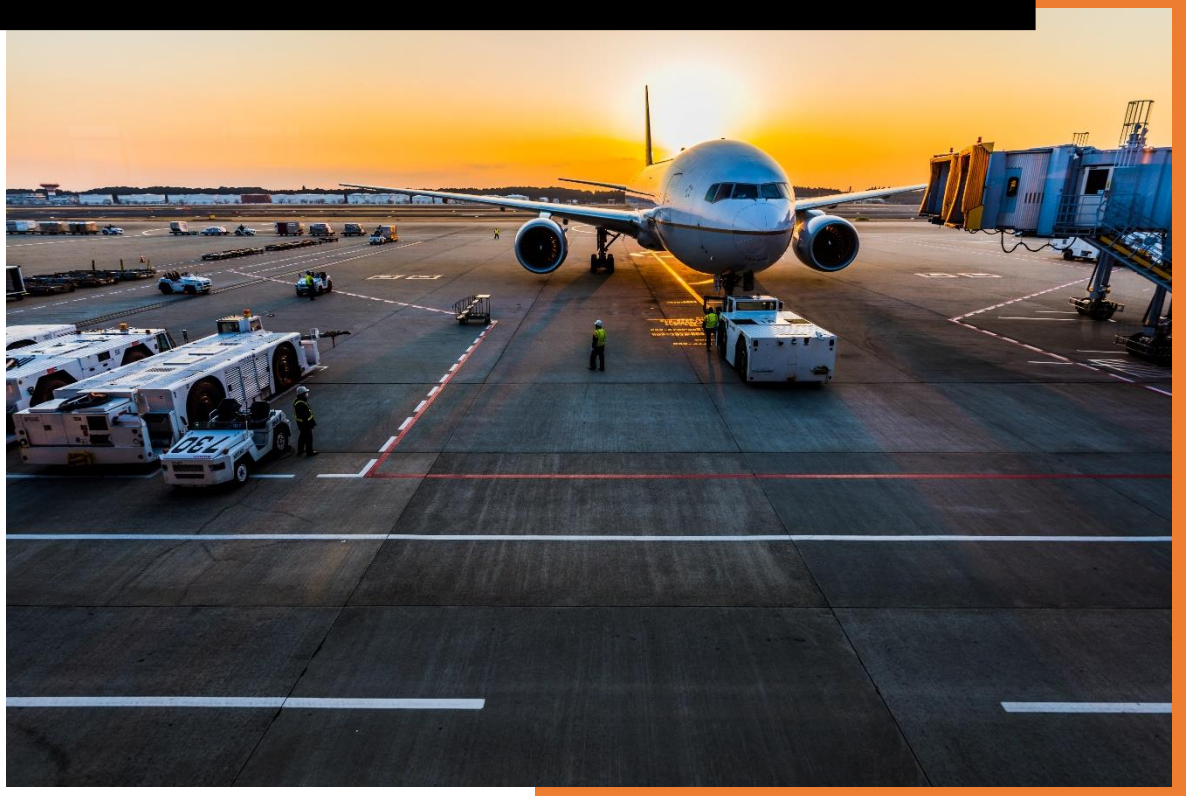
TIME SERIES FORECASTING – SAN FRANCISCO INTERNATIONAL AIRPORT AIR TRAFFIC



Submitted by:

RAHEL GHEBREKIDAN

VANSHIKA GUPTA

DEEPTHI MANDALAPARTY

KARAN MALIK

Cal State East Bay

7/25/2021

Executive Summary

For this project, data from SFO Air traffic and Cargo Statistics has been used to predict the monthly air traffic of the passengers for the upcoming fiscal year. The analysis is based on 13 years with monthly data. From the visualization we identified that dataset has an upward trend with seasonality. Also, the data is highly auto correlated, as the autocorrelation coefficient in all 12 lags are significant. Regression-based models, advanced exponential smoothing models and, autoregressive integrated moving average models (ARIMA) were utilized for this project. Additional variations of the regression and advanced exponential smoothing models were also constructed to ensure good results. Model evaluation was based on the RMSE and MAPE accuracy metrics.

We have used some of the most comparable models of Time Series Forecasting to compare Air traffic travelling from/to San Francisco Airport and thereby chosen some of the best models which can most accurately predict Air traffic and help them set most realistic idea of the Air traffic for the next year. Models like Multi-Linear Regression, Arima, Holt-Winter, etc form the base of our analysis. We have constructed these models using R as the statistical tool and have tried to learn some of the unexplored relationships in the historical data, via various Graphs and Time Series components. A correctly implemented Forecasting model would be of utmost help to SF Airport to figure out the Air Traffic for the upcoming year. We are hopeful that our model and our findings can be of some value and reference to similar real-world scenarios.

Introduction

San Francisco International Airport (SFO), California, USA is one of the busiest airports in the United States and also an important transpacific gateway for international passenger. In 2017 it was the seventh-busiest airport in the United States and the 24th-busiest in the world by passenger count. This explores the passenger and cargo traffic data in and out of SFO airport and the impact of COVID-19 on air travel from/to SFO. The airport is owned and operated by the City and County of San Francisco, although it is outside of San Francisco in unincorporated San Mateo County. As the world is still trying to get out of the COVID-19 pandemic, we all know that transportation industry is one of the heavily impacted ones. It would be good learning experience to explore a dataset from the industry. SFO being one of the major hubs in USA, exploring this airport statistics may be a good sample to understand how the industry is doing.

Eight Steps of Forecasting

Step 1: Define Goal

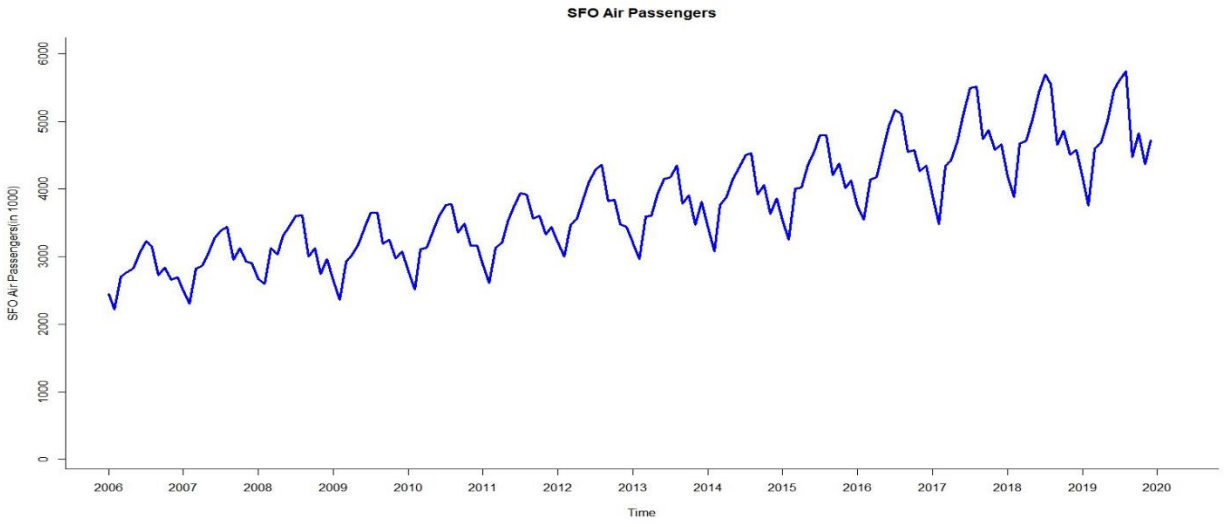
The goal of this project is to study and forecasts the monthly travel pattern of the passenger to/from the San Francisco International Airport for the upcoming one year. The objective is to create a predictive model which will properly consider both the trend and seasonal components of the historical data and effectively forecast the desired monthly data. Naturally, the model with the highest accuracy will be considered the model of choice. We will leverage R to execute and compare performance of various forecasting

models and chose the most accurate model for forecasting the passengers travelling to/from the SF international Airport for the next 12 months.

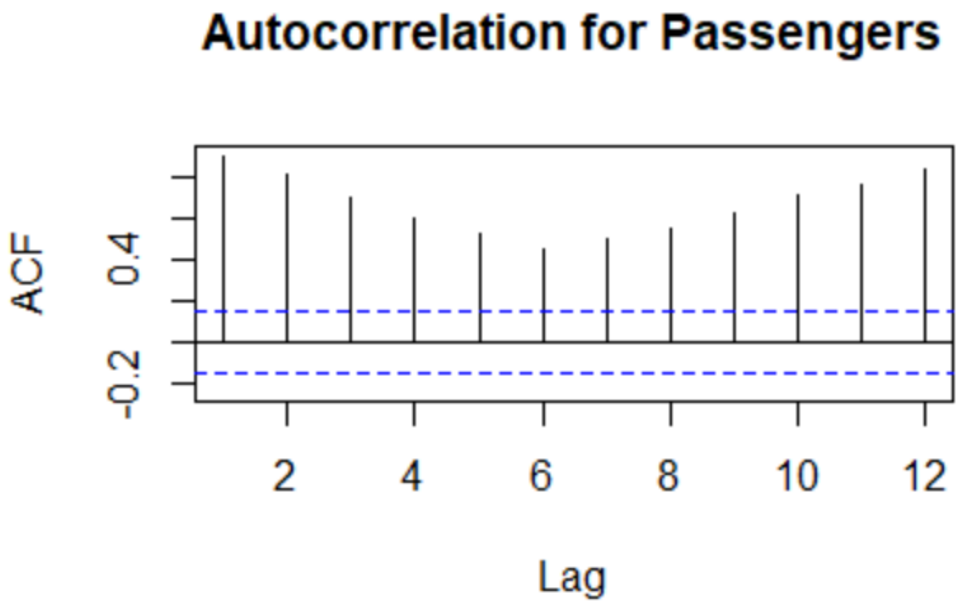
Step 2: Get data

This report will focus on the time series dataset provided by San Francisco International Airport. Though, we relied on the Kaggle Database to extract data from July 2005 to Dec 2020 for the same. The time period for the dataset we have worked on for this project is from Jan 2006 to Dec 2019. Though the original dataset also had data for Cargo and data was given for different airlines but for the purposes of this project we have summed up the data month wise and only considered data for passenger traveling to/from San Francisco International Airport. Also, due to covid-19 which was an exception, we have dropped that period and taken data only till 2019. After doing all this, the total of 168 Data points has been taken into consideration.

