# **<u>Air Passengers data forecasting using Time Series Analytics</u>**


Prachiti Jadhav (on2732)

Sukhman Legha (vy3059)

Manohar Chakrapu (hx6672)

Chandratej Kurella (ja8071)

Manideep Elasagaram (is6496)

**Summary:**

The air passenger data forecasting project using time series analytics involves analyzing historical data on the number of air passengers and forecasting future passenger traffic based on data patterns and trends. The project's main goal is to create a predictive model that accurately predicts the number of air passengers for the next 12 months, allowing airlines to plan and optimize their operations accordingly.

Several key steps are involved in the project, including optimal forecasting technique, exploratory data analysis, model selection, and model evaluation. This project's dataset includes monthly data on the number of air passengers from 1948 to 1960. To ensure that the data is ready for analysis, it is preprocessed and transformed, which includes checking for missing values, removing outliers, and running statistical tests to identify trends and patterns in the data.

To build predictive models and forecast the number of air passengers for the next 12 months, various time series forecasting techniques such as ARIMA, MA, and Multiple Regression are used. Each model's accuracy is assessed using various metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

Finally, the project comes to a close with a discussion of the findings and their implications for the airline industry. This project's findings can help airlines make more informed decisions about scheduling, pricing, and capacity planning over the next 12 months, resulting in increased efficiency and profitability.

**Introduction:**

Air travel has become an important mode of transportation for both business and pleasure, with the number of air passengers increasing rapidly over the years. Forecasting air passenger traffic accurately is critical for airlines to efficiently plan operations, manage resources, and maximize profitability. According to studies, air passenger demand will double in the next 20 years, with an average annual growth rate of 3.5%.

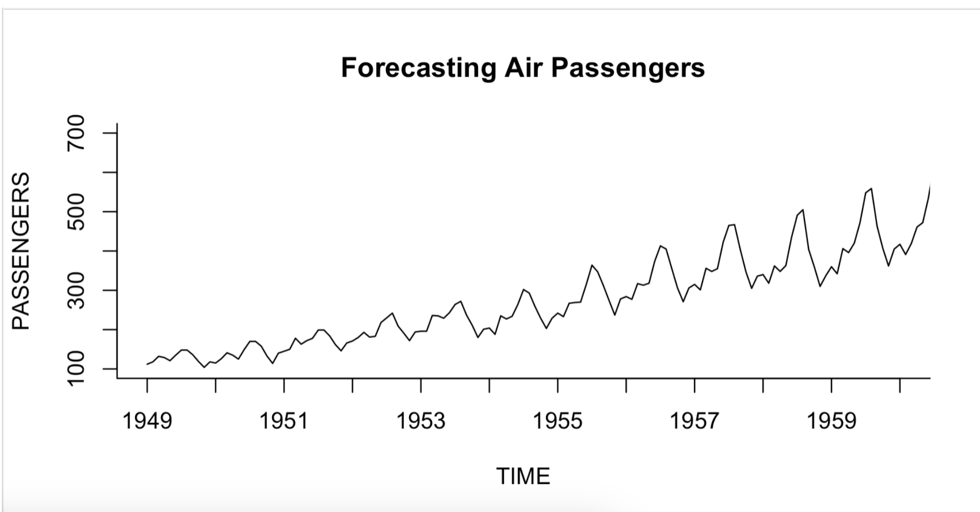
Forecasting air passenger traffic accurately can assist airports and government agencies in planning infrastructure development, such as the expansion of airport facilities, parking lots, and public transportation. This can help ensure that airport infrastructure keeps up with rising demand and that passengers have a better travel experience.

It also helps to improve safety by planning for contingencies and disruptions, such as flight cancellations or delays caused by weather. This can help to improve safety by lowering the risk of accidents caused by last-minute flight changes or overbooking. The project also emphasizes the importance of investing in airport infrastructure, technology, and sustainable aviation fuels to meet rising air travel demand while reducing the industry's environmental impact.

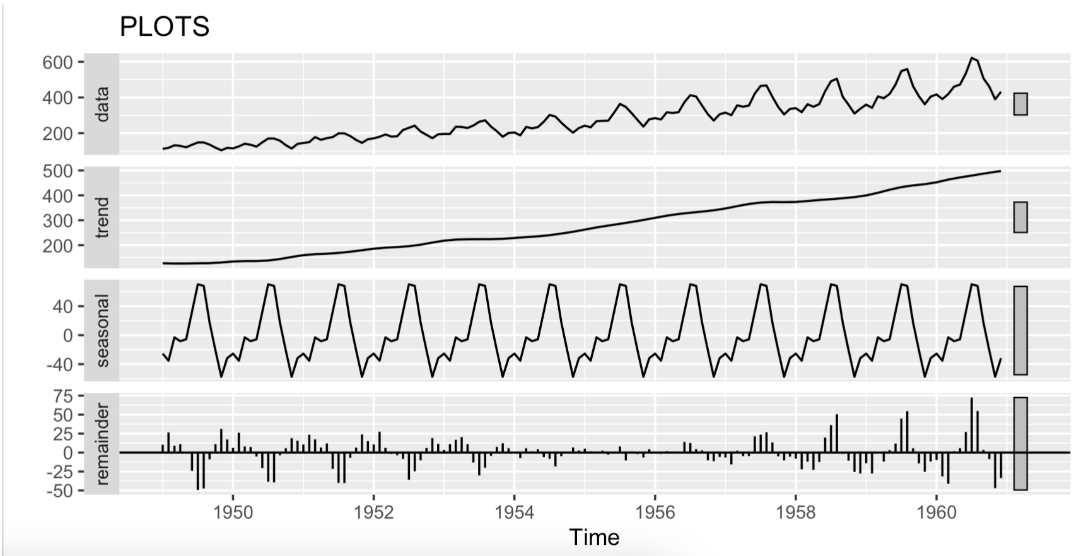
This also benefits the environment by lowering harmful gas emissions. Airlines can reduce the number of empty seats on flights by accurately forecasting demand. We can predict future demand by identifying trends and patterns and making informed decisions about infrastructure, technology, and sustainability thanks to this time series project. It also demonstrates the significance of including a variety of scenarios in the forecasting model in order to provide a comprehensive view of the future of air passenger demand.

