Introduction

Heavy Haulers Inc., a leading manufacturer of motor trucks, has been analyzing its sales data to improve its forecasting capabilities for the next few years. As a Business Analyst of the company, I have been entrusted with the task of developing a predictive model that will help the company forecast its sales accurately. The objective of this project is to create a sales forecasting model for the next three years, utilizing historical sales data for the past 55 years. The project will use time series forecasting methods such as ARIMA and Exponential Smoothing to build the model and evaluate its performance using the Mean Absolute Percentage Error (MAPE) and Mean Absolute Scaled Error (MASE) metrics.

The development of this sales forecasting model for Heavy Haulers Inc. will enable the company to plan its production schedules, manage its inventory efficiently, and make informed decisions regarding sales and marketing strategies.

Problem Presentation

Heavy Haulers Inc. is facing challenges in accurately forecasting its sales of motor trucks. The company's historical sales data shows significant fluctuations, making it difficult to predict future sales with certainty. Inaccurate sales forecasting can result in several negative impacts on the company, including inefficient production and inventory management, ineffective marketing strategies, and decreased customer satisfaction. Furthermore, it can lead to missed revenue opportunities and a competitive disadvantage in the market.

To address these challenges, the company needs a new approach that can leverage its existing data and resources to build a robust sales forecasting model. Developing a sales forecasting model is critical for Heavy Haulers Inc. to remain competitive in the market, increase revenue, and maintain customer satisfaction.

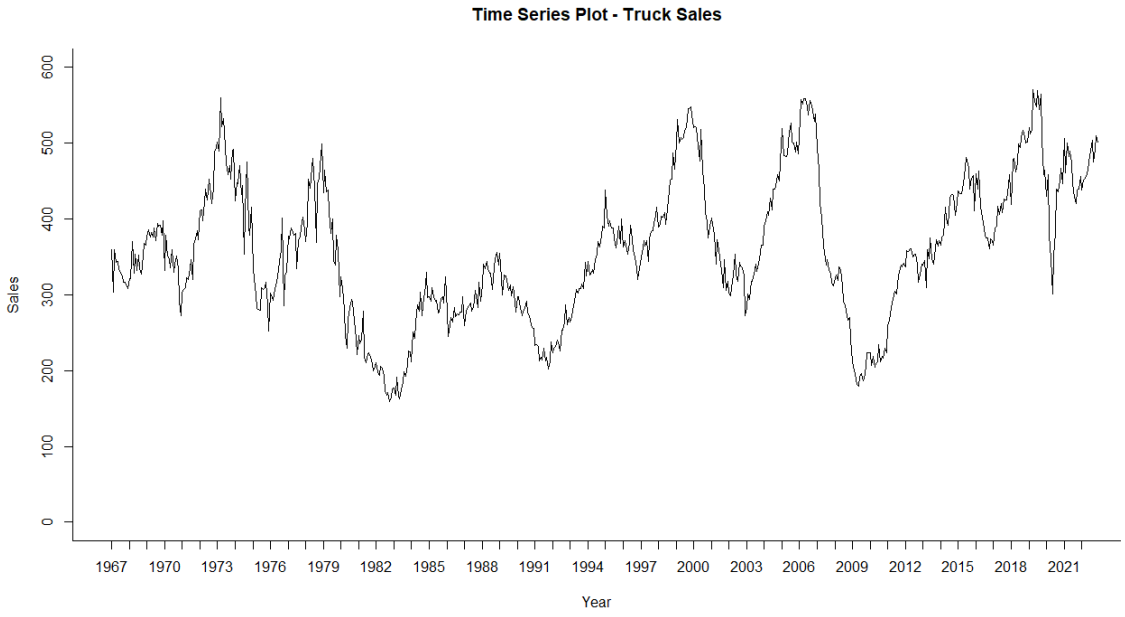
This report presents an analysis of historical truck sales data from the year 1967 to 2022. The purpose of this analysis is to identify trends and patterns in the data and use this information to develop a sales forecasting model for the next 3 years.

Data Description

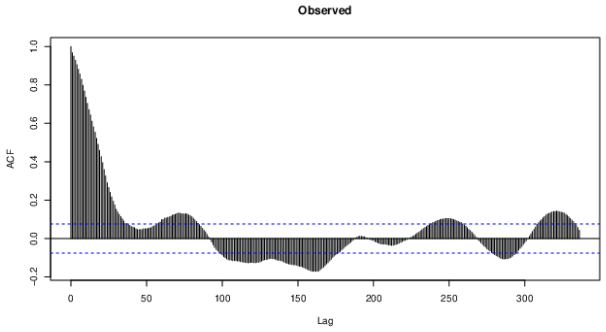
The data used for this analysis was sourced from the Federal Reserve Bank of St. Louis and contains monthly truck sales data from January 1967 to December 2022. The data consists of 672 observations and has three variables: Year, Month, and Truck Sales.

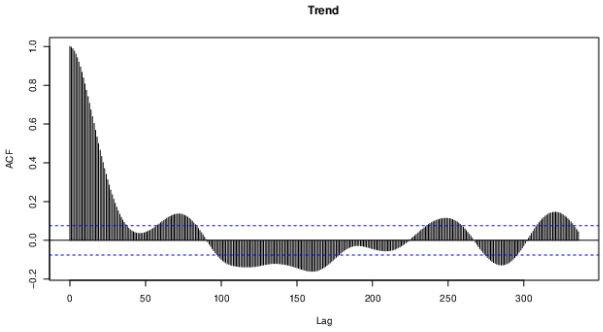
Data Analysis

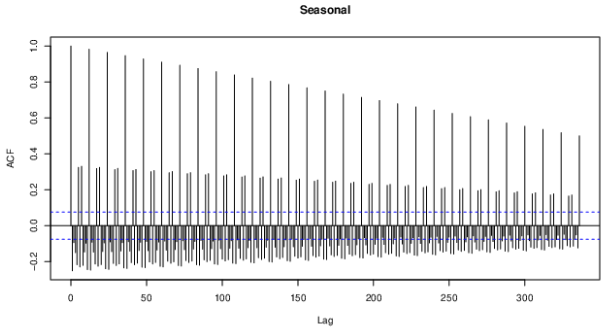
After importing the data into R, the date variable was created by combining the Year and Month variables. The data was then converted into a time series object and the time series plot was generated.