Step 3: Explore and Visualize Series



The data plot above shows the respective traffic of air passenger for 168 months series. The time series appear to have an upward trend with multiplicative seasonality. The seasonality is apparent with the low travels at the first and fourth quarters of each year and peak travels at the second and third quarters of each year. The series seems to have more of a linear trend.



From the plot above, we can see that the data's autocorrelation is quite high, as the autocorrelation coefficients in all the lags are substantially higher than the horizontal threshold (significantly greater than zero) states that statistically significant. For the lags, the autocorrelation coefficients have a positive value. A positive autocorrelation coefficient in lag 1 is substantially higher than the horizontal threshold, which is indicative of an upward trend component. A positive autocorrelation coefficient in lag 12,

which is also statistically significant, points to a seasonal component being present in the data.

Identification of time series data Predictability –

```
Series: passenger.ts
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1       mean
      0.9122  3754.4010
s.e.  0.0318   279.0172

sigma^2 estimated as 114622:  log likelihood=-1216.82
AIC=2439.63   AICc=2439.78   BIC=2449.01

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE        ACF1
Training set 11.67264 336.5367 275.8766 -0.4803855 7.495373 1.528565 -0.06077122
```

Hypothesis Testing: Z- Test

Null hypothesis Ho: $\beta 1 = 1$

Alternative hypothesis H1: $\beta 1 \neq 1$

z-statistic = $(\beta 1 - 1)/(s.e.) = (0.9122 - 1)/0.0318 = -2.761006$

p-value for z-statistic = 0.002881178

Based on the p-value of 0.0029 (< 0.05), we can reject the null hypothesis that $\beta 1 = 1$

Therefore, the time series data for SFO Air Passengers, passenger.ts, is predictable (not random walk).

Step 4: Data Preprocessing

The original data from the San Francisco International Airport dataset, located in the "Data 1" tab of the excel spreadsheet. The data of Air traffic passenger was kept and the rest Cargo Stats, was deleted from the original file. The amount of observations per series

are 168 for 13 years series, respectively. Kaggle Database chosen for air traffic of different airlines travelling to/from SF International Airport for the time period July 2005 to Dec 2020 is a good quality database, which does not have any missing values or any data entry errors but we had to sum up the traffic of the airlines month wise to work on this data. Thus, not much pre-processing was required for this database. We have worked on the data starting from Jan 2006 and ending from Dec 2019 for the sake of better forecast results.

Step 5: Partition Series

Partitioning the time series data is an important preliminary step to be considered before applying any forecasting method. The main scope of this step is to divide the time series data into training and validation datasets. We will develop forecasting models based on training data and test the models' performance using the validation data. We created a data partition of 132 records for the training period and 36 records for validation period These partitioned validation and training data sets (2006 -2019). In below R code, train.ts.ad represents train data and valid.ts.ad represents validation data.

Step 6 & 7: Apply Forecasting & Comparing Performance

Regression Based Models

The technique of various regression-based models has been applied on the time-series of SFO air passengers and that, each type of model forecasts different than the other. The regression models are few of the simplest forecasting methods. We have applied the regression model with linear trend, quadratic trend, linear trend and seasonality, quadratic trend and seasonality. Additionally,

we have tried to enhance the model with relatively better accuracy by using two-level models. Two-level models such as, regression model combined with trailing moving average of the residuals, and regression model combined with autoregressive model of the residuals. Accuracies of their performance measures are compared. For each method, we first evaluate the prediction accuracy for the validation data using the forecast model trained using training data. Few of the methods with lower error measures are used to forecast using the entire dataset.

Forecasting for Validation period, using the Training data

On observing the plot for the time-series, we have trend and seasonality present in the data; Based on which we apply 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: -

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

Residuals:
    Min      1Q  Median      3Q     Max
-215.36  -86.57  -13.09   73.50  340.37

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2211.3978    39.4772  56.017  < 2e-16 ***
trend         13.0605     0.2724  47.944  < 2e-16 ***
season2     -253.5151    50.6431  -5.006 1.95e-06 ***
season3      310.1516    50.6453   6.124 1.21e-08 ***
season4      341.8184    50.6489   6.749 5.74e-10 ***
season5      588.1215    50.6540  11.611  < 2e-16 ***
season6      803.4246    50.6606  15.859  < 2e-16 ***
season7      958.5458    50.6687  18.918  < 2e-16 ***
season8      961.3035    50.6782  18.969  < 2e-16 ***
season9      438.3338    50.6892   8.647 2.92e-14 ***
season10     528.0915    50.7016  10.416  < 2e-16 ***
season11     192.0309    50.7155   3.786 0.000241 ***
season12     287.6977    50.7309   5.671 1.01e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 118.8 on 119 degrees of freedom
Multiple R-squared:  0.9679,    Adjusted R-squared:  0.9646
F-statistic: 298.6 on 12 and 119 DF,  p-value: < 2.2e-16
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-233.262  -59.398   -9.929   53.480  289.445

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.361e+03  3.752e+01  62.918  < 2e-16 ***
trend        6.338e+00  8.928e-01   7.099 1.01e-10 ***
I(trend^2)   5.054e-02  6.501e-03   7.774 3.13e-12 ***
season2     -2.530e+02  4.136e+01  -6.118 1.27e-08 ***
season3      3.111e+02  4.136e+01   7.521 1.17e-11 ***
season4      3.430e+02  4.136e+01   8.293 2.03e-13 ***
season5      5.895e+02  4.137e+01  14.252  < 2e-16 ***
season6      8.049e+02  4.137e+01  19.456  < 2e-16 ***
season7      9.601e+02  4.138e+01  23.202  < 2e-16 ***
season8      9.627e+02  4.139e+01  23.262  < 2e-16 ***
season9      4.395e+02  4.139e+01  10.618  < 2e-16 ***
season10     5.290e+02  4.140e+01  12.776  < 2e-16 ***
season11     1.925e+02  4.142e+01   4.649 8.77e-06 ***
season12     2.877e+02  4.143e+01   6.944 2.21e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 96.99 on 118 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9764
F-statistic:   418 on 13 and 118 DF,  p-value: < 2.2e-16
```

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: -

$$\mathbf{y_t = 2211.398 + 13.061\ t - 253.515\ D_2 + 310.152\ D_3 + \ldots + 192.031\ D_{11} + 287.698\ D_{12}}$$

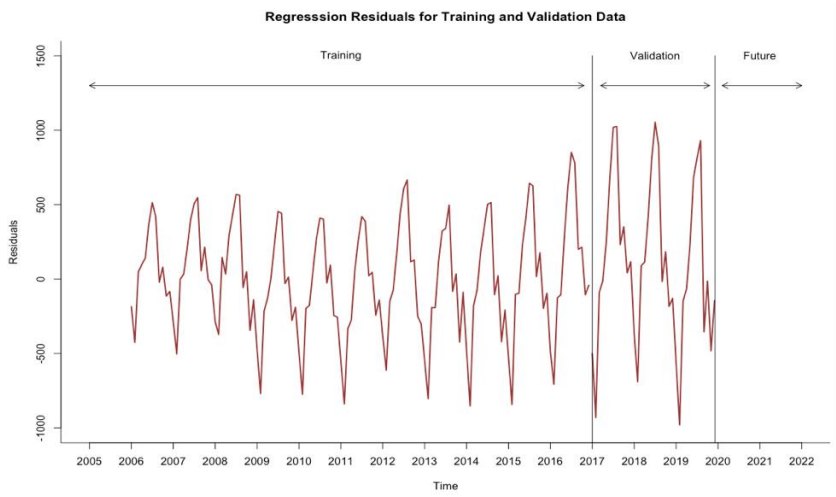
The Regression model with Quadratic Trend and Seasonality contains 13 independent variables: trend index (t), squared trend index (t2), and 11 seasonal dummy variables for respective seasons (season2 – D2, season3 – D3 and so on). The regression model equation is: -

$$\mathbf{y_t = 2360.603 + 6.338\ t + 0.051\ t^2 - 253.010\ D_2 + 311.061\ D_3 + \ldots + 192.536\ D_{11} + 287.698\ D_{12}}$$

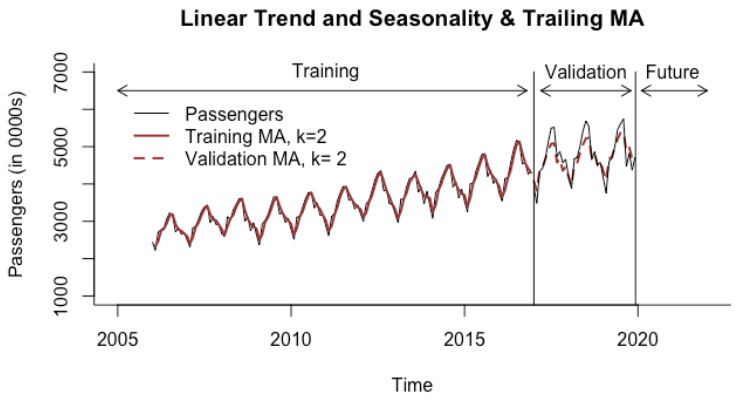
According to the model summary of both the models, all the independent variables (trend and season) for respective model are statistically significant (all their p-values are all much below 0.05 or 0.01). The R-squared and Adjusted R-squared values are high, which represents that the models are a good 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 or 0.01). 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.

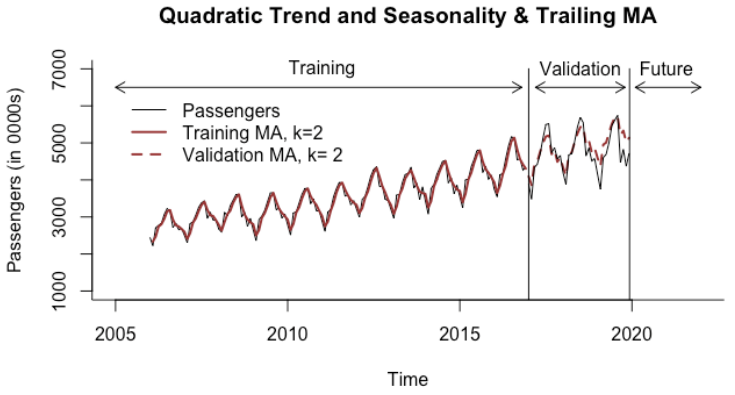
```
> # Accuracy of regression model with linear trend and seasonality (training data)
> round(accuracy(training.lin.trend.seas.pred$mean, validation.ts), 3)
                ME    RMSE     MAE   MPE  MAPE ACF1 Theil's U
Test set 143.046 265.57 214.246 2.588 4.342 0.57     0.516
> # Accuracy of regression model with quadratic trend and seasonality (training data)
> round(accuracy(training.quad.trend.seas.pred$mean, validation.ts), 3)
                ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -145.66 299.095 231.879 -3.524 5.133 0.678     0.625
```

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”.



If a model is adequate, there shouldn’t be any obvious patterns or systematic structure of residuals. If the plots are not evenly distributed vertically, they have an outlier, or they have a shape to them. As we can see a clear pattern / trend in our residuals plot, thus, the model has room for improvement. Further, we will be enhancing the regression models.





Above two plots show us the combined model forecast for validation period; 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 the both the graphs we see that forecast models are overall fitting well with the validation data. On comparing, first combined forecast model (regression model with Linear Trend and Seasonality + Trailing Moving Average of its residuals) looks to fit better than the other.

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