The goal of this project is to forecast export data for the USA for the upcoming quarters of 2023 and the upcoming fiscal year. The goal is to create a time series forecast predictive model that will properly consider the components of the historical data and effectively forecast the desired quarters. Naturally, the model with the highest accuracy will be considered the model of choice. The resulting forecasts will be used to monitor the forecasting of USA exports. The forecasting models developed for this project were done via the R language.

**Step 3: Explore and Visualize Series**




**Time Series Components:**




The time series seems to be trending upward.

There is a trend component, as seen in the plot above.

**Step 4: Data Preprocessing**

| | |
|---|---|
| airpassenger.ts | Time-Series [1:144] from 1949 to 1961: 112 118 132 129 121 135 … |

There are two columns: one lists the Time, and the other lists the Passengers.

We have a total of 144 records in total.

Checking the predictability of data

We tested the dataset for predictability to see if there were any random walks in the data, or if it was predictable.

**Step 5: Partition Series**

Data is divided into 80:20 split partitions, 80 split is for training dataset and 20 split is for validation dataset.

These partitioned data sets are: Training data:train.ts (120 records), Validation data:valid.ts(24 records)

| | |
|---|---|
| airpassenger.ts | Time-Series [1:144] from 1949 to 1961: 112 118 132 129 121 135 … |
| nTrain | 120 |
| nValid | 24 |
| train.ts | Time-Series [1:120] from 1949 to 1959: 112 118 132 129 121 135 … |
| valid.ts | Time-Series [1:24] from 1959 to 1961: 360 342 406 396 420 472 5… |

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

**<u>Regression based Models.</u>**

Regression-based models are the following method applied to the time series analysis.

Depending on the time series plot, a different type of model will be used. This kind of model was taken into consideration because it is straightforward to apply and offers reliable results because it takes seasonality and trend into account. Additionally, autoregressive components and a tail moving average for the residuals may be added to this model to further improve it. The model was first tested on the training and validation partitions before being run on the complete data set.

**a. Regression model with linear trend**

```
> summary(train.lin)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-81.861 -23.544  -2.859  18.331 120.624 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  94.9661     7.1032   13.37   <2e-16 ***
trend         2.4949     0.1019   24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.66 on 118 degrees of freedom
Multiple R-squared:  0.8356,	Adjusted R-squared:  0.8342 
F-statistic: 599.6 on 1 and 118 DF,  p-value: < 2.2e-16

> 
```

The model is a regression model with a linear trend and seasonality for the 12-year span. Since the F-Statistic p-value is so low (2.2e16), significantly lower than an alpha of 5%, the model is statistically significant. The model's R-Square is 83.56%, which indicates that the predictors can account for 83.56% of the variation in the data related to air passengers. The model's adjusted R-square is 83.42%. This model has only 1 trend predictor. Since none of these predictors had p-values higher than an alpha of 5%, they are all significant.

**Forecasting for validation data**

```
> train.lin.pred
         Point Forecast     Lo 0     Hi 0
Jan 1959       396.8506 396.8506 396.8506
Feb 1959       399.3455 399.3455 399.3455
Mar 1959       401.8404 401.8404 401.8404
Apr 1959       404.3353 404.3353 404.3353
May 1959       406.8302 406.8302 406.8302
Jun 1959       409.3251 409.3251 409.3251
Jul 1959       411.8200 411.8200 411.8200
Aug 1959       414.3150 414.3150 414.3150
Sep 1959       416.8099 416.8099 416.8099
Oct 1959       419.3048 419.3048 419.3048
Nov 1959       421.7997 421.7997 421.7997
Dec 1959       424.2946 424.2946 424.2946
Jan 1960       426.7895 426.7895 426.7895
Feb 1960       429.2844 429.2844 429.2844
Mar 1960       431.7793 431.7793 431.7793
Apr 1960       434.2743 434.2743 434.2743
May 1960       436.7692 436.7692 436.7692
Jun 1960       439.2641 439.2641 439.2641
Jul 1960       441.7590 441.7590 441.7590
Aug 1960       444.2539 444.2539 444.2539
Sep 1960       446.7488 446.7488 446.7488
Oct 1960       449.2437 449.2437 449.2437
Nov 1960       451.7386 451.7386 451.7386
Dec 1960       454.2336 454.2336 454.2336
>
```

### b. Regression model with quadratic trend

```
> summary(train.quad)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-97.372 -20.805  -5.131  18.724 108.826 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.121e+02  1.060e+01  10.574  < 2e-16 ***
trend       1.650e+00  4.046e-01   4.079 8.29e-05 ***
I(trend^2)  6.980e-03  3.239e-03   2.155   0.0332 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.08 on 117 degrees of freedom
Multiple R-squared:  0.8418,	Adjusted R-squared:  0.8391 
F-statistic: 311.4 on 2 and 117 DF,  p-value: < 2.2e-16

>
```

The model is a regression model with a linear trend and seasonality for the 12-year span. Since the F-Statistic p-value is so low (2.2e16), significantly lower than an alpha of 5%, the model is statistically significant. The model's R-Square is 84.18%, which indicates that the predictors can account for 95.97% of the variation in the data related to air passengers. The model's adjusted

R-square is 84.18%. This model has 2 trend predictors. Since none of these predictors had p-values higher than an alpha of 5%, they are all significant.

