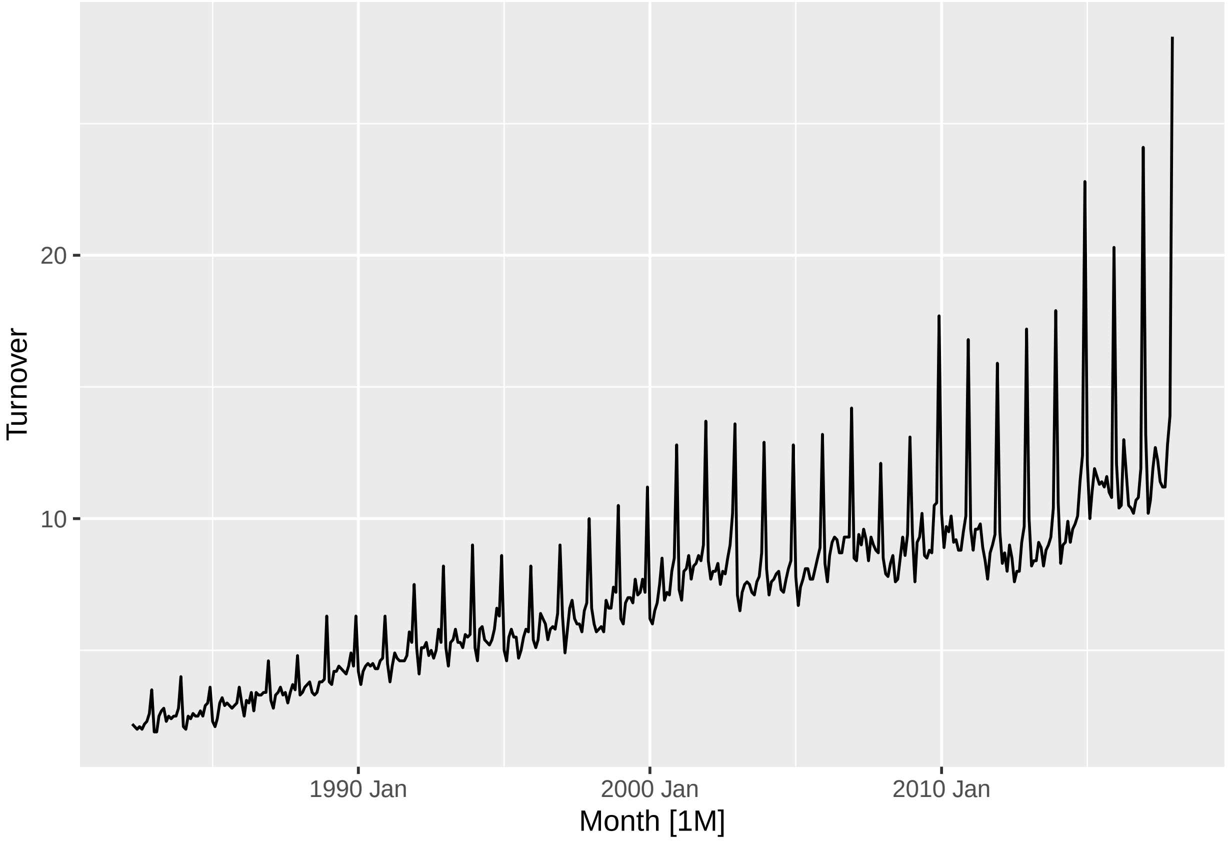
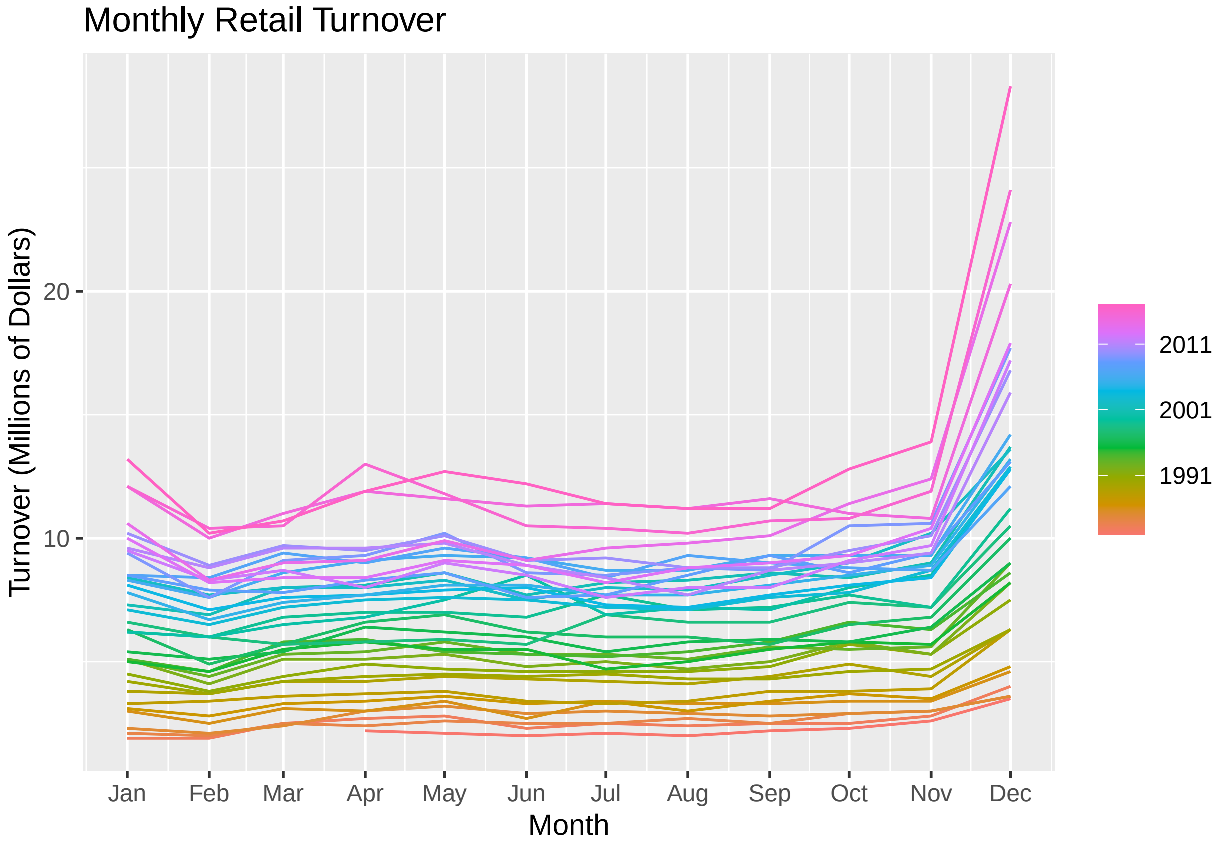
Forecast Analysis - Australian Retail Turnover

Sivaram Ainkaran

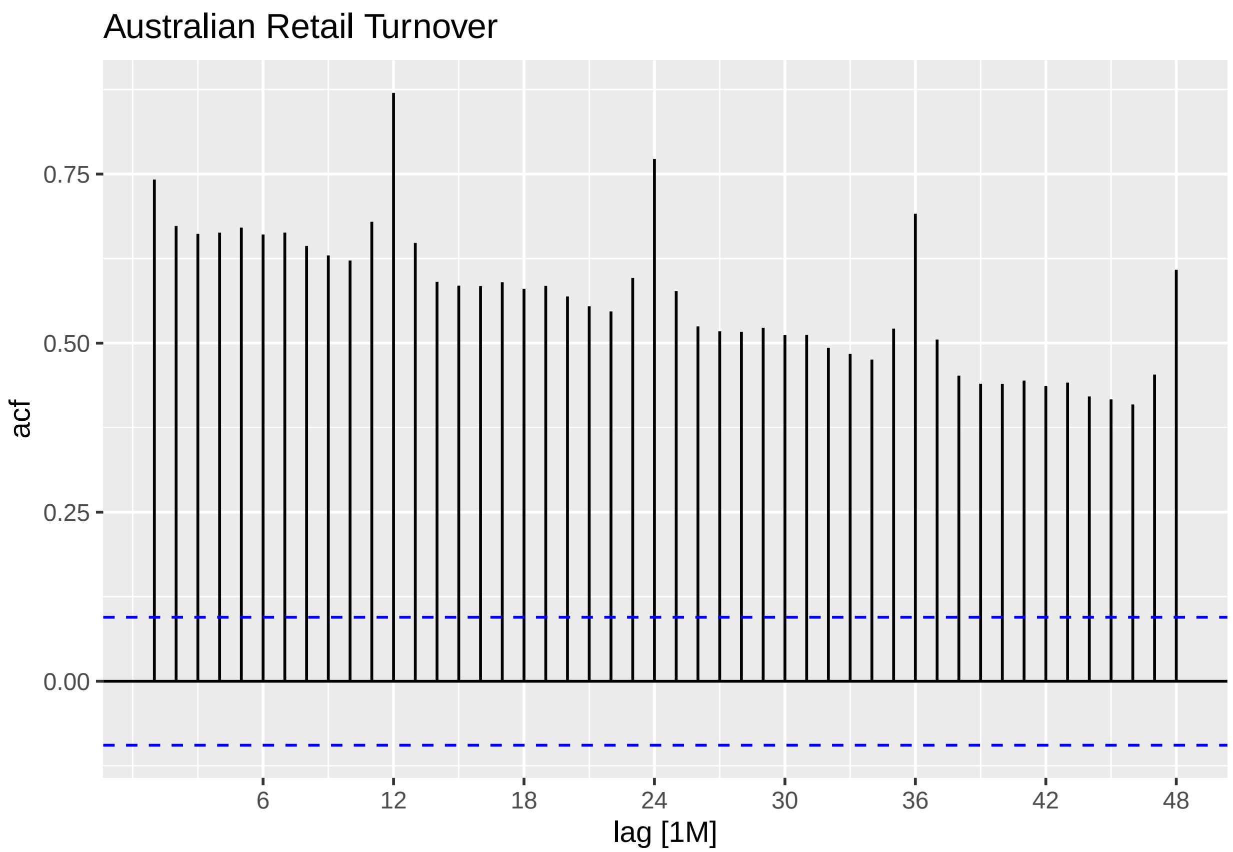
Australian Retail is a highly seasonal business. As we tend towards the end of the year, retail turnover tends to increase, and due to the increasing amount of promotion, stores and people willing to buy, these sales have been steadily increasing over the last few decades. A glimpse of this increasing trend and seasonality can be