1. Introduction

The Monthly Cars Sales dataset illustrates the sales of cars in Quebec, Canada. It comprises 108 observations, with units representing the count of sales. Given that Canada experiences peak winters from November to February and summers typically begin in April or May, the data reveals a trend where a majority of car purchases occur during the summer months. The study will help to understand the months with the highest sales, detect any repeating patterns, and understand how consumers behaved in terms of car purchases. This will in turn help to find the key factors influencing the trends. Car companies can plan when to make cars by predicting how many they'll sell. This helps them make the manufacturing process work better, spend less money on production, and avoid making too many cars or not enough.

2. Gathering Data

We have selected the dataset named "Monthly car sales in Quebec" .We picked this dataset from the source : https://www.kaggle.com/datasets/dinirimameev/monthly-car-sales-in-quebec-1960/data.

The Monthly Car Sales dataset provides information on the sales of cars in Quebec, Canada from January 1960 to December 1968. There are a total of 108 observations.

2.1. Descriptive Analytics

There are two columns: "Month" and "Monthly Car Sales in Quebec 1960-1968."

Month	Monthly.car.sales.in.Quebec.1960.1968
1960-01	6550
1960-02	8728
1960-03	12026
1960-04	14395
1960-05	14587
1960-06	13791

The "Month" column contains the date in the format YYYY-MM. We have separated the ' Month' column into 'Year' and 'Month' columns where the format is YYYY and MM respectively. The "Year" and "Month" columns represent the year and month components of the original "Month" column, respectively. Each row still corresponds to a specific month. This separation makes it easier to analyze and filter the data based on the year and month components.

The second column, "Monthly Car Sales in Quebec 1960-1968," contains the corresponding monthly car sales figures. These figures represent the count of car sales for each specific month in the dataset. The data type of this column is numeric. We have renamed this column name to 'CarSales' for easier access.

Year	Month	CarSales
1960	01	6550
1960	02	8728
1960	03	12026
1960	04	14395
1960	05	14587
1960	06	13791

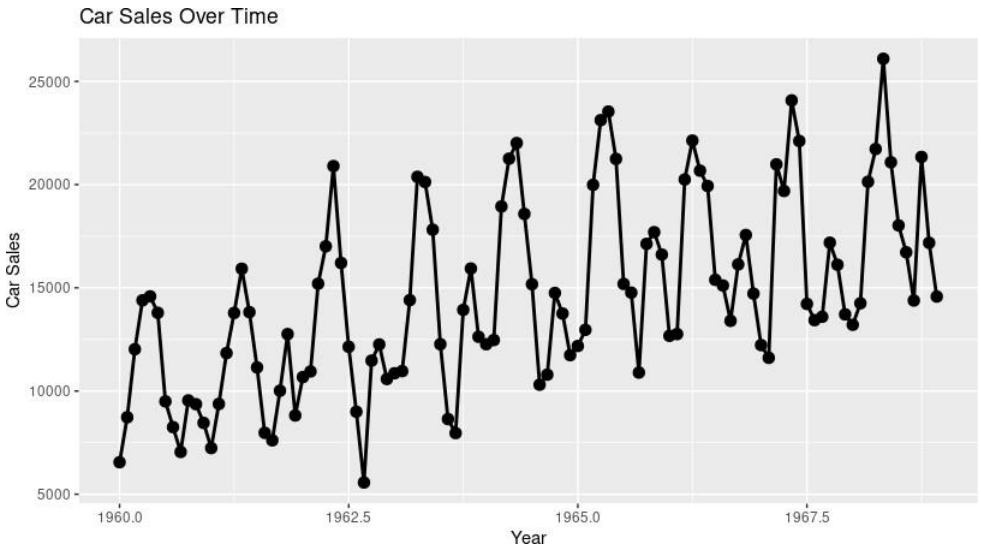
The dataset doesn't have any missing values.

3. Visualization

We now plot Time series plot, ACF and PACF plots in order to understand the data better.

Time Plot

We transform the data frame into a time series format and create a visual representation to make observations about trends, seasonality etc

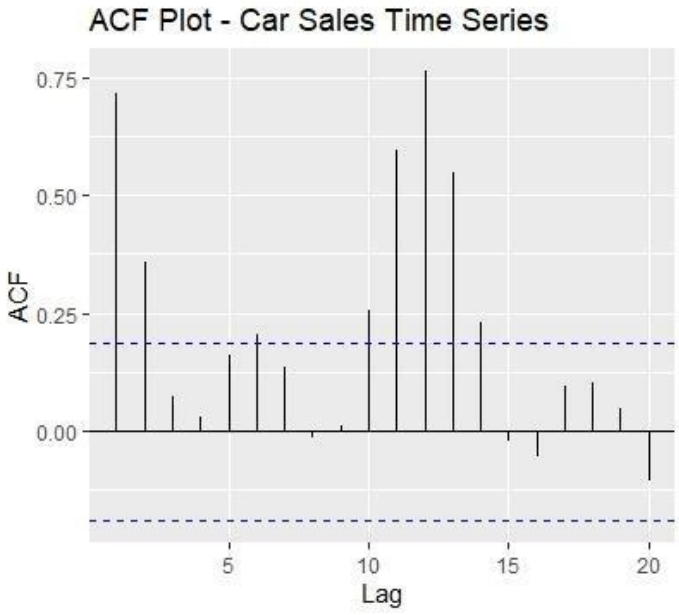


From the time plot, we can observe a pronounced upward trend. Additionally, there is a recurring pattern in the data where values rise and fall consistently at regular intervals, indicating a seasonal cycle.

ACF plot

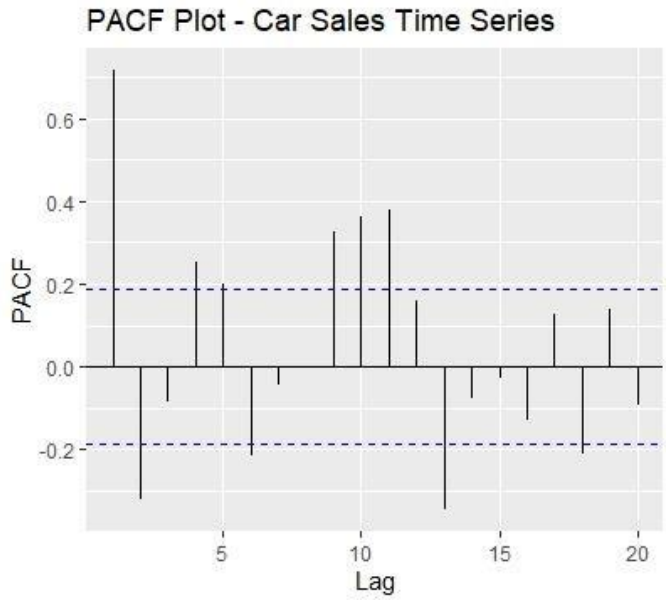
The relationship between a variable and its historical values is known as autocorrelation. The blue line shows significance and if a lag is above the blue line, it is statistically significant. The ACF plot for the car sales time series shows that there is a strong positive autocorrelation at lags

1 and 2. This means that the number of cars sold in a given month is positively correlated with the number of cars sold in the previous month and the month before that. This suggests that there is a seasonal pattern in car sales, with higher sales in certain months of the year.