### Forecasting for validation data

```
> train.quad.pred
         Point Forecast     Lo 0     Hi 0
Jan 1959       414.0230 414.0230 414.0230
Feb 1959       417.3694 417.3694 417.3694
Mar 1959       420.7298 420.7298 420.7298
Apr 1959       424.1042 424.1042 424.1042
May 1959       427.4925 427.4925 427.4925
Jun 1959       430.8948 430.8948 430.8948
Jul 1959       434.3110 434.3110 434.3110
Aug 1959       437.7412 437.7412 437.7412
Sep 1959       441.1853 441.1853 441.1853
Oct 1959       444.6435 444.6435 444.6435
Nov 1959       448.1155 448.1155 448.1155
Dec 1959       451.6016 451.6016 451.6016
Jan 1960       455.1016 455.1016 455.1016
Feb 1960       458.6155 458.6155 458.6155
Mar 1960       462.1434 462.1434 462.1434
Apr 1960       465.6853 465.6853 465.6853
May 1960       469.2411 469.2411 469.2411
Jun 1960       472.8109 472.8109 472.8109
Jul 1960       476.3946 476.3946 476.3946
Aug 1960       479.9924 479.9924 479.9924
Sep 1960       483.6040 483.6040 483.6040
Oct 1960       487.2296 487.2296 487.2296
Nov 1960       490.8692 490.8692 490.8692
Dec 1960       494.5228 494.5228 494.5228
>
```

### c. Regression model with seasonality

```
> summary(train.season)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-156.80  -71.28   -8.85   76.17  200.20 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   212.40      29.55   7.187 9.03e-11 ***
season2        -3.70      41.80  -0.089   0.9296    
season3        29.30      41.80   0.701   0.4848    
season4        22.40      41.80   0.536   0.5931    
season5        24.60      41.80   0.589   0.5574    
season6        60.90      41.80   1.457   0.1480    
season7        92.20      41.80   2.206   0.0295 *  
season8        92.40      41.80   2.211   0.0292 *  
season9        53.40      41.80   1.278   0.2041    
season10       20.70      41.80   0.495   0.6214    
season11       -8.20      41.80  -0.196   0.8448    
season12       18.10      41.80   0.433   0.6658    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.46 on 108 degrees of freedom
Multiple R-squared:  0.1205,	Adjusted R-squared:  0.03094 
F-statistic: 1.345 on 11 and 108 DF,  p-value: 0.2098

> 
```

The model is a regression model with a linear trend and seasonality for the 12-year span. Since the F-Statistic p-value is so high (9.03e-11), and has high alpha of 5%, the model is statistically insignificant. The model's R-Square is 12.05%, which indicates that the predictors can account for 12.05% of the variation in the data related to air passengers. The model's adjusted R-square is 3.094%. No trend predictors and 11 dummy variables that indicate seasonality make up this model. Since all these predictors had p-values higher than an alpha of 5%, they are all in-significant.

**Forecasting for validation data**

```
> train.season.pred
         Point Forecast  Lo 0  Hi 0
Jan 1959          212.4 212.4 212.4
Feb 1959          208.7 208.7 208.7
Mar 1959          241.7 241.7 241.7
Apr 1959          234.8 234.8 234.8
May 1959          237.0 237.0 237.0
Jun 1959          273.3 273.3 273.3
Jul 1959          304.6 304.6 304.6
Aug 1959          304.8 304.8 304.8
Sep 1959          265.8 265.8 265.8
Oct 1959          233.1 233.1 233.1
Nov 1959          204.2 204.2 204.2
Dec 1959          230.5 230.5 230.5
Jan 1960          212.4 212.4 212.4
Feb 1960          208.7 208.7 208.7
Mar 1960          241.7 241.7 241.7
Apr 1960          234.8 234.8 234.8
May 1960          237.0 237.0 237.0
Jun 1960          273.3 273.3 273.3
Jul 1960          304.6 304.6 304.6
Aug 1960          304.8 304.8 304.8
Sep 1960          265.8 265.8 265.8
Oct 1960          233.1 233.1 233.1
Nov 1960          204.2 204.2 204.2
Dec 1960          230.5 230.5 230.5
> 
```

### d. Regression model with linear trend and seasonality

```
> summary(train.lin.trend.season)

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

Residuals:
   Min     1Q Median     3Q    Max 
-35.72 -15.64  -2.60  10.79  65.05 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  74.74491    7.55463   9.894  < 2e-16 ***
trend         2.50282    0.05747  43.551  < 2e-16 ***
season2      -6.20282    9.70399  -0.639 0.524057    
season3      24.29436    9.70450   2.503 0.013811 *  
season4      14.89154    9.70535   1.534 0.127892    
season5      14.58872    9.70654   1.503 0.135790    
season6      48.38590    9.70807   4.984 2.41e-06 ***
season7      77.18308    9.70994   7.949 2.05e-12 ***
season8      74.88026    9.71215   7.710 6.87e-12 ***
season9      33.37744    9.71470   3.436 0.000842 ***
season10     -1.82538    9.71759  -0.188 0.851356    
season11    -33.22820    9.72082  -3.418 0.000893 ***
season12     -9.43102    9.72439  -0.970 0.334317    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.7 on 107 degrees of freedom
Multiple R-squared:  0.953,	Adjusted R-squared:  0.9478 
F-statistic: 180.9 on 12 and 107 DF,  p-value: < 2.2e-16
```

The model is a regression model with a linear trend and seasonality for the 12-year span. Since the F-Statistic p-value is so low (2.2e16), significantly lower than an alpha of 5%, the model is statistically significant. The model's R-Square is 95.97%, which indicates that the predictors can account for 95.97% of the variation in the data related to air passengers. The model's adjusted R-square is 95.48%. 2 trend predictors and 11 dummy variables that indicate seasonality make up this model. Since none of these predictors had p-values higher than an alpha of 5%, they are all significant.

**Forecasting for validation data**

```
> train.lin.trend.season.pred
         Point Forecast     Lo 0     Hi 0
Jan 1959       377.5861 377.5861 377.5861
Feb 1959       373.8861 373.8861 373.8861
Mar 1959       406.8861 406.8861 406.8861
Apr 1959       399.9861 399.9861 399.9861
May 1959       402.1861 402.1861 402.1861
Jun 1959       438.4861 438.4861 438.4861
Jul 1959       469.7861 469.7861 469.7861
Aug 1959       469.9861 469.9861 469.9861
Sep 1959       430.9861 430.9861 430.9861
Oct 1959       398.2861 398.2861 398.2861
Nov 1959       369.3861 369.3861 369.3861
Dec 1959       395.6861 395.6861 395.6861
Jan 1960       407.6199 407.6199 407.6199
Feb 1960       403.9199 403.9199 403.9199
Mar 1960       436.9199 436.9199 436.9199
Apr 1960       430.0199 430.0199 430.0199
May 1960       432.2199 432.2199 432.2199
Jun 1960       468.5199 468.5199 468.5199
Jul 1960       499.8199 499.8199 499.8199
Aug 1960       500.0199 500.0199 500.0199
Sep 1960       461.0199 461.0199 461.0199
Oct 1960       428.3199 428.3199 428.3199
Nov 1960       399.4199 399.4199 399.4199
Dec 1960       425.7199 425.7199 425.7199
>
```

