ETF3231/5231 Individual Assignment 3

Darren Luwi (ID 29051754)

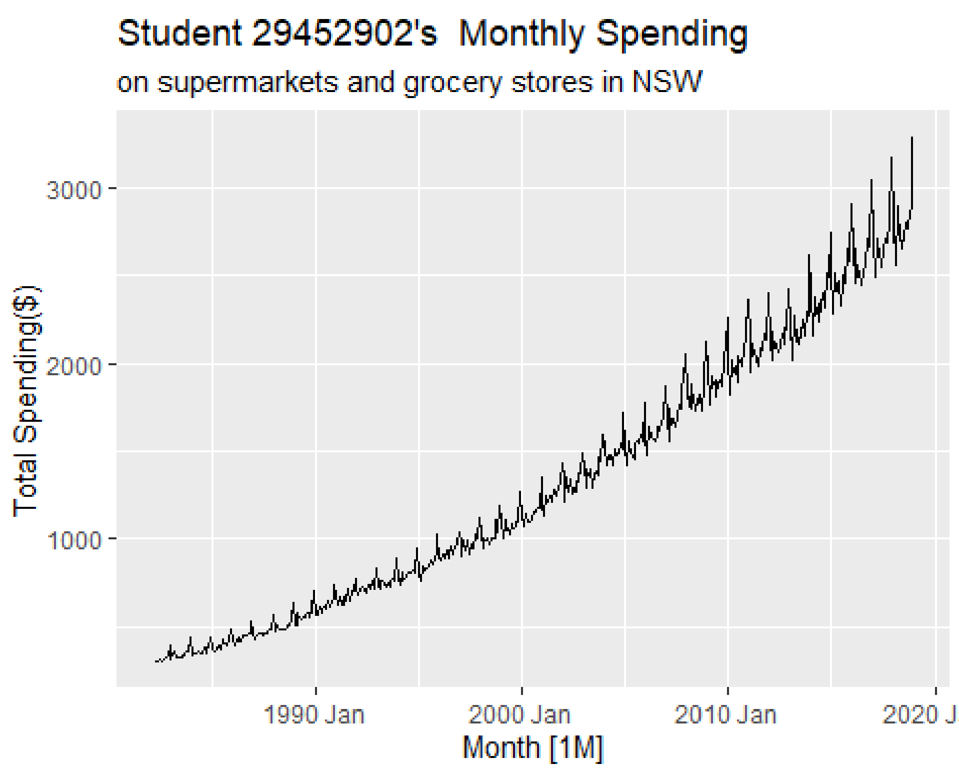
For the tasks that follow you will be modelling the **original data without any transformations** performed (even if you previously deemed it necessary).

```
# Read in and tidy up your data
# Make sure you select the column with your student ID

# First three rows contain metadata, read them in separately
meta <- read_csv("Undergrad_Data.csv", col_names = TRUE, n_max = 3)

##
## -- Column specification ------------------------------------------------
------
## cols(
##   .default = col_character()
## )
## i Use `spec()` for the full column specifications.

# meta
# The data follows after the third row, we skip the metadata and read the
data.
# Note: Skipping the first row skips the column names, we add them back from
the
#       metadata.
dat <- read_csv("Undergrad_Data.csv",
                # use column names from the metadata
                col_names = colnames(meta),
                # skip 4 rows as we also skip column names, specified above
                skip = 4,
                # The automatic column types correctly guess all columns but
the
                # date, we specify the date format manually here to correctly
                # get dates.
                col_types = cols("Student ID" = col_date("%b-%y")))

my_series <- dat %>%
  # feel free to rename your series appropriately
  rename(Month = "Student ID", y ="29452902") %>%
  select(Month, y) %>%
  mutate(Month=yearmonth(Month)) %>%
  as_tsibble(index = Month)
```

