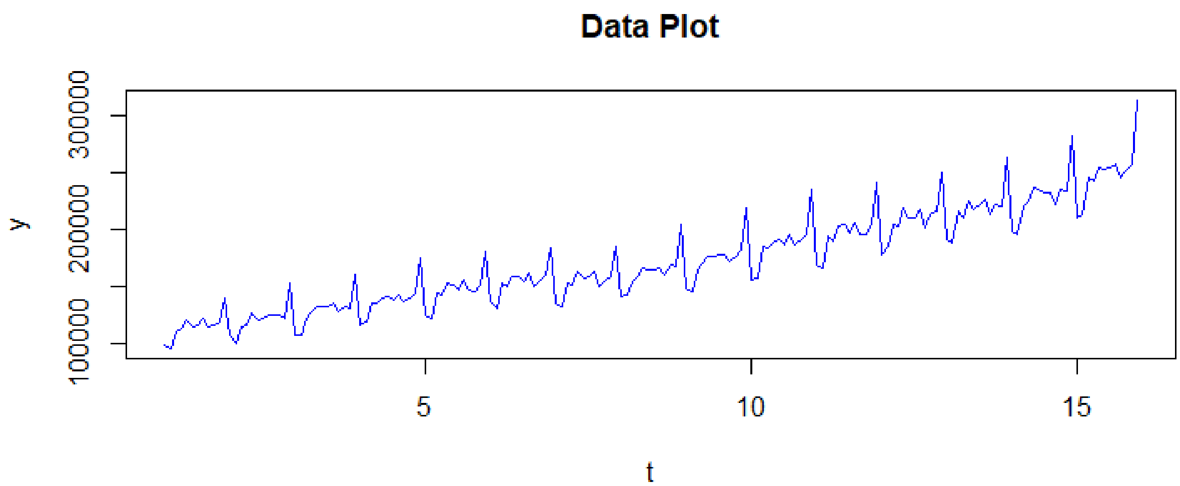
Introduction

In this essay, I have been given a time series monthly data to analyse. I will be identifying the pattern of the data, finding a model that captures the pattern best, estimating the model and making forecasts for 12 periods.

Pattern of the data

Figure 1

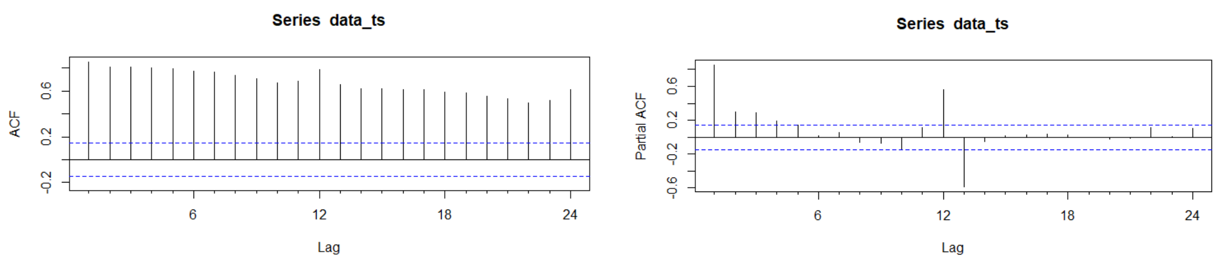


The plot of the data given can be observed in figure 1 above. From figure 1, we can observe that the data contains a trend with seasonal variation. In the data, we observe peaks every 12 months provided with dips after the peak occurs. Towards the end of the time series, we observe that the seasonal variation is growing with larger peaks.

Best model fit for the data

In order to determine the best model for the data, we focus on the ACF and PACF of the data as follows but before that, we must first examine the assumption required by the model which is stationarity.

Figure 2: ACF and PACF of the original data



As observed in the ACF of the data, it dies down very slowly. Thus, we might suspect that the original data is non-stationary. Conducting an ADF test will indicate its stationarity and it is as follows.