PACF plot

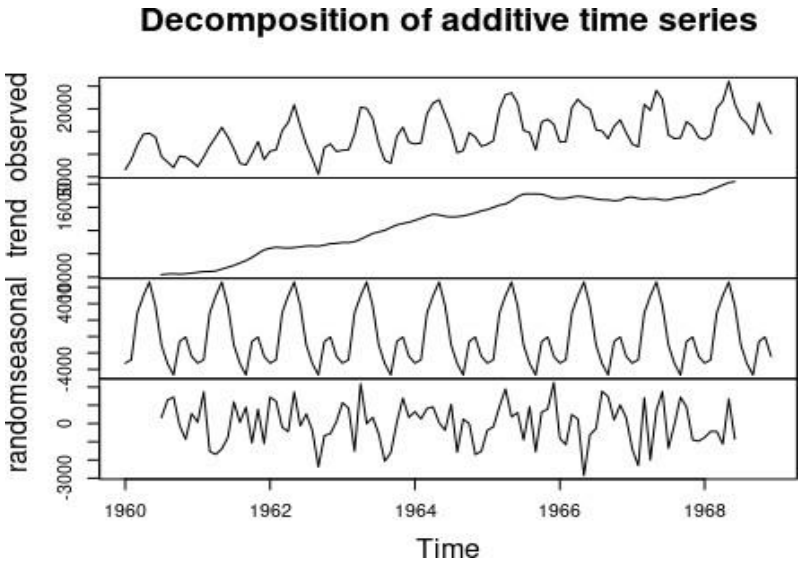
The PACF is related to the autocorrelation function (ACF), which measures the correlation between observations in a time series at different lags, without controlling for the effects of intermediate lags.



Both the Autocorrelation Function (ACF) plot and the Partial Autocorrelation Function (PACF) plot indicate the potential non-stationarity of the data. Consequently, we employ these plots to obtain insights into the parameters for modelling.

4. Decomposition

We employ the STL decomposition technique to separate the trend and seasonal components within the given data series. This approach enhances the accuracy of our forecasts. The breakdown of the STL decomposition is as follows:

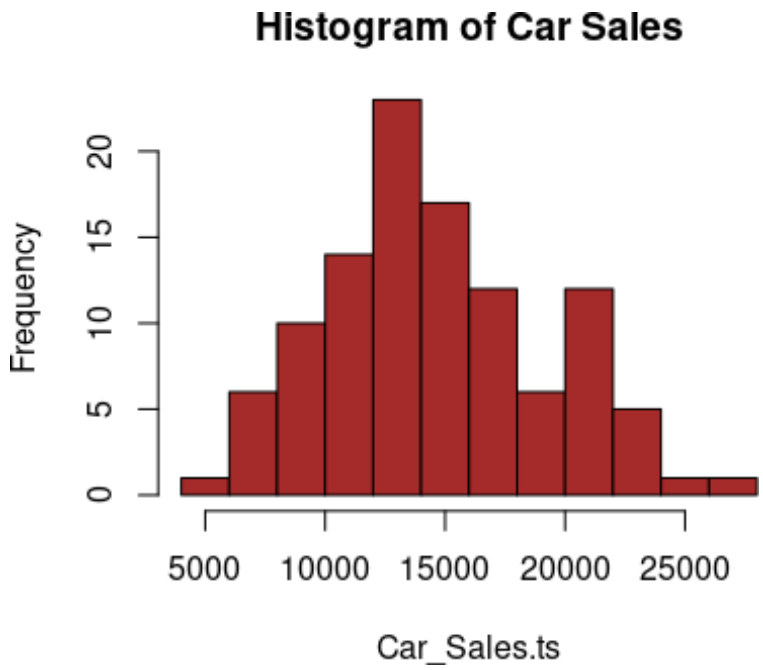


The observed component of the data reveals a combination of upward trend and cyclic behavior, characterized by a long-term increasing pattern with shorter-term cycles. This aligns with the overall ascending trend observed in the trend component. The seasonal component displays a cyclic pattern, though it is not strictly periodic and exhibits variability. Lastly, the random (remainder) component is centered around the mean, with residuals clustering around zero. This overall decomposition indicates a well-fitted model that effectively captures the different elements contributing to the data series.

5. Transformations

Transformations, such as the Box-Cox transformation, are applied to statistical data to stabilize variance, reduce skewness, and make the data more suitable for analysis.

We evaluated the data's distribution by inspecting a histogram.



The findings from the histogram suggest a close approximation to a normal distribution. As a result, considering the data's proximity to normality, there is very little need for transformation techniques such as Box-Cox. Therefore, it is advisable to skip the data transformation step.

6. Differencing and KPSS Test

The term "differencing" typically refers to a statistical or mathematical procedure that computes the variance between consecutive values in a sequence. Employed to stabilize variance and render time series data stable, differentiation aids in analysis and modelling, particularly for achieving stationary time series.

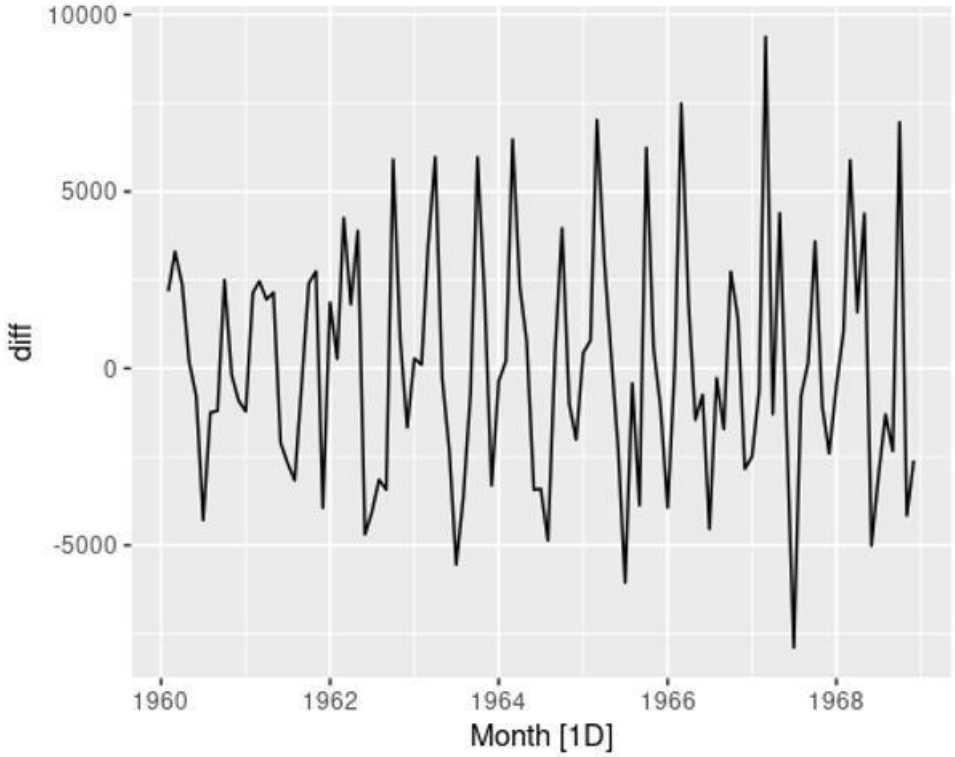
In the context of the "CarSales" time series data, the difference function is applied to the "carsales" column, creating a new column labelled "diff" to store the resulting variances. This process aims to eliminate any inherent trend or seasonality in the data.

```r
#omit na values in diff.ts
diff.ts <- na.omit(diff.ts)
head(diff.ts)
```

