TIME SERIES ANALYSIS AND FORECASTING FOR FRONTIER AIRLINES FLIGHT DELAYS

Table of Contents

SUMMARY:

Departure of flights on time plays a vital role in choosing any airline especially with a larger customer base as well as workforce relying on the convenience of access on-demand information. Tracking departure delays can have an intense and helpful impact on evaluating an airline's scope of improvement sectors and take appropriate measures in order to emerge out as a leader. Objective of our project is to extract data from Kaggle and choose the best forecasting model to predict the future departure delay and win the lost trust of disappointed customers of Frontier Airlines.

The data for departure delay on Frontier Airlines was available from www.kaggle.com, it contained around 144 records and 2 attributes. In our analysis we have decided to run three models in order to understand the departure delays of available dataset to forecast into 12 periods (2020) ahead. Before running the different models, we divided time series data into two partitions: Training and Validation data sets to be able to test how well any selected model performs with the new data not included in the model development. We have explored regression, two-level, smoothing and ARIMA models with historical data partitioning as well as on the entire dataset in order to arrive at the best accurate model. Each model is then compared particularly on two performance measures, MAPE and RMSE. Post comparison of models, we will conclude our analysis by identifying the best forecast model.

_____**********_____

INTRODUCTION:

Frontier Airlines is a domestic low-cost airline which is the eight-largest commercial airline in the US. They operate flights from over 100 airports throughout the United States with a fleet of 96 flights. Frontier was ranked in an airline quality rating report by Embry-Riddle Aeronautical University and Wichita State University in 2015 (ref.) as one of the five worst airlines in the United States. The main reason for the low ratings was the poor on-time performance of its flights and the confusion it was causing for customers. Frontier realized that they have started to incur losses as their customers are preferring other airline options due to the delays. Frontier would like to investigate the causes and are willing to come up with new features for customers to improve their travel experience. They are consulting with data analysts to identify and predict delay patterns based on airline flight delay data. We have presumed these flight delays are within the airline company's limitations and external factors are involved. The monthly departure delay data is given for a 12-year period, from January 2008 through December of 2019, and measured in hours and minutes. The goal is to increase customer satisfaction by identifying the best forecasting method to predict departure delay of Frontier Airlines in 12 months of 2020.

Based on these forecasts generated by the best accurate model, will be used to take proactive measures to minimize the departure delays, communicating with customers in advance, corrective measures if there are any workforce issues in order to improve the overall operations and customer satisfaction in 2020.

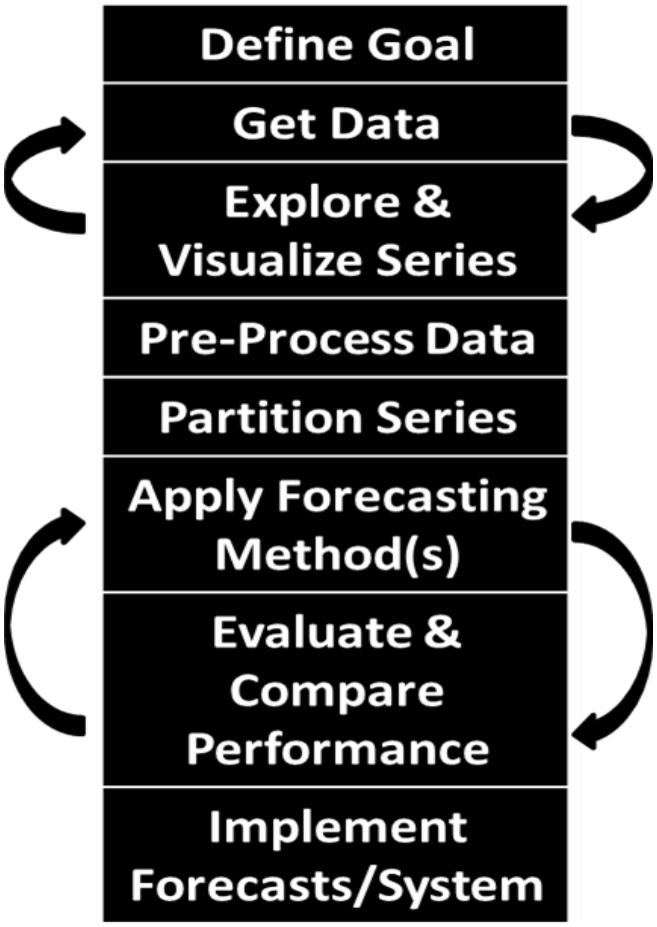
_____**********_____

MAIN CHAPTER:

The time series analysis will help us to describe and summarize time series data, fit dimensional models and make forecasts.

The time series forecasting is not a straightforward function to apply on the time series data. It includes an entire process of steps, which are not sequential always and sometimes we need to re-execute the steps based on the findings.

The forecasting process of a time series data typically includes the following:



DEFINE GOAL:

Our goal is to improve departure delay of Frontier Airlines flights by using forecasting methods like Regression, smoothing models and ARIMA models, which will help us in increasing customer satisfaction. We will be performing descriptive time series forecasting as it will incorporate analysis basis on understanding the correlation of successive data points, time series data patterns like trend and seasonality and any relationship to external factors.

Next important decision was to decide the forecast horizon and we reached the conclusion to forecast 12 periods into the future. The stakeholders for the forecast usage are the Frontier Airlines company and the decision making of flights schedule with minimal delays in year 2020 will be based on our forecasting results. Our dataset consists of 144 data points and we have modeled them using above mentioned models. The future scope of our model would be that it can be used for forecasting delays in future periods after year 2020 also by updating the model and the data accordingly.

DATA COLLECTION:

Considerations at the data collection level for forecasting results:

- Data quality
- Temporal frequency
- Series granularity

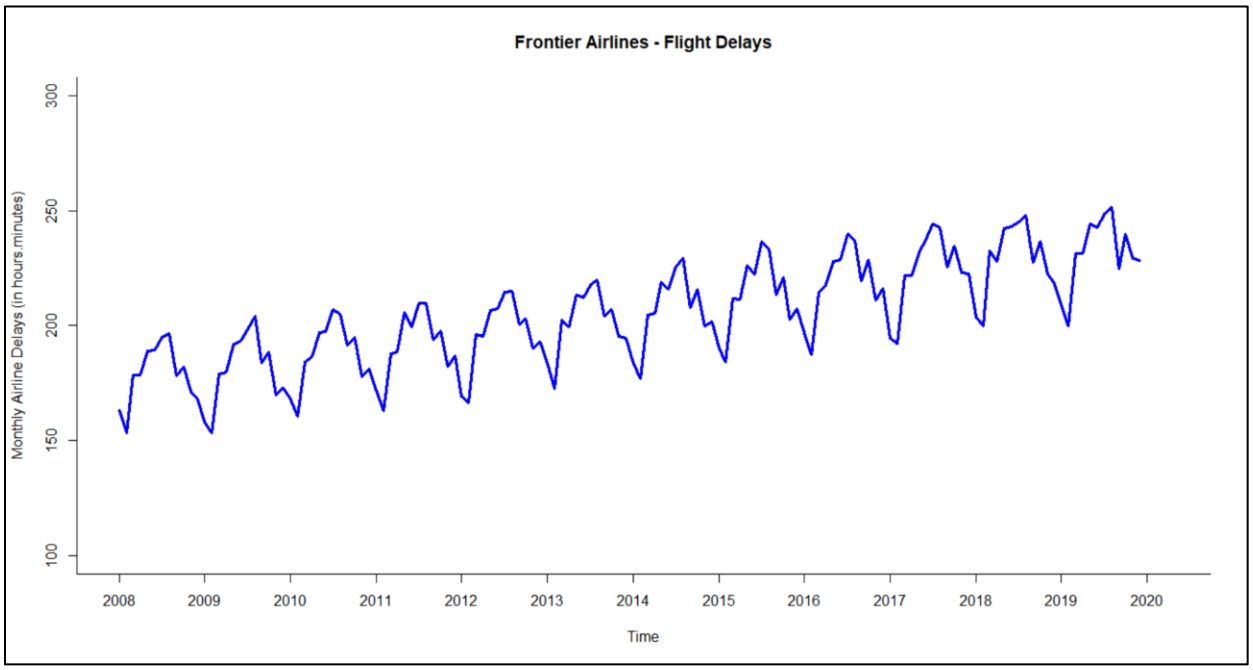
Above points were taken into consideration while collecting data. Depending on our forecast goal, we chose our series granularity to be monthly data of total departure delays. The data for departure delay of Frontier Airlines was available from https://www.kaggle.com/usdot/flight-delays#flights.csv

and it contained around 144 records and 2 attributes. In our analysis we have decided to run three methods in order to understand the delays trend of the past 12 years (2008 - 2019).

EXPLORE AND VISUALIZE TIME SERIES:

Explore time series step is used to do preliminary time series analysis of the data. For this purpose, we examined two charts namely, time plot to identify various components of the historical time series data patterns and correlogram to investigate various correlations between periods.

