Code ▾

Assignment 4: Suggested Solutions & Feedback

ETF3231/5231

Read and tidy up the data using previous code.

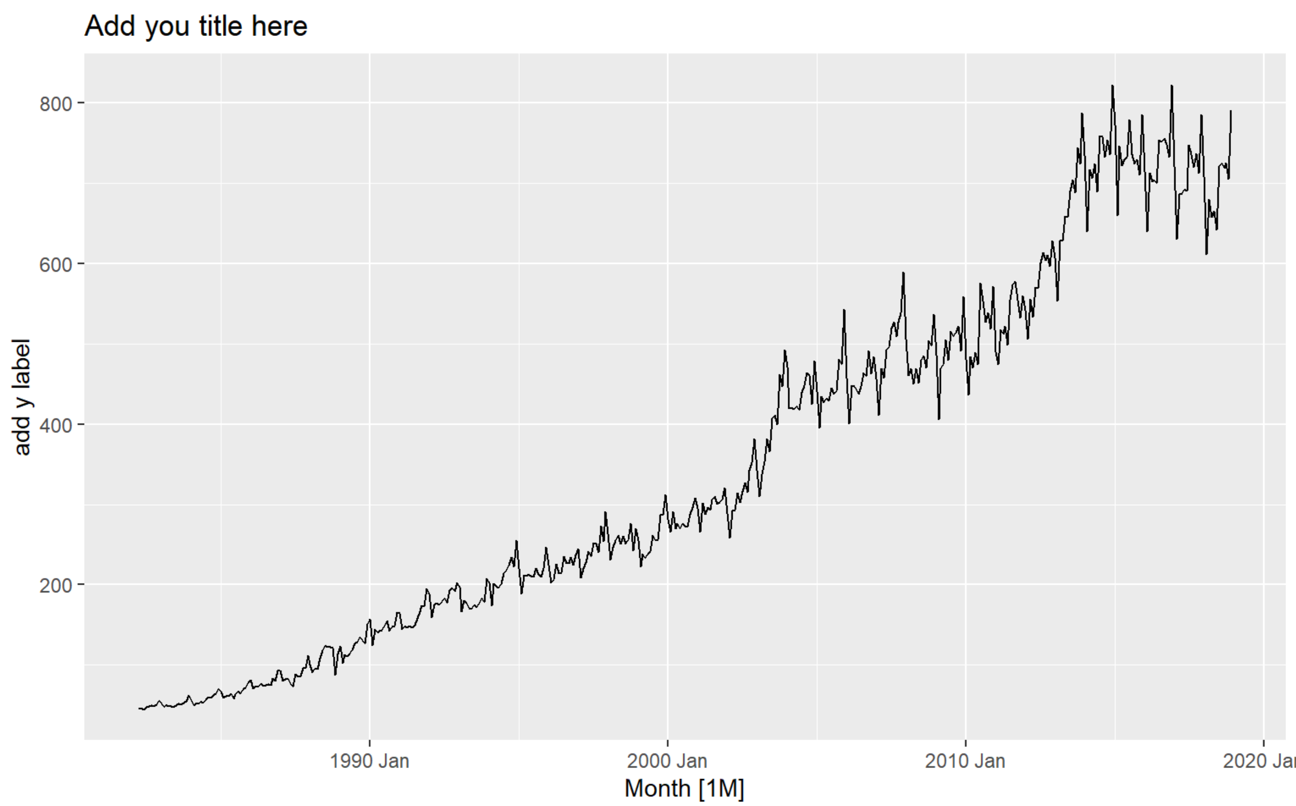
> Your aim in the first part of the assignment is to build an ARIMA model and use this to forecast. Recall the first step in ARIMA modelling is to stabilize the variance of your data. If you decided that your data required a transformation in the previous assignments you will be required to use the same transformation for what follows (unless you have a reason to change your mind - please do so if you think it is necessary).

Question 1

> Visually inspect your data and decide on the transformation and what differencing is required to achieve stationarity. Plot the data at every step and comment on each plot justifying your actions. (No more than 50 words per plot). (6 marks)

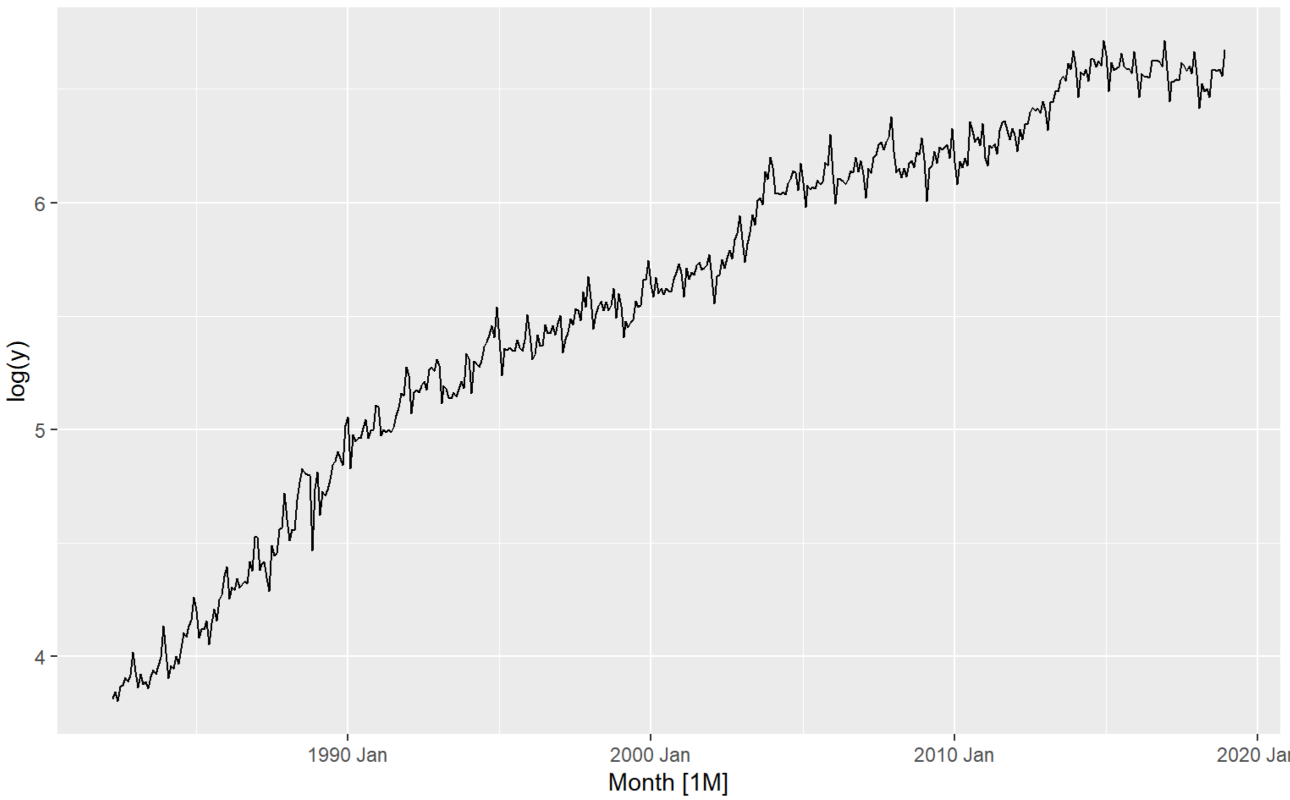
Hide

```
myts %>% autoplot(y) + ggtitle("Add you title here") +
  ylab("add y label")
```



