UNIVERSITAT DE BARCELONA

2122PA_3

Time Series Analysis

Final Project

Prepared By: Mert Mecit

Abstract

In this assignment, different types of time series models are used for forecasting US airline passengers in 1961. Dataset that is used is from Bureau of Transportation Statistics (BTS) from the US. Monthly data from 1949 to 1960 is used. Before model selection and forecast some analyses are made. Descriptive statistics show a relatively right skewed but not much dispersed distribution. For decomposition, a multiplicative model is used. Decomposition shows a clear trend and seasonal component that will be considered later. Also, decomposition shows that residuals exhibit a serial correlation problem. Then, stationarity of the series is checked. We see that, they are non-stationary. Finally, we checked autocorrelation and partial autocorrelations. We see strong autocorrelation with 1st and 12nd lags. Also, for some lags, we see strong partial autocorrelations as well. Then, we came to the most important part. Model selection is done by selecting best model that we can get by using AR, ARIMA and SARIMA. Then, we compared the best of the best by comparing them via RMSE. We see that SARIMA does the best job among three and we used SARIMA for our 1961 forecasts. We see that the trend and seasonal components continues with this model.

1. Introduction

This project is done for University of Barcelona's 2122PA_3 Time Series Analysis course. In this project, I will utilize a univariate time series dataset and apply different analyses on the data.

First, I will introduce my dataset. Then, descriptive statistics of the dataset will be provided. Before models, I will check different considerations in time series analyses: decomposition to see whether there are trends, seasonality and whether residuals are iid. I will check stationarity and autocorrelation and partial autocorrelation to understand mechanics of our dataset. Finally, I will look at different models and select the best among them to forecast.

2. Dataset

The dataset that will be utilized throughout this assignment is called "Air Passengers" dataset. This dataset provides monthly totals of a US airline passengers from 1949 to 1960. The link is here. Air passengers dataset consists of two columns:

- Month: Indicator of the month. I preferred to use a monthly data to understand the effects of any possible seasonality in the data.

- Passengers: Total number of passengers travelled on that particular month. For visualization purposes, they are in 100 thousands.

Figure 1 provides a visualization of the dataset.

Airport demand forecasting is crucial for understanding and changing strategic plans for operations and infrastructure for airports. Especially in post-pandemic period, it become rather harder for airports to correctly forecast passenger demand and act accordingly. Although this dataset will not grasp the mechanics of the pandemic era as it covers a timespan between 1949 and 1960, it will give some insights about what kind of methods can be utilized for this purpose.

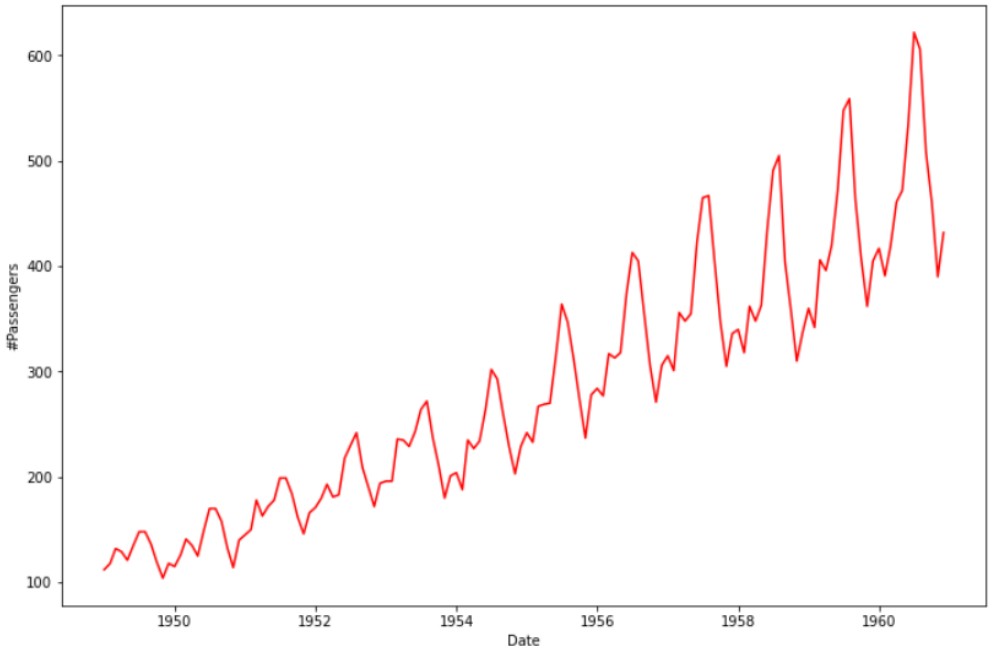

Figure 1. Visualization of the Dataset

3. Descriptive Statistics

Before delving into our analysis, it is always a good choice to start the analysis with descriptive statistics as it can give us some insights.