```
> # Accuracy of regression model with linear trend and seasonality (training data)
> round(accuracy(training.lin.trend.seas.pred$mean, validation.ts), 3)
                ME   RMSE     MAE   MPE  MAPE ACF1 Theil's U
Test set 143.046 265.57 214.246 2.588 4.342 0.57     0.516
> # Accuracy of two-level Model
> # regression model with linear trend & seasonality and trailing MA for residuals (training data)
> round(accuracy(valid.forecast.2level.linTS.ma, validation.ts), 3)
              ME    RMSE     MAE  MPE  MAPE  ACF1 Theil's U
Test set 57.089 232.676 186.571 0.74 3.884 0.571     0.457
> # Accuracy of regression model with quadratic trend and seasonality (training data)
> round(accuracy(training.quad.trend.seas.pred$mean, validation.ts), 3)
                ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -145.66 299.095 231.879 -3.524 5.133 0.678     0.625
> # Accuracy of two-level Model
> # regression model with quadratic trend & seasonality and trailing MA for residuals (training data)
> round(accuracy(valid.forecast.2level.quadTS.ma, validation.ts), 3)
                 ME  RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -141.937 297.3 230.638 -3.445 5.101 0.678     0.621
```

Based on the values of RMSE and MAPE, the two-level model (regression model with Linear Trend and Seasonality + Trailing Moving Average of its residuals) 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:

```
Series: training.lin.trend.seas$residuals
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1     mean
      0.7889   9.2927
s.e.  0.0546  28.4709

sigma^2 estimated as 5064:  log likelihood=-749.76
AIC=1505.51   AICc=1505.7   BIC=1514.16

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE     MASE       ACF1
Training set -1.929432 70.62284 54.76192 0.2090481 145.917 0.553718 -0.1389355
> z.stat <- (0.7889 - 1)/0.0546
> p.val <- pnorm(z.stat)
> p.val
[1] 5.524942e-05
```

```
Series: training.lin.trend.seas$residuals
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1     ar2     mean
      0.6619  0.1674  14.3061
s.e.  0.0853  0.0873  34.5112

sigma^2 estimated as 4963:  log likelihood=-747.95
AIC=1503.9   AICc=1504.21   BIC=1515.43

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE         ACF1
Training set -2.46304 69.64021 53.46016 10.74294 137.3236 0.5405554 -0.003031787
> z.stat <- (0.6619 - 1)/0.0853
> p.val <- pnorm(z.stat)
> p.val
[1] 3.690504e-05
```

Based on the coefficients of both the models, we observed that 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:

	Passenger.Valid	Reg.LinTS.Forecast	Combined.Forecast.AR(2)
1	3898	3948.450	4055.643
2	3481	3707.995	3801.458
3	4335	4284.723	4366.969
4	4426	4329.450	4401.974
5	4698	4588.814	4653.025
6	5134	4817.177	4874.260
7	5497	4985.359	5036.331
8	5517	5001.177	5046.912
9	4736	4491.268	4532.514
10	4869	4594.086	4631.484
11	4573	4271.086	4305.186
12	4661	4379.814	4411.086
13	4190	4105.177	4134.025
14	3882	3864.722	3891.493
15	4674	4441.449	4466.440
16	4713	4486.177	4509.641
17	5026	4745.540	4767.697
18	5427	4973.904	4994.939
19	5693	5142.086	5162.159
20	5546	5157.904	5177.154
21	4649	4647.995	4666.539
22	4862	4750.813	4768.751
23	4509	4427.813	4445.233
24	4576	4536.540	4553.515
25	4157	4261.903	4278.497
26	3753	4021.448	4037.715
27	4599	4598.176	4614.163
28	4693	4642.903	4658.650
29	5008	4902.267	4917.808
30	5467	5130.630	5145.995
31	5612	5298.812	5314.026
32	5742	5314.630	5329.714
33	4471	4804.721	4819.694
34	4825	4907.539	4922.417
35	4370	4584.539	4599.335
36	4721	4693.267	4707.993

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:

```
Series: training.quad.trend.seas$residuals
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1     mean
      0.6700   0.8860
s.e.  0.0639  17.6241

sigma^2 estimated as 4673:  log likelihood=-744.27
AIC=1494.54   AICc=1494.72   BIC=1503.18

Training set error measures:
                     ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set -0.5700347 67.84308 52.98653 85.39124 149.6926 0.5498048 -0.07759028
> z.stat <- (0.6700 - 1)/0.0639
> p.val <- pnorm(z.stat)
> p.val
[1] 1.206578e-07
```

```
Series: training.quad.trend.seas$residuals
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1     ar2     mean
      0.5901  0.1187   1.4012
s.e.  0.0859  0.0862  19.7207

sigma^2 estimated as 4642:  log likelihood=-743.33
AIC=1494.66   AICc=1494.97   BIC=1506.19

Training set error measures:
                     ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set -0.7360019 67.35489 52.16582 85.35996 146.3344 0.5412889 0.01328572
> z.stat <- (0.5901 - 1)/0.0859
> p.val <- pnorm(z.stat)
> p.val
[1] 9.128082e-07
```

Based on the coefficients of both the models, we observed that 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:

	Passenger.Valid	Reg.QuadTS.Forecast	Combined.Forecast.(AR1)
1	3898	4097.655	4082.793
2	3481	3864.479	3854.814
3	4335	4448.484	4442.302
4	4426	4500.490	4496.640
5	4698	4767.132	4764.845
6	5134	5002.774	5001.534
7	5497	5178.234	5177.696
8	5517	5201.330	5201.262
9	4736	4698.699	4698.946
10	4869	4808.796	4809.254
11	4573	4493.074	4493.673
12	4661	4609.080	4609.773
13	4190	4342.327	4343.085
14	3882	4110.364	4111.164
15	4674	4695.583	4696.411
16	4713	4748.801	4749.649
17	5026	5016.656	5017.516
18	5427	5253.511	5254.380
19	5693	5430.184	5431.059
20	5546	5454.494	5455.372
21	4649	4953.076	4953.957
22	4862	5064.386	5065.268
23	4509	4749.877	4750.761
24	4576	4867.096	4867.980
25	4157	4601.556	4602.441
26	3753	4370.806	4371.691
27	4599	4957.238	4958.123
28	4693	5011.669	5012.555
29	5008	5280.737	5281.623
30	5467	5518.805	5519.691
31	5612	5696.692	5697.578
32	5742	5722.214	5723.100
33	4471	5222.009	5222.895
34	4825	5334.532	5335.418
35	4370	5021.236	5022.122
36	4721	5139.668	5140.554

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