Hide

```
myts %>% autoplot(log(y))
```

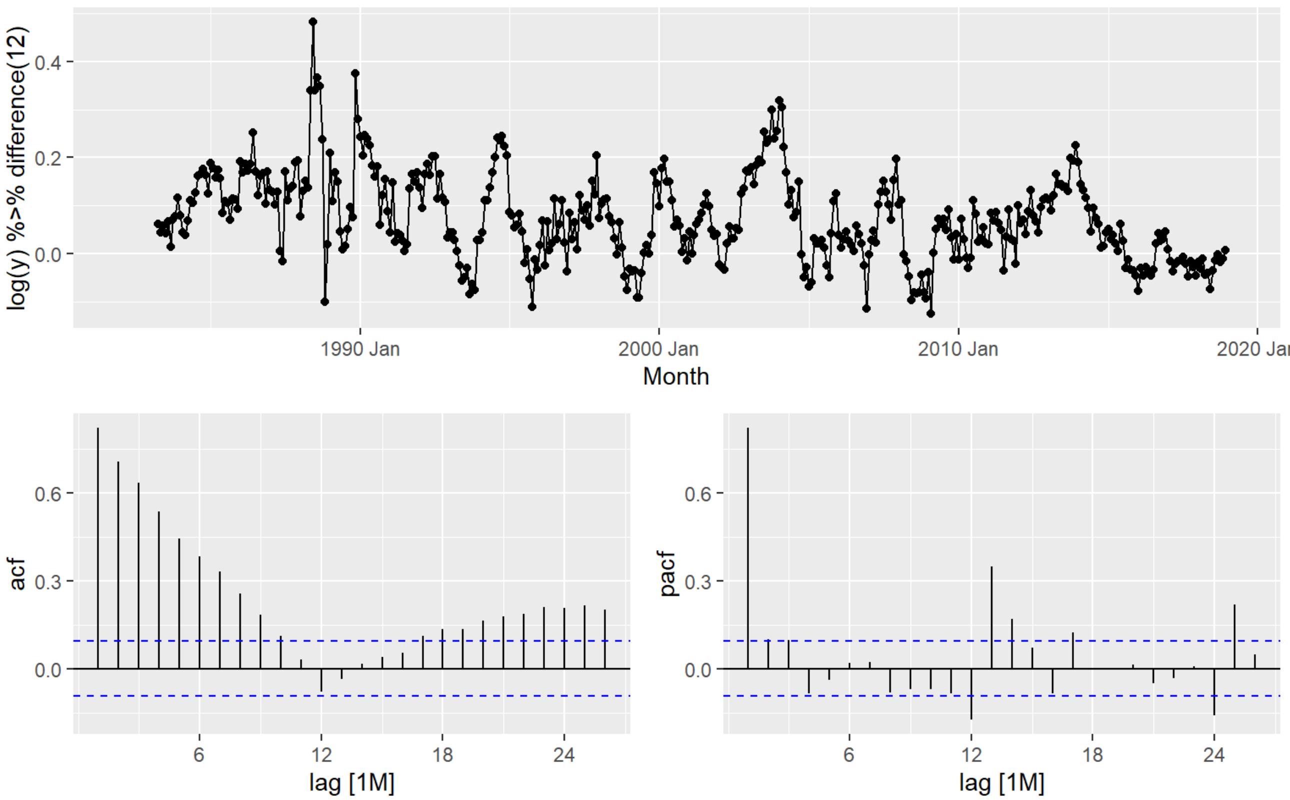


Hide

```
myts %>% gg_tsdisplay(log(y) %>% difference(12), plot_type = 'partial
')
```

```
## Warning: Removed 12 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```

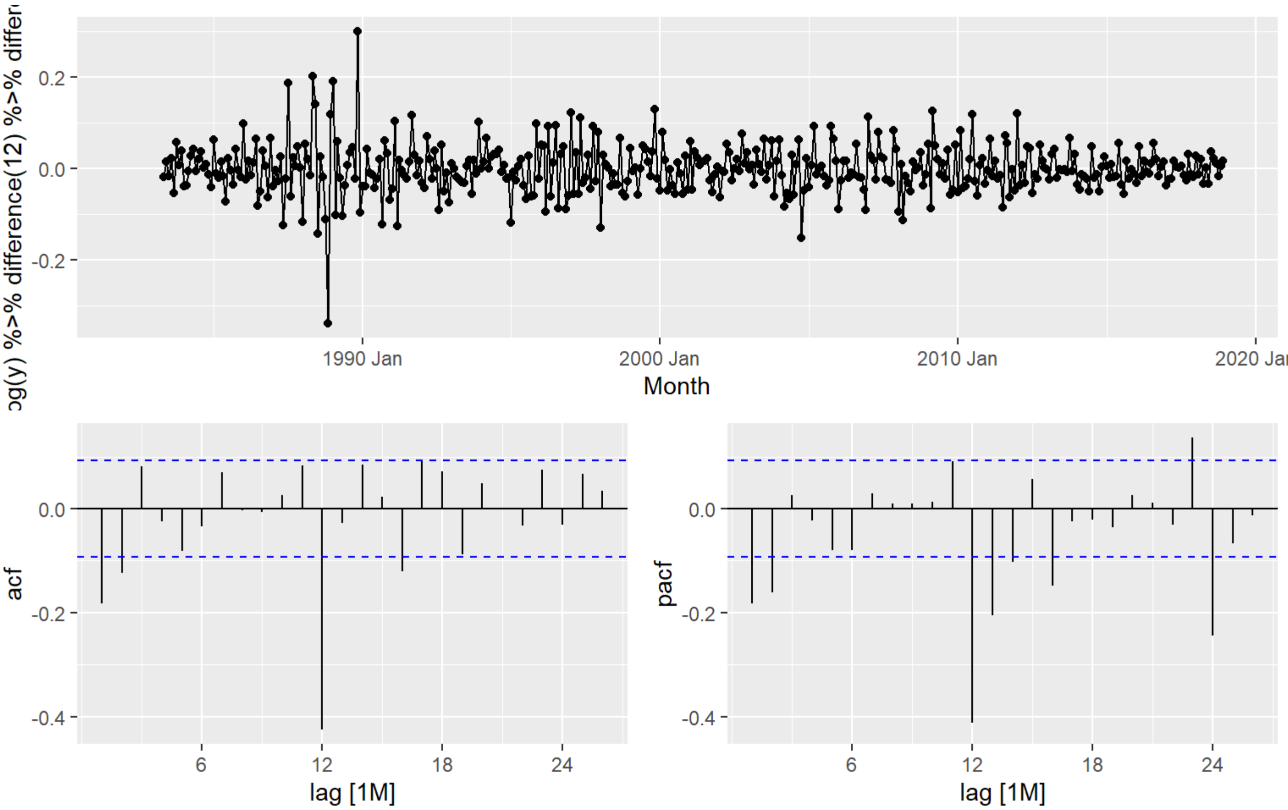


Hide

```
myts %>% gg_tsdisplay(log(y) %>% difference(12)%>% difference(), plot
_type = 'partial')
```

```
## Warning: Removed 13 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```



Marking guide:

1. Transformation and discussion (2m)
    a. If not needed, explain
2. Appropriate plots and explanation
    a. Original plot (1m)
        i. Explain what action needs to be taken to achieve stationarity (1m)
    b. Stationary plot (1m)
        i. Explain whether the series is stationary or not by providing the discussion on ACF (possibly unit root test), constant mean or any predictable pattern (1m)

Expectations

I. Transformation (or appropriate discussion)
II. Appropriate plots
III. Appropriate justification of differences

Common errors:

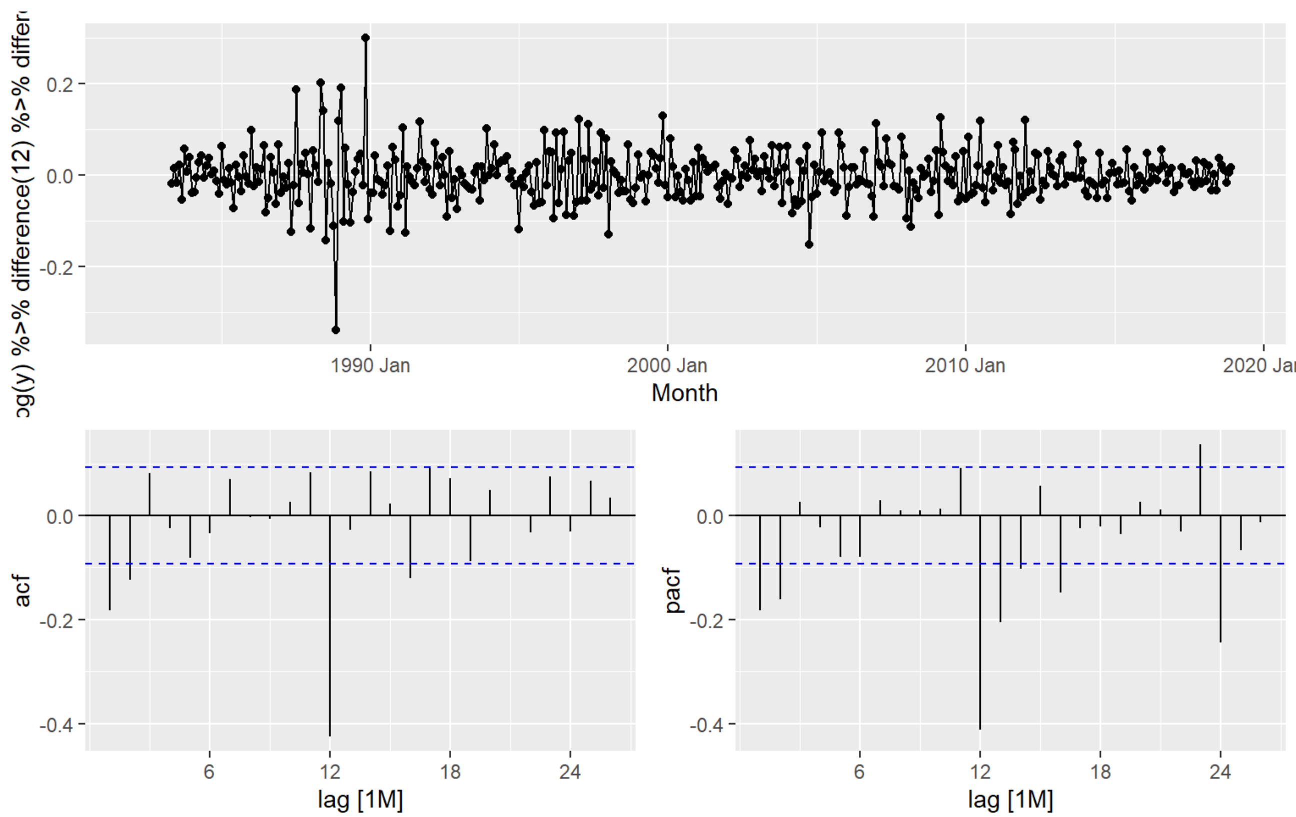
1. Fail to discuss the plot in each step.
2. Some students decided that only one seasonal difference was required, although a first order difference was still required due to a slowly decaying ACF() plot. (FOR US - LETS NOT BE TOO HARSH ON THIS - IF THEY JUSTIFY AND THEY POSSIBLE REVISE FURTHER BELOW IT IS OK - BUT IF CLEARLY NEEDING ANOTHER DIFFERENCE THAT IS FINE - PENALISE)

Question 2

> Plot the ACF and PACF of the stationary data. Reading from these choose an appropriate ARIMA model. Make sure you justify your choice. (No more than 70 words in total – do not revise the theory – describe what you see on your plots and decide what ARIMA orders may be appropriate. Also note that it is highly likely that the ACF and PACF plots will be very messy. Do the best you can.). (6 marks)

Hide

```
myts %>% gg_tsdisplay(log(y) %>% difference(12) %>% difference(), plo
t_type = 'partial')
```



Oh yes very messy. I would go for an AR(2) due to the non seasonal spikes in the PACF and a seasonal MA(1) due to the large spike at 12 in the ACF. So the suggested model is an ARIMA(2,1,0)(0,1,1)[12]

Marking guide:

1. ACF and PACF plot for stationary data (1m)

2. Identification of any AR/SAR or MA/SMA term with explanation (5m)

Expectation I. Appropriate plots provided II. Identification of potential models from the plots III. Graph based decision of most appropriate model

Common errors:

1. Did not identifying the appropriate model based on the plot, instead opting to select a model based on AICc.
2. Did not specify the ARIMA model based on the ACF/PACF plot but just briefly explain there were spikes outside the bounds.

Question 3

> Check the whiteness of the residuals from the fitted ARIMA model. Based on these evaluate and if necessary review the ARIMA model specified in Q2. (No more than 50 words). (4 marks)

Hide

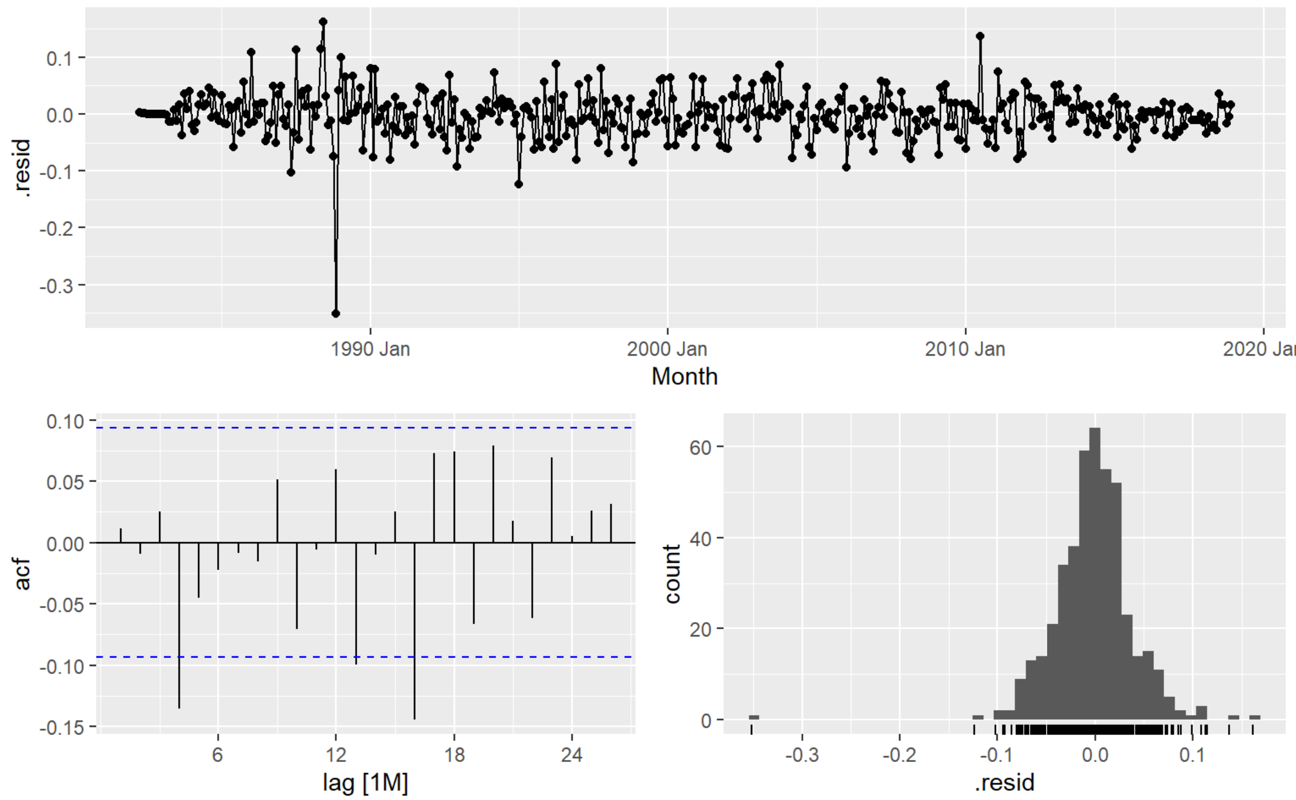
```
arima_210_011 <- myts %>% model(
  ARIMA(log(y)~pdq(2,1,0)+PDQ(0,1,1))
  )

arima_210_011 %>% report()
```

```
## Series: y
## Model: ARIMA(2,1,0)(0,1,1)[12]
## Transformation: log(y)
##
## Coefficients:
##          ar1      ar2     sma1
##      -0.2698  -0.2054  -0.8711
## s.e.   0.0474   0.0472   0.0313
##
## sigma^2 estimated as 0.001727:  log likelihood=747
## AIC=-1486   AICc=-1485.91   BIC=-1469.77
```