**e. Regression model with quadratic trend and seasonality.**

```
> summary(train.quad.trend.season)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-45.381 -11.495  -0.419  11.620  52.024 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  92.338400   8.185328  11.281  < 2e-16 ***
trend         1.631070   0.214586   7.601 1.24e-11 ***
I(trend^2)    0.007205   0.001717   4.195 5.69e-05 ***
season2      -6.130774   9.029011  -0.679 0.498612    
season3      24.424042   9.029523   2.705 0.007962 ** 
season4      15.064449   9.030356   1.668 0.098227 .  
season5      14.790448   9.031498   1.638 0.104460    
season6      48.602037   9.032941   5.381 4.48e-07 ***
season7      77.399217   9.034682   8.567 9.26e-14 ***
season8      75.081988   9.036720   8.309 3.47e-13 ***
season9      33.550350   9.039058   3.712 0.000330 ***
season10     -1.695697   9.041706  -0.188 0.851594    
season11    -33.156153   9.044673  -3.666 0.000387 ***
season12     -9.431019   9.047975  -1.042 0.299628    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.19 on 106 degrees of freedom
Multiple R-squared:  0.9597,	Adjusted R-squared:  0.9548 
F-statistic: 194.3 on 13 and 106 DF,  p-value: < 2.2e-16

>
```

This model is a regression model with a quadratic trend and seasonality for the 12-year span and it is almost the same as the model with a linear trend and seasonality. Since the F-Statistic p-value is so low (2.2e16), significantly lower than an alpha of 5%, the model is statistically significant. The model's R-Square is 95.97%, which indicates that the predictors can account for 95.97% of the variation in the data related to air passengers. The model's adjusted R-square is 95.48%. 2 trend predictors and 11 dummy variables that indicate seasonality make up this model. Since none of these predictors had p-values higher than an alpha of 5%, they are all significant.

**Forecasting for validation data**

```
> train.quad.trend.season.pred
         Point Forecast     Lo 0     Hi 0
Jan 1959       395.1796 395.1796 395.1796
Feb 1959       392.4306 392.4306 392.4306
Mar 1959       426.3816 426.3816 426.3816
Apr 1959       420.4326 420.4326 420.4326
May 1959       423.5836 423.5836 423.5836
Jun 1959       460.8346 460.8346 460.8346
Jul 1959       493.0856 493.0856 493.0856
Aug 1959       494.2366 494.2366 494.2366
Sep 1959       456.1876 456.1876 456.1876
Oct 1959       424.4386 424.4386 424.4386
Nov 1959       396.4896 396.4896 396.4896
Dec 1959       423.7406 423.7406 423.7406
Jan 1960       436.7119 436.7119 436.7119
Feb 1960       434.1358 434.1358 434.1358
Mar 1960       468.2597 468.2597 468.2597
Apr 1960       462.4836 462.4836 462.4836
May 1960       465.8075 465.8075 465.8075
Jun 1960       503.2314 503.2314 503.2314
Jul 1960       535.6553 535.6553 535.6553
Aug 1960       536.9793 536.9793 536.9793
Sep 1960       499.1032 499.1032 499.1032
Oct 1960       467.5271 467.5271 467.5271
Nov 1960       439.7510 439.7510 439.7510
Dec 1960       467.1749 467.1749 467.1749
>
```

**Comparison of Accuracies:**

```
> round(accuracy(train.lin.pred$mean, valid.ts),3)
             ME   RMSE    MAE  MPE   MAPE  ACF1 Theil's U
Test set 26.708 74.788 54.938 3.75 11.215 0.701     1.326
> round(accuracy(train.quad.pred$mean, valid.ts),3)
             ME   RMSE    MAE    MPE   MAPE  ACF1 Theil's U
Test set -1.434 69.128 56.336 -2.533 12.289 0.694     1.307
> round(accuracy(train.season.pred$mean, valid.ts),3)
              ME    RMSE     MAE    MPE   MAPE  ACF1 Theil's U
Test set 206.342 211.388 206.342 45.276 45.276 0.748     4.148
> round(accuracy(train.lin.trend.season.pred$mean, valid.ts),3)
             ME   RMSE    MAE  MPE MAPE  ACF1 Theil's U
Test set 26.139 47.944 34.638 4.59 6.88 0.698     0.838
> round(accuracy(train.quad.trend.season.pred$mean, valid.ts),3)
            ME   RMSE    MAE    MPE MAPE  ACF1 Theil's U
Test set -2.91 38.628 31.233 -1.893 6.87 0.675     0.741
>
```

By taking MAPE and RMSE values into consideration, we can say that, Regression model of Quadratic trend and seasonality is the first best regression model with least MAPE (6.87) and RMSE (38.628) values.

**FORECASTING ENTIRE DATASET:**

A model must be built utilizing the complete dataset to generate forecasts. Along with variants of the models, both the regression with linear trend and seasonality and the regression with quadratic trend and seasonality were used. The changes include applying an autoregressive model and using a trailing moving average for the model's residuals. Theoretically, the adoption of these multilevel models will lead to more precise forecasts.