Time Plot for Frontier Airlines - Flight Delays Using plot() function in R

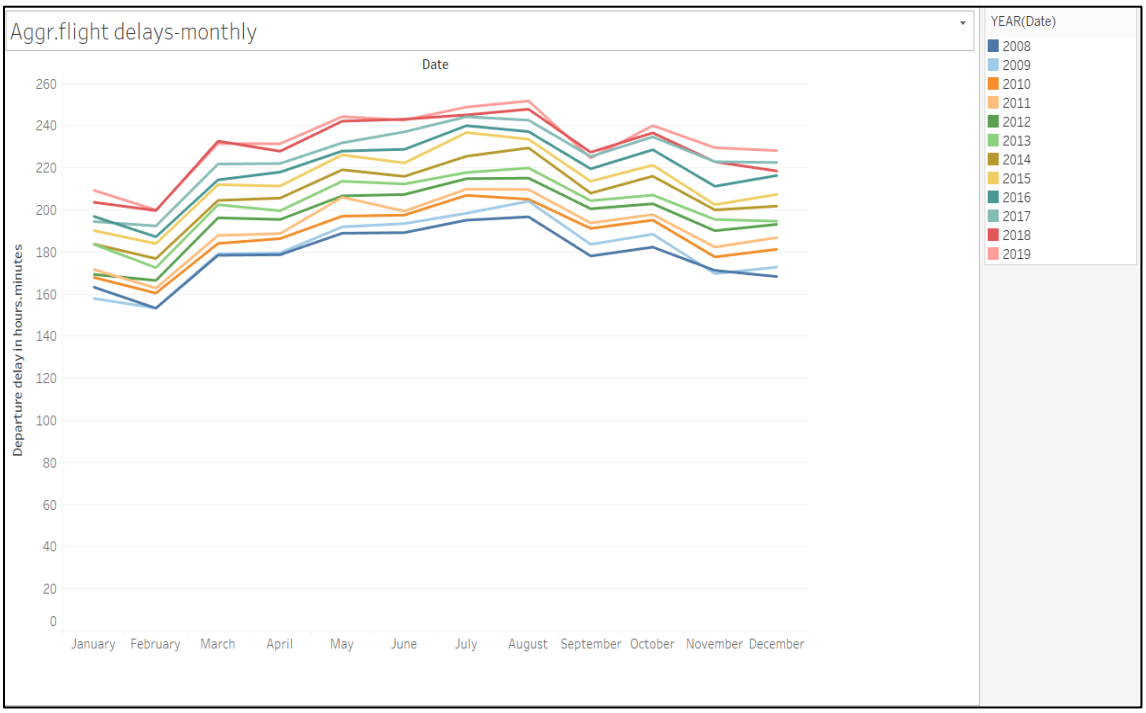


Time plot is a very informative and helps us to visualize the time series data. Visualizing the time series data will help us to identify the patterns or issues in the data such as outliers or extreme values, unequal spacing, breaks in the data that is not defined and missing values.

- The X-axis of the plot represents “Time” and Y-axis represents “Monthly Airline departure Delays (in hours. Minutes)”.

- The time series data is consistent, and we do not see any breaks or missing values or extreme or unusual values.
- The data plot above shows mostly leveled data with an upward trend and seasonality from 2008 through 2019.
- The seasonality picks after February and is highest in August of every year respectively. The below plot generated using Tableau.

Time Plot for Frontier Airlines - Flight Delays Using plot() function in R

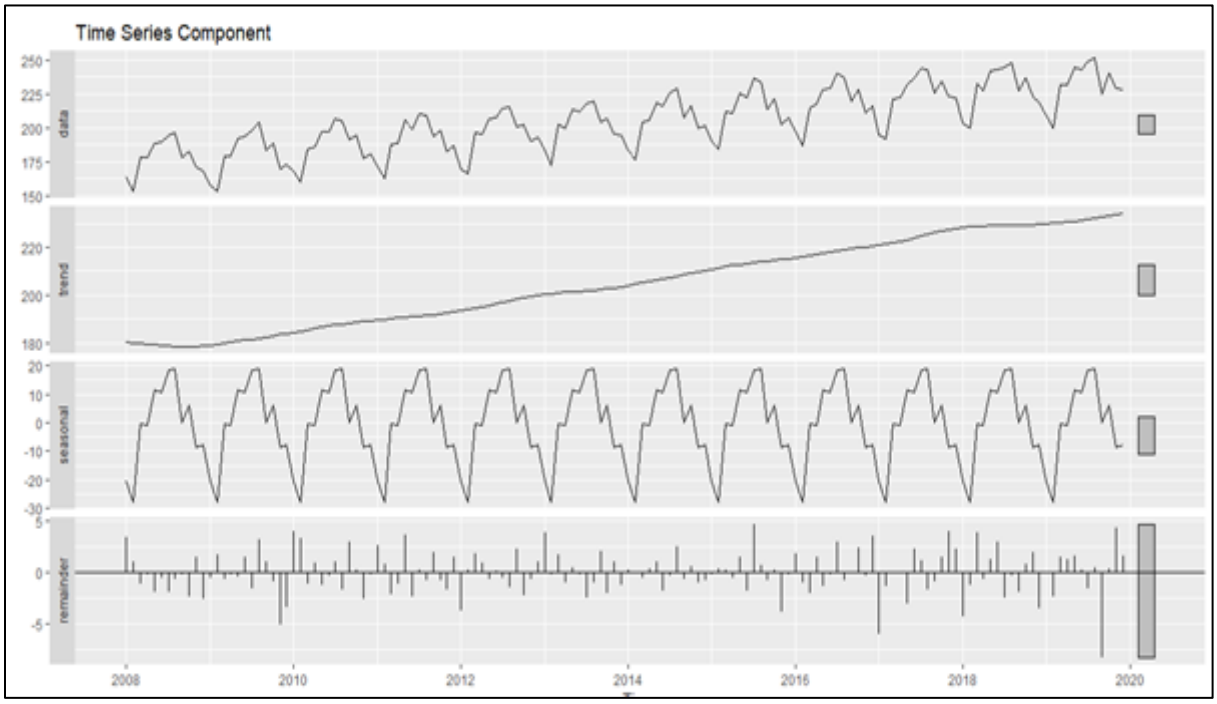


TIME SERIES COMPONENTS:

The stl() function in R, decomposes the regional data into trend, seasonality, reminder (level + noise). These time series components can also be identified using autocorrelation function (which will be discussed in the later sections).

The time series plot in general consists of Systematic part components and non-systemic part components.

Time Series Components Using stl() function in R



Systemic part components:

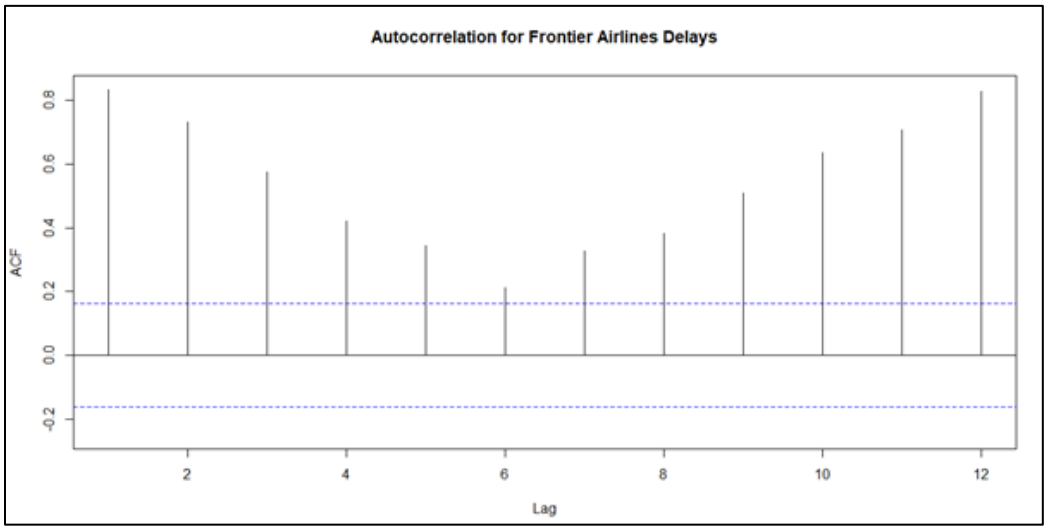
- **Level (stationary)** describes the average value of the series
- **Trend** describes the upward or downward change in the series from one period to the next, in our case it is an upward trend.
- **Seasonality** is a short-term cyclical behavior observed several times in a time series and the flight delays tome series data has monthly seasonality.

Non-systematic part component:

- **Noise (randomness)** – random variations resulting from measurement error or other causes not accounted for, which is shown in the plot (reminder) along with level.