Hide

```
arima_210_011 %>% gg_tsresiduals()
```



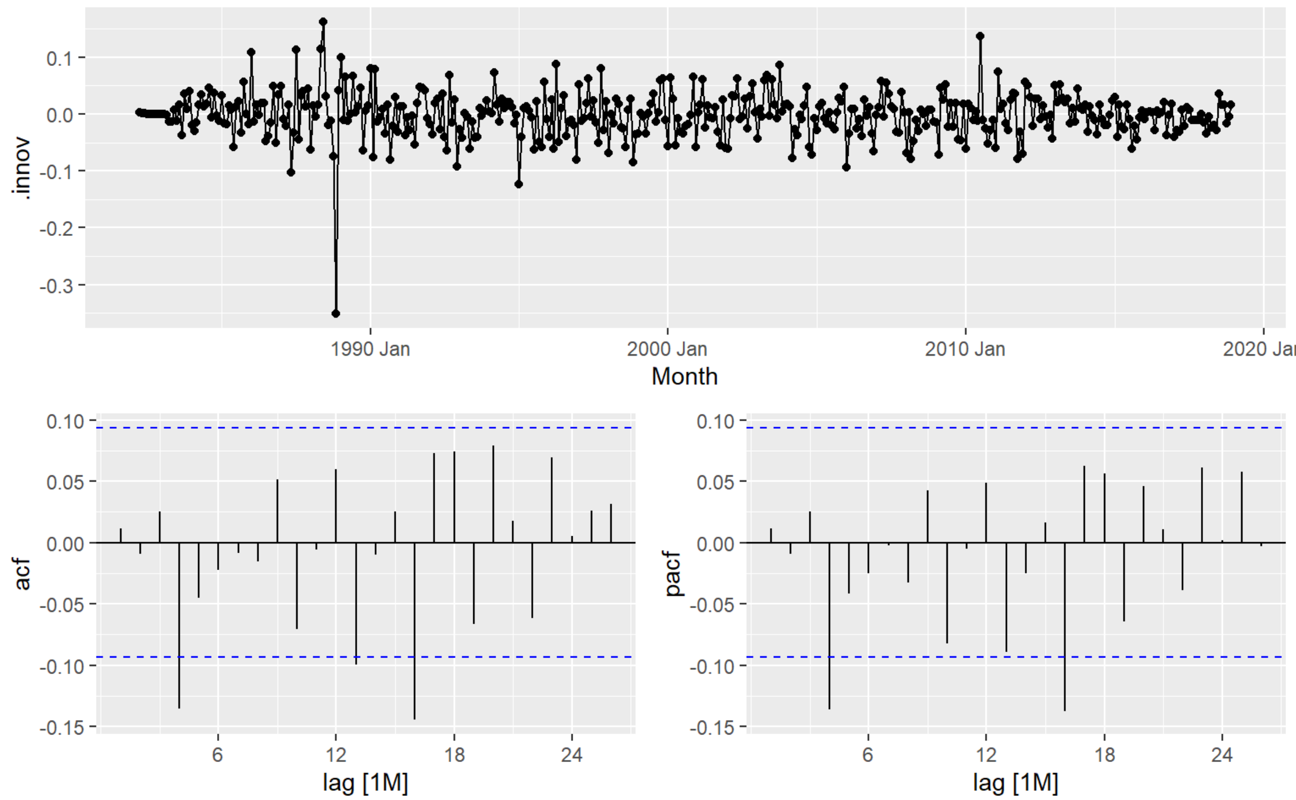
Hide

```
augment(arima_210_011) %>% features(.innov, ljung_box, lag=24, dof=3)
```

```
## # A tibble: 1 x 3
##   .model                                          lb_stat lb_pvalue
##   <chr>                                             <dbl>     <dbl>
## 1 ARIMA(log(y) ~ pdq(2, 1, 0) + PDQ(0, 1, 1))     43.5   0.00267
```

Hide

```
augment(arima_210_011) %>% gg_tsdisplay(.innov, plot_type = 'partial
')
```



There are a couple of large residuals (see the long tails in the histogram). Histogram looks close to normal. The residuals are definitely not WN. The Ljung-Box test with 3 dof verifies the non-whiteness of the residuals as the Null of joint zero autocorrelation for the first 24 lags of the residuals is easily rejected with a p-value=0 at any reasonable level of significance.

However, I don't think there is much I can do manually to improve the model by reading the ACF/PACF of the residuals. A combination of the non-seasonal AR and MA orders is probably needed and I will try those next.

Marking guide:

1. Plot/Comment on the residuals (1m)
2. Plot/Comment on the histogram of the residuals (1m)
3. Plot/Comment on the ACF/PACF plot (1m)
4. Perform LB test, with the Null hypothesis and correct dof. Comment on the result (1m)

Expectation:

I. Analysis of ACF/PACF or residuals
II. Ljung-box test with appropriate dof
III. Review of ARIMA model

Common errors:

1. Incorrectly used the Ljung-Box test, often miscalculating dof or not interpreting the

test correctly.
2. Fail to review the existing model by just providing the conclusion of whiteness of the residuals.
3. No analysis on ACF/PACF plot in detail.

Question 4

> Consider three (up to five if you think you need them) alternative ARIMA models based on your choice in Q2 and Q3. (Very briefly justify each choice with no more than 1 or 2 lines each). Use information criteria to choose the best model you have considered so far. (Very briefly justify each choice with no more than 1 or 2 lines each). (6 marks)

Hide

```
arima.models <- myts %>% model(
  arima_210_011 = ARIMA(log(y)~pdq(2,1,0)+PDQ(0,1,1)),
  arima_012_011 = ARIMA(log(y)~pdq(0,1,2)+PDQ(0,1,1)),
  arima_410_011 = ARIMA(log(y)~pdq(4,1,0)+PDQ(0,1,1)),
  arima_212_011 = ARIMA(log(y)~pdq(2,1,2)+PDQ(0,1,1))
  )

arima.models %>%  glance() %>% select(.model, AIC, AICc, BIC)
```

```
## # A tibble: 4 x 4
##   .model          AIC   AICc    BIC
##   <chr>         <dbl>  <dbl>  <dbl>
## 1 arima_210_011 -1486. -1486. -1470.
## 2 arima_012_011 -1483. -1483. -1467.
## 3 arima_410_011 -1489. -1488. -1464.
## 4 arima_212_011 -1487. -1487. -1463.
```

My ACF/PACF plots are very messy so I don't have many alternatives here. My models are slight variations of the original model. Notice in `arima_012_011` I select an MA(2) as an alternative to the AR(2) for the two spikes in the ACF. I also use an AR(4) for the large spike in the PACF of the residuals and then a combination of ARs and MAs.

I have a conflict in my information criteria with the lowest AICc selecting the `arima_410_011` while the lowest BIC pointing to an `arima_210_011`.

Marking guide:

1. State at least 3 alternative ARIMA models and justify. (4m)
2. Choose the best model using the information criteria. (2m)

Expectation

I. Generation and justification of alternative ARIMA models
II. Choice of information criterion
III. Best model selection

Common errors:

1. Fail to elaborate on how they created the alternative ARIMA models. AICc can be used for comparison of models only with the same orders of differencing.

Question 5

> Let the `ARIMA()` function choose a model. How does this compare with your chosen model from Q4? If you need to, make `ARIMA()` search harder exploring all possible options within the function. Perform a residual diagnostics analysis for your chosen model. (No more than 100 words). (6 marks).

Hide

```
arima_models_auto <- myts %>%
  model(arima_auto1 = ARIMA(log(y)),
        arima_auto2 = ARIMA(log(y), stepwise = FALSE, approximation =
FALSE)
  )

arima_models_auto %>% select(arima_auto1) %>%  report()
```

```
## Series: y
## Model: ARIMA(1,1,2)(0,1,1)[12]
## Transformation: log(y)
##
## Coefficients:
##           ar1     ma1      ma2     sma1
##       -0.6452  0.3949  -0.3279  -0.8705
## s.e.   0.1439  0.1409   0.0486   0.0314
##
## sigma^2 estimated as 0.001711:  log likelihood=749.42
## AIC=-1488.84   AICc=-1488.7   BIC=-1468.55
```

Hide

```
arima_models_auto %>% select(arima_auto2) %>%  report()
```

```
## Series: y
## Model: ARIMA(0,1,4)(0,1,1)[12]
## Transformation: log(y)
##
## Coefficients:
##           ma1      ma2     ma3      ma4     sma1
##       -0.2618  -0.1477  0.1212  -0.1514  -0.8693
## s.e.   0.0482   0.0492  0.0502   0.0492   0.0308
##
## sigma^2 estimated as 0.001697:  log likelihood=751.76
## AIC=-1491.51   AICc=-1491.31   BIC=-1467.16
```