```
> # Accuracy of regression model with linear trend and seasonality (training data)
> round(accuracy(training.lin.trend.seas.pred$mean, validation.ts), 3)
              ME   RMSE     MAE   MPE  MAPE ACF1 Theil's U
Test set 143.046 265.57 214.246 2.588 4.342 0.57     0.516
> # Accuracy of two-level regression model with linear trend & seasonality and
> # trailing MA for residuals (training data)
> round(accuracy(valid.forecast.2level.linTS.ma, validation.ts), 3)
             ME    RMSE     MAE  MPE  MAPE  ACF1 Theil's U
Test set 57.089 232.676 186.571 0.74 3.884 0.571     0.457
> # Accuracy of regression model with linear trend (training data) and
> # AR(2) model for residuals (training data)
> round(accuracy(valid.2level.linTS.ar2.pred, validation.ts), 3)
              ME    RMSE     MAE   MPE  MAPE ACF1 Theil's U
Test set 110.821 252.065 201.595 1.875 4.126 0.57      0.49
> # Accuracy of regression model with quadratic trend and seasonality (training data)
> round(accuracy(training.quad.trend.seas.pred$mean, validation.ts), 3)
              ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -145.66 299.095 231.879 -3.524 5.133 0.678     0.625
> # Accuracy of regression model with quadratic trend (training data) and
> # AR(1) model for residuals (training data)
> round(accuracy(quad_valid.two.level.pred, validation.ts), 3)
              ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set -145.22 298.849 231.189 -3.51 5.114 0.679     0.625
> # Accuracy of Seasonal Naive forecast model (training data)
> round(accuracy(training.snaive.pred$mean, validation.ts), 3)
              ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set 324.306 365.045 331.861 6.674 6.865 0.317     0.758
```

On comparing the accuracies of various models as shown above, relatively lowest RMSE value is 232.676 and relatively lowest MAPE is 3.884 %. Hence, the Two-Level Model using the Regression model with Linear Trend and Seasonality + the Trailing Moving Average of it Residuals is the best model.

The top three single or multiple Regression Models are as shown in the below table:

Rank	**Model**	**MAPE Value**	**RMSE Value**
#1	Two-Level Regression Model (Regression Model with Linear Trend and Seasonality + Trailing Moving Average Model of its Residuals)	3.884	232.676

#2	Two-Level Regression Model (Regression Model with Linear Trend and Seasonality + Auto-Regressive (order 2) Model of its Residuals)	4.126	252.065
#3	Regression Model with Linear Trend and Seasonality	4.342	265.57

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.

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

Residuals:
    Min      1Q  Median      3Q     Max
-389.71  -91.95  -14.13   79.46  399.95

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2143.2887    44.7163  47.931  < 2e-16 ***
trend         13.8589     0.2414  57.417  < 2e-16 ***
season2     -283.4303    57.2010  -4.955 1.88e-06 ***
season3      333.8536    57.2025   5.836 3.02e-08 ***
season4      371.1375    57.2051   6.488 1.11e-09 ***
season5      625.3501    57.2087  10.931  < 2e-16 ***
season6      883.4911    57.2132  15.442  < 2e-16 ***
season7     1057.0608    57.2188  18.474  < 2e-16 ***
season8     1055.8447    57.2255  18.451  < 2e-16 ***
season9      430.7001    57.2331   7.525 4.01e-12 ***
season10     547.6269    57.2417   9.567  < 2e-16 ***
season11     201.1252    57.2514   3.513 0.000581 ***
season12     308.8377    57.2621   5.393 2.54e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 151.3 on 155 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9642
F-statistic: 375.8 on 12 and 155 DF,  p-value: < 2.2e-16
```

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: -

$$\mathbf{y_t = 2143.289 + 13.859\ t - 283.430\ D_2 + 333.854\ D_3 + \ldots + 201.125\ D_{11} + 308.838\ D_{12}}$$

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 (0.9668) and Adjusted R-squared value (0.9642) are high, and all the regression coefficients are statistically significant (p-value < 0.05 or 0.01). Overall, the regression model is very good fit for the historical dataset and thus, can be used for forecasting SFO's air passengers count.

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

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2020	4311.302	4041.730	4672.873	4724.016	4992.087	5264.087	5451.516	5464.159	4852.873	4983.659	4651.016	4772.587

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): -

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2020	4422.146	4169.970	4812.258	4872.270	5147.165	5424.440	5615.944	5631.736	5022.883	5155.549	4824.358	4947.052

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

```
> # Accuracy of regression model with linear trend and seasonality(entire data)
> round(accuracy(passenger.lin.trend.seas.pred$fitted, passenger.ts), 3)
         ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set  0 145.365 111.468 -0.022 3.048 0.707      0.42
> # Accuracy of regression model with linear trend and seasonality (entire data) and
> # trailing MA for residuals (entire data)
> round(accuracy(passenger.lin.trend.seas.pred$fitted +
+                   ap.ma.trail.lin.trend.seas.res.pred$fitted, passenger.ts), 3)
             ME    RMSE    MAE    MPE  MAPE  ACF1 Theil's U
Test set -3.973 110.512 80.412 -0.154 2.161 0.234     0.301
> # Accuracy of regression model with linear trend and seasonality (entire data) and
> # AR(2) model for residuals (entire data)
> round(accuracy(passenger.lin.trend.seas.pred$fitted +
+                   passenger.lin.trend.seas.res.ar2.pred$fitted, passenger.ts), 3)
             ME    RMSE    MAE    MPE  MAPE  ACF1 Theil's U
Test set -2.124 101.103 74.336 -0.079 2.025 0.001     0.274
> # Accuracy of seasonal naive forecast (baseline model)
> round(accuracy((snaive(passenger.ts))$fitted, passenger.ts), 3)
              ME   RMSE     MAE   MPE  MAPE ACF1 Theil's U
Test set 154.391 209.22 180.481 4.008 4.754 0.72      0.57
```

Based on the accuracy performance measures, comparatively lowest MAPE value is 2.025 % and lowest RMSE value is 101.103. 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.

Arima and auto-arima:

ARIMA is an acronym for “autoregressive integrated moving average.” It's a model used in statistics and econometrics to measure events that happen over a period of time. The model is used to understand past data or predict future data in a series.It is a Complex structure, may include up to 6 parameters to be identified for each model (unless using an automated option).ARIMA (p, d, q) model is used to forecast data with level and trend components – non-seasonal ARIMA model.

- *p* = order of autoregressive model *AR(p)* – number of autocorrelation lags included
- *d* = order of differencing in *AR* model – indicates how many rounds of lag-1 differencing are performed to remove certain trend
- *q* = order of moving average *MA(q)* – number of residuals' autocorrelation lags included

ARIMA (p, d, q) (P, D, Q)$_m$ model is used to forecast data with *level, trend, and seasonality components. In addition to the (p, d, q) parameters, it includes seasonal parameters:*

- *P = order of autoregressive seasonal model AR(P) – number of autocorrelation lags included*
- *D = order of differencing in AR seasonal model – indicates how many rounds of lag-1 differencing the are performed to remove certain trend*
- *Q = order of moving average MA(Q) – number of residuals' autocorrelation lags included*
- *m = number of seasons*

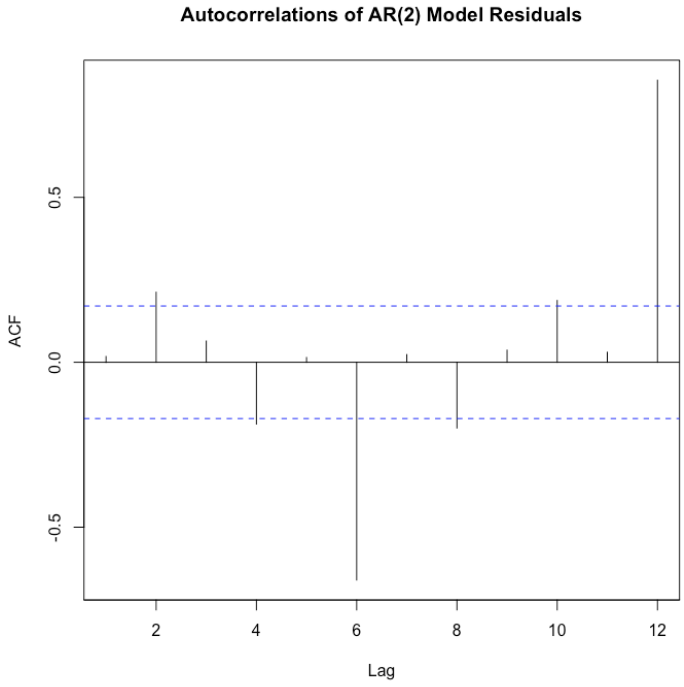
The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary. It is easier to predict when the series is stationary. Differencing is a method of transforming a non-stationary time series into a stationary one.

ARIMA(2,0,0) function is first used to fit the AR(2) model for the training data set. The equation for the model is obtained as below.

$y_t = 3493.4 + 0.85\, y_{t-1} - 0.050\, y_{t-2}$

The model is then used to forecast the data for validation data set. The autocorrelation for the residuals is as given below.

We can see that significant relationships exist but are less as compared to the other forecasting models.

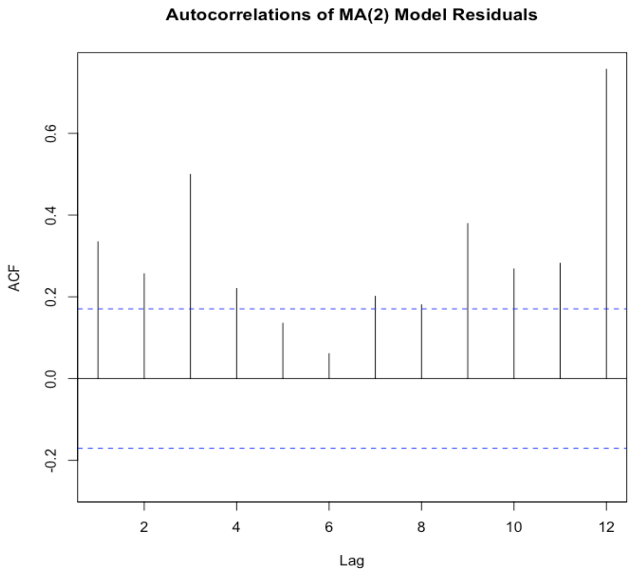


The components are highly above the threshold values for lags 2,6 and 12. Hence there exists autocorrelation among them.

Now the ARIMA(0,0,2) function is used to fit the MA(2) model for the training data set.In this model, the model equation we obtained is as below.

