FORECASTING

Assignment Report

APRIL 30, 2022

HARIKRISHNAN GOPALAKRISHNAN

EXERCISE-1

Data description:

Airpass_BE

This dataset contains the number of monthly air passengers traveling from reporting country Belgium(BE) and EU partner country during the time period January 2003 until July 2021.

The data set contains 2 columns

Date: Information of month and year of travel

Airpass_BE: Information about the number of air passengers travelled from Belgium.

Objective :

The objective is to Analyze two different sets of Time Series Data through Data Description and also split the data into a Train Set & Test Set to apply different forecasting models and make predictions. Additionally, we select the best model based on analyzing the results and comparing the performance of the models. This selected model is to predict the forecasts and final inferences have to provided.

Note:

The time-series data is split into the training set and test set.

Training set: January 2003-December 2017

Test set: January 2018 – February 2020

The values after February 2020 are not considered in the time-series test set because impact the efficiency of the forecasting models and it will distort the assessment for forecast accuracy. So we can keep them for later reference.

Explanatory Data Analysis

The given data set is converted into time series data using the ts () function.Let us explore the raw time-series data and review the data with summary statistics and plots in R.

The frequency of the time series is 12 i.e data with monthly seasonality.

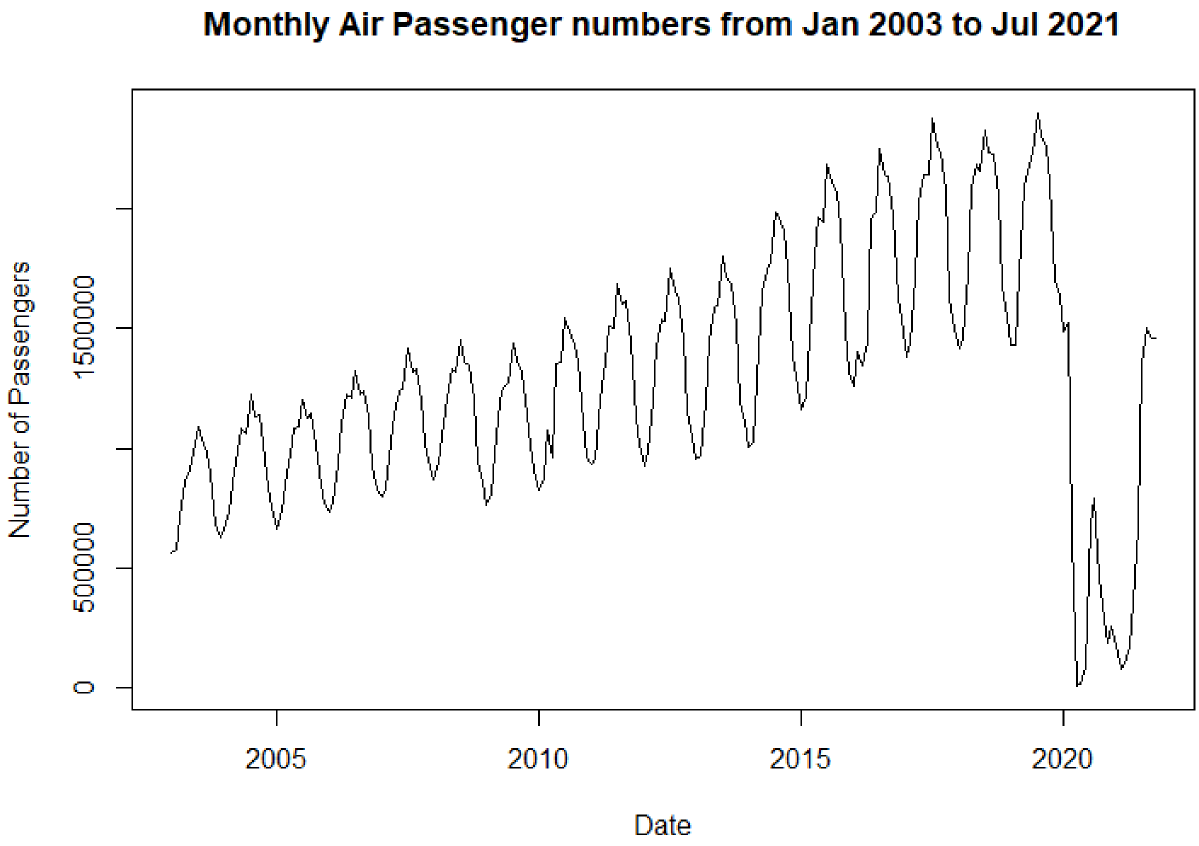
Data summary :

The summary table of the data suggests that the Minimum number of passengers traveled is 9390 and the Maximum number of passengers traveled is 1258060. The average number of passengers traveled is 1298757.

Airpass_BE	Values
Min.	9390
1st Qu.	964807
Median	1258060
Mean	1298757
3rd Qu.	1612030
Max.	2397567

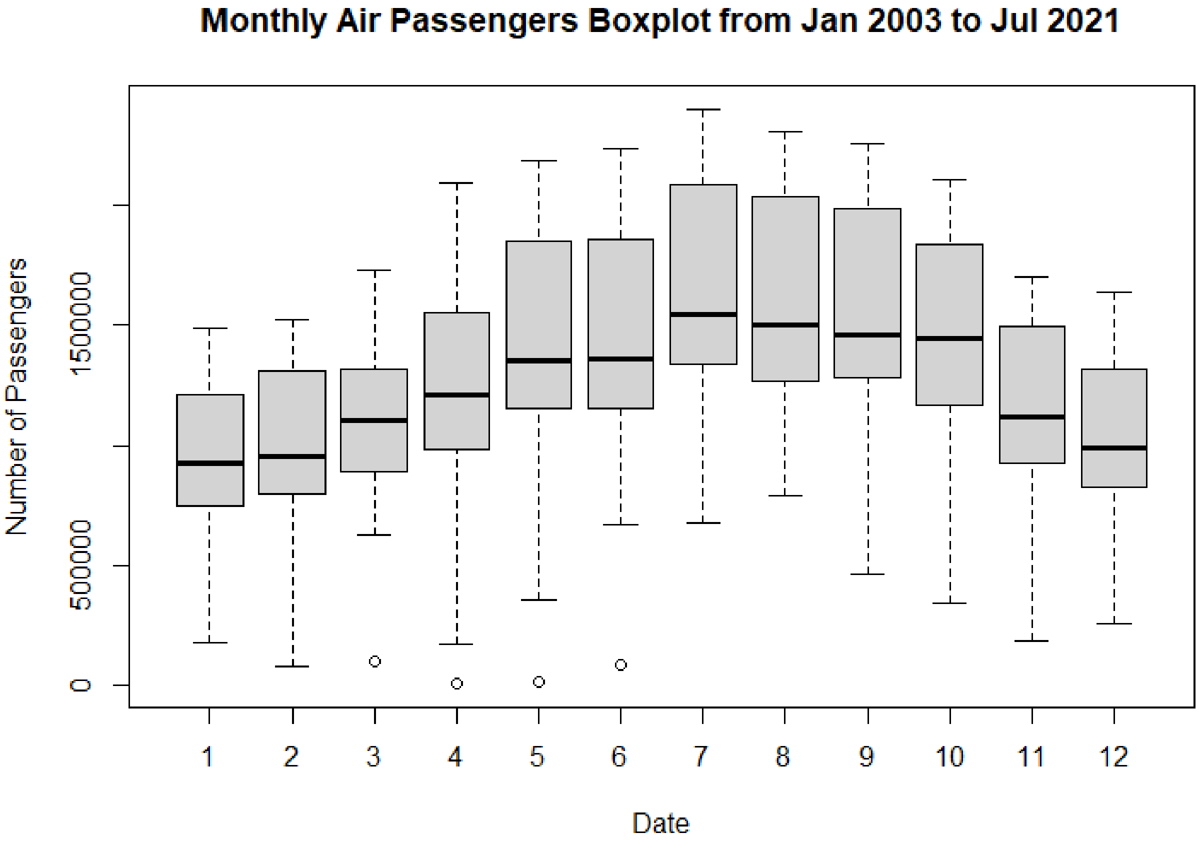
Visualizing Seasonality and Trend:

1.Base plot function



From the base plot of the raw data, it can be observed that the time series have a strong seasonality component that may be either linear or quadratic. Moreover, the magnitude of the seasonal variation increases as the number of passenger arrivals increases. So there is an upward trend throughout the time series data till 2020 February. After Feb 2020, there is a huge decline in the number of passengers due to the covid-19 outbreak which lead to the cancellation of many flight services.

2.Box plot function:



From the above Box plot,we can see the average number of passengers traveled monthwise from jan 2003 to jul 2021.

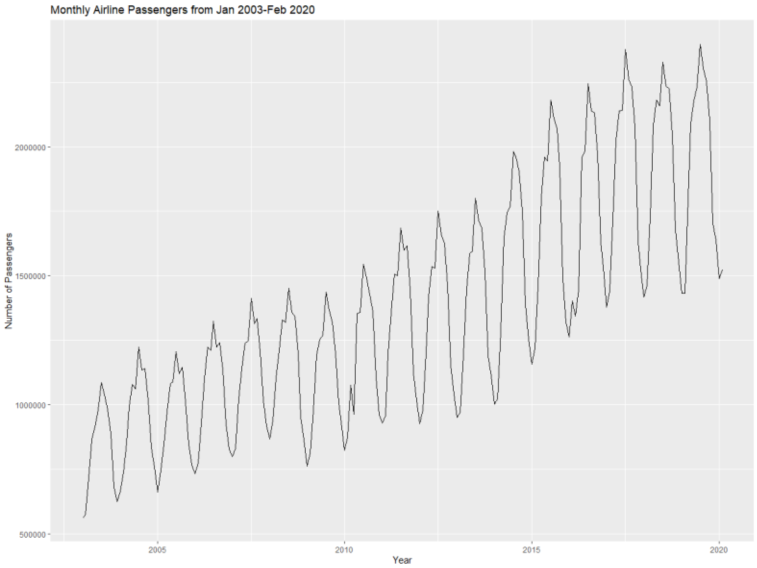
Inferences:

The number of passengers increases over the month which indicates that there is an increasing linear trend.

Months 7 to 10 have higher means and higher variances shows that Many passengers are traveling in these particular months when compared to other months. The reason could be people taking holidays and travel to other EU countries.

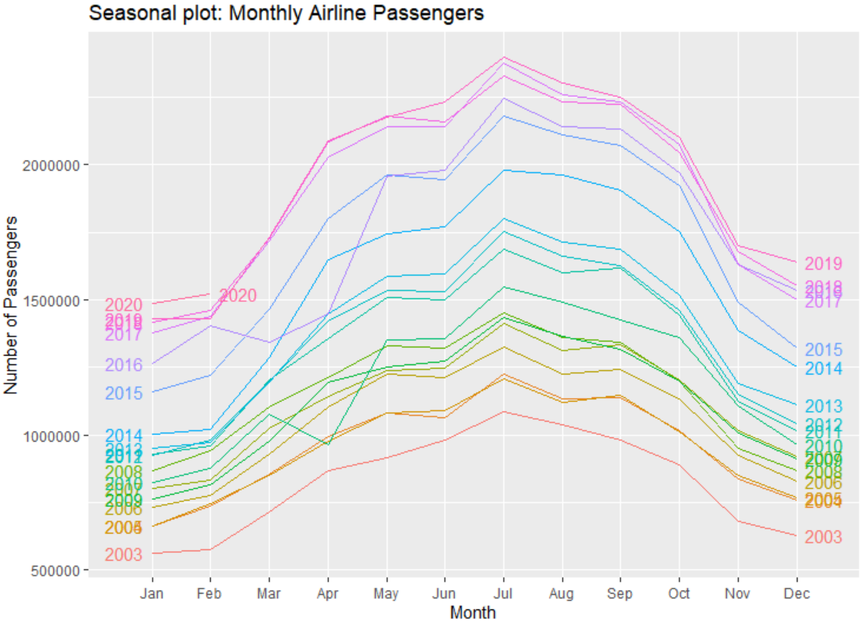
No data cleaning is required as No outliers and missing values are detected.