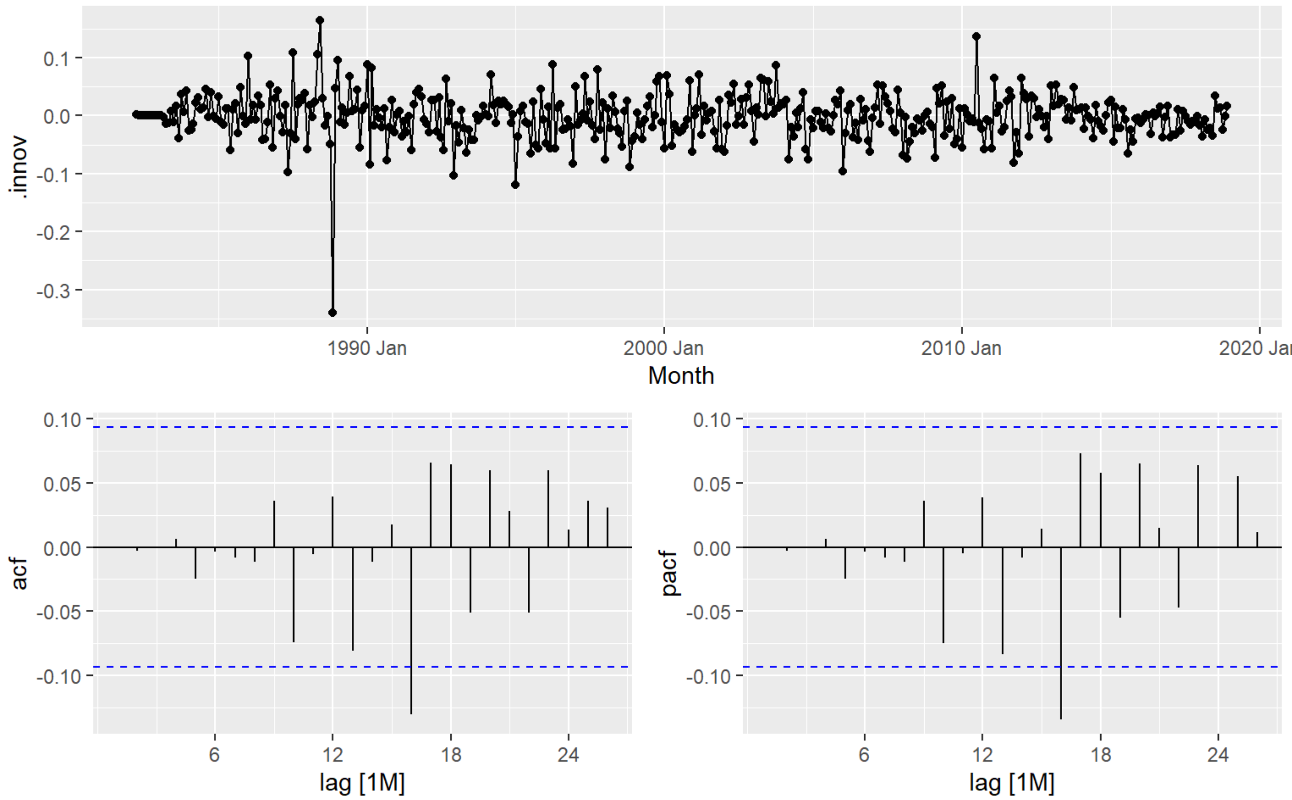
Hide

```
arima_models_auto %>%
  glance() %>%
  select(.model, AIC, AICc, BIC)
```

```
## # A tibble: 2 x 4
##   .model         AIC   AICc    BIC
##   <chr>        <dbl>  <dbl>  <dbl>
## 1 arima_auto1 -1489. -1489. -1469.
## 2 arima_auto2 -1492. -1491. -1467.
```

Hide

```
arima_models_auto %>%
  select(arima_auto2) %>%
  augment() %>%
  gg_tsdisplay(.innov, plot_type = 'partial')
```



Hide

```
arima_models_auto %>%
  select(arima_auto2) %>%
  augment() %>%
  features(.innov, ljung_box, lag=24, dof=5)
```

```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 arima_auto2    25.4     0.148
```

The two models from the automated process (even with the wider search process as compared to the stepwise which skips some combinations) are not too different from the model I manually selected. I could have easily selected an MA(4) in place of the AR(4). The `arima_auto2` is the model with the lowest AICc so I will go with this. The residuals look pretty good with only one significant spike at lag 16, and Ljung-Box test does not reject the null of the first 24 autocorrelations of the innovation residuals being jointly zero.

Marking guide:

1. Run ARIMA() and explain how the model is chosen (using min AICc). (2m)
2. Compare with Q4 with the chosen ARIMA and state what is the different. If different number of differencing is used than AICc is not comparable. (2m)
3. Residual diagnostics (2m)

Expectation

I. Appropriate use of ARIMA()
II. Comparison with Q4 model, note that different differences makes the model IC are incomparable if these are different.
III. Residual diagnostics

Common errors: 1. Fail to compare models and state the difference. 2. Fail to discuss the residuals in detail. 3. Fail to mention how the ARIMA select the best model.

Question 6

> Remember that you cannot use information criteria to compare between models with different orders of differencing. If necessary use an appropriate test set to choose the ARIMA model you want to use for forecasting. Which model have you selected and why? (No more than 50 words). (4 marks)

My models are all of the same order so I do not have to do another comparison.

Full marks if you didn't need to do a test comparison and just added a comment.

Marking guide:

1. Perform test set RMSE based on appropriate test set. (2m)
2. Select the correct model (2m)

Expectation

I. Appropriate selection method for forecasting
II. Model selection and justification

Common errors:

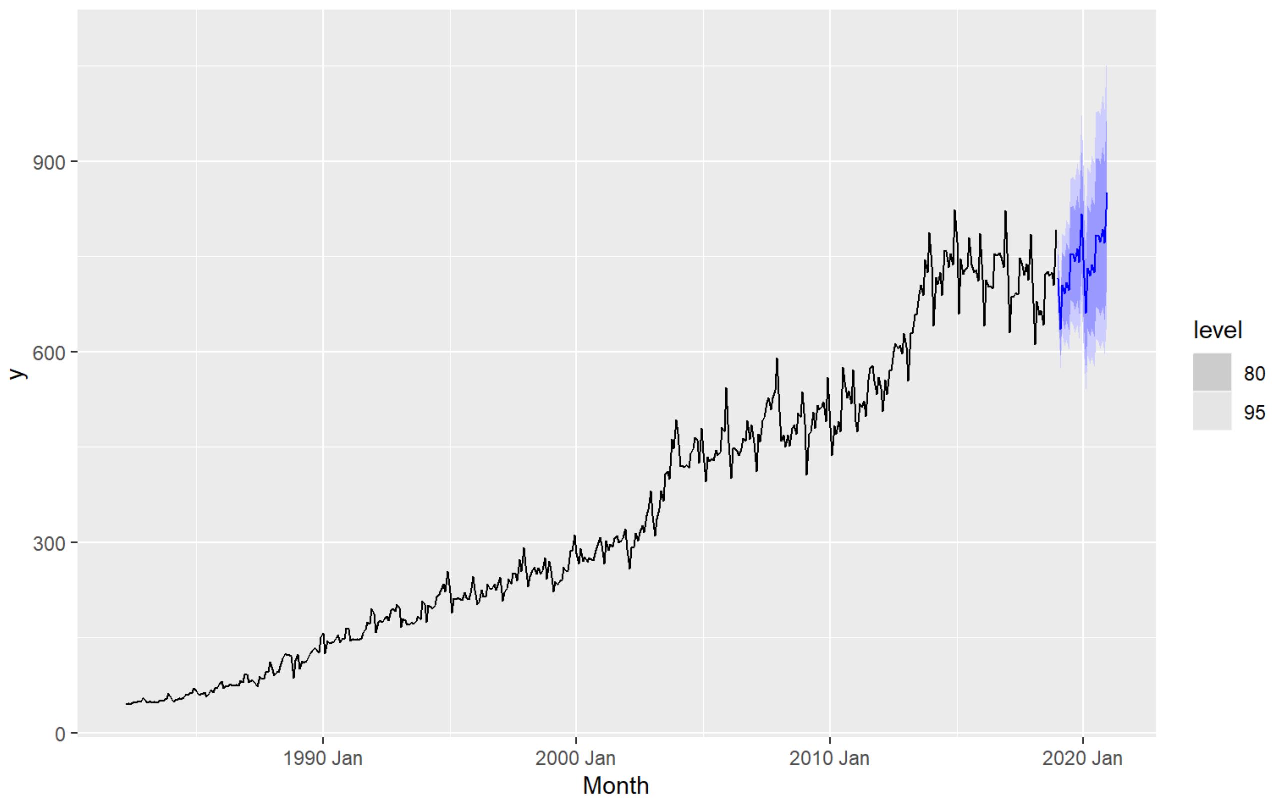
1. Missing explanation on why RMSE test set is not needed in the case when all differencing are the same but straight away provided the answers on choosing the model with lowest AICc.