seen in the figure below which shows Australian retail turnover from 1982 to 2017.



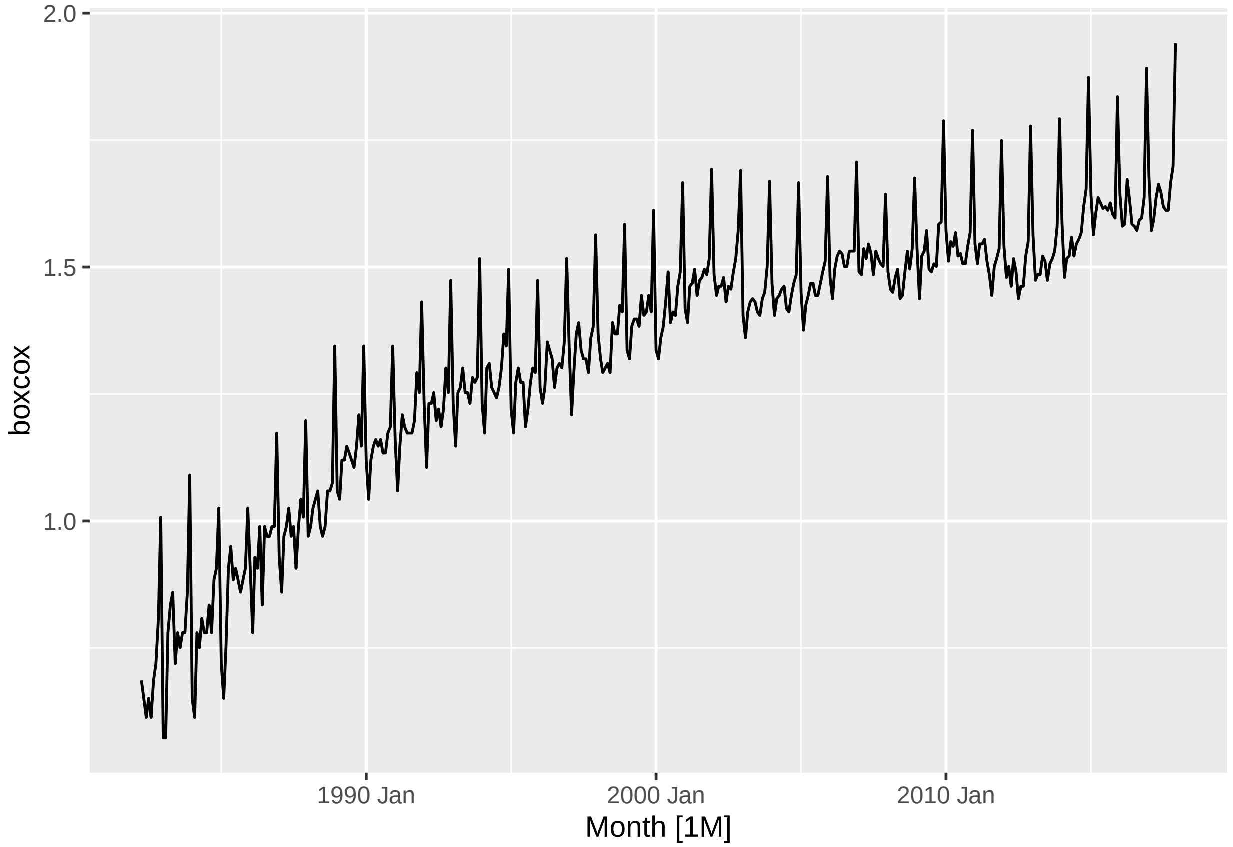
There is very strong seasonality in this data. The figure below shows a rather stable year, with slight troughs, especially in February, possibly due to families recovering from Christmas and New Years shopping. There tend to be small peaks in May, most likely due to Mother's Day and enormous peaks follow, from November to January, during Christmas time shopping.



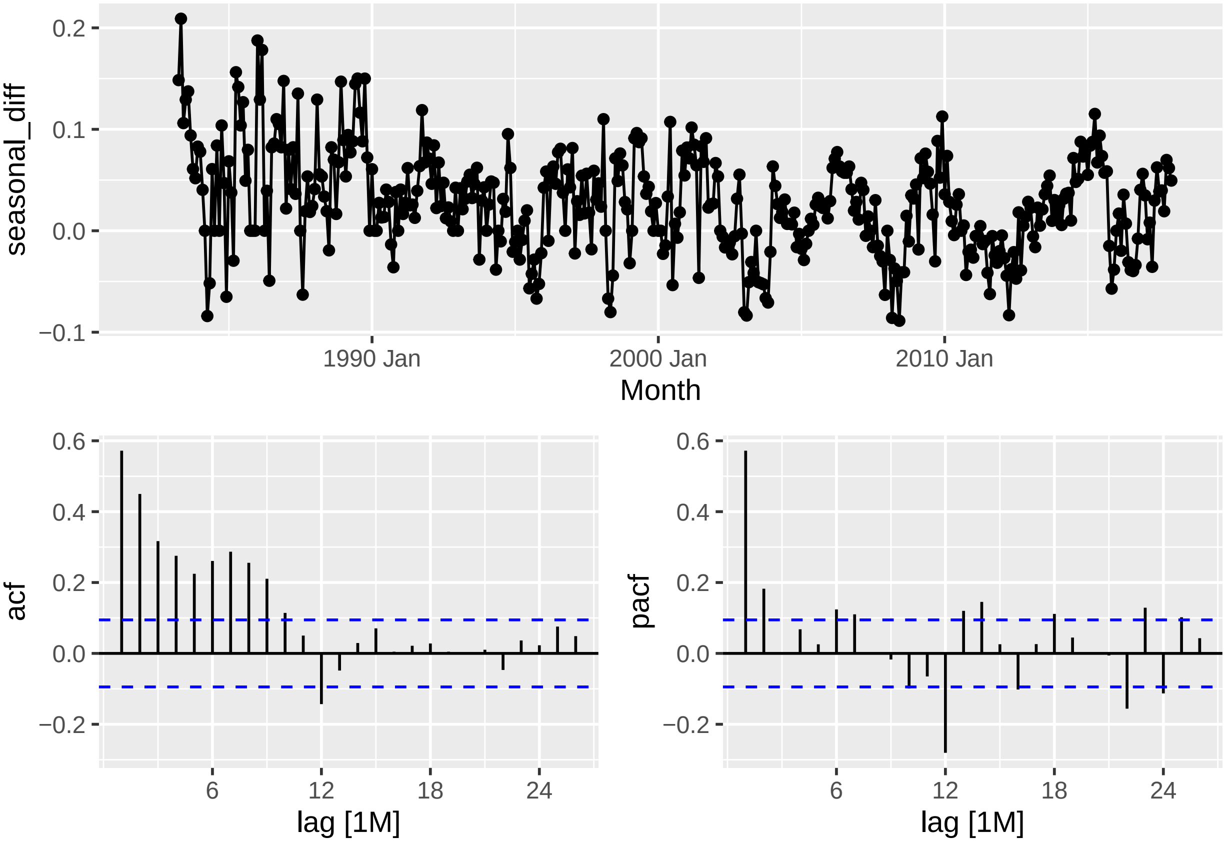