**A. Regression model with linear trend and seasonality.**

```
> summary(tlm.season)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-42.121 -18.564  -3.268  15.189  95.085 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  63.50794    8.38856   7.571 5.88e-12 ***
trend         2.66033    0.05297  50.225  < 2e-16 ***
season2      -9.41033   10.74941  -0.875 0.382944    
season3      23.09601   10.74980   2.149 0.033513 *  
season4      17.35235   10.75046   1.614 0.108911    
season5      19.44202   10.75137   1.808 0.072849 .  
season6      56.61502   10.75254   5.265 5.58e-07 ***
season7      93.62136   10.75398   8.706 1.17e-14 ***
season8      90.71103   10.75567   8.434 5.32e-14 ***
season9      39.38403   10.75763   3.661 0.000363 ***
season10      0.89037   10.75985   0.083 0.934177    
season11    -35.51996   10.76232  -3.300 0.001244 ** 
season12     -9.18029   10.76506  -0.853 0.395335    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.33 on 131 degrees of freedom
Multiple R-squared:  0.9559,	Adjusted R-squared:  0.9518 
F-statistic: 236.5 on 12 and 131 DF,  p-value: < 2.2e-16

> 
```

The model is a regression model with a linear trend and seasonality for the 12-year span. Since the F-Statistic p-value is so low (2.2e16), significantly lower than an alpha of 5%, the model is statistically significant. The model's R-Square is 95.59%, which indicates that the predictors can account for 95.59% of the variation in the data related to air passengers. The model's adjusted R-square is 95.18%. 2 trend predictors and 11 dummy variables that indicate seasonality make up this model. Since none of these predictors had p-values higher than an alpha of 5%, they are all significant.

**Forecast data with linear trend and seasonality data in 12 future periods.**

```
> lin.season.pred
         Point Forecast     Lo 0     Hi 0
Jan 1961       449.2557 449.2557 449.2557
Feb 1961       442.5057 442.5057 442.5057
Mar 1961       477.6723 477.6723 477.6723
Apr 1961       474.5890 474.5890 474.5890
May 1961       479.3390 479.3390 479.3390
Jun 1961       519.1723 519.1723 519.1723
Jul 1961       558.8390 558.8390 558.8390
Aug 1961       558.5890 558.5890 558.5890
Sep 1961       509.9223 509.9223 509.9223
Oct 1961       474.0890 474.0890 474.0890
Nov 1961       440.3390 440.3390 440.3390
Dec 1961       469.3390 469.3390 469.3390
>
```

**B. Regression model with quadratic trend and seasonality.**

```
> summary(quad.season)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-46.122 -13.394   0.825  12.733  75.773 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  88.557838   8.799762  10.064  < 2e-16 ***
trend         1.625504   0.191722   8.478 4.34e-14 ***
I(trend^2)    0.007137   0.001281   5.573 1.38e-07 ***
season2      -9.338962   9.694487  -0.963 0.337172    
season3      23.224469   9.694859   2.396 0.018021 *  
season4      17.523627   9.695469   1.807 0.073012 .  
season5      19.641845   9.696310   2.026 0.044842 *  
season6      56.829122   9.697379   5.860 3.59e-08 ***
season7      93.835460   9.698673   9.675  < 2e-16 ***
season8      90.910857   9.700192   9.372 2.91e-16 ***
season9      39.555314   9.701939   4.077 7.90e-05 ***
season10      1.018831   9.703917   0.105 0.916544    
season11    -35.448592   9.706131  -3.652 0.000376 ***
season12     -9.180288   9.708591  -0.946 0.346115    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.75 on 130 degrees of freedom
Multiple R-squared:  0.9644,	Adjusted R-squared:  0.9608 
F-statistic: 270.8 on 13 and 130 DF,  p-value: < 2.2e-16

> 
```

The model is a regression model with a linear trend and seasonality for the 12-year span. Since the F-Statistic p-value is so low (2.2e16), significantly lower than an alpha of 5%, the model is statistically significant. The model's R-Square is 96.64%, which indicates that the predictors can account for 96.64% of the variation in the data related to air passengers. The model's adjusted R-square is 96.08%. 2 trend predictors and 11 dummy variables that indicate seasonality make up this model. Since none of these predictors had p-values higher than an alpha of 5%, they are all significant.

***Forecasting data to predictions with quadratic trend and seasonality data in 12 future periods:***

```
> quad.season.pred
         Point Forecast    Lo 0     Hi 0
Jan 1961       474.3056 474.3056 474.3056
Feb 1961       468.6689 468.6689 468.6689
Mar 1961       504.9489 504.9489 504.9489
Apr 1961       502.9789 502.9789 502.9789
May 1961       508.8422 508.8422 508.8422
Jun 1961       549.7889 549.7889 549.7889
Jul 1961       590.5689 590.5689 590.5689
Aug 1961       591.4322 591.4322 591.4322
Sep 1961       543.8789 543.8789 543.8789
Oct 1961       509.1589 509.1589 509.1589
Nov 1961       476.5222 476.5222 476.5222
Dec 1961       506.6355 506.6355 506.6355
>
```

```
> round(accuracy(lin.season.pred$fitted, airpassenger.ts),3)
         ME   RMSE    MAE   MPE  MAPE  ACF1 Theil's U
Test set  0 25.114 19.774 0.604 8.592 0.763      1.09
> round(accuracy(lin.pred$fitted, airpassenger.ts),3)
         ME   RMSE    MAE    MPE   MAPE  ACF1 Theil's U
Test set  0 45.736 34.406 -1.291 12.319 0.728     1.372
> round(accuracy(quad.season.pred$fitted, airpassenger.ts),3)
         ME   RMSE    MAE   MPE  MAPE  ACF1 Theil's U
Test set  0 22.562 17.579 0.096 7.273 0.712     0.935
> round(accuracy((naive(airpassenger.ts))$fitted, airpassenger.ts), 3)
            ME  RMSE   MAE   MPE  MAPE  ACF1 Theil's U
Test set 2.238 33.71 25.86 0.378 9.019 0.303         1
> round(accuracy((snaive(airpassenger.ts))$fitted, airpassenger.ts), 3)
             ME   RMSE   MAE    MPE   MAPE  ACF1 Theil's U
Test set 31.773 36.316 32.03 11.124 11.249 0.746     1.132
>
```

**Conclusion:**

From the above accuracy measures, we can conclude that out of all the regression models, the regression model with Quadratics trend and Seasonality is the best model with lowest RSME (22.562) and MAPE (7.273) values.