Table 1: ADF test of the original data

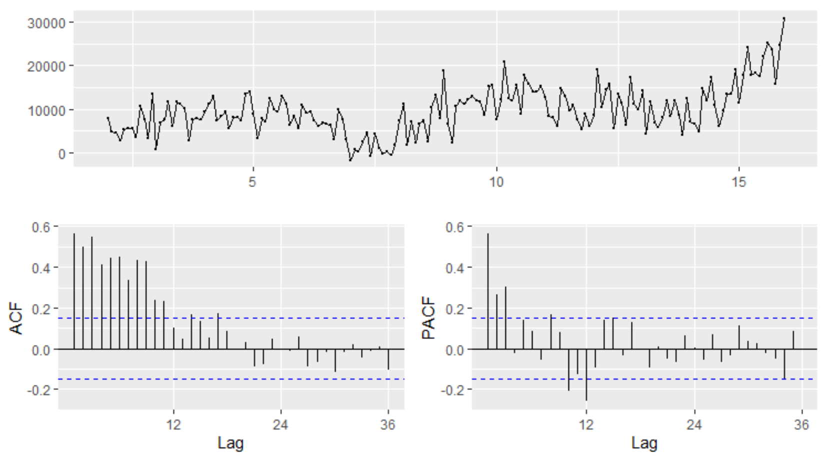
```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 0.145   0.685
[2,]   1 0.704   0.845
[3,]   2 1.302   0.951
[4,]   3 1.858   0.983
[5,]   4 2.599   0.990
Type 2: with drift no trend
     lag    ADF p.value
[1,]   0 -2.642  0.0906
[2,]   1 -1.577  0.4935
[3,]   2 -0.667  0.8133
[4,]   3 -0.129  0.9410
[5,]   4  0.448  0.9829
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -11.14  0.0100
[2,]   1  -9.22  0.0100
[3,]   2  -6.74  0.0100
[4,]   3  -5.07  0.0100
[5,]   4  -3.42  0.0521
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Since the lag 0 for type 1 and 2 of the ADF test yields p-values greater than 0.05, we do not reject the null hypothesis hence, the data is non-stationary. In order to convert the data into a stationary process, we should difference the model accordingly.

Through the pattern of the data as explained after figure 1, we observe seasonality in the data. Thus, we will first take a seasonal difference with a lag of 12 since we have monthly data.

Figure 3: Plot, ACF and PACF of the seasonally differenced data (12 months)



As observed in the ACF of the seasonally differenced data, it still appears to be non-stationary since the ACF dies down slowly thus, we take an additional first difference.

Figure 4: Plot, ACF and PACF of the seasonally differenced (12 months) and first differenced data

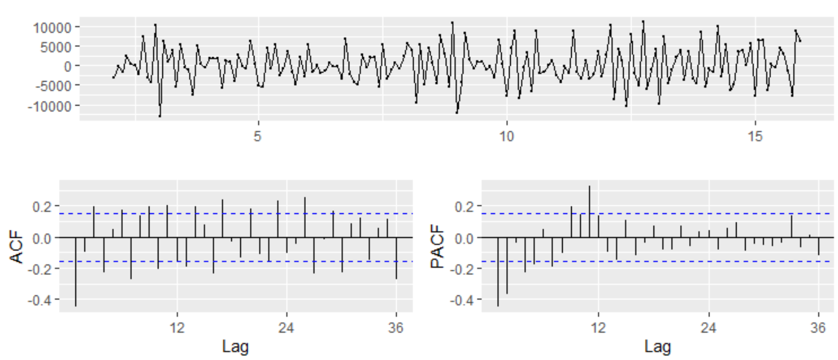


Table 2: ADF test of the differenced data

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag    ADF p.value
[1,]   0 -20.72    0.01
[2,]   1 -16.09    0.01
[3,]   2 -10.42    0.01
[4,]   3 -10.10    0.01
[5,]   4  -9.49    0.01
Type 2: with drift no trend
     lag    ADF p.value
[1,]   0 -20.69    0.01
[2,]   1 -16.09    0.01
[3,]   2 -10.45    0.01
[4,]   3 -10.16    0.01
[5,]   4  -9.59    0.01
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -20.67    0.01
[2,]   1 -16.10    0.01
[3,]   2 -10.46    0.01
[4,]   3 -10.20    0.01
[5,]   4  -9.67    0.01
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

By conducting an ADF test, we ensure that the seasonally differenced and first-differenced data is stationary. Since the p-values are all smaller than 0.05, we reject the null hypothesis of non-stationarity hence, the data is now stationary.

Turning our attention to figure 4, we observe the ACF and PACF to find a model that captures the pattern best. If we perceive the PACF to die down, the significant spike at lag 1 in the ACF suggests a non-seasonal MA(1) component, and the significant spike at lag 12 and 36 in the ACF suggests a seasonal MA(1) component. Therefore, a model that could represent the pattern would be the ARIMA(0,0,1)(0,0,1)[12] model.

Additionally, if we perceive the ACF to die down, the significant spikes at lag 1 suggest a non-seasonal AR(1) model and the significant spikes at lag 2 suggest a non-seasonal AR(2) model. Hence, we can consider the ARIMA(1,0,0)(0,0,1)[12] and ARIMA(2,0,0)(0,0,1)[12] models. Lastly, since the ACF shows significant spikes at lag 1 and the PACF shows significant spikes at lags 1 and 2, we can also consider the ARIMA(1,0,1)(0,0,1)[12] and ARIMA(1,0,2)(0,0,1)[12] models. We now compare the AICs of these models and select the model with the smallest AIC to yield the best fitting model.

Figure 5: AIC values of the ARIMA models