As could be seen in the original plot of all the data, there does seem to be a positive trend in the data over time. This ACF plot below shows the seasonality of the data, through larger spikes in acf every 12 months but it also shows that there is a trend in the data as the ACf plot is decreasing over time. This trend is clearly positive and has been greatly increasing over the last few years especially. This plot does also show the slight increase in retail turnover in the middle of the year as stated before as it has a slight "W" shape between every 12 months.



this Australia retail turnover data has a very positive trend with variation which has been increasing over each consequent year. This can be simply fixed with a logarithmic transformation such as a Box-Cox transformation. This will allow the seasonal variance to be the same or much more similar over all the years of data. Using the guerrero feature, we can choose an optimal lambda for the Box-Cox transformation for this data set which maximally reduces variation in the data.



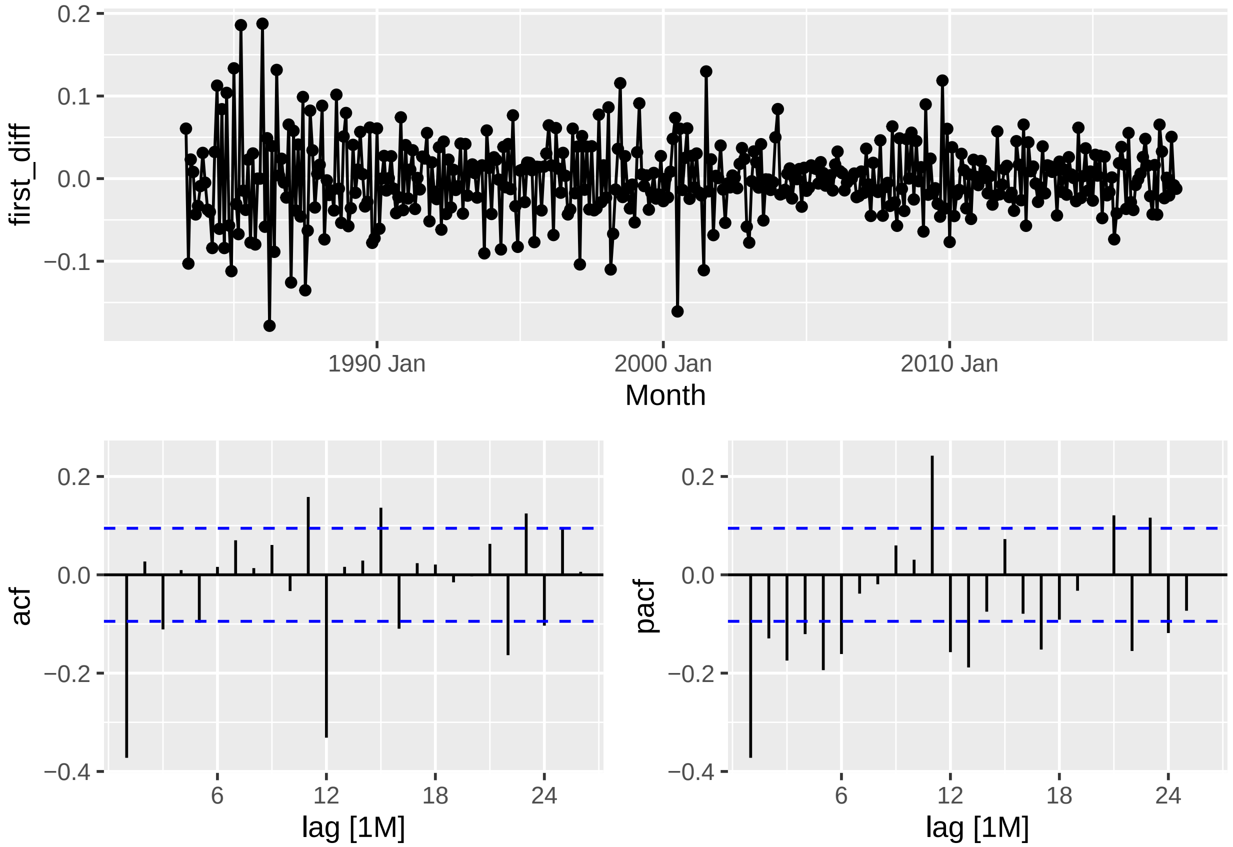
As can be seen from this Box-Cox plot, with lambda= - 0.3611761, the variation in seasonal differences over the years has been greatly reduced, making the data more easy to work with. This data however is still trending upwards with an initial curve. Using differencing we can stabilise this to a more steady mean. By simply taking the difference between a value and the previous value from the same season, we are able to produce seasonally differenced data.