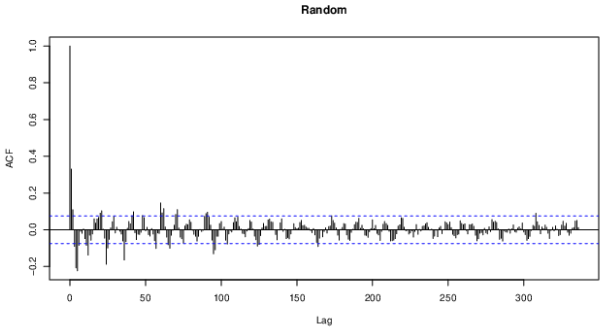


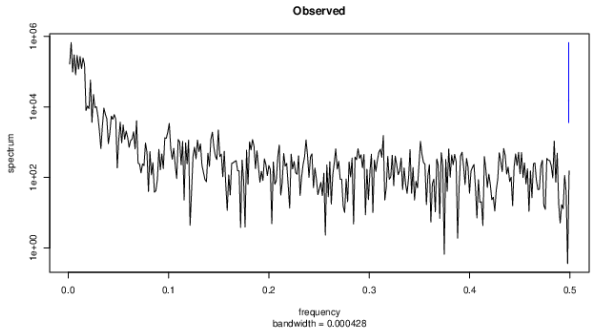
Next, the time series was decomposed into its components using the additive method.