Time-series plot



Similar to the base function plot, The time series plot is shown only for the time period between Jan 2003-Feb 2020. From the time-series plot, it can be observed that the time series have a strong seasonality component that may be either linear or quadratic. Moreover, the magnitude of the seasonal variation increases as the number of passenger arrivals increases. So there is an upward trend throughout the time series data.

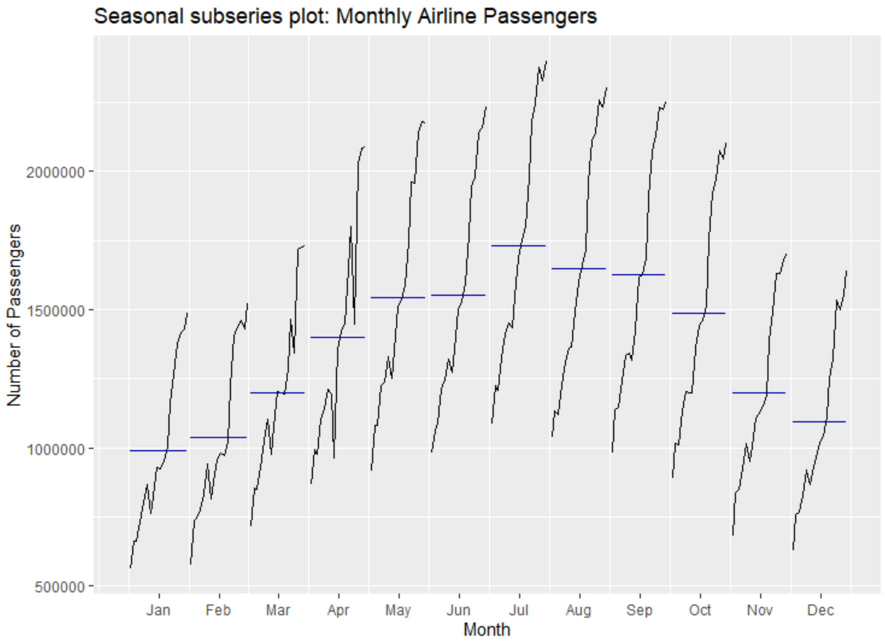
Seasonal plot



This plot allows the underlying seasonal pattern to be seen more clearly and is especially useful in identifying the years in which the pattern changes. It can be observed that there is huge spike in number of passenger travelling during the month of July every year .The reason is these months are considered as popular months for going on holidays for BE residents, especially during summer period.The number of passengers travelling in relatively less during months January and February every year when compared to other months.

Seasonal fluctuations are pervasive in the tourism system due to climactic and socio-structural cycles of both destinations and markets.

Seasonal Subqueries plot

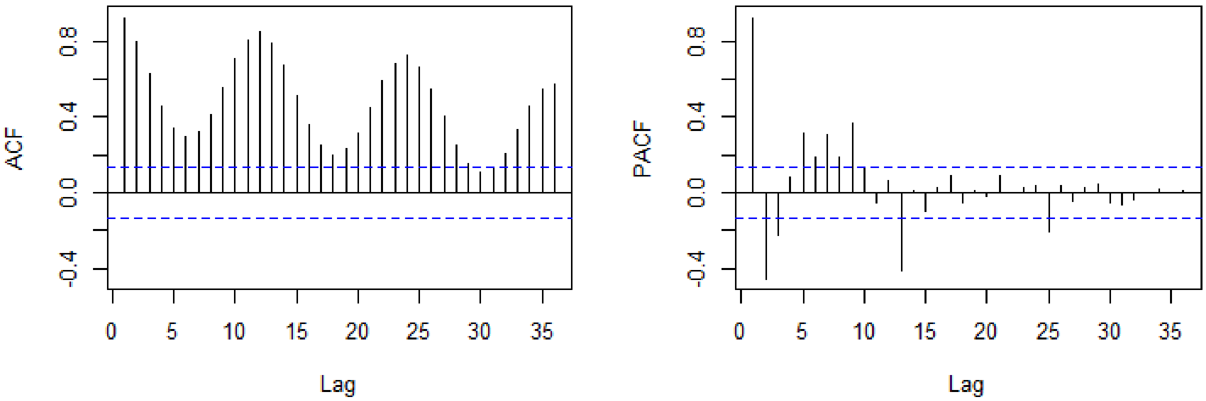


It is an alternative plot that emphasizes the seasonal patterns where the data for each season are collected together in separate mini time plots.

It enables the underlying seasonal pattern to be seen clearly and changes in seasonality over time to be visualized.

It is especially useful in identifying changes within particular seasons.

PACF plot



From the above plot, we can observe that the pattern suggests that the Large spike at lag 1 is followed by a damped wave that alternates between positive and negative correlations.

PACF shows significant autocorrelation at lags 1,2 and amplitude increase over time.

R14 is relatively large and significant.It also indicates some seasonality.

2.Transformation

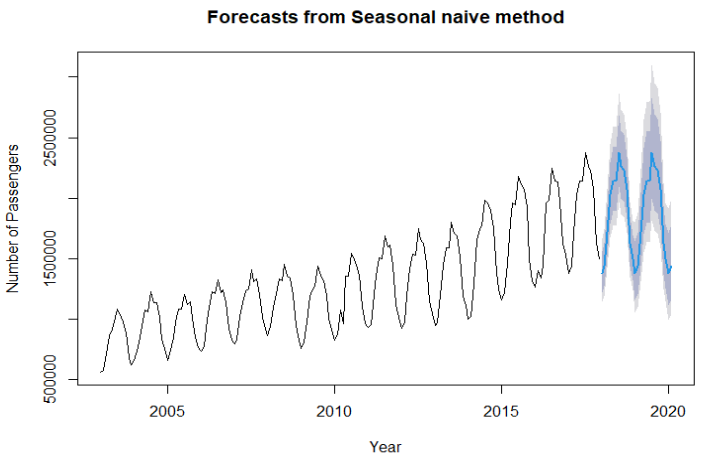
For the given time series data, The Box-Cox lambda value for the time series is around 0.014. So we assign the Optimal Lambda value as zero, λ = 0 -> Log transformation.This can make explanations and interpretations easier.Though their forecasting results are relatively insensitive to the value of λ,it has a large impact on the prediction intervals.There is a need for back transformation for the forecast In case we apply any transformation to the given time series data.So In order to perform automatic transformation and back transformations,we can directly imput the λ = 0 for the forecast models in the subsequent further analysis process.

3.Seasonal naïve forecast

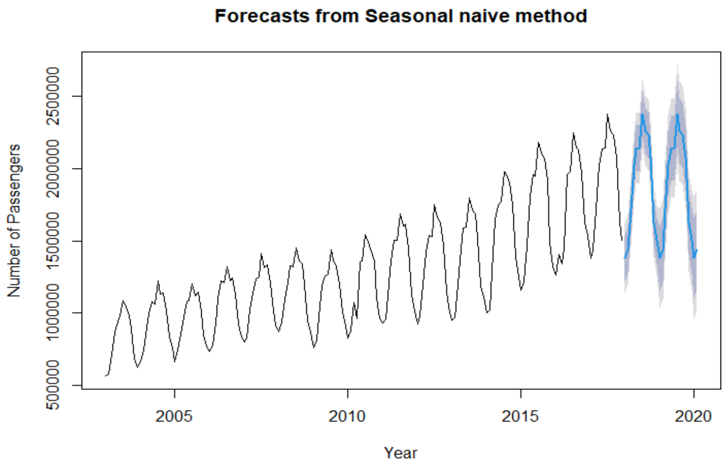
In this method,we set each forecast to be equal to the last observed value from the same season(For ex: the same month of the previous year).

The seasonal naïve forecasts are done in two ways (with and without box cox lambda value). These two methods are compared and necessary inferences are explained.

Seasonal naïve-with λ = 0



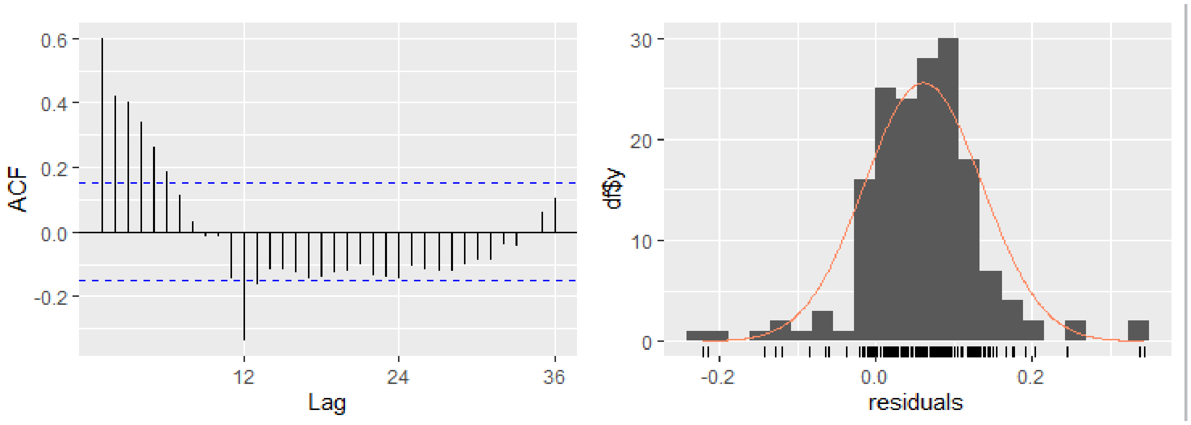
Seasonal naïve-without λ

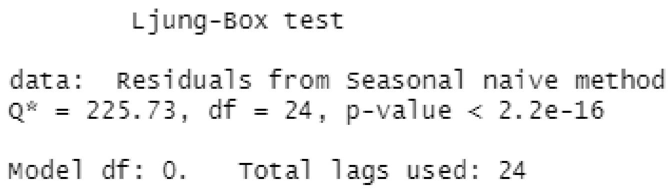


Though both the forecasts look similar, it is slightly different in terms of prediction intervals(gray region). The prediction interval for the seasonal naïve forecast with lambda value is dense when compared to the other method. This is due to the Log transformation performed in the model when imputed with lambda value.

