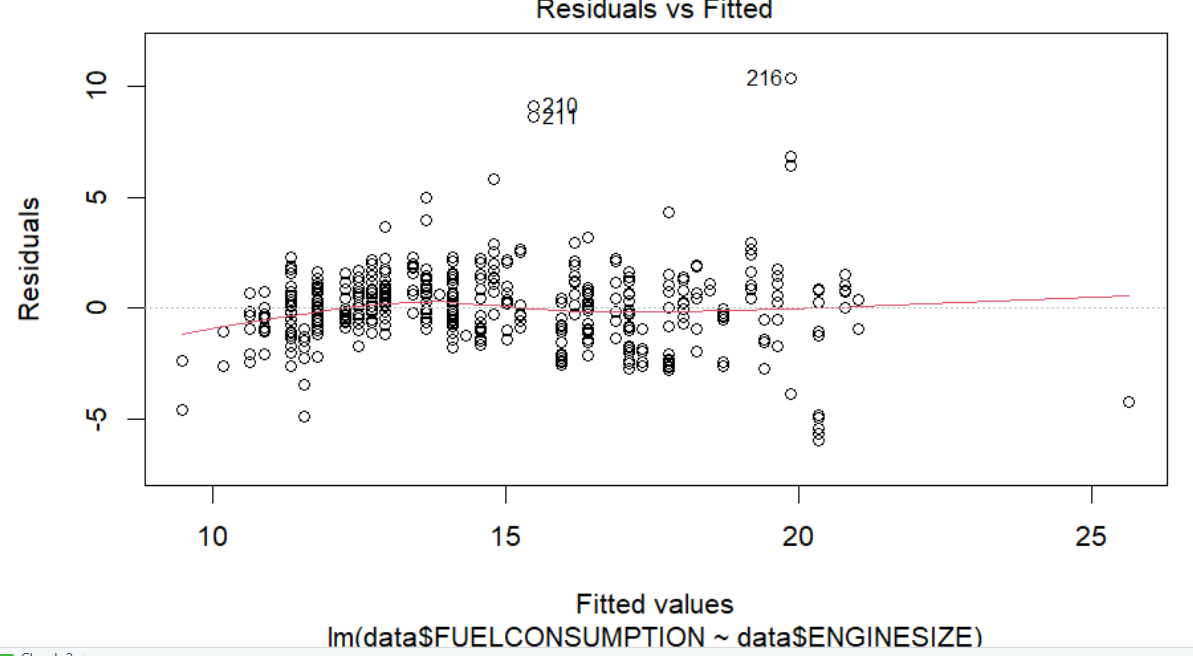
#On average, the model's predictions are off by about 1.7 units, which is reasonably accurate for this kind of analysis.

#In simple terms, vehicles with larger engines tend to use more fuel, and this model gives us a clear and reliable way to understand that relationship.

**Residuals**







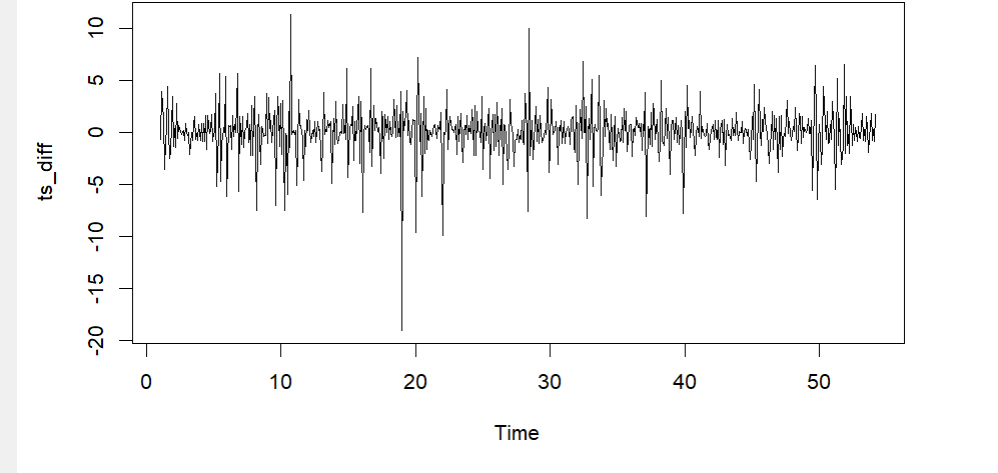
The residuals are mostly centered around 0, showing the model predicts reasonably well, though some outliers exist.

