Time Series Analysis on United Air Revenue.

BAN 673_01

Team -

Manisha Boyina

Dileep Lingamallu

Abhinay Parasa

Krushi Teja Reddy Padamati

Executive Summary

The goal of this time series project is to analyze the quarterly revenue data of United Air from 2000 to 2019 and develop a forecasting model to predict future revenue. The dataset includes quarterly revenue for each year from 2000 to 2019.

In the first step of the project we will visualize the data to understand the trend, seasonality, and other patterns. Next, we will use various time series techniques, such as autocorrelation to explore the data and identify the appropriate forecasting model. Different time series models such as Regression-based models, advanced exponential smoothing models and, autoregressive integrated moving average models (ARIMA) were utilized for this project.

The dataset is divided into training and validation sets, where the training set will be used to build and validate the model, and the validation set will be used to evaluate the model's performance. To achieve successful outcomes, additional regression and advanced exponential smoothing models were built. A trailing moving average for residuals and an autoregressive model for residuals were added to the regression models as needed. When necessary, the same improvements were made to the advanced exponential smoothing models. The RMSE and MAPE accuracy metrics were used as the basis for model evaluation.

Finally, the forecasting model is used to predict future quarterly revenue for United Air. The results and findings of the project will be presented in a report that will provide insights to the stakeholders to make better decisions regarding United Air's future revenue.

Introduction

United Airlines is one of the largest airlines in the world and has been providing passenger and cargo transportation services since 1926. The company operates more than 4,900 flights daily to 356 airports across five continents, making it a major player in the global aviation industry.

Time series analysis is an important tool for airlines because it helps them to better understand and forecast demand for their services. Airlines operate in a highly dynamic and complex environment, where demand for air travel is influenced by a variety of factors, such as economic conditions, fuel prices, exchange rates, weather patterns, and geopolitical events. These factors can have a significant impact on airline revenue and profitability, making it essential for airlines to be able to forecast demand and adjust their operations accordingly.

Time series analysis provides airlines with a range of analytical techniques that can be used to identify patterns and trends in historical data, and to forecast future demand based on these patterns. In this project, we will be utilizing regression-based models, advanced exponential smoothing models, and autoregressive integrated moving average models (ARIMA) to analyze the revenue data of United Airlines. As time series data often exhibits patterns and trends that can be difficult to identify with simple statistical analysis, these models will allow us to more accurately forecast future revenue and identify potential factors that influence revenue fluctuations over time. This analysis can provide valuable insights for decision-making and planning for United Airlines.

This, in turn, enables airlines to optimize their capacity and pricing strategies, and to make more informed decisions about route planning, fleet management, and other key operational areas. Ultimately, time series analysis is an important tool for airlines seeking to improve their revenue management and competitiveness in the highly competitive aviation industry.

Eight Steps of Forecasting

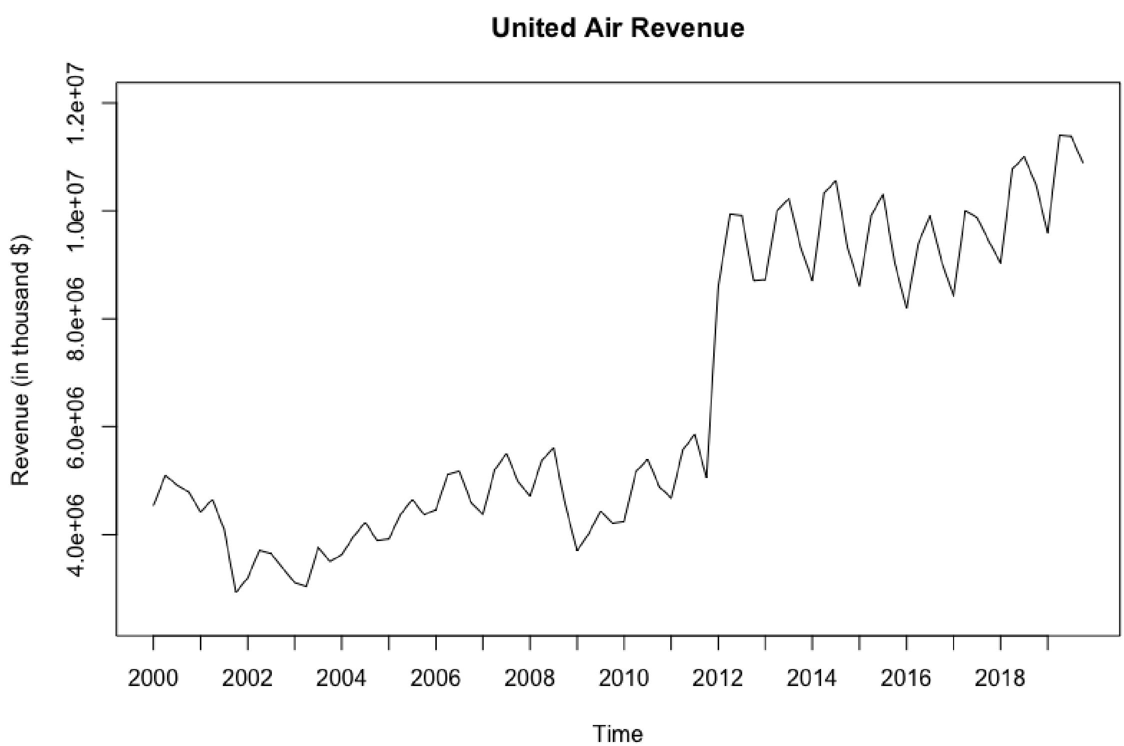
Step 1: Define the Goal

The goal of this project 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 quarters for future periods. Using the revenue data of United Airlines from 2000 to 2019 considering various time series models, the model with the highest accuracy will be considered the model of choice. This analysis can help United Airlines make data-driven decisions and plan for the future, potentially improving revenue and overall business performance. The forecasting models developed for this project were done using the R programming language.

Step 2: Get Data

This report will focus on the time series dataset provided by United Air representing the total quarterly revenues of United Airlines. The time period for the dataset ranges from Q1 of 2000 to Q4 of 2019. The data is measured in $1000. For example, for Q1 of 2000, the revenue of 4532976 means $4,532,976,000 or $4.532 Billion.

Step 3: Explore and Visualize Data



	A	B
1	Quarter_Year	Revenue
2	Q1 2000	4532976
3	Q2 2000	5097939
4	Q3 2000	4915785
5	Q4 2000	4784580
6	Q1 2001	4417662
7	Q2 2001	4648239
8	Q3 2001	4096978
9	Q4 2001	2924490
10	Q1 2002	3195591
11	Q2 2002	3706075
12	Q3 2002	3644641
13	Q4 2002	3369294

The above are the data plots of the United Air Quarterly time series data. Time series appear to have a normal up trend in the beginning years and graph drastically moved upward trend with seasonality in the year 2012 and maintained the constant trend till the end of 2019. The above is the data is used as input for the time series data set. Data set consists of quarter, year and revenue.

Step 4: Data Preprocessing

From the original data from United Air, we choose 10 years quarterly data. Doing so, only the most relevant data will be considered for the analysis. There are a total of 80 data points each year containing 4 quarters. Below is the time series data set of it.

	Qtr1	Qtr2	Qtr3	Qtr4
2000	4532976	5097939	4915785	4784580
2001	4417662	4648239	4096978	2924490
2002	3195591	3706075	3644641	3369294
2003	3107212	3034759	3757353	3498405
2004	3625990	3962484	4225317	3887378
2005	3916450	4373958	4647980	4365766
2006	4459678	5111033	5175873	4587275
2007	4374154	5196411	5504770	4973759
2008	4711208	5370526	5606838	4548676
2009	3693574	4020052	4435164	4210448
2010	4243006	5162802	5395978	4880534
2011	4675619	5570539	5856265	5052795
2012	8603978	9940572	9912178	8703439
2013	8723110	10003024	10229894	9331106
2014	8696283	10328256	10563671	9312309
2015	8608479	9913459	10305905	9036289
2016	8195291	9395665	9913185	9051740
2017	8420115	9999723	9877966	9438506
2018	9031936	10776602	11003046	10491647
2019	9589287	11401369	11380482	10887402