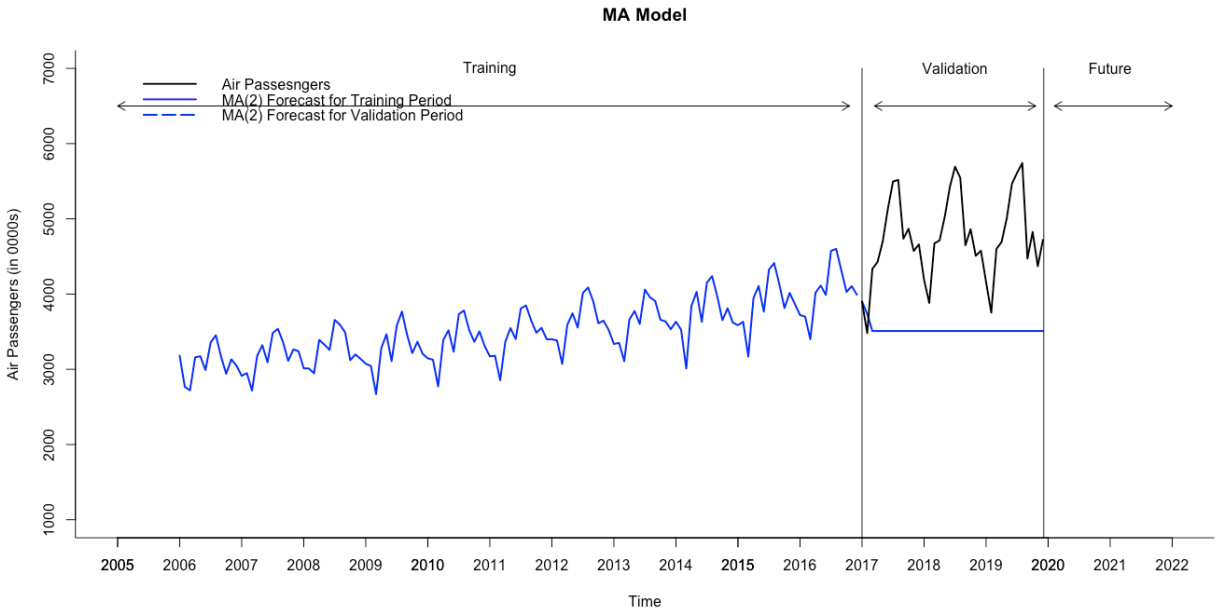
$y_t = 3508.95 + 0.823\, \varepsilon_{t-1} + 0.644\, \varepsilon_{t-2}$

The ACF plot for the above MA model of ARIMA is as below.



Here we can see high levels of components going beyond the threshold values and hence the autocorrelation exists for this plot of residuals.

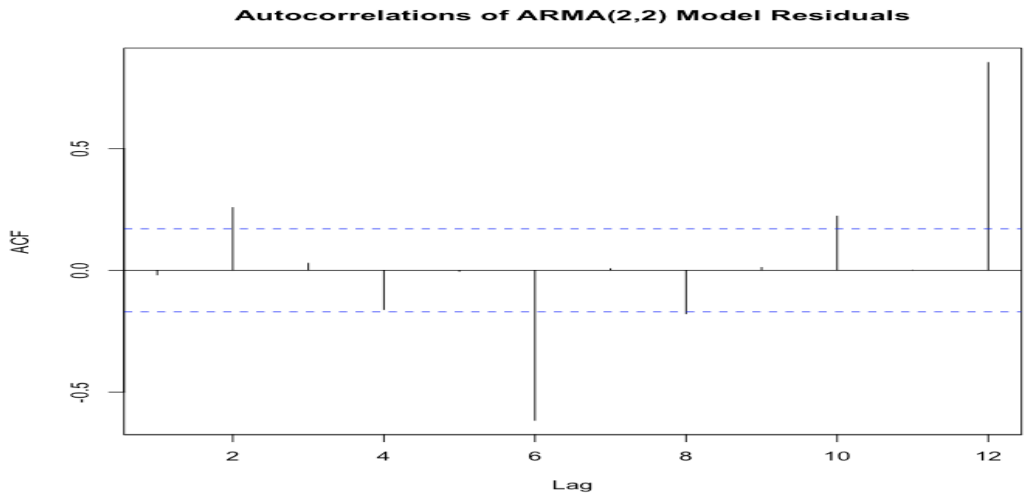
The plot for the MA(2) model is as below.



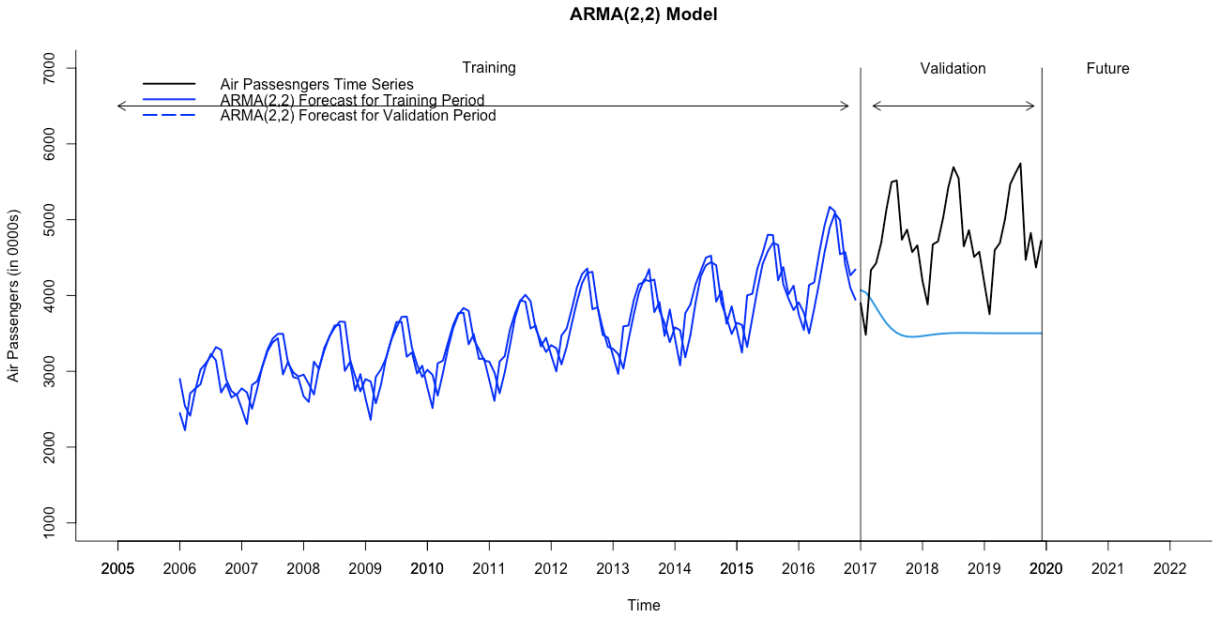
Arima() function is used to fit the ARMA(2,2) model.The ARIMA model of order = c(2,0,2) gives an ARMA(2,2) model.Applying this to the training data and predicting the values for validation data set.The model equation for this model obtained is as below.

$$y_t = 3501.95 + 1.47\, y_{t-1} - 0.60\, y_{t-2} - 0.74\, \varepsilon_{t-1} + 0.56\, \varepsilon_{t-2}$$

The correlogram for this model obtained is as below.



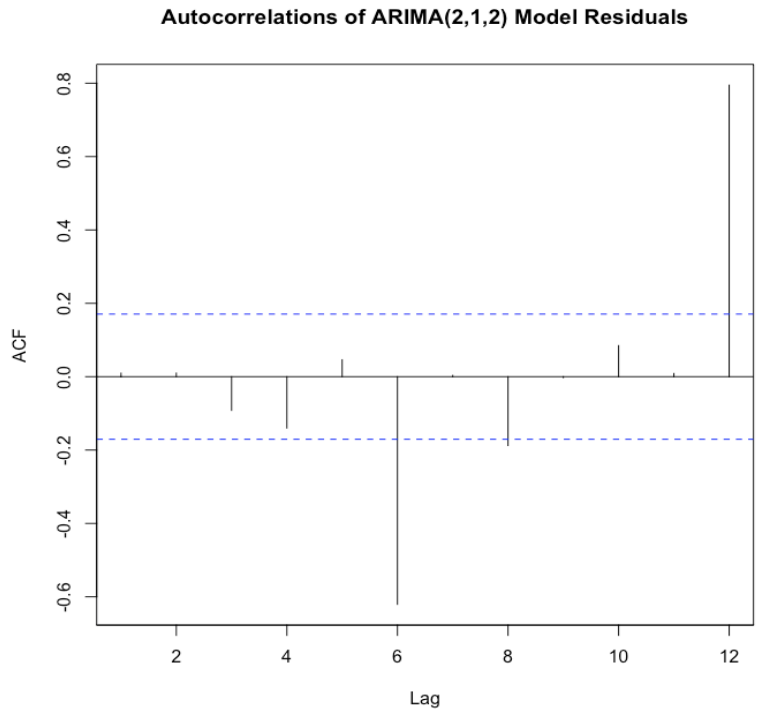
The plot for this model is as below:



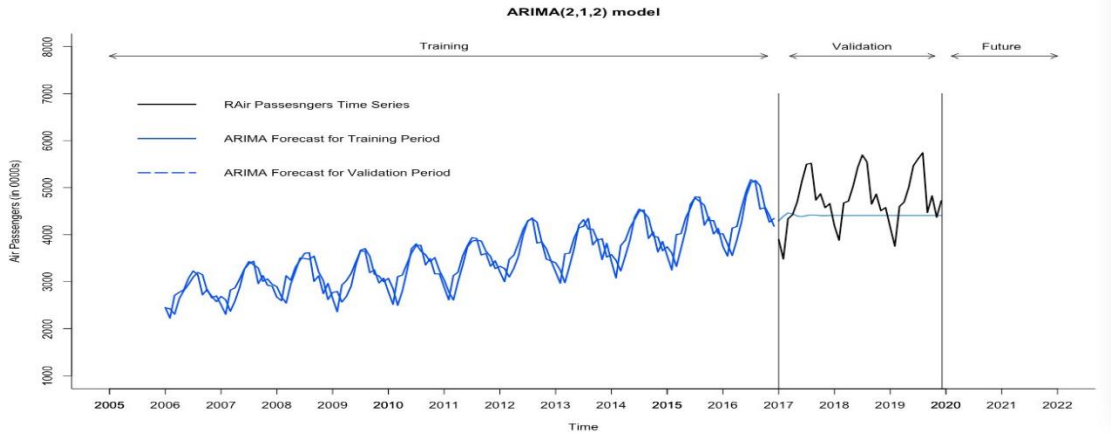
Using the Arima() function to fit the ARIMA(2,1,2) model.The validation data is forecasted and the correlogram for the residuals is obtained.The model equation is also obtained as below.

$$y_t - y_{t-1} = 0.45\ (y_{t-1} - y_{t-2}) - 0.53\ (y_{t-2} - y_{t-3}) - 0.64\ \varepsilon_{t-} - 0.99\ \varepsilon_{t-2}$$

The correlogram of the residuals for this model is below.