As table 1 shows, we have a total of 144 observations. This will be enough to utilize all the tests or models that we need to use because generally they require 100 as the sample size at most. On average of those 12 years, we have 280 hundred thousand passengers each month. Our standard deviation of 119 indicates that, compared to our mean, we do not have a relatively dispersed distribution. We can also see that our min value is 104 and max value is 622. Our middle value (or median) of 265 means that our data is relatively skewed towards right.

Count	**144**
Mean	280.3
Std	119.97
Min	104
25%	180
50%	265
75%	360
Max	622

Table 1. Descriptive Statistics

4. Decomposition

We can understand ongoing trends and seasonality in our data by decomposing it. First, we need to determine whether our data is showing additive or multiplicative behavior. To have an idea about it, we can check figure 1. In figure 1, it is clear that the variation in the seasonal pattern appears to be proportional to the level of the time series. Or in other words, as we progress in time, we see that the seasonal effect is affecting the whole model more and more. So, we can assume that our model is **multiplicative**. In multiplicative decomposition, we have:

$$y_t = S_t * T_t * R_t$$

Where y_t is the data, S_t is the seasonal component, T_t is the trend component and R_t is the residual.

When we decompose our model, it looks like this:

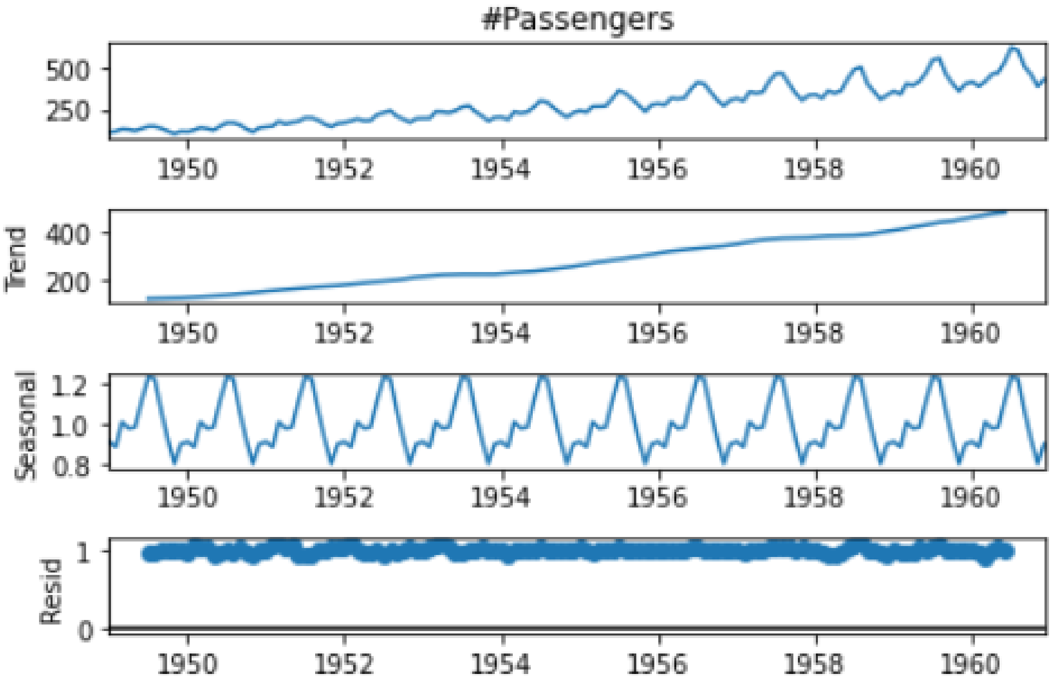

Figure 2. Decomposition

On the top of the figure 2, we have our model. In the trend part, as it can be clearly seen, we have an increasing trend. This was expected because as the time progress, more and more people are using airport system in the US. In the seasonal part, it is also expected as in the spring and summer days, more people tend to use airports then winter and autumn months. Residual part is not so clear in this picture, and it requires a different type of attention. So, I re-draw it in figure 3.