This command eliminates rows containing missing values (NA) in the differenced time series, a crucial step due to the introduction of NA values in the first observation during the differencing process.

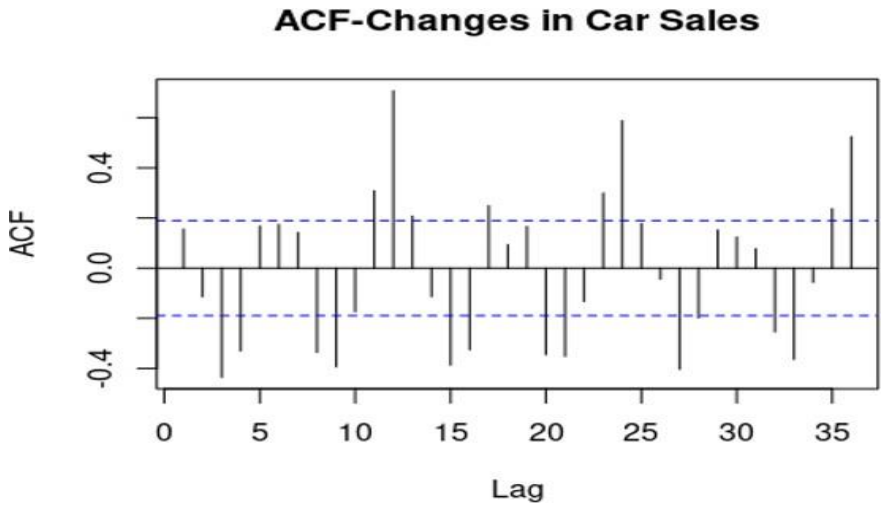
```
## # A tsibble: 6 x 3 [1D]
##   Month      CarSales  diff
##   <date>        <int> <int>
## 1 1960-02-01     8728  2178
## 2 1960-03-01    12026  3298
## 3 1960-04-01    14395  2369
## 4 1960-05-01    14587   192
## 5 1960-06-01    13791  -796
## 6 1960-07-01     9498 -4293
```

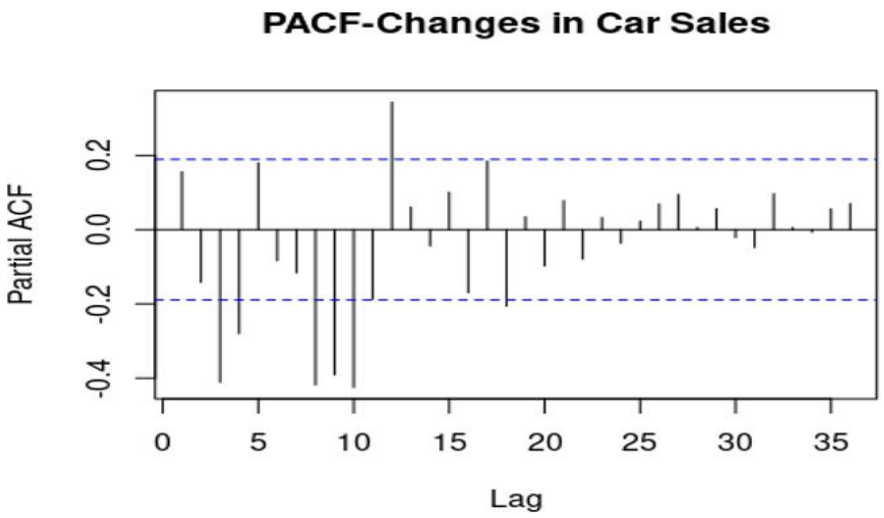
We've graphed the differenced series, using the plot to visually assess the impact of differencing on achieving stationarity.



From the plot, it appears that there is no discernible trend or seasonality, as the data is centered around the mean.

We have plotted ACF and PACF plots as well:





These plots show the autocorrelation and partial autocorrelation at different lags in the differenced series. The presence of significant lags at frequent points suggests that differencing has made the data stationary. We can derive potential values for the p and q parameters for the ARIMA model based on the plots.

We performed the KPSS test to formally evaluate stationarity for both the original and differenced series. The original series exhibited a p-value below 0.05, indicating non-stationarity.

```
#2)kpss
diff.ts %>% features(CarSales,unitroot_kpss)
```

```
## # A tibble: 1 × 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1      1.27        0.01
```

Conversely, the differenced series demonstrated a p-value exceeding 0.05, affirming stationarity post-differencing.

```
diff.ts %>% features(diff,unitroot_kpss)

## # A tibble: 1 × 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1    0.0410         0.1
```

Combining the outcomes of visual inspection, autocorrelation analysis, and the KPSS test, it is established that the first-order differenced Car Sales data series is stationary. This stationary condition renders the series conducive to subsequent time series modelling.