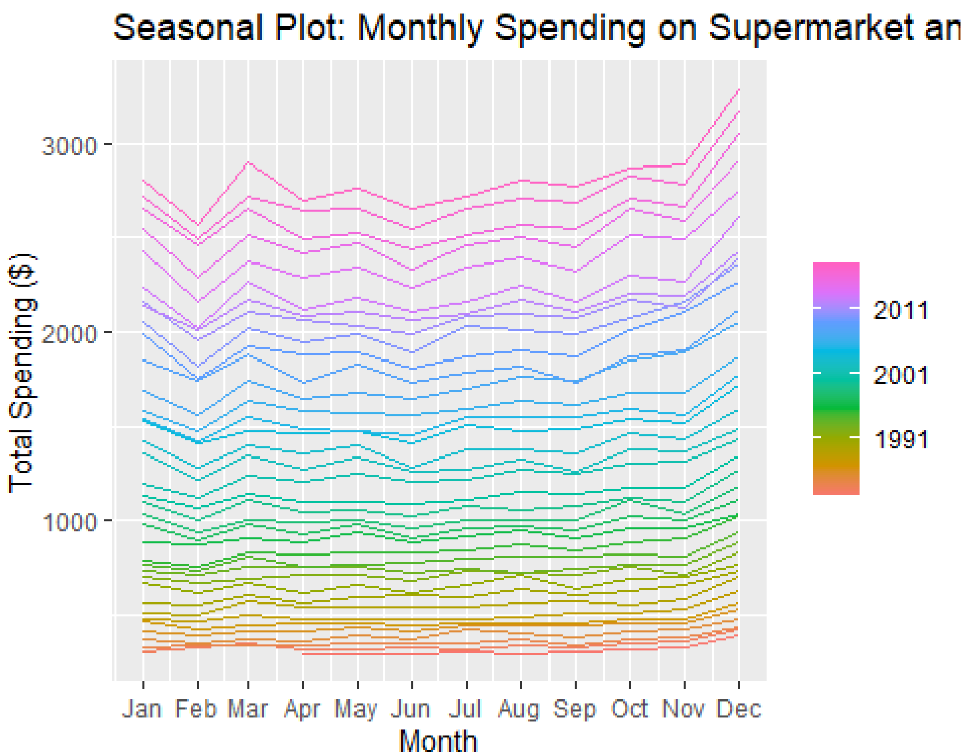
Question 1

Plot your time series. By observing the plot and describing its components select an ETS model you think is appropriate for forecasting. Make sure you justify your choice (no more than 150 words). (8 marks)

```
autoplot(my_series) + labs(title = "Student 29452902's  Monthly Spending",
subtitle = "on supermarkets and grocery stores in NSW", y= "Total
Spending($)")

## Plot variable not specified, automatically selected `.vars = y`
```




```
##There is a non-linear upward trend in spending.
##Next, notice that after some time, the magnitude of seasonality increases.
For instance, compare seasonality in 2005 and 2010, and see how the amplitude
of seasonality in 2010 is larger. This implies that a Multiplicative Method
might be suitable.
##Since Trend and Seasonality exists, a model which takes both into account
should be used.
gg_season(my_series) + labs(title = "Seasonal Plot: Monthly Spending on
Supermarket and Grocery Stores in NSW", y = "Total Spending ($)")

## Plot variable not specified, automatically selected `y = y`
```

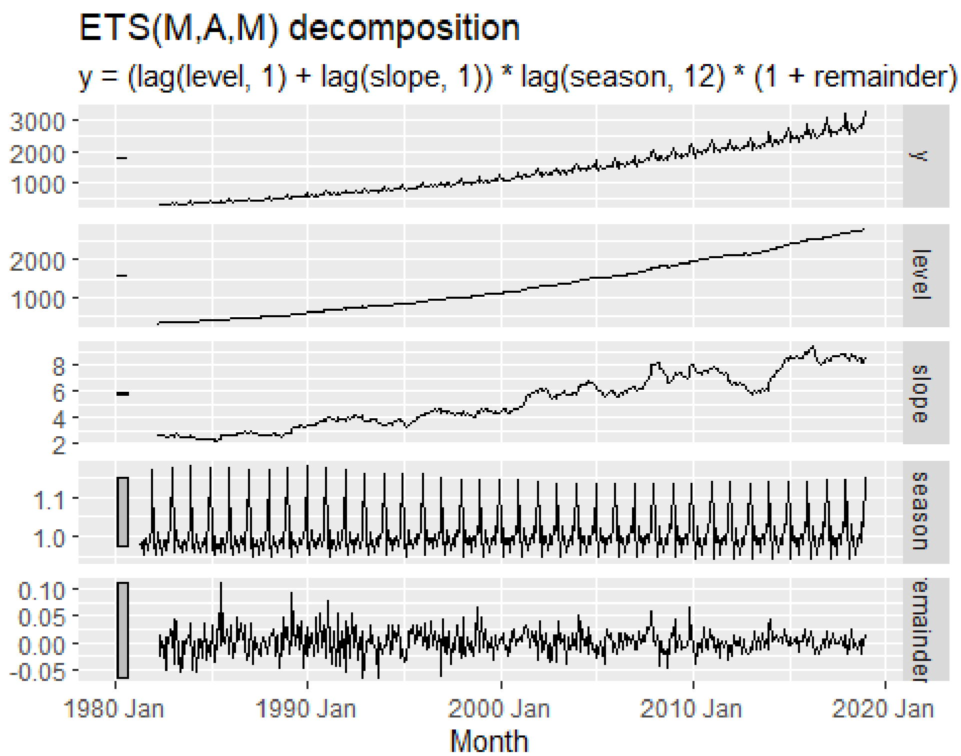



```
##The gg_season() method confirms that as level(Time) increases, amplitude of
seasonal activity also increases. Therefore we use the Holt-Winters
Multiplicative Method with Multiplicative Errors.
Ets <- my_series %>%
  model(
    MAM = ETS(y ~ error("M") + trend("A") + season("M"))
  )
```

