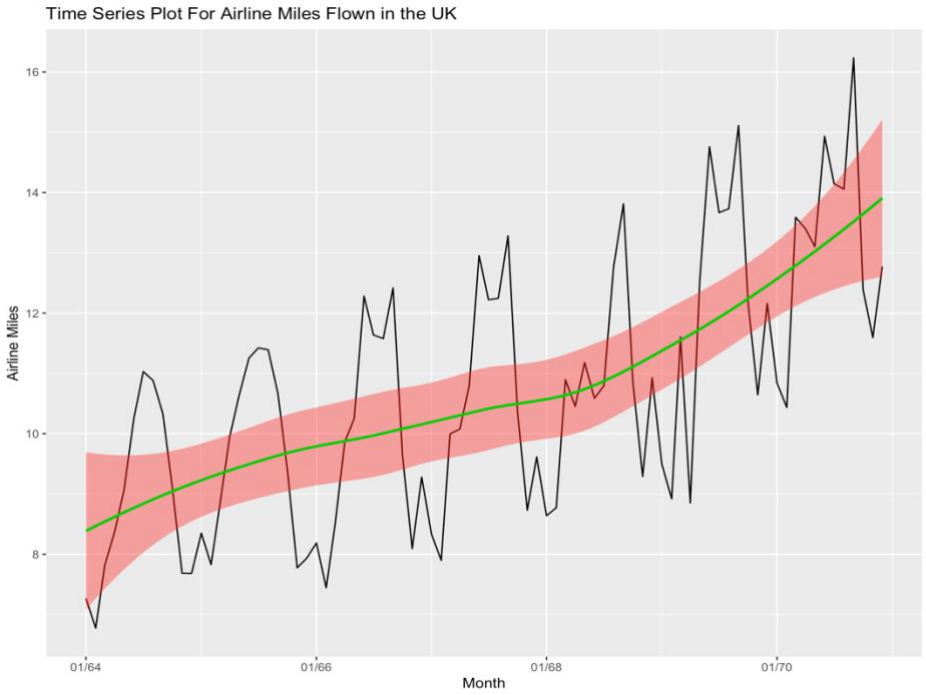
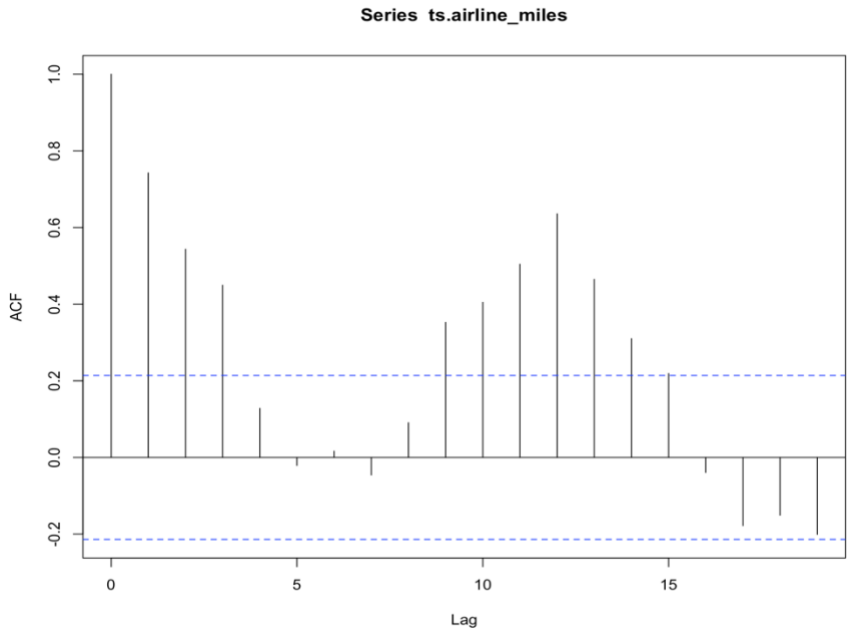
DSC 275/475: Time Series Analysis and Forecasting (Fall 2019)
Project-1 Chunlei Zhou

