**Project Report**

Amtrak Ridership Forecasting

****

**Group Members:**

1. Sonia Sunny (NetID: fh8914)

2. Anshuman Pradhan (NetID: qb8291)

3. Jai Ashokkumar Bhala (NetID: uw9410)

4. Yash Manish Dodia (NetID: ww8176)

5. Rahuldeep Shenoy (NetID: pt6377)

**Executive Summary**

For this project, data from the [United States Department of Transportation](https://www.transportation.gov/) : Bureau of Transportation Statistics has been used to predict monthly ridership for the 12 months-for the year 2014. The analysis is based on a single time series: 22-years. From the visualization we identified that the dataset has an upward trend and a seasonality, and the data is highly auto correlated, as the autocorrelation coefficients in all 12 lags are significant.

The models described below were implemented in this project:

1.Autoregressive Integrated Moving Average (ARIMA) model.

2.Advanced Smoothing Method- Holt Winters model.

3.Moving Average model.

4. ​​Regression model with Linear Trend and Seasonality + the Auto-Regressive model of its residuals

5. Regression model with Quadratic Trend and Seasonality + the Auto-Regressive model of its residuals

The model evaluation was performed based on the RMSE and MAPE accuracy metrics.

The study indicated that the best model to use was the **Centered Moving Average**.

**Introduction**

The National Railroad Passenger Corporation (Amtrak) officially began service in May 1971.

Amtrak serves more than 500 stations in 46 states and operates over a network of more than 21,000 route miles.

It is the nation’s only high-speed intercity passenger rail provider, operating at speeds up to 150 mph (241 kph) over current infrastructure. The company has more than 20,000 employees.

On an average day, nearly 85,700 passengers ride more than 300 Amtrak trains.

Ridership is highly seasonal, with July and August being the highest volume months.

**Eight Steps of Forecasting**

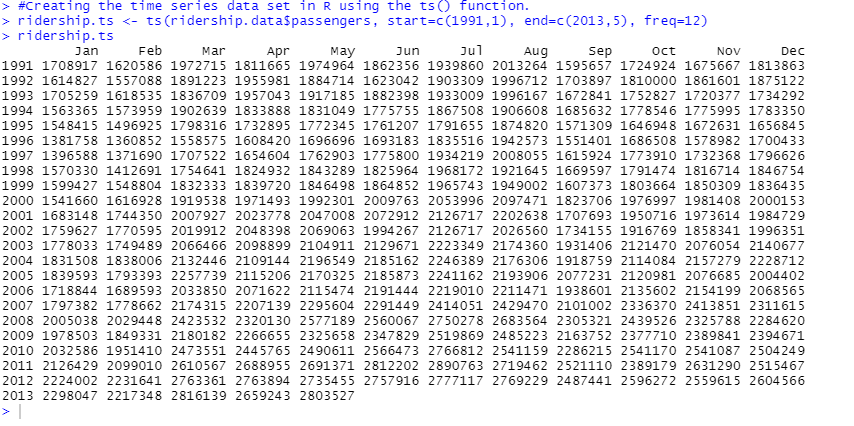
**Step 1: Define Goal**

The goal of this project is to create numeric forecasts of the Amtrak ridership for the upcoming 12 months. The objective is to create a predictive model which will properly consider both the trend and seasonal components of the historical data and effectively forecast the near future. Of course, the model with the highest accuracy will be the model of choice. The generated forecasts will be used to keep track of Amtrak ridership. The forecasting models for this project were created using the R programming language.

**Step 2: Get data**

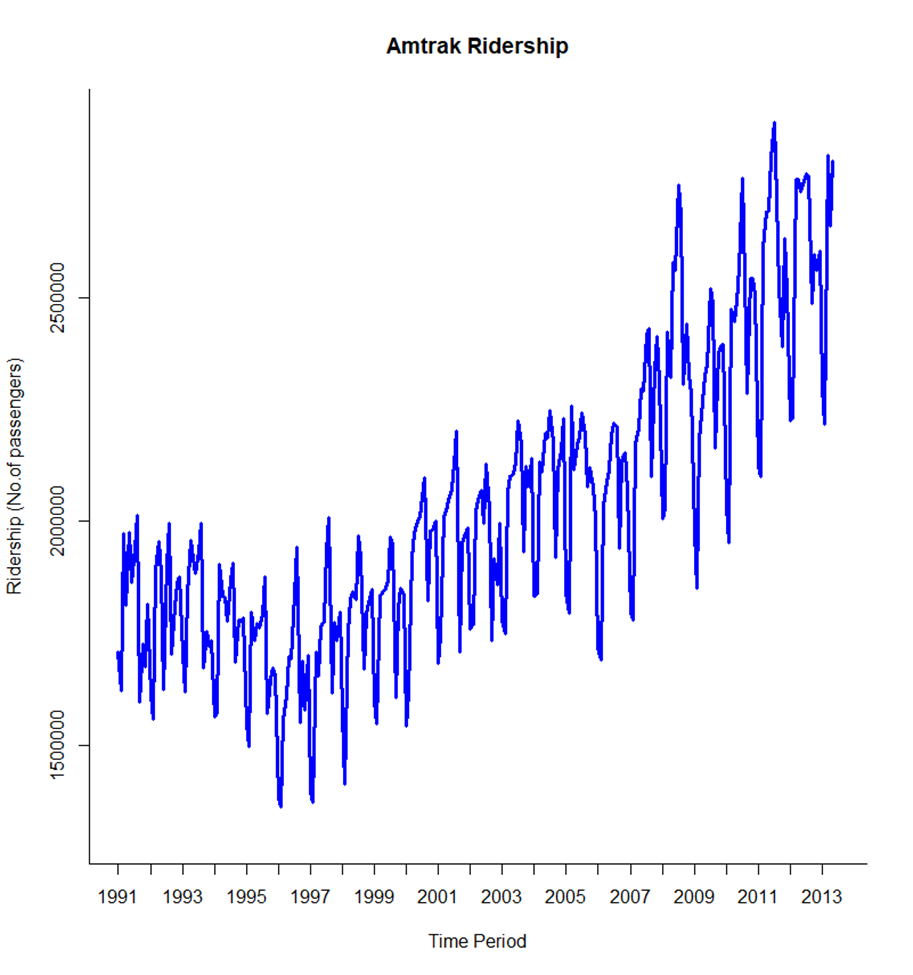
The time series dataset provided by the United States Department of Transportation: Bureau of Transportation Statistics, which records monthly Amtrak ridership, will be the focus of this research. The dataset is being investigated for a time period from January 1991 to May 2013 (a 22-year duration).

The Amtrak Ridership data was extracted and a time series data set was formed using the same.

