





KOLEJ UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

Assignment

BMMS2074 Statistics for Data Science

May/June 2020 Semester

Student Name	Student Signature	Contribution Mark	Total Mark
1) Tan Yi Hong		25	
2) Tan Teoh Xin Ee		25	
3) Tan Wei Siong		25	
4) Nigel Lee Jian Hsee		25	
Programme :	RDS	Tutorial Group:	Group 2
Lecture and Tutor :	Dr. Chin Wan Yoke		
Comments :			

(b) Introduction

Nowadays, the population in the world is becoming bigger and bigger. For a small country like Singapore which has 5703600 residents. During the last 10 years, the number of Air Passengers around worldwide has increased rapidly and Singapore is not exceptional. Singapore is one of the world leading countries and one of the famous tourism spots around the world that attracted a lot of tourists. Tourism in Singapore is one of the main industries and contributors to the Singaporean economy. According to Wikipedia, Singapore attracted approximately 17.4 million international tourists in 2017, which were more than 3 times of Singapore's actual population.

On top of that, the government is always promoting and encouraging more tourists from other countries to come over Singapore for their vacation during the holidays. Therefore, during the end of the year or some special festival holidays, there will be a lot of people in the theme park such as Sentosa, Universe Studio and so on. So that it caused the number of tourists entered to Singapore is large and sales of the theme park is good.

For this assignment given by Dr Chin, we are required to do time series analysis on a dataset we found from the internet. The dataset we found is regarding population on air passengers which means the number of passengers in the Singapore Changi airport daily. We were given 5 weeks time to finish this assignment with a good and complete report.

Currently in the market, many mathematical tools and formulas were used for prediction in the stock market and forecasting in airlines is also not exceptional. For airline companies, accurate forecasts play a significant role in the revenue management of the company. Accurate forecasts help to reduce the loss and increase the company incomes by objectively evaluating the demand of the air transportation business in each period such as month, season or year. Which allows airline companies to plan and utilize their flights schedule at different periods.

(c) Objective

Time series analysis is the collection of data at the specific intervals in a period of time. The main purpose of this analysis is to identify the trends, cycles, and seasonal variances to aid in the forecasting of the future event. The observed outcome is the data that is measurable. In time series analysis, the data is to be measured over time at consistent intervals to identify patterns from its trends, cycles, and seasonal variances. ([Time Series Analysis & Its Applications, 2019](#)). Time series analysis helps us to understand the essential forces that lead to a particular trend in the time series data points. This analysis is also able to forecast and monitor the results by fitting appropriate models to it.

The main objective for this study is to discover the most suitable approach for conducting short-term forecasting for the air passengers of the Singapore Changi Airport. There are various approaches that can be used for forecasting. A linear method in the form of an autoregressive integrated moving average (ARIMA) model is a common forecasting method. It has been used to predict the airport passenger traffic in Hong Kong([Z.W., 2018](#)). To have a comprehensive view for the approaches, this study will invoke other mathematics knowledge to support the analysis.

So, based on the dataset of the air passengers of the Singapore Changi Airport, we will carry out the time series analysis to study the trends that are consecutive increases or decreases in a measurement over time, and the seasonal variances that are measured over 10 years that associated with a specific time of the year and interpret about the seasonal growth. Also, the analysis will be forecasted by using appropriate models or approaches to predict the future values of the time series variable.

(d) Methodology

Data

The total number of air passengers in Changi Airport per month was used for doing this time series analysis. Total number of air passengers used was the total summation of air passengers that include arrival, departure, and also transit. The data range from January 2009 to October 2019 and is observed by time plot. The time series analysis for this dataset is conducted and followed by forecasting. The main tool of this analysis is using R programming language.

Selecting Time Series Analysis Techniques

Time series analysis is useful to extract information from a time series and to predict the changes of a time series. It can obtain descriptive or statistical measures of a time series like trends or seasonal cycle. We are able to observe the variation of a time series to explain the variation of it to help to understand what is happening to the time series. Besides, one of the most important steps of the time series analysis is the prediction or forecasting for the future value. From the forecasting of the time series, we are able to observe the values and predict the future variation of a time series by using different techniques. There are several techniques for time series analysis for example, Holt-Winters and Box-Jenkins Methods.

For Holt-Winters model, smoothing methods are used to reduce the effect of random oscillations in time series. They provide the chance to receive pure values that consist only of deterministic components. While the Box-Jenkins method which is also called ARIMA model, it stands for Auto-Regressive Integrated Moving Average. This technique required the series to be stationary to observe the “autoregressive” and “moving average” terms. Between these two techniques, we decide to choose the Box-Jenkins methods to conduct our time series analysis since it provides a better forecasting result and the ease of use.

Box–Jenkins methodology aka ARIMA (autoregressive integrated moving average)

Box - Jenkins is a type of mathematical tools that were named after statisticians George Box and Gwilym Jenkins. These mathematical tools were designed to forecast future value based on an input from specific time series dataset. Box - Jenkins methodology consists of a few step process of identifying, selecting, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models. Below shows the major few step of Box - Jenkins methodology:

1. The first step is to establish the stationarity of your time series. We need to convert our data into stationary if the data series we obtain is not stationary by successively difference the time series until it attains stationarity. Stationarity of a time series can be detected from an autocorrelation plot. We can conclude that the time series is non stationary if the autocorrelation plot has very slow decay and does not decrease to zero. Iterate the steps until the ACF looks like a stationary series.
2. The second step is parameter estimation and selection. After step 1, we specify our ARIMA model parameter based on the PACF and ACF graph after normal differencing and seasonal differencing. We specify the parameter based on the pattern of the graph. ARIMA models can be viewed as consisting of AR (Autoregressive) + I (Integrated) + MA (moving average).

Below is the equation of the ARIMA model:

$$Y_t = c + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \dots + \theta_q e_{t-q} + e_t$$

3. We will test multiple ARIMA models with different parameters based on the ACF and PACF graph that was plotted and find out which parameter has the lowest AIC score. The lower the AIC score the better. Also, we will compare a few models with low AIC score by its significance for the P and Q value.
4. After choosing the parameter with the low AIC score and checking for its significance, a best model is obtained and we can use the model to perform forecasting over a future time horizon and variation.

(e) Data Sources

As a brief description, airlines is one of the world's largest industries in transportation and most crucial transport in the global transportation system. While the airline industry provides long distance travel over the world in a short period of time. Other than that, the airline industry also plays an important role in economic development which improves performance of business operations by providing quick access for input supplies and stimulates innovative activities by facilitating face-to-face meetings. Hence, the demand for air travel has increased day by day.

On the other hand, the dataset we used for this assignment is able to be found from kaggle website. This dataset consists of passengers passing through Singapore Changi Airport started from year 2009 to year 2019. Almost more than 100 airlines fly to 380 cities and 100 countries worldwide served by SINGAPORE Changi Airport. Within a week, around 7,400 flights land or depart from Changi. Counting day by day, more than 65.6 million passengers pass through the airport annually. By analyzing the passengers passing through the airport, it would be a good tutorial for time series analysis which is suitable for our assignment.