7. Model & Forecasting

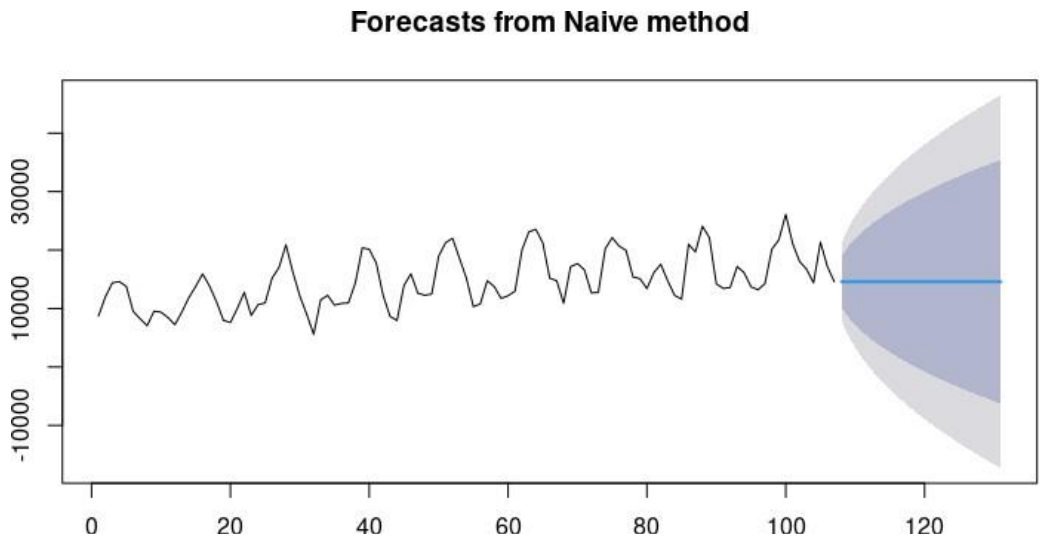
We plan to employ fundamental techniques like naïve, seasonal naïve and drift methods to predict the next 24 months. Subsequently, we will explore ETS, ARIMA, and SARIMA models. The forecasted data, along with the actual data, will be graphically presented for each model.

7.1. Naïve Model

The Naive method, a straightforward time series forecasting approach, employs the most recent observed value as the forecast for the future. The Mean Absolute Error (MAE) for this method is calculated as 2633.896.

The prediction Interval for Naïve is as below:

```
Forecasts:
    Point Forecast      Lo 80    Hi 80        Lo 95    Hi 95
108          14577 10319.0844 18834.92   8065.08002 21088.92
109          14577  8555.3980 20598.60   5367.75445 23786.25
110          14577  7202.0738 21951.93   3298.02374 25855.98
111          14577  6061.1688 23092.83   1553.16004 27600.84
112          14577  5056.0113 24097.99     15.90426 29138.10
113          14577  4147.2794 25006.72  -1373.88120 30527.88
114          14577  3311.6142 25842.39  -2651.92082 31805.92
115          14577  2533.7960 26620.20  -3841.49111 32995.49
116          14577  1803.2532 27350.75  -4958.75994 34112.76
117          14577  1112.2886 28041.71  -6015.49908 35169.50
118          14577   455.0915 28698.91  -7020.59524 36174.60
```



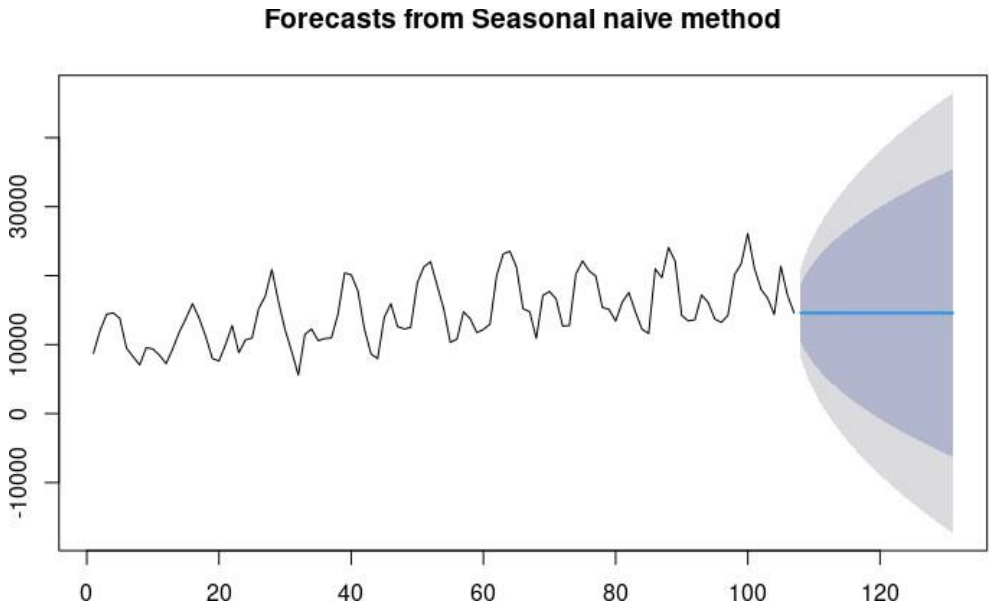
On the other hand, a straight-line forecast predicts a consistent change for the next 24 months without accounting for any underlying trends or seasonality in the data.

7.2. Seasonal Naïve Model

The Seasonal Naive model is a time series forecasting approach that relies on the last observed value from the same season in the previous year as the forecast for the corresponding future season. The Mean Absolute Error (MAE) for this method is calculated as 2637.02.

The prediction Interval for Seasonal Naïve is as below:

```
Forecasts:
    Point Forecast      Lo 80    Hi 80        Lo 95    Hi 95
108          14577 10319.0844 18834.92   8065.08002 21088.92
109          14577  8555.3980 20598.60   5367.75445 23786.25
110          14577  7202.0738 21951.93   3298.02374 25855.98
111          14577  6061.1688 23092.83   1553.16004 27600.84
112          14577  5056.0113 24097.99     15.90426 29138.10
113          14577  4147.2794 25006.72  -1373.88120 30527.88
114          14577  3311.6142 25842.39  -2651.92082 31805.92
115          14577  2533.7960 26620.20  -3841.49111 32995.49
116          14577  1803.2532 27350.75  -4958.75994 34112.76
117          14577  1112.2886 28041.71  -6015.49908 35169.50
118          14577   455.0915 28698.91  -7020.59524 36174.60
```