Question 2

Estimate the ETS model you described in Question 1 and show the estimated model output. Describe and comment on the estimated parameters and components. Include any plots you see necessary (no more than 150 words). (14 marks)

```
Ets %>% tidy()

## # A tibble: 16 x 3
##    .model term   estimate
##    <chr>  <chr>     <dbl>
##  1 MAM    alpha   0.239
##  2 MAM    beta    0.00613
##  3 MAM    gamma   0.142
##  4 MAM    l     304.
##  5 MAM    b       2.66
##  6 MAM    s0      1.01
##  7 MAM    s1      0.953
##  8 MAM    s2      0.984
##  9 MAM    s3      1.17
## 10 MAM    s4      1.02
## 11 MAM    s5      1.00
## 12 MAM    s6      0.962
## 13 MAM    s7      0.996
## 14 MAM    s8      0.990
## 15 MAM    s9      0.955
## 16 MAM    s10     0.984

##alpha value is 0.239, suggesting that "weight" given to each observation
are relatively uniform.
##The extremely small beta value (0.006) means trend is increasing slowly
over time.
##Concurrently, the gamma value of 0.142 implies that more recent seasonal
components are weighted less than the older ones.
##Finally, the initial Level, Trend and Seasonal States were also given
(303.863,2.659 and 1.008 respectively).
components(Ets) %>% autoplot()

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

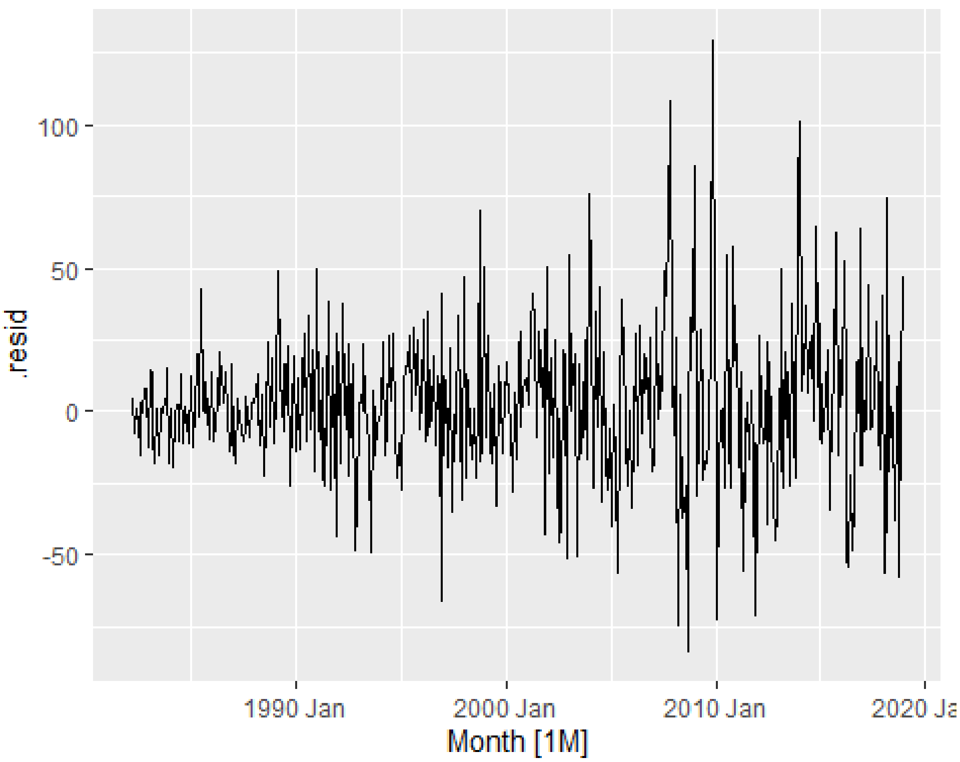



```
##Level and Slope follows the data fairly closely.
##From the slope graph, we can see how trend increases at a relatively slow
rate. We can also see regular troughs across different levels, not a
persistent upward movement (where in this case the beta value will be large).
##Finally, notice how the season plot becomes smaller in size as level
increases. This is due to the less weight assigned to the more recent
observations (and hence supporting my hypothesis in terms of gamma).
```

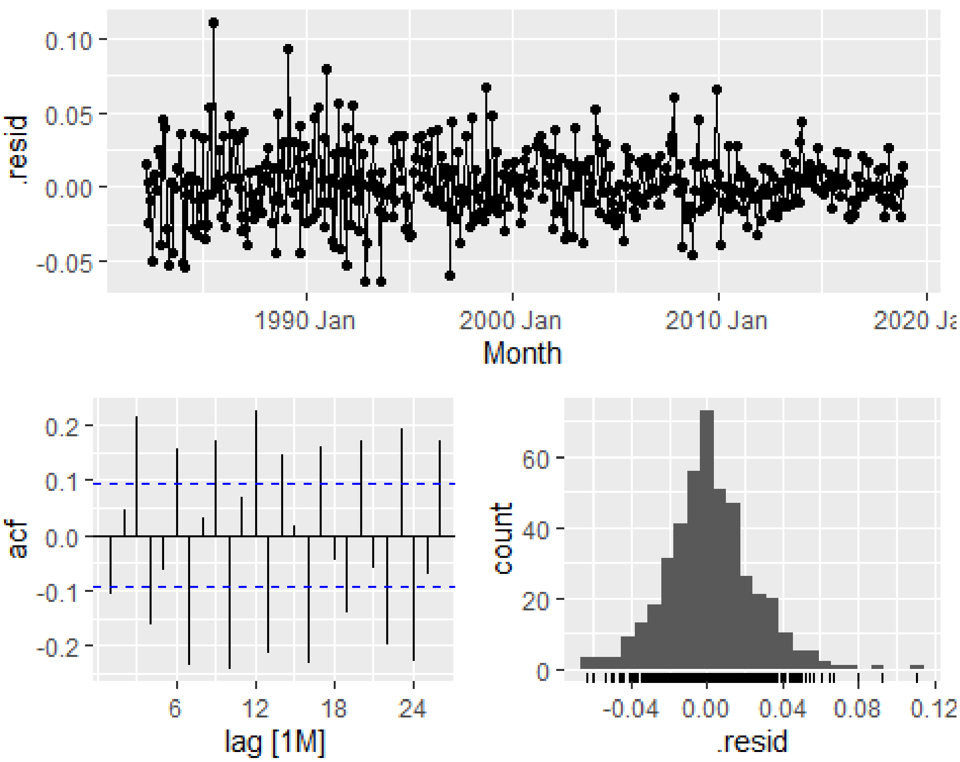
Question 3

Plot the residuals from the model and comment on these (no more than 50 words). Perform some diagnostic checks and comment on whether you are satisfied with the fit of the model. (Make sure you state all relevant information for any hypothesis test you perform such as the null hypothesis, the degrees of freedom, the decision, etc.). (10 marks)

```
residets <- residuals(Ets, type = "response")
autoplot(residets)

## Plot variable not specified, automatically selected `.vars = .resid`
```




```
##distribution seems non-normal.
Ets %>% gg_tsresiduals()
```




```
##lag = 26
##However residuals distribution here looks normal.
Ets %>%
  augment() %>%
  features(.innov, ljung_box, dof = 16, lag = 26)

## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 MAM       311.         0

##Null = No lack of fit
##dof = no. of variables from tidy() = 11
##Test returns lb_stat of 311.156 and p-value of 0.
##Do not reject null = no lack of fit
##Model fits well = satisfied.
```

Question 4

Let R select an ETS model. What model has been chosen and how has this model been chosen? (No more than 100 words). (6 marks)

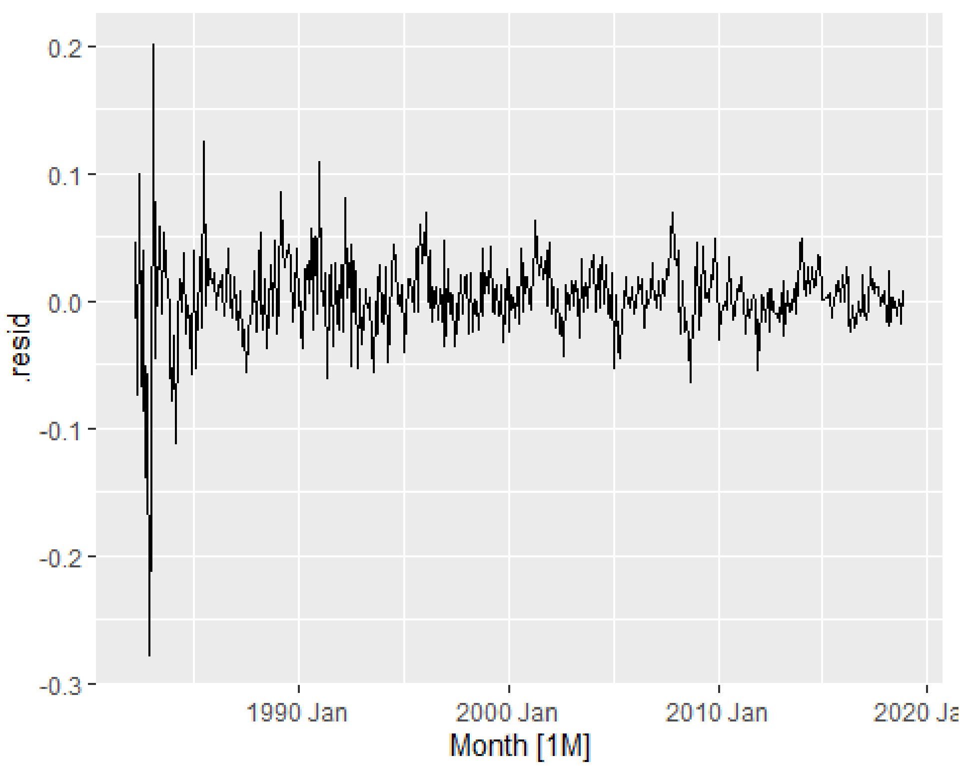
```
rmodel <- my_series %>%
  model(ETS(y))
##I assume that R takes into account the existince of an upward trend in the data.
##In addition, I believe R noticed how seasonal activity increases along with level.
##Intuitively I am sure that R used the same ETS model as me (ETS(MAM)).
##To confirm:
rmodel %>% tidy()

## # A tibble: 16 x 3
##    .model term   estimate
##    <chr>  <chr>     <dbl>
##  1 ETS(y) alpha   0.239
##  2 ETS(y) beta    0.00613
##  3 ETS(y) gamma   0.142
##  4 ETS(y) l     304.
##  5 ETS(y) b       2.66
##  6 ETS(y) s0      1.01
##  7 ETS(y) s1      0.953
##  8 ETS(y) s2      0.984
##  9 ETS(y) s3      1.17
## 10 ETS(y) s4      1.02
## 11 ETS(y) s5      1.00
## 12 ETS(y) s6      0.962
## 13 ETS(y) s7      0.996
## 14 ETS(y) s8      0.990
## 15 ETS(y) s9      0.955
## 16 ETS(y) s10     0.984

##Notice that the coefficients of each smoothing parameters and initial states are exactly the same as my model's.
##Therefore, the model R selected is also the Holt-Winters Multiplicative Method with Multiplicative Errors.
```

Question 5

Comment on how the model chosen by R is different to the model you have specified (no more than 50 words). Which of the two models would you choose and why? (no more than 50 words). (Hint: think about model selection here but also check your residuals).

If the models are identical specify a plausible alternative. Give a brief justification for your choice (no more than 100 words). (Hint: also check the residuals from this model). (12 marks)

```
##Alternatively, based on the time series plotted in question 1, notice that
within certain levels, seasonal magnitudes are identical (constant). For
instance, we see identical patterns from 1990 to 1995, and another identical
group from 2000-2005. This suggests that an Additive Trend + Additive
Seasonality Method might be ideal.
altmodel <- my_series %>%
  model(ETS(y~ error("M") + trend("A") + season("A")))
residuals(altmodel, type = "Response") %>%
  autoplot()