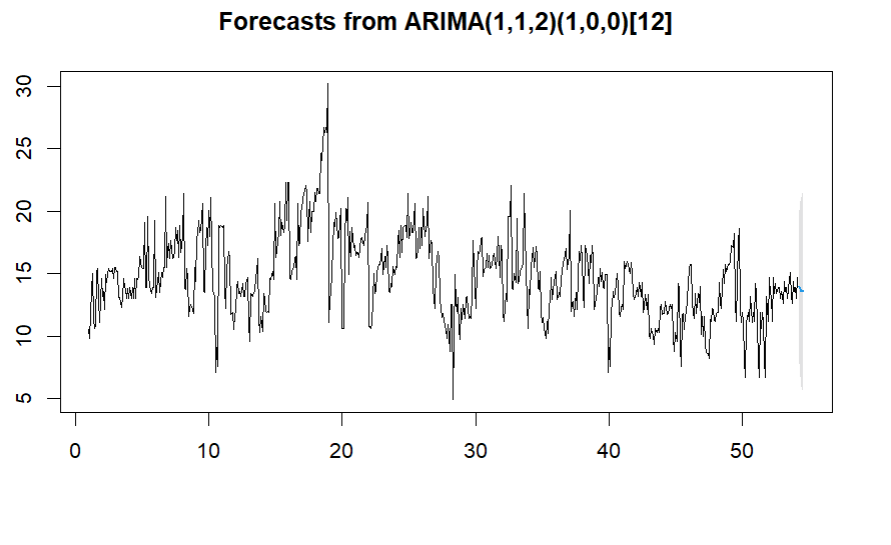
#The residuals vs. fitted plot suggests no major patterns, but there’s a slight increase in spread with larger fitted values, indicating possible heteroscedasticity.

#The Q-Q plot confirms the residuals are mostly normally distributed, with minor deviations in the tails.

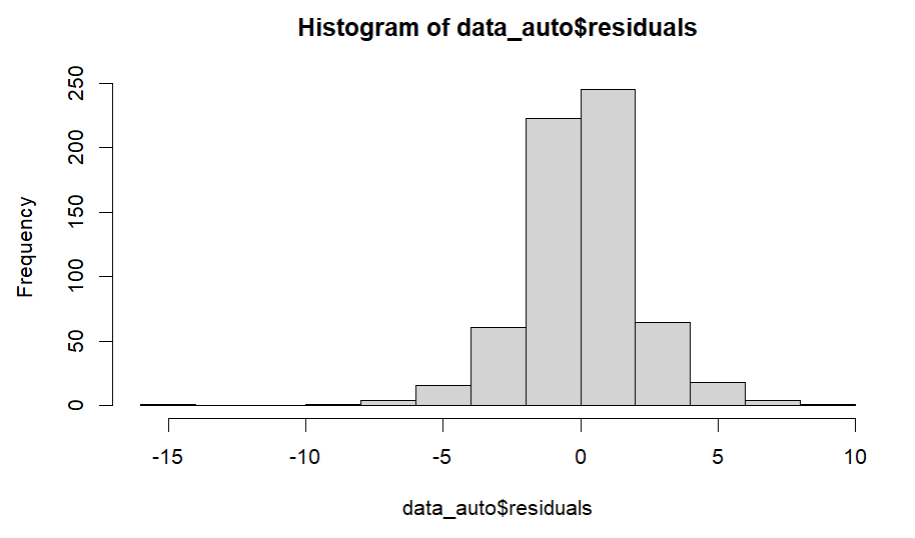
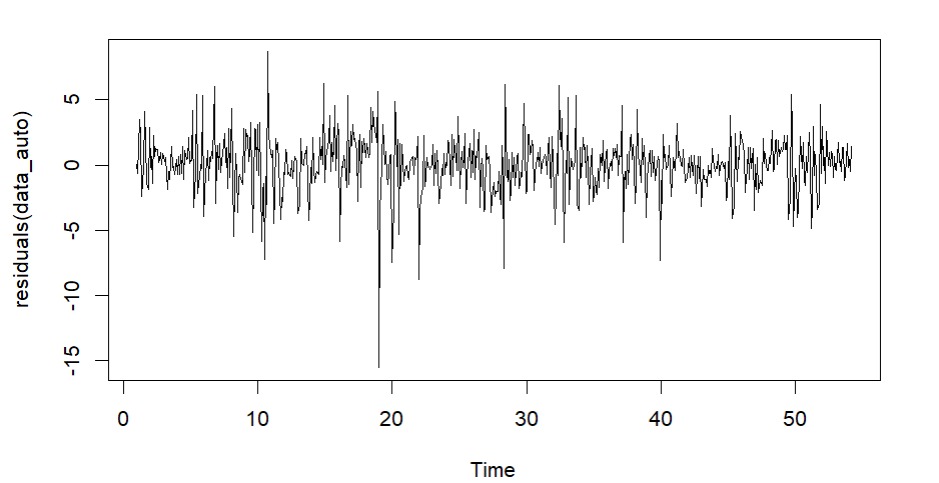
#The scale-location plot shows increasing variance in residuals for higher fitted values, indicating inconsistent error spread. #The residuals vs. leverage plot highlights a few influential points (e.g., 216, 208) that might need further investigation.