1. **Create a time series of the plot of the data provided.**



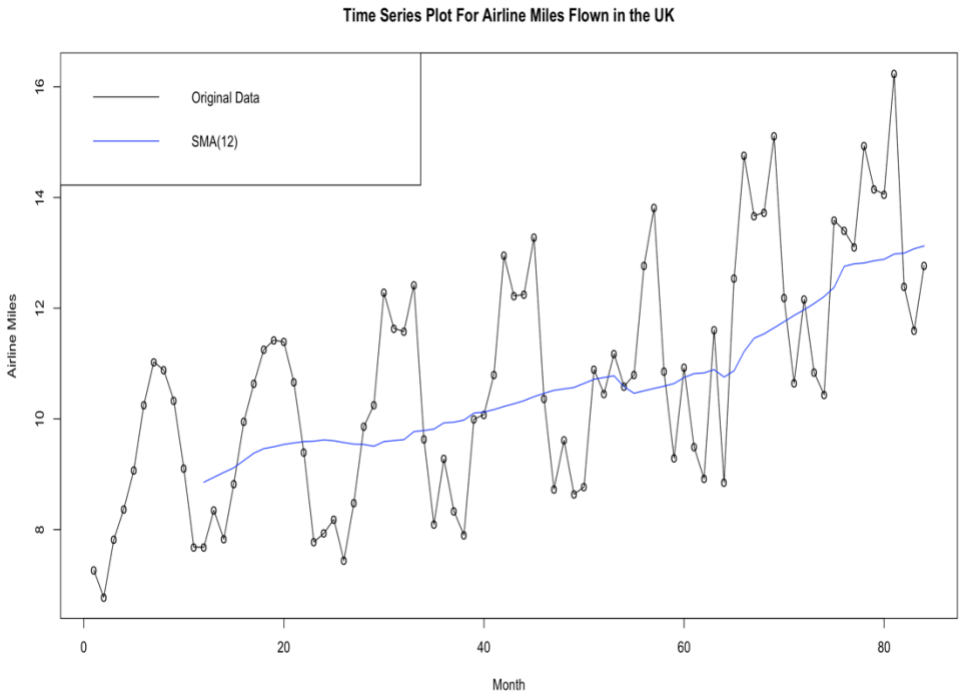
2. **Compute the autocorrelation function and display it in a plot. What is the seasonal period?**



The seasonal period is 12, because the lag between two peaks is 12.

3. **Compute a moving average for the data and overlay on the original time-series plot. What is a suitable choice for the moving average window length? Why?**

   The suitable moving average window length is 12, because the seasonal period is 12. Choosing 12 as the moving average window length will remove the effect of seasonality and retain the effect of trend.



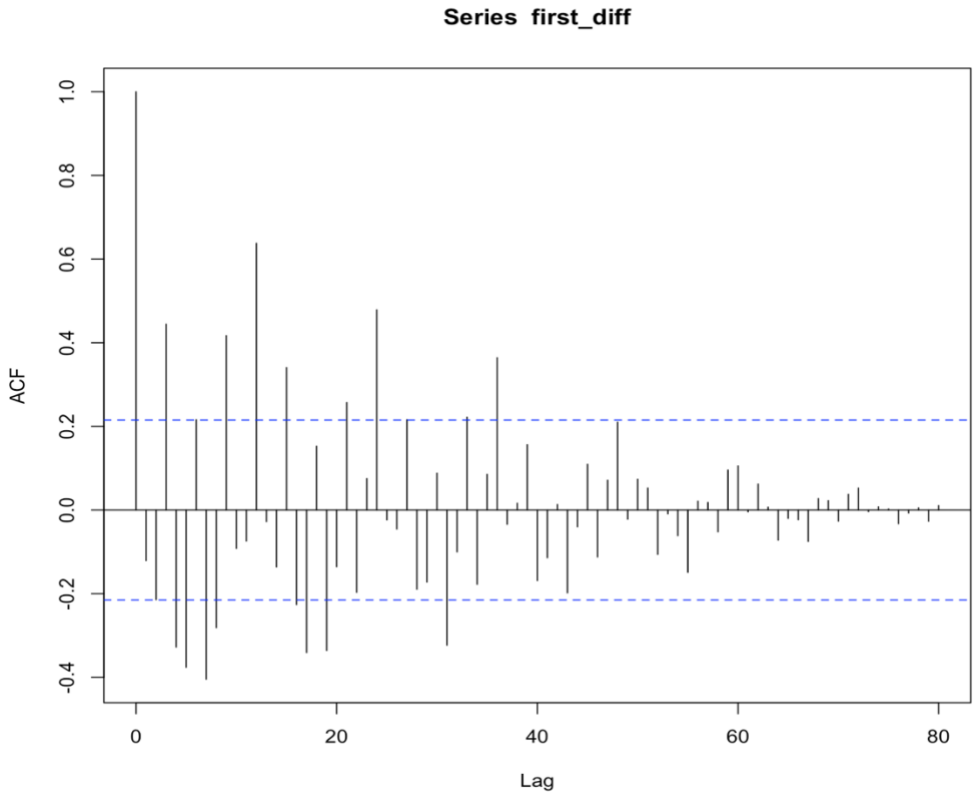
4. **Based on Q3, describe the trend line. What does it indicate, i.e. is it increasing / decreasing?**