AUTOCORRELATION:

Correlogram Using Acf() Function in R



In the time series context, values in neighboring periods tend to be correlated. Such correlation is called autocorrelation and is informative and can help to improve the forecasts.

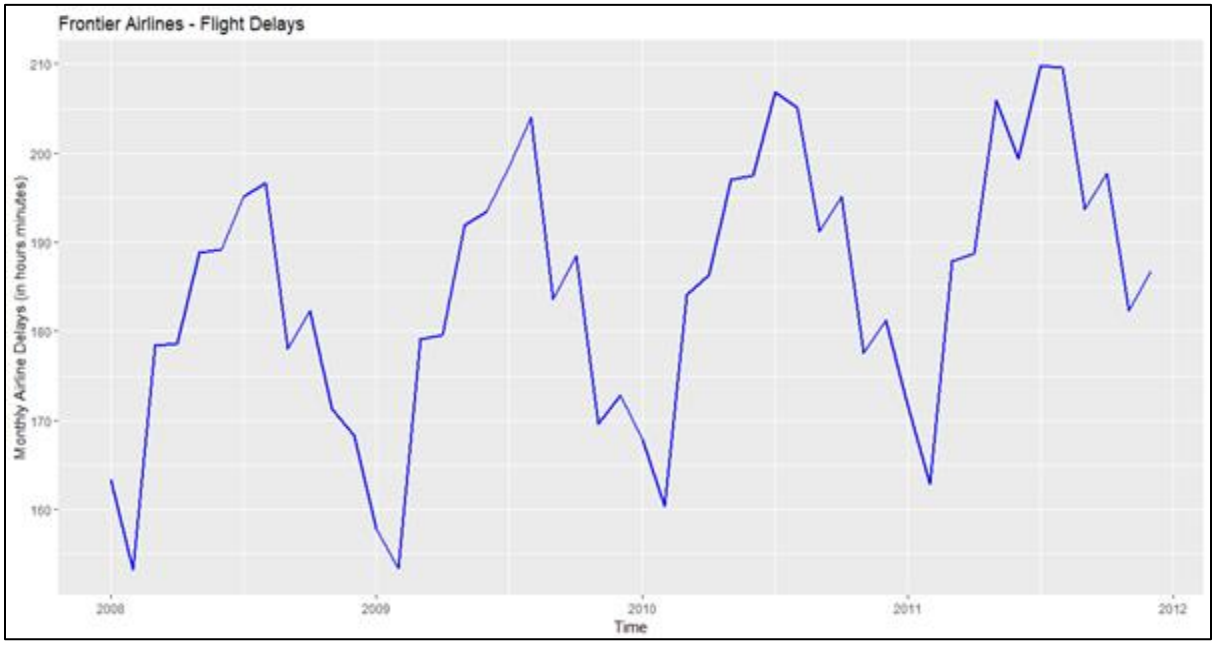
A positive autocorrelation coefficient at lag 1 through lag 12 is substantially higher, than the horizontal threshold which means they are statistically significant. This is typically indicative of an upward trend component. A high positive autocorrelation coefficient at lag 12, which is also statistically significant (greater than zero), points to monthly seasonality.

After running the model, we will try to eliminate autocorrelation which may overall improve the forecast.

All autocorrelation coefficients are statistically significant (monthly seasonality) because all the periods have ACF value beyond the two horizontal lines showing F-statistic value to be closer to 1 and having predictability nature. Therefore, we can confirm that is not a random walk and is predictable.

ZOOM-IN PLOT:

Zoom-In Time Plot Using autoplot() Function in R



This is a zoom-in version of Time plot starting from 2008 to 2012 for the monthly seasonality data. This zoom in plot helps us to observe the details of time plot in depth.

PRE-PROCESS DATA:

Pre-process data is used to detect potential issues and challenges within the data set. Initially, when we utilized the stl() function in R to analyze the components of time series data set, it was highly random and didn't incorporate any trend or seasonality. The data set was not clean enough to be ready to use for forecasting. The problem with the dataset was:

a. It showed no patterns of the trend as the data points were highly random.
b. The yearly departure delay values were not close to each other resulting in no seasonality pattern in data.
c. There were some missing and irrelevant periods in the data set.

We removed the above-mentioned flaws in the data by accumulating and consolidating data from different, yet similar data set from:

https://drive.google.com/file/d/0B76mgCvz1US5M1FrbmxfV2FXT3c/view

Following is the partial snippet of our cleaned data set:

A	B
Date	Departure_delay
01/01/2008	163.28
02/01/2008	153.25
03/01/2008	178.42
04/01/2008	178.68
05/01/2008	188.88
06/01/2008	189.16
07/01/2008	195.09
08/01/2008	196.67
09/01/2008	178.07
10/01/2008	182.27
11/01/2008	171.23
12/01/2008	168.29
01/01/2009	157.88
02/01/2009	153.34
03/01/2009	179.06
04/01/2009	179.52
05/01/2009	191.91
06/01/2009	193.45
07/01/2009	198.37
08/01/2009	204.05
09/01/2009	183.58
10/01/2009	188.43
11/01/2009	169.68

PARTITION SERIES:

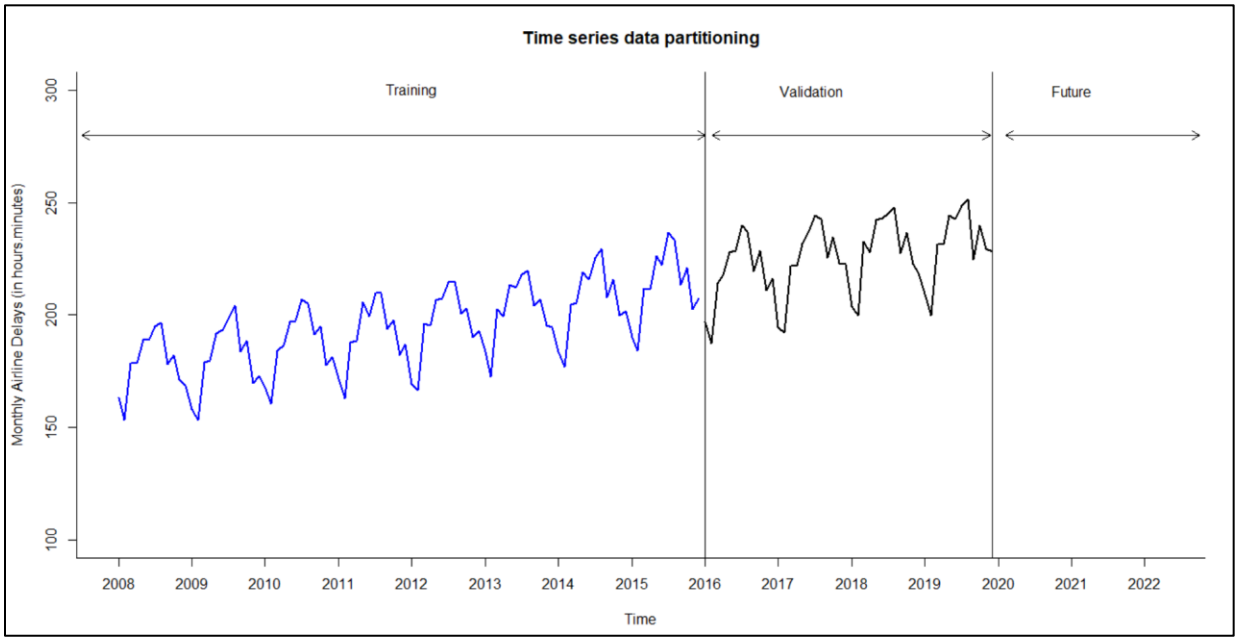
A crucial initial step for allowing the evaluation of predictive performance is data partitioning. This means that the forecasting method is applied only to a subset of the series. It is therefore important to understand why and how partitioning is carried out before applying any forecasting method.

We have partitioned our data set into training and validation data periods, which is a non-random partition. After which we developed a forecasting model based on training data, and test/validate model performance using the validation data.

The main reason for using partitioning is the problem of overfitting. Overfitting is a very potential problem and it means that a forecasting model is not only fitting the systematic components of the data (trend and seasonality), but also the noise. Over fitted model is likely to perform poorly on the new data to be forecasted.

The first 96 data periods are part of the training dataset (2008 to 2015) and the next 48 data periods (2016 till 2019) are part of the validation data set.

The below graph represents time series historical flight delays data partition into training, validation periods. The future period represents the forecasting yet to be performed using the best accurate model.