```
> # auto.arima(data_ts)
> arima1 = arima(diff_data_ts, order=c(0,0,1), seasonal=c(0,0,1))
> arima2 = arima(diff_data_ts, order=c(1,0,0), seasonal=c(0,0,1))
> arima3 = arima(diff_data_ts, order=c(1,0,0), seasonal=c(0,0,1))
> arima4 = arima(diff_data_ts, order=c(1,0,1), seasonal=c(0,0,1))
> arima5 = arima(diff_data_ts, order=c(1,0,2), seasonal=c(0,0,1))
> c(AIC(arima1), AIC(arima2), AIC(arima3), AIC(arima4), AIC(arima5))
[1] 3231.851 3265.797 3265.797 3231.900 3215.813
```

As observed in figure 5, the model with the smallest AIC is “arima5” which is the ARIMA(1,0,2)(0,0,1)[12] model. Thus, the model that captures the pattern best in the seasonally differenced (12 months) and first differenced data is the ARIMA(1,0,2)(0,0,1)[12] model. Therefore, in terms of our original data, the model would be ARIMA(1,1,2)(0,1,1)[12] since we account for seasonal differencing and first differencing.

We now conduct a Ljung-Box test on the differenced data and original data to determine if the residuals of the ARIMA model contains any autocorrelations.

Figure 6: Box test for autocorrelation in the residuals of the differenced (left) and original (right) ARIMA model

```
> Box.test(y_res, lag=1, type="Ljung-Box")

        Box-Ljung test

data:  y_res
X-squared = 0.33651, df = 1, p-value = 0.5619
> Box.test(y_res, lag=2, type="Ljung-Box")

        Box-Ljung test

data:  y_res
X-squared = 0.60127, df = 2, p-value = 0.7403
> Box.test(y_res, lag=3, type="Ljung-Box")

        Box-Ljung test

data:  y_res
X-squared = 2.7716, df = 3, p-value = 0.4282
```

```
> Box.test(y_res_og, lag=1, type="Ljung-Box")

        Box-Ljung test

data:  y_res_og
X-squared = 0.2564, df = 1, p-value = 0.6126
> Box.test(y_res_og, lag=2, type="Ljung-Box")

        Box-Ljung test

data:  y_res_og
X-squared = 0.65007, df = 2, p-value = 0.7225
> Box.test(y_res_og, lag=3, type="Ljung-Box")