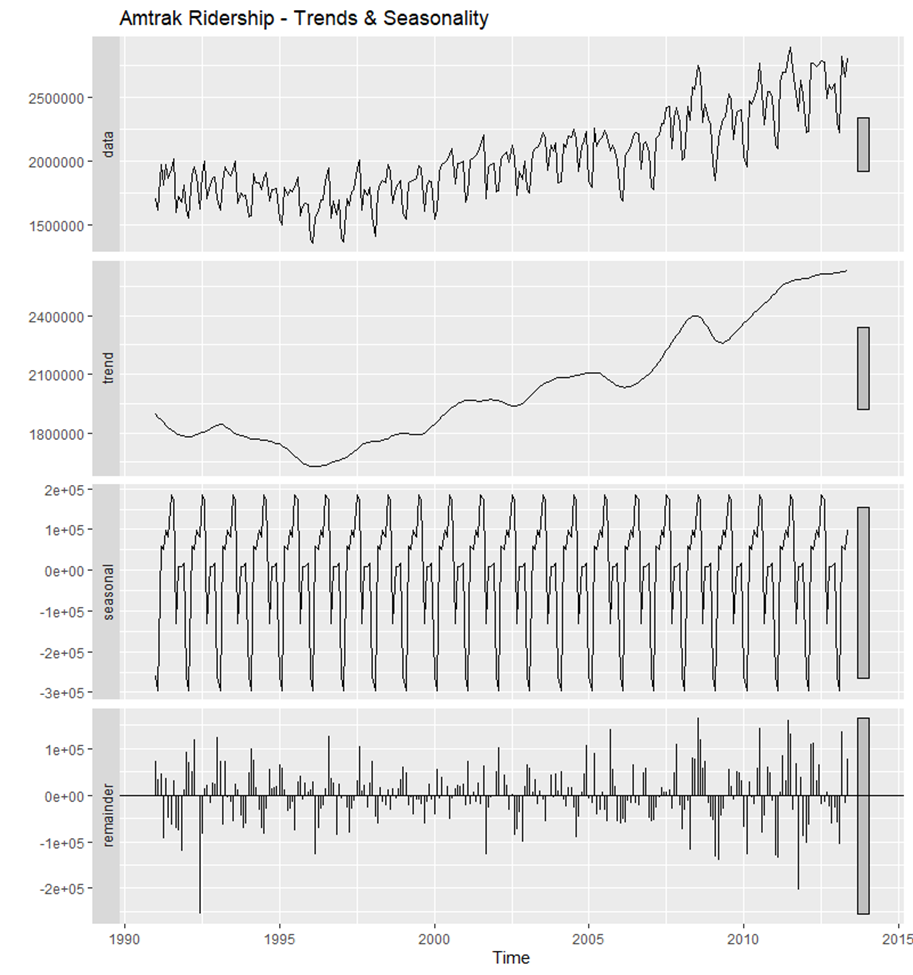


**Step 3: Explore and Visualize Series**

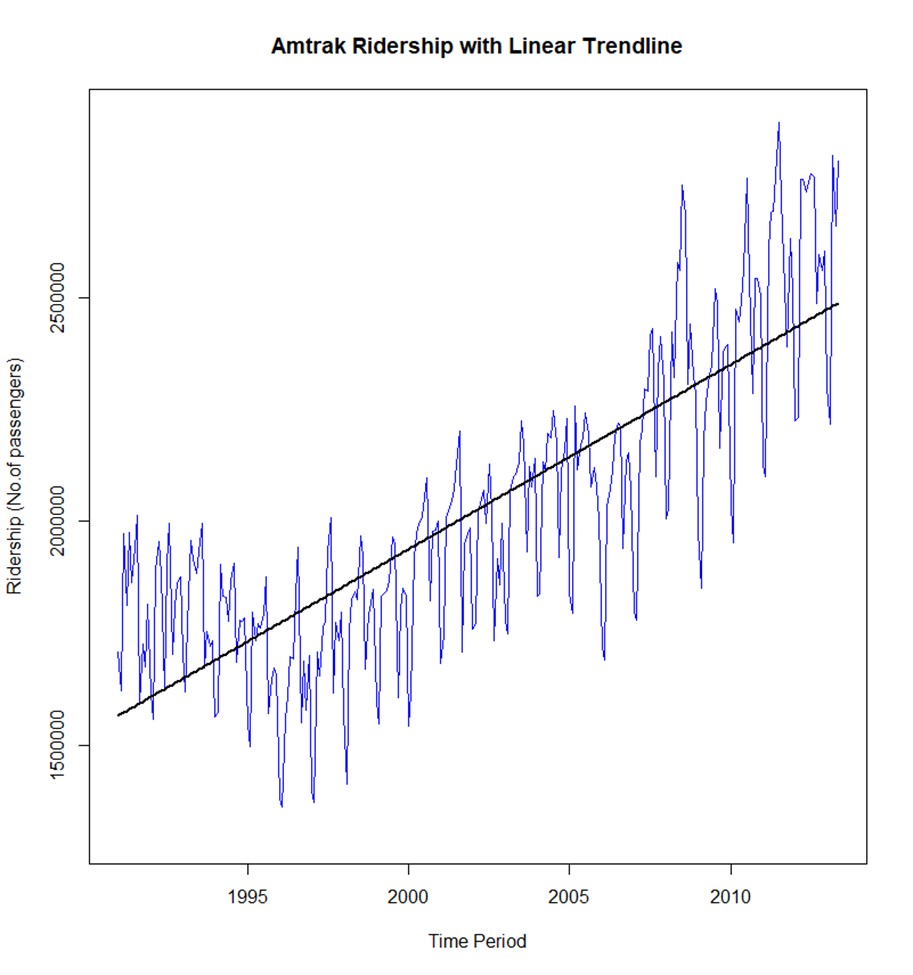
The time series data so populated was visualized using the plot function. The data plotted exhibits a linear upward trend and additive seasonality.



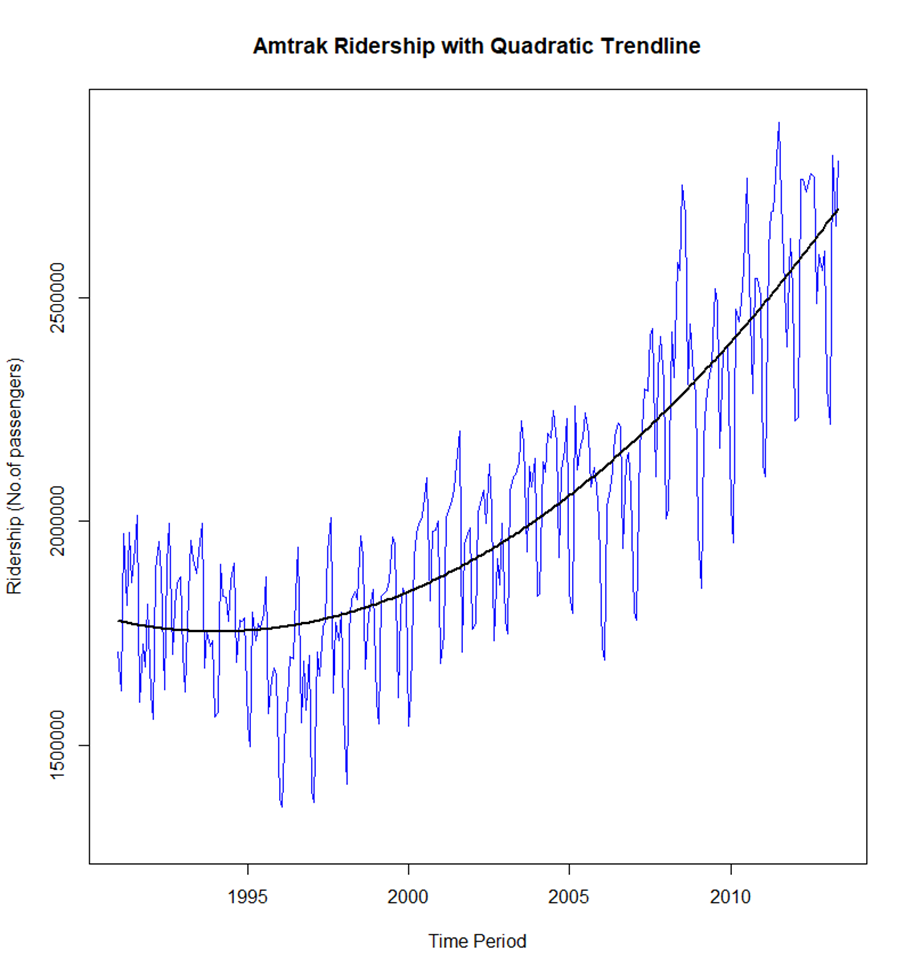
**A plot describing the trend & seasonality in the data:**

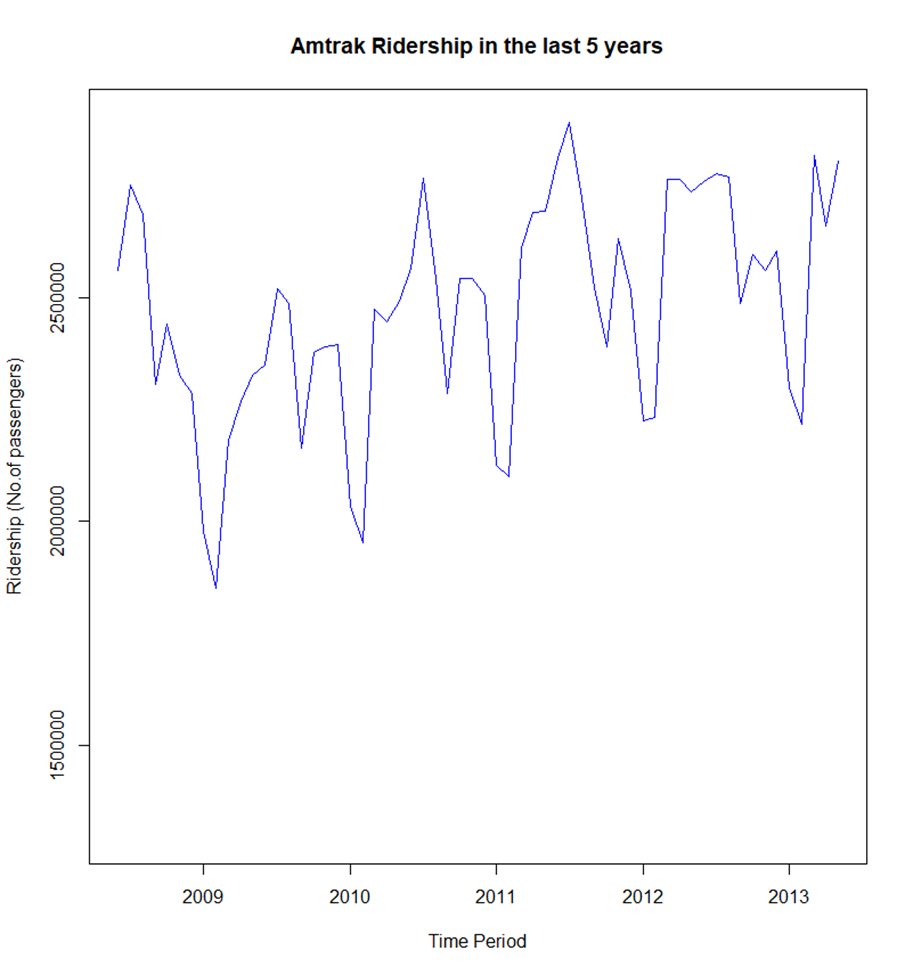


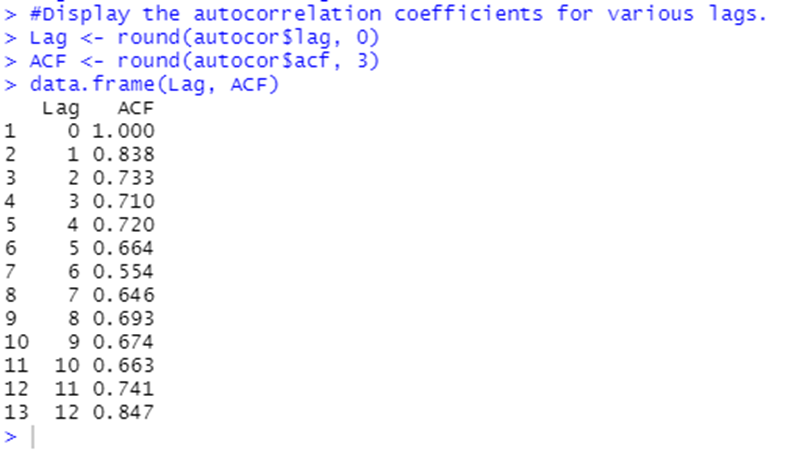
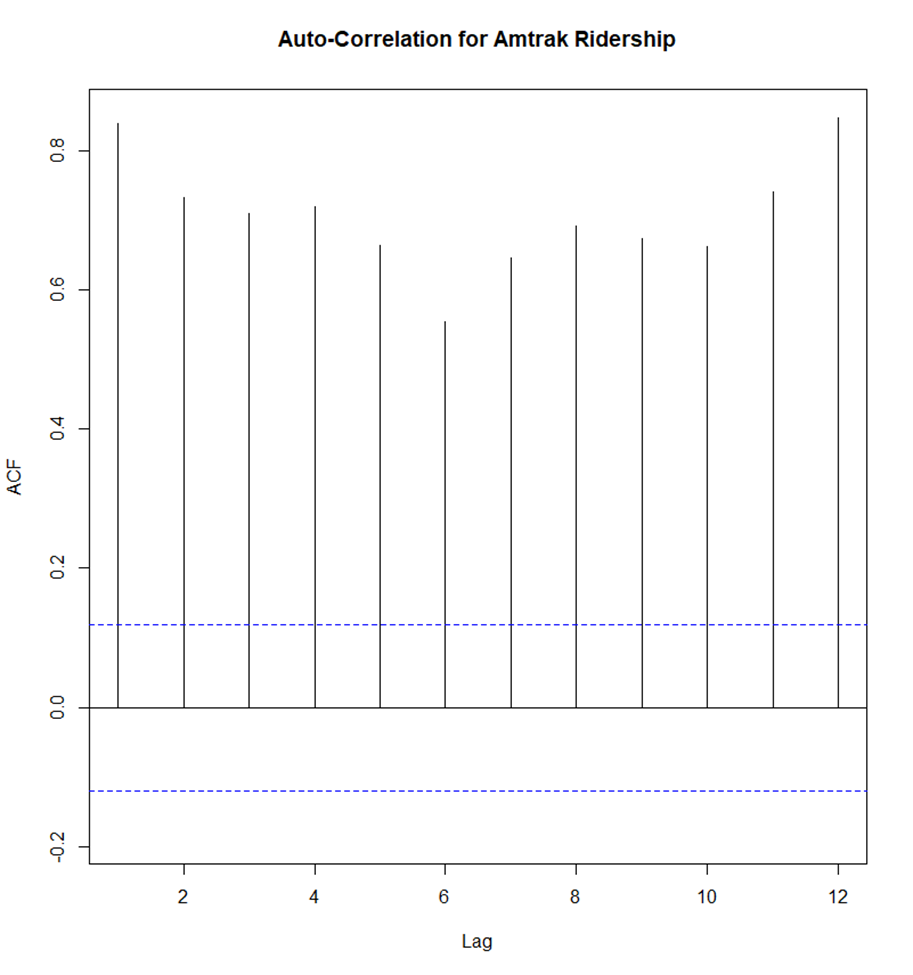
**A plot describing the linear trendline in the data:**