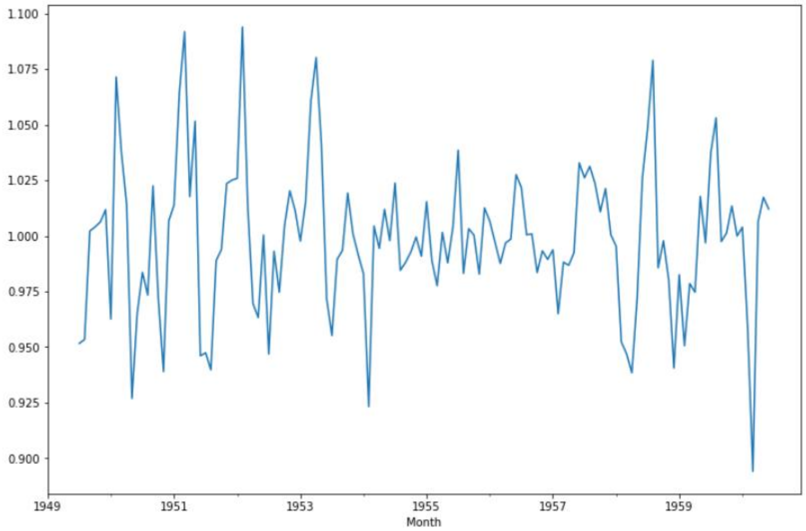

Figure 3. Residuals

Ideally, we want our residuals to be independent. To test this, we can use Ljung-Box test. It has following hypotheses:

$$H_0 = The\ residuals\ are\ iid.$$
$$H_1 = The\ residuals\ exhibit\ serial\ correlation.$$

So, ideally, we would like to fail to reject the null hypothesis. When I perform the test, as results are shown in table 2, we reject the null hypothesis and conclude that, in this dataset, residuals are showing a serial correlation problem.

LB_Stat	LB_Pvalue
51.15	0.0008

Table 2. Ljung-Box Test

This may cause some problems in our future analysis. So, while making comments on our findings, we should always be careful. For example, for OLS estimators, the standard errors will be smaller than the true standard errors, which will make parameter estimates more precise than what they really are.

5. Stationarity

Apart from decomposition, stationarity is also a key part of time series analysis. Stationarity is assumed in many types of forecasting methods, and it means that the time series is not dependent on the time at which the series is observed.

I will utilize Dickey-Fuller test to check stationarity. I will use Akaike Information Criterion (AIC). This test has:

$$H_0 = No\ stationarity$$
$$H_1 = Stationarity$$

The results are given in the table 3. Although it was obvious from the figure 1 that there is a trend and seasonality in our data, Dickey-Fuller test also supports that and shows that there is no stationarity in our data.

Test Stat	p-value	No of lags used
0.82	0.99	13

Table 3. Dickey-Fuller Test

6. Autocorrelation and Partial Autocorrelation

The final aspect before making models and predictions on our dataset is to check autocorrelation and partial autocorrelation. Autocorrelation is the correlation of time series observations with the previous observations. We can give an example interpretation like: if there is a strong autocorrelation in our dataset and we observe a dataset with a high number of passengers, we are very likely to get a high number of passengers in the next time as well. Figure 4 shows that autocorrelation in our data is very high. For a 1-month lag, our autocorrelation is 0.96. This number then declines, however, when we check the 12-month lag, it again reaches almost the same level as 0.96. This is expected as generally in the same month of the previous year, we are likely to experience similar level of passengers.

On the other hand, partial autocorrelation is similar, however this time, the intermediate lags are removed. We only care about the lags. I calculated lags for only first 50 lags. It shows that for the first lags or lags that are multiples of 12, we have strong partial autocorrelations compared to other lags.