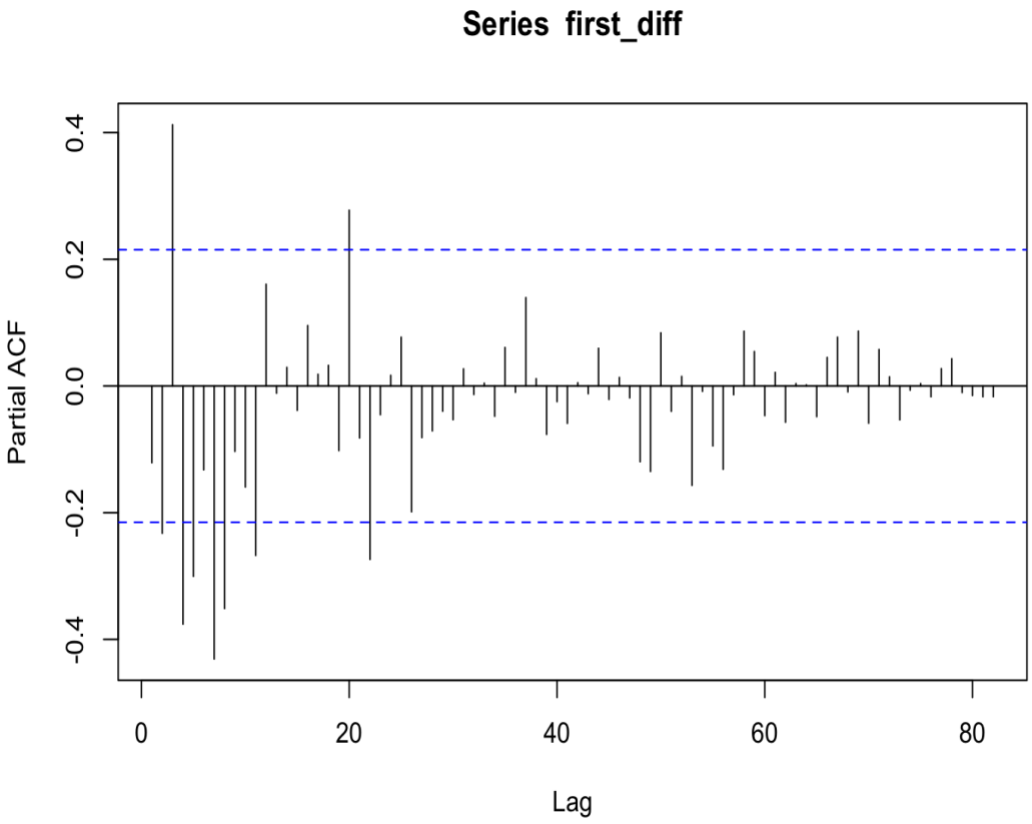
   There is an increasing trend.