**A plot describing the quadratic trendline in the data:**



**A plot zooming-in on the last 5 year’s trend in the data:**

**A plot describing the auto-correlation in the data:**

The data is highly correlated, since the autocorrelation coefficients in all lags are significantly higher than the horizontal threshold as shown in the autocorrelation plot, and also the autocorrelation coefficients are significantly greater than zero, as shown above.

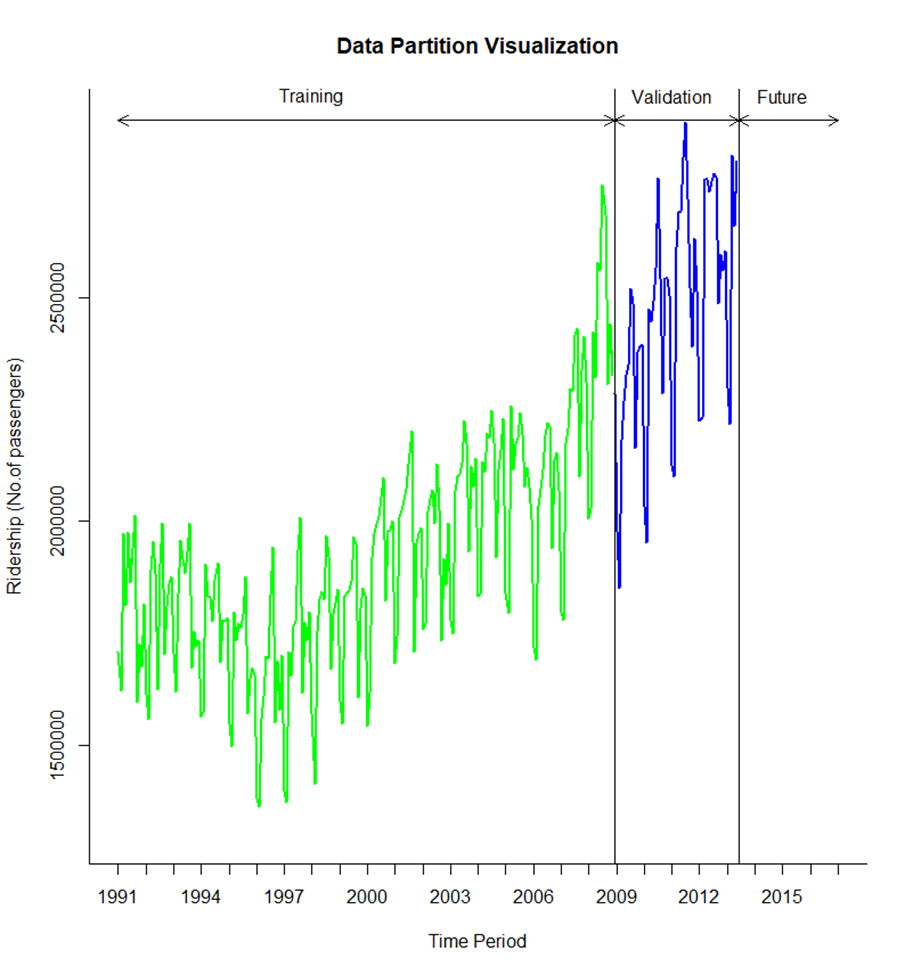
In lag 1, a positive autocorrelation coefficient is significantly higher than the horizontal threshold, indicating an upward trend component. The presence of a seasonal component in the data is shown by a positive autocorrelation coefficient in lag 12, which is also statistically significant.

**Step 4: Data Preprocessing**

In the "CSV" tab of the excel file, the original data from the Bureau of Transportation: Statistics comprises the passengers volume for various months beginning in January 1991 and ending in May 2013. The only sheet that was retained was this one; the others were discarded. All information on the CSV tab was erased except for the month and the quantity of passengers. Only the most relevant data will be evaluated for analysis in this manner. The total number of months observed was 269 months.

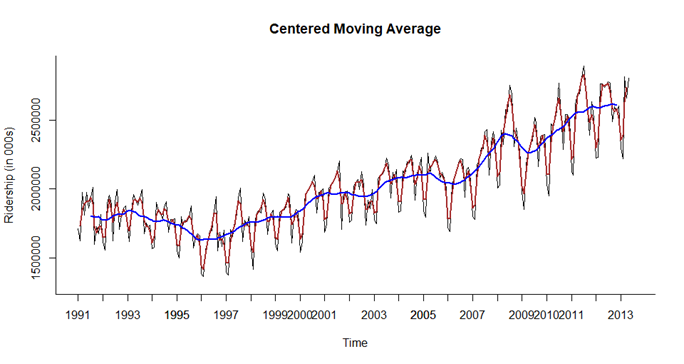
**Step 5: Partition Series**

Data partition is a crucial component to consider before applying any forecasting methodology. We established a data partition of 215 records for the training period and 54 records for the validation period off from 269 records (months).

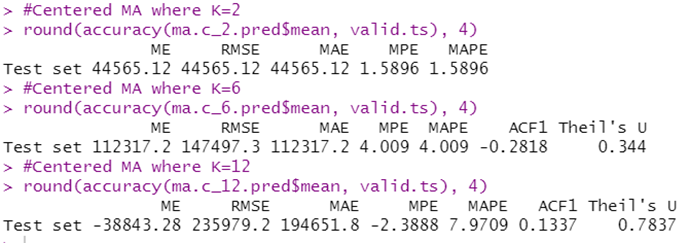
****

**Step 6 & 7: Apply Forecasting & Comparing Performance**

1. **Centered Moving Average**

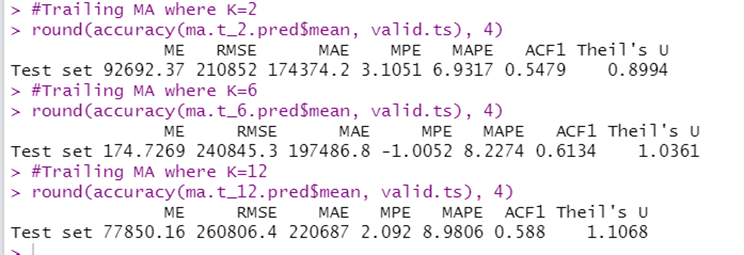
****

1. **Accuracy for Centered MA**

****

The MAPE and RMSE values for the moving average windows of 2, 6, 12 are 1.5896 and 44565.12, 4.009 and 147497.3 and 7.9709 and 234979.2 respectively. The fit seems to be for the window where k= 2.

1. **Accuracy for Trailing MA**

****

The MAPE and RMSE values for the Trailing MA with windows of K = 2, 6 and 12 are 6.9317 & 210852, 8.2274 &240845.3 and 8.9806 and 77850.16 respectively. The best fit from above seems to be when k =12.

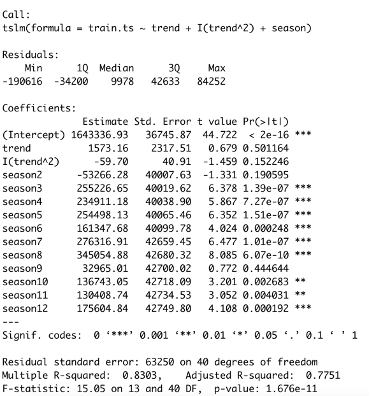
Between the Centered MA and the Trailing MA the better model is when Centered MA for k = 2.

**Regression Based Models**

We have applied various techniques for predicting regression models such as models with linear trend, quadratic trend, linear trend and seasonality, quadratic trend and seasonality. On top of it, we have tried to dig deep and predict the model with relatively better accuracy by using two-level models which are a regression model combined with trailing moving average of the residuals, and a regression model combined with an autoregressive model of the residuals. Accuracies of their performance measures are compared. In all the models, we have first evaluated the prediction accuracy for the validation data using the forecast model trained using training data. At the end, we have predicted models for the entire dataset having lower error rates.