TRAINING DATASET:

```
> train.ts
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep
2008 163.28 153.25 178.42 178.68 188.88 189.16 195.09 196.67 178.07
2009 157.88 153.34 179.06 179.52 191.91 193.45 198.37 204.05 183.58
2010 167.89 160.43 184.03 186.32 196.99 197.49 206.88 205.02 191.16
2011 171.66 162.82 187.84 188.71 205.95 199.39 209.81 209.60 193.74
2012 169.31 166.45 196.19 195.41 206.60 207.28 214.78 215.05 200.51
2013 183.62 172.52 202.41 199.53 213.56 212.28 217.72 219.87 204.32
2014 183.74 176.82 204.47 205.56 219.00 215.87 225.44 229.39 207.91
2015 190.13 183.95 211.95 211.29 226.08 222.25 236.71 233.50 213.55
        Oct    Nov    Dec
2008 182.27 171.23 168.29
2009 188.43 169.68 172.77
2010 195.10 177.62 181.25
2011 197.76 182.30 186.80
2012 202.86 190.07 193.09
2013 206.99 195.41 194.59
2014 215.97 199.94 201.75
2015 221.22 202.42 207.32
>
```

VALIDATION DATASET:

```
> valid.ts
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep
2016 196.87 187.17 214.22 217.92 227.90 228.73 239.94 237.14 219.46
2017 194.44 192.33 221.74 221.98 231.81 237.02 244.20 242.57 225.30
2018 203.58 199.64 232.63 227.81 242.11 243.01 245.08 247.77 227.33
2019 209.26 199.91 231.49 231.37 244.25 242.58 248.79 251.69 224.57
        Oct    Nov    Dec
2016 228.52 211.18 216.30
2017 234.66 222.83 222.46
2018 236.55 222.80 218.44
2019 240.00 229.48 228.09
```

FORECASTING METHODS:

The team decided to build Regression-based models with trend and seasonality to forecast Frontier's monthly flight delays in the year 2020 because we have observed strong positive trend and seasonality components in the autocorrelation chart for the historic data discussed earlier.

Using the training data periods (January of 2008 through December of 2015), we have developed two regression-based models with trend and seasonality. Below are the models and their forecasts:

I. **Regression Model with Linear Trend and Seasonality**

Summary of the model:

```
Call:
tslm(formula = train.ts ~ trend + season)


Residuals:
    Min      1Q  Median      3Q     Max
-6.6425 -1.3632 -0.2301  1.3773  7.4378


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 155.423257   0.863337 180.026  < 2e-16 ***
trend         0.418965   0.008261  50.714  < 2e-16 ***
season2      -7.660215   1.112827  -6.884 1.03e-09 ***
season3      18.769570   1.112919  16.865  < 2e-16 ***
season4      18.431855   1.113072  16.559  < 2e-16 ***
season5      31.006640   1.113287  27.851  < 2e-16 ***
season6      29.112675   1.113562  26.144  < 2e-16 ***
season7      37.147460   1.113900  33.349  < 2e-16 ***
season8      37.772245   1.114298  33.898  < 2e-16 ***
season9      19.814530   1.114757  17.775  < 2e-16 ***
season10     24.115565   1.115277  21.623  < 2e-16 ***
season11      8.455351   1.115858   7.577 4.49e-11 ***
season12     10.185136   1.116500   9.122 3.72e-14 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 2.226 on 83 degrees of freedom
Multiple R-squared:  0.9872,    Adjusted R-squared:  0.9854
F-statistic: 534.5 on 12 and 83 DF,  p-value: < 2.2e-16
```

Explanation:

The regression model with linear trend and seasonality contains 12 independent variables: trend index (t), and 11 seasonal dummy variables for February (season2) through December (season12).

The regression equation is:

$$Y_t = 155.423 + 0.419\,t - 7.660\,D_2 + 18.769\,D_3 + 18.431\,D_4 + 31.007\,D_5 + 29.112\,D_6 + 37.147\,D_7 + 37.772\,D_8 + 19.814\,D_9 + 24.115\,D_{10} + 8.455\,D_{11} + 10.185\,D_{12}$$

The model has a very high R-squared of 0.9872 (adjusted R-squared of 0.9854). The regression coefficients for all the seasons (i.e., season2 through season12) are statistically significant for a significance of 99% or alpha = 0.01 (p-value for each of them is lower than 0.01).

Overall the model is statistically significant (F-statistic's p-value is much lower than 0.01) and can be a good model for forecasting Frontier's monthly flight delays data.

Forecast for validation data partition (i.e., 48 periods):

	Point Forecast	Lo 0	Hi 0		Point Forecast	Lo 0	Hi 0
Jan 2016	196.0629	196.0629	196.0629	Jan 2018	206.1180	206.1180	206.1180
Feb 2016	188.8216	188.8216	188.8216	Feb 2018	198.8768	198.8768	198.8768
Mar 2016	215.6704	215.6704	215.6704	Mar 2018	225.7255	225.7255	225.7255
Apr 2016	215.7516	215.7516	215.7516	Apr 2018	225.8068	225.8068	225.8068
May 2016	228.7454	228.7454	228.7454	May 2018	238.8005	238.8005	238.8005

Jun 2016	227.2704	227.2704	227.2704	Jun 2018	237.3255	237.3255	237.3255
Jul 2016	235.7241	235.7241	235.7241	Jul 2018	245.7793	245.7793	245.7793
Aug 2016	236.7679	236.7679	236.7679	Aug 2018	246.8230	246.8230	246.8230
Sep 2016	219.2291	219.2291	219.2291	Sep 2018	229.2843	229.2843	229.2843
Oct 2016	223.9491	223.9491	223.9491	Oct 2018	234.0043	234.0043	234.0043
Nov 2016	208.7079	208.7079	208.7079	Nov 2018	218.7630	218.7630	218.7630
Dec 2016	210.8566	210.8566	210.8566	Dec 2018	220.9118	220.9118	220.9118
Jan 2017	201.0904	201.0904	201.0904	Jan 2019	211.1456	211.1456	211.1456
Feb 2017	193.8492	193.8492	193.8492	Feb 2019	203.9043	203.9043	203.9043
Mar 2017	220.6979	220.6979	220.6979	Mar 2019	230.7531	230.7531	230.7531
Apr 2017	220.7792	220.7792	220.7792	Apr 2019	230.8343	230.8343	230.8343
May 2017	233.7729	233.7729	233.7729	May 2019	243.8281	243.8281	243.8281
Jun 2017	232.2979	232.2979	232.2979	Jun 2019	242.3531	242.3531	242.3531
Jul 2017	240.7517	240.7517	240.7517	Jul 2019	250.8068	250.8068	250.8068
Aug 2017	241.7954	241.7954	241.7954	Aug 2019	251.8506	251.8506	251.8506
Sep 2017	224.2567	224.2567	224.2567	Sep 2019	234.3118	234.3118	234.3118
Oct 2017	228.9767	228.9767	228.9767	Oct 2019	239.0318	239.0318	239.0318
Nov 2017	213.7354	213.7354	213.7354	Nov 2019	223.7906	223.7906	223.7906
Dec 2017	215.8842	215.8842	215.8842	Dec 2019	225.9393	225.9393	225.9393

II. Regression Model with Quadratic Trend and Seasonality

Summary of the model:

```
Call:
tslm(formula = train.ts ~ trend + I(trend^2) + season)


Residuals:
    Min      1Q  Median      3Q     Max
-5.7053 -1.3449 -0.1922  1.3374  5.6850


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.573e+02  9.303e-01 169.074  < 2e-16 ***
trend        3.033e-01  3.062e-02   9.907 1.14e-15 ***
I(trend^2)   1.192e-03  3.057e-04   3.901 0.000195 ***
```

```
season2     -7.648e+00  1.028e+00  -7.438 8.96e-11 ***
season3      1.879e+01  1.028e+00  18.273  < 2e-16 ***
season4      1.846e+01  1.028e+00  17.949  < 2e-16 ***
season5      3.104e+01  1.029e+00  30.174  < 2e-16 ***
season6      2.915e+01  1.029e+00  28.328  < 2e-16 ***
season7      3.718e+01  1.029e+00  36.126  < 2e-16 ***
season8      3.781e+01  1.030e+00  36.718  < 2e-16 ***
season9      1.984e+01  1.030e+00  19.264  < 2e-16 ***
season10     2.414e+01  1.031e+00  23.422  < 2e-16 ***
season11     8.467e+00  1.031e+00   8.212 2.65e-12 ***
season12     1.019e+01  1.032e+00   9.873 1.33e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 2.056 on 82 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.9875
F-statistic: 579.1 on 13 and 82 DF,  p-value: < 2.2e-16
```

Explanation:

The regression model with quadratic trend and seasonality contains 13 independent variables: trend index (t), squared trend index (t^2), and 11 seasonal dummy variables for February (season2) through December (season12). The regression equation is:

$$\mathbf{Y_t = 157.3 + 0.303\ t + 0.001\ t^2 - 7.648\ D_2 + 18.79\ D_3 + 18.46\ D_4 + 31.04\ D_5 + 29.15\ D_6 + 37.18\ D_7 + 37.81\ D_8 + 19.84\ D_9 + 24.14\ D_{10} + 8.467\ D_{11} + 10.19\ D_{12}}$$

The model has a very high R-squared of 0.9892 (adjusted R-squared of 0.9875). The regression coefficients for all the seasons (i.e., season2 through season12) are statistically significant for a significance of 99% or alpha = 0.01 (p-value for each of them is lower than 0.01).