## Autoregressive Integrated Moving Average Models:

The ARIMA model is a versatile tool that is suitable for making forecasts on data that exhibit level, trend, and seasonal patterns. Given that if data displays all three of these characteristics, using an ARIMA model for analysis is a suitable approach. We generated an optimal ARIMA model by utilizing the auto.arima() function, which automatically selects the (p,d,q) and (P, D ,Q) parameters based on the AIC score.

**Below is the output for the ARIMA Model for the validation data set.**

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

Coefficients:
          ar1
      -0.2397
s.e.   0.0935

sigma^2 = 103.6:  log likelihood = -399.64
AIC=803.28   AICc=803.4   BIC=808.63

Training set error measures:
                      ME     RMSE      MAE         MPE    MAPE      MASE       ACF1
Training set -0.01614662 9.567988 7.120167 -0.03346415 2.90195 0.2491828 0.00821521
```

The validation data set coefficients consist of an AR-1 coefficient with a value of -0.2397.

**Below is the output for the ARIMA Model for the entire data set.**

```
> # use summary() to show auto ARIMA model and its parameters for entire data set.
> auto.arima <- auto.arima(airpassenger.ts)
> summary(auto.arima)
Series: airpassenger.ts
ARIMA(2,1,1)(0,1,0)[12]

Coefficients:
         ar1     ar2      ma1
      0.5960  0.2143  -0.9819
s.e.  0.0888  0.0880   0.0292

sigma^2 = 132.3:  log likelihood = -504.92
AIC=1017.85   AICc=1018.17   BIC=1029.35

Training set error measures:
                 ME     RMSE     MAE      MPE     MAPE     MASE        ACF1
Training set 1.3423 10.84619 7.86754 0.420698 2.800458 0.245628 -0.00124847
> 
```

The entire series coefficients consist of an AR-1 coefficient, an AR-2 coefficient and an order 1 moving average component with values of 0.5960, 0.2143, and -0.9819, respectively lagged one period.

The table below explains the unique model structure for each respective time series data.

| | Training Dataset | | Entire Dataset | |
|---|---|---|---|---|
| p | 2 | autoregressive model of 2nd order (AR2) | 1 | autoregressive model of 1st order (AR1) |
| d | 1 | order 1 differencing to remove linear trend | 1 | order 1 differencing to remove linear trend |
| q | 1 | order 1 moving average (MA1) model for error lags | 0 | No moving average model for error lags |
| P | 0 | No autoregressive model (AR) for seasonality | 0 | No autoregressive model (AR) for seasonality |
| D | 1 | order 1 differencing to remove a linear trend | 1 | order 1 differencing to remove a linear trend |
| Q | 0 | No moving average model for error lags | 0 | No moving average model for error lags |
| m | 12 | For Monthly seasonality | 12 | For Monthly seasonality |

**Below is the forecast for validation dataset:**

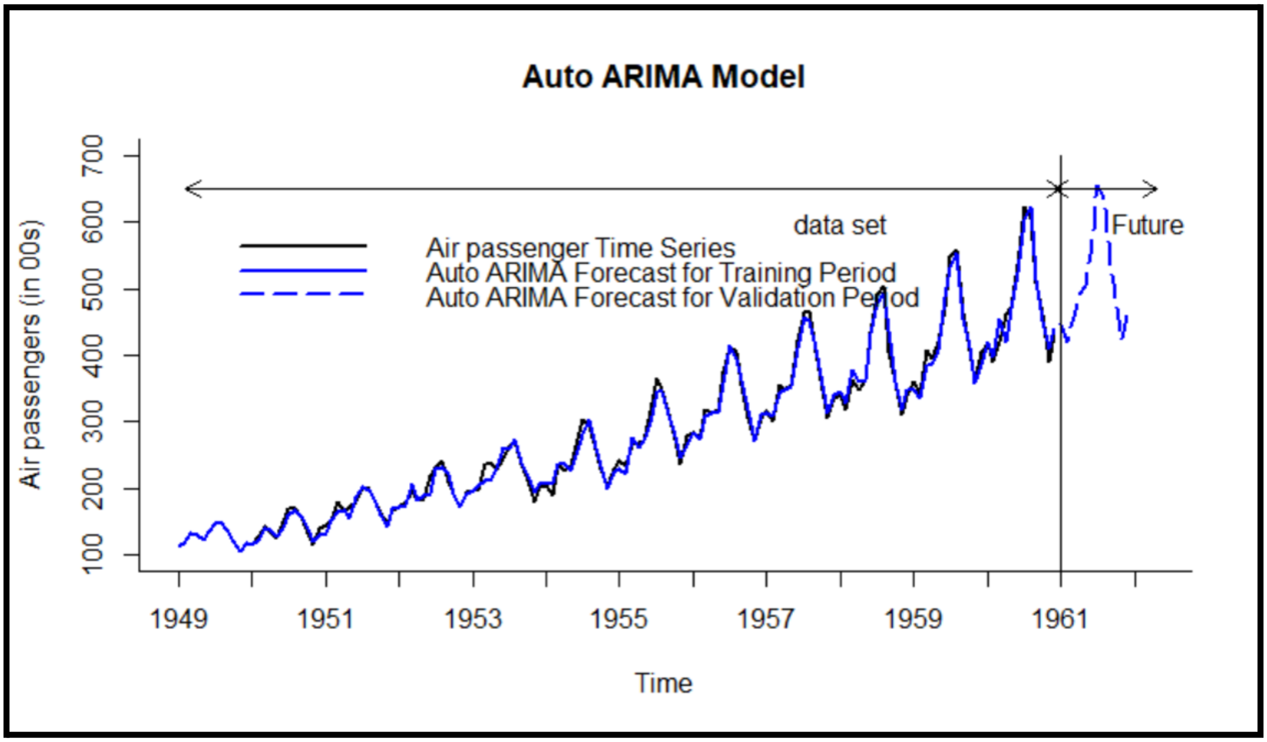
| | Point Forecast | Lo 0 | Hi 0 |
|---|---|---|---|
| Jan 1959 | 341.9589 | 341.9589 | 341.9589 |
| Feb 1959 | 319.7290 | 319.7290 | 319.7290 |
| Mar 1959 | 363.7842 | 363.7842 | 363.7842 |
| Apr 1959 | 349.7709 | 349.7709 | 349.7709 |
| May 1959 | 364.7741 | 364.7741 | 364.7741 |
| Jun 1959 | 436.7734 | 436.7734 | 436.7734 |
| Jul 1959 | 492.7735 | 492.7735 | 492.7735 |
| Aug 1959 | 506.7735 | 506.7735 | 506.7735 |
| Sep 1959 | 405.7735 | 405.7735 | 405.7735 |
| Oct 1959 | 360.7735 | 360.7735 | 360.7735 |
| Nov 1959 | 311.7735 | 311.7735 | 311.7735 |
| Dec 1959 | 338.7735 | 338.7735 | 338.7735 |
| Jan 1960 | 343.7324 | 343.7324 | 343.7324 |
| Feb 1960 | 321.5025 | 321.5025 | 321.5025 |
| Mar 1960 | 365.5577 | 365.5577 | 365.5577 |
| Apr 1960 | 351.5444 | 351.5444 | 351.5444 |
| May 1960 | 366.5476 | 366.5476 | 366.5476 |
| Jun 1960 | 438.5468 | 438.5468 | 438.5468 |
| Jul 1960 | 494.5470 | 494.5470 | 494.5470 |
| Aug 1960 | 508.5470 | 508.5470 | 508.5470 |
| Sep 1960 | 407.5470 | 407.5470 | 407.5470 |
| Oct 1960 | 362.5470 | 362.5470 | 362.5470 |
| Nov 1960 | 313.5470 | 313.5470 | 313.5470 |
| Dec 1960 | 340.5470 | 340.5470 | 340.5470 |

The forecasts below represent the 12 future period predictions of the ARIMA model for the year 1961.