Question 7

> Generate and plot forecasts and forecast intervals from your chosen ARIMA model for two years following the end of your sample. Comment on these. (No more than 50 words). (3 marks)

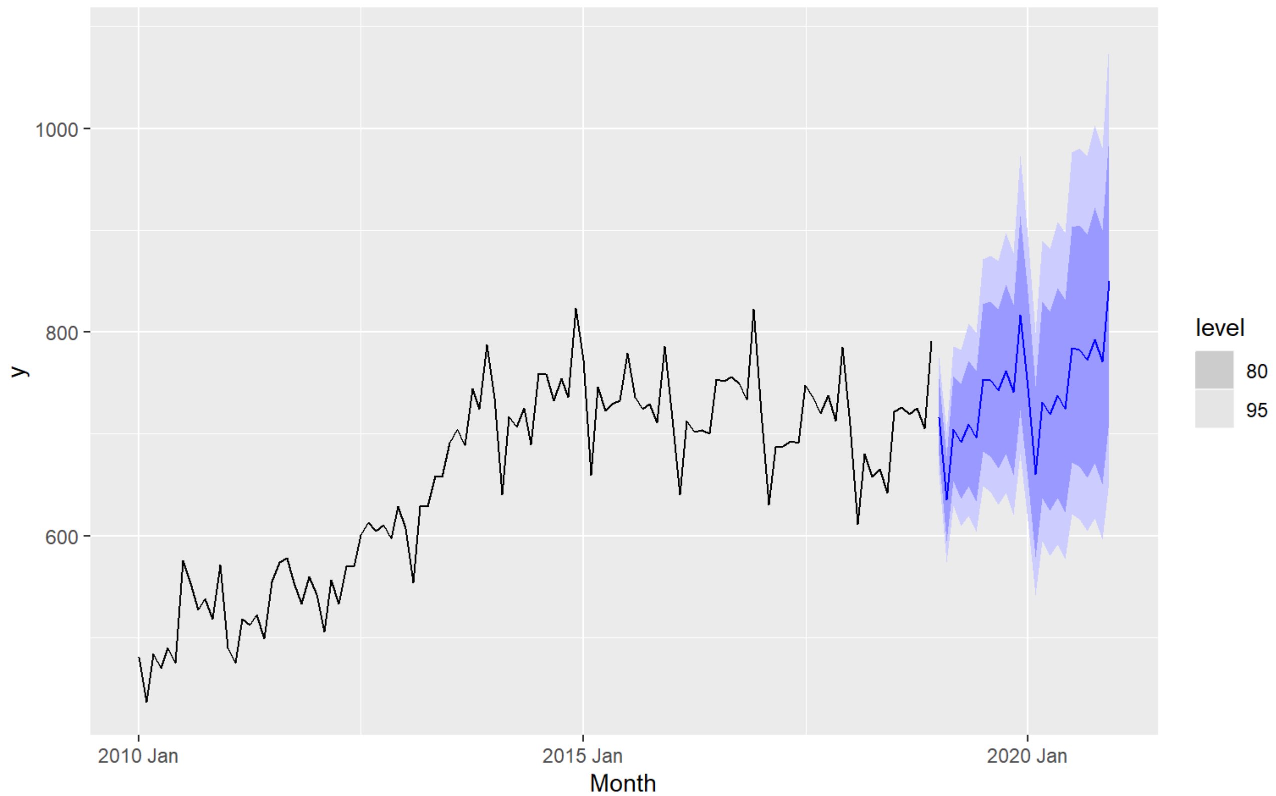
Hide

```
arima_models_auto %>%
  select(arima_auto2) %>%
  forecast() %>%
  autoplot(myts)
```



Hide

```
arima_models_auto %>%
  select(arima_auto2) %>%
  forecast() %>%
  autoplot(filter(myts,year(Month)>=2010))
```



Forecasts seem reasonable both in terms of trend and variance. Prediction intervals seem reasonably wide given the cyclical nature of the data.

Marking guide:

1. Plot forecasts and forecast intervals (2m)
2. Comment on the forecast and intervals (1m)

Expectation

I. Generate forecasts
II. Plot forecasts
III. Comment

Common errors:

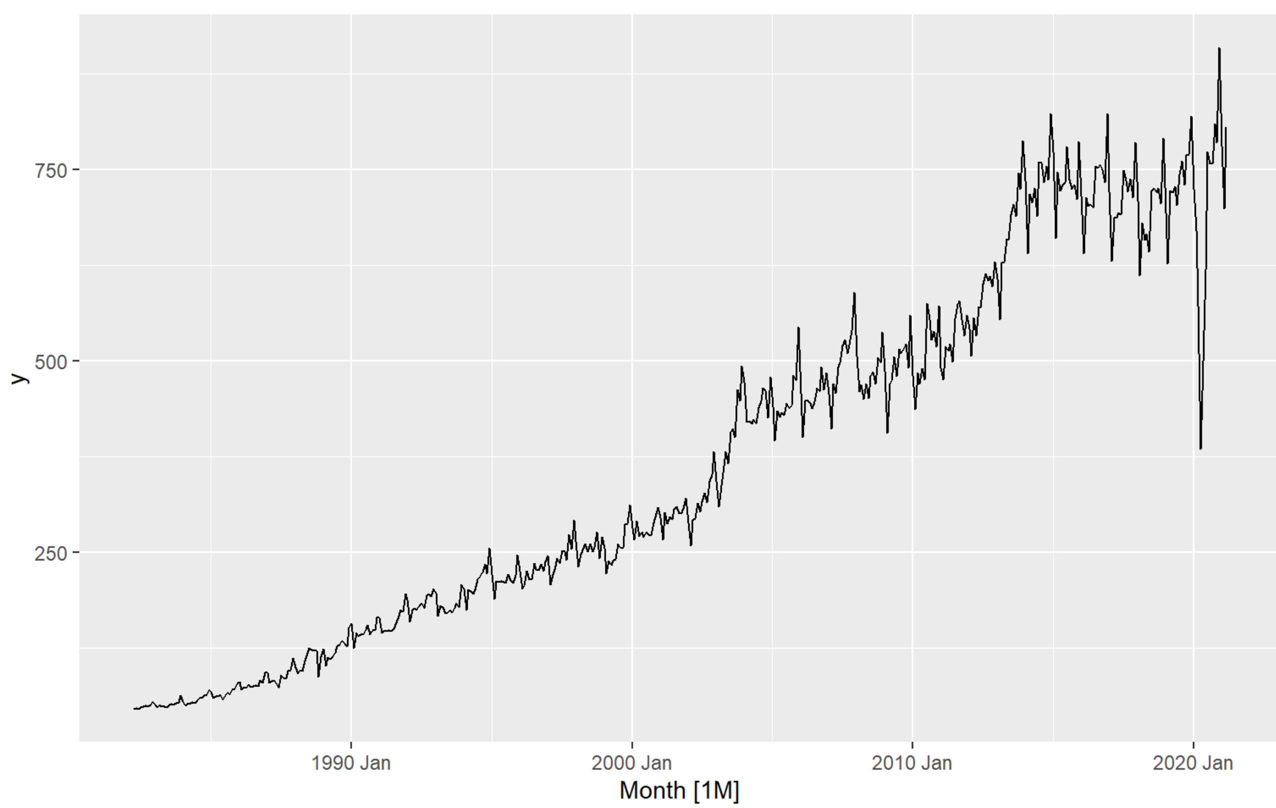
1. Fail to comment on the forecast.
2. Just mentioning the model forecast well is not sufficient, detail explanation such as discussion of the pattern is needed.