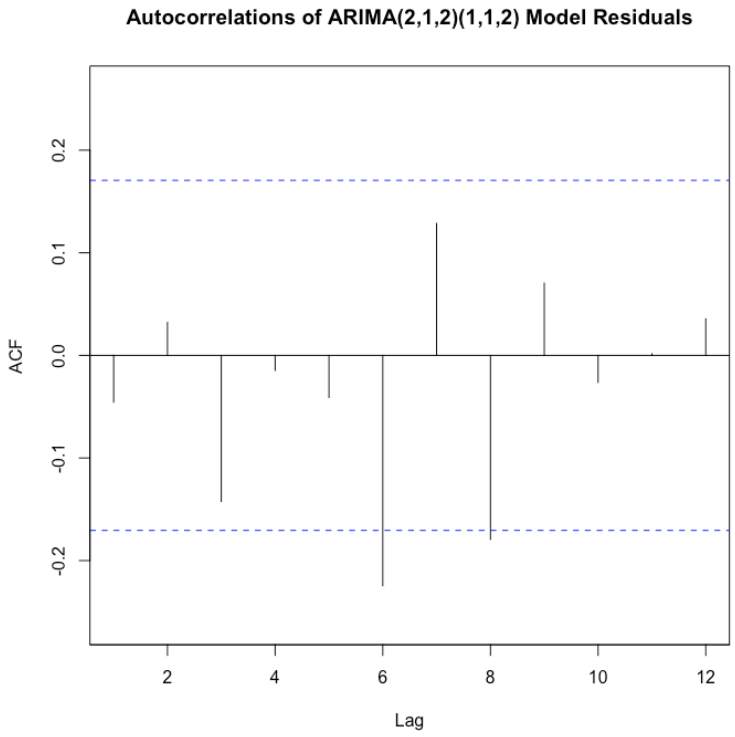
The plot for this data is as below.



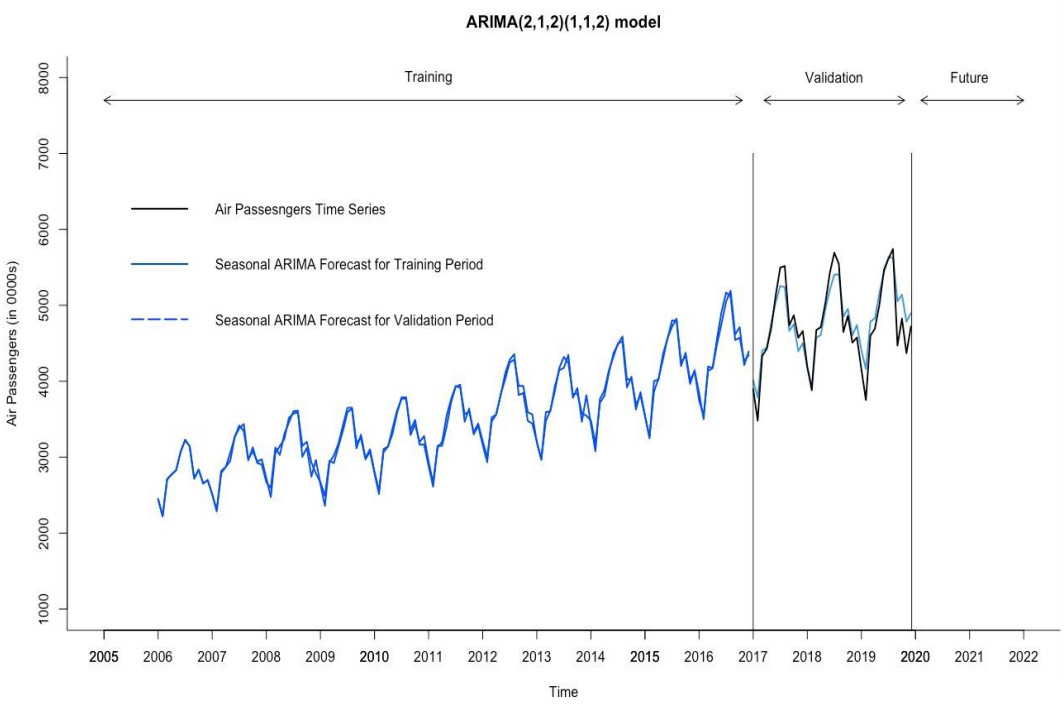
Using the Arima() function to fit the ARIMA(2,1,2)(1,1,2) model for trend and seasonality. The model equation is obtained as below.

$$y_t - y_{t-1} = 0.142\,(y_{t-1} - y_{t-2}) - 0.59\,(y_{t-2} - y_{t-3}) - 0.42\,\varepsilon_{t-1} + 0.6\,\varepsilon_{t-2} - 0.88\,(y_{t-1} - y_{t-13}) - 0.62\rho_{t-2} + 0.32\rho_{t-1}$$

The plot for residuals obtained for the validation data set is as below.



The plot for this model is as below.



ARIMA models are rather complex with a number of parameters involved and complex relationships between the model parts.It is hard (but not impossible) to clearly identify what specific parameters should be used in the model.

Visualizing the historical data and Applying ACF and PACF (partial autocorrelation function) charts, may not produce optimal results.

auto.arima() function in R is used to identify the optimal ARIMA model and its respective *(p, d, q)(P, D, Q)* parameters. It does not require to input any of these parameters into the function

Using the auto ARIMA model,the model is first used for trained data and predicted values for validation data set is obtained.The model obtained is as below.

```
Series: train.ts
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:
          ma1     sma1
      -0.3500  -0.5855
s.e.   0.0894   0.0918

sigma^2 estimated as 6238:  log likelihood=-690.35
AIC=1386.71   AICc=1386.92   BIC=1395.04

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE         ACF1
Training set 3.6942 74.35593 56.00482 0.03337747 1.604804 0.3081421 -0.00923631
```

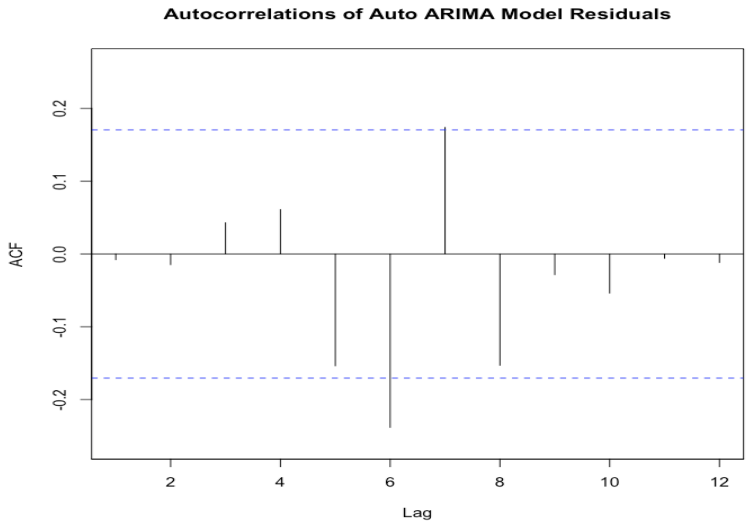
The ARIMA model(0,1,1)(0,1,1)(12) is obtained.It means

- ***p = 0,* order 0 autoregressive model *AR(2)-means no autoregressive***

- ***d = 1*, order 1 differencing to remove linear trend**
- ***q = 1*, order 2 moving average *MA(2)* for error lags**
- ***P = 0*, order 0 autoregressive model *AR(1) for seasonality-means no auto regression***
- ***D = 1*, order 1 differencing to remove linear trend**
- ***Q = 1*, order 2 moving average *MA(2)* for error lags**
- ***m = 12*, for monthly seasonality**
- The model equation isa s below.

$$y_t - y_{t-1} = -0.35\ \varepsilon_{t-1} - 0.58\ \rho_{t-1}$$

The plot for residuals is obtained as below.