**Forecasting for Validation period using the Training data**

We have applied the following regression models: “Linear Trend and Seasonality” and “Quadratic Trend and Seasonality” on the training data. Below are the snaps of the simultaneous summaries: -

****

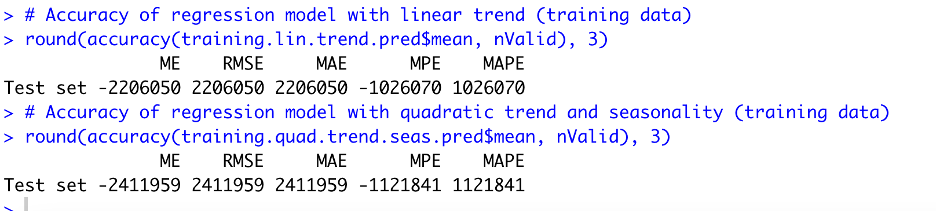
The Regression model with Linear Trend and Seasonality contains 12 independent variables: trend index (t) and 11 seasonal dummy variables for respective seasons (season2 – D2, season3 – D3 and so on). The regression model equation is: -

**yt = 1670920.0 - 1710.5 t - 53027.5 D2 + 255584.9 D3 + … + 137095.5 D11 + 182052.8 D12**

The Regression model with Quadratic Trend and Seasonality contains 13 independent variables: trend index (t), squared trend index (t2), and 11 seasonal dummy variables for respective seasons (season2 – D2, season3 – D3 and so on). The regression model equation is: -

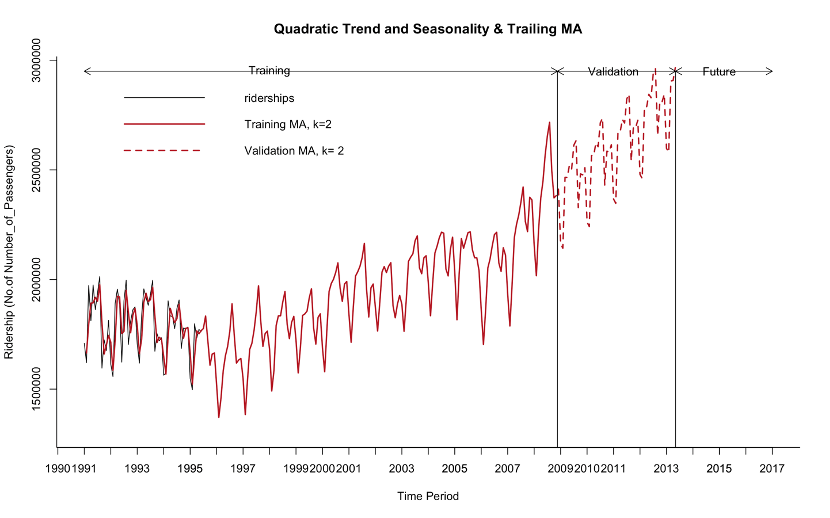
**yt = 1643336.93 + 1573.16 t - 59.70 t2 - 53266.28 D2 + 255226.65 D3 + … + 130408.74 D11 + 175604.84 D12**

According to the model summary of both the models, all the independent variables (trend and season) for respective model are statistically significant (all their p-values are all much below 0.05). The R-squared and Adjusted R-squared values are high, which represents that the models are a good fit for the training data. The intercept and coefficient for the trend (t) variable are statistically significant (p-values are much lower than 0.05). In addition, the F-statistic is also statistically significant (p-value is much lower than 0.05). Therefore, these models may be used for time series forecasting.



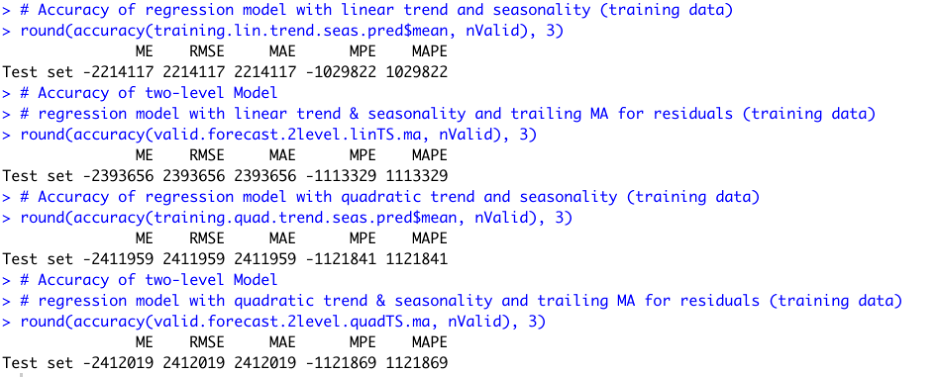
Among these two models, on comparing MAPE and RMSE values, we find the regression model with linear trend and seasonality to be a better fit as a forecasting model. We will be enhancing these models further by combining with the following: “Trailing MA model of residuals” and “Auto-Regressive model of residuals”.





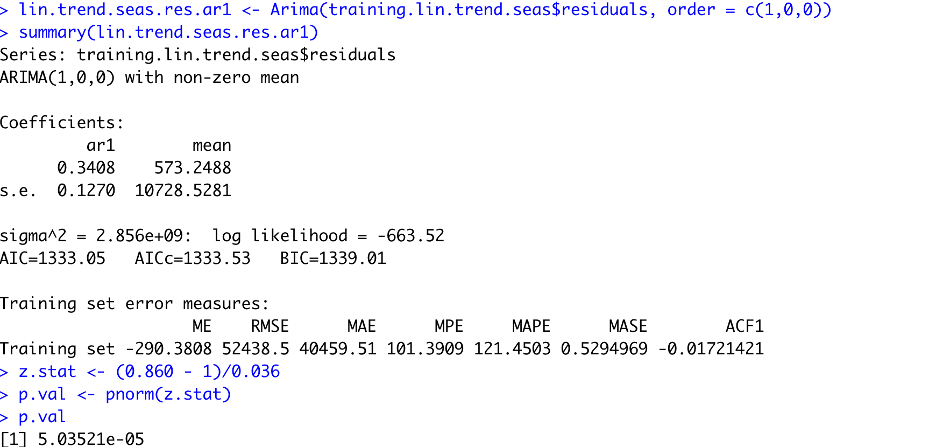
Out of above 2 plots, one is the regression model with Linear Trend and Seasonality + Trailing Moving Average of its residuals, and another is the regression model with Quadratic Trend and Seasonality + Trailing Moving Average of its residuals. In both the graphs we see that forecast models are overall fitting well with the validation data.

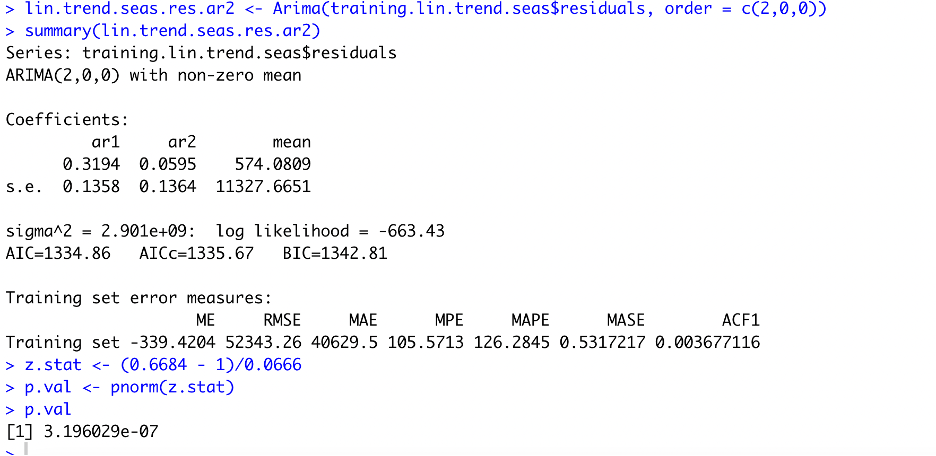
We can compare the accuracies of all the four-forecast model for the validation period as shown below:



Based on the values of RMSE and MAPE, the two-level model (regression model with Linear Trend and Seasonality) has best performance measures, as it shows relatively lower error rates than other models.

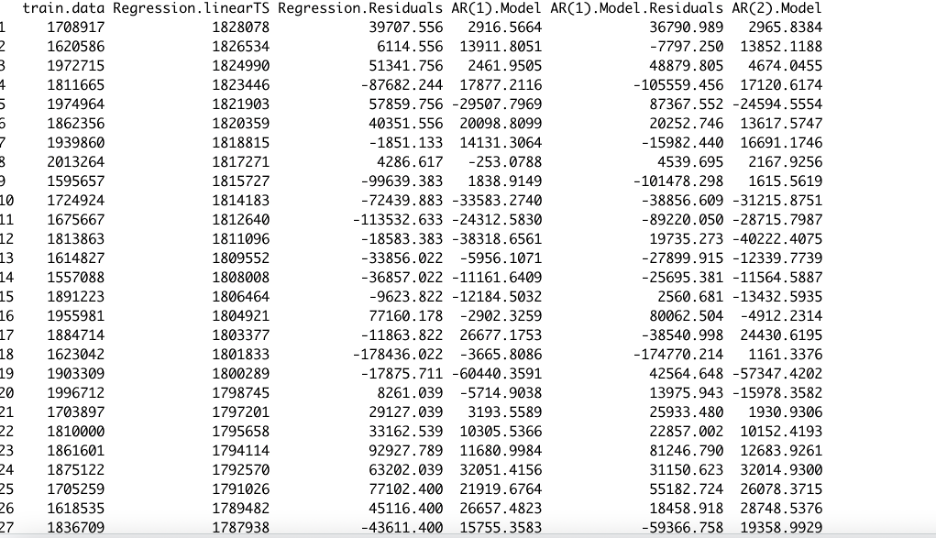
Further, we developed a combined forecast model using the regression model with Linear Trend and Seasonality + Autoregressive model of its residuals. Below are shown the summaries of AR(1) model of residuals and AR(2) model of residuals:



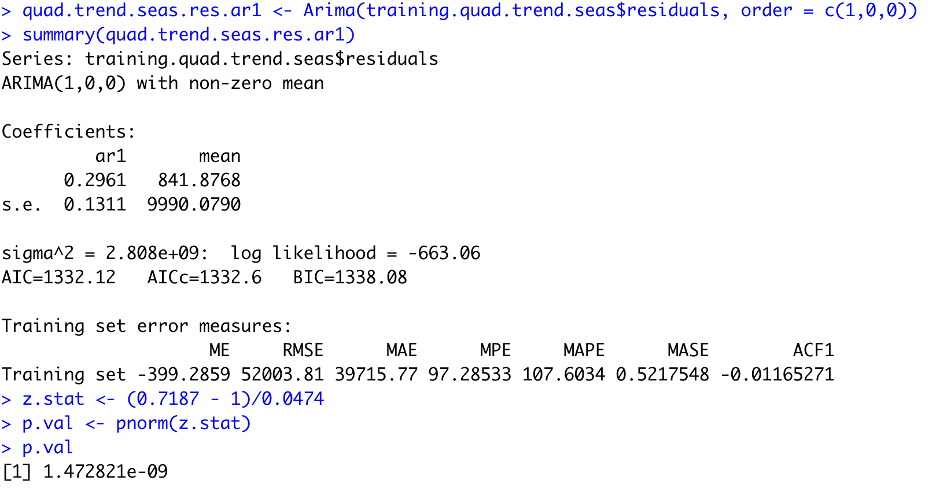


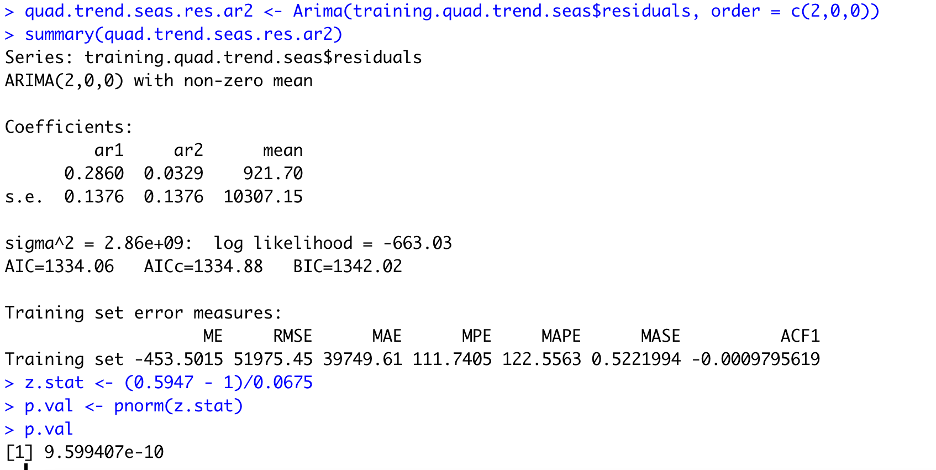
Based on the coefficients of both the models, we observed that the AR(2) model of residuals seems to be more significant than AR(1) model of residuals. Hence, we will develop a two-level model using the regression model with Linear Trend and Seasonality + Autoregressive (order 2) model of its residuals to forecast for the validation period.

Following is a table showing the validation data, regression with linear trend and seasonality forecast data and the combined forecast data:



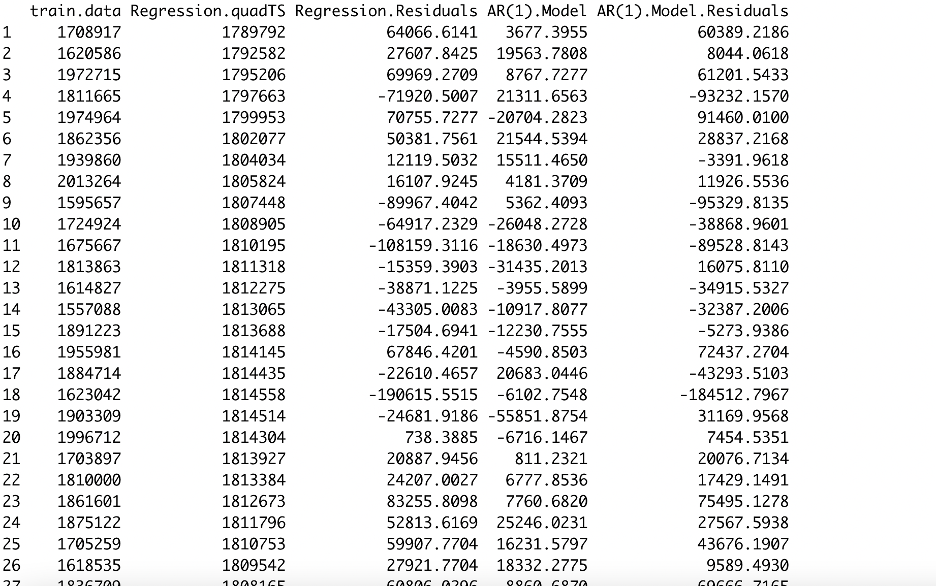
We now developed a combined forecast model using the regression model with Quadratic Trend and Seasonality + Autoregressive model of its residuals. Below are shown the summaries of AR(1) model of residuals and AR(2) model of residuals:



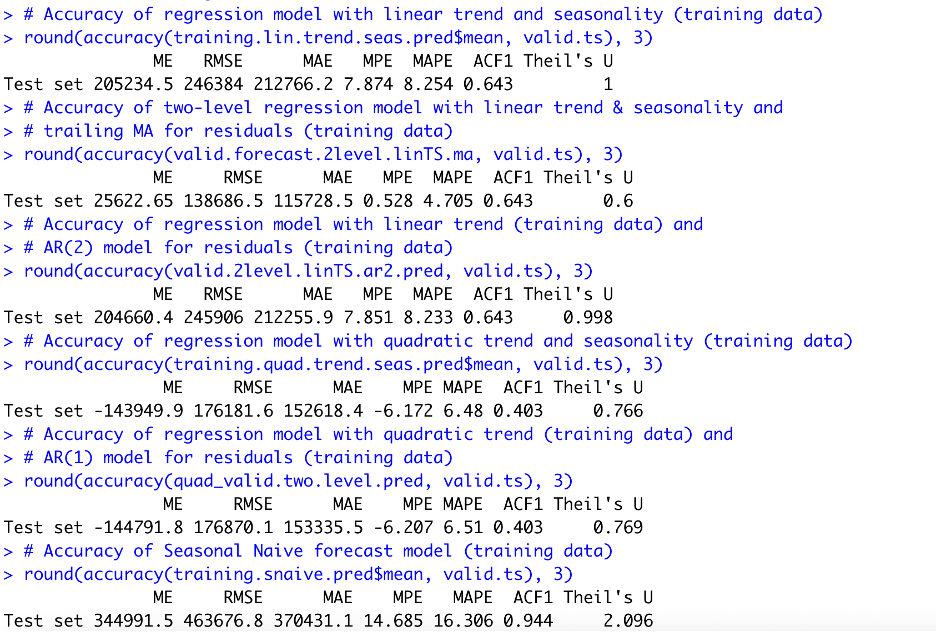


Based on the coefficients of both the models, we observed that the AR(1) model of residuals seems to be more significant than AR(2) model of residuals. Hence, we will develop a two-level model using the regression model with Quadratic Trend and Seasonality + Autoregressive (order 1) model of its residuals to forecast for the validation period.

Following is a table showing the validation data, regression with quadratic trend and seasonality forecast data and the combined forecast data:

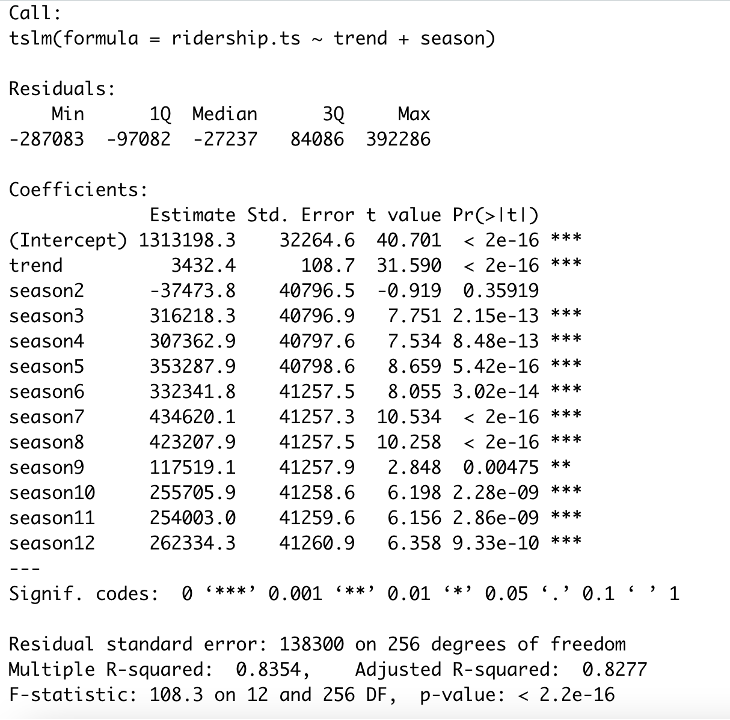


Let us now compare the accuracy performance measures for various regression model (simple and multiple) along with a baseline model – Seasonal Naïve forecast model:



​​**Forecasting for future 12 periods, using the Entire Dataset**

To perform the prediction for future periods using the entire dataset, we will create forecast models with the top three models recognized while forecasting using the training dataset for the validation period.

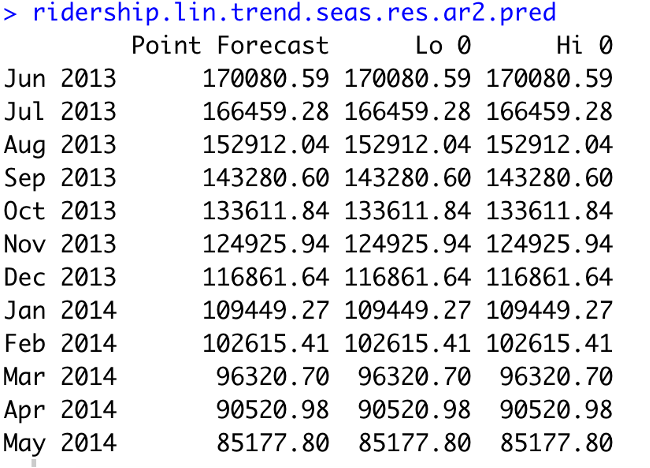


The Regression model with Linear Trend and Seasonality contains 12 independent variables: trend index (t) and 11 seasonal dummy variables for respective seasons (season2 – D2, season3 – D3 and so on). The regression model equation is: -

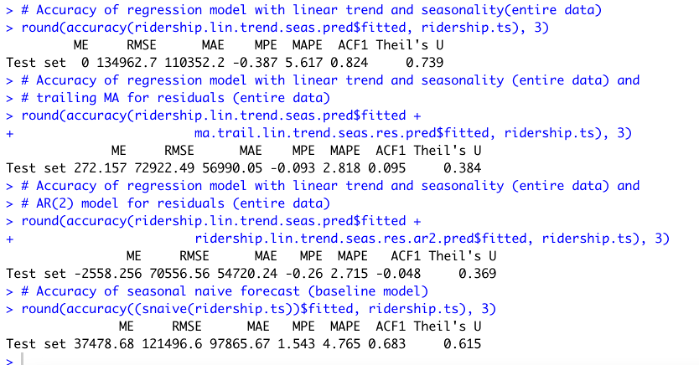
**yt = 1313198.3 + 3432.4 t - 37473.8 D2 + 316218.3 D3 + … + 254003.0 D11 + 262334.3 D12**

As per the model summary, the regression model with linear trend and seasonality is statistically significant. F-statistic value is high and F-statistic’s p-value for the model is 2.2 \* 10 ^ -16, which is much lower than the level of significance (0.05). Additionally, R-squared value (0.8354) and Adjusted R-squared value (0.8277) are high, and all the regression coefficients are statistically significant (p-value < 0.05). Overall, the regression model is very good fit for the historical dataset and thus, can be used for forecasting Amtrak Ridership data.