Question 8

> You have now considered several modelling frameworks and built several models for your data set. In this part of the assignment you will evaluate these.
>
> Find your data on the ABS website and update your series till the end of the current sample. Explore your updated time series and comment on the effects of the COVID-19 pandemic. Provide any necessary plots to support your analysis. Some States and Industries are unfortunately affected more than others. (6 marks)*

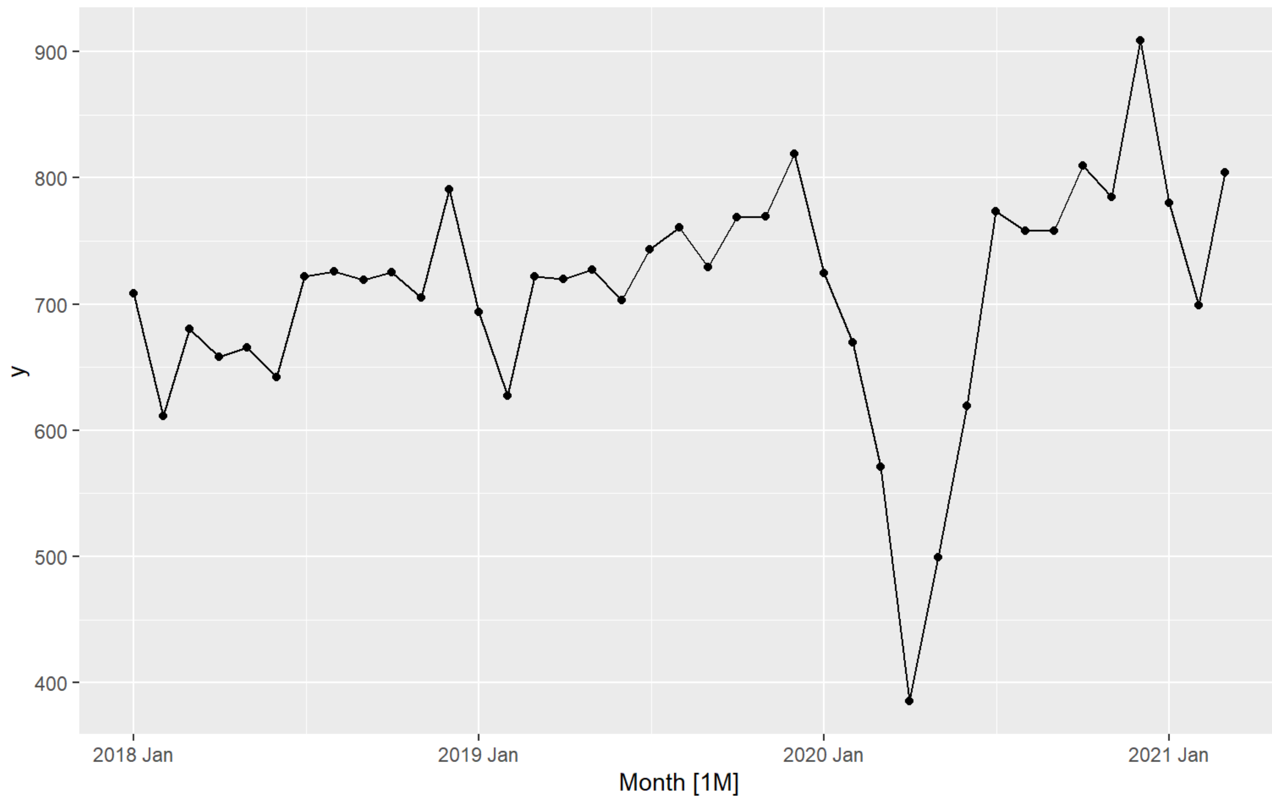
Hide

```
myts_full %>%
  autoplot(y)
```



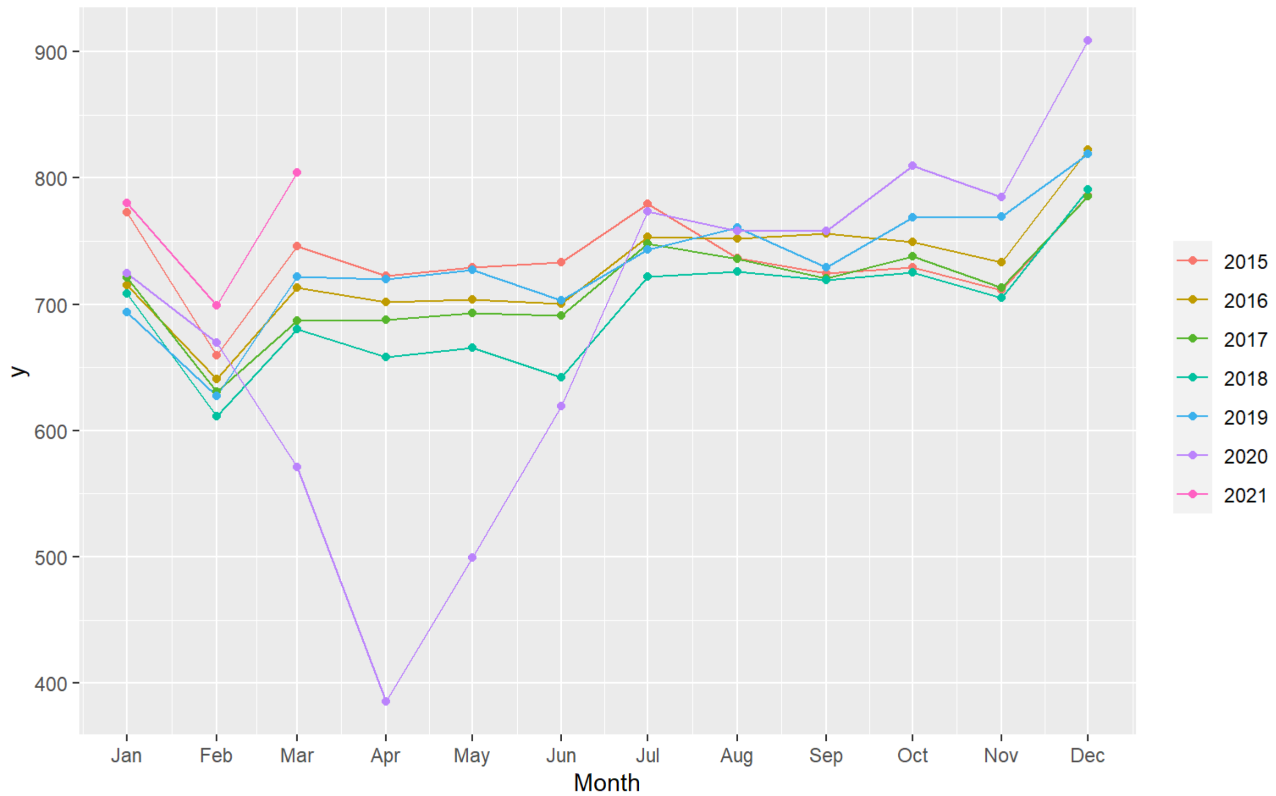
Hide

```
myts_full %>%
  filter(year(Month)>=2018) %>%
  autoplot(y) +
  geom_point()
```



Hide

```
myts_full %>%
  filter(year(Month)>=2015) %>%
  gg_season(y) +
  geom_point()
```



- Successfully updates series (2) (do not deduct marks if series was not updated by ABS).
- Plot data and visualise as necessary (at least time plot and season plot) (2)
- Comments (2marks)
- My series showed significant drop between March and April 2020 and then followed a rapid recovery.
- Full recovery seems to had been returned to by July.
- Bonus marks to whoever commented on lockdowns, etc.(2 marks)