5. **Compute the first difference of the data and plot the ACF and PACF for the differenced data. What are the significant lags based on the ACF and PACF?**

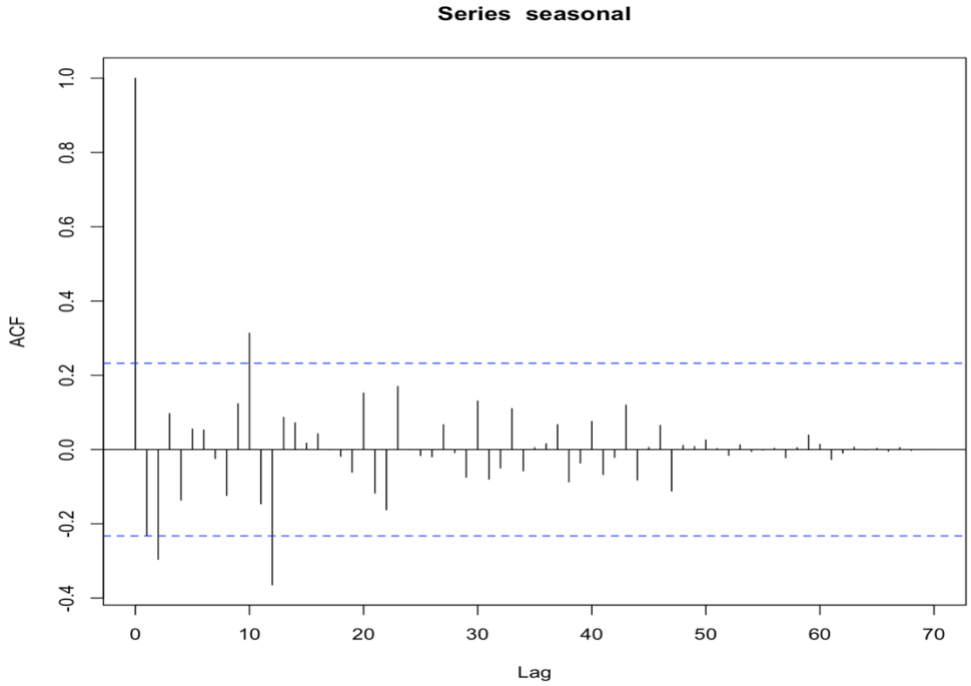
   Based on the ACF plot, significant lags including lag = 12, 9, 7, 3, and etc., among which lag = 12 is the most significant one because we have seasonality equals to 12. Also, lag = 24 and lag = 36 are also ones of most significant lags.

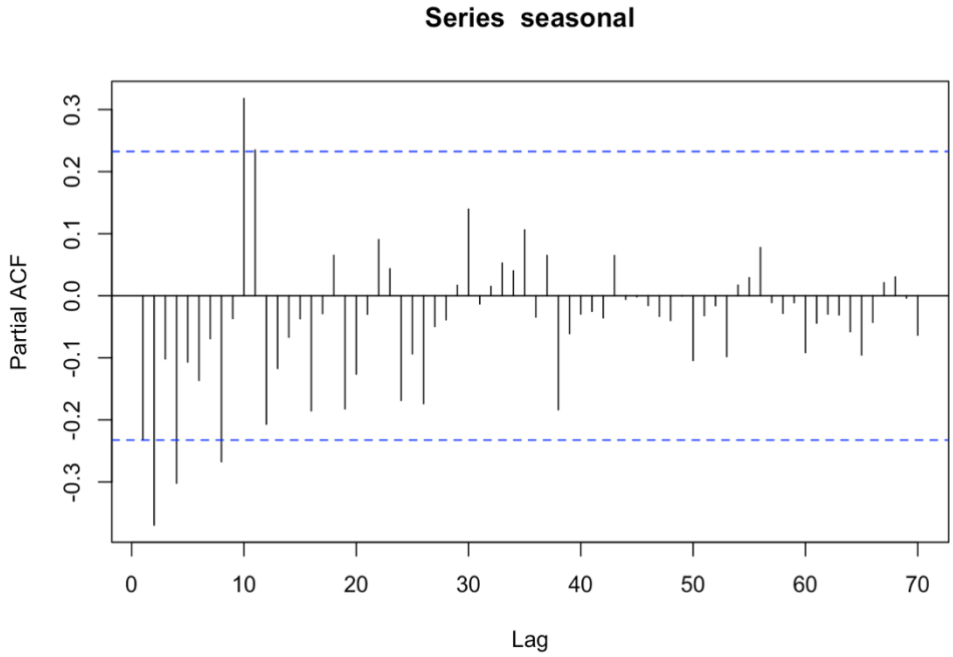
   Based on the PACF plot, there are also a lot of significant lags including lag = 3, 4, 7, 20, and etc.