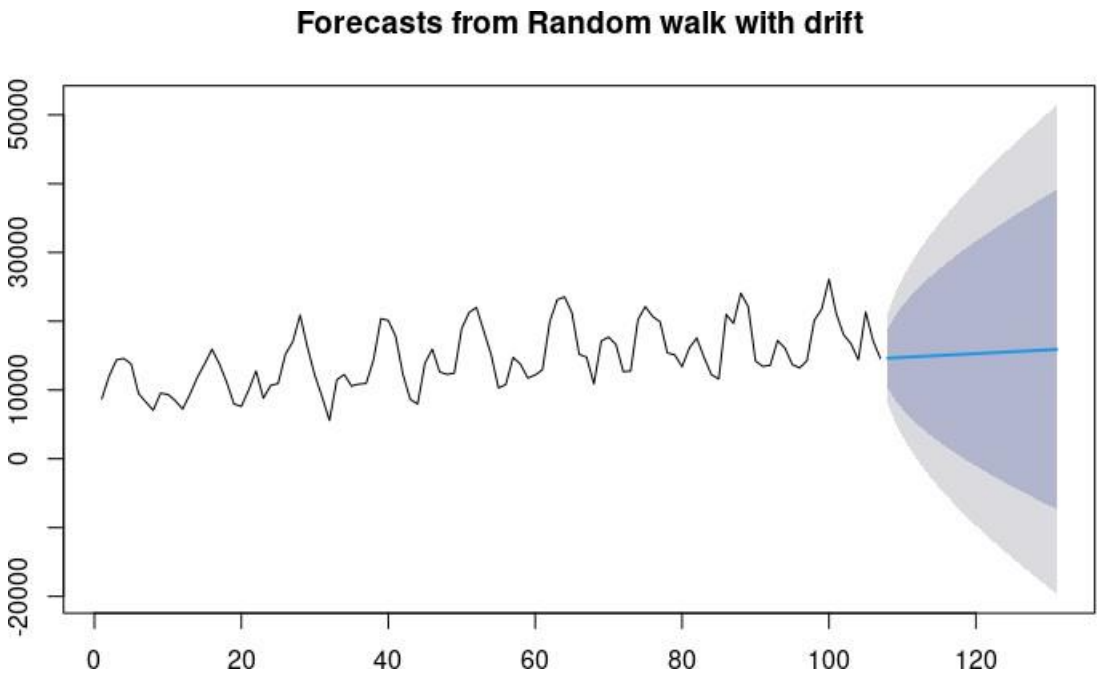
The forecast presents a straight line, signifying that the model predicts a consistent change throughout the observed period.

7.3. Drift Model

In this particular model, each forecast for a future period is determined by adding a constant drift term to the last observed value. The Mean Absolute Error (MAE) for this approach is calculated at 2637.02.

The prediction Interval for Drift is as below:

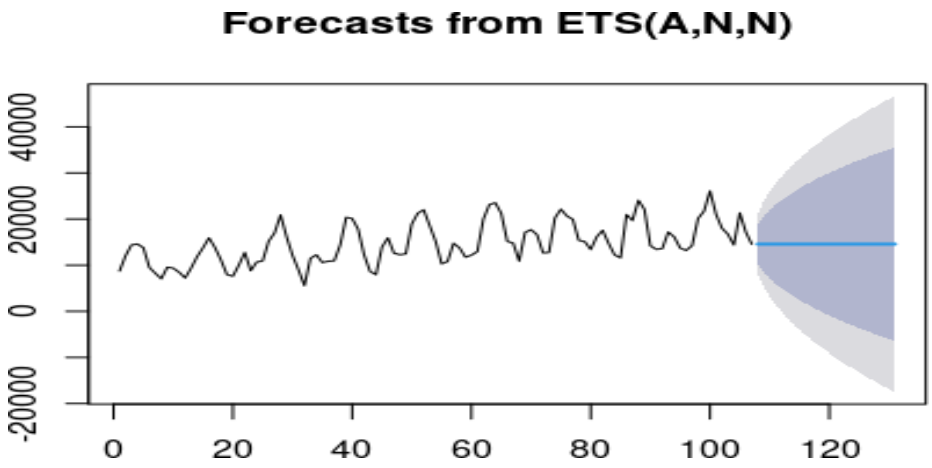
```
Forecasts:
    Point Forecast      Lo 80    Hi 80       Lo 95    Hi 95
108       14632.18 10334.4962 18929.86   8059.4402 21204.92
109       14687.36  8581.1817 20793.54   5348.7670 24025.95
110       14742.54  7229.4861 22255.59   3252.3168 26232.76
111       14797.72  6082.6880 23512.75   1469.2303 28126.20
112       14852.90  5065.0083 24640.78   -116.3866 29822.18
113       14908.08  4137.7920 25678.36  -1563.6517 31379.80
114       14963.25  3278.2003 26648.31  -2907.4939 32834.00
115       15018.43  2471.4334 27565.43  -4170.5475 34207.42
116       15073.61  1707.2680 28439.96  -5368.4477 35515.67
117       15128.79   978.3020 29279.28  -6512.5151 36770.10
118       15183.97   278.9789 30088.96  -7611.2477 37979.19
```



Notably, the model generates a forecast represented by a straight line, suggesting its anticipation of a consistent change over the observed time span.

7.4. ETS Model

The ETS model is configured as ETS(A, N, N), denoting additive error with no trend or seasonality. The model's AIC and BIC values are recorded at 2240.747 and 2248.765, respectively, while the Mean Absolute Error (MAE) is calculated as 2632.194.



Significantly, opting for a high smoothing parameter (alpha) in the ETS(A, N, N) model indicates a tendency to produce a forecast with relative constancy, underscoring the potential stability in the predicted results.

7.5. Auto ARIMA Model

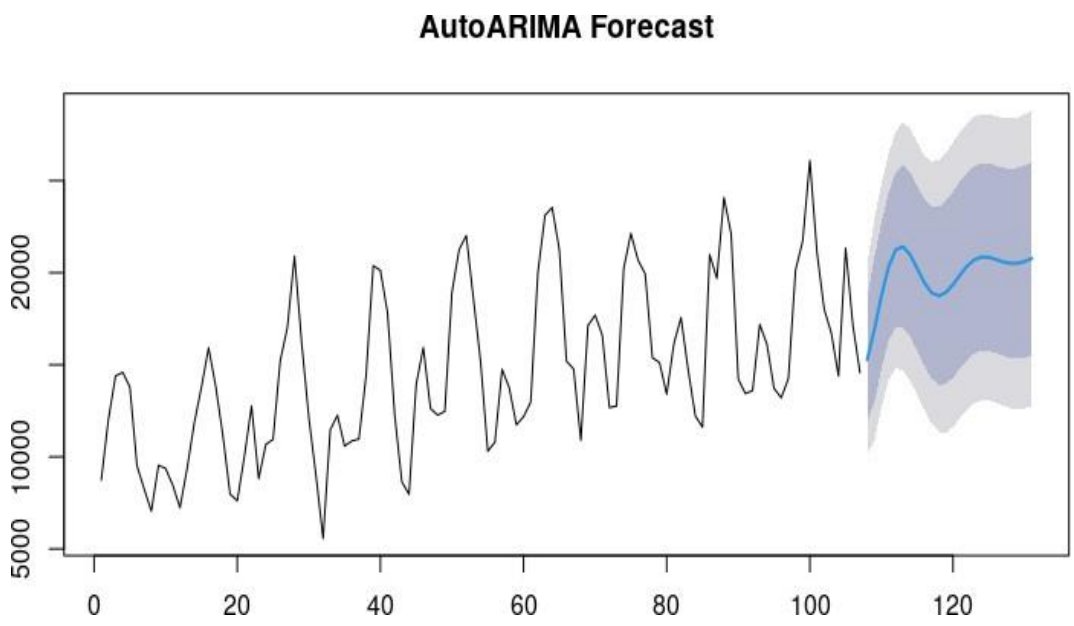
The auto.arima function is used to automatically select an ARIMA model based on the provided time series data.