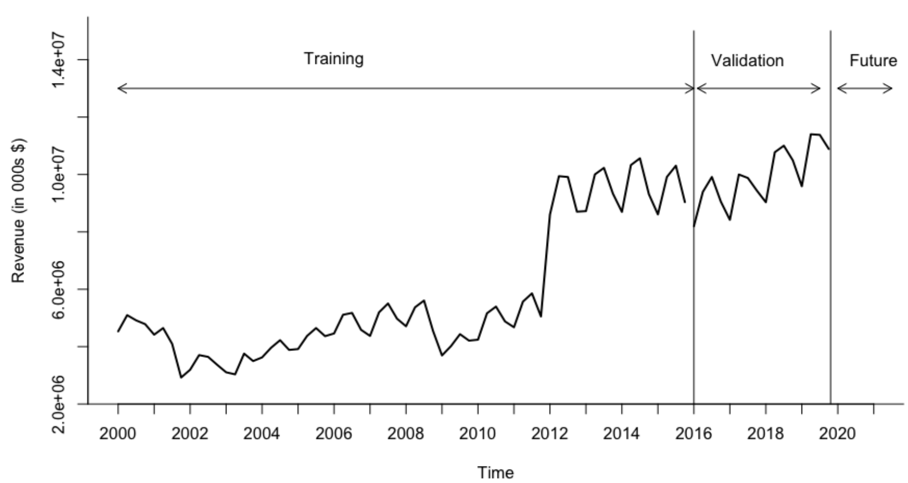
Step 5: Partition Series

We created a data partition of 64 records for the training period and 16 records for the validation period. These partitioned validation and training data sets are (2000-2015) and (2016-2019) respectively are below named as valid.ts and train.ts.

```
> valid.ts
         Qtr1     Qtr2     Qtr3     Qtr4
2016  8195291  9395665  9913185  9051740
2017  8420115  9999723  9877966  9438506
2018  9031936 10776602 11003046 10491647
2019  9589287 11401369 11380482 10887402
```

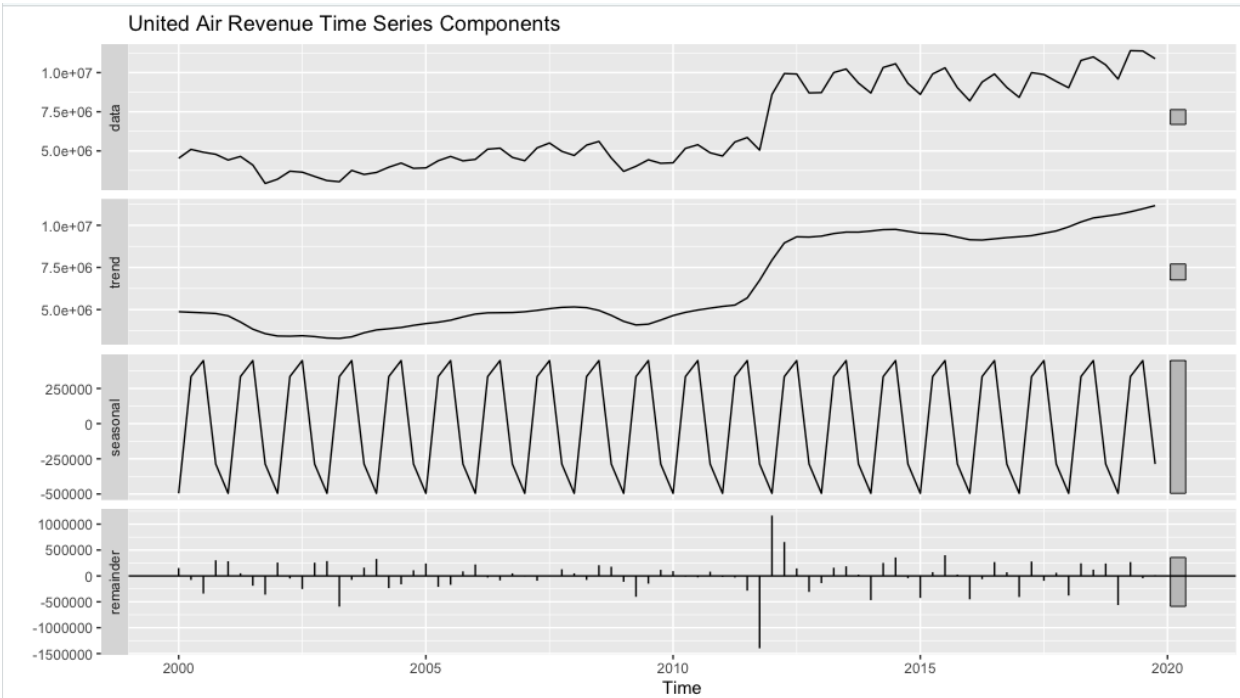
```
> train.ts
         Qtr1     Qtr2     Qtr3     Qtr4
2000  4532976  5097939  4915785  4784580
2001  4417662  4648239  4096978  2924490
2002  3195591  3706075  3644641  3369294
2003  3107212  3034759  3757353  3498405
2004  3625990  3962484  4225317  3887378
2005  3916450  4373958  4647980  4365766
2006  4459678  5111033  5175873  4587275
2007  4374154  5196411  5504770  4973759
2008  4711208  5370526  5606838  4548676
2009  3693574  4020052  4435164  4210448
2010  4243006  5162802  5395978  4880534
2011  4675619  5570539  5856265  5052795
2012  8603978  9940572  9912178  8703439
2013  8723110 10003024 10229894  9331106
2014  8696283 10328256 10563671  9312309
2015  8608479  9913459 10305905  9036289
```

Visual representation of the training and validation partitions of the data.

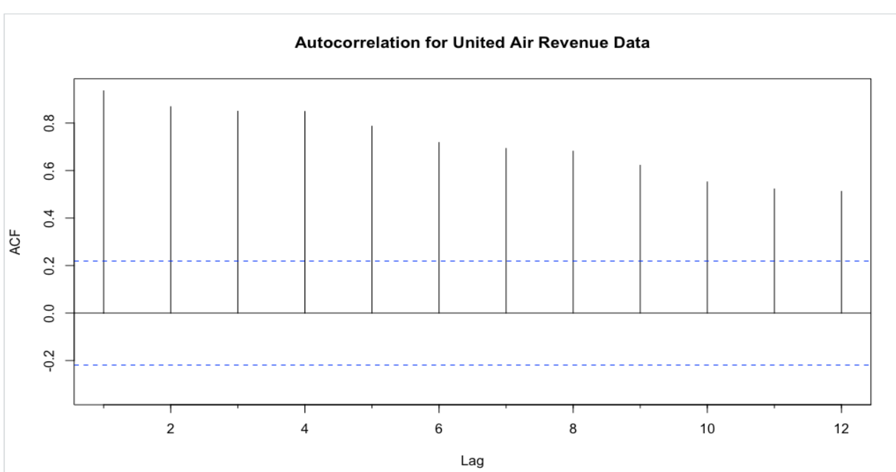


Step 6 & 7: Apply Forecasting & Comparing Performance

With the below time series component we can tell the United airlines revenue has an overall upward trend and also an additive seasonality.



Also, the autocorrelation of the data appears to be statistically significant at all lags implying that there is strong autocorrelation in the data. For all the lags, the ACF is above the upper threshold making the data significant. Which means by further processing of the data with forecasting models we would get better results. Also at lag 1 which represents the trend the ACF is substantially higher than the other lags. And at lag 12 which represents the seasonality the ACF is lower but still significant which tells us that the data has trend and seasonality.

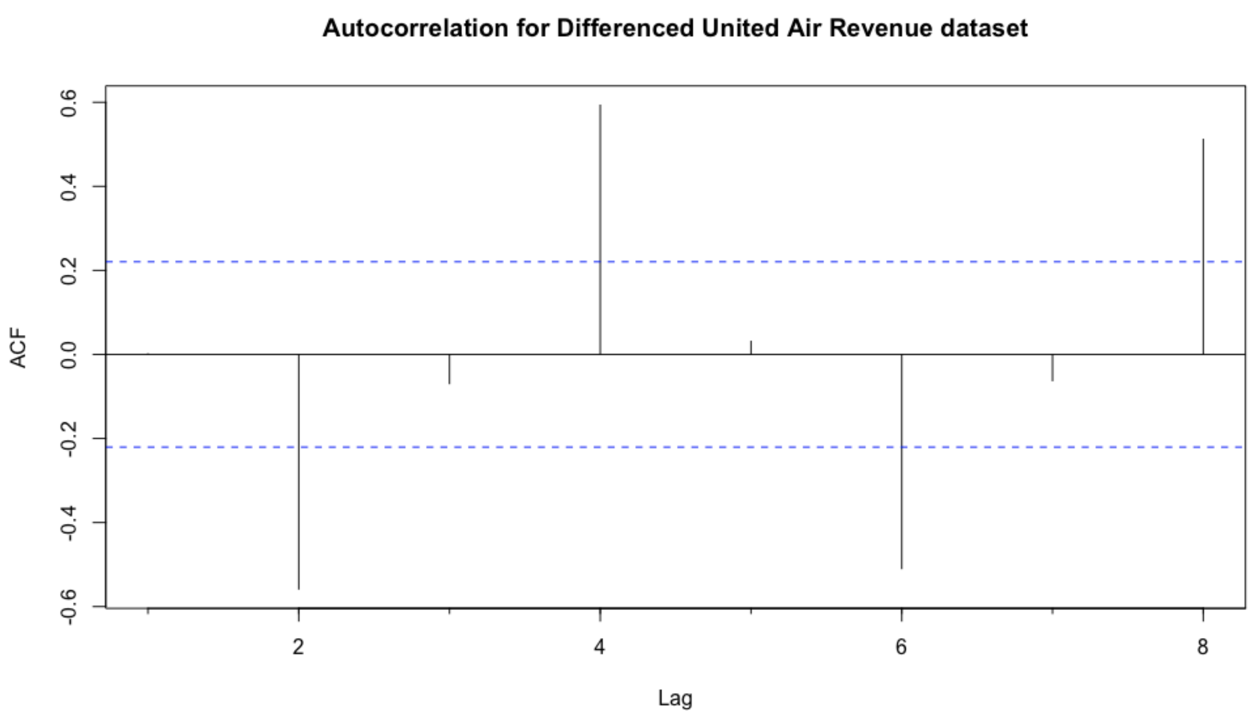


Predictability Test for the Dataset using Lag 1 Differencing -

Below is the data after lag 1 differencing:

	Qtr1	Qtr2	Qtr3	Qtr4
2000		564963	-182154	-131205
2001	-366918	230577	-551261	-1172488
2002	271101	510484	-61434	-275347
2003	-262082	-72453	722594	-258948
2004	127585	336494	262833	-337939
2005	29072	457508	274022	-282214
2006	93912	651355	64840	-588598
2007	-213121	822257	308359	-531011
2008	-262551	659318	236312	-1058162
2009	-855102	326478	415112	-224716
2010	32558	919796	233176	-515444
2011	-204915	894920	285726	-803470
2012	3551183	1336594	-28394	-1208739
2013	19671	1279914	226870	-898788
2014	-634823	1631973	235415	-1251362
2015	-703830	1304980	392446	-1269616
2016	-840998	1200374	517520	-861445
2017	-631625	1579608	-121757	-439460
2018	-406570	1744666	226444	-511399
2019	-902360	1812082	-20887	-493080

Below is the correlogram generated using the lag 1 differencing data from above and maximum lag of 8.



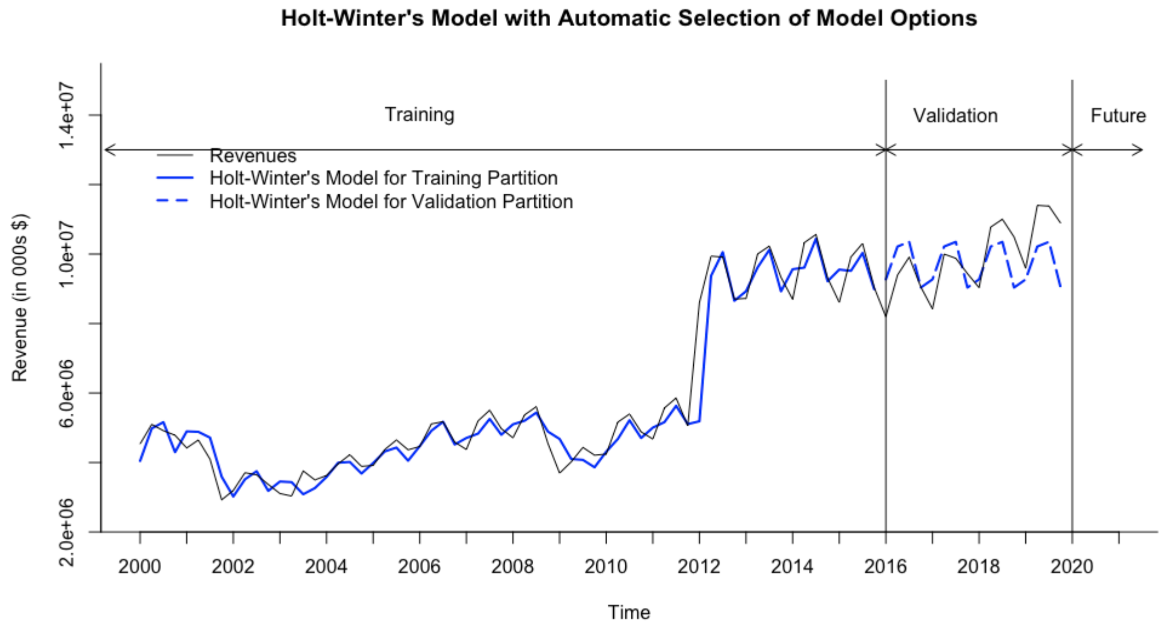
From the correlogram of Lag 1 differencing method, we can say that the data is not a random walk as the ACF at lag 2,4,6,8 are above the level of significance.

Holts Winter Model for training set:

Holt Winter's Model for prediction is used for time series that contains trend and seasonality. Here we have used the automated selection of model options(Z, Z, Z) and the optimal parameters by using ETS().

```
ETS(M,N,M)

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

  Smoothing parameters:
    alpha = 0.9713
    gamma = 1e-04

  Initial states:
    l = 4239339.2048
    s = 0.9296 1.0651 1.0513 0.954

  sigma:  0.1159

     AIC     AICc      BIC
1978.219 1980.219 1993.331
>
```

- The optimal Holt - Winters model obtained is a model of (M, N, M) which represents multiplicative error, multiplicative seasonality and no trend.
- The optimal smoothing parameters for the model are:
    - Alpha (smoothing constant for exponential smoothing) = 0.9713
    - Beta (smoothing constant for trend estimate) = 0
    - Gamma (smoothing constant for seasonality estimate) = 0.0001



The above plot shows us that there's an underestimate of this model.

Auto ARIMA model for training set

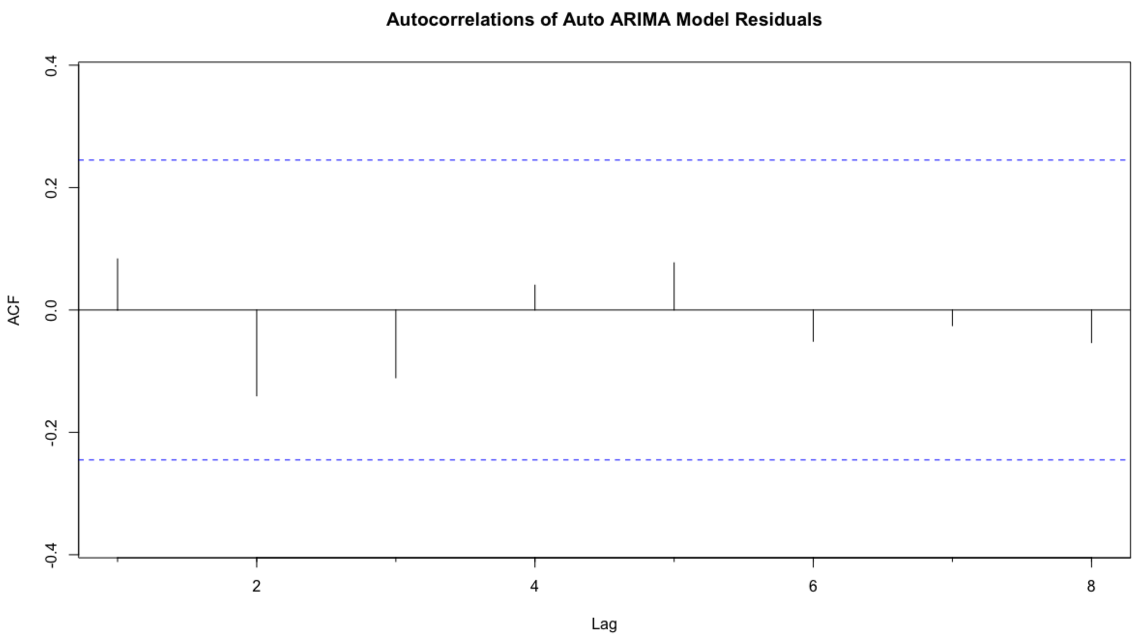
The Autoregressive Integrated Moving Average (ARIMA) model is a flexible model that can be used for forecasting on data with level, trend, and seasonal components. Since our data consists of all three, this model is appropriate to use for analysis. We generated an optimal ARIMA model with automatic selection of (p,d,q) (P,D,Q) parameters using the auto.arima() function. Auto Arima model which uses automatic selection of the optimal parameters. The Auto ARIMA model summary is as below:

```
Series: train.ts
ARIMA(0,1,0)(0,1,1)[4]

Coefficients:
         sma1
      -0.7812
s.e.   0.1097

sigma^2 = 4.073e+11:  log likelihood = -873.71
AIC=1751.43   AICc=1751.64   BIC=1755.58

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 50470.07 607531.6 344854.4 0.7072178 6.283895 0.4844089 0.08345988
```