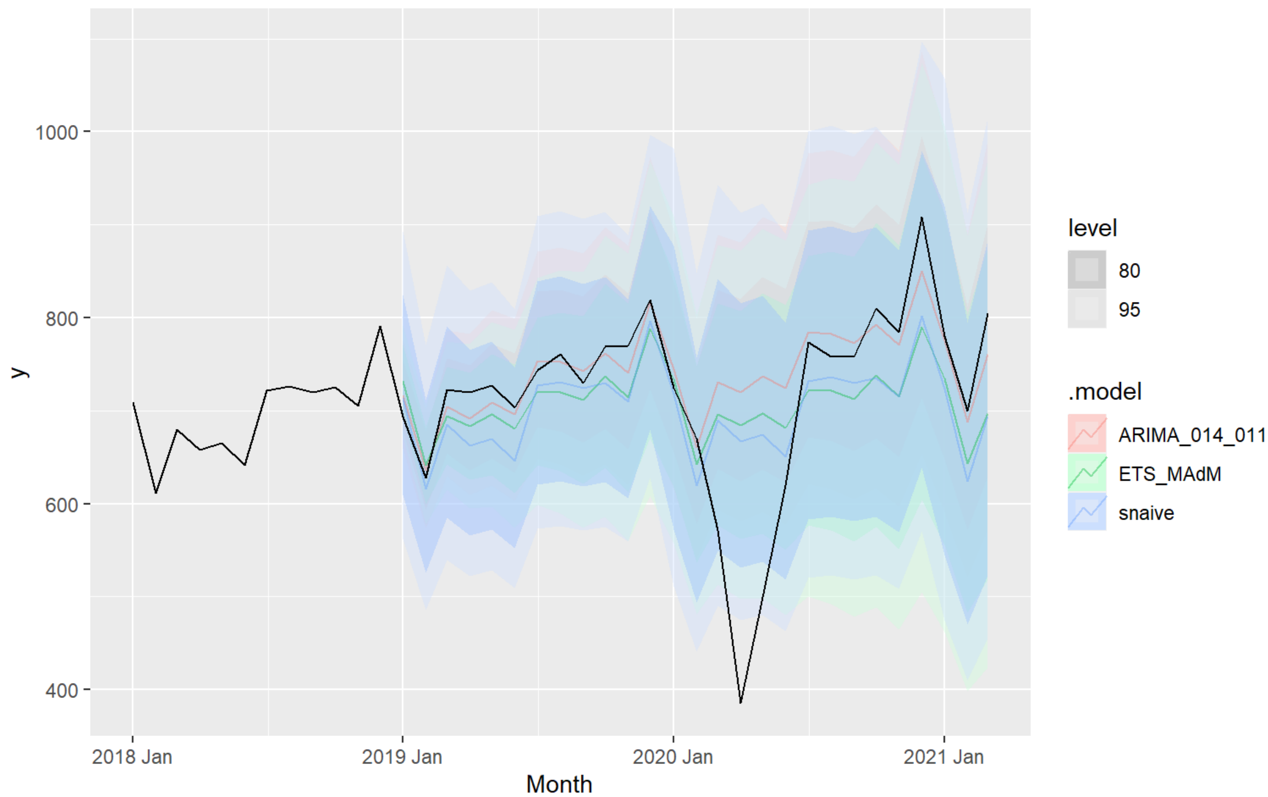
Question 9

> Generate forecasts for the period post 2019 until the end of the sample, from the models considered "best" in all assignments. More specifically, generate forecasts from the best benchmark, the best ETS and best ARIMA model. Plot the forecasts (both point and prediction intervals) together with the observed data and comment on these. (Make sure you can clearly visualise these. You may choose to plot on multiple graphs.) (6 marks)

Hide

```
models_final <- myts %>% model(
  snaive = SNAIVE(log(y)),
  ETS_MAdM=ETS(y~error("M")+trend("Ad")+season("M")),
  ARIMA_014_011=ARIMA(log(y)~pdq(0,1,4)+PDQ(0,1,1))
)

models_final %>%
  forecast(h=27) %>%
  autoplot(myts_full %>% filter(year(Month)>=2018), alpha=0.4)
```



- All three sets of forecasts seem to do extremely well.
- It seems that my series returns to natural/predicted levels after the effect of the few months due to COvid.

Marking guide

- generate forecasts from all three model (2 marks)
- plots and visualize (2 marks)
- Comment on what you see or what you would have expected to see (2)

Question 10

> Evaluate the accuracy of the point forecasts over the period post 2019. A table with accuracy measures will be necessary to be presented here. Comment on which forecasts are the most accurate. (4 marks)

Hide

```
models_final %>%
  forecast(h=27) %>%
  accuracy(myts_full) %>% select(.model,.type, ME, RMSE,MAPE, MASE, R
MSSE)
```

```
## # A tibble: 3 x 7
##   .model        .type     ME  RMSE  MAPE  MASE RMSSE
##   <chr>         <chr>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 ARIMA_014_011 Test  -25.4   89.4  8.58  1.77  2.48
## 2 ETS_MAdM      Test    7.51  87.9 10.4   2.42  2.44
## 3 snaive        Test   15.1   84.4 10.1   2.37  2.34
```

COMMENT

- ARIMA for me seems to do the best

Marking guide

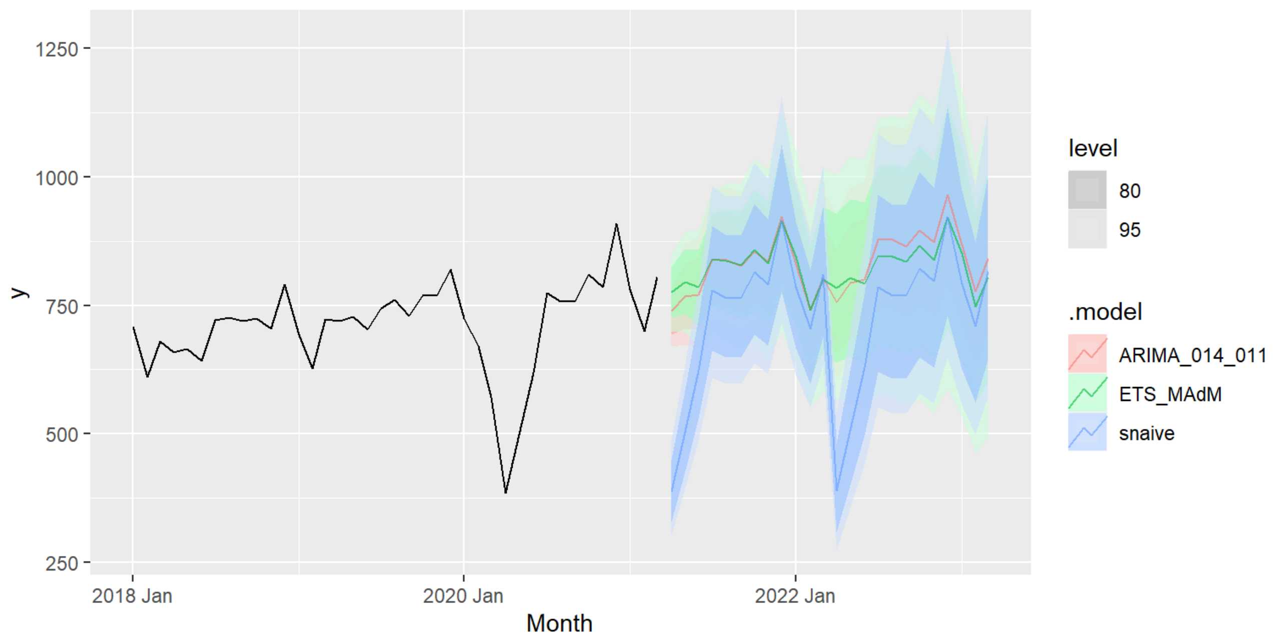
- Table (2)
- Commenting on the observed or expected differences. (2)

Question 11

> Use all three models to forecast the next 24-months of your updated series. Generate the necessary plots and comment on the forecasts. Make sure you can clearly visualise these. How have your models faired amidst the effects of COVID-19? (9 marks)

Hide

```
myts_full %>% model(
  snaive = SNAIVE(log(y)),
  ETS_MAdM=ETS(y~error("M")+trend("Ad")+season("M")),
  ARIMA_014_011=ARIMA(log(y)~pdq(0,1,4)+PDQ(0,1,1))
) %>% forecast(h=24) %>%
  autoplot(myts_full %>% filter(year(Month)>=2018), alpha=0.6)
```



COMMENT:

- the forecasts highlight the inappropriateness of the snaive projecting the effect of Covid-19 on the series

Marking guide

- generate forecasts from all three model (2 marks)
- plots and visualise (3 marks)
- comment on points forecasts for each model (3 marks)
- comment on forecast intervals (1 mark)