Overall the model is statistically significant (F-statistic's p-value is much lower than 0.01) and can also be accurate model for forecasting Frontier's monthly flight delays data.

Forecast for validation data partition (i.e., 48 periods):

	Point Forecast	Lo 0	Hi 0		Point Forecast	Lo 0	Hi 0
Jan 2016	197.9301	197.9301	197.9301	Jan 2018	211.4479	211.4479	211.4479
Feb 2016	190.8176	190.8176	190.8176	Feb 2018	204.3927	204.3927	204.3927
Mar 2016	217.7952	217.7952	217.7952	Mar 2018	231.4275	231.4275	231.4275
Apr 2016	218.0052	218.0052	218.0052	Apr 2018	231.6947	231.6947	231.6947
May 2016	231.1277	231.1277	231.1277	May 2018	244.8745	244.8745	244.8745
Jun 2016	229.7815	229.7815	229.7815	Jun 2018	243.5855	243.5855	243.5855
Jul 2016	238.3640	238.3640	238.3640	Jul 2018	252.2252	252.2252	252.2252
Aug 2016	239.5366	239.5366	239.5366	Aug 2018	253.4550	253.4550	253.4550
Sep 2016	222.1266	222.1266	222.1266	Sep 2018	236.1023	236.1023	236.1023
Oct 2016	226.9754	226.9754	226.9754	Oct 2018	241.0083	241.0083	241.0083
Nov 2016	211.8629	211.8629	211.8629	Nov 2018	225.9530	225.9530	225.9530
Dec 2016	214.1404	214.1404	214.1404	Dec 2018	228.2878	228.2878	228.2878
Jan 2017	204.5173	204.5173	204.5173	Jan 2019	218.7219	218.7219	218.7219
Feb 2017	197.4335	197.4335	197.4335	Feb 2019	211.6953	211.6953	211.6953
Mar 2017	224.4396	224.4396	224.4396	Mar 2019	238.7587	238.7587	238.7587
Apr 2017	224.6783	224.6783	224.6783	Apr 2019	239.0546	239.0546	239.0546
May 2017	237.8294	237.8294	237.8294	May 2019	252.2630	252.2630	252.2630
Jun 2017	236.5118	236.5118	236.5118	Jun 2019	251.0026	251.0026	251.0026
Jul 2017	245.1229	245.1229	245.1229	Jul 2019	259.6710	259.6710	259.6710
Aug 2017	246.3241	246.3241	246.3241	Aug 2019	260.9293	260.9293	260.9293
Sep 2017	228.9427	228.9427	228.9427	Sep 2019	243.6052	243.6052	243.6052
Oct 2017	233.8201	233.8201	233.8201	Oct 2019	248.5398	248.5398	248.5398
Nov 2017	218.7363	218.7363	218.7363	Nov 2019	233.5132	233.5132	233.5132
Dec 2017	221.0424	221.0424	221.0424	Dec 2019	235.8766	235.8766	235.8766

The below table shows the accuracy measures based on validation data :

No.	Model	RMSE	MAPE
1.	Regression Model with Linear Trend and Seasonality	2.069	1.227
II.	Regression Model with Quadratic Trend and Seasonality	6.251	2.209

Based on the above MAPE and RMSE measures, the Regression Model with Linear Trend and Seasonality (MAPE = 1.227% and RMSE = 3.618) is more accurate than the Regression Model with Quadratic Trend and Seasonality (MAPE = 2.209% and RMSE = 6.251). Therefore, Regression Model with Linear Trend and Seasonality is the most accurate model and can be used to forecast the Frontier's monthly flight delays in 2020.

Regression Model with Linear Trend and Seasonality using entire data set:

Summary of the model:

```
Call:
tslm(formula = delays.ts ~ trend + season)


Residuals:
    Min      1Q  Median      3Q     Max
-9.4217 -1.5731 -0.0786  1.2608  8.8417


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.011061   0.841621 182.993  < 2e-16 ***
trend         0.427273   0.005314  80.401  < 2e-16 ***
season2      -7.346439   1.078484  -6.812 3.16e-10 ***
season3      20.211288   1.078524  18.740  < 2e-16 ***
season4      19.754848   1.078589  18.315  < 2e-16 ***
season5      31.905909   1.078681  29.579  < 2e-16 ***
season6      30.934470   1.078799  28.675  < 2e-16 ***
season7      38.365530   1.078942  35.558  < 2e-16 ***
season8      38.730758   1.079113  35.891  < 2e-16 ***
season9      19.735152   1.079309  18.285  < 2e-16 ***
season10     26.043712   1.079531  24.125  < 2e-16 ***
season11     11.002273   1.079780  10.189  < 2e-16 ***
season12     11.924167   1.080054  11.040  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.642 on 131 degrees of freedom
Multiple R-squared:  0.9879,    Adjusted R-squared:  0.9868
F-statistic: 891.5 on 12 and 131 DF,  p-value: < 2.2e-16
```

Explanation:

The regression model with linear trend and seasonality contains 12 independent variables: trend index (t), and 11 seasonal dummy variables for February (season2) through December (season12).

The regression equation is:

$$Y_t = 154.011 + 0.427\,t - 7.346\,D_2 + 20.211\,D_3 + 19.754\,D_4 + 31.905\,D_5 + 30.934\,D_6 + 38.365\,D_7 + 38.730\,D_8 + 19.735\,D_9 + 26.043\,D_{10} + 11.002\,D_{11} + 11.924\,D_{12}$$

The model has a very high R-squared of 0.9879 (adjusted R-squared of 0.9868). The regression coefficients for all the seasons (i.e., season2 through season12) are statistically significant for a significance of 99% or alpha = 0.01 (p-value for each of them is lower than 0.01). Overall the model is statistically significant (F-statistic's p-value is much lower than 0.01) and can be a good model to forecast monthly flight delays in 2020.

Therefore, we have applied the above model to forecast flight delays for Frontier Airlines in 2020.

Forecast using Regression model with linear trend and seasonality for flight delays in 2020:

	Point Forecast	Lo 0	Hi 0
Jan 2020	215.9656	215.9656	215.9656
Feb 2020	209.0464	209.0464	209.0464
Mar 2020	237.0314	237.0314	237.0314
Apr 2020	237.0023	237.0023	237.0023
May 2020	249.5806	249.5806	249.5806
Jun 2020	249.0364	249.0364	249.0364

Jul 2020	256.8948	256.8948	256.8948
Aug 2020	257.6873	257.6873	257.6873
Sep 2020	239.1189	239.1189	239.1189
Oct 2020	245.8548	245.8548	245.8548
Nov 2020	231.2406	231.2406	231.2406
Dec 2020	232.5898	232.5898	232.5898

We also wanted to develop models using Smoothing Methods to compare with the Regression Model with Linear Trend and Seasonality. We developed a Two-Level forecasting model (combining the Regression model with Linear Trend and Seasonality + trailing MA for residuals) and the Holt-Winter's model with automated selection of model options (i.e., an advanced exponential smoothing model).

Below are the models and their forecasts:

III. Two-level (Regression model with linear trend and seasonality + Trailing MA for residuals) :

In this section we implemented multilevel forecasting model by combining the original time series to generate forecast of future values and then use the forecast errors to generate forecast of future forecast errors. This will correct the first level forecasts. Since we have the best predictive accuracy with linear trend and seasonality from the previous sections, we further forecasted the future forecast errors using the second method of trailing moving average (TMA). The two-level combined forecasting methods can lead to improved predictive performance based on the business use case.

As we know trailing MA forecasting method is generally used for time series data with no trend and no seasonality.

We are implementing the first part of two-level forecasting i.e. de-trending and de-seasonalizing the time series data and forecasting the value for the next 12 months (2020) using regression forecasting model.

The 2nd part of the trailing moving average for the residuals and combining these two parts in the two-level forecasting model will be discussed below.

The first level regression model with linear trend and seasonality is the same as discussed in the regression model in above sections.