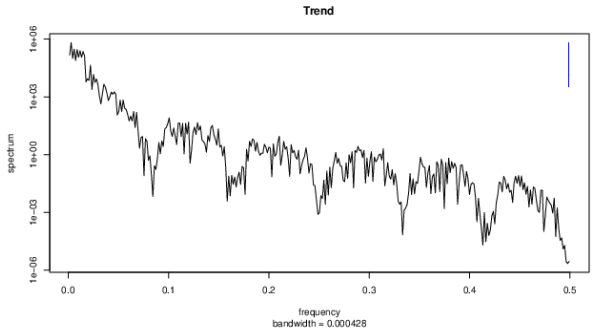

Observed
ACF
Lag

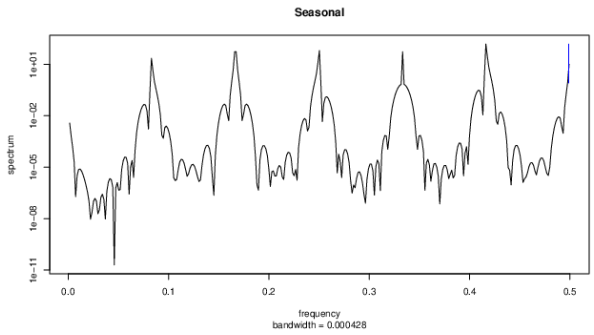

Trend
ACF
Lag

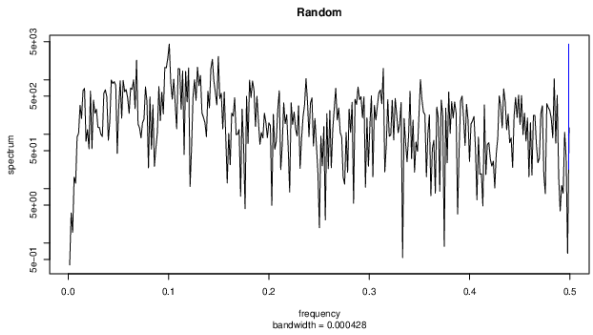

Seasonal
ACF
Lag

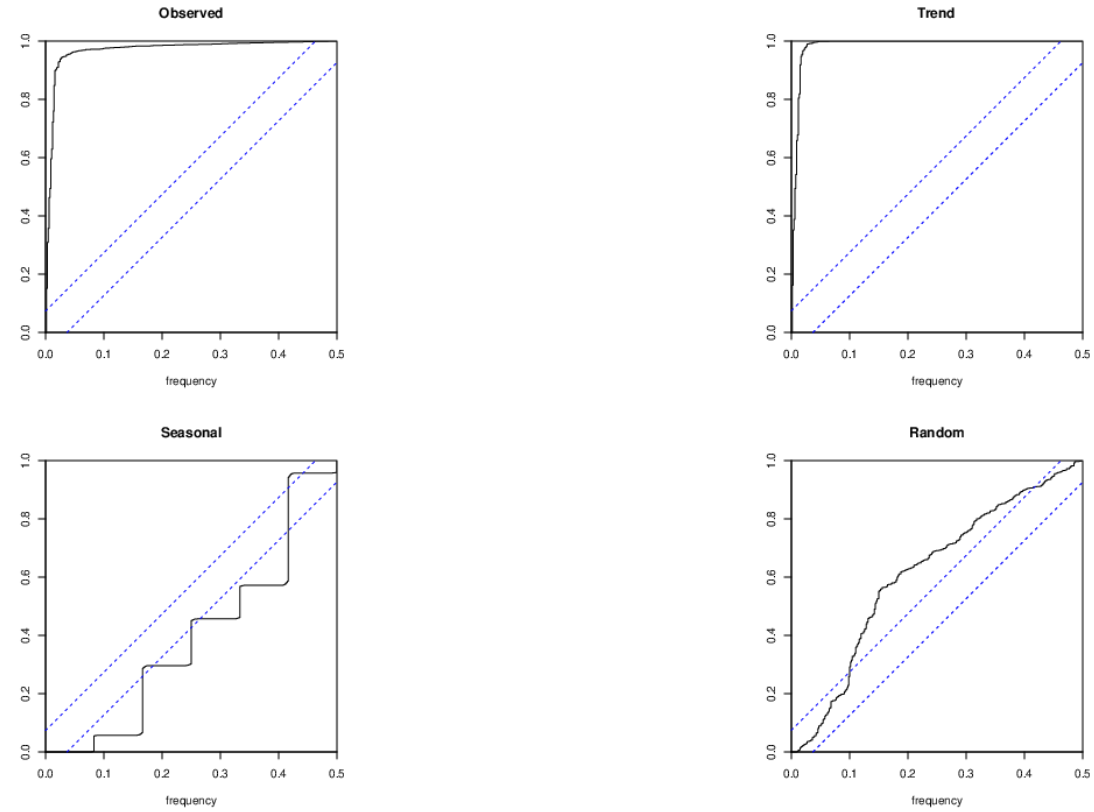

Random
ACF
Lag

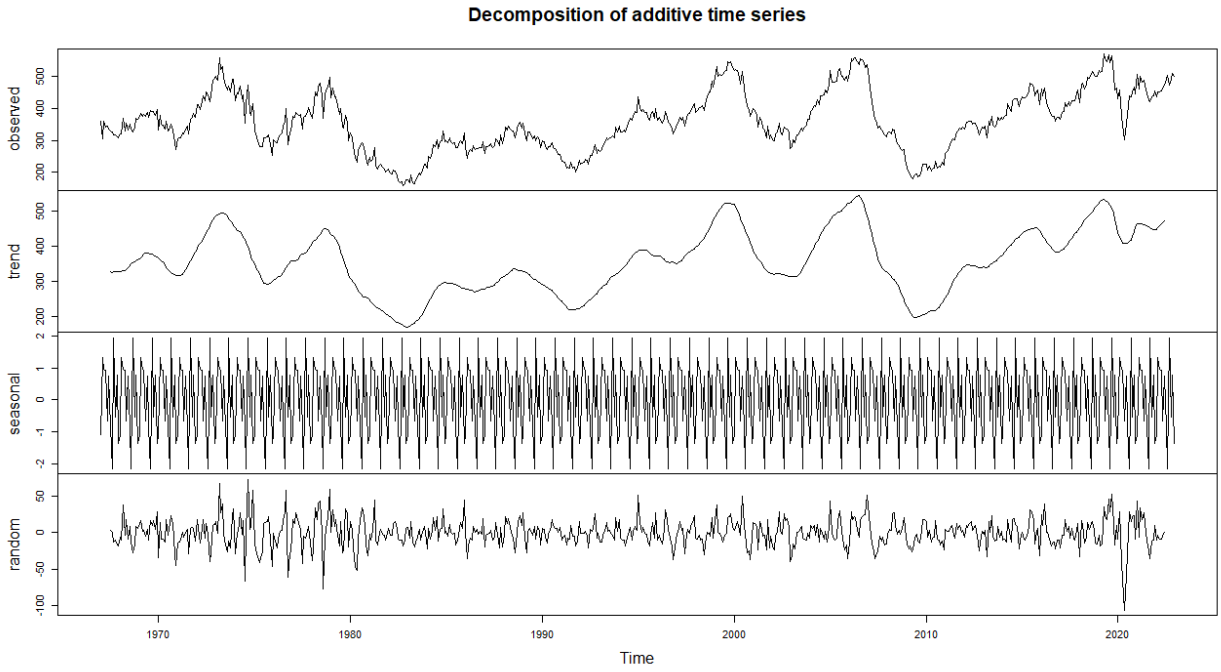

Observed
spectrum
frequency
bandwidth = 0.000428


Trend
spectrum
frequency
bandwidth = 0.000428


Seasonal
spectrum
frequency
bandwidth = 0.000428


Random
spectrum
frequency
bandwidth = 0.000428





Observations from the decomposition of Time-series data are as follows:

- The trend component reveals that the sales of trucks exhibited a range of fluctuations over the years, reaching their peak performance around 2008-09. Ironically, this is duing the period where the actual sales took a dip due to recession. This tells us that there can be a number of factors beyond the historical sales data that determines the trend of the sales. However, the analysis has not considered any other factors while predicting the future sales values.
- The random component shows irregular or unexpected fluctuations during 2020. It is reasonable to assume that these fluctuations were caused by the Covid-19 pandemic.

The Augmented Dickey-Fuller (ADF) test is conducted to check the stationarity of the time series data.

```
> adf.test(ts_data)

        Augmented Dickey-Fuller Test

data:  ts_data
Dickey-Fuller = -3.5247, Lag order = 8, p-value = 0.03993
alternative hypothesis: stationary
```

The results of the ADF test indicate a p-value of 0.01, which is less than the significance level of 0.05, indicating that we can reject the null hypothesis of non-stationarity. In other words, the data is stationary.

This suggests that we can proceed with time series modeling techniques that assume stationary data, such as ARIMA and Exponential Smoothing.