(f) Data Analysis

To perform data analysis, the first thing that is required to do is read the dataset file into the R programming. Then use the plot function to show the graph based on the dataset.

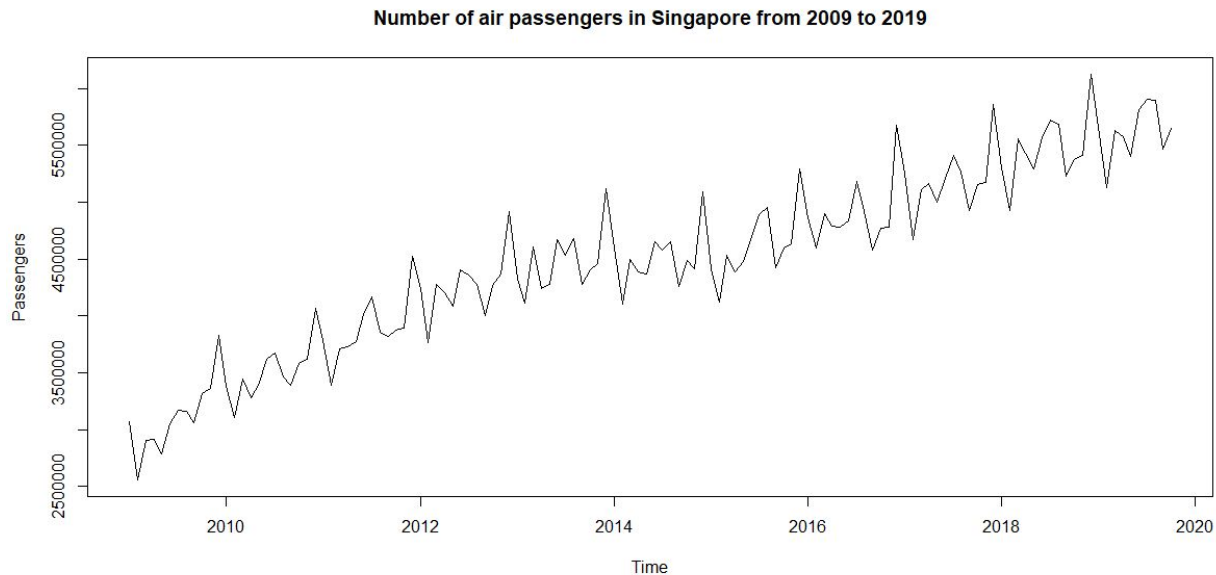


Image 1

The original dataset has been defined as a time series model and a time plot is being plotted. The image 1 above indicates the time series graph based on the number of air passengers in Singapore from 2009 to 2019. The graph has shown a trend variation in the number of passengers per year. We can see that the number of the passengers is around 3,000,000 in 2009 and it was gradually increased with same size peaks to around 5,500,000 in 2019. Therefore, we also can know that it is an additive time series model because the variance of the time series has no big difference and remains the same throughout the whole time plot.

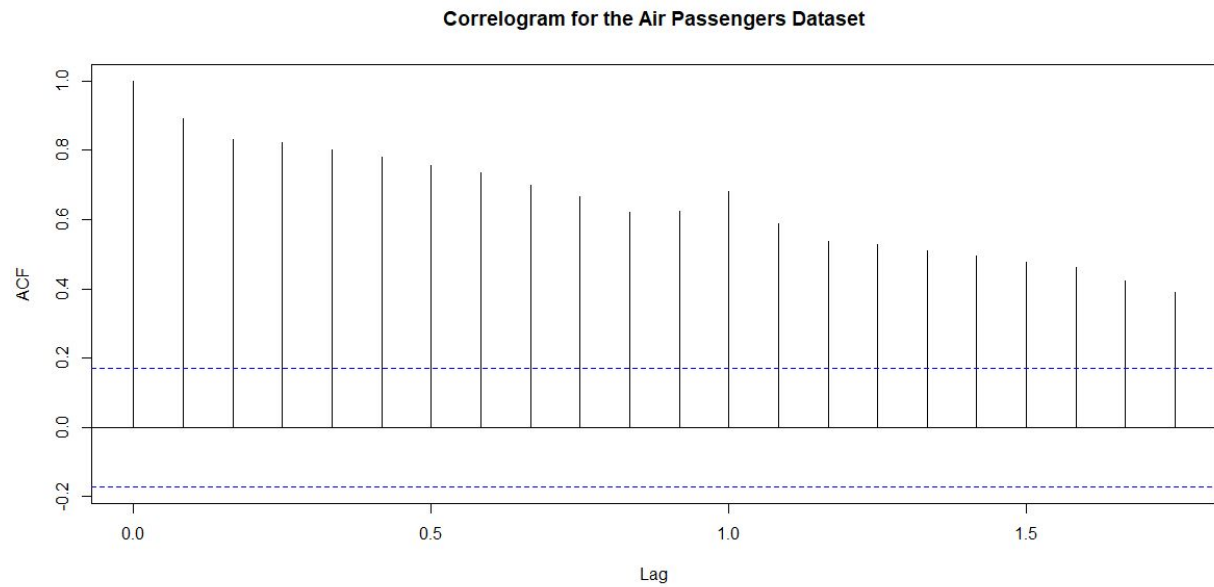


Image 2

The image 2 shows the correlogram for the air passengers dataset. We can know that it is a non-stationary time series from the observation of this autocorrelation plot that is having a very slow decay and does not decrease to zero, so a differencing is needed for our next time analysis step to obtain a stationary time series model.