The forecast below are for the next one year using Two-Level Forecast Regression Model (Regression Model with Linear Trend and Seasonality + Auto-Regressive (order 2) Model of its Residuals): -



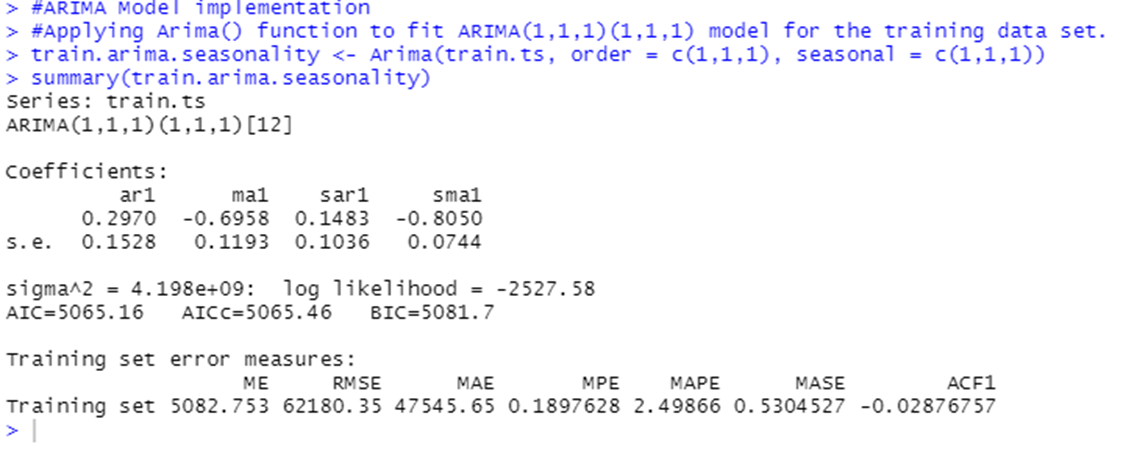
We compare the accuracy performance measures for the three regression models (simple and multiple) along with a baseline model – Seasonal Naïve forecast model: -



​​Based on the accuracy performance measures, comparatively lowest MAPE value is 2.715 % and lowest RMSE value is 70556.56. Therefore, the best model among the four models is the two-level model using the Regression model with Linear Trend and Seasonality + the Auto-Regressive (order 2) model of its residuals.

**Autoregressive Integrated Moving Average (ARIMA) Model**

The summary for the ARIMA(1,1,1)(1,1,1)4 model for the training data set is given below:

****

The parameters for the ARIMA() function ARIMA(p, d, q)(P, D, Q)m, are given below:

• p=1, order 1 autoregressive model AR(1)

• d=1, order 1 differencing to remove linear trend

• q=1, order 1 moving average MA(1) for error lags

• P=1, order 1 autoregressive model AR(1) for seasonality

• D=1, order 1 differencing for seasonality

• Q=1, order 1 moving average MA(1) for error lags

• m=4, for quarterly seasonality

The ARIMA model equation is given by:

**yt - yt-1 = 0.2970 (yt-1 - yt-2) – 0.6958 et - 1 + 0.1483 (yt-1 - yt-5) - 0.8050 ρt-1**

yt - yt-1 is the first differencing model because d = 1.

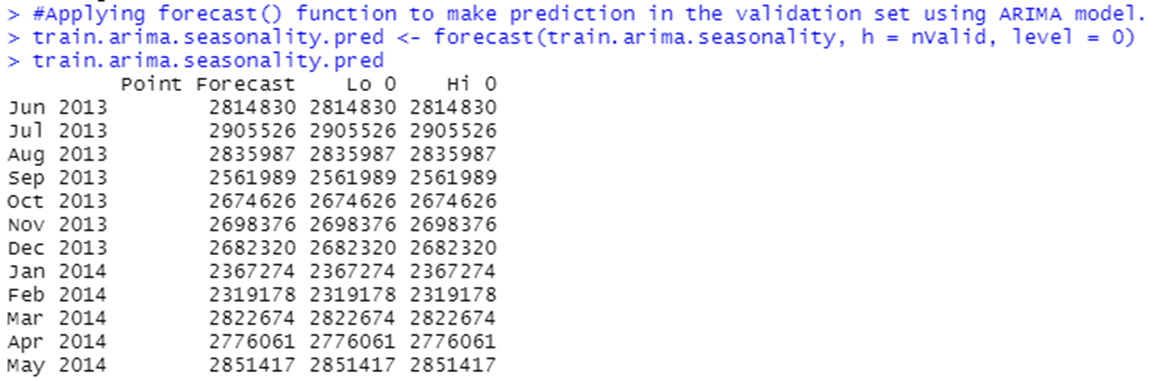
ar1 = 0.2970 is the autoregression.

ma1 = -0.6958 is the moving average for lags.

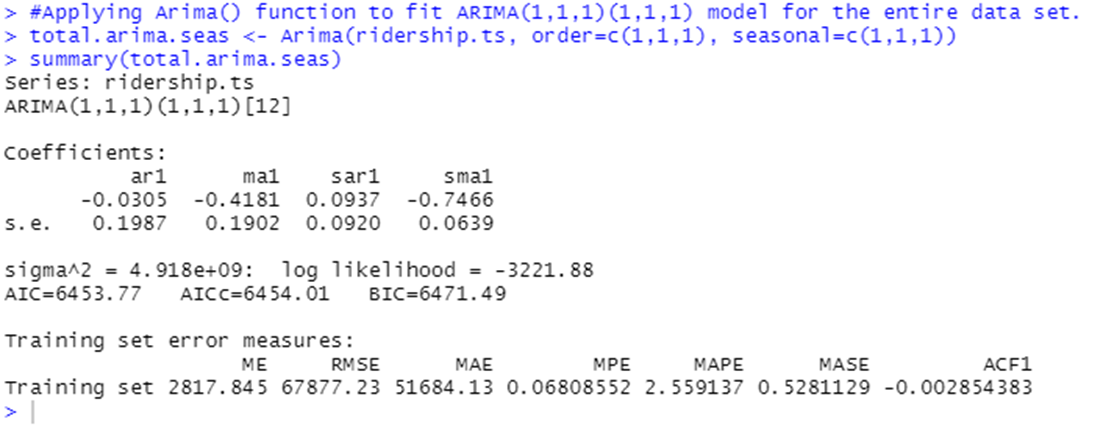
sar1 = 0.1483 is the autoregressive model for seasonality.

sma1 = -0.8050 is moving average for seasonality.

Forecasting for the ARIMA model is given below:

****

The summary for the ARIMA model for the entire data set is given below :

****

The ARIMA model equation is given by:

**yt - yt-1 = -0.0305 (yt-1 - yt-2) - 0.4181 et - 1 + 0.0937 (yt-1 - yt-5) - 0.7466 ρt-1**

yt - yt-1 is the first differencing model because d = 1.

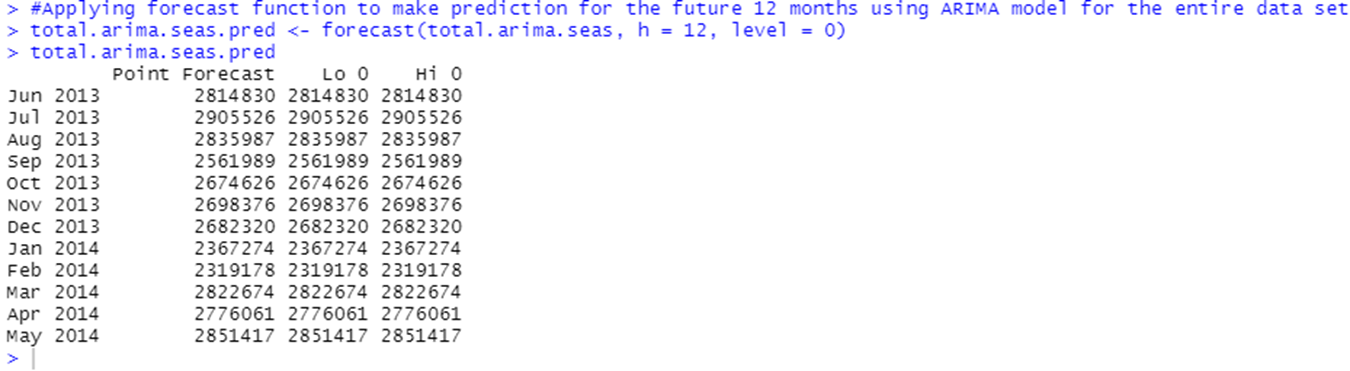
ar1 = -0.0305 is the autoregression.

ma1 = -0.4181 is the moving average for lags.