Residual diagnostics:

with λ = 0



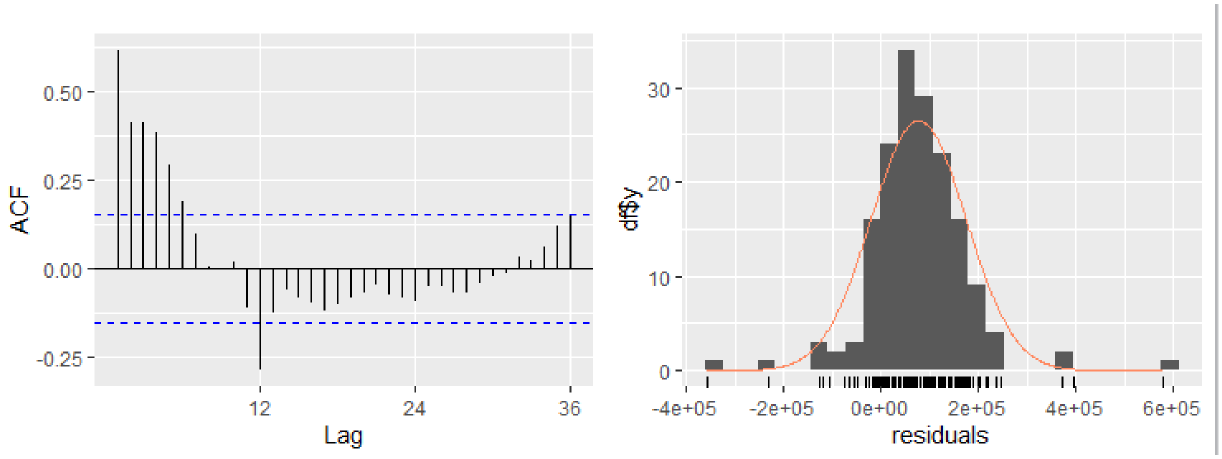


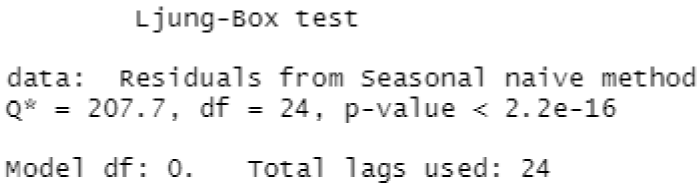
```
        Ljung-Box test

data:  Residuals from Seasonal naive method
Q* = 225.73, df = 24, p-value < 2.2e-16

Model df: 0.   Total lags used: 24
```

without λ





```
        Ljung-Box test

data:  Residuals from Seasonal naive method
Q* = 207.7, df = 24, p-value < 2.2e-16

Model df: 0.   Total lags used: 24
```

From the above plots,we can see that for both forecast PACF plots look similar but the residual plot for the seasonal naïve forecast model without transformation has a better equally distributed residual curve.

The P-value for both the models is 2.2e-16 (significantly very low) i.e model need a meaningful addition and it establishes a pattern.

Forecast Accuracy:

Mean absolute scale error(MASE) is the one important measure of forecast accuracy to consider for the given time series data.So, the lower the MASE value the better the model. Both the forecast models have the same MASE value(0.4753).

Hence from all the above inferences,we can conclude that for seasonal naïve forecast,the model without lambda value(log transformation) is better.

4.STL decomposition forecast method:

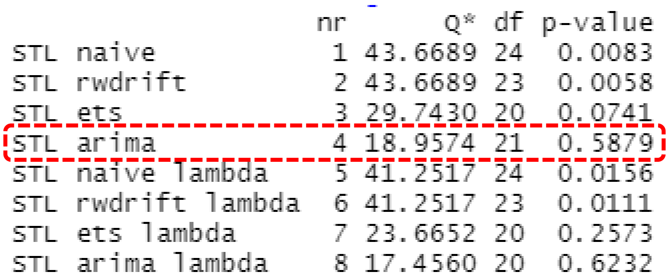
Seasonal and Trend decomposition using Loess (STL) method is used for decomposing the time series data.It is best suited for seasonality data.

Decompositions between additive and multiplicative can be obtained using a Box-Cox transformation of the data. For the given time series data, A value of λ=0 corresponds to the multiplicative decomposition.

Usage of stlf() function seasonally adjust the data.The methods naïve,rwdrift,ets,arima are fitted with and without lambda values.

Residual Diagnostics:

The residual diagnostics for the seasonally adjusted data for all the STL methods are shown below:

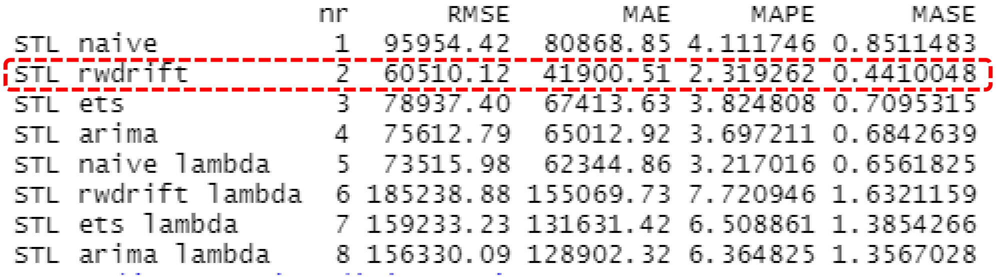


```
                     nr       Q* df  p-value
STL naive             1 43.6689 24   0.0083
STL rwdrift           2 43.6689 23   0.0058
STL ets               3 29.7430 20   0.0741
STL arima             4 18.9574 21   0.5879
STL naive lambda      5 41.2517 24   0.0156
STL rwdrift lambda    6 41.2517 23   0.0111
STL ets lambda        7 23.6652 20   0.2573
STL arima lambda      8 17.4560 20   0.6232
```

From the above model summary, we can see that the model STL Arima has a high p-value (0.5879) when compared to other models. the model has a better learning rate from the time serie.In terms of p-value, this is chosen as a better model.

Forecast accuracy:

The measures of the forecast summary for all the STL models are shown below:

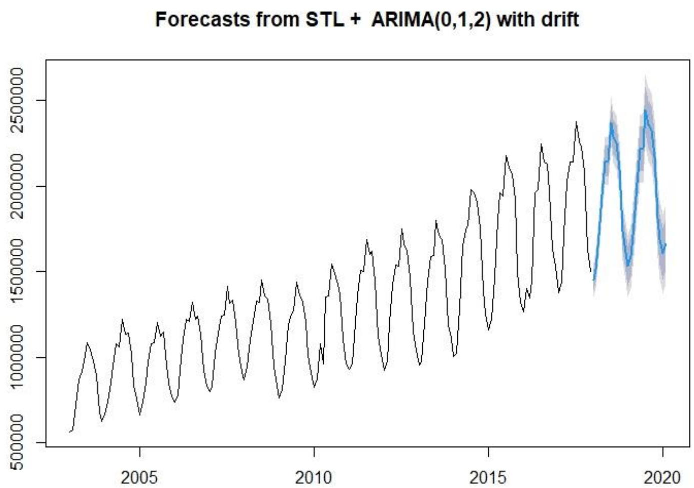


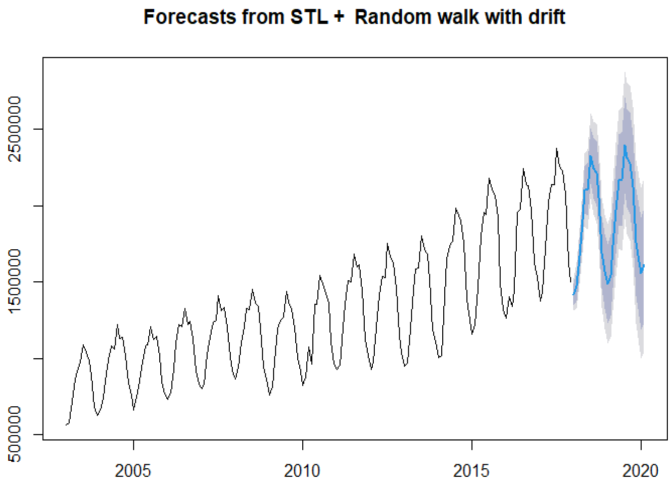
```
                     nr       RMSE       MAE     MAPE      MASE
STL naive             1   95954.42  80868.85 4.111746 0.8511483
STL rwdrift           2   60510.12  41900.51 2.319262 0.4410048
STL ets               3   78937.40  67413.63 3.824808 0.7095315
STL arima             4   75612.79  65012.92 3.697211 0.6842639
STL naive lambda      5   73515.98  62344.86 3.217016 0.6561825
STL rwdrift lambda    6  185238.88 155069.73 7.720946 1.6321159
STL ets lambda        7  159233.23 131631.42 6.508861 1.3854266
STL arima lambda      8  156330.09 128902.32 6.364825 1.3567028
```

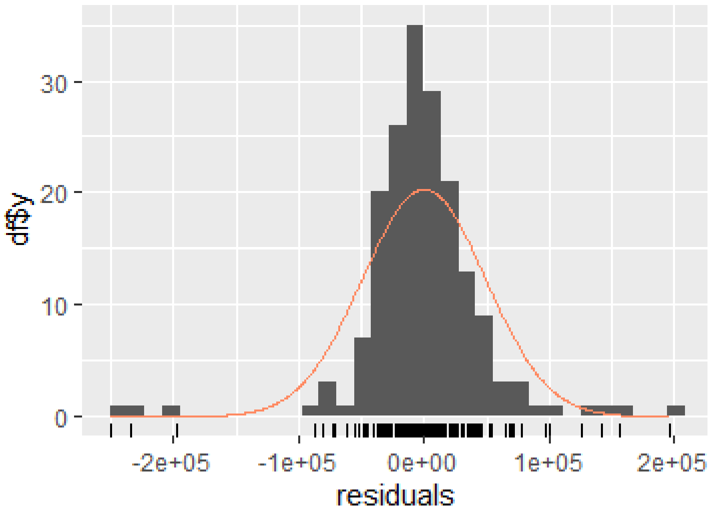
From the above table values, we can see that the model STL rwdrift has a low MASE-value (0.441) when compared to other models. In terms of MASE measure, this is chosen as a better model.

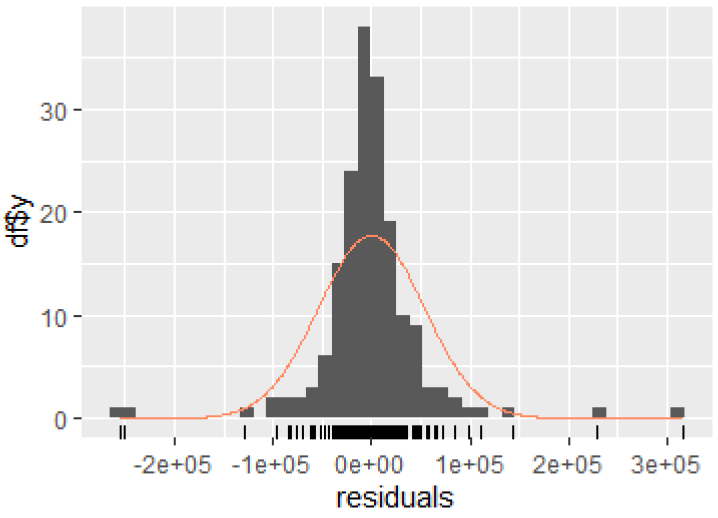
Final inferences:

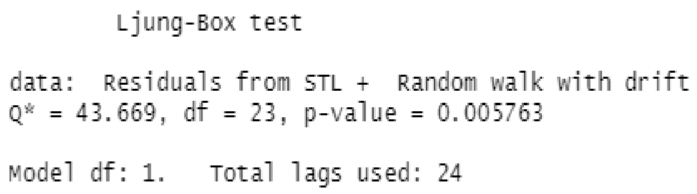
Forecasting is done on 2 selected models based on p-value and MASE values.their respective Forecasting Plots are shown below:


Forecasts from STL + ARIMA(0,1,2) with drift


Forecasts from STL + Random walk with drift

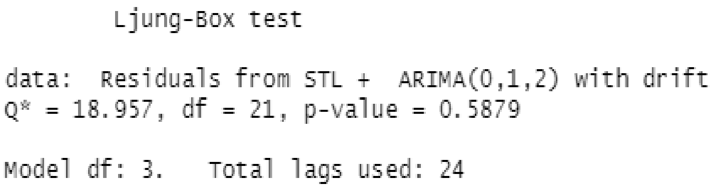





```
        Ljung-Box test

data:  Residuals from STL +  Random walk with drift
Q* = 43.669, df = 23, p-value = 0.005763

Model df: 1.   Total lags used: 24
```



```
        Ljung-Box test

data:  Residuals from STL +  ARIMA(0,1,2) with drift
Q* = 18.957, df = 21, p-value = 0.5879

Model df: 3.   Total lags used: 24
```