## Plot variable not specified, automatically selected `.vars = .resid`
```



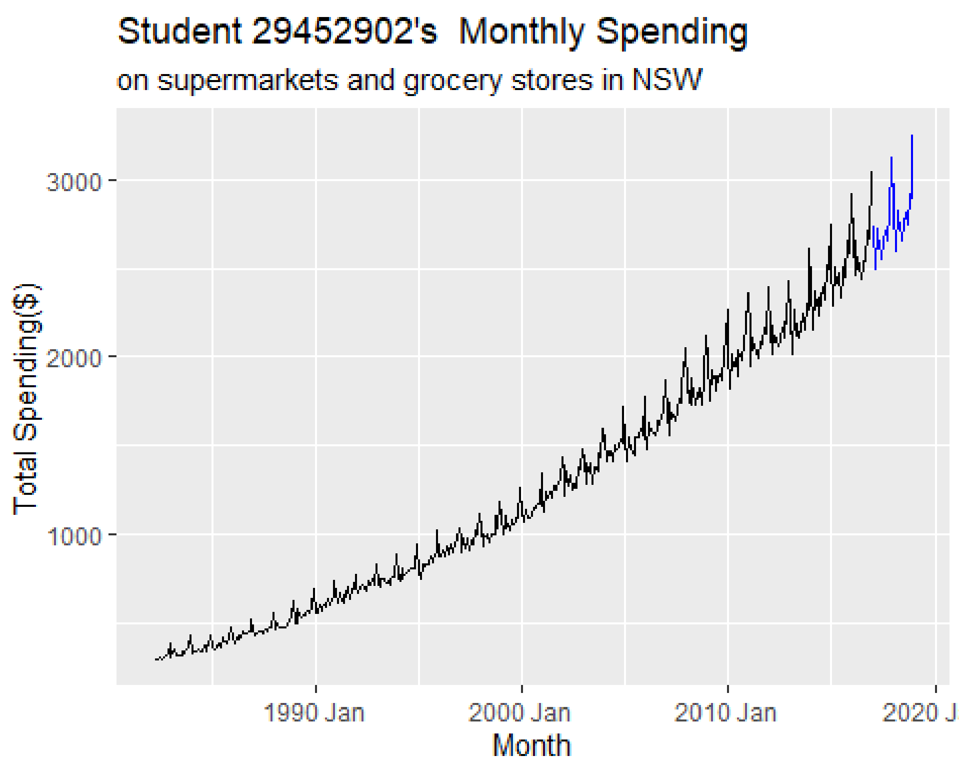

```
##Residuals distribution seems normal (with some outliers).
##To confirm:
altmodel %>%
  augment() %>%
  features(.innov, ljung_box, dof = 16, lag = 26)
```

```
## # A tibble: 1 x 3
##   .model                                                  lb_stat lb_pvalue
##   <chr>                                                     <dbl>     <dbl>
## 1 "ETS(y ~ error(\"M\") + trend(\"A\") + season(\"A\"))"    119.          0

##Null = No lack of fit
##Test suggests lb_stat of 119.129 and p-value of 0.
##We do not reject null, and assume a good fit.
##However it is hard to tell which model fits better.
```

Question 6

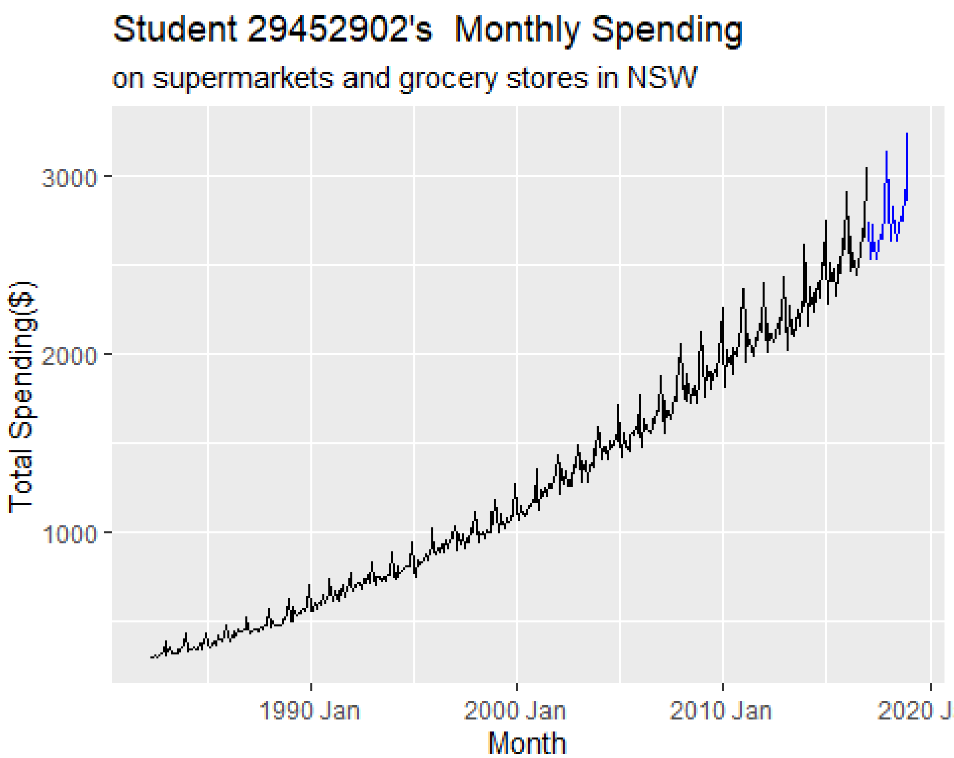
Generate forecasts for the last two years of your sample using both alternative ETS models (you will need to re-estimate both models over the appropriate training sample). Plot the forecasts and forecast intervals. Briefly comment on these. Which model does best? (No more than 100 words). (8 marks)