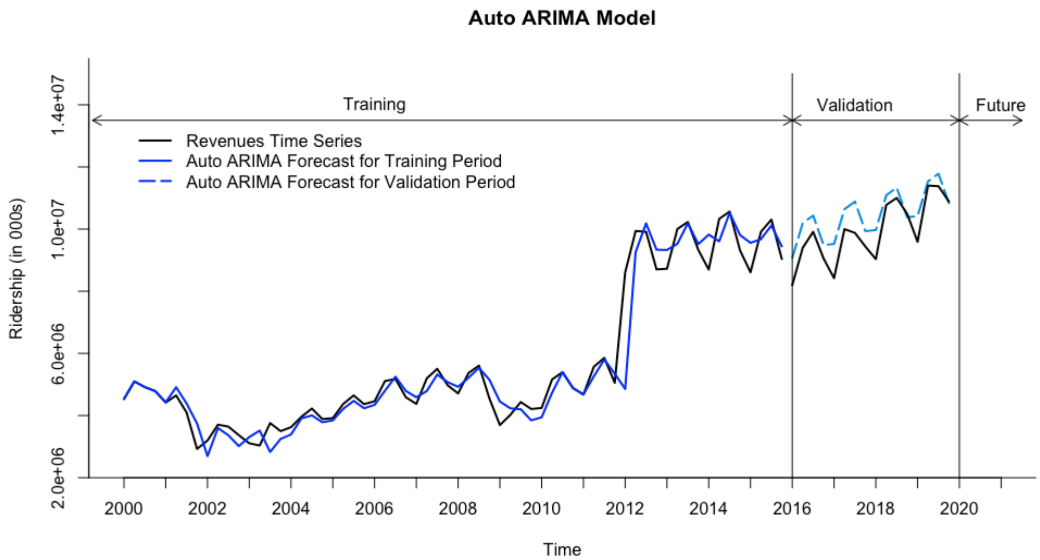
The optimal Auto ARIMA model obtained is *ARIMA (p, d, q) (P, D, Q)[m] -*

- *p = 0*, order 0 autoregressive model *AR (0)*
- *d = 1*, first differencing
- *q = 0*, order 0 moving average *MA (1)* for error lags
- *P = 0*, order 0 autoregressive model *AR (0)* for the seasonal part
- *D = 1*, first differencing for the seasonal part
- *Q = 1*, order 1 moving average *MA (1)* for the seasonal error lags
- *m = 4*, for quarterly seasonality.

The ACF plot below has been generated using the residuals AUTO ARIMA model above.



- Based on the ACF plot we can see that the model has captured the trend, seasonality and any other patterns that existed in the original dataset and has incorporated them into the model.
- The autocorrelation at all lags now fall within the levels of significance signifying there are no more patterns in the residuals.



Regression Models:

Regression model with linear trend: Used to fit a global trend that is applied to the training set of time series and will apply in the forecasting period.

The equation of the below model is:

$$Yt = 2421200 + 100867t$$

- The model has an r^2 of 64.76% which can be considered as good fit.
- Also, all regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.

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

Residuals:
     Min       1Q   Median       3Q      Max 
-2459799  -706129  -213922  1087401  2475930 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2421300     353242   6.855 3.84e-09 ***
trend         100867       9449  10.675 1.13e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1396000 on 62 degrees of freedom
Multiple R-squared:  0.6476,	Adjusted R-squared:  0.6419 
F-statistic: 113.9 on 1 and 62 DF,  p-value: 1.127e-15
```

Regression model with Quadratic trend:

The equation of the below model is:

$$Yt = 4684477 - 104877\,t + 3165\ t\text{^}2$$

- The model has an r^2 of 82.16% which can be considered as good fit.
- Also, all regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1901087  -712515    36404   597680  2586715 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4684476.6   387697.7  12.083  < 2e-16 ***
trend       -104876.5    27523.2  -3.810 0.000325 ***
I(trend^2)     3165.3      410.4   7.713 1.37e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1002000 on 61 degrees of freedom
Multiple R-squared:  0.8216,	Adjusted R-squared:  0.8158 
F-statistic: 140.5 on 2 and 61 DF,  p-value: < 2.2e-16
```

Regression Model with Seasonality:

The equation of the below model is:

$$yt = 5224061 + 740947D2 + 918101D3 + 242598D4$$

- The model has an r^2 of 2.5% which is very low.
- The trend coefficient is statistically significant for this model. However, given the very low r^2 we can determine that this model is not a good fit for the given dataset.

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

Residuals:
     Min       1Q   Median       3Q      Max
-2930249 -1582223  -804302   594772  4421509

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5224061     590161   8.852 1.75e-12 ***
season2       740947     834614   0.888    0.378
season3       918101     834614   1.100    0.276
season4       242598     834614   0.291    0.772
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2361000 on 60 degrees of freedom
Multiple R-squared:  0.02549,	Adjusted R-squared:  -0.02323
F-statistic: 0.5232 on 3 and 60 DF,  p-value: 0.668
```

Regression model with linear trend and seasonality:

The model equation is as below:

$$yt = 2096007 + 100905\,t + 640042\,D2 + 716291\,D3 - 60117\,D4$$

- The model has an r^2 of 67.12% which can be considered as good fit.
- Also, all numeric non-seasonal regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.

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

Residuals:
     Min       1Q   Median       3Q      Max
-2550386  -720280  -216973  1236376  2345070

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2096007     451586   4.641 1.98e-05 ***
trend         100905       9374  10.765 1.51e-15 ***
season2       640042     488953   1.309    0.196
season3       716291     489223   1.464    0.148
season4       -60117     489672  -0.123    0.903
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1383000 on 59 degrees of freedom
Multiple R-squared:  0.6712,	Adjusted R-squared:  0.6489
F-statistic: 30.11 on 4 and 59 DF,  p-value: 1.165e-13
```

Regression Model with quadratic trend and seasonality:

The regression model with quadratic trend and seasonality contains 5 independent variables: trend index (t), squared trend index (t2), and 3 seasonal dummy variables for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4).

The equation for this model is presented below:

$$yt = 4358758 - 105087.8\,t + 3169.1\,t2 + 646380.7\,D2 + 722629.6\,D3 - 60116.6\,D4$$

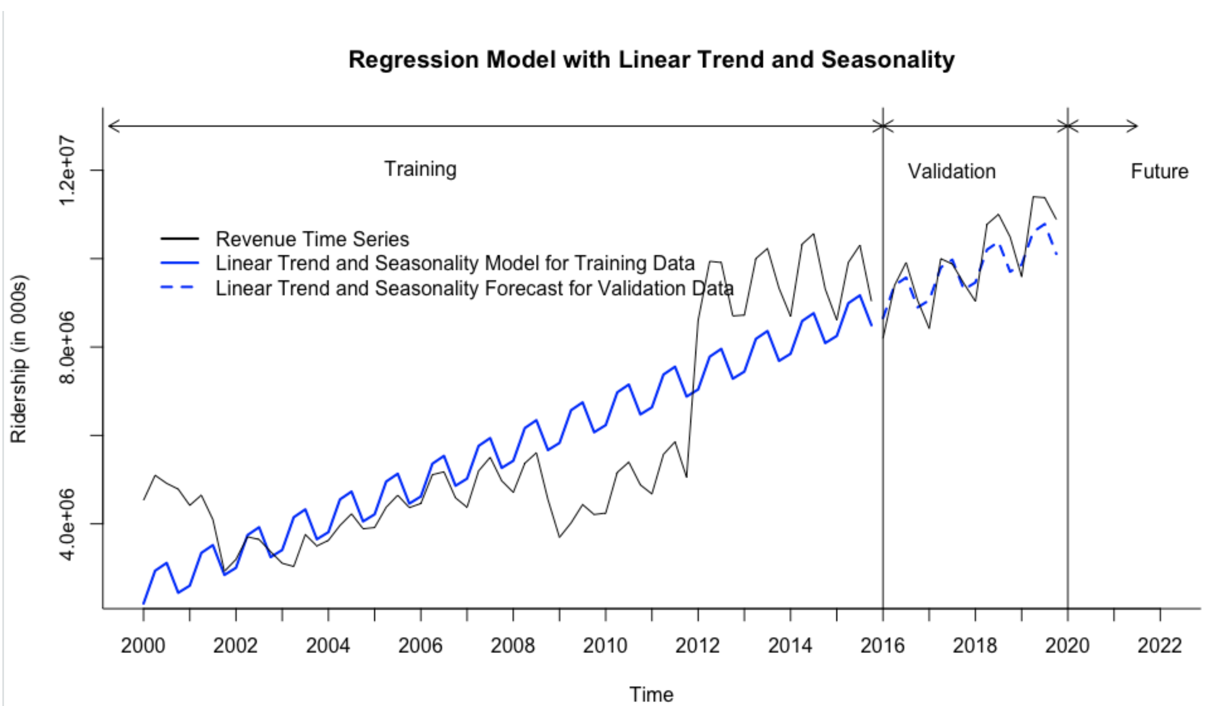
- The model has an r^2 of 84.56% which can be considered as good fit.
- Also, all numeric non-seasonal regression coefficients (trend, intercept) are statistically significant making this model a good fit for the dataset.