Data Partitioning

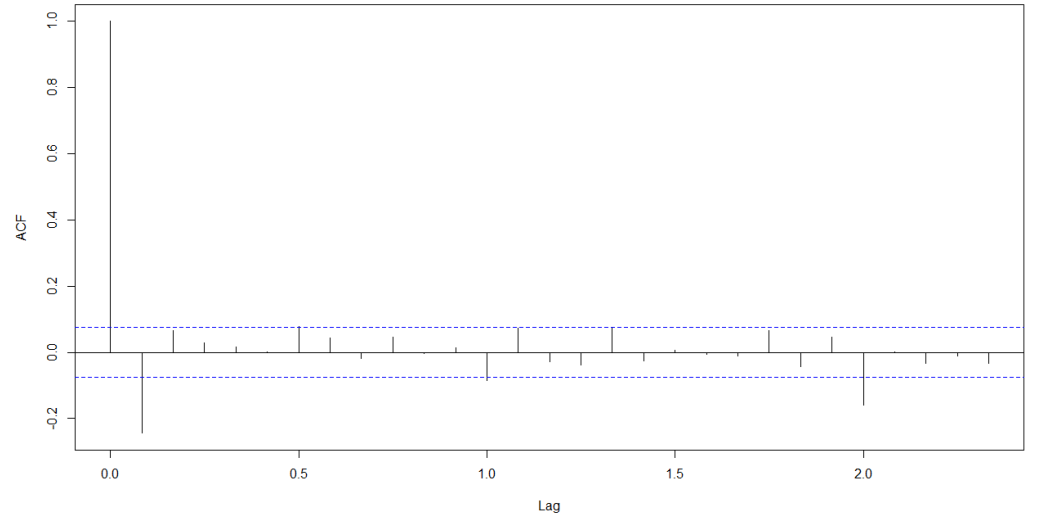
To build and validate a forecasting model, partitioning the time series data into training and validation sets is an important step. In this regard, the time series data is partitioned into two sets - a training set and a validation set.

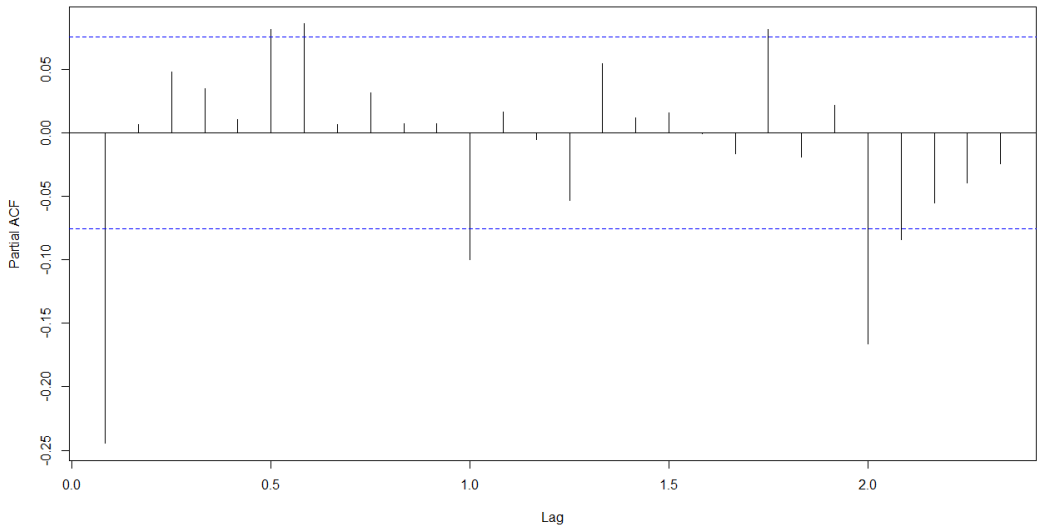
The length of the validation set is defined to be 12 months (to mean Jan 2022 to Dec 2022). Then, the training set is defined to include all observations from the start of the time series in January 1967 to the end of December 2021. Summary statistics of the training and validation sets is as follows:

```
nValid = 12
nTrain = length(Trucks.ts) - nValid
train_Trucks.ts = window(Trucks.ts, start = c(1967, 1), end = c(2021, 12))
valid_Trucks.ts = window(Trucks.ts, start = c(2022, 1), end = c(2022, 12))
summary(train_Trucks.ts)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
159.0   295.8   348.5   356.7   420.2   571.0
summary(valid_Trucks.ts)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
438.0   455.0   475.0   475.8   494.2   510.0
```

We can notice that the average sales during the validation period is around 476 trucks.

Before building a model, autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were generated to investigate the presence of correlation between the different lags in the time series data.





The ACF plot indicates that there is a significant correlation at lags 0, 1, and 2, as these lags have crossed the significance level. This finding suggests that the time series may have an autoregressive component.

The PACF plot revealed that lags at 1 and 2 have surpassed the significance level, implying a strong correlation between the current observation and those at lags 1 and 2. This indicates the possibility of the existence of a moving average component in the time series.

Overall, the presence of both autoregressive and moving average components in the time series suggests that an ARMA model might be appropriate for modeling the data.

The parameters of the ARIMA model are selected based on the below observations:

- Autoregressive (AR) component: 2 (as lags 1 and 2 in ACF plot have crossed significance level)
- Integrated (I) component: 0 (as the stationarity test indicated that the data is already stationary)
- Moving Average (MA) component: 1 or 2 (as lags 1 and 2 in PACF plot have crossed significance level)