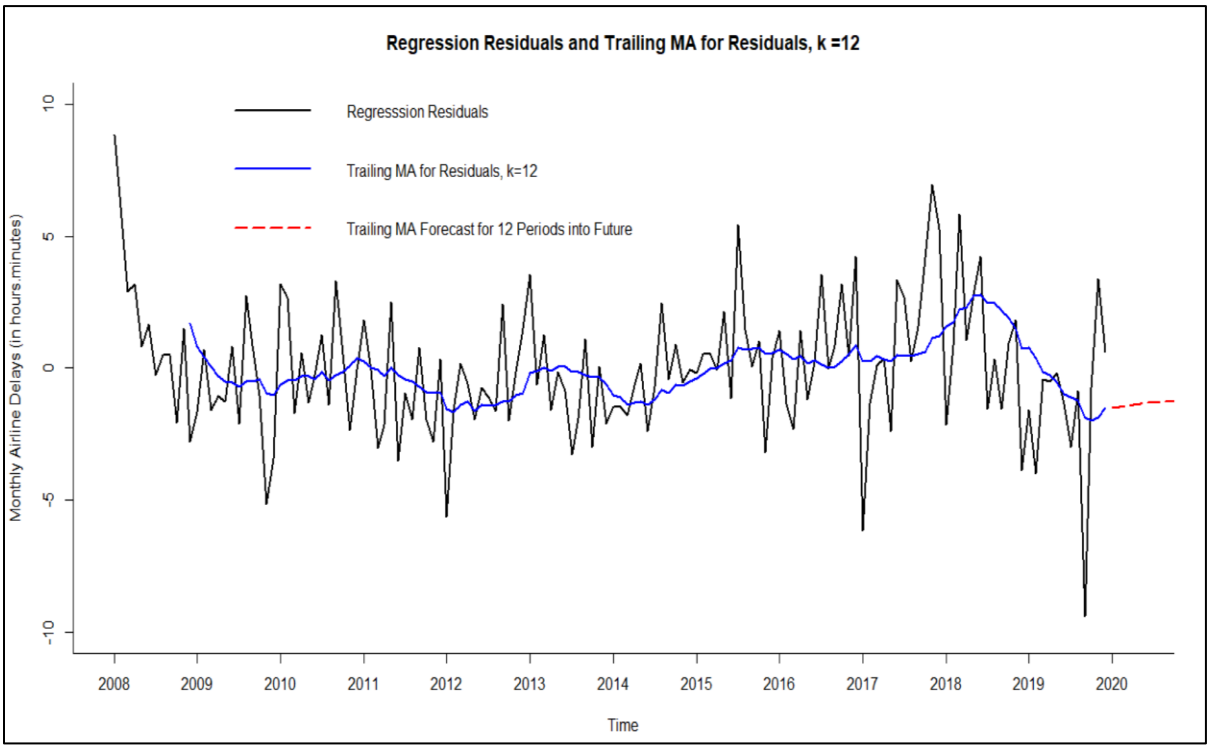
We developed trailing MA smoothing method for the residuals (with window size 12) from regression model with linear trend and seasonality. The two- level forecasting model is generated by combining the forecast developed for regression model and the residuals forecast using trailing MA method. A forecast of residuals for the next 12 months (2020) with point estimate is generated.

The below-mentioned table has three columns representing the regression forecast, trailing MA forecast and the total forecast in 2020 respectively.

	delays_reg_seas_pred.mean	ma.trailing.res_12.pred.mean	ts.forecast.12
1	215.9656	-1.525798	214.4398
2	209.0464	-1.465221	207.5812
3	237.0314	-1.416760	235.6147
4	237.0023	-1.377991	235.6243
5	249.5806	-1.346976	248.2336
6	249.0364	-1.322163	247.7143
7	256.8948	-1.302314	255.5925
8	257.6873	-1.286434	256.4008
9	239.1189	-1.273730	237.8452
10	245.8548	-1.263567	244.5912
11	231.2406	-1.255437	229.9852

12	232.5898	-1.248932	231.3408

The residuals generated form first level regression model and second level residuals can be plotted together to get more insight of the model on the entire dataset.



The residuals forecasted from trailing moving average smoothing method has been added to the first level regression model which improved/corrected the overall prediction accuracy of the combined model.

Forecast using Two-level (Regression model with linear trend and seasonality + trailing MA for residuals) for flight delays in 2020:

	ts.forecast.12
1	214.4398
2	207.5812
3	235.6147
4	235.6243

5	248.2336
6	247.7143
7	255.5925
8	256.4008
9	237.8452
10	244.5912
11	229.9852
12	231.3408

IV. Holt-Winter's Model with Automated Selection of Model Options using entire data:

The Holt-Winter's model (Winter's model) is an advanced exponential smoothing method and is used for time series data that contains trend and seasonality. Holt-Winter's model is developed with automated selection of error, trend, and seasonality options (Z,Z,Z) and automated selection of smoothing parameters (no parameters specified in the function).

Summary of the model:

```
ETS(A,A,A)

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

  Smoothing parameters:
    alpha = 0.2051
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 178.9869
    b = 0.3939
    s = -8.0946 -9.0189 6.1656 0.237 18.7203 18.3006
           11.0424 11.831 -0.4952 0.1214 -27.9659 -20.8437
```

```
  sigma:  2.6198

     AIC     AICc      BIC
1010.062 1014.919 1060.549
```

Explanation:

The best possible model and the best possible parameters (α, β, γ) developed by ets() are as follows:

Based on the automatic definition of the model, the model that we received is ETS (A,A,A) i.e. Additive error, additive trend and additive seasonality. The parameters that are optimized by the ets model are alpha = 0.2051, Beta = 1e-04 (0.0001) and gamma = 1e-04 (0.0001).

The alpha value of this model indicates that the model's level component tends to be more global, which also represents slow learning of trend component while additive seasonality is globally adjusted as gamma is close to zero. The latter is also indicating that, according to this model, the seasonality does not change over time.

Forecast using Holt-Winter's Automated selection of options model for flight delays in 2020:

	Point Forecast	Lo 0	Hi 0
Jan 2020	214.0751	214.0751	214.0751
Feb 2020	207.3460	207.3460	207.3460
Mar 2020	235.8261	235.8261	235.8261
Apr 2020	235.6033	235.6033	235.6033
May 2020	248.3227	248.3227	248.3227
Jun 2020	247.9274	247.9274	247.9274
Jul 2020	255.5794	255.5794	255.5794
Aug 2020	256.3925	256.3925	256.3925
Sep 2020	238.3021	238.3021	238.3021
Oct 2020	244.6247	244.6247	244.6247
Nov 2020	229.8339	229.8339	229.8339

```
Dec 2020       231.1516 231.1516 231.1516
```

These models have been implemented on the entire data set and accuracy measures are compared between the Two-level combined forecast and Holt-Winter's forecast.

The below table shows the accuracy measures based on entire data set:

No.	Model	RMSE	MAPE
III.	Two-level model (Regression + trailing MA for residuals)	2.276	0.834
IV.	Holt-Winter's model	2.47	0.939

Based on the lowest values of MAPE and RMSE of the compared forecasting models, the most accurate (best) model for forecasting the Frontier airlines monthly flight delays in 2020 is the Two-level model (Regression with linear trend and seasonality + trailing MA for residuals). Therefore, can be applied to forecast Frontier Airline's monthly flight delays in 2020.

Forecast using Two-level (Regression model with linear trend and seasonality + trailing MA for residuals) for flight delays in 2020:

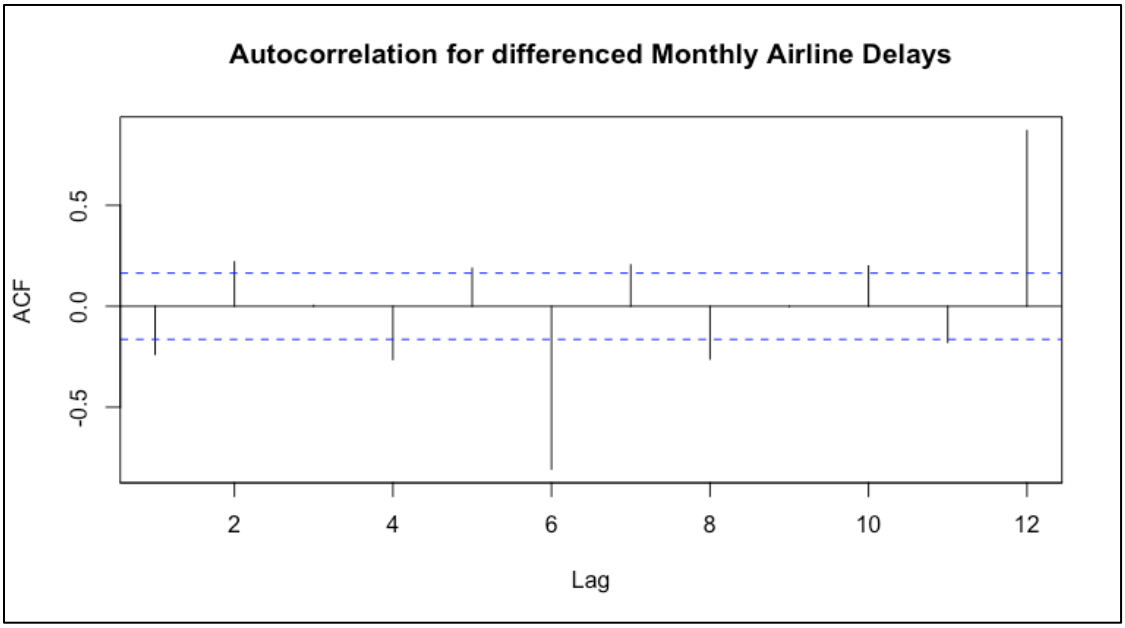
	ts.forecast.12
1	214.4398
2	207.5812
3	235.6147
4	235.6243
5	248.2336
6	247.7143
7	255.5925
8	256.4008
9	237.8452
10	244.5912
11	229.9852
12	231.3408

Further, we developed two more forecasting models that are based on Autocorrelation and Autoregressive models, i.e., Two-Level forecasting model (combining the Regression model with Linear trend and Seasonality + AR(1) for Residuals) and Auto ARIMA model.