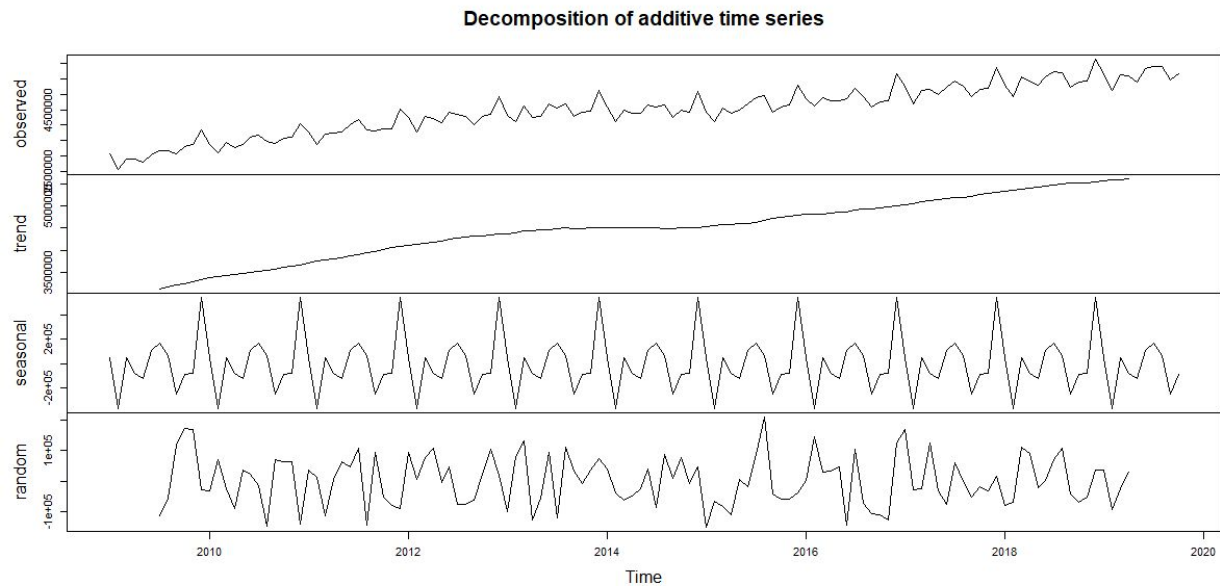


Image 3

Next, a decomposition is performed to the dataset to constituent components that are within the graph. Offently are trend component and random (Irregular) component, if there is seasonal, a seasonal component will also provided. Based on the image 3, we can see that there is a trend, seasonal and random in the additive time series. The observed is the original plot graph without decomposition. From the trend, we can see that there is only an upward trend in this dataset. While we can also see that there is a seasonal cycle for every one year from the seasonal component.

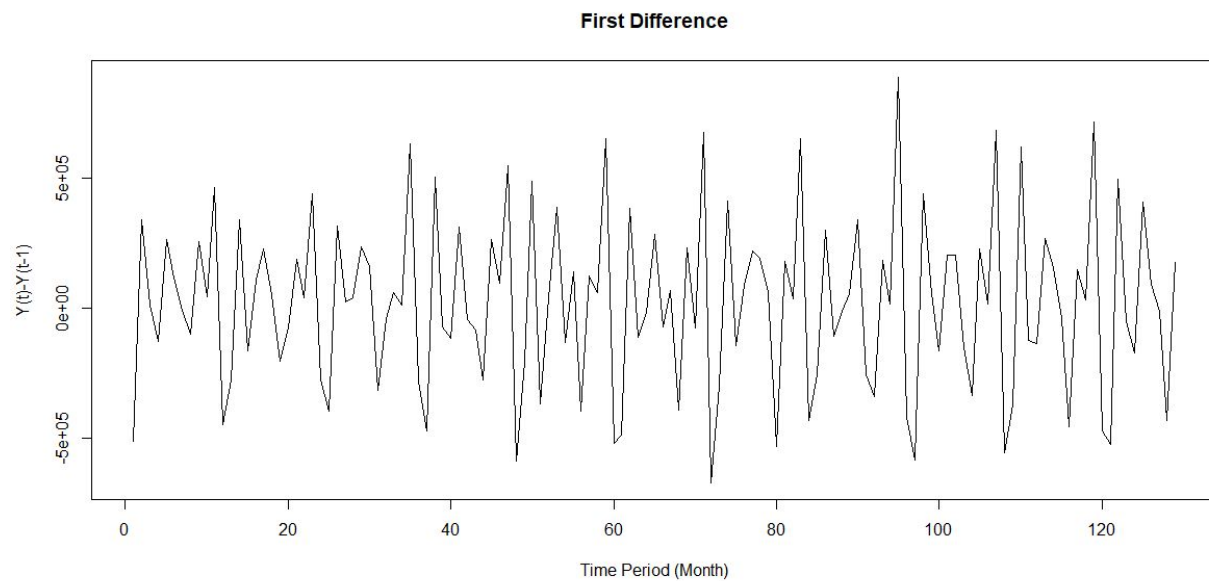


Image 4

Next, a differencing is performed for the time series data to change the model into a stationary series. Image 4 shows the time plot of the time series data after the first differencing. From the image above we can see that after first differencing has been performed, we obtained a stationary time plot from its original model. The variance and mean of the series in the graph remains constantly. After that, ACF and PACF are being plotted.

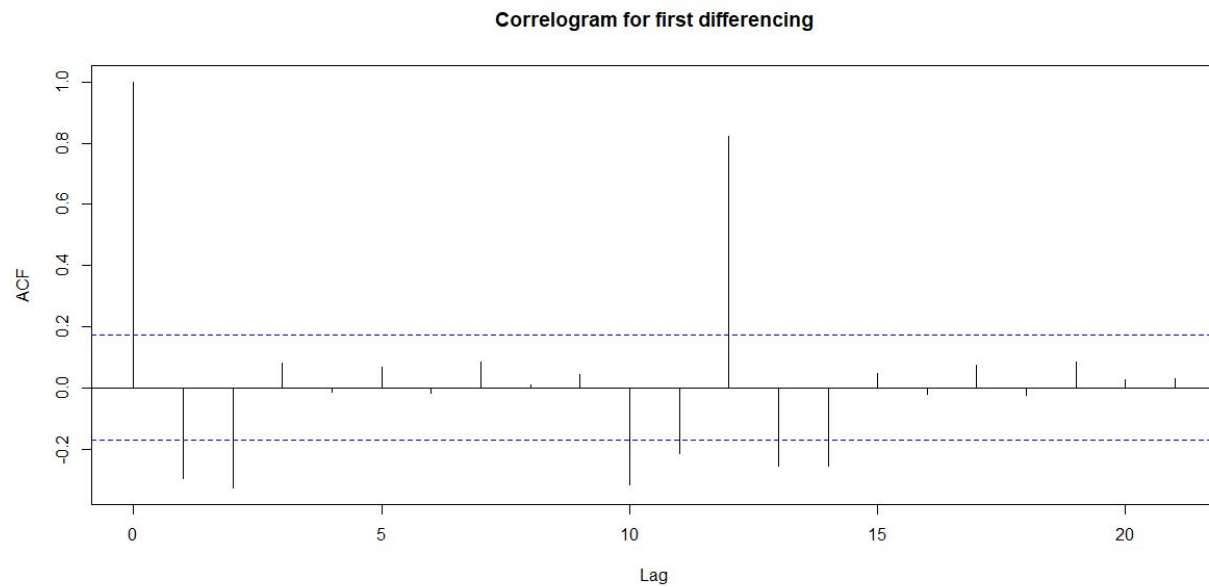


Image 5

Image 5 above shows the ACF plot of the time series after performing the first differencing. From the plot above, we can see that the autocorrelation exceeds the significance bounds at lag 1, lag 2, lag 10, lag 11, lag 12, lag 13, and lag 14 are negative. The autocorrelation after the lag 1 and lag 2 tail off to zero, this means that there is a possibility of a non-seasonal MA model for the time series after the first difference. Besides, the autocorrelation at lag 12 is relatively high, this means that there is a seasonal effect for this time series and a seasonal differencing has to be carried out. The autocorrelation at lag 10 until lag 14 has been affected by the seasonal effect so that they exceed the significance bounds. From this ACF plot, we guess that it is a non-seasonal MA(2) model.