```
train <- my_series %>%
  filter(year(Month) <= 2016)
train2 <- my_series %>%
  filter(year(Month) > 2010)

benchmark1 <- train %>%
  model(
    MAM = ETS(y ~ error("M") + trend("A") + season("M"))
  )
benchmark_fc1 <- benchmark1 %>%
  forecast(h = 24)
benchmark_fc1%>% autoplot(train, level = NULL) + labs(title = "Student
29452902's  Monthly Spending", subtitle = "on supermarkets and grocery stores
in NSW", y= "Total Spending($)")
```

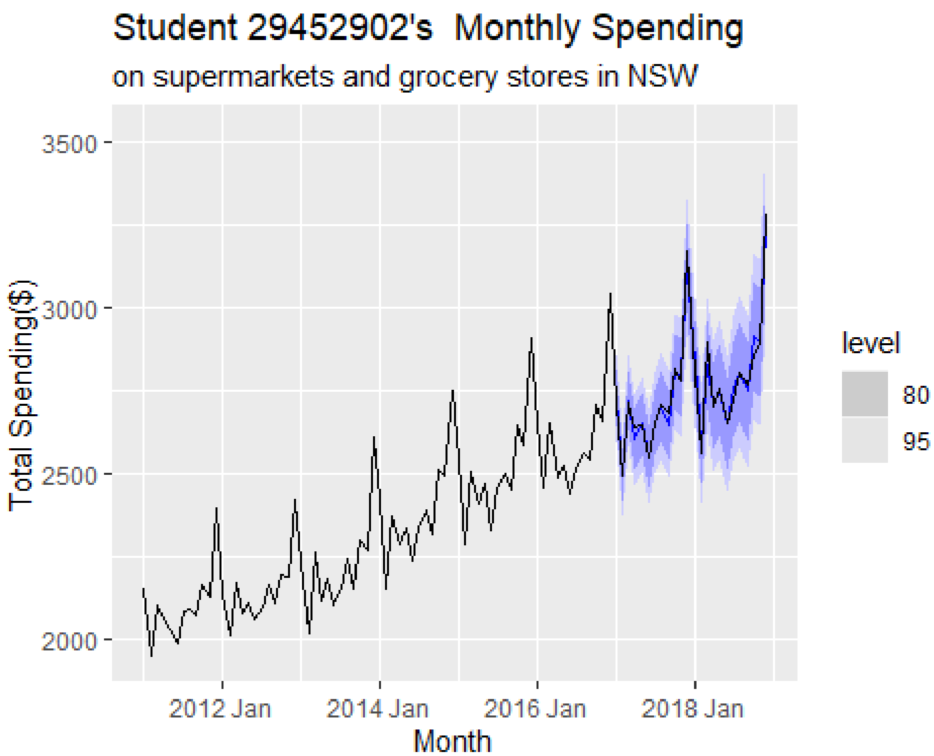



```
benchmark2 <- train %>%
  model(
    MAA = ETS(y ~ error("M") + trend("A") + season("A"))
  )
```

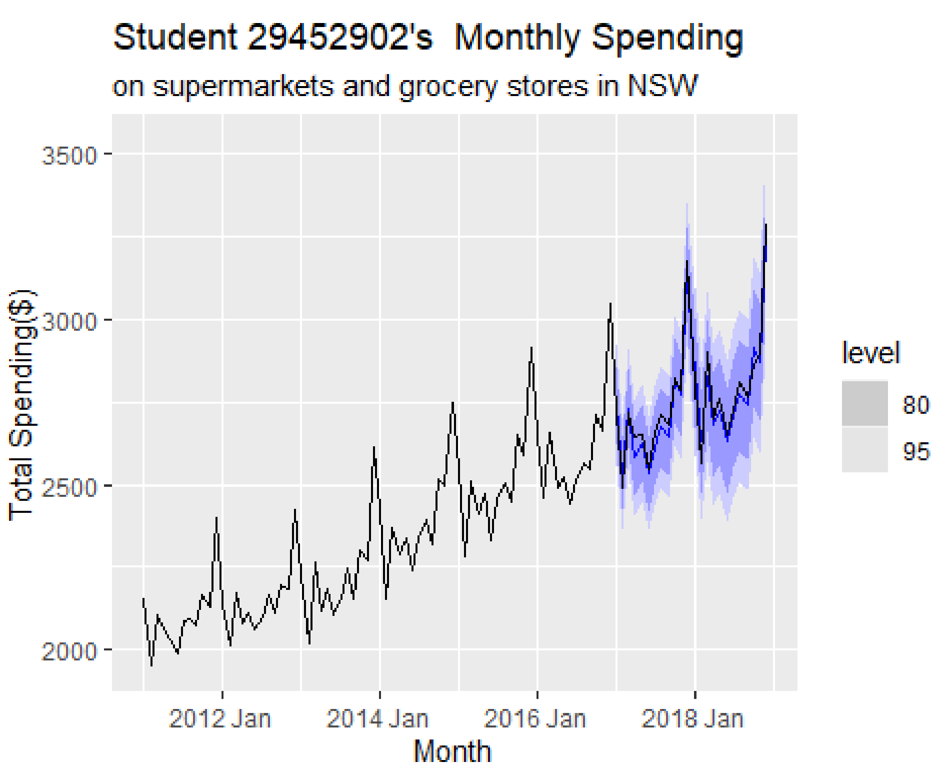
```
benchmark_fc2 <- benchmark2 %>%
  forecast(h = 24)
benchmark_fc2%>% autoplot(train, level = NULL) + labs(title = "Student 
29452902's  Monthly Spending", subtitle = "on supermarkets and grocery stores 
in NSW", y= "Total Spending($)")
```



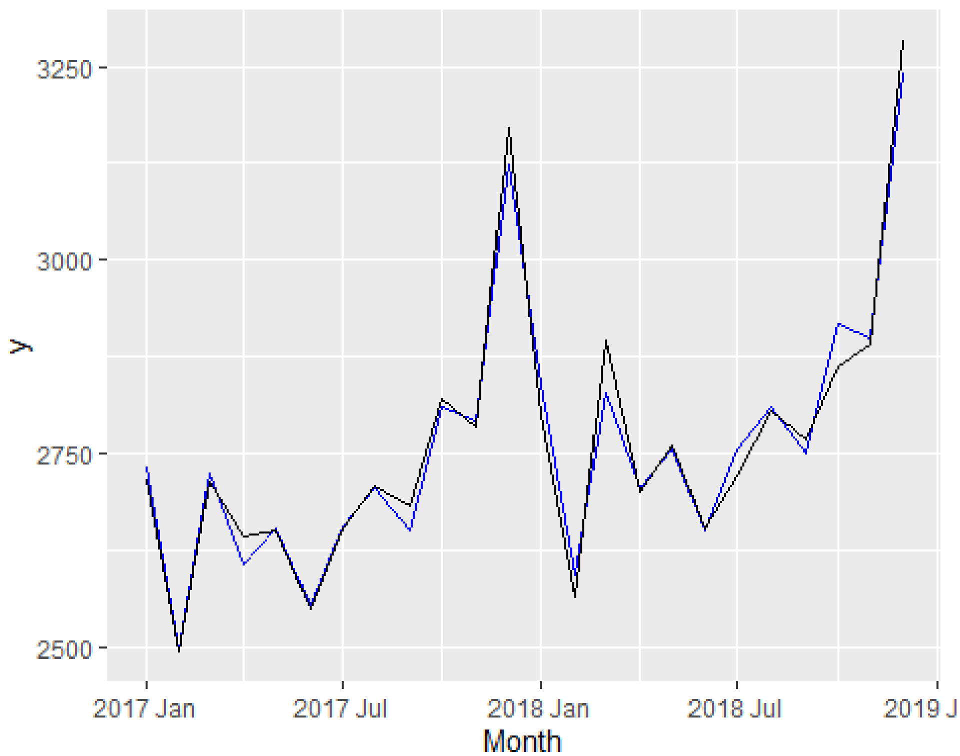