Series first_diff
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
20
40
60
80
Lag


Series first_diff
Partial ACF
0.4
0.2
0.0
-0.2
-0.4
0
20
40
60
80
Lag

6. **Using the output from Q5 above, perform a first seasonal difference with the seasonal period you identified in Q2, and plot the ACF and PACF again. What are the significant lags based on the ACF and PACF?**





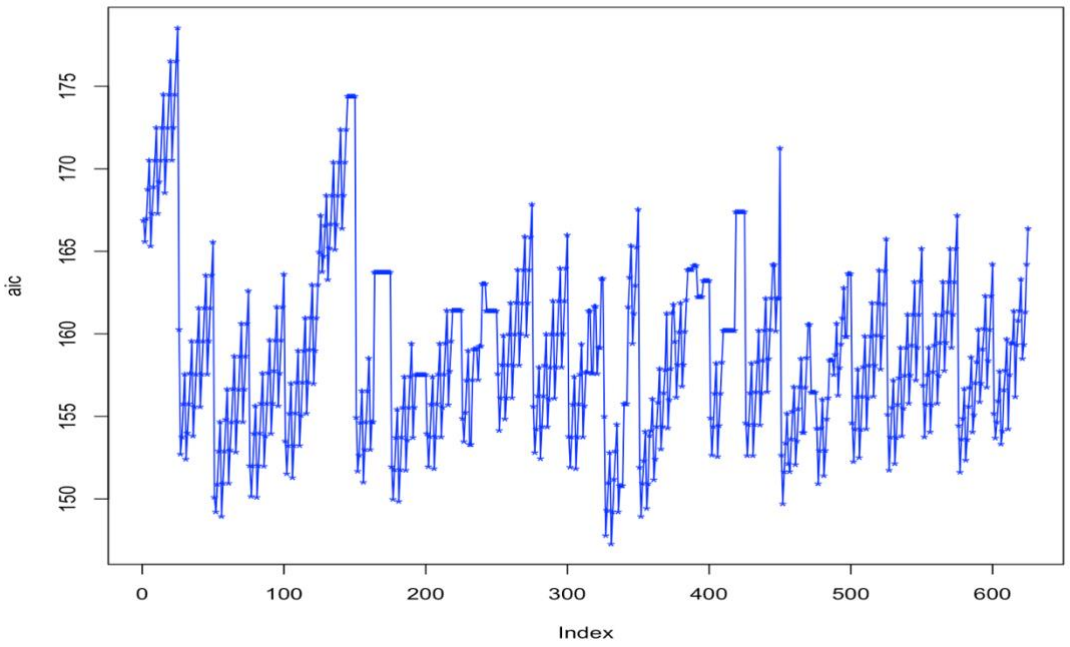
Based on the ACF plot, the significant lags are lag = 12, 10, and 2, among which the most significant one is lag = 12.

Based on the PACF plot, the significant lags are lag = 10, 8, 4, 2.

7. **Develop a suitable SARIMA model that can be applied on the time series. Use the first 6 years of data only to develop the model.**

    1. To develop the model, vary the model parameters for the non-seasonal *(p,d,q)* and seasonal components *(P,D,Q)* and calculate the output for each combination of parameters.

       The output of each model is the corresponding AIC. The AIC plot is as follow:



    2. Use an evaluation criteria such as AIC or sum squared error or mean squared error to determine the ***best choice of parameters*** *(p,d,q,P,D,Q)*. *Note:* AIC is a metric that is readily output by the ARIMA model. When comparing two models, the model with the lower AIC value is better.

       The best choice of parameters is as follow:
       p = 2, d = 1, q = 3, P = 1, D = 1, Q = 0.
       This model has the smallest AIC among all models.