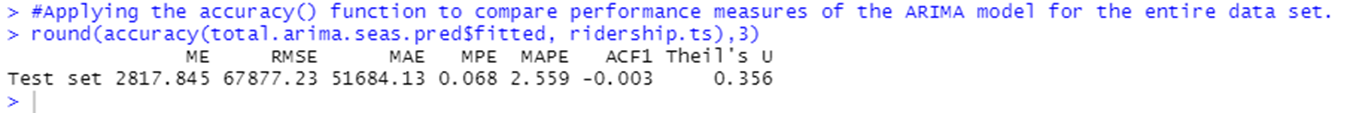
sar1 = 0.0937 is the autoregressive model for seasonality.

sma1 = -0.7466 is moving average for seasonality.

Forecasting for the ARIMA model on the entire data set is given below:

****

The performance measures of the ARIMA model is given below:



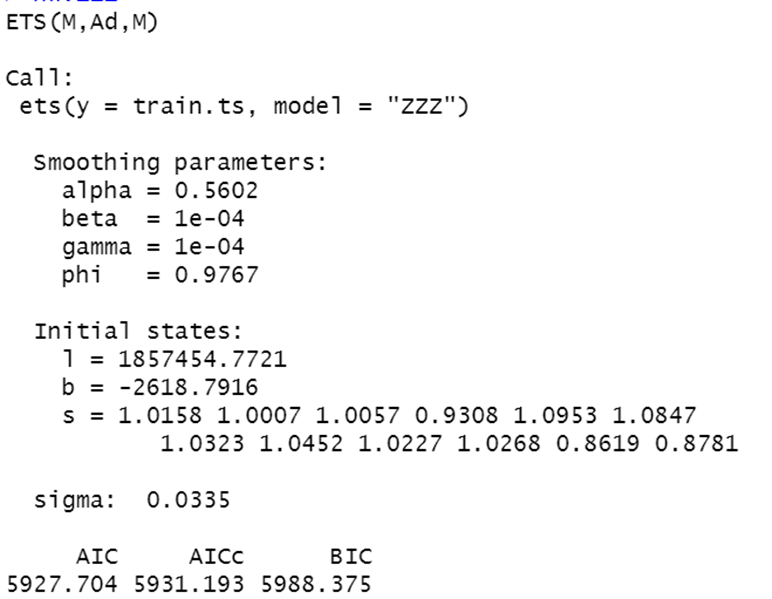
Upon comparing the MAPE, RMSE measures, it can be observed that the **MAPE = 2.559** and the **RMSE = 67877.23** for the ARIMA (1,1,1)(1,1,1) model.

**Holt-Winters Model**

Holt-Winter’s is one of the advanced exponential Smoothing models which is ideal for data sets that have trend and seasonal variation. The Amtrak Ridership historical data set has seasonality and trend components. Hence, we developed the Holt-Winters Model with the automated selection of error, trend and seasonality options. The model was first developed on the training partition and applied to forecast using the validation partitions. Then, it was re-runed on the entire data set.

**Holt-Winters Model on Training Partition**

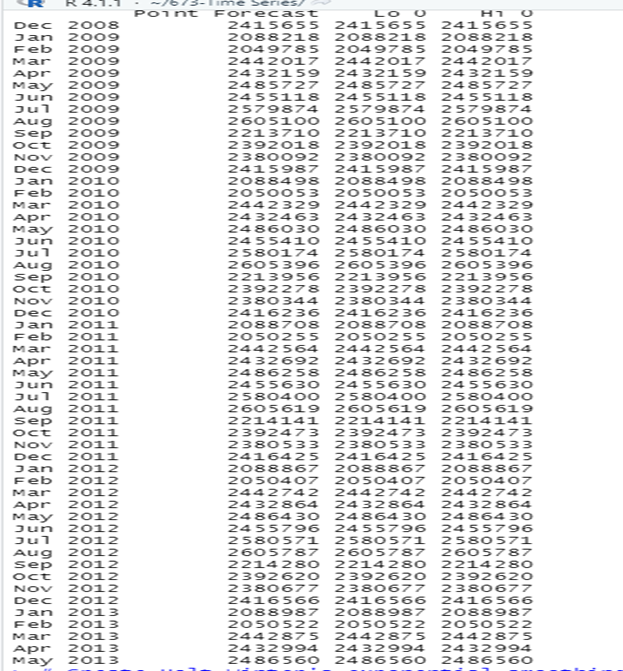
The automated selection of error, trend and seasonality option of Holt’s model on the training partition was applied and below is the output of the model’s summary.



The summary shows that the optimal model for the training data set is (M, Ad, M) which indicates multiplicative error, additive trend and multiplicative seasonality. As displayed in the output summary, the optimal value for exponential smoothing constant (alpha) is 0.5602 and smoothing constant for trend estimate (beta) is 0.0001, smoothing constant for seasonality estimate (gamma) is 0.0001. The alpha value of this model indicates that the model’s level component tends to be more local, on the other hand, the smoothing constant for trend (beta) and smoothing constant for seasonality estimate (gamma) are close to zero which indicates that the trend and seasonal components are changing slowly over time.

**Forecast on Validation Period**

Forecast was made on the validation period using the Holt-Winters Model with automated selection of error, trend and seasonality component with no confidence level.

The HW model’s forecast in the validation period is presented below:

Forecast plot chart is displayed below:

The plot shows that the forecast model is a good fit. The lines for the historical data set and the lines for when the model is used on the training data set and the forecast on the validation are very close to each other.

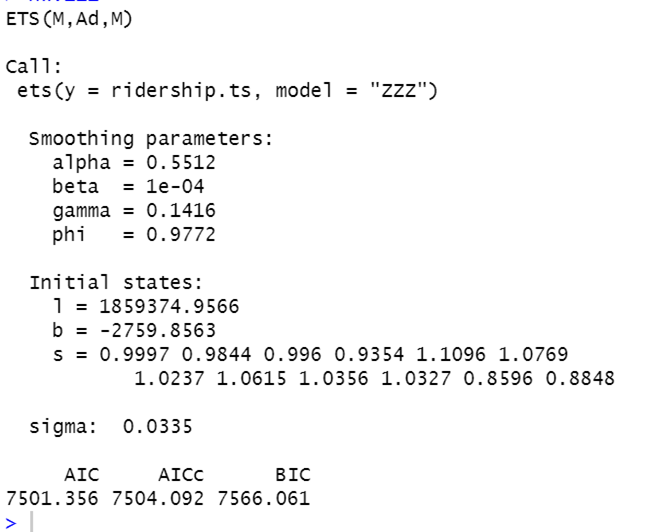
**Accuracy comparison for automated Holt-Winters Model, Seasonal Naïve and Naïve Models**

| Model | RMSE | MAPE(%) |
| --- | --- | --- |
| Automated Holt-Winters | 188621.1 | 6.318 |
| Seasonal Naïve | 195523.9 | 6.577 |
| Naïve | 288404.3 | 9.772 |

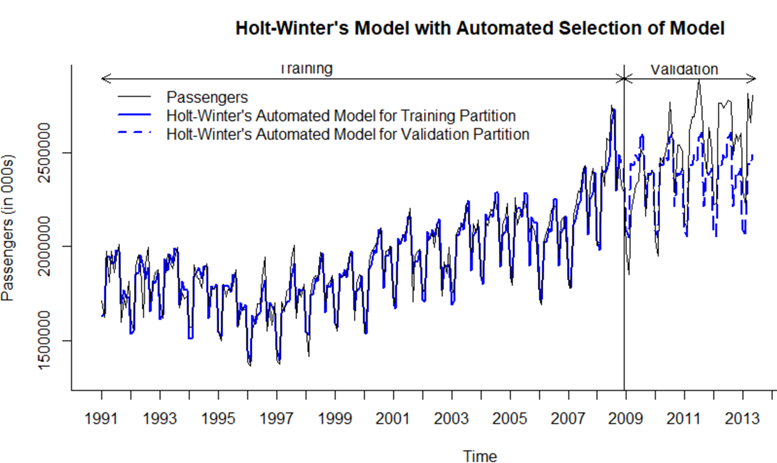
Upon comparing the automated Holt’s model with Seasonal Naïve and Naïve, Holt’s model has the lowest RMSE (188621) and MAPE (6.318%).

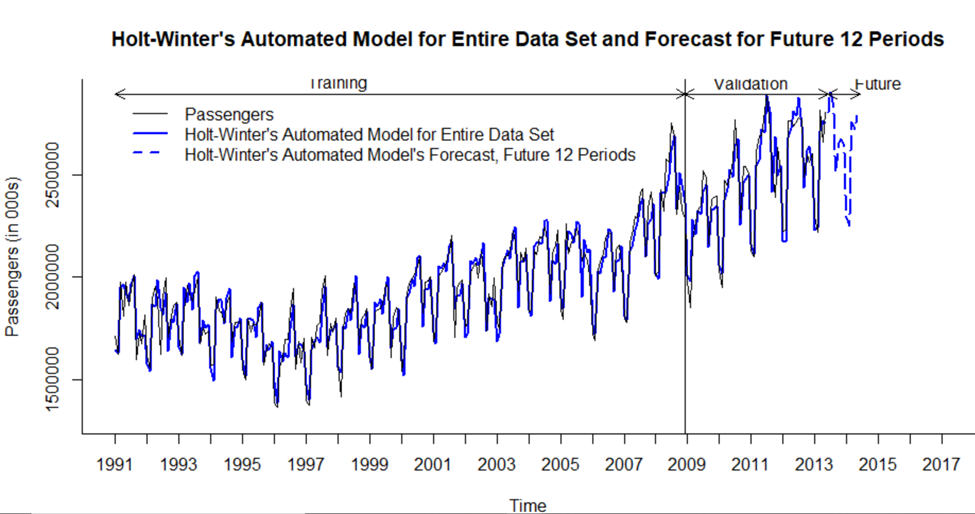
**Holt-Winters Model with Automated Selection of Error, Trend and Seasonality on the Entire Set**

The model performed well in forecasting using the validation period. It needs to be re-run on the entire data set before we use it to forecast for the future period. The following summary output was obtained when the model was used on the entire data set.