From the above plots,

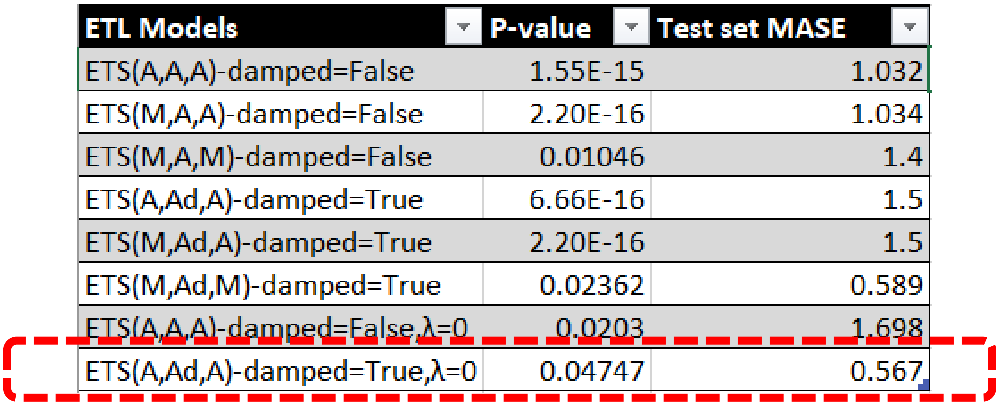
When compared to these two models , we can see that the STL Arima model has a better prediction interval when compared to STL random drift model which has a denser prediction interval.For residual plots,STL arima has more centered values around the zero.Additionally, it has significant p-value of the ljung Box test (0.5879).Hence, the STL arima model is selected as the best STL model.

5.ETS Models

It is one type of exponential smoothing method where forecasts produced using these methods are weighted averages of past observations. with the weights decaying exponentially as the observations get older.

For the given time series data, different models with and without damped trends are imputed and checked.The damped trend is used to flatten the curve over time.The lambda value (log transformation) imputation is done For two models ETS (A,Ad.A) with damped and without damped trends.

The p-values obtained from the Ljung Box test and Test set MASE values are shown below:

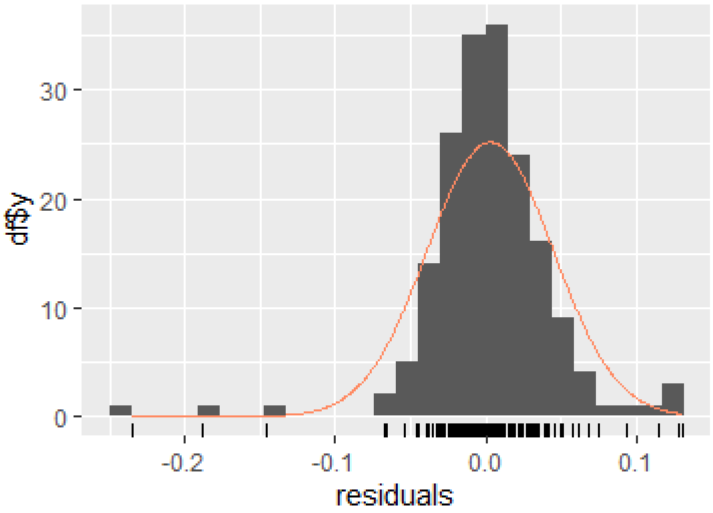


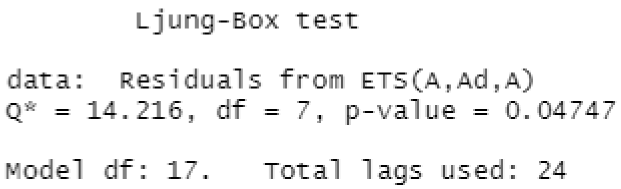
ETL Models	P-value	Test set MASE
ETS(A,A,A)-damped=False	1.55E-15	1.032
ETS(M,A,A)-damped=False	2.20E-16	1.034
ETS(M,A,M)-damped=False	0.01046	1.4
ETS(A,Ad,A)-damped=True	6.66E-16	1.5
ETS(M,Ad,A)-damped=True	2.20E-16	1.5
ETS(M,Ad,M)-damped=True	0.02362	0.589
ETS(A,A,A)-damped=False,λ=0	0.0203	1.698
ETS(A,Ad,A)-damped=True,λ=0	0.04747	0.567

Inferences:

Based on the performance of the models in terms of Residual Diagnostics and forecast accuracy,we can clearly see that Model - ETS(A,Ad,A) is better when compared to other ETS models.It has a low MASE value of 0.567 and a High significant p-value of 0.047 close to 0.05, which is considered as significant. Hence ETS(A,Ad,A) is selected as the best model.

Selected model-ETS(A,Ad,A)





```
                Ljung-Box test

data:  Residuals from ETS(A,Ad,A)
Q* = 14.216, df = 7, p-value = 0.04747

Model df: 17.   Total lags used: 24
```

Smoothing parameters:

Description:

Alpha = parameter for controlling the rate at which the influence of the observations at prior time steps decays exponentially.

Beta = parameter for change in trend over time.

Gamma = Parameter for Change in Seasonality Overtime

Phi = Damping Coefficient of the model

ETS selected Model smoothing parameter values:

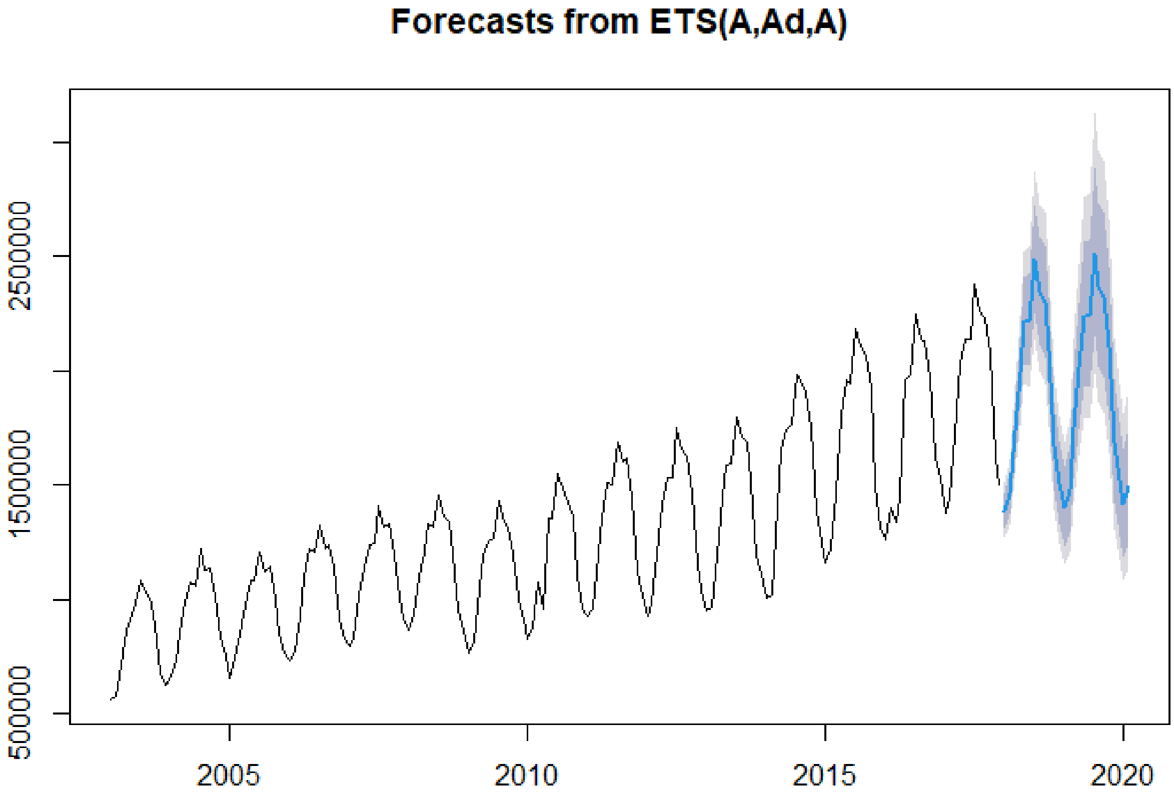
alpha = 0.5585, this value suggests that almost 50% of the past data is being used to forecast the future.

beta = 0.0013, The small value of Beta explains that the change in the Trend of the Data is very minimal.

gamma = 1e-04, The Small Value of Gamma explains that there is no change in the Seasonality of the data.

phi = 0.98

The forecasts for the selected model is shown below:



ARIMA:

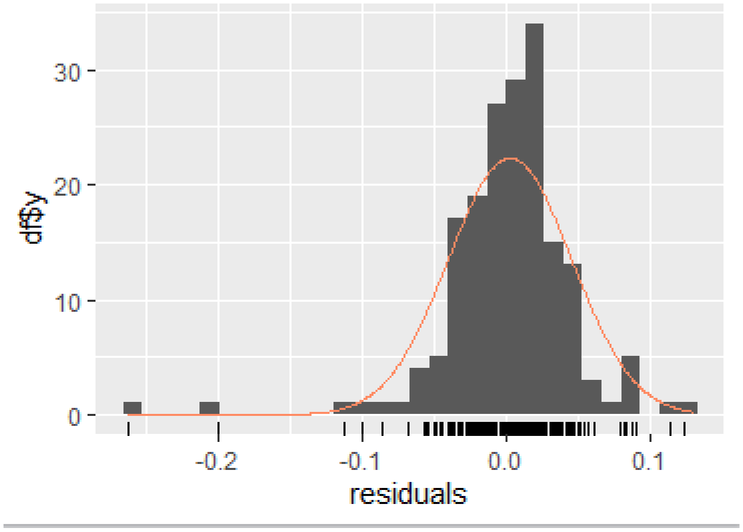
Initially, the auto Arima model gives a relatively high MASE value-1.48.In order to get the best MASE values, the different parameters are imputed inside the model.