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

Residuals:
     Min       1Q   Median       3Q      Max
-1567959  -741787   129494   639375  2267025

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4358758.0   418978.8  10.403 7.03e-15 ***
trend       -105087.8    26259.3  -4.002 0.000181 ***
I(trend^2)     3169.1      391.5   8.095 4.25e-11 ***
season2      646380.7   337922.6   1.913 0.060714 .
season3      722629.6   338108.8   2.137 0.036804 *
season4      -60116.6   338418.1  -0.178 0.859625
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 955600 on 58 degrees of freedom
Multiple R-squared:  0.8456,	Adjusted R-squared:  0.8323
F-statistic: 63.54 on 5 and 58 DF,  p-value: < 2.2e-16
```

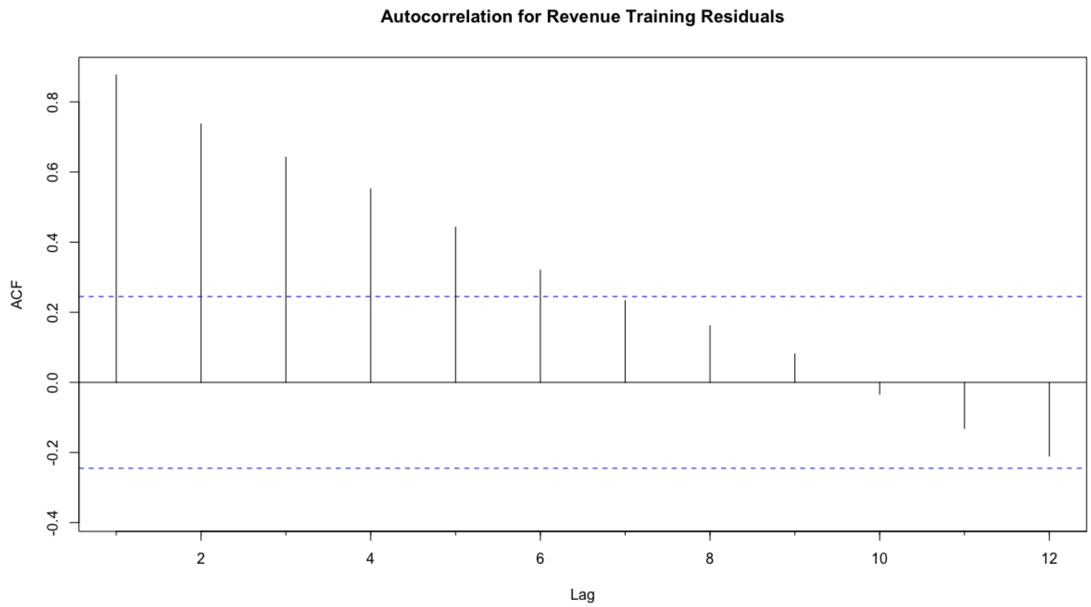
Accuracy measures for Regression models above: Below are the accuracy measures for all the regression models developed on the training partition:

```
> round(accuracy(train.lin.pred$mean, valid.ts),3)
               ME     RMSE    MAE  MPE  MAPE  ACF1 Theil's U
Test set 194227 704232.5 640285 1.36 6.442 0.044     0.689
> round(accuracy(train.quad.pred$mean, valid.ts),3)
                 ME    RMSE     MAE     MPE   MAPE  ACF1 Theil's U
Test set -3857333 4006997 3857333 -38.954 38.954 0.538     3.896
> round(accuracy(train.season.pred$mean, valid.ts),3)
                ME    RMSE     MAE    MPE   MAPE  ACF1 Theil's U
Test set 4228900 4300147 4228900 42.283 42.283 0.636     4.254
> round(accuracy(train.lin.season.pred$mean, valid.ts),3)
                 ME     RMSE      MAE   MPE  MAPE  ACF1 Theil's U
Test set 192701.7 501208.6 430589.7 1.539 4.268 0.078     0.479
> round(accuracy(train.quad.season.pred$mean, valid.ts),3)
                 ME    RMSE     MAE     MPE   MAPE ACF1 Theil's U
Test set -3863771 3981770 3863771 -38.823 38.823 0.69     3.901
```

- A model with the least RMSE and least MAPE is considered to be the best model.
- Based on the above accuracy measures for the training partition, the best model is Linear trend and seasonality with the lowest MAPE of 4.2% and RMSE of 3981770.
- Below is the Plot for the same.



Below, the ACF plot of the residuals from the level 1 forecasting model(Regression model with Linear Trend and Seasonality)



- From the above correlogram it can be seen that for most of the lags the ACF is still above the upper threshold.
- This signifies that there are still certain patterns (trend) existing in the residuals which need to be captured to further improve our forecasts.
- We intend to do this with a 2-Level forecasting model.

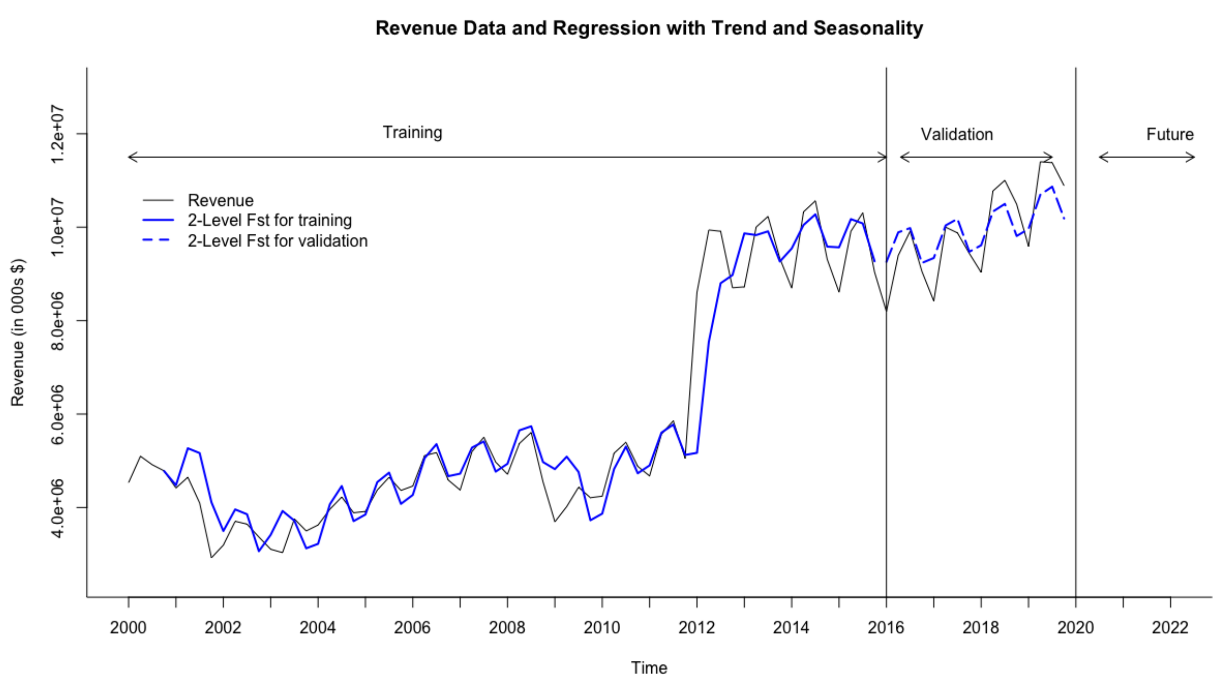
2-LEVEL FORECASTING REGRESSION + MA

Given that the best model is Regression model with Linear trend with Seasonality, we choose the same for Level 1 forecasting of our 2-Level forecasting model.

Below are the steps involved in applying the 2-Level forecast.
- Identified regression residuals for training partition (differences between actual and regression values in the same periods).
- Display the Autocorrelation correlogram for the Revenue training residuals.
- Apply trailing MA for residuals with window width k = 4 for training partition.
- Create residuals forecast for validation period.
- Develop a two-level forecast for validation period by combining regression forecast and trailing MA forecast for residuals.
- Create and represent a table for validation period: validation data, regression forecast, trailing MA for residuals and total forecast.