Autoregression (AR) looks at how a variable changes by comparing it to its own past values.

Integrated (I) involves making raw observations more stable by calculating the differences between consecutive data points.

Moving Average (MA) considers the connection between an observation and the error left over from a moving average model applied to earlier observations.

In our scenario, the function automatically chose the ARIMA model with parameters (p, d, q) specifically denoted as ARIMA(2,1,2).



With AIC at 1980.36 and BIC at 1996.34, coupled with an MAE of 2000.547, the observed pattern of a subtle rise and fall appears to be driven by the interplay of Autoregressive (AR), Moving Average (MA), and drift terms in the model.

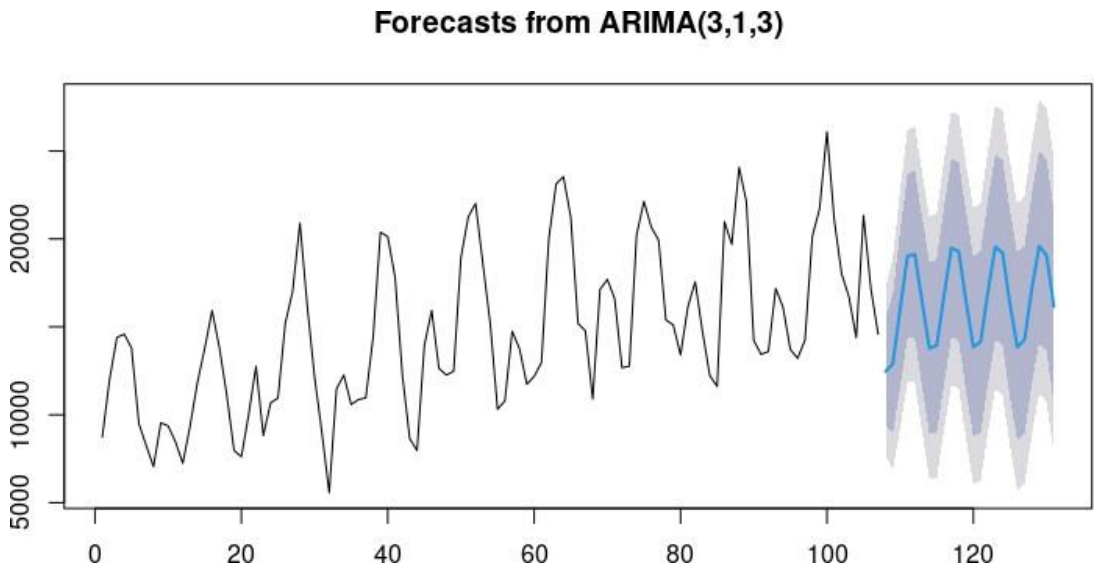
7.6. ARIMA Model

7.6.1. ARIMA(3,1,3) Model

In selecting the appropriate parameters for the ARIMA model, we considered, Partial Autocorrelation Function (PACF) plot, determining an autoregressive order (p) of 3 based on the most significant lags observed.

For the differencing order (d), first-order differencing was performed, resulting in a value of 1. The choice for the moving average order (q) was informed by the Autocorrelation Function (ACF) plot, with 3 being selected based on the most significant lags identified. Consequently, our finalized ARIMA(p,d,q) model is denoted as ARIMA(3,1,3).

The associated information criteria values, AIC at 1965.24 and BIC at 1983.89, along with a Mean Absolute Error (MAE) of 1782.656, provide insights into the model's performance.



Additionally, the plotted data visually exhibits a consistent rise and fall cycle, suggesting the presence of a discernible seasonal pattern within the time series.

7.6.2. ARIMA(4,1,4) Model

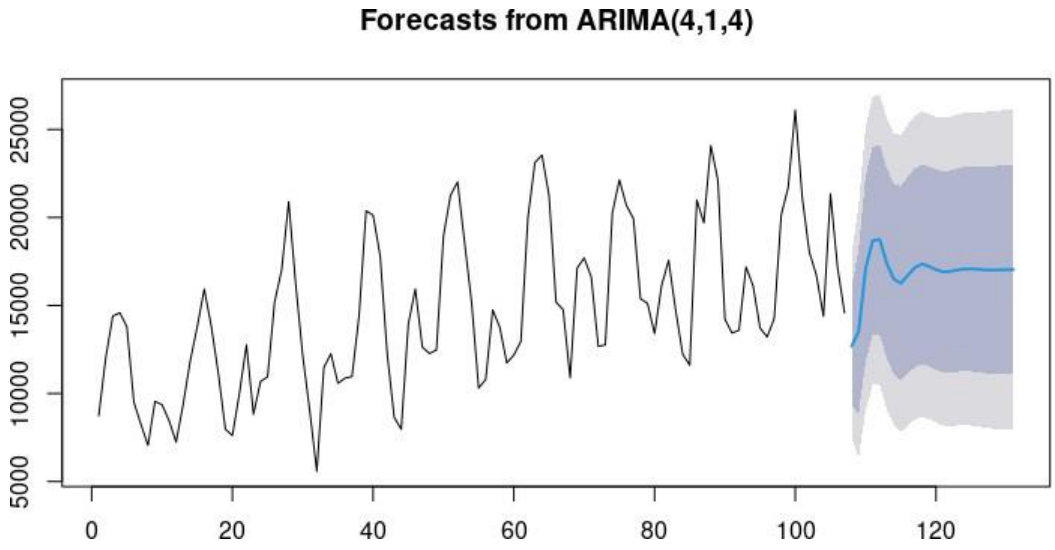
Similarly, we determined parameters for another ARIMA model as follows:

An autoregressive order (p) of 4, chosen based on the second most significant lags observed in the Partial Autocorrelation Function (PACF) plot

A differencing order (d) of 1, indicating first-order differencing

A moving average order (q) of 4, selected based on the second most significant lags identified in the Autocorrelation Function (ACF) plot.

The resulting ARIMA model is represented as ARIMA(4,1,4). The associated information criteria values include AIC at 1983.22 and BIC at 2007.19, with a Mean Absolute Error (MAE) of 1994.245 providing insights into the model's accuracy.



The forecast plot exhibits a pattern of rise, fall, and stabilization, indicating that the model successfully captures the underlying patterns inherent in the data.

7.7. SARIMA Model

SARIMA, or Seasonal Autoregressive Integrated Moving Average, is a time series forecasting model that extends ARIMA to account for seasonality, involving additional parameters denoting seasonal patterns.

In the pursuit of model selection based on AIC values for various SARIMA combinations, the objective is to choose the model with the lowest AIC.