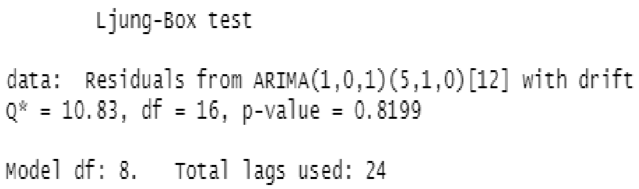
Best Arima model:

The best ARIMA model parameters are as follows:

stepwise = TRUE, approximation = FALSE,lambda = l,max.p = 10,max.q = 10,max.P = 10,max.Q =10, max.order = 10,max.d = 10, max.D = 10

Using the above parameters, the model forecast accuracy and residual diagnostics are as follows:



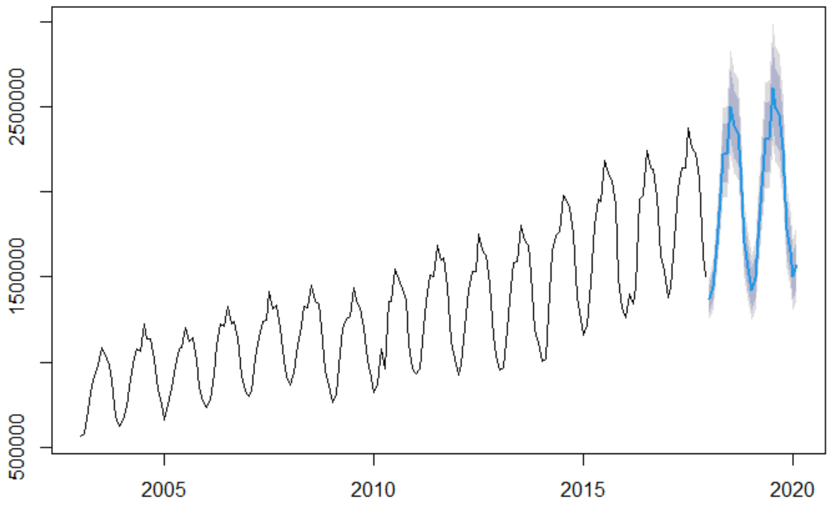


```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,1)(5,1,0)[12] with drift
Q* = 10.83, df = 16, p-value = 0.8199

Model df: 8.   Total lags used: 24
```

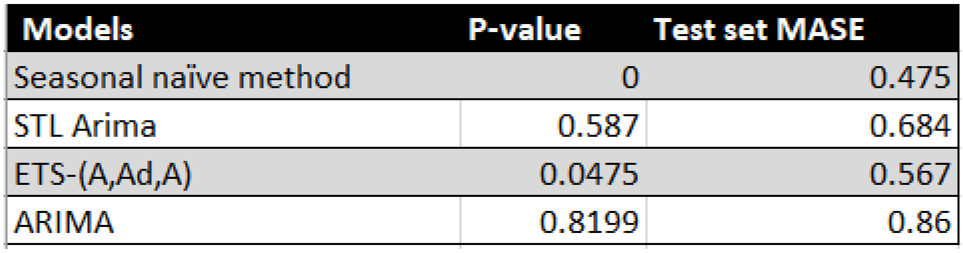
Forecasts from ARIMA(1,0,1)(5,1,0)[12] with drift



The p-value of the selected arima model is 0.819 and MASE is reduced to 0.86 after the selecting the best parameters for the model.

7. Models-Comparison

The best models from the seasonal naïve, STL, ETS, and ARIMA are selected and compared in terms of residual diagnostics and forecast accuracy.



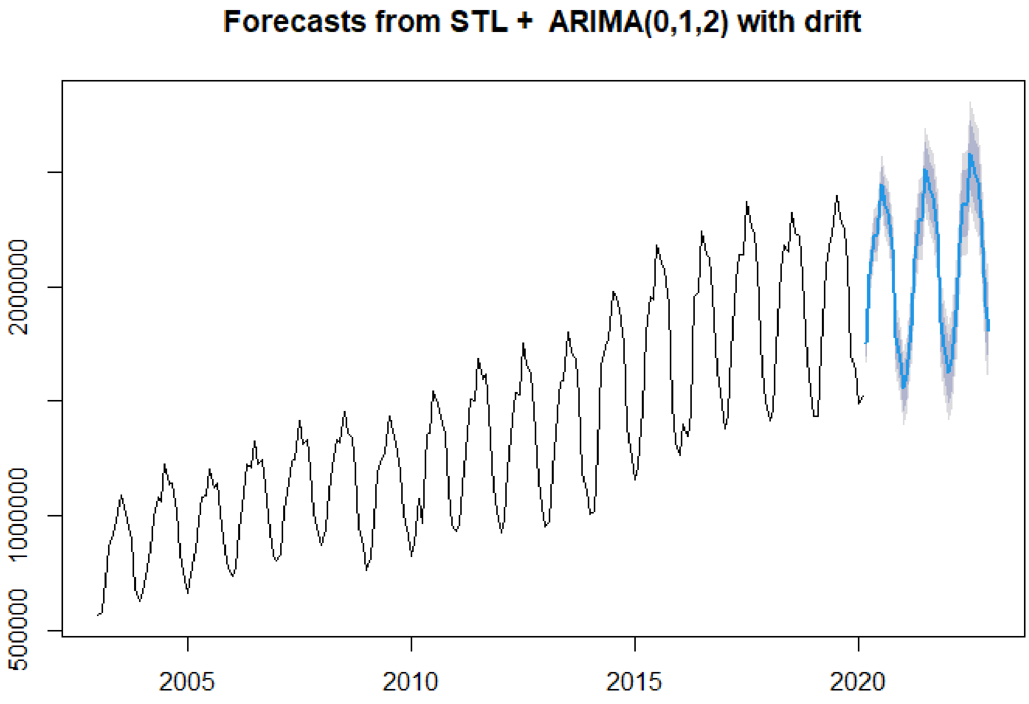
Models	P-value	Test set MASE
Seasonal naïve method	0	0.475
STL Arima	0.587	0.684
ETS-(A,Ad,A)	0.0475	0.567
ARIMA	0.8199	0.86

From the above model summary table, Though the ARIMA model has a significant High p-value when compared to the other models,It still has a high test MASE value(Lower the MASE better the model).So Keeping both measures in mind,**STL Arima** with a significant p-value of 0.587 and Test set MASE of 0.684 is selected as the best model.

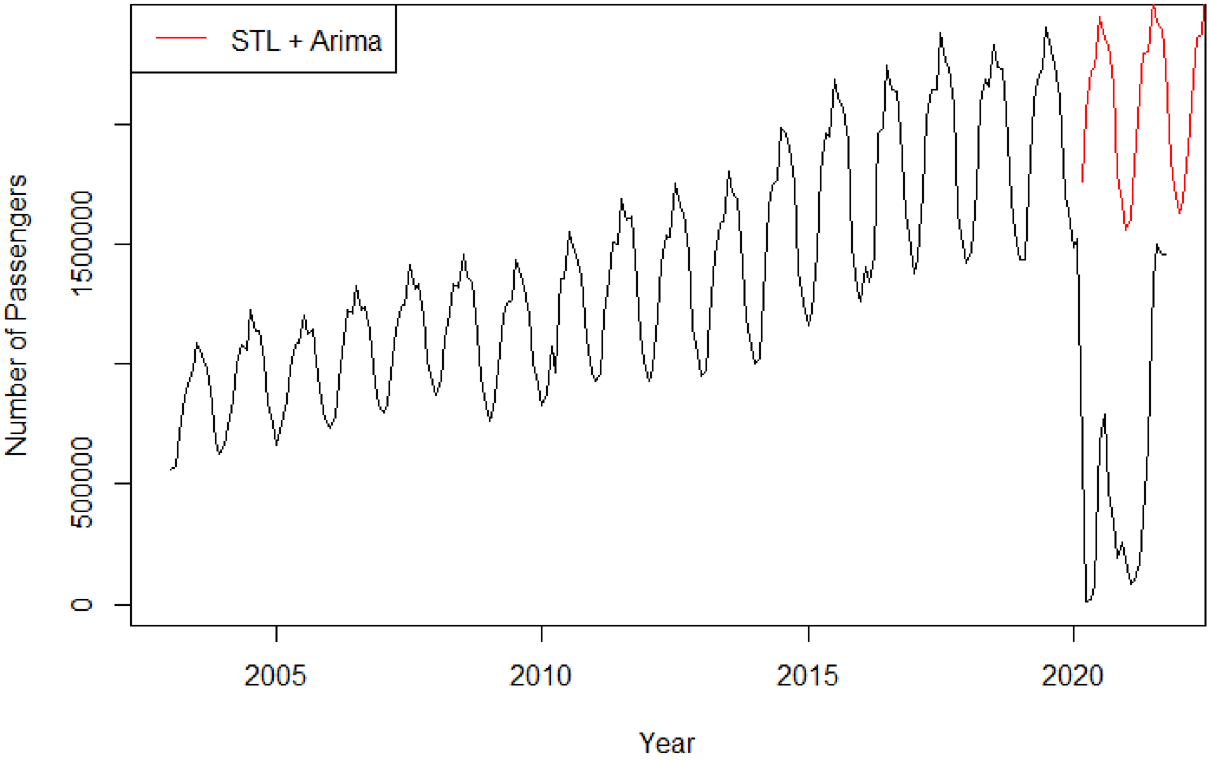
8. Forecast using the selected model

STL Arima

Using STL-Arima model, the forecast up to December 2022 is done and the respective plot is shown below:



From the above plot,we can observe that there is an increasing trend over time till December 2022.the prediction interval is also relatively less which shows that it is the best for predicting the number of passengers in future.



Final inferences:

From the plot, we can observe that there is a huge declining trend in the number of passengers traveling between Belgium and other EU countries after 2020 February due to the Covid-19 breakdown. This downward trend is due to many reasons. Flight bans, flight booking cancellations, closing country borders, Imposing travel restrictions, and quarantine requirements are the instances that made traveling very difficult for the passengers, and hence there is a huge number decrease in passengers traveling from Belgium to other EU countries. Another drop in the number of passengers during Feb-2021 is maybe a spike of Covid cases and the Government imposing any new lockdown measures. The current numbers in July 2021 show that things are getting normal and airlines are started functioning like before.

Based on the STL Arima model forecasted plot, The predictions suggest that time series have a strong seasonality component and good trend over time. The magnitude of the seasonal variation increases as the number of passenger arrivals increases. So there is an upward trend throughout the time series data. So if Covid doesn't happened, The airlines would have functioned well and many passengers would have traveled from Belgium to other Eu countries.

EXERCISE -2

Data description:

Unemployment data

This dataset is from the Current Population Survey (Household Survey) conducted by the Bureau of Labor Statistics. Labor force flows show the movements that underlie the net over-the-month changes in employment, unemployment, or not in the labor force.

This dataset contains the number of monthly air passengers traveling from reporting country Belgium(BE) and EU partner country during the time period January 2003 until July 2021.

The data set contains 2 columns

Date: Information of month and year of flow from employed to unemployed

LNU07100000 : Information about the number of unemployed people.

Objective :

The objective is to Analyze two different sets of Time Series Data through Data Description and also split the data into a Train Set & Test Set in order to be able to apply different forecasting techniques. Moreover, we try to understand the choose the best model and try to interpret the results of our forecast using the parameters of the model.

Note:

The time-series data is split into the training set and test set.

Training set: February 1990-December 2017

Test set: January 2018 – March 2020

The values after March 2020 are not considered in the time-series test set because impact the efficiency of the forecasting models and it will distort the assessment for forecast accuracy. So we can keep them for later reference.

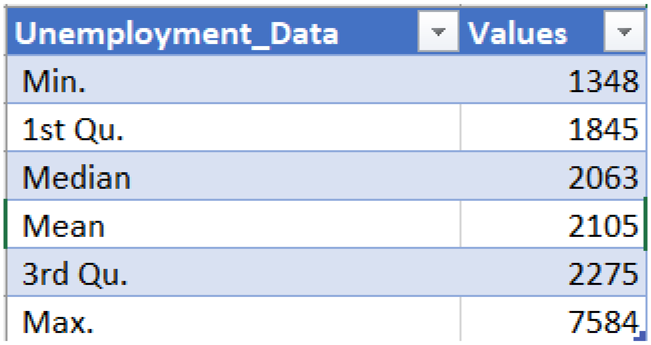
Explanatory Data Analysis:

The given data set is converted into time series data using the ts () function. Let us explore the raw time-series data and review the data with summary statistics and plots in R.

The frequency of the time series is 12 i.e data with monthly seasonality.