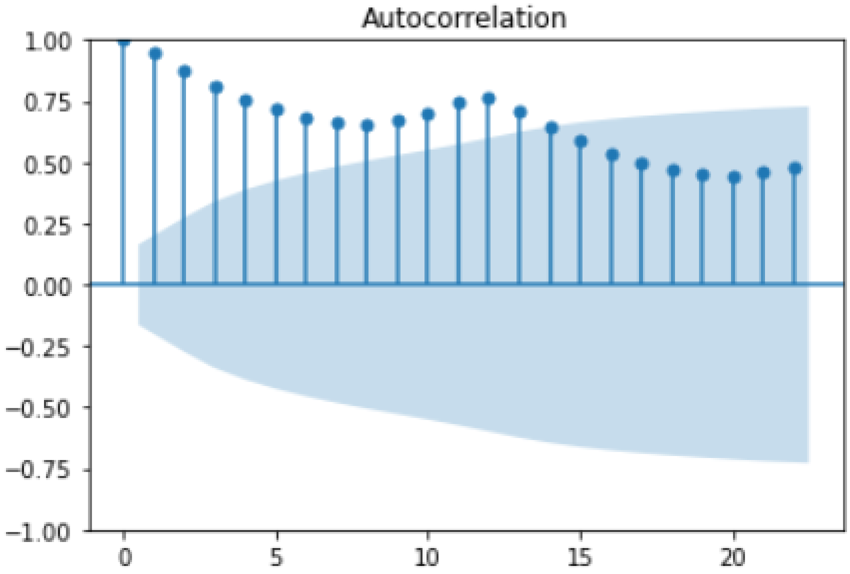

Figure 4. Autocorrelation

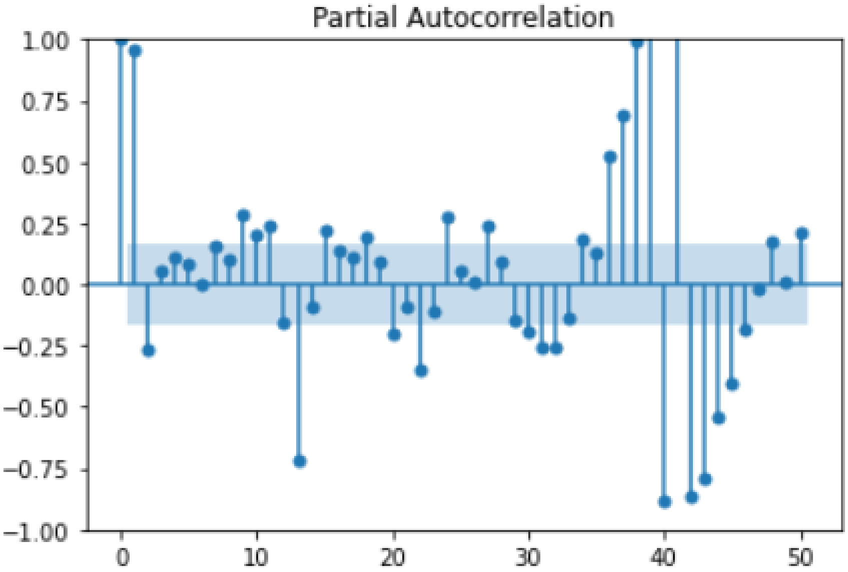

Figure 5. Partial Autocorrelation

7. Model Selection & Forecasting

Now, I will utilize some models to understand our model. For that purpose, I will divide my dataset into training and test dataset. For this purpose, I will select the first 11 years as my training data and the last year for test data. Figure 6 shows how they are divided. Red line will be used for test and black line will be used for training. I will compare the model performances based on how well they perform on those test data.

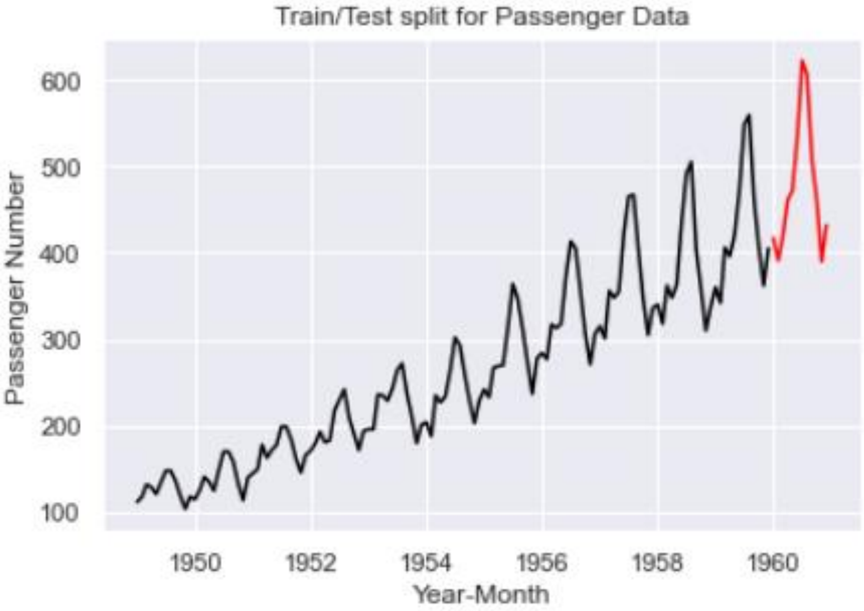

Figure 6. Test Train Split

7.1 AR Model

The first model I will utilize is an AR model. It is a representation of a type of random process. We forecast the variable of interest using linear combination of past values of the time series. Here, I used up to 10 lags and performed their performances. The best came out to be using 2 lags. When we use 2 lags, predictions on test data are as follows:

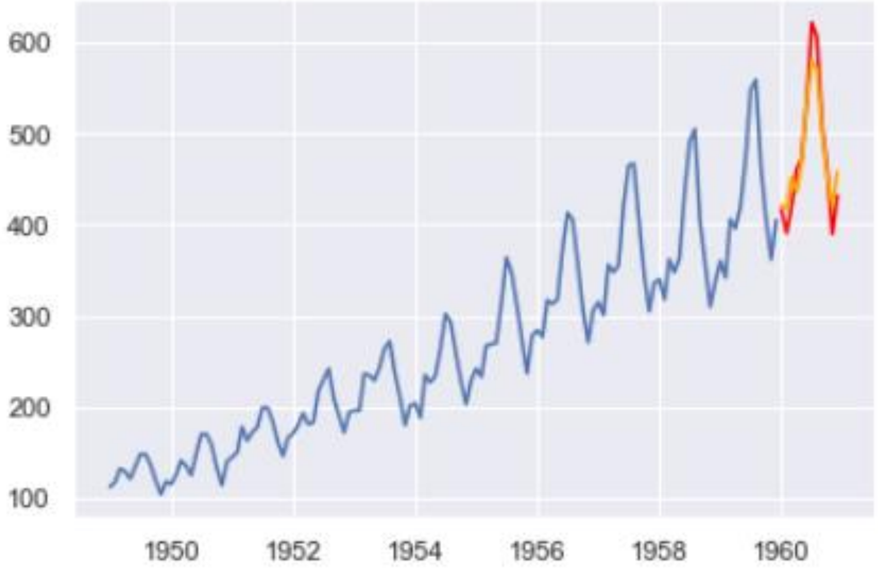

Figure 7. AR Model Performance

I used AIC while I choose among different lags of AR. However, to see results, I used RMSE. RMSE of the AR (2) model is 24.385.

7.2 ARIMA Model

ARIMA models are applied generally when non-stationarity behavior is observed in our data. Here, we include differencing step to eliminate this non-stationary behavior. When I check models up to 10 and used AIC as my selection criteria, the best model comes out to be ARIMA (2,1,2) with a RMSE of 33.29. Predictions are shown with an orange line in figure 8. This is quite interesting as I would expect this to perform better than AR model as our data is non-stationary.

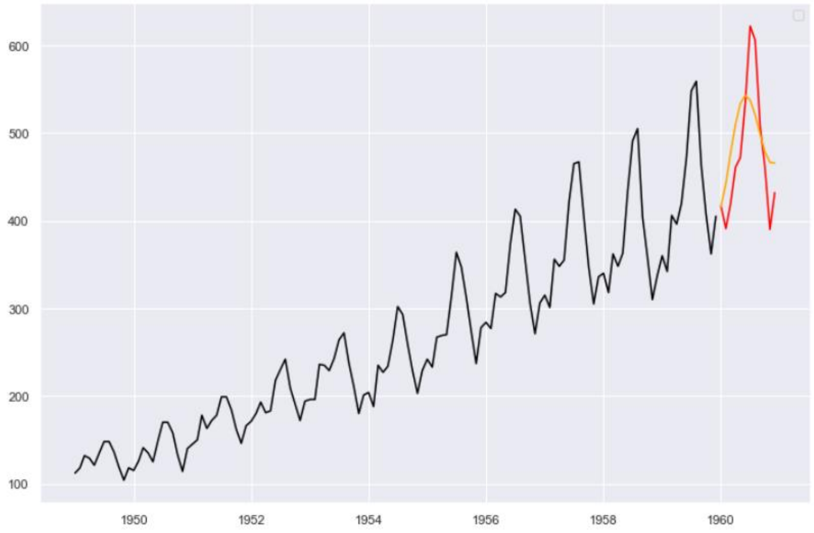

Figure 8. ARIMA Model Performance

7.3 SARIMA Model

There are different types of ARIMA models to be applied for different situations. One of the models that are widely used is SARIMA, which stands for seasonal ARIMA. It is widely used when a seasonal effect is observed, which is the case in our model. When I apply a SARIMA model with (1,1,1) $(1,1,1)_{12}$, I get the best result I can get both among SARIMA models and among overall. We have a RMSE of 21.19. Figure 9 shows that its almost a perfect fit.