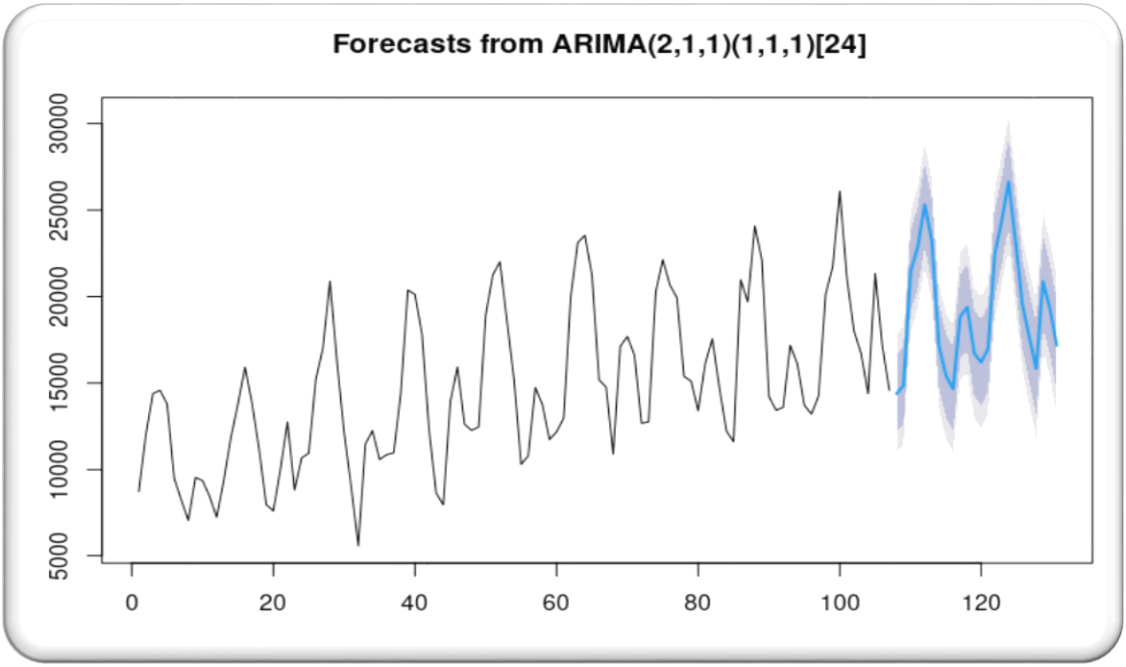
	params	AIC
1	(2,1,1)(1,1,1)	1473.574
2	(3,1,1)(1,1,1)	1475.545
3	(1,1,2)(1,1,1)	1474.287
4	(2,1,2)(1,1,1)	1475.562
5	(3,1,2)(1,1,1)	1477.489
6	(1,1,3)(1,1,1)	1475.740
7	(2,1,3)(1,1,1)	1477.615
8	(3,1,3)(1,1,1)	1474.698
9	(3,1,2)(2,1,1)	1487.548
10	(1,1,3)(2,1,1)	1474.102
11	(1,1,2)(3,1,1)	1487.308
12	(2,1,1)(1,1,2)	1472.269
13	(3,1,1)(1,1,2)	1474.257
14	(1,1,2)(1,1,2)	1472.669
15	(2,1,2)(1,1,2)	1474.272
16	(3,1,2)(1,1,2)	1476.220
17	(1,1,3)(1,1,2)	1474.352
18	(2,1,3)(1,1,2)	1476.249

From the available combinations, the SARIMA model characterized by parameters (2,1,1)(1,1,1) is favored as it exhibits the lowest AIC, presenting a balanced trade-off between model accuracy and complexity.

In this instance, an autoregressive order of 2, a differencing order of 1, and a moving average order of 1 are applied to the non-seasonal part.

The seasonal autoregressive order of 1, a seasonal differencing order of 1, and a seasonal moving average order of 1 are implemented for the seasonal aspect.

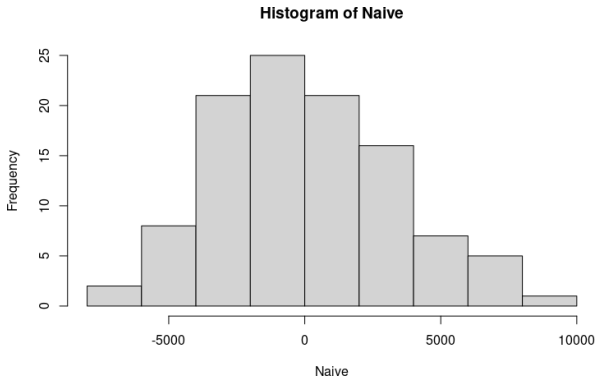
The associated metrics include AIC at 1473.57, BIC at 1488.01, and a Mean Absolute Error (MAE) of 926.418.

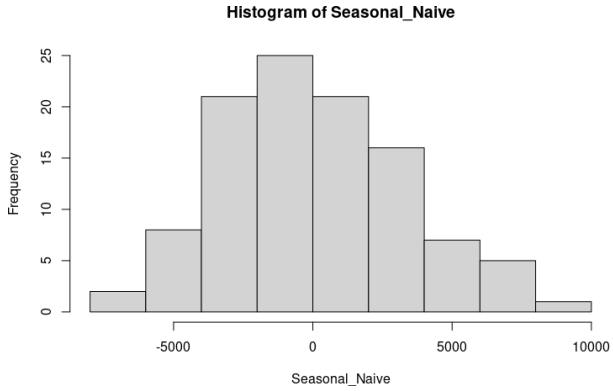


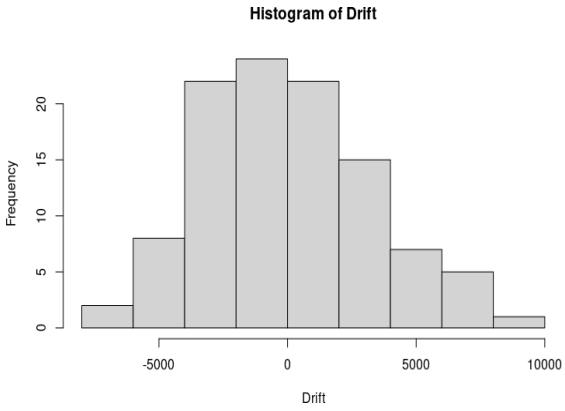
The forecast plot generated by this model suggests a continuation of the existing data pattern, featuring a consistent rise and fall with an observable linear trend.