```
> valid.df
   Revenues Regression.Fst MA.Residuals.Fst Combined.Fst
1   8195291        8654829        603490.39      9258320
2   9395665        9395777        494750.20      9890527
3   9913185        9572931        407758.00      9980689
4   9051740        8897428        338164.20      9235592
5   8420115        9058449        282489.13      9340938
6   9999723        9799397        237949.05     10037346
7   9877966        9976550        202316.96     10178867
8   9438506        9301048        173811.27      9474859
9   9031936        9462069        151006.71      9613076
10 10776602       10203016        132763.05     10335780
11 11003046       10380170        118168.11     10498338
12 10491647        9704667        106492.15      9811160
13  9589287        9865689         97151.38      9962840
14 11401369       10606636         89678.76     10696315
15 11380482       10783790         83700.66     10867491
16 10887402       10108287         78918.18     10187205
```

Plot for 2-Level forecasting for the validation period.

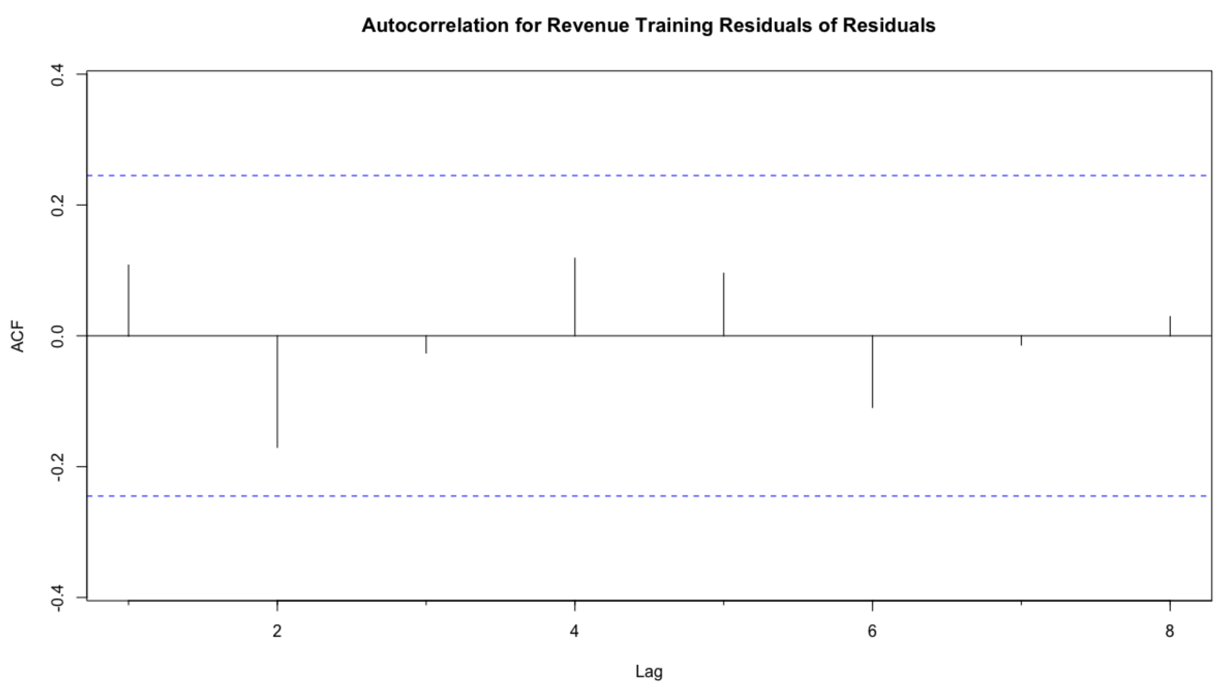


2-LEVEL FORECASTING REGRESSION + AR model

In this model, in addition to Level - 1 regression forecasting, we will use an AutoRegressive model of order 1 to forecast the residuals from the level-1 model and improve our forecasts. The summary of the AR (1) developed is as below:

```
Series: train.lin.season.res
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1      mean
      0.9079  335933.5
s.e.  0.0508  703459.5

sigma^2 = 3.435e+11:  log likelihood = -940.66
AIC=1887.32   AICc=1887.72   BIC=1893.8

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE    ACF1
Training set -46169.2 576820.7 368537.6 7.383986 72.13319 0.5341419 0.10807
```

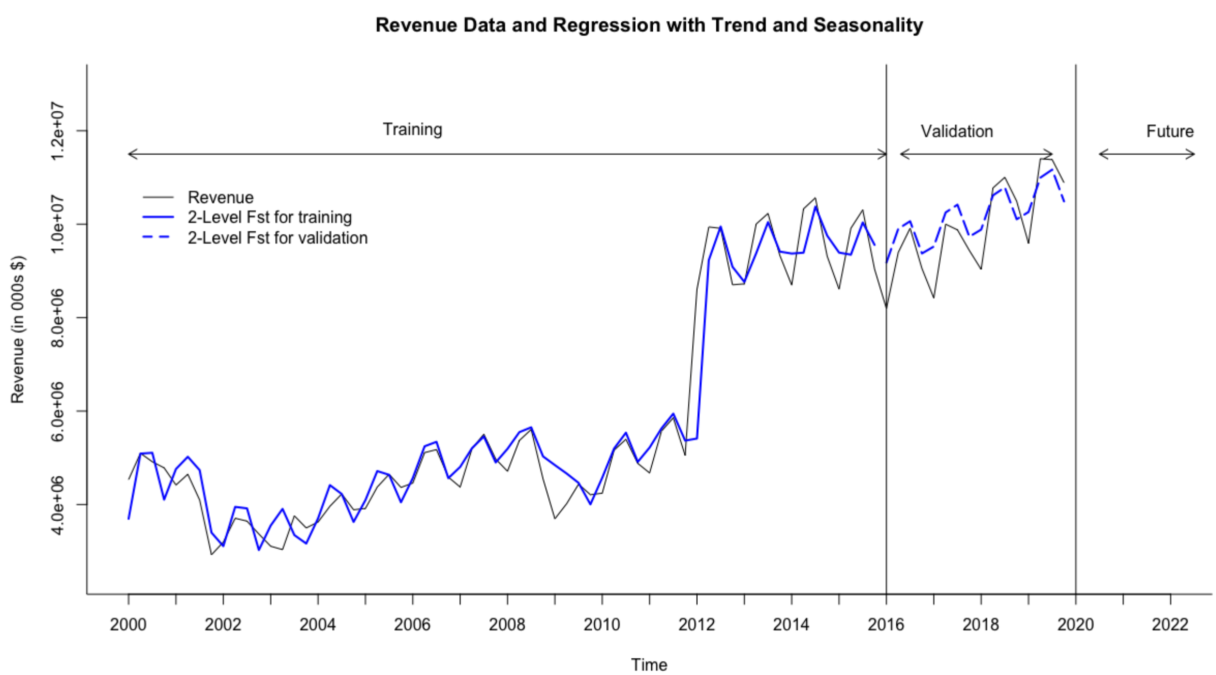
- *ARIMA (1, 0, 0)* is an autoregressive (AR) model with order 1, no differencing, and no moving average model.



- Autocorrelations of residuals of residuals produced by the AR (1) model can be inferred from this correlogram to be random. Hence, significant autocorrelation in all lags has been observed by the AR (1) model for residuals.

- Data table with validation data, regression forecast for validation period, AR (1) residuals for validation, and two level model results.

	Revenues	Regression.Fst	AR(1).Residuals.Fst	Combined.Fst
1	8195291	8654829	523451.9	9178281
2	9395665	9395777	506175.8	9901953
3	9913185	9572931	490491.4	10063422
4	9051740	8897428	476251.9	9373680
5	8420115	9058449	463324.4	9521774
6	9999723	9799397	451587.8	10250984
7	9877966	9976550	440932.6	10417483
8	9438506	9301048	431259.0	9732307
9	9031936	9462069	422476.6	9884546
10	10776602	10203016	414503.4	10617520
11	11003046	10380170	407264.8	10787435
12	10491647	9704667	400693.0	10105360
13	9589287	9865689	394726.7	10260416
14	11401369	10606636	389310.1	10995946
15	11380482	10783790	384392.5	11168183
16	10887402	10108287	379928.0	10488215



Below are the accuracy measures for all the forecasting models considered above namely: (1) Holt Winter's model; (2) Auto ARIMA model; (3) Regression Model with Linear Trend and

Seasonality; (4) 2 Level Forecasting (Regression Model with Linear Trend and Seasonality and Moving Averages); (5)2 Level Forecasting (Regression Model with Linear Trend and Seasonality and AR (1))