Data summary :

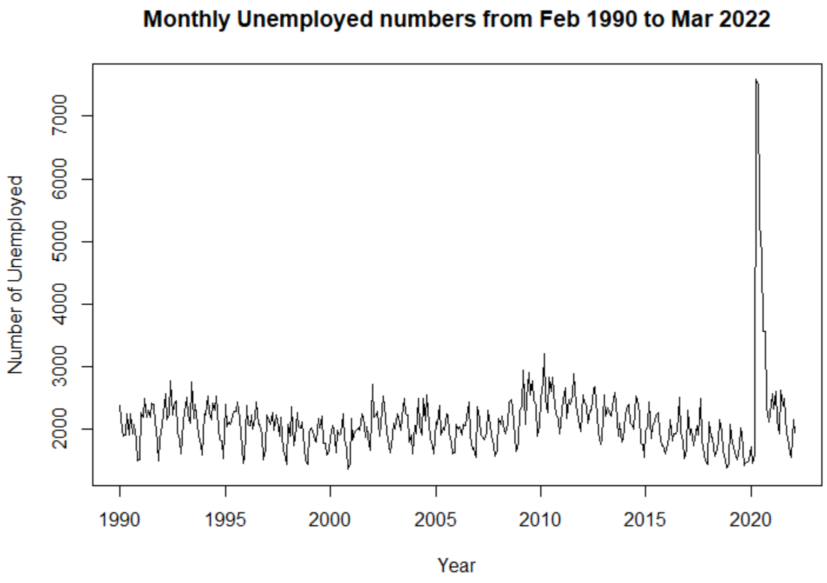
The summary table of the data suggests that the Minimum number of unemployed is 1348 and the Maximum number of passengers traveled is 7584. The average number of passengers traveled is 2105.

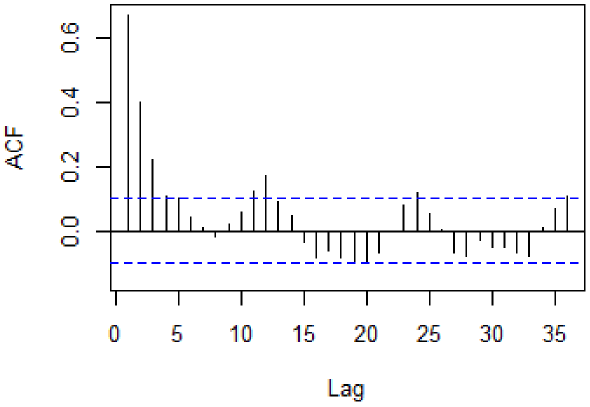


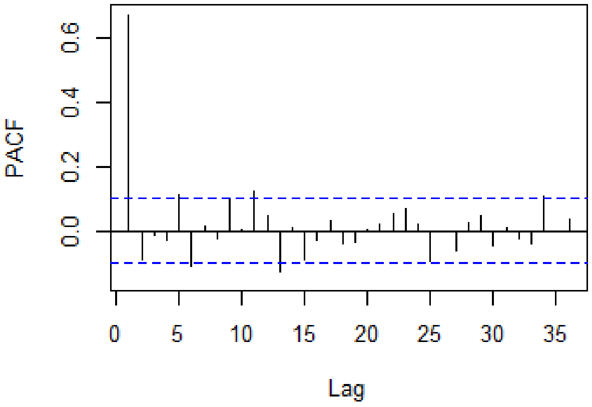
Unemployment_Data	Values
Min.	1348
1st Qu.	1845
Median	2063
Mean	2105
3rd Qu.	2275
Max.	7584

Visualizing Seasonality and Trend:

1. **Base plot function**

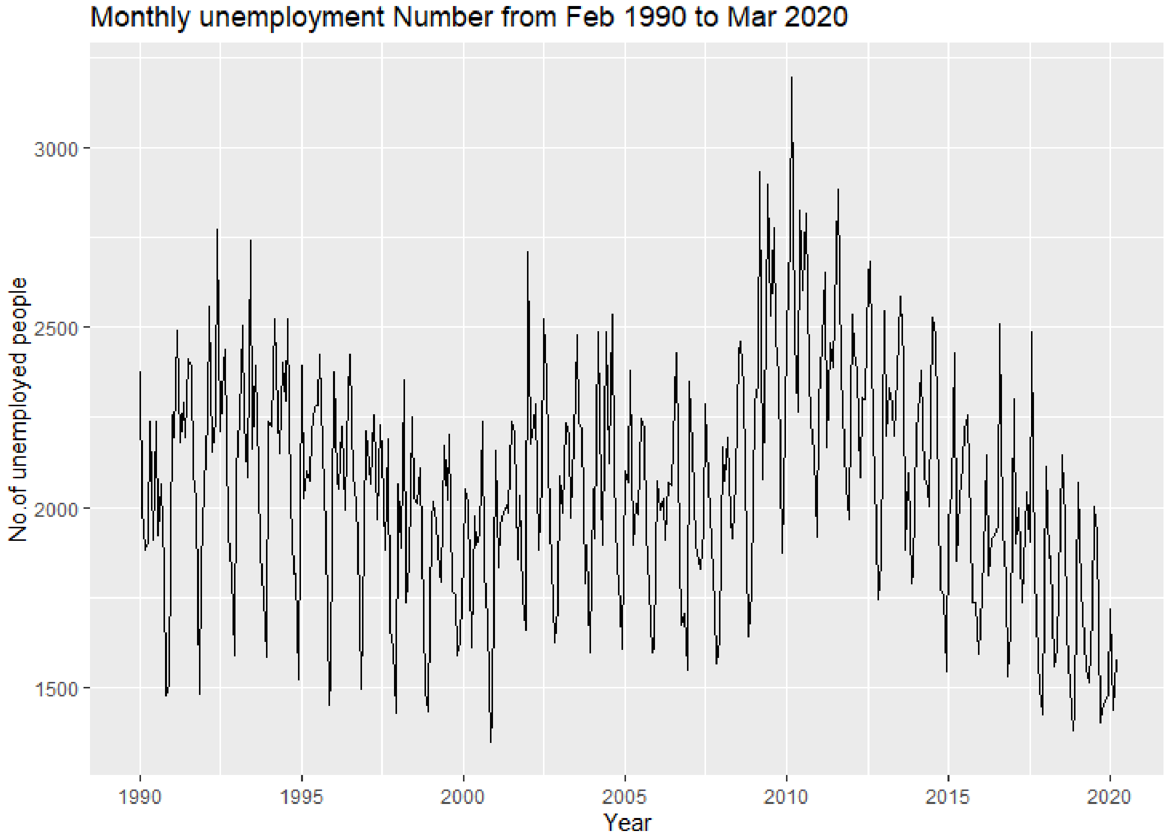

Monthly Unemployed numbers from Feb 1990 to Mar 2022





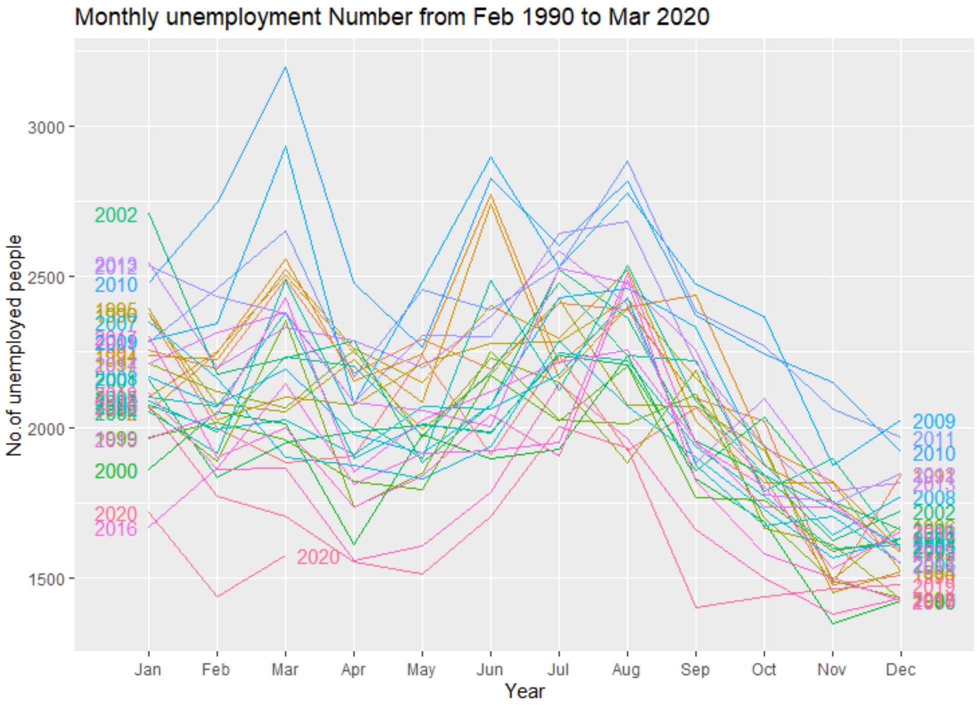
From the above plots of the raw data, it can be observed that the time series have a seasonality component. There is no visible trend throughout the time series data till 2020 March. After Mar 2020, there is a huge spike in the number of unemployed due to the covid-19 outbreak.

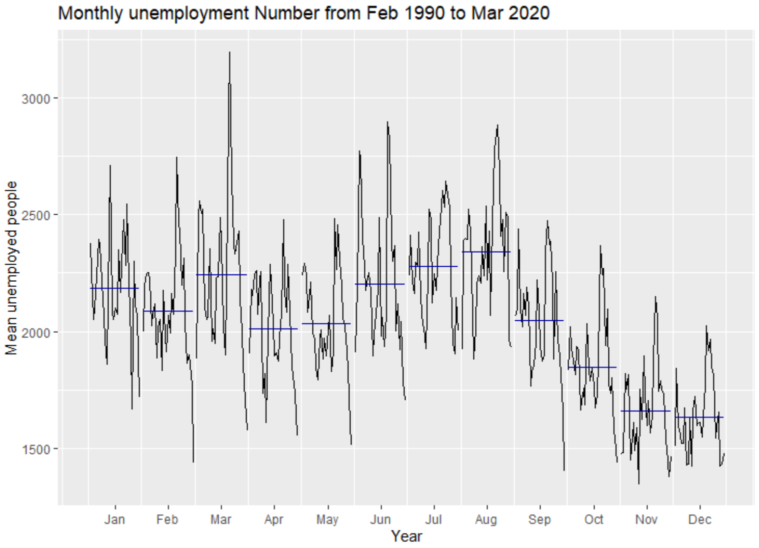
Time-series plot



Similar to the base function plot, The time series plot is shown only for the time period between Feb 1990-Mar 2020. From the time-series plot, it can be observed that the time series have a seasonality component. There is a decreasing trend of the time series data after the year 2010.So the final model alpha parameter has to be cross-checked, which influences the observation of the prior time. So correct past data has to be selected to make predictions.

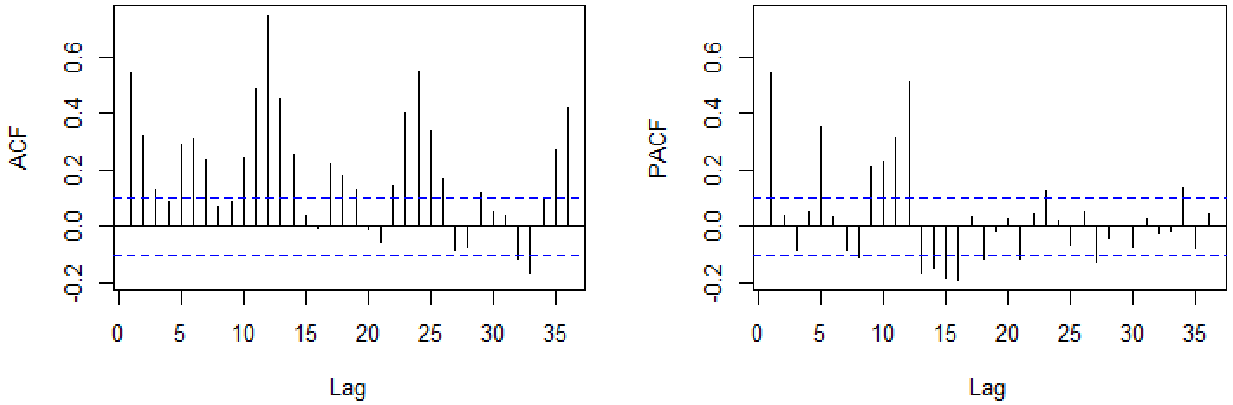
Seasonal & Seasonal Subqueries plot





The above plots allow the underlying seasonal pattern to be seen more clearly and is especially useful in identifying the years in which the pattern changes. It can be observed that the average unemployed is high during the month of August every year and relatively less number of people unemployed during the month of December every year.