        Box-Ljung test

data:  y_res_og
X-squared = 2.7857, df = 3, p-value = 0.4259
```

From the p-value of the lag 1, 2 and 3 Box test for the differenced and original data, we determined that the residuals of the ARIMA model do not contain any remaining autocorrelation since the p-value is not smaller than 0.05. Thus, the ARIMA(1,0,2)(0,0,1)[12] model is adequate for the differenced data and the ARIMA(1,1,2)(0,1,1)[12] is adequate for the original data.

Estimate of the model

The following figures represent the estimate of the ARIMA models I have deemed to best represent the data provided.

Figure 7: ARIMA model of the original data

```
Call:
arima(x = data_ts, order = c(1, 1, 2), seasonal = c(0, 1, 1))

Coefficients:
         ar1      ma1     ma2     sma1
      0.8097  -1.7599  0.8701  -0.4661
s.e.  0.0739   0.0431  0.0403   0.0784

sigma^2 estimated as 12349397:  log likelihood = -1602.73,  aic = 3215.46
```

Figure 8: ARIMA model of the data after seasonally differencing and first differencing

```
Call:
arima(x = diff_data_ts, order = c(1, 0, 2), seasonal = c(0, 0, 1))

Coefficients:
         ar1      ma1     ma2     sma1  intercept
      0.7939  -1.7561  0.8660  -0.4744   105.6516
s.e.  0.0753   0.0446  0.0415   0.0782    80.0262

sigma^2 estimated as 12221704:  log likelihood = -1601.91,  aic = 3215.81
```

Figures 7 and 8 represent the estimate of the ARIMA(1,0,2)(0,0,1)[12] and ARIMA(1,1,2)(0,1,1)[12] models generated from R. The following are the equations for the models.

Original Data: $Y_t = 0.8097Y_{t-1} - 1.7599 + 0.87601 - 0.4661\varepsilon_{t-12}$

Differenced Data: $Y_t = 0.7939Y_{t-1} - 1.7561\varepsilon_{t-1} + 0.8860\varepsilon_{t-2} - 0.4744\varepsilon_{t-12} + 105.6516$

Forecasting for 12 periods

Forecasting for 12 periods in terms of the original data

Figure 9: 12 periods ahead forecasts for the original data

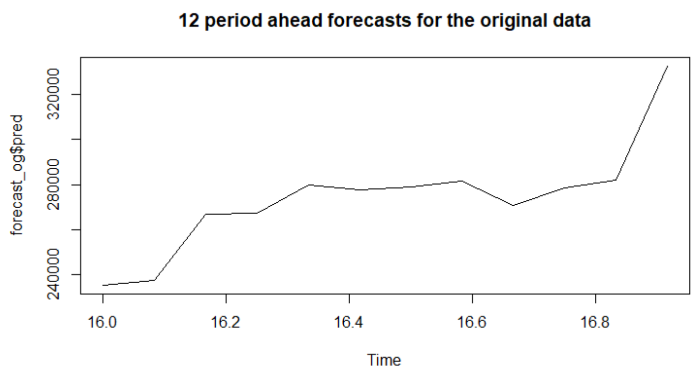
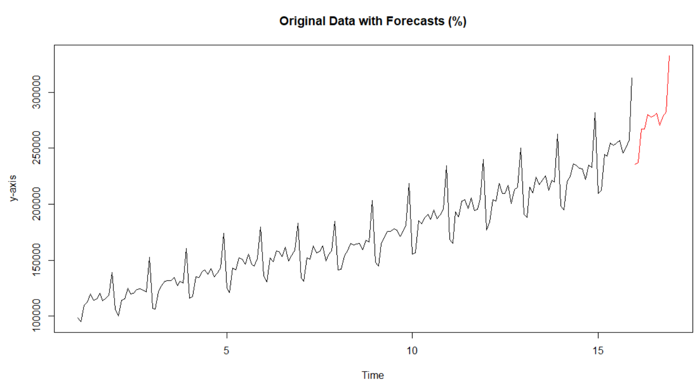


Figure 10: Original Data + 12 periods ahead forecasts for the original data



Figures 9 and 10 represent the forecasting for 12 periods ahead in terms of the original data.

The point forecasts along with its standard error are determined from R as follows.