```
> round(accuracy(hw.ZZZ.pred$mean,valid.ts), 3)
              ME     RMSE      MAE  MPE  MAPE  ACF1 Theil's U
Test set 209736 869494.8 724341.2 1.45 7.203 0.508     0.796
> round(accuracy(train.auto.arima.pred$mean,valid.ts), 3)
                ME   RMSE      MAE    MPE  MAPE  ACF1 Theil's U
Test set -540639.2 647861 561785.2 -5.763 5.962 0.008     0.648
> round(accuracy(train.lin.season.pred$mean, valid.ts),3)
               ME     RMSE      MAE   MPE  MAPE  ACF1 Theil's U
Test set 192701.7 501208.6 430589.7 1.539 4.268 0.078     0.479
> round(accuracy(fst.2level.ma, valid.ts), 3)
                ME     RMSE      MAE    MPE  MAPE  ACF1 Theil's U
Test set -32211.27 558703.3 475243.7 -0.859 4.894 0.266     0.498
> round(accuracy(fst.2level.ar, valid.ts), 3)
                ME   RMSE      MAE    MPE  MAPE  ACF1 Theil's U
Test set -243346.4 546017 465582.4 -2.929 4.954 0.145     0.512
```

- The model with the least RMSE at 501208.6 and MAPE at 4.268% is the Regression model with Linear trend and Seasonality.
- The other models which have slightly higher MAPE and RMSE but still performing good include the 2-Level forecasting models.
- Thus, we would like to consider applying these 3 models on the entire dataset.

Now fitting the optimal model on the Entire Dataset

The chosen models to apply on the entire dataset include:

1. **Regression model with Linear Trend and Seasonality**
2. **2-Level Forecasting Regression + MA**
3. **2-Level Forecasting Regression + AR () model**

1. Regression Model with Linear trend and Seasonality for entire dataset
Below is the summary of the same.

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

Residuals:
     Min       1Q   Median       3Q      Max 
-2623468  -608681  -138752   657098  2527823 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1901576     364079   5.223 1.53e-06 ***
trend         103577       6027  17.185  < 2e-16 ***
season2       806017     393242   2.050   0.0439 *  
season3       874229     393381   2.222   0.0293 *  
season4       114981     393612   0.292   0.7710    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1243000 on 75 degrees of freedom
Multiple R-squared:  0.8024,	Adjusted R-squared:  0.7919 
F-statistic: 76.15 on 4 and 75 DF,  p-value: < 2.2e-16
```

- The regression model contains 4 independent variables: trend index (t) and 3 seasonal dummy variables for Q2 (season2 – D2), Q3 (season3 – D3) and Q4 (season4 – D4).
- All numeric coefficients are below 0.05 making these coefficients significant.
- R-squared and Adj.R-squared are at 80.24% and 79.19%, which is very good.
- Thus, this model equation is significant, making it a good fit for forecasting.

Below is the future 12 periods prediction using Regression model with Linear Trend and Seasonality.

	Point Forecast	Lo 0	Hi 0
2020 Q1	10291316	10291316	10291316
2020 Q2	11200910	11200910	11200910
2020 Q3	11372699	11372699	11372699
2020 Q4	10717027	10717027	10717027
2021 Q1	10705624	10705624	10705624
2021 Q2	11615218	11615218	11615218
2021 Q3	11787007	11787007	11787007
2021 Q4	11131336	11131336	11131336
2022 Q1	11119932	11119932	11119932
2022 Q2	12029526	12029526	12029526
2022 Q3	12201315	12201315	12201315
2022 Q4	11545644	11545644	11545644

Later, using trailing MA residuals forecast for the future 12 periods.

	Point Forecast	Lo 0	Hi 0
2020 Q1	329735.3	329735.3	329735.3
2020 Q2	326759.9	326759.9	326759.9
2020 Q3	324379.6	324379.6	324379.6
2020 Q4	322475.3	322475.3	322475.3
2021 Q1	320951.9	320951.9	320951.9
2021 Q2	319733.2	319733.2	319733.2
2021 Q3	318758.2	318758.2	318758.2
2021 Q4	317978.2	317978.2	317978.2
2022 Q1	317354.2	317354.2	317354.2
2022 Q2	316855.0	316855.0	316855.0
2022 Q3	316455.7	316455.7	316455.7
2022 Q4	316136.2	316136.2	316136.2

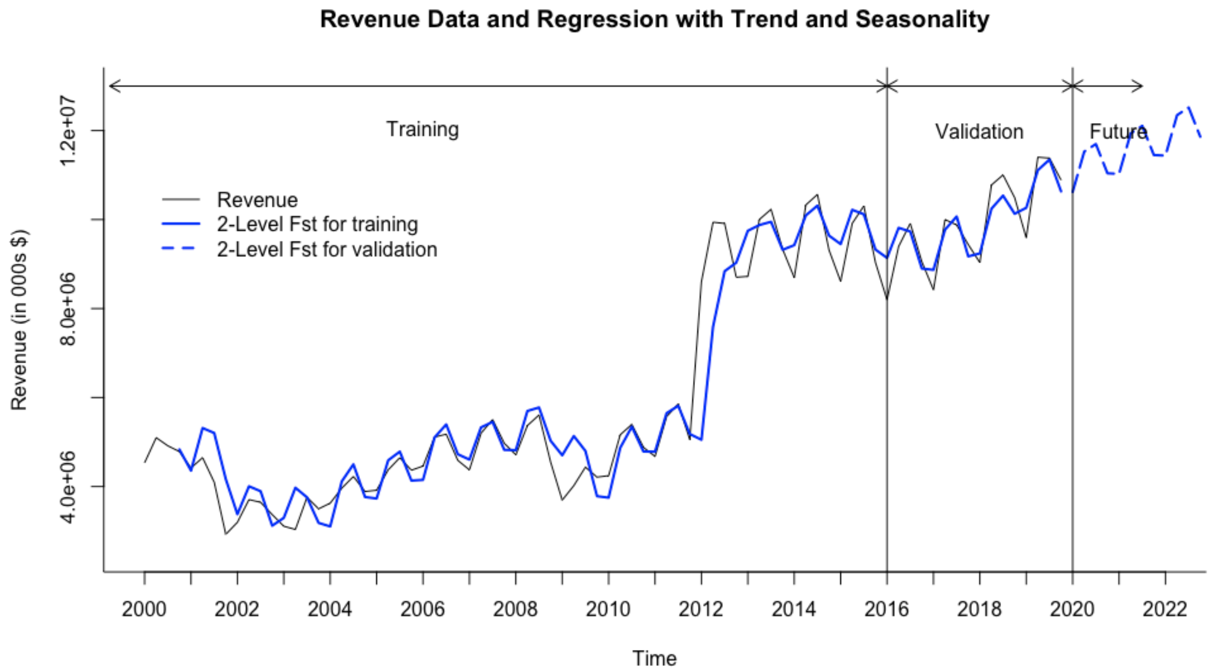
Now Developing a 2-Level forecast for the entire dataset by combining regression forecast and trailing MA forecast for residuals and presented the future 12 periods prediction below.

	Qtr1	Qtr2	Qtr3	Qtr4
2020	10621051	11527670	11697079	11039503
2021	11026576	11934951	12105765	11449314
2022	11437286	12346381	12517771	11861780

After that, created a table that shows the Regression Forecast, MA.Residuals Forecast, Combined Forecast for future 12 periods.

	Regression.Fst	MA.Residuals.Fst	Combined.Fst
1	10291316	329735.3	10621051
2	11200910	326759.9	11527670
3	11372699	324379.6	11697079
4	10717027	322475.3	11039503
5	10705624	320951.9	11026576
6	11615218	319733.2	11934951
7	11787007	318758.2	12105765
8	11131336	317978.2	11449314
9	11119932	317354.2	11437286
10	12029526	316855.0	12346381
11	12201315	316455.7	12517771
12	11545644	316136.2	11861780

Below is the Regression with trailing MA and seasonality for the entire revenue dataset:



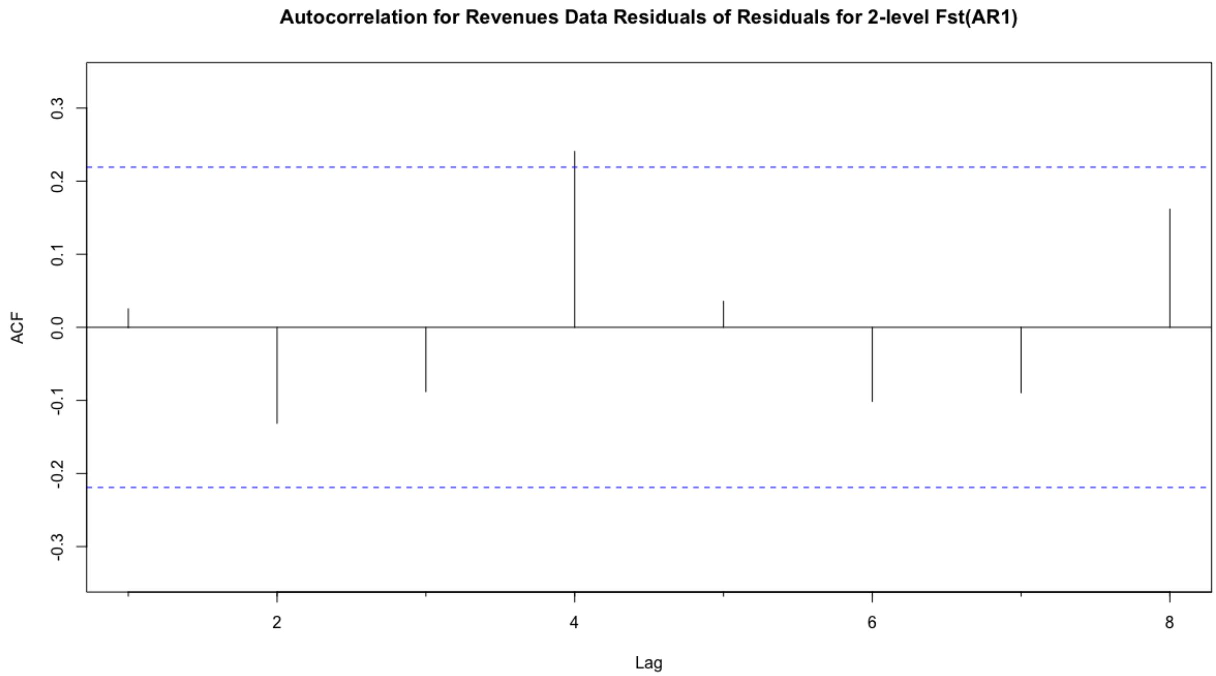
2-Level forecasting model with Regression and AR model for entire time series dataset.