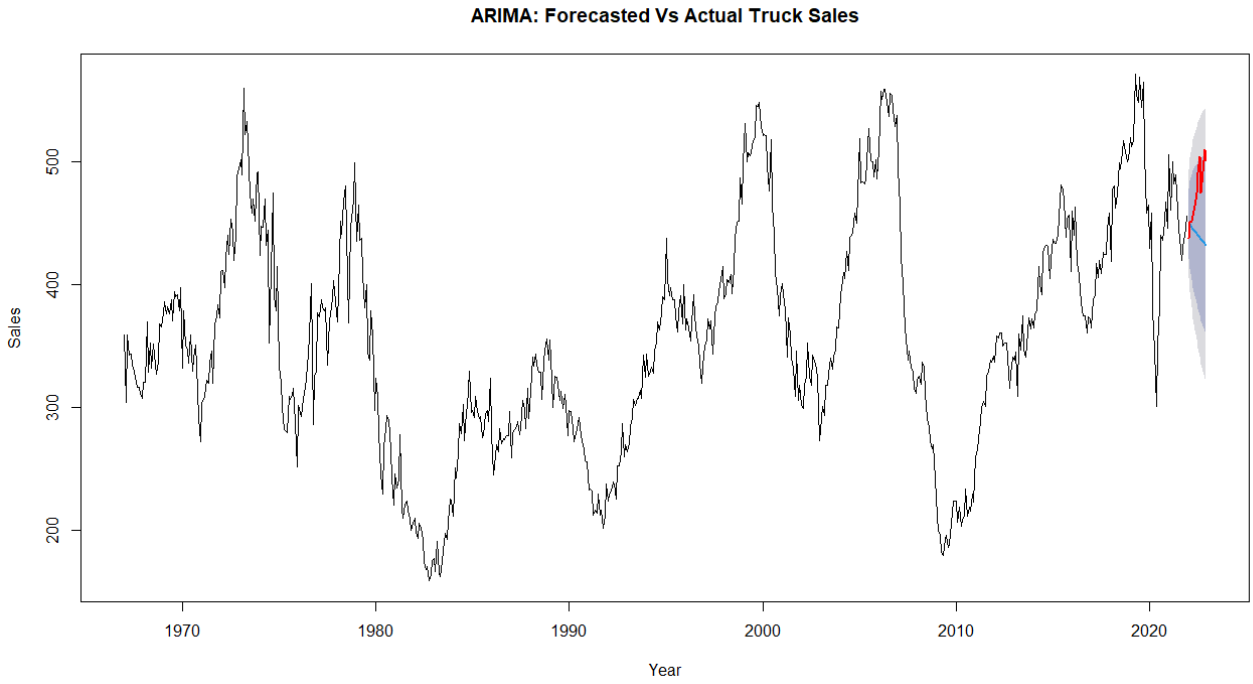
Therefore, a possible ARIMA model would be ARIMA(2,0,1) or ARIMA(2,0,2).

ARIMA (2,0,1)

An ARIMA model of order (2, 0, 1) based on the training data of the time series is created to forecast future values of the time series data. A summary of the model is given below:

```
> summary(model_AR)
Series: train_Trucks.ts
ARIMA(2,0,1) with non-zero mean

Coefficients:
         ar1     ar2     ma1      mean
      0.6859  0.2874  0.0581  360.0979
s.e.  0.1348  0.1308  0.1387   32.1331

sigma^2 = 497.6:  log likelihood = -2985.16
AIC=5980.32   AICc=5980.41   BIC=6002.78

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE        ACF1
Training set 0.08559392 22.23929 16.66876 -0.4081234 4.831096 0.2810658 0.00380381
```

Forecasting for the validation set and future time periods is made using the ARIMA (2,0,1) model. The forecasted values are plotted against the actual truck sales data in a line chart. The actual sales data is represented by the red line, while the blue line shows the forecasted values.



The accuracy of the model has been verified and the results are given below:

```
> print(accuracy_AR)
                    ME      RMSE      MAE         MPE      MAPE      MASE       ACF1 Theil's U
Training set  0.08559392 22.23929 16.66876 -0.4081234 4.831096 0.2810658 0.00380381        NA
Test set     34.74217112 44.70557 36.68542  7.0351873 7.478852 0.6185833 0.69139465  3.171143
```

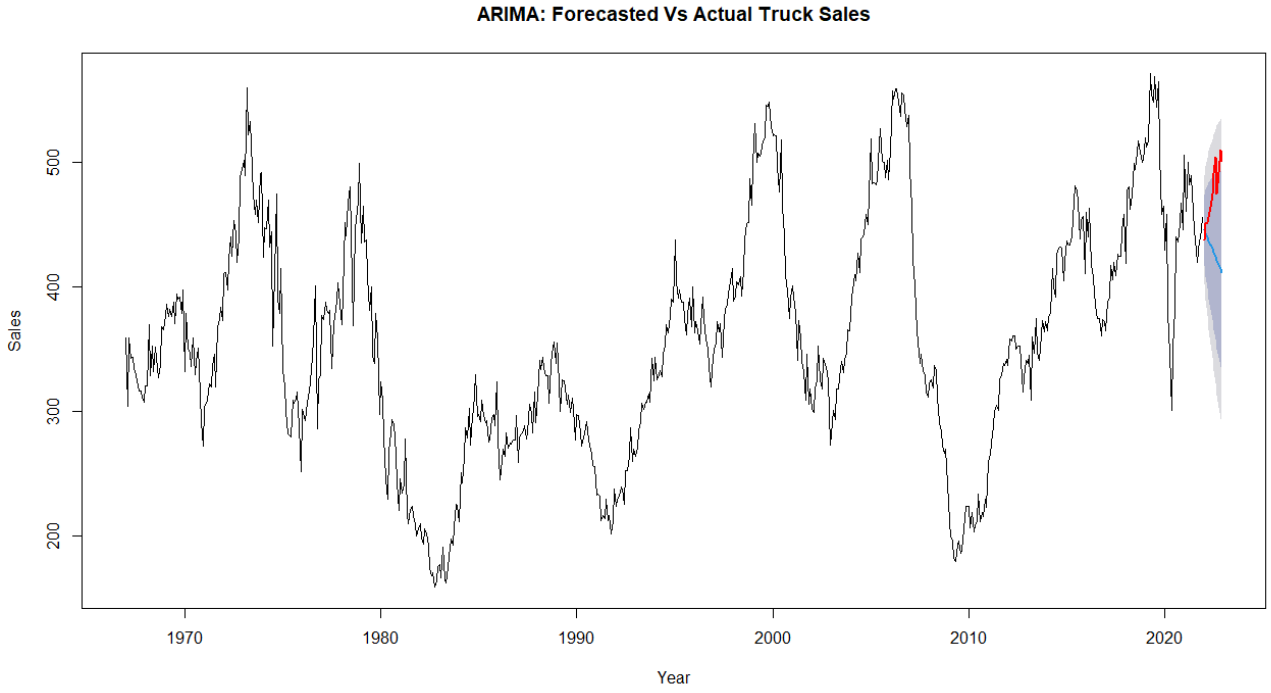
The MAPE is 4.83 for the training set and 7.47 for the test set. The MASE is 0.28 for the training set, which indicates that the model has a better fit than a naïve model. The MASE for the test set is 0.62, which indicates that the model is less accurate when applied to new data. So, an ARIMA model of the order (2,0,2) is built.