```
> auto.arima.pred
```

| | Point Forecast | Lo 0 | Hi 0 |
|---|---|---|---|
| Jan 1961 | 445.6349 | 445.6349 | 445.6349 |
| Feb 1961 | 420.3950 | 420.3950 | 420.3950 |
| Mar 1961 | 449.1983 | 449.1983 | 449.1983 |
| Apr 1961 | 491.8399 | 491.8399 | 491.8399 |
| May 1961 | 503.3945 | 503.3945 | 503.3945 |
| Jun 1961 | 566.8624 | 566.8624 | 566.8624 |
| Jul 1961 | 654.2602 | 654.2602 | 654.2602 |
| Aug 1961 | 638.5975 | 638.5975 | 638.5975 |
| Sep 1961 | 540.8837 | 540.8837 | 540.8837 |
| Oct 1961 | 494.1266 | 494.1266 | 494.1266 |
| Nov 1961 | 423.3327 | 423.3327 | 423.3327 |
| Dec 1961 | 465.5076 | 465.5076 | 465.5076 |

The plot below is based on the ARIMA model for the 12-year series From the plot, it is apparent that the model seems to be fitting well into the historical data. Trend and seasonality seem to be taken into consideration.




**<u>Accuracy Measures:</u>**

```
> round(accuracy(auto.arima.pred$fitted, airpassenger.ts), 3)
            ME   RMSE   MAE   MPE MAPE   ACF1 Theil's U
Test set 1.342 10.846 7.868 0.421  2.8 -0.001     0.376
> round(accuracy((snaive(airpassenger.ts))$fitted, airpassenger.ts), 3)
             ME   RMSE   MAE    MPE   MAPE  ACF1 Theil's U
Test set 31.773 36.316 32.03 11.124 11.249 0.746     1.132
> round(accuracy((naive(airpassenger.ts))$fitted, airpassenger.ts), 3)
            ME  RMSE   MAE   MPE  MAPE  ACF1 Theil's U
Test set 2.238 33.71 25.86 0.378 9.019 0.303         1
>
```

From the above accuracy score, we can see that auto arima is performing better as RMSE and MAPE values of auto arima are less than naive and seasonal naive, 10.846 and 33.71 respectively.

**Advanced Exponential Smoothing:**

In our time series analysis, advanced exponential smoothing is the next model employed, specifically using the Holt-Winters model. This model is particularly beneficial because it takes into account both the trend and seasonality elements when generating forecasts. To ensure accurate results, we first evaluated the model's performance by testing it on the training and validation sections before running it on the complete dataset.

### Automated Holt-Winters Model

We utilized an automated Holt-Winters Model with appropriate training and validation partitions. The code automatically selects the error, trend, and seasonality components through the c(Z, Z,Z) parameter. By not specifying default values for alpha (error), beta (trend), or gamma (seasonality) in the code, the model will provide us with "optimized" values.

The screenshot table displays the smoothing values of the Holt-Winters Model for the training dataset.

```
> hw.ZZZ
ETS(M,Ad,M)

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

  Smoothing parameters:
    alpha = 0.7459
    beta  = 0.0189
    gamma = 3e-04
    phi   = 0.9793

  Initial states:
    l = 120.667
    b = 1.7375
    s = 0.8978 0.7964 0.919 1.0576 1.2072 1.218
           1.1113 0.9779 0.9838 1.0253 0.8973 0.9084

  sigma:  0.0381

     AIC     AICc      BIC
1110.450 1117.222 1160.625
```

The notation (M, Ad, M) represents the seasonal decomposition of the Error-Trend-Seasonality model. The first component, denoted as "M," refers to the error term or the random fluctuations that cannot be explained by the trend or seasonality. The second component, "Ad," represents the trend component, which captures the long-term upward or downward movement in the time series data. The third component, also denoted as "M," represents the seasonal component, which captures the repeating patterns or cycles within the time series data.

Smoothing parameters:

alpha = 0.7459

beta = 0.0189

gamma = 3e-04

phi = 0.9793

The presence of small beta and gamma values in the model indicates that the trend and seasonal components are changing at a relatively slow pace over time. This suggests that the overall trend and seasonality in the data are being gradually adjusted or modified. Consequently, we can conclude that the trend and seasonal components of the models are globally adjusted, meaning they exhibit a more gradual and stable pattern of change rather than sudden or rapid fluctuations.

**Forecast data for the validation period using this Holt-Winters model**