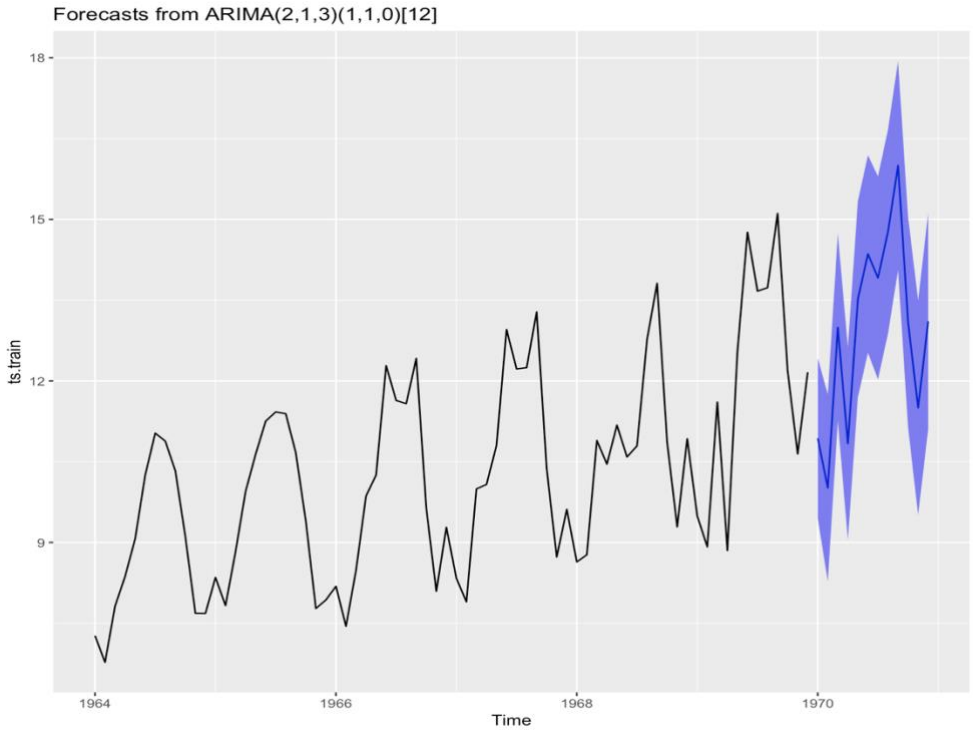
```
> best_model
Series: ts.train
ARIMA(2,1,3)(1,1,0)[12]

Coefficients:
          ar1      ar2     ma1      ma2      ma3     sar1
      -1.4507  -0.5748  1.0774  -0.5941  -0.8339  -0.4318
s.e.   0.1288   0.1342  0.1255   0.2108   0.1220   0.1561

sigma^2 estimated as 0.5626:  log likelihood=-66.64
AIC=147.28   AICc=149.48   BIC=161.83
```

8. Use the model parameters determined in Q7 above to forecast for the 7th year. Compare the forecast with actual values. Provide your own insight on how you believe the forecast can be improved?

The forecasted data for 7th year is listed below:




```
> forecast_1970
         Point Forecast     Lo 95    Hi 95
Jan 1970       10.93529  9.446556 12.42403
Feb 1970       10.01423  8.276107 11.75235
Mar 1970       12.99511 11.257048 14.73318
Apr 1970       10.83643  9.046620 12.62625
May 1970       13.51972 11.699746 15.33968
Jun 1970       14.35735 12.522887 16.19180
Jul 1970       13.91275 12.028325 15.79717
Aug 1970       14.77387 12.879174 16.66857
Sep 1970       15.99880 14.057874 17.93972
Oct 1970       13.08965 11.135069 15.04423
Nov 1970       11.50112  9.510293 13.49194
Dec 1970       13.10544 11.095900 15.11498
```

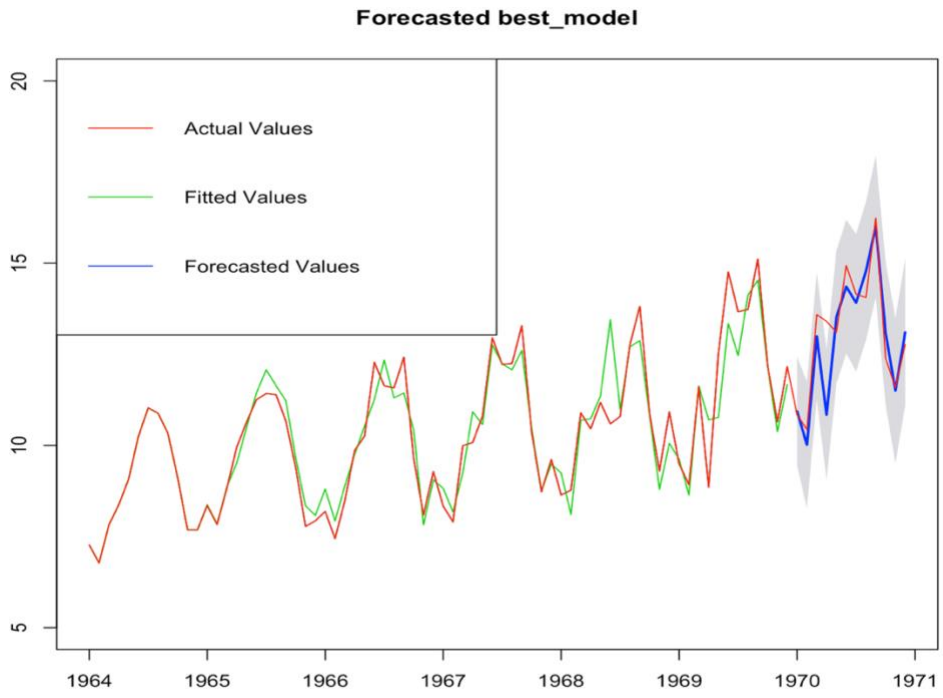
The actual values are:

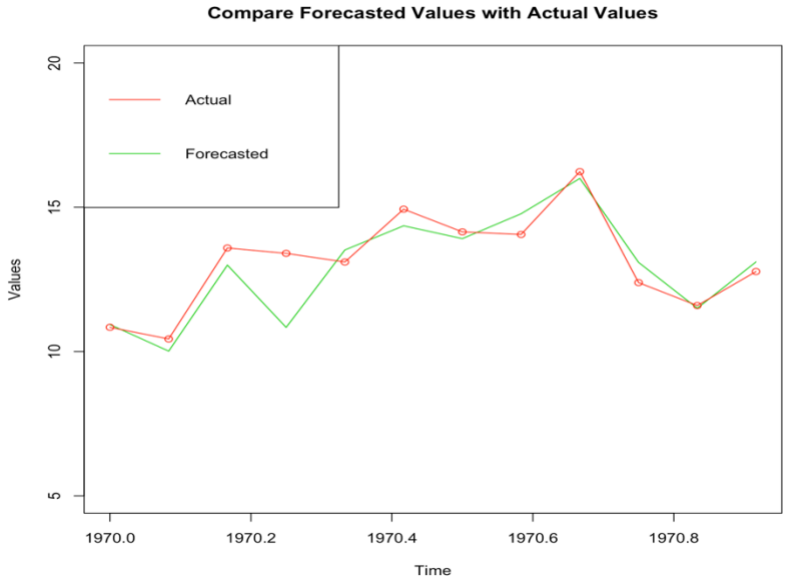
```
ts.test
[1] 10.840 10.436 13.589 13.402 13.103 14.933 14.147 14.057
[9] 16.234 12.389 11.594 12.772
```

The accuracy of the forecast is:

```
> accuracy(forecast_1970,x=ts.test)
                     ME      RMSE       MAE        MPE
Training set 0.009917203 0.6435498 0.4054589 -0.2880924
Test set     0.204687120 0.8593434 0.5818486  1.4908710
                  MAPE      MASE        ACF1
Training set 3.821305 0.3622356 0.004067347
Test set     4.395081 0.5198214          NA
```

Based on the plot, we can conclude that the forecasted values are very close to the actual values with slight difference.





To find the best model and to improve the accuracy of forecasting, I should choose the model by considering more evaluation criteria. For example, I should consider AIC, BIC and MSE together. Listed the five models that gives lowest value for each one of evaluation criteria listed above and choose the model that appears in all three lists. If there are multiple models that appear in all three lists at the same time, then choose the one with lower value for all three criteria. By doing so, the model I choose may perform better.

To improve the accuracy of the forecasting, we can apply several methods on the same dataset. Methods such as ARIMA, ETS, STL-ETS, NNAR, and TBATS may perform differently in different dataset. Some methods may be more accurate than others in dataset A but less accurate in dataset B. One way to improve the accuracy of forecasting is to perform all these methods and choose the one that is most accurate to use in forecasting. Another way is to combine all methods by assigning weight to each method and use the combined method to perform forecasting activity.

As for this project, I think we can combine different ARIMA models. Some model may predict the value of time A more accurate than others, but less accurate in predicting the value of time B. If we choose several models and assign different weights to different models using deep learning algorithms, we may get a perfect model that can predict every value accurately.