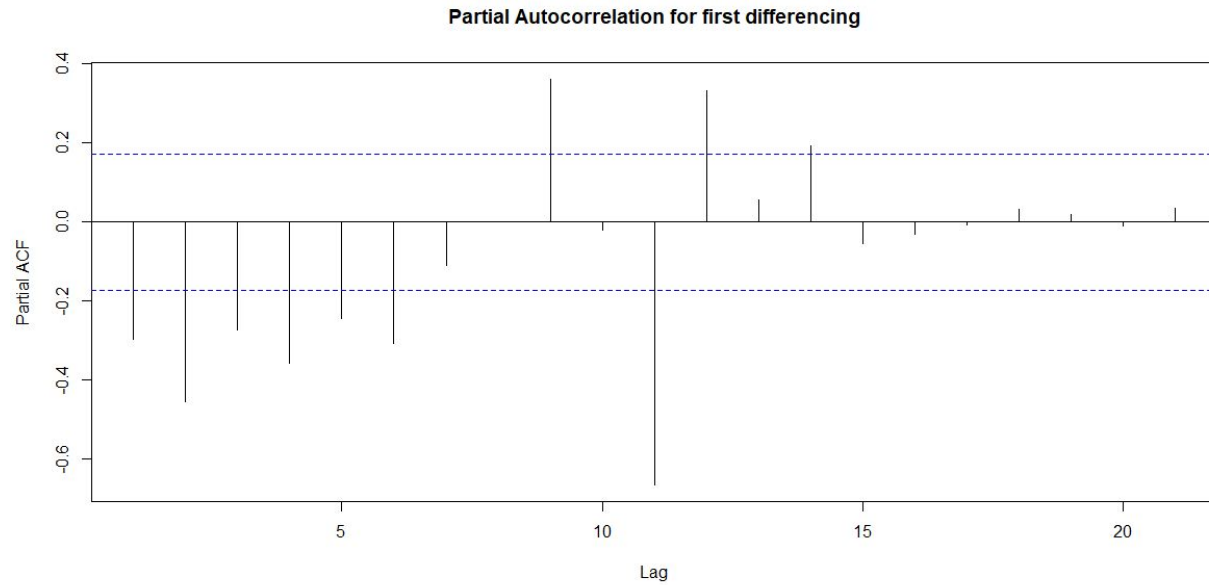


Image 6

Image 6 shows the PACF plot of the time series after the first differencing. The partial autocorrelations at lag 1 until lag 6, lag 9, lag 11, lag 12, and lag 14 exceed the significance bounds. The partial autocorrelation only tailed off to zero after lag 7. This could mean that there may or may not be a non-seasonal AR model for the time series after the first differencing. This is to be determined by comparing different ARMA models with different order of the P and Q. On top of that, the partial autocorrelations that exceed the significance bounds at lag 11 and lag 12 also means that there is a seasonal effect and they are affected by it. From here we still take a guess that it is a non-seasonal AR(2) model.

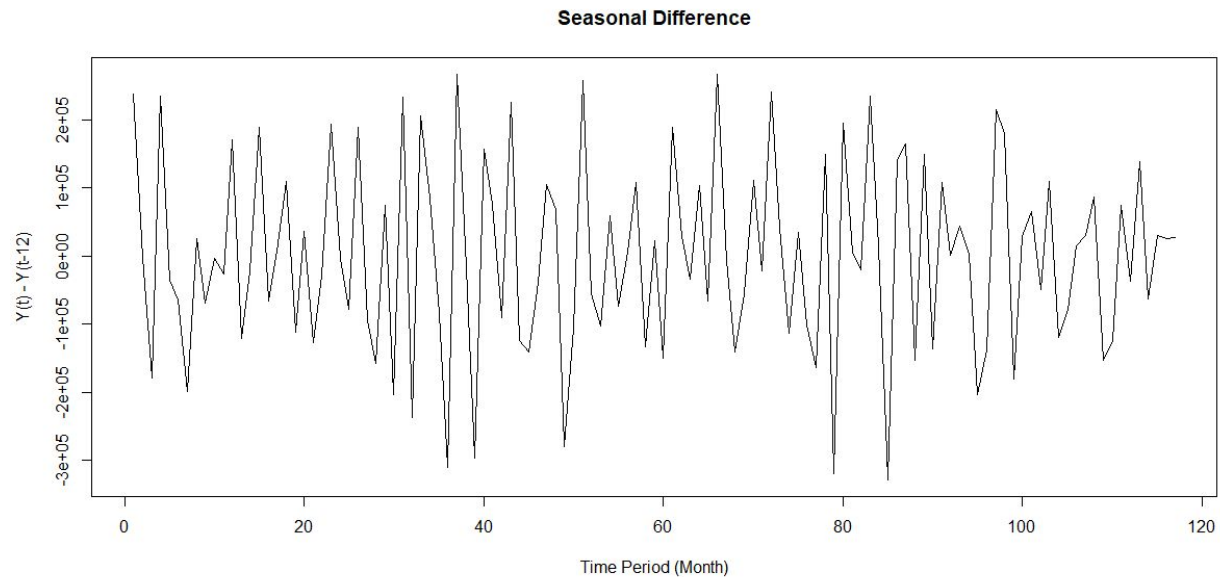


Image 7 : Seasonal Difference time plot

Image 7 above shows the time plot of the time series after the first seasonal differencing. Seasonal difference is the difference between the observation and the previous observation from the same season. After we performed the first order differencing, seasonal differencing is required to be performed for the time series in order to change the model into a stationary series. Previously from the decomposition graph, we observed that there is a seasonal effect and the ACF and PACF graphs show that it is non-stationary. After we performed the first seasonal differencing, we can observe from the image above that we obtained a stationary time plot from its original time plot. We can also observe that we obtain a constant variance and mean time series from the time plot above.