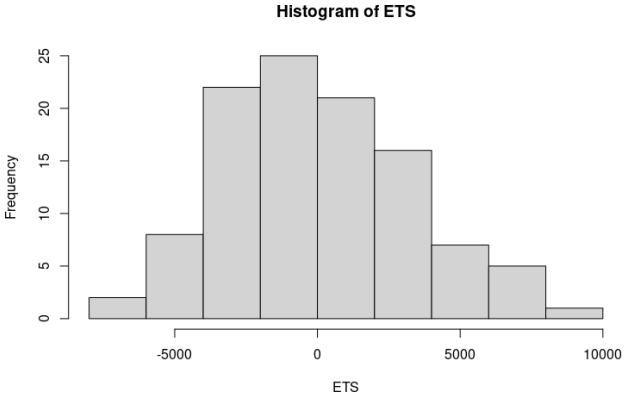
8. Residual Analysis

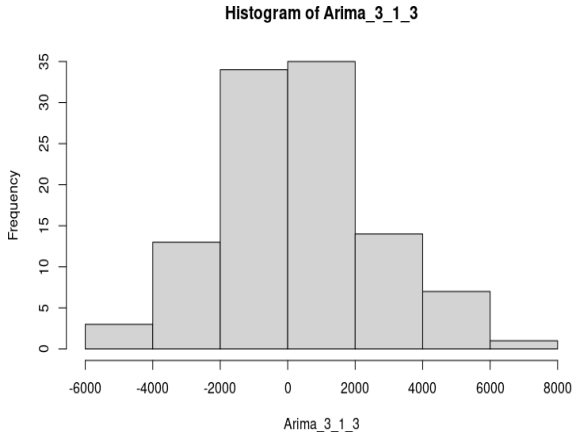
Residuals are the differences between what we predicted and what actually happened. When these differences are very small, it means our prediction is good and matches the real data closely. In a normal distribution, small differences suggest that our mistakes in prediction are spread evenly around zero, showing that our model is working well and is accurate.

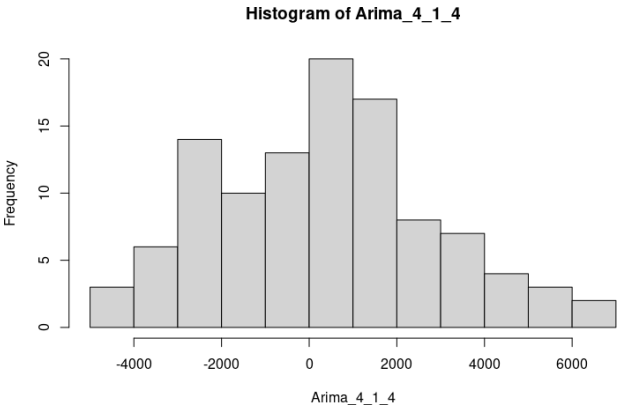

Histogram of Naive
Frequency
0
5
10
15
20
25
-5000
0
5000
10000
Naive

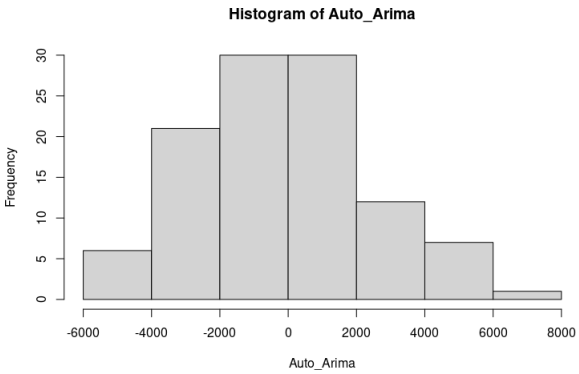

Histogram of Seasonal_Naive
Frequency
0
5
10
15
20
25
-5000
0
5000
10000
Seasonal_Naive

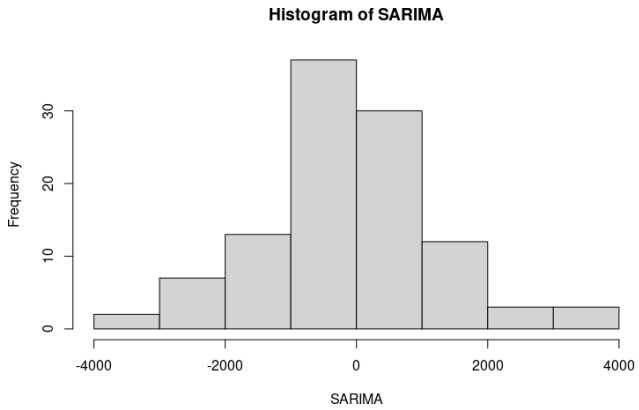

Histogram of Drift
Frequency
0
5
10
15
20
-5000
0
5000
10000
Drift


Histogram of ETS
Frequency
0
5
10
15
20
25
-5000
0
5000
10000
ETS


Histogram of Arima_3_1_3
Frequency
0
5
10
15
20
25
30
35
-6000
-4000
-2000
0
2000
4000
6000
8000
Arima_3_1_3


Histogram of Arima_4_1_4
Frequency
0
5
10
15
20
-4000
-2000
0
2000
4000
6000
Arima_4_1_4





Looking at all the graphs, we see that the residuals in all the models are somewhat spread out like a normal distribution. However, the graph of ARIMA(3,1,3) and SARIMA models are perfectly shaped like a normal distribution when compared to the other models.

9. Model Selection

We've got eight models available for predicting the given data. We compute the metrics for each of these models, and the summarized results are presented in the table below.

Sl.No	MODEL	ME	RMSE	MAE	AIC & BIC
1	NAÏVE	55.17925	3322.469	2633.896	
2	SNAIVE	55.17925	3322.469	2633.896	
3	DRIFT	-6.863629e-13	3322.011	2637.02	
4	ETS	31.83847	3315.377	2632.194	

5	AUTOARIMA	-39.66704	2538.026	2000.547	AIC=1980.36 BIC=1996.34
6	ARIMA(3,1,3)	317.6336	2304.763	1782.656	AIC=1965.24 BIC=1983.89
7	ARIMA(4,1,4)	382.1822	2513.158	1994.245	AIC=1983.22 BIC=2007.19
8	SARIMA	-113.6167	1320.009	926.418	AIC=1473.57 BIC=1488.01

Looking at the table above, it's evident that the SARIMA forecasting method outperforms the others with the lowest values for RMSE, MAE, ME, AIC and BIC. Considering these metrics as criteria, we would choose the SARIMA forecasting method as the best model from the options available and apply it for forecasting in the upcoming months/years.

10. Conclusion

We have ultimately predicted car sales for the upcoming ten years using the model that emerged as the most effective among the various options considered. Organizations can utilize information derived from the SARIMA model to comprehend and foresee cyclic patterns in car sales. The forecasting capability assists in optimizing production schedules, managing inventory

efficiently, and planning workforce requirements for both car manufacturers and dealerships. The model's proficiency in capturing seasonal patterns offers valuable insights for adapting to changes in consumer demand throughout the year.

11. References

Dataset - https://www.kaggle.com/datasets/dinirimameev/monthly-car-sales-in-quebec-1960/data

ARIMA and SARIMA - https://www.visual-design.net/post/time-series-analysis-arma-arima-sarima

Timeseries Forecasting - https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b