The accuracy measures for all the models of ARIMA used till now to predict the validation data set are as below.

```
> round(accuracy(train.ar2.pred, valid.ts), 3)
                    ME     RMSE      MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set   12.453  291.580  247.104 -0.347  7.273 1.360 0.007        NA
Test set     1038.150 1224.734 1096.982 20.555 22.185 6.036 0.674     2.418
> round(accuracy(train.ma2.pred, valid.ts), 3)
                    ME     RMSE      MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set    3.151  345.512  279.149 -1.394  8.314 1.536 0.257        NA
Test set     1223.561 1358.610 1237.922 24.606 25.018 6.811 0.617     2.706
> round(accuracy(train.arma2.pred, valid.ts), 3)
                    ME     RMSE      MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set    5.490  272.625  212.751 -0.497  6.351 1.171 0.053        NA
Test set     1189.935 1350.140 1230.557 23.789 24.927 6.771 0.659     2.677
> round(accuracy(train.arima.pred, valid.ts), 3)
                   ME    RMSE     MAE   MPE   MAPE  MASE ACF1 Theil's U
Training set  11.791 259.958 206.540 0.088  6.189 1.136 0.01        NA
Test set     343.677 650.173 518.045 5.923 10.520 2.850 0.62     1.265
> round(accuracy(train.arima.seas.pred, valid.ts), 3)
                   ME    RMSE     MAE    MPE  MAPE  MASE   ACF1 Theil's U
Training set   2.877  72.344  55.211  0.031 1.575 0.304 -0.047        NA
Test set     -43.688 208.630 168.135 -1.253 3.648 0.925  0.592     0.426
> round(accuracy(train.auto.arima.pred, valid.ts), 3)
                   ME    RMSE     MAE    MPE  MAPE  MASE   ACF1 Theil's U
Training set   3.694  74.356  56.005  0.033 1.605 0.308 -0.009        NA
Test set     -98.778 228.986 176.477 -2.403 3.885 0.971  0.611     0.471
```

We can say that the RMSE and MAPE values for ARIMA(2,1,2)(1,1,2) model and Auto ARIMA model are less when compared to all other models. These are the 2 models with least MAPE values. Among these 2, The ARIMA(2,1,2)(1,1,2) model has the least values for RMSE and MAPE as well; 72.34 and 1.57 respectively. Hence ARIMA(2,1,2)(1,1,2) model is the best fit for validation data set prediction.

Now applying the auto ARIMA model for the entire data set, we obtain the prediction for future 12 periods as below.

	Point Forecast	Lo 0	Hi 0
Jan 2020	4218.612	4218.612	4218.612
Feb 2020	3881.506	3881.506	3881.506
Mar 2020	4715.446	4715.446	4715.446
Apr 2020	4806.439	4806.439	4806.439
May 2020	5130.180	5130.180	5130.180
Jun 2020	5578.317	5578.317	5578.317
Jul 2020	5775.865	5775.865	5775.865
Aug 2020	5829.579	5829.579	5829.579
Sep 2020	4703.602	4703.602	4703.602
Oct 2020	5007.992	5007.992	5007.992
Nov 2020	4595.417	4595.417	4595.417
Dec 2020	4856.321	4856.321	4856.321

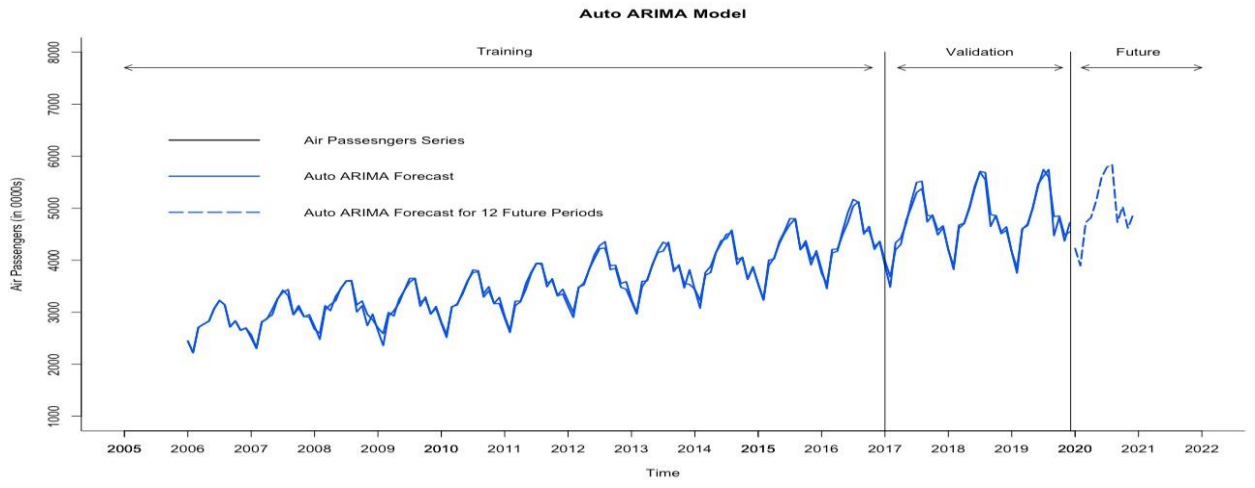
The low and high confidence intervals are showing the same values as we have considered the level=0 for this.

Applying Auto arima with optimal parameters and values; the model we obtained is below.

```
Series: airpass.ts
ARIMA(2,0,0)(0,1,1)[12] with drift

Coefficients:
         ar1     ar2     sma1    drift
      0.5261  0.2799  -0.3405  12.7279
s.e.  0.0778  0.0795   0.0827   2.0526

sigma^2 estimated as 8249:  log likelihood=-923.86
AIC=1857.73   AICc=1858.13   BIC=1872.98

Training set error measures:
                    ME     RMSE      MAE          MPE     MAPE      MASE
Training set 1.009497 86.39025 63.04943 -0.06090535 1.683112 0.3493416
                     ACF1
Training set -0.0003117407
```

We can see that ARIMA(2,0,0)(0,1,1)[12] has been obtained for the entire data set and the MAPE and RMSE values are 1.6 and 86.39 respectively. The ACF1 is the very least for this model;it means the autocorrelation is best taken care of for this model.

The plot for the entire data is as below.



When the accuracy measures are compared for the entire data set with other forecasting models, the auto ARIMA is the best suitable model for this data set. The results for accuracy measures of different forecasting models is as below.

```
> # (1) Seasonal ARIMA (2,1,2)(1,1,2) Model
> round(accuracy(arima.seas.pred$fitted, airpass.ts), 3)
             ME RMSE    MAE   MPE MAPE   ACF1 Theil's U
Test set 3.616 86.4 62.451 0.037 1.66 -0.003     0.241
> # (2) Auto ARIMA Model
> round(accuracy(auto.arima.pred$fitted, airpass.ts), 3)
             ME  RMSE    MAE    MPE  MAPE ACF1 Theil's U
Test set 1.009 86.39 63.049 -0.061 1.683    0     0.242
> # (3) Seasonal naive forecast
> round(accuracy((snaive(airpass.ts))$fitted, airpass.ts), 3)
              ME   RMSE     MAE   MPE  MAPE ACF1 Theil's U
Test set 154.391 209.22 180.481 4.008 4.754 0.72      0.57
> # (4) Naive forecast
> round(accuracy((naive(airpass.ts))$fitted, airpass.ts), 3)
              ME    RMSE     MAE    MPE  MAPE   ACF1 Theil's U
Test set 13.605 343.662 272.754 -0.032 7.405 -0.109         1
```

We can see that among all the models; the auto arima model has the least values for MAPE and RMSE. Hence we can conclude that the auto arima model is the most suitable model for this data set.

Holt-Winter's Model

Holt-Winter's is one of the advanced exponential Smoothing models which is ideal for data set that has trend and seasonal variation. The SFO Air passengers' historical data set has seasonality and trend components. Hence, we developed the Holt-Winters Model with the automated selection of error, trend and seasonality option. 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-Winter's Model on Training Partition

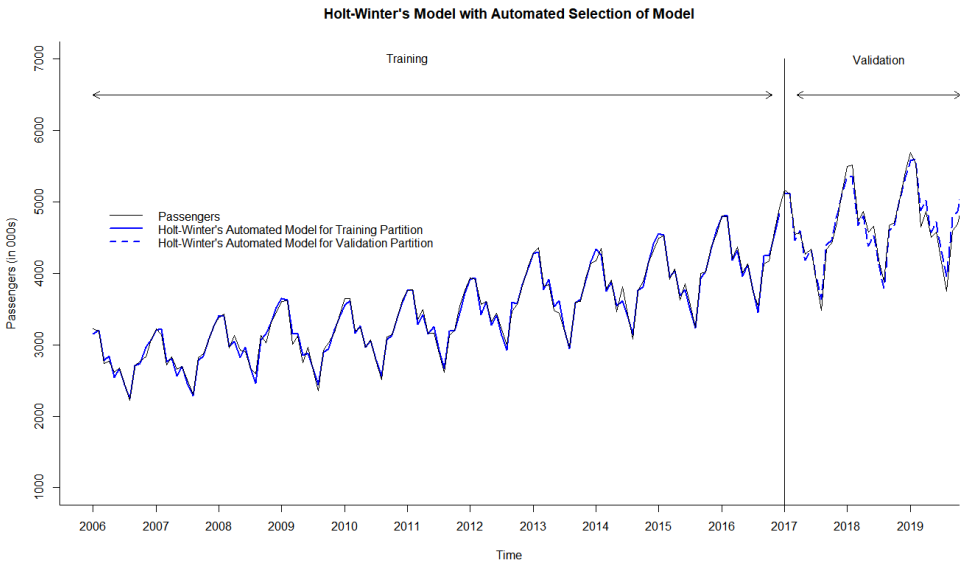
The automated selection of error, trend and seasonality option of Holt's model on the training partition was applied and below is the output of the model's summary.

```
ETS(M,A,M)

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

  Smoothing parameters:
    alpha = 0.5529
    beta  = 0.0179
    gamma = 1e-04

  Initial states:
    l = 2720.3463
    b = 7.2272
    s = 1.1084 1.0467 0.9756 0.9653 0.7996 0.8727
           0.9595 0.9326 1.0279 1.002 1.1546 1.1552

  sigma:  0.0203

     AIC     AICc      BIC
1776.892 1782.260 1825.899
```

The summary show that the optimal model for the training data set is (M, A, M) which indicates multiplicative error, additive trend ad multiplicative seasonality. As display in the output summary, the optimal value for exponential smoothing constant (alpha) is 0.5529 and smoothing constant for trend estimate (beta) is 0.0179, 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 and the 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	262.16	4.29
Seasonal Naïve	365.05	6.87

Naïve	688.589	11.26

Upon comparing the automated Holt's model with Seasonal Naïve and Naïve, it has the lowest RMSE (262.16) and MAPE (4.29%).