```
##Both forecasts are identical in terms of trend and seasonality.

benchmark_fc1%>% autoplot(train2) + labs(title = "Student 29452902's  Monthly 
Spending", subtitle = "on supermarkets and grocery stores in NSW", y= "Total 
Spending($)")
```



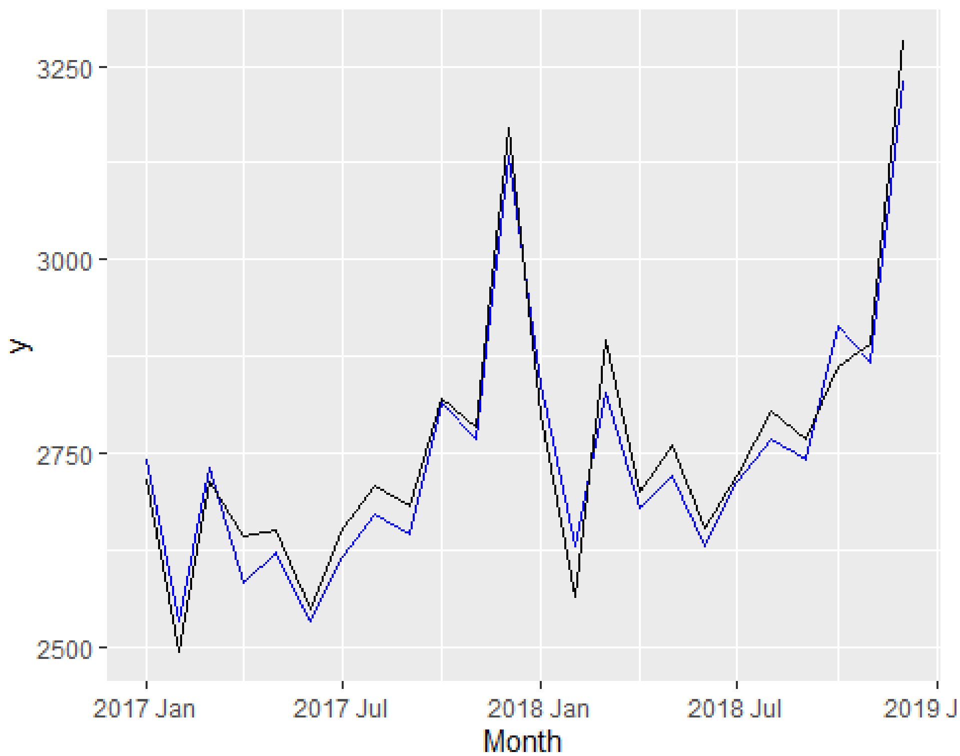

```
benchmark_fc2%>% autoplot(train2) + labs(title = "Student 29452902's  Monthly
Spending", subtitle = "on supermarkets and grocery stores in NSW", y= "Total
Spending($)")
```



```r
##ETS(MAA) have slightly wider forecast intervals.
##We can compare both forecasts against the test set:
test <- my_series %>%
  filter(year(Month) > 2016)
benchmark_fc1 %>% autoplot(test, level = NULL)
```