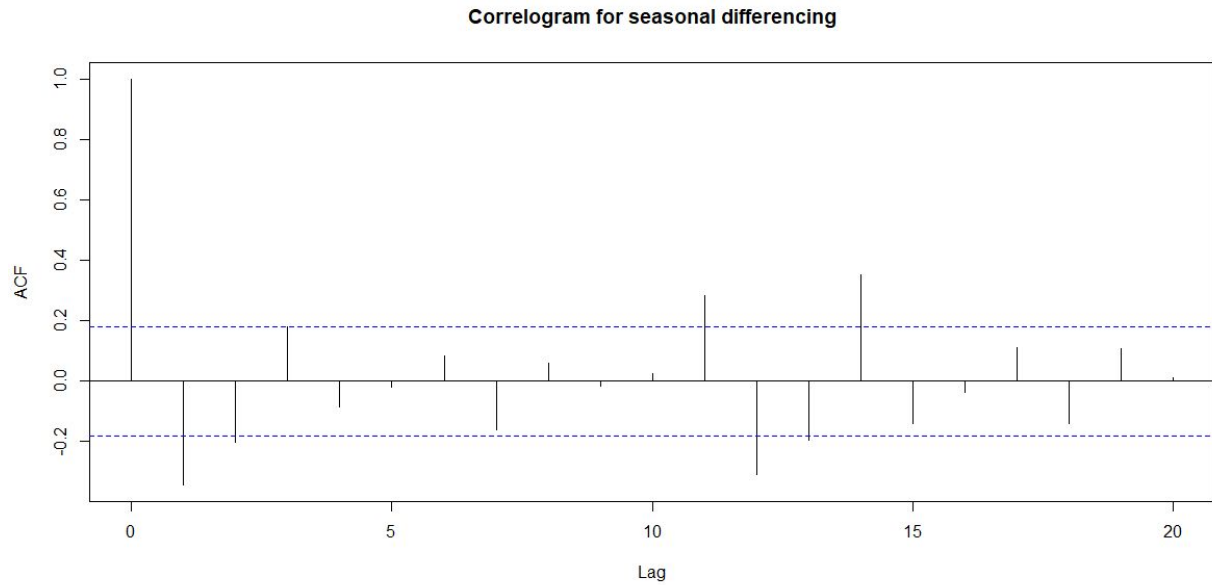


Image 8 : ACF plot after the first seasonal differencing

Image 8 above shows the ACF plot of the time series after performing the first seasonal differencing. From the time plot above, we can observe that the autocorrelation exceeds the significance bounds at lag 1, lag 2, lag 11, lag 12, lag 13 and lag 14. The ACF plot above shows that there is a spike in lag 12. Hence we can conclude that there is a seasonal effect from the time series. The autocorrelation at lag 13 and lag 14 were affected by the seasonal effect in lag 12. Hence, this causes lag 13 and lag 14 exceed the significance bounds. The autocorrelation after the lag 1 and lag 2 tail off to zero, hence we conclude that there is a possibility of a seasonal MA model after the first seasonal difference. Hence, we assume that the parameter for seasonal MA is 2.

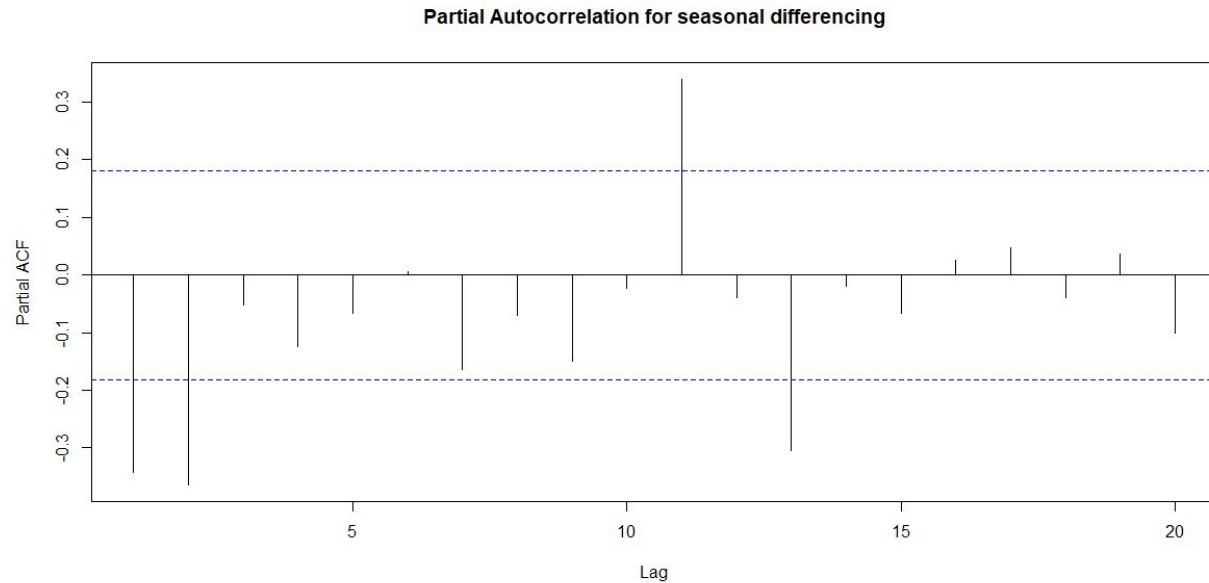


Image 9 : PACF plot after the first seasonal differencing

Image 9 shows the PACF plot of the time series after the first seasonal differencing. The partial autocorrelations at lag 1, lag 2, lag 11 and lag 13 exceed the significance bounds. The partial autocorrelations tailed off inside the significance bounds after lag 2. This could mean that there is a seasonal AR model for the time series after the first seasonal differencing. Hence, we assume that the AR for seasonal differencing parameters is 2 which is the P. Beside that we also observe that there is a spike on lag 11 and lag 13, this might be due to the seasonal effects of the time series.

As a conclusion, throughout the observation of the ACF and PACF plot of the first difference and seasonal difference, we can guess that the most likely ARIMA model for this series is $ARIMA(2,1,2)(2,1,2)$ [12]. But it may not be the best ARIMA model. Hence, we will compare this ARIMA model with some other nearest order of the P and Q to find out which ARIMA model is the best by looking at their AIC score. So, several models of ARIMA have been tested and their AIC scores are being tabled.

Arima Model	AIC
ARIMA(0,1,1)(2,1,1)[12]	3061.99
ARIMA(0,1,2)(2,1,1)[12]	3063.9
ARIMA(0,1,1)(1,1,1)[12]	3054.91
ARIMA(0,1,2)(1,1,1)[12]	3056.08
ARIMA(0,1,1)(1,1,2)[12]	3055.69
ARIMA(0,1,2)(1,1,2)[12]	3057.29
ARIMA(0,1,1)(0,1,1)[12]	3054.95
ARIMA(0,1,2)(0,1,1)[12]	3056.85
ARIMA(0,1,1)(0,1,2)[12]	3054.84
ARIMA(1,1,2)(2,1,1)[12]	3065.63
ARIMA(1,1,2)(1,1,1)[12]	3056.95
ARIMA(1,1,2)(0,1,2)[12]	3056.97
ARIMA(1,1,1)(1,1,1)[12]	3056.32
ARIMA(1,1,1)(1,1,2)[12]	3057.37
ARIMA(2,1,2)(0,1,1)[12]	3059.62
ARIMA(2,1,2)(0,1,2)[12]	3058.52
ARIMA(2,1,2)(1,1,2)[12]	3060.48
ARIMA(2,1,2)(2,1,2)[12]	3061.63
ARIMA(2,1,2)(2,1,1)[12]	3060.43

From the table above, we can observe that different orders of p, d and q will generate different AIC which brings score meaning for the arima models. The lower the AIC score, the more suitable arima model to be used. From the table above, we can see that our guessing model which is the ARIMA(2,1,2)(2,1,2)[12] model obtained the AIC score of 3061.63. But this is not the lowest AIC score and we can find that there are many other ARIMA models that can get the

AIC score lower than the score of the guessing model. Therefore by comparing the several different models, we are able to conclude that ARIMA(0,1,1)(1,1,1)[12], ARIMA(0,1,1)(0,1,1)[12] and ARIMA(0,1,1)(0,1,2)[12] had a respectively low AIC score, which is around 3054. Hence, we decided to compare these models by doing coefficient tests to check whether it is significant of the order of P and Q in the ARIMA model.

```
z test of coefficients:

      Estimate Std. Error z value Pr(>|z|)
ma1  -0.610491   0.078336 -7.7933 6.530e-15 ***
sma1  -0.541186   0.104924 -5.1579 2.498e-07 ***
sma2  -0.150348   0.108582 -1.3847  0.1662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Image 10 : Z-test of coefficients for ARIMA(0,1,1)(0,1,2)[12] model

The image above shows the z test of coefficients for ARIMA(0,1,1)(0,1,2)[12] model. There is 1 MA value in the first differencing and 2 MA values in the seasonal differencing of time series dataset. Therefore, there will be a 3 z test of coefficients in this model. We can see that the MA1 and SMA1 of this model are significant and labeled with 3 stars. The P-value for MA1 and SMA1 are lower than 0.001. However, the result in SMA2 is not significant. This is because the P-value for SMA2 is more than 0.05. We conclude that this arima model order is not significant and not recommended to be used.