```
> hw.ZZZ.pred
         Point Forecast     Lo 0     Hi 0
Jan 1959       345.4758 345.4758 345.4758
Feb 1959       342.0246 342.0246 342.0246
Mar 1959       391.6908 391.6908 391.6908
Apr 1959       376.6639 376.6639 376.6639
May 1959       375.2408 375.2408 375.2408
Jun 1959       427.3115 427.3115 427.3115
Jul 1959       469.3286 469.3286 469.3286
Aug 1959       466.1152 466.1152 466.1152
Sep 1959       409.1393 409.1393 409.1393
Oct 1959       356.2006 356.2006 356.2006
Nov 1959       309.2821 309.2821 309.2821
Dec 1959       349.2780 349.2780 349.2780
Jan 1960       354.0668 354.0668 354.0668
Feb 1960       350.3343 350.3343 350.3343
Mar 1960       400.9889 400.9889 400.9889
Apr 1960       385.4007 385.4007 385.4007
May 1960       383.7459 383.7459 383.7459
Jun 1960       436.7762 436.7762 436.7762
Jul 1960       479.4876 479.4876 479.4876
Aug 1960       475.9757 475.9757 475.9757
Sep 1960       417.5986 417.5986 417.5986
Oct 1960       363.3989 363.3989 363.3989
Nov 1960       315.3912 315.3912 315.3912
Dec 1960       356.0217 356.0217 356.0217
```

**The accuracy measures for Holt-Winters model (for the validation period)**

```
> ## ACCURACY MEASURES FOR HW MODEL WITH AUTOMATED SELECTION OF MODEL OPTIONS.
> round(accuracy(hw.ZZZ.pred$mean, valid.ts), 3)
             ME   RMSE    MAE    MPE   MAPE  ACF1 Theil's U
Test set 63.211 72.548 63.213 13.303 13.303 0.746     1.357
```

The RMSE is a commonly used measure of the model's prediction accuracy. It quantifies the average magnitude of the prediction errors. In this case, the RMSE is 72.548, which suggests that, on average, the predictions differ from the actual values by approximately 72.548 units. The MAPE calculates the average absolute percentage difference between the predicted values and the actual values. It provides a measure of the model's average prediction error magnitude relative to the actual values, irrespective of the direction of the errors. In this case, the MAPE is 13.303%, indicating that, on average, the predictions differ from the actual values by approximately 13.303%.

The screenshot table displays the smoothing values of the Holt-Winters Model for the entire dataset. The model was fitted using the ETS (Error-Trend-Seasonality) framework.

```
> HW.ZZZ
ETS(M,Ad,M)

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

  Smoothing parameters:
    alpha = 0.7096
    beta  = 0.0204
    gamma = 1e-04
    phi   = 0.98

  Initial states:
    l = 120.9939
    b = 1.7705
    s = 0.8944 0.7993 0.9217 1.0592 1.2203 1.2318
           1.1105 0.9786 0.9804 1.011 0.8869 0.9059

  sigma:  0.0392

     AIC     AICc      BIC
1395.166 1400.638 1448.623
```

Below mentioned are the exponential smoothing parameters for entire Airpassenger data set:

alpha = 0.7096

beta = 0.0204

gamma = 1e-04

phi   = 0.98

These metrics also provide insights into the model's fit and can be used to compare different models.

**Forecast data for future 12 months using this Holt-Winters model.**

```
> HW.ZZZ.pred
         Point Forecast     Lo 0     Hi 0
Jan 1961       441.8018 441.8018 441.8018
Feb 1961       434.1186 434.1186 434.1186
Mar 1961       496.6300 496.6300 496.6300
Apr 1961       483.2375 483.2375 483.2375
May 1961       483.9914 483.9914 483.9914
Jun 1961       551.0244 551.0244 551.0244
Jul 1961       613.1797 613.1797 613.1797
Aug 1961       609.3648 609.3648 609.3648
Sep 1961       530.5408 530.5408 530.5408
Oct 1961       463.0332 463.0332 463.0332
Nov 1961       402.7478 402.7478 402.7478
Dec 1961       451.9694 451.9694 451.9694
>
```

**The accuracy measures for Holt-Winters model (for the entire dataset) comparision**

```
> round(accuracy(HW.ZZZ.pred$fitted, airpassenger.ts), 3)
            ME   RMSE   MAE   MPE  MAPE  ACF1 Theil's U
Test set 1.567 10.747 7.792 0.436 2.858 0.039     0.352
> round(accuracy((naive(airpassenger.ts))$fitted, airpassenger.ts), 3)
            ME  RMSE   MAE   MPE  MAPE  ACF1 Theil's U
Test set 2.238 33.71 25.86 0.378 9.019 0.303         1
> round(accuracy((snaive(airpassenger.ts))$fitted, airpassenger.ts), 3)
             ME   RMSE   MAE    MPE   MAPE  ACF1 Theil's U
Test set 31.773 36.316 32.03 11.124 11.249 0.746     1.132
```

From the above accuracy scores, we can see that automated Holt-Winters Model is performing better as RMSE (10.747) and MAPE (2.858) values of automated Holt-Winters Model are less than naive and seasonal naive.

**Step 8: Conclusion**

```
> # Identify performance measures compare.
> round(accuracy(quad.season.pred$fitted, airpassenger.ts), 3)
         ME   RMSE    MAE   MPE  MAPE  ACF1 Theil's U
Test set  0 22.562 17.579 0.096 7.273 0.712     0.935
> round(accuracy(auto.arima.pred$fitted, airpassenger.ts), 3)
            ME   RMSE   MAE   MPE MAPE   ACF1 Theil's U
Test set 1.342 10.846 7.868 0.421  2.8 -0.001     0.376
> round(accuracy(HW.ZZZ.pred$fitted, airpassenger.ts), 3)
            ME   RMSE   MAE   MPE  MAPE  ACF1 Theil's U
Test set 1.567 10.747 7.792 0.436 2.858 0.039     0.352
>
```

From the above performance measures, we can see that automated Holt-Winters Model is performing better as RMSE (10.747) and MAPE (2.858) values of automated Holt-Winters Model are less than quadratic trend and seasonality and the Auto ARIMA model.