PACF plot



From the above plot, we can observe that the pattern suggests that the Large spike at lag 1 and lag 12 is followed by a damped wave that alternates between positive and negative correlations.it suggests that there is some seasonality component that exists in the data.

2.Transformation

For the given time series data, The Box-Cox lambda value for the time series is -0.042. So we assign the Optimal Lambda value as zero, λ = 0 -> Log transformation.This can make explanations and interpretations easier.Though their forecasting results are relatively insensitive to the value of λ,it has a large impact on the prediction intervals.There is a need for back transformation for the forecast In case we apply any transformation to the given time series data.So In order to perform automatic transformation and back transformations,we can directly impute the λ = 0 for the forecast models in the subsequent further analysis process.

Models:

1.STL decomposition forecast method:

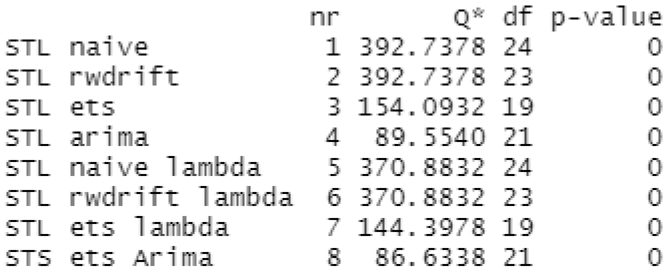
Seasonal and Trend decomposition using Loess (STL) method is used for decomposing the time series data.It is best suited for seasonality data.

Decompositions between additive and multiplicative can be obtained using a Box-Cox transformation of the data. For the given time series data, A value of λ=0 corresponds to the multiplicative decomposition.

Usage of stlf() function seasonally adjust the data.The methods naïve,rwdrift,ets,arima are fitted with and without lambda values.

Residual Diagnostics:

The residual diagnostics for the seasonally adjusted data for all the STL methods are shown below:

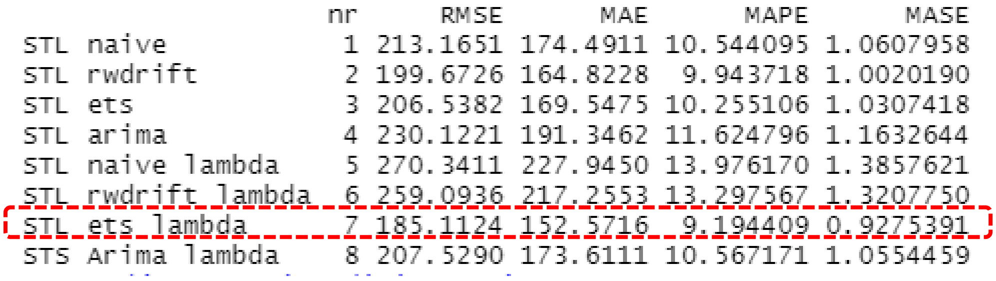


```
                    nr        Q* df p-value
STL naive            1 392.7378 24       0
STL rwdrift          2 392.7378 23       0
STL ets              3 154.0932 19       0
STL arima            4  89.5540 21       0
STL naive lambda     5 370.8832 24       0
STL rwdrift lambda   6 370.8832 23       0
STL ets lambda       7 144.3978 19       0
STS ets Arima        8  86.6338 21       0
```

From the above model summary, we can see that all models have a p-value of zero.So, the model should be selected based on the MASE measure.

Forecast accuracy:

The measures of the forecast summary for all the STL models are shown below:

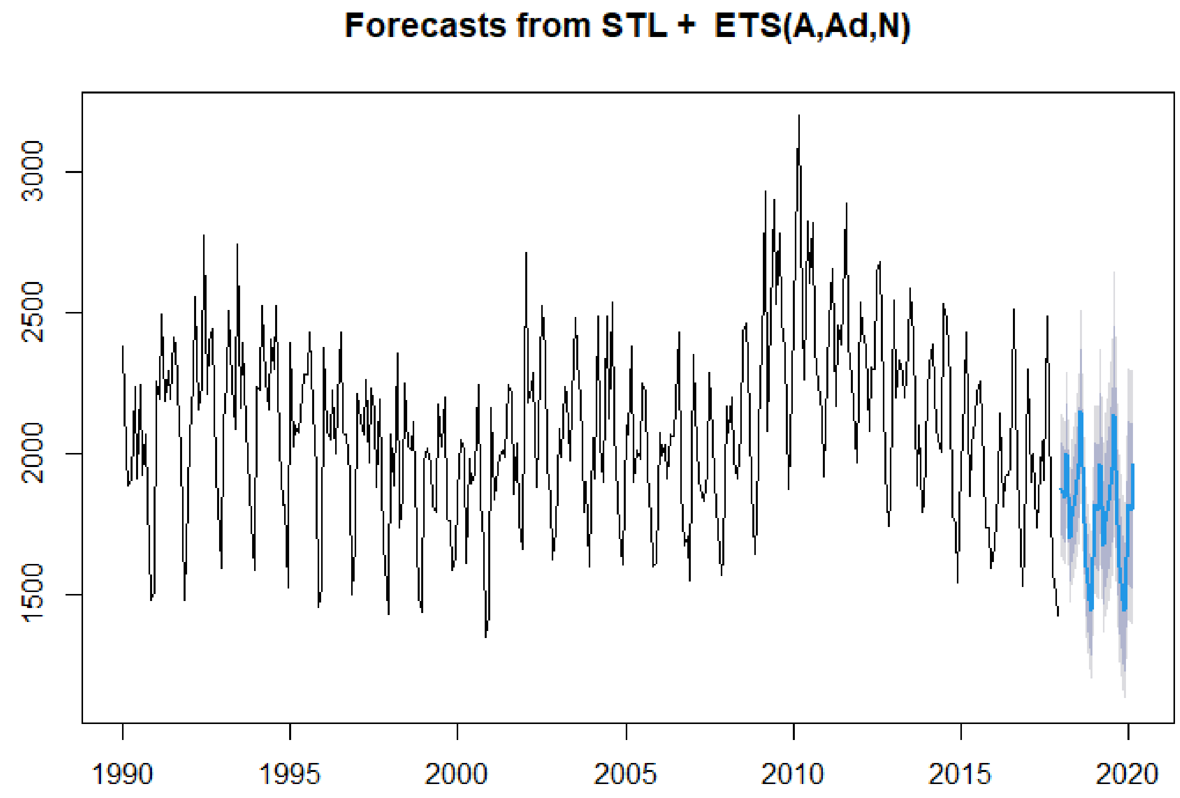


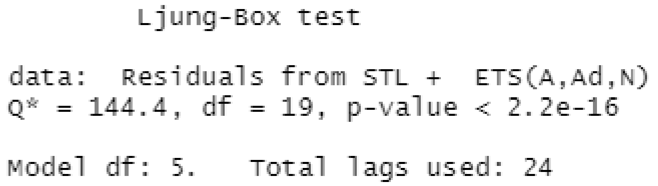
```
                   nr     RMSE      MAE      MAPE      MASE
STL naive           1 213.1651 174.4911 10.544095 1.0607958
STL rwdrift         2 199.6726 164.8228  9.943718 1.0020190
STL ets             3 206.5382 169.5475 10.255106 1.0307418
STL arima           4 230.1221 191.3462 11.624796 1.1632644
STL naive lambda    5 270.3411 227.9450 13.976170 1.3857621
STL rwdrift lambda  6 259.0936 217.2553 13.297567 1.3207750
STL ets lambda      7 185.1124 152.5716  9.194409 0.9275391
STS Arima lambda    8 207.5290 173.6111 10.567171 1.0554459
```

From the above table values, we can see that the model STL Ets has a low MASE-value (0.92) when compared to other models. In terms of MASE measure, this is chosen as a better model.

Final inferences:

Forecasting is done on selected model-STL ets based on MASE values.their respective Forecasting Plots and residual plots are shown below:

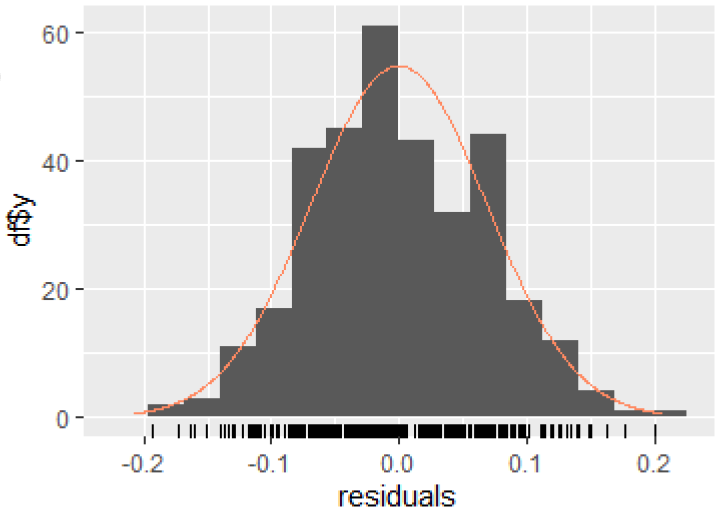




```
        Ljung-Box test

data:  Residuals from STL +  ETS(A,Ad,N)
Q* = 144.4, df = 19, p-value < 2.2e-16

Model df: 5.   Total lags used: 24
```



From the above plots,

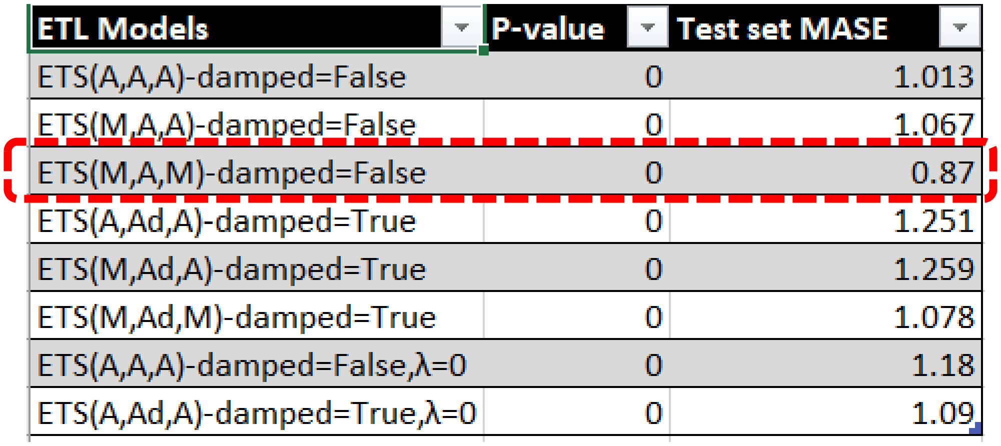
we can see that the STL ETS model has a better prediction interval when compared to STL random drift model which has a denser prediction interval.For residual plots,STL ETS has more centered values around the zero. Hence, the STL arima model is selected as the best STL model.

ETS Models

It is one type of exponential smoothing method where forecasts produced using these methods are weighted averages of past observations. with the weights decaying exponentially as the observations get older.