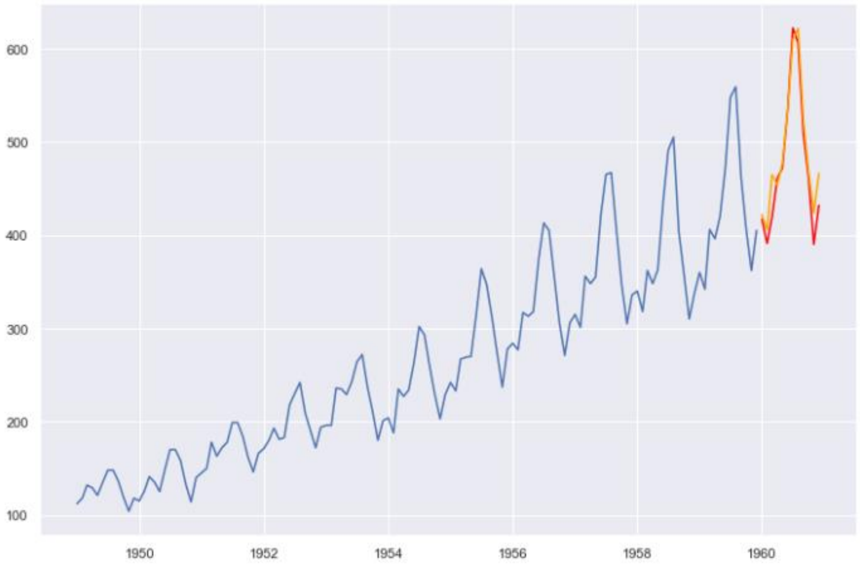

Figure 9. SARIMA Model Performance

As I have chosen my best model, now it is time to do some forecasting with it. When I do my forecast with SARIMA model for the next year, figure 10 shows that, as expected, our multiplicative effect will continue, and we will have similar trend and seasonal effects.

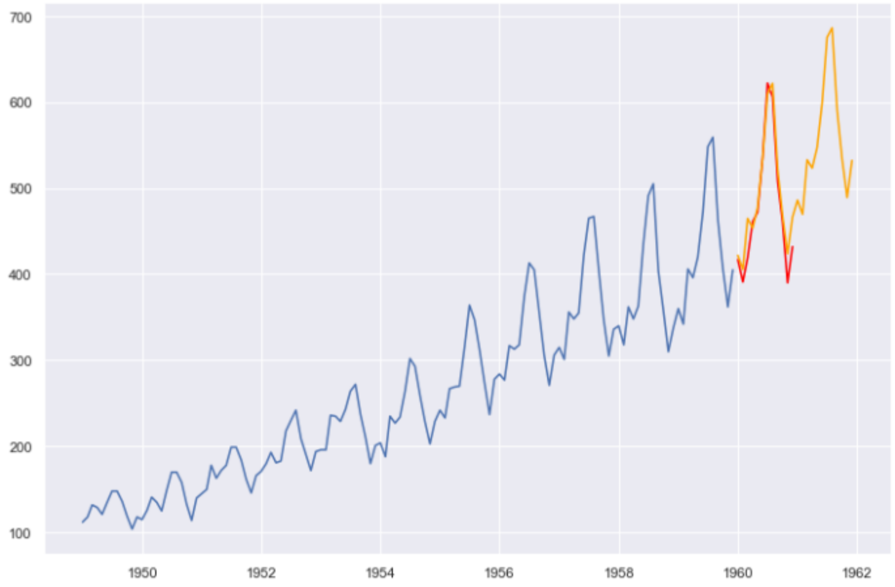

Figure 10. SARIMA Forecast

8. Conclusion

In this assignment, I tried to forecast US flights in 1961. I used monthly data from BTS to model time series models. This was relatively hard because our model has some inherited problems. Its residuals are not in iid and it was non-stationary. Among the 3 models I have tried (AR, ARIMA, SARIMA), SARIMA comes out to be the best according to RMSE. Then, forecasts are done via SARIMA. It suggests that our trend and seasonal affects will remain.