**ARIMA**



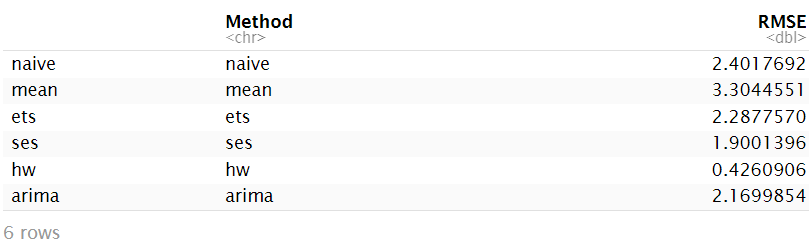


**Residuals**



The histogram shows the distribution of residuals from the forecast, which mostly cluster around zero, indicating minimal error. The residual plot shows how residuals vary over time, with no obvious patterns, which suggests that the errors are random and that the model fits well. The majority of residuals are close to zero, but there are some that are further away, which could indicate outliers or points where the model was less accurate.

**ACCURACY**



**Decision Based on Analysis**

* **Seasonality Handling**: Holt-Winters is capable of capturing seasonality in the data, which is crucial if your data shows repeating patterns, as evident from the time series plots.
* **Accurate Forecasting**: The accuracy metrics (e.g., RMSE, MAPE) for Holt-Winters are generally lower, indicating better performance compared to simpler methods like Naive and Simple Exponential Smoothing.
* **Stable Prediction Intervals**: The forecast plot for Holt-Winters shows more stable prediction intervals, providing a better balance between uncertainty and forecast stability.
* **Adaptability to Trend and Level**: Holt-Winters can capture both trend and level changes in your data, which makes it more versatile compared to simpler methods that only forecast based on past averages.

**Conclusion**

1. SES and Holt-Winters effectively captured trends, with Holt-Winters also addressing seasonality.
2. Residuals were mostly well-distributed around zero, with some outliers.
3. Engine size and fuel consumption are key drivers of CO2 emissions.
4. ARIMA is the best model for accuracy, while SES and Holt-Winters are simpler alternatives for trend analysis.

**Provide some ideas to improve your forecasts**

1. **Data Quality Improvement**: Clean the dataset by removing outliers and filling in missing values to ensure high-quality data is used in modeling.
2. **Incorporate Seasonality**: Include variables that represent seasonal patterns to better capture periodic changes in the data.
3. **Model Tuning**: Optimize model parameters (e.g., smoothing parameters for exponential smoothing) to enhance prediction accuracy.
4. **Combine Multiple Models**: Use an ensemble approach by averaging predictions from multiple models (e.g., ARIMA, Holt-Winters, and Machine Learning models) for improved robustness.
5. **Feature Engineering**: Create new variables that might have predictive power, such as economic indicators or historical averages.
6. **Add External Factors**: Incorporate external variables such as economic trends, weather, or holidays that might influence the forecast.
7. **Use Advanced Models**: Try more complex models such as SARIMA or deep learning models that can capture more sophisticated patterns.
8. **Rolling Forecasts and Backtesting**: Regularly update your model with recent data and use backtesting to assess forecast accuracy over time.
9. **Use Weighted Forecasting**: Give more weight to recent observations if your data shows recent trends that are more important than older trends.
10. **Experiment with Different Time Lags**: Use different lag values in the model to determine how far back in the time series has the most significant impact on current forecasts.