For the given time series data, different models with and without damped trends are imputed and checked.The damped trend is used to flatten the curve over time.The lambda value (log transformation) imputation is done For two models ETS (A,Ad.A) with damped and without damped trends.

The p-values obtained from the Ljung Box test and Test set MASE values are shown below:

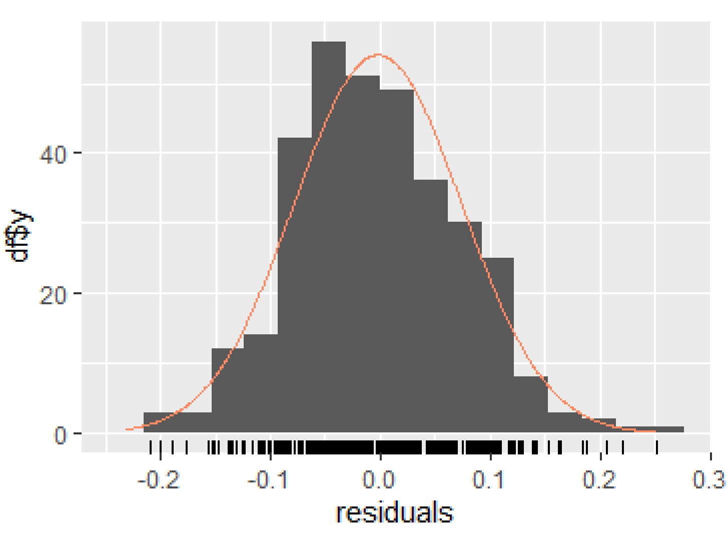


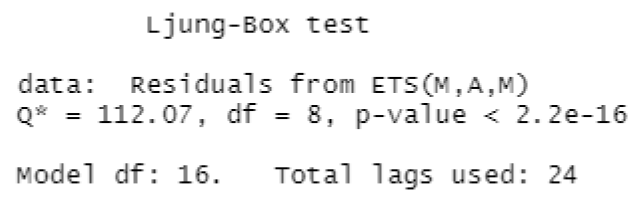
ETL Models	P-value	Test set MASE
ETS(A,A,A)-damped=False	0	1.013
ETS(M,A,A)-damped=False	0	1.067
ETS(M,A,M)-damped=False	0	0.87
ETS(A,Ad,A)-damped=True	0	1.251
ETS(M,Ad,A)-damped=True	0	1.259
ETS(M,Ad,M)-damped=True	0	1.078
ETS(A,A,A)-damped=False,λ=0	0	1.18
ETS(A,Ad,A)-damped=True,λ=0	0	1.09

Inferences:

Based on the performance of the models in terms of Residual Diagnostics and forecast accuracy,we can clearly see that Model - ETS(M,A,M) is better when compared to other ETS models.It has a low MASE value of 0.87. Hence ETS(M,A,M) is selected as the best model.

Selected model-ETS(M,A,M)





```
        Ljung-Box test

data:  Residuals from ETS(M,A,M)
Q* = 112.07, df = 8, p-value < 2.2e-16

Model df: 16.   Total lags used: 24
```

Smoothing parameters:

Description:

Alpha = parameter for controlling the rate at which the influence of the observations at prior time steps decays exponentially.

Beta = parameter for change in trend over time.

Gamma = Parameter for Change in Seasonality Overtime

Phi = Damping Coefficient of the model

ETS selected Model smoothing parameter values:

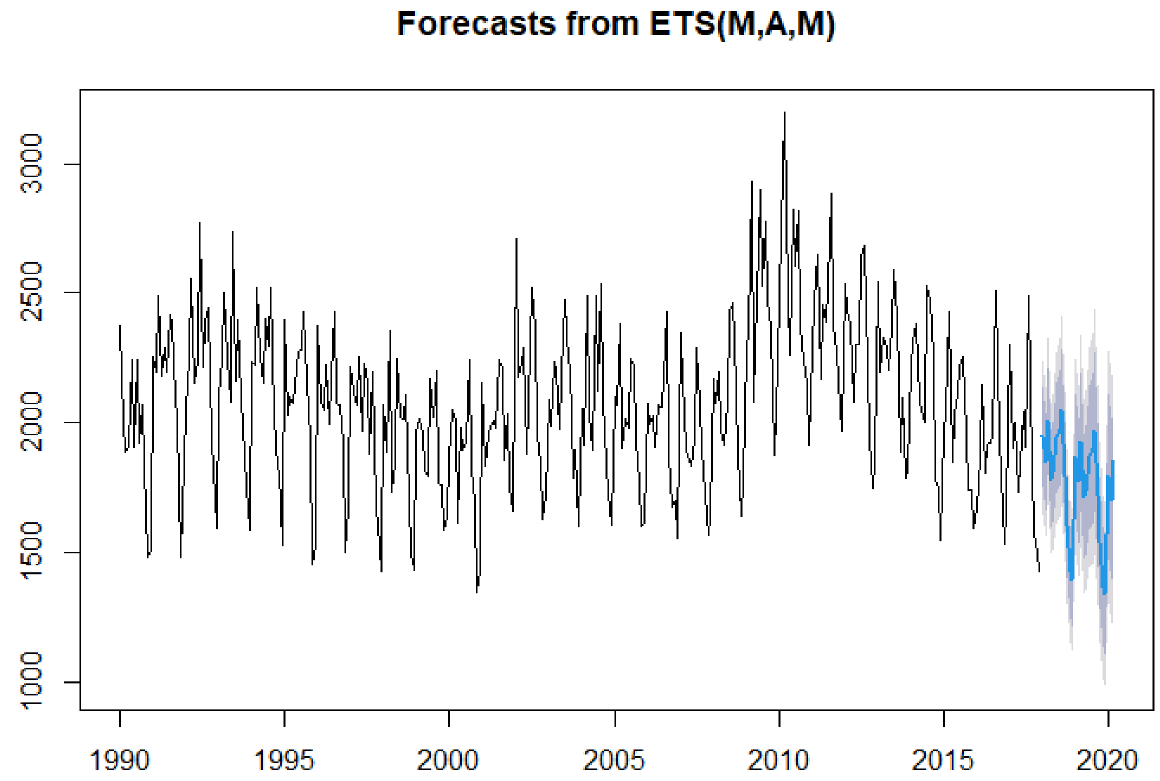
alpha = 0.1973, this value suggests that almost 20% of the past data is being used to forecast the future.

beta = 0.0068, The small value of Beta explains that the change in the Trend of the Data is very minimal.

gamma = 1e-04, The Small Value of Gamma explains that there is no change in the Seasonality of the data.

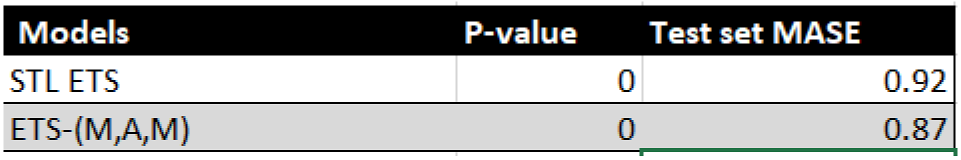
phi = 0.98

The forecasts for the selected model is shown below:



Models-Comparison

The best models from the STL and ETS are selected and compared in terms of residual diagnostics and forecast accuracy.



Models	P-value	Test set MASE
STL ETS	0	0.92
ETS-(M,A,M)	0	0.87

From the above model summary table,The STL ETS (A, Ad, N) Test set MASE of 0.87 lesser than the ETS (M,A,M) model .So STL ETS (A,Ad,N) is selected as the best model.

Forecast using the selected model

STL ETS (A, Ad, N)

parameters:

Description:

Alpha = parameter for controlling the rate at which the influence of the observations at prior time steps decays exponentially.

Beta = parameter for change in trend over time.

Gamma = Parameter for Change in Seasonality Overtime

Phi = Damping Coefficient of the model

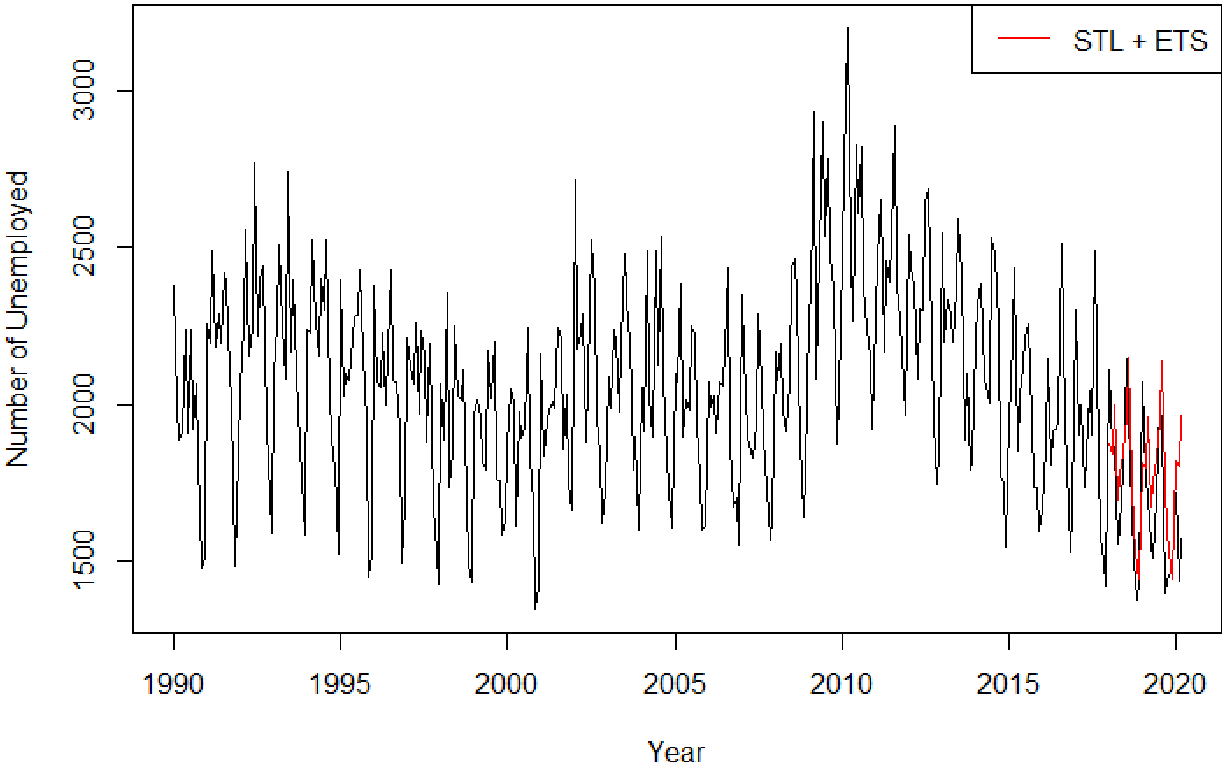
ETS selected Model smoothing parameter values:

alpha = 0.0776, this value suggests that almost 7% of the past data is being used to forecast the future.

beta = 0.067, The small value of Beta explains that the change in the Trend of the Data is very minimal.

phi = 0.8

Using STL ETS (A, Ad, N) model, the forecast is done and the respective plot is shown below:



From the above plot,we can observe the comparison between the test data and pmodel prediction. the red line indicates the predictions by STL + ETL model.

References:

https://ec.europa.eu/eurostat/cache/metadata/en/avia_pa_esms.htm

http://rstudio-pubs-static.s3.amazonaws.com/311446_08b00d63cc794e158b1f4763eb70d43a.html

https://www.eca.europa.eu/lists/ecadocuments/ap21_04/ap_air_passenger_rights_en.pdf

https://otexts.com/fpp2/stl.html

https://fred.stlouisfed.org/series/LNU07100000