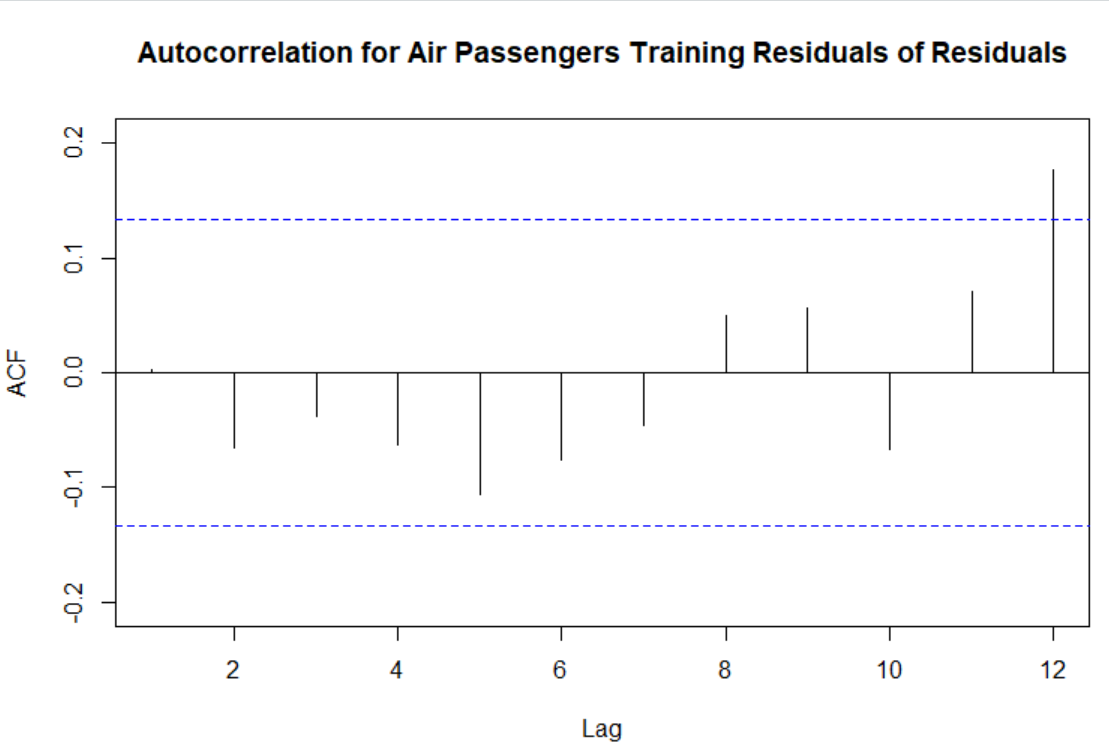
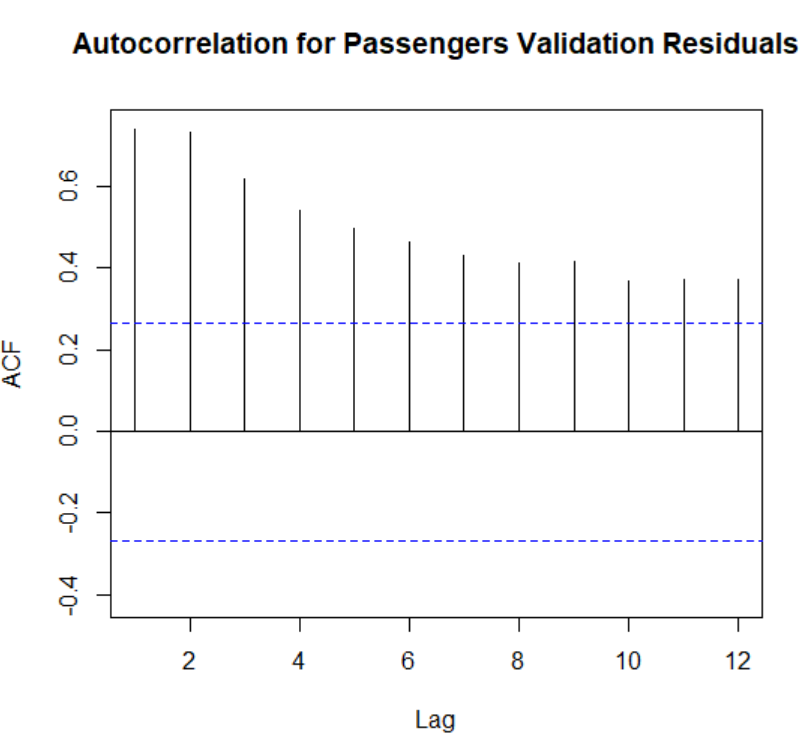
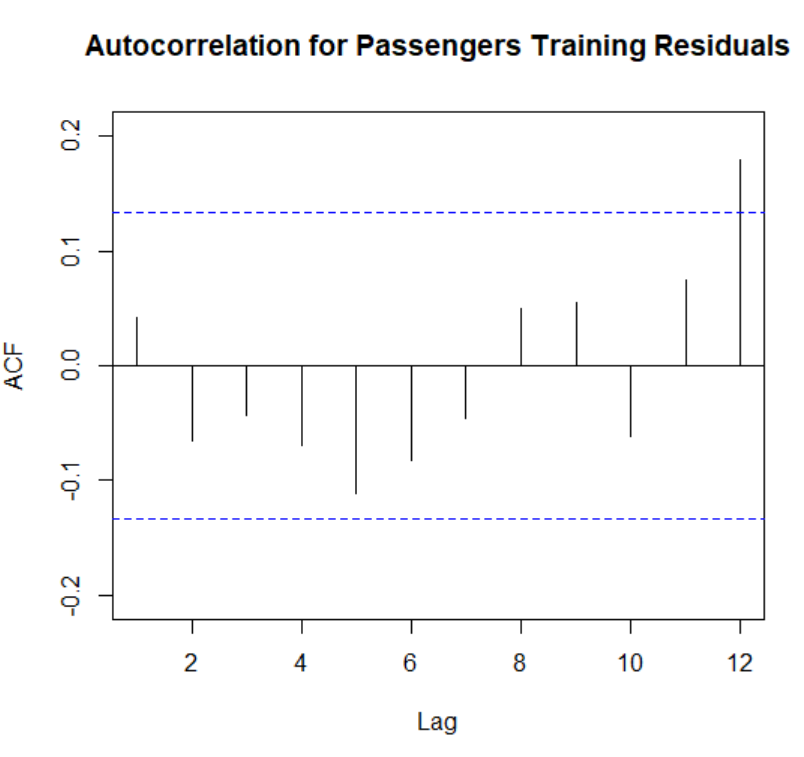
Similar to output for the model on the training data set, this HW model has the (M, Ad, M) options which indicates multiplicative error, additional trend, and multiplicative seasonality. The optimal value for exponential smoothing constant (alpha) is 0.5512, suggesting that the model’s level component tends to be more local. Smoothing constant for trend estimate (beta) is 0.0001, smoothing constant for seasonality estimate (gamma) is 0.1416. The alpha value of this model indicates that the model’s level component tends to be more local, on the other hand, the smoothing constant for trend (beta) and smoothing constant for seasonality estimate (gamma) are close to zero which indicates that the trend and seasonal components are changing slowly over time.



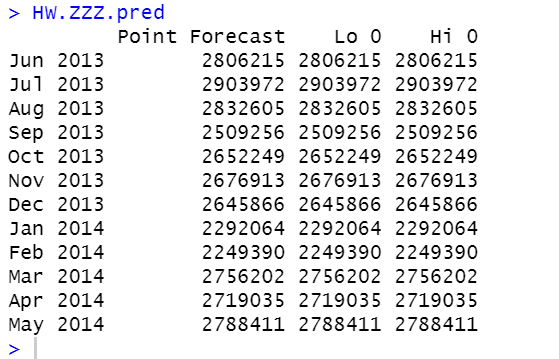


The accuracy of the automated Holt-Winters model is having very low RMSE(66102.04) and MAPE(2.604%) when compared with the Seasonal Naïve model.

**Automated Holt-Winters Model with an Autoregressive, AR(1) Model for Residuals**

Before applying a model to smooth the residual of the Automated Holt-Winters the Af() function was used to develop autocorrelation plots developed to review if relation exists between the residuals and below are the plots made on validation and training.

The autocorrelation for the training residuals on almost all the lags are insignificant, except for lag 12 which is not also that strong. But the autocorrelation for the validation residuals on all lags are significant, specially lag 1 is very strong indicating. Hence, AR(1) Model on the residuals to incorporate the autocorrelation which were not handled by the Holt-Winters model. The AR(1) model was used both in the training partition and in the entire data set and a combined two-level model developed accordingly. Below is a forecast table for forecasts made in the 12 months in future .



| **Model** | **RMSE** | **MAPE (%)** |
| --- | --- | --- |
| **Automated Holt-Winter’s** | 66102.04 | 2.604 |
| **Combine Model (HW(ZZZ) +AR (1)** | 188621.1 | 6.318 |
| **Seasonal Naïve** | 121496.6 | 4.765 |

It can be concluded that the automated Holt-Winters model is performing well without adding additional models for the residuals of the model.

**Step 8: Implement Forecast**

Rank Best Model Per Methodology RMSE MAPE

| **Model** | **RMSE** | **MAPE (%)** |
| --- | --- | --- |
| **Automated Holt-Winter’s** | 66102.04 | 2.604 |
| **Centered Moving Average** | 44565.12 | 1.5896 |
| **ARIMA (1,1,1)(1,1,1)** | 67877.23 | 2.559 |
| **Regression model with Linear Trend and Seasonality + Auto Regressive (Order 2) model of residual** | 70556.56 | 2.715 |

As seen from the above comparison table it is evident that Centered moving average gives the best predictions with an RMSE and MAPE values of 44565.12 and 1.5896% respectively. This is the recommended model to implement when forecasting for Amtrak Ridership for passengers traveling.

**Conclusion**

The model of choice as per the above computations is Centered Moving Average model which gives the best predictions. Additionally, it can also be observed that Holt Winteralso provides a good forecast but the least RMSE and MAPE is something to look for to choose the best model and with this we can forecast with minimum error with the kind of ridership data we have.

**Bibliography**

* Data(xls File) **-** [**https://www.bts.gov/sites/bts.dot.gov/files/legacy/amtrak\_ridership\_3.xls**](https://www.bts.gov/sites/bts.dot.gov/files/legacy/amtrak_ridership_3.xls)
* Website (Amtrak Ridership**) -**[**https://www.bts.gov/archive/publications/multimodal\_transportation\_indicators/2013\_08/passenger/amtrak\_ridership**](https://www.bts.gov/archive/publications/multimodal_transportation_indicators/2013_08/passenger/amtrak_ridership)