```
z test of coefficients:

      Estimate Std. Error z value Pr(>|z|)
ma1  -0.602563   0.077101 -7.8152 5.488e-15 ***
sma1  -0.604810   0.101680 -5.9482 2.712e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Image 11 : Z-test of coefficients for ARIMA(0,1,1)(0,1,1)[12] model

The image above shows the z test of coefficients for ARIMA(0,1,1)(0,1,1)[12] model. There is 1 MA value in the first differencing and 1 MA values in the seasonal differencing of time series

dataset. Therefore, there will be a 2 z test of coefficients in this model. We can see that the MA1 and SMA1 of this model are significant and labeled with 3 stars. The P-value for MA1 and SMA1 are lower than 0.001. Therefore, we conclude that this arima model order is significant and recommended to be used.

```
z test of coefficients:

      Estimate Std. Error z value Pr(>|z|)
ma1  -0.610643   0.078976 -7.7320 1.058e-14 ***
sar1   0.227768   0.161038  1.4144  0.1573
sma1  -0.777013   0.148468 -5.2335 1.663e-07 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Image 12 : Z-test of coefficients for ARIMA(0,1,1)(1,1,1)[12] model

The image above shows the z test of coefficients for ARIMA(0,1,1)(1,1,1)[12] model. There is 1 MA value in the first differencing and 1 MA values in the seasonal differencing of time series dataset. Therefore, there will be a 3 z test of coefficients in this model. We can see that the MA1 and SMA1 of this model are significant. The P-value for MA1 and SMA1 are lower than 0.001. However, the result in SAR1 is not significant. This is because the P-value for SAR1 is greater than 0.05. Therefore, we conclude that this arima model order is not significant and not recommended to be used.

From the z test of coefficients, we are able to conclude that the best model tested is the ARIMA(0,1,1)(0,1,1)[12] model. Although the AIC value for this model is slightly higher than other 2 models, the other models have extra 1 parameters which is not significant. Extra parameters will decrease the AIC value but it is not meaningful when the extra parameters are not significant.

Best Model : ARIMA(0,1,1)(0,1,1)[12]

```
Call:
arima(x = Y, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ma1          sma1
      -0.6026   -0.6048
s.e.    0.0771    0.1017