Firstly, we used the autocorrelation of the first differencing for delays.ts data to:

- Check time series forecast residuals for independence (or dependence)
- Model remaining information (residuals) if residual dependency exists
- Evaluate predictability, i.e., whether the series is a random walk or can be predictable

Autocorrelation of the first differencing for delays.ts:



From the above plot, we noticed that several autocorrelation coefficients of the first differenced data are statistically significant, particularly in lag 12 (for monthly seasonality). Therefore, using the first differencing, we can confirm that delays.ts is not a random walk and can be predictable.

Below are the two models based on Autocorrelation and Autoregressive models using the entire dataset:

V. Two-level forecasting (Regression model linear trend and seasonality + AR(1) for Residuals):

We implemented multilevel forecasting model by combining the original time series to generate forecast of future values and then use the forecast errors to generate forecast of future forecast errors. This will correct the first level forecasts. Since we have the best predictive accuracy with linear trend and seasonality from the previous sections, we further forecasted the future forecast errors using the second method of AR(1) for residuals. The two-level combined forecasting methods can lead to improved predictive performance based on the business use case.

We are implementing the first part of two-level forecasting i.e. the first level regression model with linear trend and seasonality for entire dataset is the same as discussed in the regression model in above sections.

The 2nd part of the AR(1) model for the residuals and combining these two parts in the two-level forecasting model will be discussed below.

AR(1) model for Regression Residuals:

Developed AR(1) model for the residuals from regression model with linear trend and seasonality. The two- level forecasting model is generated by combining the forecast developed for regression model and the residuals forecast using AR(1) model. A forecast of residuals for the next 12 months (2020) with point estimate is generated.

Summary of the model:

```
Series: entiredata.lineartrend.season$residuals
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1    mean
      0.2137  0.0178
s.e.  0.0850  0.2610

sigma^2 estimated as 6.165:  log likelihood=-334.31
AIC=674.61   AICc=674.78   BIC=683.52

Training set error measures:
                      ME      RMSE      MAE       MPE      MAPE      MASE         ACF1
Training set -0.01452417 2.465691 1.83832 110.5167 136.7467 0.7485062 -0.03436167
```

Explanation:

From the above AR(1) model for flight delays data residuals, we have the order as c(1, 0, 0) in which the first parameter (i.e., p=1) describes the order of Autoregressive model, the second parameter (i.e., d=0) describes the order of differencing and the third parameter (i.e., q=0) describes the order of Moving Average.

The coefficient of ar1 (β_1) = 0.2137, the intercept ($\propto$) = 0.0178 and the standard error = 0.0850.

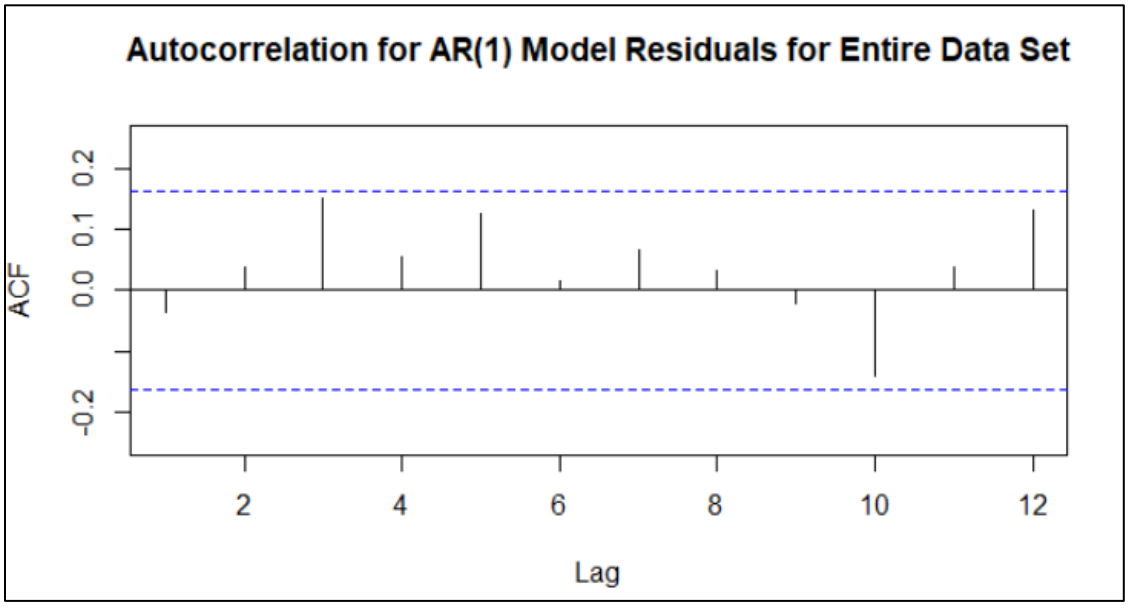
AR(1) model equation for residuals:

$$\boldsymbol{e_t = \alpha + \beta_1 e_{t-1} + \varepsilon_t} \quad => \quad \boldsymbol{e_t} = 0.0178 + 0.2137\, \boldsymbol{e_{t-1}}$$

Below is the prediction of residuals into the future 12 months of 2020:

	Point Forecast	Lo 0	Hi 0
Jan 2020	0.14813569	0.14813569	0.14813569
Feb 2020	0.04567575	0.04567575	0.04567575
Mar 2020	0.02377584	0.02377584	0.02377584
Apr 2020	0.01909492	0.01909492	0.01909492
May 2020	0.01809441	0.01809441	0.01809441
Jun 2020	0.01788057	0.01788057	0.01788057
Jul 2020	0.01783486	0.01783486	0.01783486
Aug 2020	0.01782509	0.01782509	0.01782509
Sep 2020	0.01782300	0.01782300	0.01782300
Oct 2020	0.01782255	0.01782255	0.01782255
Nov 2020	0.01782246	0.01782246	0.01782246
Dec 2020	0.01782244	0.01782244	0.01782244

Autocorrelation for AR(1) Model Residuals for Entire Data Set



The autocorrelation chart above of the AR(1) model residuals for entire dataset shows that all the autocorrelations in lags 1 to lags 12 are within horizontal thresholds (i.e., insignificant), which

means that our AR(1) model absorbed all the autocorrelation relationships pf the regression model's residuals.

The table below provides 3 forecasts for the future 12 months of 2020 associated with: regression model with linear trend and seasonality (Reg.Forecast), AR(1) model for the regression residuals (AR(1)Forecast), and two-level combined forecast (Combined.Forecast) as a sum of the regression and AR(1) models' forecasts.

	Reg.Forecast	AR(1)Forecast	Combined.Forecast
1	215.9656	0.14813569	216.1137
2	209.0464	0.04567575	209.0921
3	237.0314	0.02377584	237.0552
4	237.0023	0.01909492	237.0214
5	249.5806	0.01809441	249.5987
6	249.0364	0.01788057	249.0543
7	256.8948	0.01783486	256.9126
8	257.6873	0.01782509	257.7051
9	239.1189	0.01782300	239.1368
10	245.8548	0.01782255	245.8726
11	231.2406	0.01782246	231.2584
12	232.5898	0.01782244	232.6076

VI. Auto ARIMA model using entire dataset:

We developed this model to identify the optimal ARIMA model and its respective (p, d, q)(P, D, Q) parameters. Unlike, regular ARIMA models that are rather complex and require to input number of parameters, Auto ARIMA model does not require to input any of these parameters into the function.

Summary of the model:

```
Series: delays.ts
ARIMA(1,0,1)(0,1,1)[12] with drift


Coefficients:
         ar1      ma1     sma1   drift
      0.8938  -0.7046  -0.6375  0.4117
s.e.  0.0811   0.1032   0.0869  0.0239


sigma^2 estimated as 7.125:  log likelihood=-317.95
AIC=645.91   AICc=646.38   BIC=660.32


Training set error measures:
                      ME      RMSE       MAE          MPE       MAPE       MASE         ACF1
Training set -0.004052025 2.516697 1.782887 -0.03593232 0.8742909 0.3469009 -0.09291707
```

Explanation:

The above model obtained by using auto.arima() is a seasonal ARIMA model. The first three parameters of the model describe an AR component with order 1 (p=1) for trend, no differencing (d=0), and order 1 moving average components for trend. The next three parameters describe no AR seasonal component (P=0), first differencing (D=1) and MA with order 1 (Q=1) components. The model is also done for monthly seasonality (number 12 in the ARIMA model description). The drift parameter is a constant parameter (like intercept) in the MA portion of this model. The ARIMA model's equation is:

$$y_t = 0.4117 + 0.8938\, y_{t-1} - 0.7046\, \varepsilon_{t-1} - 0.6375\, \rho_{t-1}$$

The above autoregressive and ARIMA-based models have been implemented on the entire data set and below is the comparison of accuracy measures between the Two-level combined forecast

(Regression model with linear trend and seasonality + AR(1) model for residuals) and Auto ARIMA model:

No.	Model	RMSE	MAPE
V.	Two-level combined model (Regression + AR(1) for residuals)	2.466	0.913
VI.	Auto ARIMA model	2.517	0.874

Based on the lowest values of MAPE and RMSE of the compared forecasting models, the most accurate (best) model for forecasting the Frontier airlines monthly flight delays in 2020 is the Auto Arima model.