```
> forecast_og
$pred
        Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
16 235403.7 237353.6 267053.8 267285.8 280003.2 277692.3 278994.3 281557.0 270684.1 278644.6 282206.8 332582.6

$se
        Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
16 3514.171 3518.530 3558.067 3650.322 3797.644 3993.526 4227.695 4489.497 4769.599 5060.557 5356.762 5654.150
```

The 95% interval forecasts can be computed in the following manner:

$\mu_{t+1|t} - 1.96 \times \sigma_{t+1|t}$, $\mu_{t+1|t} + 1.96 \times \sigma_{t+1|t}$ where $\mu_{t+1|t}$ represents the point forecasts and $\sigma_{t+1|t}$ represents the standard errors.

Table 3: Interval Forecasts

Jan	$235403.7 \pm 1.96 \times 3514.171$
Feb	$237353.6 \pm 1.96 \times 3518.530$
Mar	$267053.8 \pm 1.96 \times 3558.067$
Apr	$267285.8 \pm 1.96 \times 3650.322$
May	$280003.2 \pm 1.96 \times 3797.644$
Jun	$277692.3 \pm 1.96 \times 3993.526$
Jul	$278994.3 \pm 1.96 \times 4227.695$
Aug	$281557.0 \pm 1.96 \times 4489.497$
Sep	$270684.1 \pm 1.96 \times 4769.599$
Oct	$278644.6 \pm 1.96 \times 5060.557$
Nov	$282206.8 \pm 1.96 \times 5356.762$
Dec	$332582.6 \pm 1.96 \times 5654.150$

In terms of the accuracy of the forecasts, we can analyse the RMSE values of our ARIMA models.

Table 4: Accuracy Values of our ARIMA models

```
> arima11 = arima(data_ts, order=c(0,1,1), seasonal=c(0,1,1))
> arima22 = arima(data_ts, order=c(1,1,0), seasonal=c(0,1,1))
> arima33 = arima(data_ts, order=c(1,1,0), seasonal=c(0,1,1))
> arima44 = arima(data_ts, order=c(1,1,1), seasonal=c(0,1,1))
> arima5 = arima(data_ts, order=c(1,1,2), seasonal=c(0,1,1))
> accuracy(arima11)
                   ME     RMSE      MAE      MPE     MAPE     MASE        ACF1
Training set 406.7666 3626.822 2762.747 0.125195 1.592525 0.218284 -0.09235709
> accuracy(arima22)
                   ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
Training set 223.2949 3993.652 3019.259 0.02531783 1.757731 0.2385509 -0.210464
> accuracy(arima33)
                   ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
Training set 223.2949 3993.652 3019.259 0.02531783 1.757731 0.2385509 -0.210464
> accuracy(arima44)
                   ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 407.4284 3603.376 2700.687 0.1239267 1.556834 0.2133806 -0.02723103
> accuracy(arima5)
                   ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set 280.3837 3385.019 2565.738 0.06407803 1.481256 0.2027183 0.03742953
```

Since “arima5”, ARIMA(1,1,2)(0,1,1)[12], has the smallest RMSE out of all the other arima models, we can conclude that the ARIMA model that has been selected yields the lowest root mean squared error when forecasting.

Conclusion

To summarise this essay, the provided time series monthly data was discovered to be non-stationary thus, in order to determine an ARIMA model for the data, we would have to seasonally difference and first difference the data. We then analysed the ACF and PACF generated from the differenced data in order to determine several ARIMA models that can be considered to model our data. Out of the models generated, we identified the model with the smallest AIC which is the model that best captures the pattern of the data. We then conducted 12 month forecasts and used the RMSE measure to ensure that our model provides the least forecasting error out of the other possible ARIMA models. Thus, the model that is most optimal for the data is the ARIMA(1,1,2)(0,1,1)[12] model.