sigma^2 estimated as 1.156e+10:  log likelihood = -1524.47,  aic = 3054.95
```

Image 13 : Best model

Next, the ARIMA(0,1,1)(0,1,1)[12] are being fitted into a model as shown below :

$$(1 - B)(1 - B^{12})Y_t = (1 - 0.6026B)(1 - 0.6048B^{12})\varepsilon_t$$

$$(1 - B - B^{12} + B^{13})Y_t = (1 - 0.6026B - 0.6048B^{12} + 0.3645B^{13})\varepsilon_t$$

$$Y_t - BY_t - B^{12}Y_t + B^{13}Y_t = \varepsilon_t - 0.6026B\varepsilon_t - 0.6048B^{12}\varepsilon_t + 0.3645B^{13}\varepsilon_t$$

$$\therefore Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} + \varepsilon_t - 0.6026\varepsilon_{t-1} - 0.6048\varepsilon_{t-12} + 0.3645\varepsilon_{t-13}$$

(g) Results

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Nov 2019	5668795	5530997	5806593	5458051	5879539
Dec 2019	6387091	6238809	6535373	6160313	6613869
Jan 2020	5895421	5737349	6053493	5653670	6137171
Feb 2020	5444211	5276920	5611501	5188362	5700059
Mar 2020	5936218	5760191	6112244	5667009	6205427
Apr 2020	5865334	5680985	6049683	5583397	6147271
May 2020	5743511	5551200	5935823	5449396	6037626
Jun 2020	6040685	5840727	6240642	5734876	6346493
Jul 2020	6182442	5975120	6389764	5865371	6499513
Aug 2020	6128186	5913753	6342619	5800239	6456133
Sep 2020	5716947	5495631	5938263	5378474	6055421
Oct 2020	5897217	5669226	6125208	5548534	6245899
Nov 2020	5920181	5667378	6172984	5533552	6306810
Dec 2020	6638477	6374379	6902575	6234574	7042380
Jan 2021	6146807	5871878	6421736	5726339	6567275
Feb 2021	5695597	5410248	5980946	5259193	6132001
Mar 2021	6187604	5892202	6483005	5735826	6639382
Apr 2021	6116720	5811597	6421843	5650074	6583366
May 2021	5994897	5680352	6309442	5513843	6475952
Jun 2021	6292071	5968379	6615763	5797027	6787115
Jul 2021	6433828	6101240	6766416	5925179	6942477
Aug 2021	6379572	6038320	6720824	5857673	6901471
Sep 2021	5968333	5618632	6318034	5433512	6503155
Oct 2021	6148603	5790652	6506554	5601164	6696042

Image 14

After choosing the best ARIMA model, we are able to proceed to the prediction and forecasting step. By using the ARIMA(0,1,1)(0,1,1)[12] model, we predict the number of air passengers for the next 24 months and the result is shown in the image above. The prediction value also included the forecast value for the 80% and 95% prediction interval.

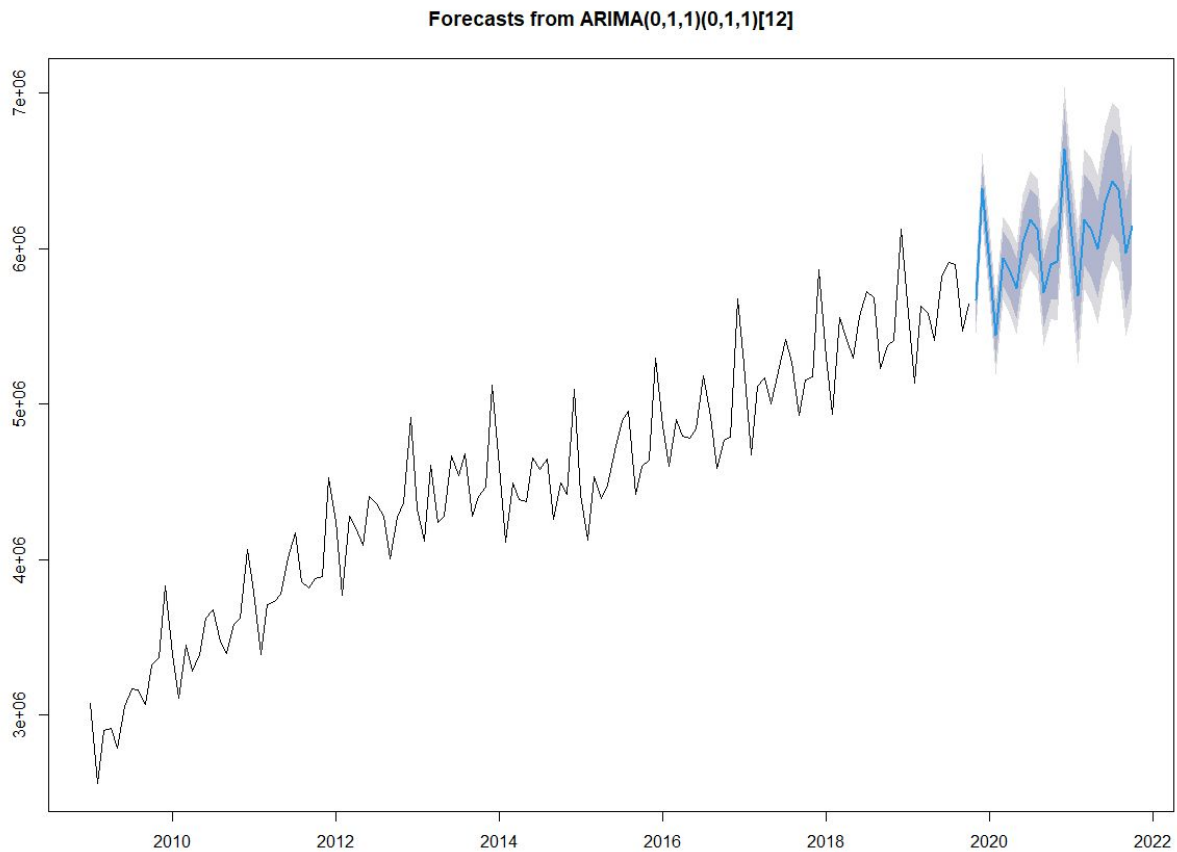


Image 15

We can plot the prediction for the next 24 months as well as the first 10 years data of the number of air passengers in Singapore using the ARIMA(0,1,1)(0,1,1)[12] model. Image 15 above shows the forecast time plot of the future variation for the next 24 months. The darker shaded region in the forecasting plot is the possible value for the 95% prediction interval while the lighter shaded region is the possible value for the 80% prediction interval. This is the overall process by which we can analyze the time series data and perform forecasting from the existing value by using ARIMA.

(h) Discussions and interpretations

From this assignment task given by Dr Chin, we are able to come out with the time series analysis. By doing a research from the internet we found out a few methodologies were suggested and recommended to be used for time series analysis. Therefore, we chose the most suitable methodology which is Box–Jenkins methodology. After we decided the methodology, we started to plot the graph for decomposition. Besides that, we do first differencing to observe the result of its ACF and PACF plot in order to find the suitable order of the non-seasonal AR and non-seasonal MA in the first difference ARIMA model, meanwhile we also do seasonal differencing to interpret the result from the observation of its ACF and PACF plot so that we able to get the suitable order for the seasonal AR and seasonal MA value for the seasonal ARIMA model. After that, we do the best ARIMA model selection. In the selection, we have to do comparison for the model, therefore we are able to observe the AIC score for each model. By having the lowest AIC score, it brings the meaning that it might be the best model. Moving on, we do the coefficient tests to get the significance score. The more the star it gives, the better the model is. After we got the best model, we started to do the forecasting. We had done the forecasting for the next two years.

For future improvements, it is recommended to obtain the air passengers data from different local and international airports. Besides, It is better to include the airline company information for each flight and their number of passengers. It can help to study which airlines are most preferred by passengers. It is interesting to know which airlines own the most passengers and how their flight planning strategy balances the market demand and supply.

Limitation of this time series forecasting is that we are not able to predict some natural disaster or pandemic that may happen in the future. Natural disaster or pandemic may cause the forecasting that we obtain from this time series inaccurate. Such disaster and pandemic is beyond our control and we fit into the forecasting model for prediction. According to Wikipedia (Impact of Covid 2019 Pandemic, 2020), airline companies could lose up to 113 billion of revenue due to

the reduced number of passengers because of the current pandemic situation that is faced by the whole world in the year of 2020.

(i) Conclusion

As a conclusion, by doing the time series analysis of the air passengers in Singapore, we knew that from the past ten years, the number of air passengers have kept increasing as it shows only the upward trend. And those data also carried with the seasonal cycle which the number of passengers are increasing when near to the mid year and the end of year, and will slightly drop when near to the start of the year. This may be because people in Singapore are having their holiday at the mid year and year end, so the air passengers at there are increasing significantly.

On top of that, based on the forecasting graph, we can conclude that the forecasted results that we obtained by using the Box-Jenkins technique shows that the number of air passengers in Singapore will keep increasing for the next two years. Hence, we suggest that Changi Airport may enlarge the airport in order to support more air passengers in the future. Based on the forecasting and the time series analysis results, Airline companies and Changi Airport are able to make better decisions for the future. For instance, airline companies are allowed to plan and utilize their flights schedule at different periods and seasons.

(j) Reference

Bao, Y., Xiong, T. and Hu, Z, 2012, *Forecasting Air Passenger Traffic by Support Vector Machines with Ensemble Empirical Mode Decomposition and Slope-Based Method*, viewed on 1 Sep 2020, <<https://www.hindawi.com/journals/ddns/2012/431512/>>

European union, 2012, *Time Series Analysis CONTENT.*, viewed on 3 Sep 2020, <https://www.espon.eu/sites/default/files/attachments/TR_Time_Series_june2012.pdf>

Ming, W., Bao, Y., Hu, Z. and Xiong, T, 2014, *Multistep-Ahead Air Passengers Traffic Prediction with Hybrid ARIMA-SVMs Models*, viewed on 3 Sep 2020, <<https://www.hindawi.com/journals/tswj/2014/567246/>>

Norhaidah, M.A and Maria, E.N, 2018, *Time Series Forecasting of the Number of Malaysia Airlines and AirAsia Passengers*, viewed on 4 Sep 2020, <https://www.researchgate.net/publication/324426592_Time_Series_Forecasting_of_the_Number_of_Malaysia_Airlines_and_AirAsia_Passengers>

Phyoe, S.M., Guo, R. and Zhong, Z.W, 2017, *An Air Traffic Forecasting Study and Simulation*, viewed on 2 Sep 2020, <<https://grdspublishing.org/index.php/matter/article/view/90/3356>>

UCLA Statistical Consulting, 2013, *Logit Regression | R Data Analysis Examples*, viewed on 5 Sep 2020, <<https://stats.idre.ucla.edu/r/dae/logit-regression/>>

Study.com, 2019, *Time Series Analysis & Its Applications*, viewed on 2 Sep 2020, <<https://study.com/academy/lesson/time-series-analysis-its-applications.html>>

StackExchange, 2016, *R - How are the significance codes determined when summarizing a logistic regression model?*, viewed on 5 Sep 2020, <<https://stats.stackexchange.com/questions/232548/r-how-are-the-significance-codes-determined-when-summarizing-a-logistic-regres>>

Wayne, F.V and Joseph, L.V, 2003, *Time Series Analysis*, viewed on 4 Sep 2020, <https://www.researchgate.net/publication/229633091_Time_Series_Analysis>

Wikipedia 2020, *Impact of the COVID-19 pandemic on aviation*, viewed on 4 Sep 2020, <https://en.wikipedia.org/wiki/Impact_of_the_COVID-19_pandemic_on_aviation#:~:text=On%205%20March%202020%2C%20the,the%20reduced%20number%20of%20passengers.&text=IATA%20further%20revised%20their%20revenue,globally%2C%20a%2044%20percent%20drop>.

Mathworks, 1994, *Box-Jenkins Methodology - MATLAB & Simulink*, viewed on 4 Sep 2020, <<https://www.mathworks.com/help/econ/box-jenkins-methodology.html>>

Vu, X.M and Zhong ,Z. W., 2018, *Forecasting Air Passengers of Changi Airport Based on Seasonal Decomposition and an LSSVM Model*, viewed on 4 Sep 2020, <https://www.researchgate.net/publication/328406271_Forecasting_Air_Passengers_of_Changi_Airport_Based_on_Seasonal_Decomposition_and_an_LSSVM_Model>

Yakovyna, V. and Bachkai, O. , 2013, *The Comparison of Holt -Winters and Box -Jenkins Methods for Software Failures Prediction*, viewed on 5 Sep 2020, <<http://ceur-ws.org/Vol-2136/10000090.pdf>>

(k) Appendix