ARIMA (2,0,2)

The ARIMA (2,0,2) model is utilized for predicting truck sales for both the validation set and future time periods. Summary of the model is as follows:

```
> summary(model_AR)
Series: train_Trucks.ts
ARIMA(2,0,2) with non-zero mean

Coefficients:
         ar1      ar2      ma1     ma2      mean
      1.912  -0.9164  -1.1842  0.2865  358.9873
s.e.  0.038   0.0371   0.0518  0.0382   19.7257

sigma^2 = 487.8:  log likelihood = -2978.2
AIC=5968.39   AICc=5968.52   BIC=5995.34

Training set error measures:
                     ME     RMSE      MAE        MPE     MAPE      MASE         ACF1
Training set -0.02732152 22.00312 16.49474 -0.4362488 4.773068 0.2781314 -0.001378928
```

The below line chart compares the forecasted values with the actual sales data. The actual sales data is shown by the red line, while the blue line shows the forecasted values.



The accuracy of the ARIMA model is evaluated using MAPE and MASE.

```
> accuracy_AR <- accuracy(forecast_AR, valid_Trucks.ts)
> print(accuracy_AR)
                     ME     RMSE      MAE        MPE     MAPE      MASE         ACF1 Theil's U
Training set -0.02732152 22.00312 16.49474 -0.4362488 4.773068 0.2781314 -0.001378928        NA
Test set     46.71173022 57.24374 48.02444  9.5086039 9.808310 0.8097799  0.708928073  4.072091
```

From the above results, we can see that the ARIMA model achieved a MAPE value of 4.77% for the training set and 9.8% for the test set. The MASE values for the training and test sets are 0.28 and 0.81, respectively, suggesting that the model outperforms the naïve model for the training set, but its performance is slightly worse for the test set.

The metrics from the above models suggest that the ARIMA (2,0,1) model is more accurate than the ARIMA (2,0,2) model, as it has lower MAPE and MASE values for both the training and test sets. However, it appears that the accuracy is not satisfactory, and therefore ETS (Exponential Smoothing) method, a popular time series forecasting technique is used for further analysis.

Exponential Smoothing [ETS]

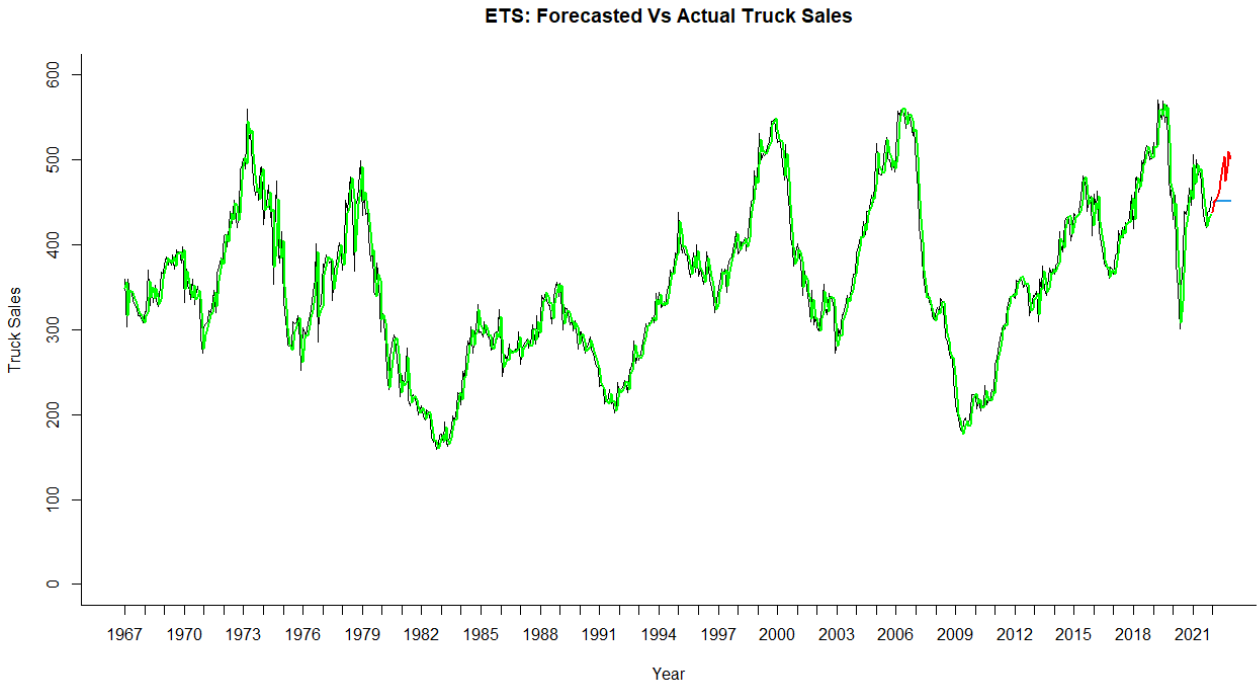
To forecast truck sales, the ETS (Error, Trend, Seasonality) model is used in this analysis. The ETS function is applied to the training data set, with the model specification of "ZZZ". This model specification allows for the automatic selection of the error, trend, and seasonality components of the time series. The resulting ETS model is, then, used to generate forecasts for future truck sales data (Including for the Validation period).