```r
benchmark_fc2 %>% autoplot(test, level = NULL)
```

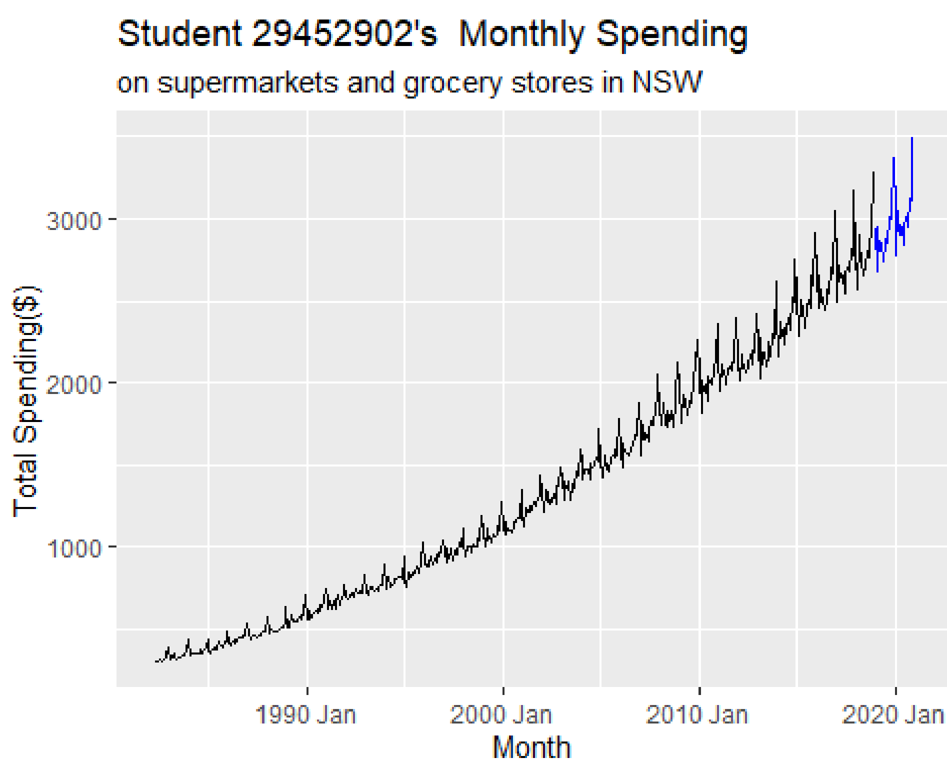



```
##Both forecasts fit the test set well, although ETS(MAM) fits slightly
better.
##Finally we check their accuracies:
my_series%>%
  filter(year(Month) > 2016) %>%
  model(
    MAM = ETS(y ~ error("M") + trend("A") + season("M")),
    MAA = ETS(y ~ error("M") + trend("A") + season("A"))
  )%>%
  accuracy()

## # A tibble: 2 x 10
##   .model .type        ME  RMSE   MAE      MPE  MAPE  MASE RMSSE    ACF1
##   <chr>  <chr>     <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 MAM    Training  0.340  30.4  24.4 -0.00574 0.884 0.261 0.305 -0.164
## 2 MAA    Training 11.4    65.9  30.0  0.402   1.09  0.321 0.661 -0.0852

##ETS(MAM) have lower RMSE and MAE values, hence it is more accurate.
##Therefore, ETS(MAM) is the better model.
```

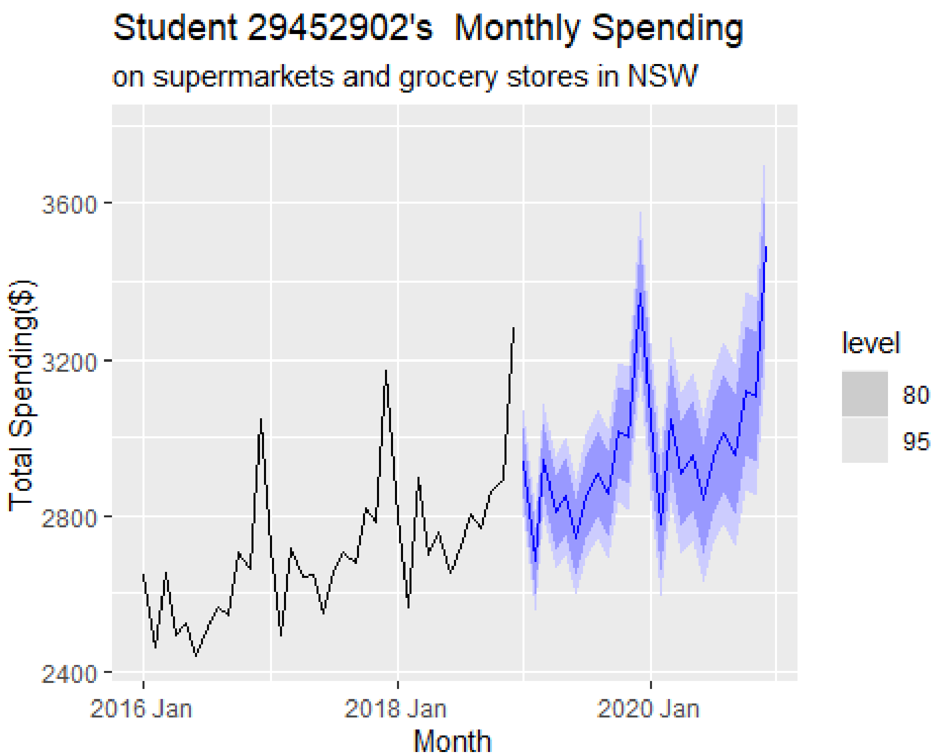
Question 7

Generate forecasts for the two years following the end of your sample using your chosen model. Plot them and briefly comment on these. (No more than 100 words). (4 marks)

```
fit2 <- my_series %>%
  model(ETS(y ~ error("M") + trend("A") + season("M")))
fc <- fit2 %>%
  forecast(h = 24)
fc %>% autoplot(my_series, level = NULL) + labs(title = "Student 29452902's
Monthly Spending", subtitle = "on supermarkets and grocery stores in NSW", y=
"Total Spending($)")
```




```
##The student's spending pattern is expected to behave similarly to its
predecessors (again with existence of seasonal variations and an increasing
trend).
##This is reasonable as groceries (food and water) are our basic necessities,
and spending will be affected by seasonal factors like Christmas and Easter.
##Meanwhile, it is likely that the student will get married/have more family
members in the future, explaining the upward trend in monthly spending.
forecastinterval <- my_series %>%
  filter(year(Month) > 2015)
fc %>% autoplot(forecastinterval) + labs(title = "Student 29452902's  Monthly
Spending", subtitle = "on supermarkets and grocery stores in NSW", y= "Total
Spending($)")
```




```
##Forecast Intervals are relatively wide.
```