Summary of Fitted Two Level forecasting with AR model:

```
Series: tot.trend.seas.res
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1      mean
      0.8954  273884.2
s.e.  0.0506  565668.1

sigma^2 = 3.298e+11:  log likelihood = -1174.18
AIC=2354.37   AICc=2354.69   BIC=2361.52

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set -40805.9 567081.3 380491.6 68.90846 141.2462 0.6058791 0.02528264
```

- *ARIMA (1, 0, 0)* is an autoregressive (AR) model with order 1, no differencing, and no moving average model.

Autocorrelation for the entire dataset using the 2-level Forecasting with regression and AR model.



- In the autocorrelation for residuals of residuals for 2-level forecast using AR (1) model all the lags are within the significance threshold except for lag 4 which is weakly significant.
- By looking at the correlogram we can say that all the patterns of residuals are considered in this model.

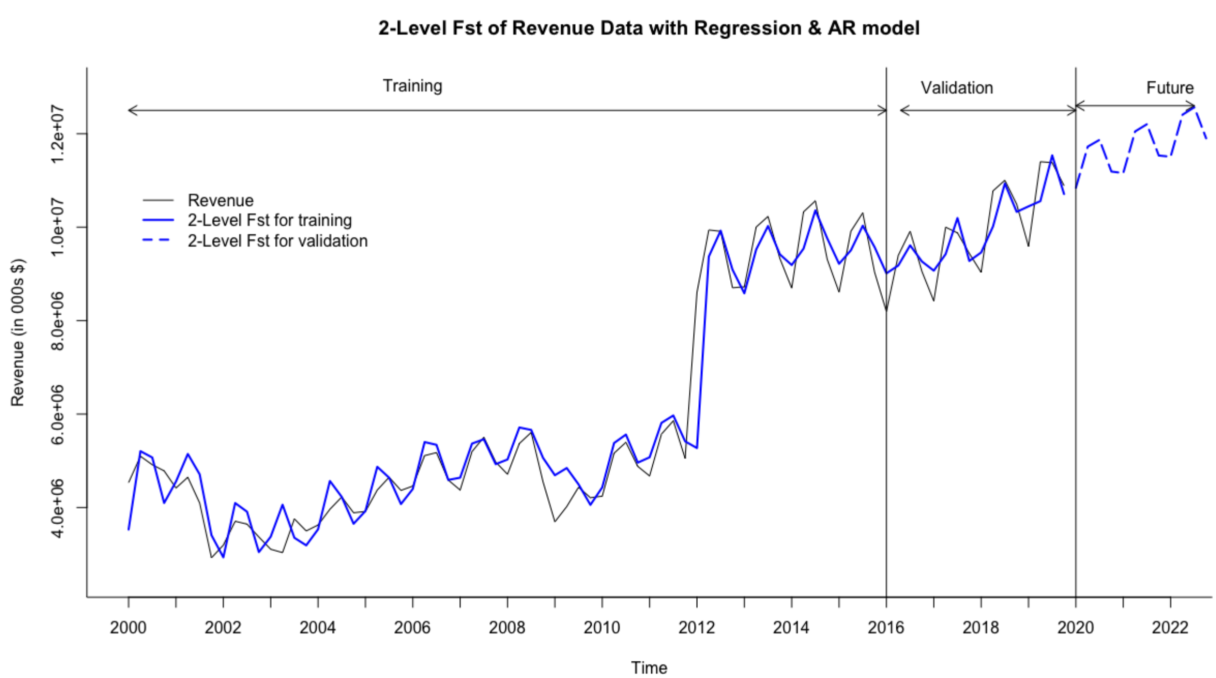
Developed a two-level model's forecast with linear trend and seasonality regression + AR (1) for residuals for future periods.

	Qtr1	Qtr2	Qtr3	Qtr4
2020	10843479	11723957	11869676	11190662
2021	11158358	12049239	12204273	11533599
2022	11508763	12406330	12567351	11902037

Data table with Future 12 periods data, regression forecast for Future 12 periods, AR (1) residuals for Future 12 periods, and 2-level model results.

	UnitedRev.Forecast	AR(1)Forecast	Combined.Forecast
1	10291316	552163.4	10843479
2	11200910	523046.7	11723957
3	11372699	496976.5	11869676
4	10717027	473634.1	11190662
5	10705624	452734.0	11158358
6	11615218	434020.7	12049239
7	11787007	417265.4	12204273
8	11131336	402263.3	11533599
9	11119932	388830.8	11508763
10	12029526	376803.8	12406330
11	12201315	366035.2	12567351
12	11545644	356393.3	11902037

Plot for 2-Level forecast of entire revenue data with Regression and AR model.



Step 8: Implement Forecast

- Below are the accuracies for all the models chosen for the entire dataset and also the Naive and Seasonal Naive models.
- The least MAPE off all the models is 6.617% and the least RMSE is 567081.3

- Based on the MAPE and RMSE the best model that forecasts the future Revenue for United Airlines appears to be 2-Level Forecasting model with Regression with Linear Trend and seasonality and Auto Regressive for AR (1)

```
> round(accuracy((naive(unitedrev.ts))$fitted, unitedrev.ts), 3)
               ME     RMSE      MAE   MPE  MAPE  ACF1 Theil's U
Test set 80435.77 815067.2 597819.7 0.374 9.274 0.001         1
> round(accuracy((snaive(unitedrev.ts))$fitted, unitedrev.ts), 3)
               ME    RMSE      MAE   MPE   MAPE  ACF1 Theil's U
Test set 314832.4 1090306 668772.6 2.703 11.386 0.773      1.32
> round(accuracy(tot.trend.seas$fitted, unitedrev.ts), 3)
         ME    RMSE      MAE    MPE   MAPE  ACF1 Theil's U
Test set  0 1203913 941453.7 -3.328 16.823 0.856     1.715
> round(accuracy(fst.2level.ma.tot, unitedrev.ts), 3)
               ME     RMSE      MAE    MPE  MAPE  ACF1 Theil's U
Test set -33334.8 660513.5 408709.1 -1.849 7.057 0.469     0.917
> round(accuracy(fst.2level.tot, unitedrev.ts), 3)
               ME     RMSE      MAE    MPE  MAPE  ACF1 Theil's U
Test set -40805.9 567081.3 380491.6 -1.709 6.617 0.025     0.818
```

Conclusion

After considering various models for forecasting for United Airlines Revenue, the best model based on least RMSE and MAPE is 2-Level Forecasting model with Regression with Linear Trend and Seasonality and Auto Regressive for AR (1). However, the other models considered specifically the 2-Level Forecasting model with Regression with Linear Trend and seasonality and MA can also be used to forecast. Thus, it is important for the forecasting team to semiannually inspect the performances of these 2 models as data gets updated with new quarters and use the right forecasting model. This constant checkup will help the team to improve the forecasting.

Appendix

1. With the help of the ARIMA () function, we test the predictability of the United Airlines Revenue dataset to fit the AR (1) model. This is tested for the entire dataset with the beta coefficient and the standard error is considered with the alpha value of 0.05.

```
#TEST predictability of United Air Revenues dataset.

# Use Arima() function to fit AR(1) model for United Air Revenues dataset.
# The ARIMA model of order = c(1,0,0) gives an AR(1) model.
unitedrev.ar1<- Arima(unitedrev.ts, order = c(1,0,0))
summary(unitedrev.ar1)

# Apply z-test to test the null hypothesis that beta
# coefficient of AR(1) is equal to 1.
ar1 <- 0.9594
s.e. <- 0.0291
null_mean <- 1
alpha <- 0.05
z.stat <- (ar1-null_mean)/s.e.
z.stat
p.value <- pnorm(z.stat)
p.value
if (p.value<alpha) {
  "Reject null hypothesis"
} else {
  "Accept null hypothesis"
}
```

```
> ar1 <- 0.9594
> s.e. <- 0.0291
> null_mean <- 1
> alpha <- 0.05
> z.stat <- (ar1-null_mean)/s.e.
> z.stat
[1] -1.395189
> p.value <- pnorm(z.stat)
> p.value
[1] 0.08147943
> if (p.value<alpha) {
+   "Reject null hypothesis"
+ } else {
+   "Accept null hypothesis"
+ }
[1] "Accept null hypothesis"
>
```