A summary of the ETS model is provided below:

```
> summary(Truck_ETS)
ETS(M,Ad,N)

Call:
 ets(y = train_Trucks.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.7042
    beta  = 0.0565
    phi   = 0.8258

  Initial states:
    l = 350.2055
    b = -1.3953

  sigma:  0.0625

     AIC     AICc      BIC
8343.425 8343.554 8370.379

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 0.1689232 22.25255 16.64834 -0.157264 4.806686 0.2807214 -0.003140927
```

From the above summary, we can see that ETS(M,Ad,N) is used to forecast future truck sales. ETS(M,Ad,N) is a variant of the ETS method that models the data with additive errors, a multiplicative trend, and no seasonality. This method is useful when there is no clear seasonal pattern in the data, but a trend is present. The ETS model is a good alternative to ARIMA models when there is no clear seasonal pattern.

Below is a plot comparing the forecasted truck sales with the actual sales data using the ETS model. The plot displays two lines: one in red represents the actual truck sales data, while the other in green depicts the fitted values obtained from the ETS model.



The forecast accuracy of the ETS model was evaluated using MAPE and MASE.

```
> accuracy(Truck_ETS.pred, valid_Trucks.ts) #
                    ME      RMSE       MAE        MPE      MAPE      MASE          ACF1 Theil's U
Training set  0.1689232 22.25255 16.64834 -0.157264 4.806686 0.2807214 -0.003140927        NA
Test set     24.4482558 33.17468 26.52301  4.923693 5.397381 0.4472264  0.644597215  2.345823
```

Based on results, the ETS model has performed well on both the training and test sets. The MAPE value for the training set is 4.81%, indicating that the model has an average error of 4.81% in its predictions. The MAPE value for the test set is 5.40%, which is also a reasonable level of accuracy.

Similarly, The MASE values were 0.28 for the training set and 0.45 for the test set, indicating that the forecast errors were small relative to the variability of the data.

Therefore, based on the accuracy metrics, we can conclude that the ETS model provides good forecasts for the Truck sales data.

AUTO ARIMA

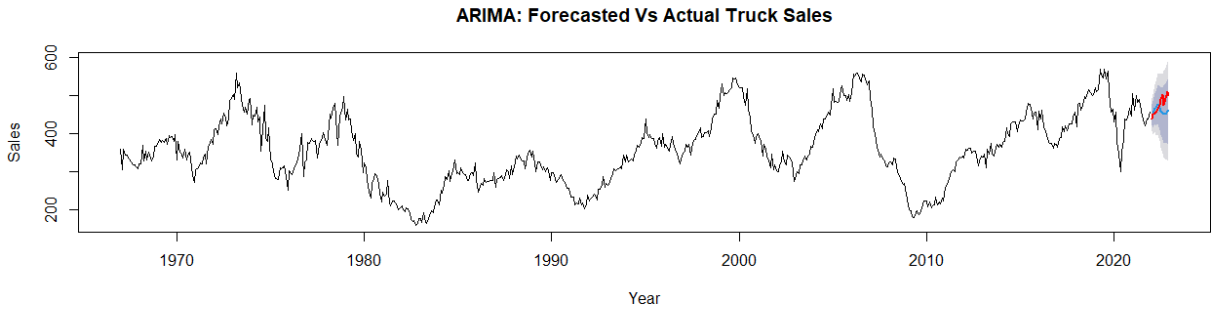
According to the peer review recommendations, an AUTO ARIMA model has also been created. The results of the model are as follows:

Summary of AUTO ARIMA

```
> summary(model_AR)
Series: train_Trucks.ts
ARIMA(1,1,2)(2,0,2)[12]

Coefficients:
         ar1      ma1     ma2    sar1     sar2     sma1    sma2
      0.8728  -1.1404  0.2887  0.7258  -0.6552  -0.8618  0.5261
s.e.  0.0694   0.0750  0.0388  0.1027   0.0829   0.1148  0.1006

sigma^2 = 452.5:  log likelihood = -2948.09
AIC=5912.17   AICc=5912.39   BIC=5948.1

Training set error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE         ACF1
Training set 0.1875922 21.14211 15.51927 -0.2127958 4.501778 0.2616833 0.006185147
```

So, the ARIMA(1,1,2)(2,0,2)[12] model has an autoregressive order of 1, a differencing order of 1, a moving average order of 2, a seasonal autoregressive order of 2, no seasonal differencing, a seasonal moving average order of 2, and a seasonal period of 12. This is the model that was fitted to the time series data by AUTO ARIMA Function.

AUTO ARIMA – Forecast Vs Actual



From the above plot, we can see the Red Line (Actual) vs Blue line (Predicted) for the validation period. It seems that the auto Arima model has predicted with plausible certainty. The accuracy of the model is also verified.

```
> print(accuracy_AR)
                    ME      RMSE       MAE         MPE      MAPE      MASE
Training set  0.1875922 21.14211 15.51927 -0.2127958 4.501778 0.2616833
Test set     15.9065390 29.67013 24.41914  3.1136263 4.994666 0.4117514
```

Based on accuracy resluts, the Auto ARIMA model has performed well on both the training and test sets. The MAPE value for the training set is 4.5%, indicating that the model has an average error of 4.5% in its predictions. The MAPE value for the test set is around 5%, which is also a reasonable level of accuracy.