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

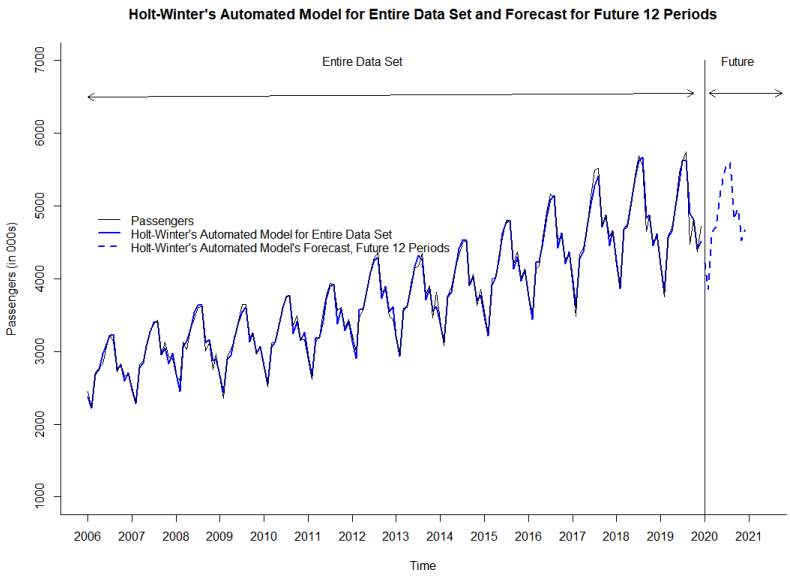
The model performed well in forecasting using the validation period. It needs to be re-run on the entire data set before we use it to forecast for the future period. The following summary output was obtained when the model was used on the entire data set.

```
ETS(M,A,M)

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

  Smoothing parameters:
    alpha = 0.5582
    beta  = 0.0178
    gamma = 5e-04

  Initial states:
    l = 2717.479
    b = 7.5507
    s = 1.1106 1.0428 0.9756 0.9628 0.7973 0.8735
           0.961 0.9379 1.0245 0.9999 1.1554 1.1588

  sigma:  0.0195

     AIC     AICc      BIC
2309.030 2313.110 2362.138
```

Similar to output for the model on the training data set, this HW model has the (M, A, M) options which indicates multiplicative error, additional trend, and multiplicative seasonality. The optimal value for exponential smoothing constant (alpha) is 0.5582, suggesting that the model's level component tends to more local. Both the smoothing constant for trend estimate (beta), and smoothing constant for seasonality estimate (gamma) are close to 1, indicating that the model's multiplicative seasonality is global and the seasonality does not change over time.

The model's forecast points in 12 months of 2020 and the chart for it are displayed below.

	Point Forecast	Lo 0	Hi 0
Jan 2020	5664.689	5664.689	5664.689
Feb 2020	5663.286	5663.286	5663.286
Mar 2020	4914.156	4914.156	4914.156
Apr 2020	5048.677	5048.677	5048.677
May 2020	4634.034	4634.034	4634.034
Jun 2020	4761.196	4761.196	4761.196
Jul 2020	4338.890	4338.890	4338.890
Aug 2020	3970.956	3970.956	3970.956
Sep 2020	4808.132	4808.132	4808.132
Oct 2020	4884.354	4884.354	4884.354
Nov 2020	5234.612	5234.612	5234.612
Dec 2020	5589.571	5589.571	5589.571



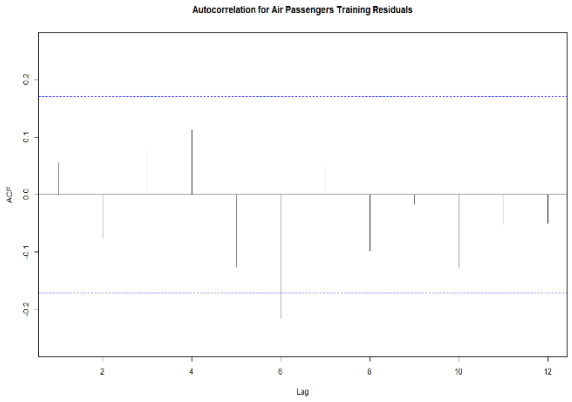
Accuracy measure on the entire dataset

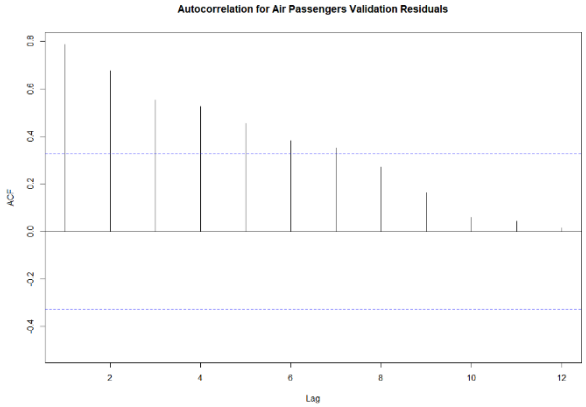
Model	RMSE	MAPE (%)
Automated Holt-Winters	78.12	1.53
Seasonal Naïve	209.22	4.75

The accuracy of the automated Holt-Winters model is having very low RMSE(78.12) and MAPE(1.53%) when compared with Seasonal Naïve model.

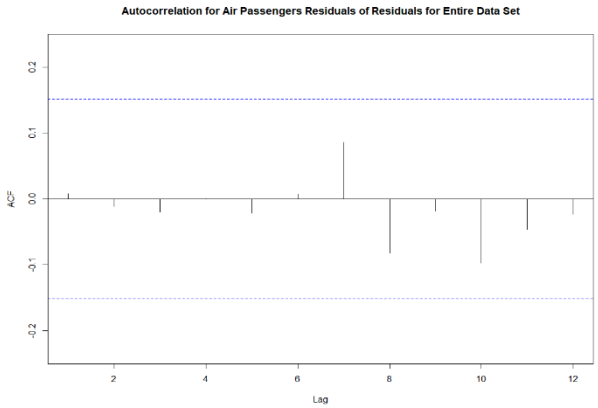
Automated Holt-Winters Model with an Autoregressive, AR(6) Model for Residuals

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





The autocorrelation plot for residual of residuals after Autoregressive model (AR(6)) on the residuals of the automated Holt-Winter's Model is used is showing below. It is evident that the autocorrelation coefficients of the residual of residuals on all the lags are insignificant.



The autocorrelation for the training residuals on almost all the lags are insignificant, except for lag 6 which is not also strong. But the autocorrelation for the validation residuals on lags 1-7 are significant, specially lag 1 is very strong indicating. Hence, AR(6) Model on the residential to incorporate the autocorrelation which were not handled by the Holt-Winter's model. The AR(6) 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 forecast made on the entire data set

using the automated Holt-Winter's Model, The AR(6) for residuals and the combined model(automated Holt-Winter's Model and the AR(6) for residuals of the automated Holt-Winter's Model), all the values are in 1000s.

	HW.Forecast	AR(6)Forecast	Combined.Forecast
1	4213.649	0.005407653	4213.655
2	3852.246	0.002728377	3852.249
3	4650.896	0.002547465	4650.898
4	4713.352	0.002535249	4713.354
5	5041.479	0.002534424	5041.481
6	5372.595	0.002534369	5372.597
7	5581.808	0.002534365	5581.810
8	5591.307	0.002534365	5591.309
9	4814.449	0.002534365	4814.451
10	4957.119	0.002534365	4957.122
11	4529.119	0.002534365	4529.121
12	4666.576	0.002534365	4666.579

The above table indicates that the autoregressive forecast for the residuals are negligible, when the accuracy measure for the automated Holt's model and the combined model (automated Holt's model and the AR(6) for the residuals are same as indicated below.

Automated Holt's Model and Trailing Moving Average for Residuals

Trailing moving average for residuals was also used to incorporate the residuals of the automated Holt-Winter's model on training partition and the entire data set. Forecast was also made but the forecast for the residuals are negligency and the accuracy for the automated Holt-Winter's, Combine Model of the Holt-Winter's with autogressive (AR(6) and Trailing Moving Average for residuals for entire data set is as shown below.

Model	RMSE	MAPE (%)
Automated Holt-Winter's	78.12	1.53
Combine Model (HW(ZZZ) +AR (6)	78.12	1.53

Combine Model (HW(ZZZ) +Trailing MA)	78.64	1.54
Seasonal Naïve	209.22	4.75

It can be concluded that the automated Holt-Winter's model is performing well with out adding additional model for the residuals of the model.

Step 8: Implement Forecast

Rank	Best Model Per Methodology	RMSE	MAPE
1	Automated Holt-Winters	78.12	1.53%
2	Seasonal ARIMA (2,1,2)(1,1,2) Model	86.4	1.66%
3	Regression with Linear Trend and Seasonality + AR(2)	101.103	2.025 %

As seen from the above comparison table it is evident that Automated Holt-Winters gives the best predictions with an RMSE and MAPE values of 78.12 and 1.53% respectively. This is the

recommended model to implement when forecasting for Air traffic passengers travelling from/to San Francisco International Airport.

Conclusion:

The model of choice as per the above computations is Automated Holt Winter's Mode which gives the best predictions. Additionally, it can also be observed that Seasonal ARIMA (2,1,2)(1,1,2) Model and Regression with Linear Trend and Seasonality + AR(2) also provides good forecast but the least RMSE and MAPE is something to look for to chose the best model and with this we can forecast with minimum error the kind of air traffic expected travelling to/from the San Francisco International Airport.

Bibliography

https://en.wikipedia.org/wiki/San_Francisco_International_Airport#Passenger

https://www.kaggle.com/rajsengo/sfo-air-traffic-eda#notebook-container