Also, it is important to point that, based on MAPE and RMSE, the Two-level combined model (Regression + AR(1) for residuals) accuracy is only marginally worse than that of the best model (Auto ARIMA model), and therefore, may be also taken into consideration for forecasting airlines monthly flight delays in 2020.

Forecast using Auto ARIMA model for flight delays in 2020:

	Point Forecast	Lo 0	Hi 0
Jan 2020	212.9836	212.9836	212.9836
Feb 2020	206.3683	206.3683	206.3683
Mar 2020	236.9742	236.9742	236.9742
Apr 2020	236.2018	236.2018	236.2018
May 2020	248.8900	248.8900	248.8900
Jun 2020	248.9567	248.9567	248.9567
Jul 2020	255.5444	255.5444	255.5444
Aug 2020	256.7851	256.7851	256.7851
Sep 2020	234.9848	234.9848	234.9848
Oct 2020	245.9499	245.9499	245.9499
Nov 2020	232.9626	232.9626	232.9626
Dec 2020	232.3662	232.3662	232.3662

So far, we have implemented three forecasting methods, i.e., regression-based models, smoothing methods and autoregressive models and identified three most accurate forecasting models (i.e., one best model from each method mentioned previously). In the next step, we have compared the performance measures of these most accurate models with the performance measures of the baseline models, i.e., Naïve forecast and Seasonal naïve forecast to choose the best model among these models to forecast monthly flight delays in 2020 for Frontier Airlines.

COMPARISON OF ACCURACY MEASURES:

The accuracy measures for the 5 models (for the entire data set) are presented below:

No.	Model	RMSE	MAPE
1	Two-level model (Regression + trailing MA for residuals)	2.276	0.834
2	Auto ARIMA model	2.517	0.874
3	Regression model with linear trend and seasonality	2.520	0.932
4	Seasonal naïve model	5.804	2.495
5	Naïve model	12.661	4.794

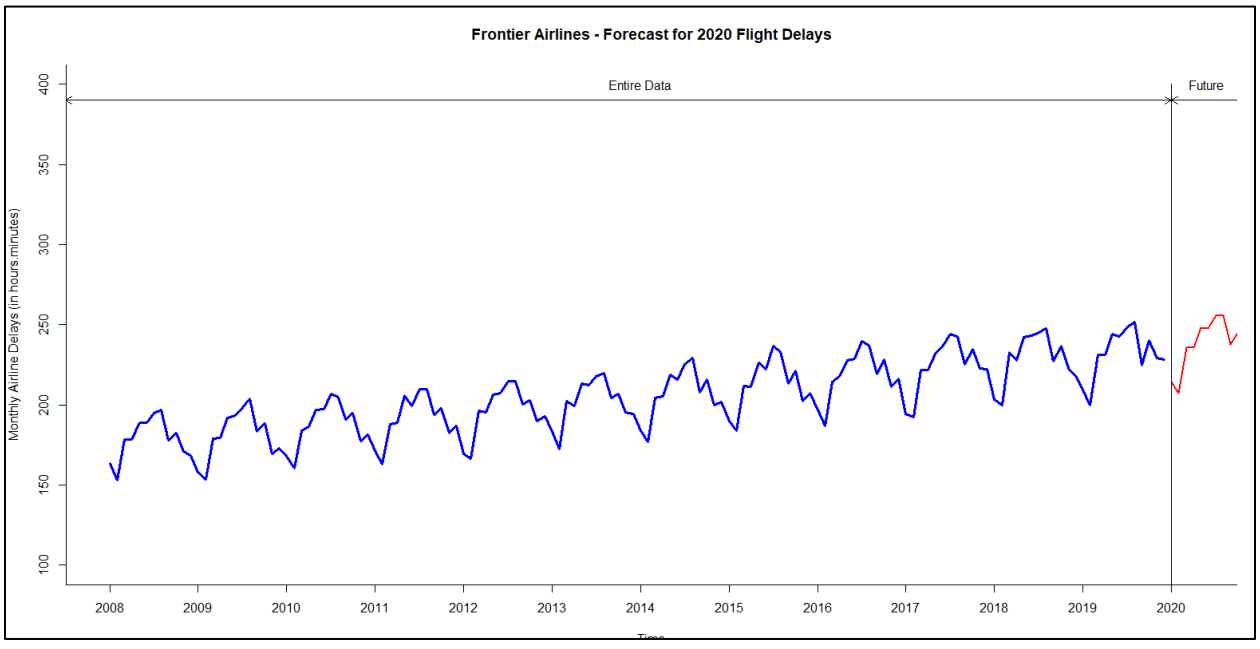
Based on the lowest values of MAPE and RMSE of the compared forecasting models, the most accurate (best) model for forecasting is the Two-level model, i.e., the combination of regression model with linear trend and seasonality and trailing MA model for residuals.

Therefore, we chose Two-level model (Regression with linear trend and seasonality + trailing MA for residuals) to forecast monthly flight delays in 2020 for Frontier Airlines.

Forecast using Two-level (Regression model with linear trend and seasonality + trailing MA for residuals) for flight delays in 2020:

	ts.forecast.12
1	214.4398
2	207.5812
3	235.6147
4	235.6243
5	248.2336
6	247.7143
7	255.5925
8	256.4008
9	237.8452
10	244.5912
11	229.9852
12	231.3408

We used the most accurate model developed to forecast flight delays that might incur for Frontier Airlines in the year 2020. Also, we created a plot to visualize the historical data and the best model predictions for future periods.



- The plot for the 2020 forecast also shows an upward trend and seasonality in 2020 like historic data.
- The seasonality picks after February of 2020 and is highest in August of 2020

CONCLUSION

In this project, R-programming is used to demonstrate the optimal models to best predict the airline departure delays that help Frontier airlines to improve their customer satisfaction. Through the forecasting models that we built, we were able to forecast the delays in advance and take proactive measures to minimize the departure delays and communicate with customers in advance about the delays which will help improve the overall operations and customer satisfaction in 2020.

The best models among all the models we built were: Two-level model (Regression + trailing MA for residuals), Regression model with linear trend and seasonality, Auto ARIMA models. Comparing these models with naïve and seasonal naïve models, the model with the lowest RMSE(2.276) and lowest MAPE(83.4%) is selected as the best model i.e., Two-level model (Regression + trailing MA for residuals). The forecasting is done for 2020 using this model and the results are displayed above section.

The forecasted delays of Frontier airlines in 2020 has the lowest value in February and highest value in August which would suggest close monitoring of resources available, rewarding the workforce who are helping the on-time departure of flights and other corrective actions by the airline company.

The forecast generated would give guidance to the Frontier Airlines and the forecast needs to be updated/automated to use for near future periods.

Limitations and options to overcome: One of the most important limitation in the time series forecasting is the problem of overfitting, the models need to be revalidated with reference to accuracy measures form the historical data partitioning whenever we get the updated data.

REFERENCES:

- Practical Time Series Forecasting with R – 2nd Edition
- Analysis of Time Series an Introduction – 5th Edition – by Chris Chatfield
- Some Blogs around Airline Industry:
    - 5 worst airlines in America: https://fortune.com/2015/04/13/five-worst-airlines-in-america/
    - Frontier Airlines Wiki: https://en.wikipedia.org/wiki/Frontier_Airlines#Customer_satisfaction_and_airline_ratings
    - Frontier Airlines Rating at Skytrax website: https://www.airlinequality.com/airline-reviews/frontier-airlines/
    - Website showing live cancelled and delayed flight statistic: https://flightaware.com/live/cancelled
- Google.Com for all day to day references and internet searches
- Kaggle.com for searching datasets
- Dr. Zinovy Radovilsky for valuable inputs and feedback throughout the project execution

****

Project Report By:

- Ayushi Shrivastava (net id : tf2272)

- Bhargavi Sankula (net id : qv4668)

- Renuka Reddy Valisekkagari – (net id : iw4227)

- Shubhangi Jain – (net id : qt9375)

- Sowmya Chintha (net id: jf8997)

_____**********_____