Similarly, The MASE values were 0.26 for the training set and 0.41 for the test set, indicating that the forecast errors were small relative to the variability of the data.

Therefore, based on the accuracy metrics, we can conclude that the Auto ARIMA model provides good forecasts for the Truck sales data, even better results than the ETS model.

Based on the final ARIMA model, the forecasted values for the years 2023, 2024, and 2025 are as follows:

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2022	450.2570	422.9966	477.5175	408.5658	491.9483
Feb 2022	450.5042	416.7138	484.2947	398.8262	502.1822
Mar 2022	466.7020	426.6675	506.7364	405.4746	527.9294
Apr 2022	465.9531	419.8902	512.0160	395.5059	536.4003
May 2022	478.1635	426.2566	530.0705	398.7787	557.5483
Jun 2022	469.1119	411.5297	526.6941	381.0476	557.1763
Jul 2022	460.5937	397.4958	523.6916	364.0938	557.0936
Aug 2022	452.9368	384.4762	521.3975	348.2353	557.6383
Sep 2022	451.7301	378.0543	525.4059	339.0527	564.4075
Oct 2022	452.8947	374.1463	531.6432	332.4595	573.3300
Nov 2022	458.9731	375.2897	542.6566	330.9904	586.9559
Dec 2022	460.3012	371.8154	548.7871	324.9739	595.6286
Jan 2023	448.6788	356.6108	540.7468	307.8730	589.4847
Feb 2023	461.4848	365.6065	557.3631	314.8516	608.1180
Mar 2023	462.0991	362.5321	561.6660	309.8246	614.3735
Apr 2023	469.5686	366.4246	572.7126	311.8235	627.3137
May 2023	473.0667	366.4484	579.6850	310.0080	636.1253
Jun 2023	472.8499	362.8523	582.8474	304.6232	641.0766
Jul 2023	475.3605	362.0721	588.6488	302.1009	648.6201
Aug 2023	474.5745	358.0777	591.0713	296.4080	652.7410
Sep 2023	479.0377	359.4095	598.6659	296.0822	661.9932
Oct 2023	470.3437	347.6564	593.0310	282.7097	657.9777
Nov 2023	468.3640	342.6855	594.0424	276.1553	660.5726
Dec 2023	468.6775	340.0719	597.2832	271.9921	665.3629

Jan 2024	460.6146	330.2903	590.9390	261.3007	659.9286
Feb 2024	469.6430	337.3431	601.9428	267.3078	671.9781
Mar 2024	459.3861	325.1947	593.5775	254.1580	664.6142
Apr 2024	465.2188	329.2086	601.2290	257.2091	673.2285
May 2024	459.6888	321.9234	597.4542	248.9948	670.3827
Jun 2024	465.4015	325.9368	604.8661	252.1087	678.6943
Jul 2024	472.7518	331.6372	613.8664	256.9356	688.5680
Aug 2024	477.1520	334.4312	619.8728	258.8794	695.4246
Sep 2024	481.1418	336.8539	625.4296	260.4725	701.8110
Oct 2024	474.0341	328.2142	619.8540	251.0218	697.0464
Nov 2024	468.5845	321.2640	615.9049	243.2773	693.8916
Dec 2024	467.9153	319.1229	616.7077	240.3570	695.4736
Jan 2025	469.6549	319.6900	619.6199	240.3034	699.0065
Feb 2025	467.7970	316.6081	618.9860	236.5734	699.0207
Mar 2025	459.9330	307.5521	612.3138	226.8865	692.9794
Apr 2025	459.2569	305.7125	612.8012	224.4310	694.0827
May 2025	452.9381	298.2556	607.6207	216.3716	689.5047
Jun 2025	457.2145	301.4162	613.0128	218.9415	695.4874
Jul 2025	460.8939	304.0000	617.7877	220.9454	700.8423
Aug 2025	464.5933	306.6220	622.5646	222.9971	706.1896
Sep 2025	464.5570	305.5247	623.5893	221.3380	707.7759
Oct 2025	465.0877	305.0092	625.1661	220.2688	709.9065
Nov 2025	462.4237	301.3127	623.5347	216.0256	708.8217
Dec 2025	461.7274	299.5963	623.8585	213.7692	709.6855

Conclusion

The presence of both autoregressive and moving average components in the time series suggested that an ARMA or ETS model might be appropriate for modeling the data. Consequently, ARIMA and ETS models were developed. As the accuracy was more satisfactory with ARIMA, an ARIMA(1,1,2)(2,0,2)[12] was used to generate forecasts for future truck sales data.

Overall, I believe that the sales forecasting model can allow Heavy Haulers Inc. to manage inventory efficiently and make informed decisions regarding sales and marketing strategies.

Time Spent - Planned Vs Actual

Week	Task	Planned Hours	Actual Hours	Variance
7	Familiarizing with dataset	1	1.5	50%
	Data cleaning and preprocessing	2	3	100%
	Exploratory data analysis	3	3.75	75%
8	Reviewing literature	2	3	100%
	Selecting the forecasting method	1	1	0%
	Developing the forecasting model	2	3	100%
9	Continuing to develop the forecasting model	3	3	0%
	Documenting progress	1	2	100%
10	Validating the model	3	3	0%
	Evaluating the model	2	2	0%
11	Refining the forecasting model	3	3	0%
	Testing the model	2	1	-100%
12	Drafting the report	2	3	100%
	Finalizing the forecasting model	3	3.75	75%