```
library(ggplot2)
library(forecast)
library(lmtest)
passengers.data <- read.csv(file = "C:/Users/HP ZBook/Desktop/Statistics/
sgpassengers.csv")
passengers.data$value

#Plot graph
Y <- c(passengers.data$value)
Y <- ts(Y, frequency = 12, start = c(2009,1))
plot(Y, ylab="Passengers", main="Number of air passengers in Singapore from
2009 to 2019")
#ACF
ACF <- acf(Y,main="Correlogram for the Air Passengers Dataset")

#Decomposition
components <- decompose(Y)
plot(components)
cbind(components$x,components$trend,components$seasonal,components$random)

#First differencing
diffY <- ts(diff(Y))
plot(diffY, xlab="Time Period (Month)", ylab="Y(t)-Y(t-1)",main="First
Difference")
#ACF
ACF <- acf(diffY,main="Correlogram for first differencing")
#PACF
pacf(diffY,main="Partial Autocorrelation for first differencing")

#seasonal difference
diffsea <- ts(diff(diffY,lag=12,differences = 1))
plot(diffsea, xlab="Time Period (Month)", ylab="Y(t) - Y(t-12)",main="Seasonal
Difference")
#ACF
ACF <- acf(diffsea,main="Correlogram for seasonal differencing")
#PACF
pacf(diffsea,main="Partial Autocorrelation for seasonal differencing")

# fit an ARIMA model and compare
arma(Y, order=c(0,1,1), seasonal = list(order = c(2,1,1), period = 12))
#ARIMA(0,1,1)(2,1,1)[12]
arma(Y, order=c(0,1,2), seasonal = list(order = c(2,1,1), period = 12))
#ARIMA(0,1,2)(2,1,1)[12]
arma(Y, order=c(0,1,1), seasonal = list(order = c(1,1,1), period = 12))
#ARIMA(0,1,1)(1,1,1)[12]
arma(Y, order=c(0,1,2), seasonal = list(order = c(1,1,1), period = 12))
#ARIMA(0,1,2)(1,1,1)[12]
arma(Y, order=c(0,1,1), seasonal = list(order = c(1,1,2), period = 12))
#ARIMA(0,1,1)(1,1,2)[12]
arma(Y, order=c(0,1,2), seasonal = list(order = c(1,1,2), period = 12))
#ARIMA(0,1,2)(1,1,2)[12]
arma(Y, order=c(0,1,1), seasonal = list(order = c(0,1,1), period = 12))
#ARIMA(0,1,1)(0,1,1)[12]
```

```

arima(Y, order=c(0,1,2), seasonal = list(order = c(0,1,1), period = 12))
#ARIMA(0,1,2) (0,1,1) [12]
arima(Y, order=c(0,1,1), seasonal = list(order = c(0,1,2), period = 12))
#ARIMA(0,1,1) (0,1,2) [12]
arima(Y, order=c(1,1,2), seasonal = list(order = c(2,1,1), period = 12))
#ARIMA(1,1,2) (2,1,1) [12]
arima(Y, order=c(1,1,2), seasonal = list(order = c(1,1,1), period = 12))
#ARIMA(1,1,2) (1,1,1) [12]
arima(Y, order=c(1,1,2), seasonal = list(order = c(0,1,2), period = 12))
#ARIMA(1,1,2) (0,1,2) [12]
arima(Y, order=c(1,1,1), seasonal = list(order = c(1,1,1), period = 12))
#ARIMA(1,1,1) (1,1,1) [12]
arima(Y, order=c(1,1,1), seasonal = list(order = c(1,1,2), period = 12))
#ARIMA(1,1,1) (1,1,2) [12]
arima(Y, order=c(2,1,2), seasonal = list(order = c(0,1,1), period = 12))
#ARIMA(2,1,2) (0,1,1) [12]
arima(Y, order=c(2,1,2), seasonal = list(order = c(0,1,2), period = 12))
#ARIMA(2,1,2) (0,1,2) [12]
arima(Y, order=c(2,1,2), seasonal = list(order = c(1,1,2), period = 12))
#ARIMA(2,1,2) (1,1,2) [12]
arima(Y, order=c(2,1,2), seasonal = list(order = c(2,1,2), period = 12))
#ARIMA(2,1,2) (2,1,2) [12]
arima(Y, order=c(2,1,2), seasonal = list(order = c(2,1,1), period = 12))
#ARIMA(2,1,2) (2,1,1) [12]

#test significance of the model
ARIMA1 <- arima(Y, order=c(0,1,1),seasonal = list(order = c(0,1,2), period =
12),method="ML")
coeftest(ARIMA1)
ARIMA2 <- arima(Y, order=c(0,1,1),seasonal = list(order = c(0,1,1), period =
12),method="ML")
coeftest(ARIMA2)
ARIMA3 <- arima(Y, order=c(0,1,1),seasonal = list(order = c(1,1,1), period =
12),method="ML")
coeftest(ARIMA3)

bestmodel <- arima(Y, order=c(0,1,1), seasonal = list(order = c(0,1,1), period
= 12)) #ARIMA(0,1,1) (0,1,1) [12]
bestmodel

#predict(bestmodel,n.ahead = 5)
passengersforecasts <- forecast(bestmodel, h=24)
passengersforecasts
plot(passengersforecasts)

```