Starting with a Seasonal difference, by simply looking at the differenced data above, we can see that the data is rather stationary, being random around a mean of 0. Looking at the ACF graph we can see that it decays over the first 12 lags rather quickly, showing stationarity. The PACF plot also does have spikes at the first and twelfth lags, but they are barely exceeding 0.5 and not close to 1 at all. Although these features all show a relatively stationary dataset doing a unit root test will be ideal in measuring this stationarity more clearly.

```
## # A tibble: 1 x 4
##   State                        Industry                     kpss_stat kpss_pvalue
##   <chr>                        <chr>                            <dbl>       <dbl>
## 1 Australian Capital Territory Footwear and other persona~       1.46        0.01
```

As we can see from the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit-root test results from above, although the p-value is reported as 0.01, the test statistice (1.46) is greater than 1 (or the 1% critical value). This means the p-value is actually less than 0.01 and we have to reject the null hypothesis that this seasonally differenced data is stationary. This indicates that we must take a first difference to obtain stationary data.



By taking the first difference of the seasonally differenced data, we can see above that acf values decay even quicker, with spikes every twelve months which also decay rather quickly. The first lag of the pacf plot is also below 0.4, and not nearing 1, indicating stationarity already. If we check the unit root test below, we can see that the p-value is 0.1 and the test statistic is 0.0307 which is much less than the 1% critical value. This indicates we can accept the null hypothesis that this seasonally and first differenced data is now stationary.

```
## # A tibble: 1 x 4
##   State                        Industry                        kpss_stat kpss_pvalue
##   <chr>                        <chr>                               <dbl>       <dbl>
## 1 Australian Capital Territory Footwear and other persona~        0.0307         0.1
```

Using the ACF and PACF plots above, we can determine the seasonal and non-seasonal components of the ARIMA model this data would follow. First of all, this data was seasonally difference and first differenced once, indicating that both d=1 and D=1. By looking at the ACF graph above, we can also see that the last significant lag is at the 3rd lag (-0.111, while 5th lag is -0.0974). This suggests a non-seasonal MA(3) component, ie. q=3. This ACf plot also shows the last significant seasonal lag at lag 24. This suggests a seasonal MA(2) component, ie. Q=2. If we look at the PACF plot, we can see that there is a very strong first lag compared to all other lags which suggests a non-seasonal AR(1) component, ie. p=1. We can also suggest p=5 since there is a larger, significant spike in the PACF there as well. The PACF also shows a significant twelfth lag and a very slightly significant 24th lag which would indicate a seasonal MA(2) component. These all point to the following models: ARIMA(p=1,d=1,q=0)(P=2,D=1,Q=0)[12] ARIMA(p=5,d=1,q=0)(P=2,D=1,Q=0)[12] ARIMA(p=0,d=1,q=3)(P=0,D=1,Q=2)[12]

```
## # A tibble: 3 x 6
##   .model       sigma2 log_lik    AIC   AICc    BIC
##   <chr>         <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
## 1 arima013012 0.00123    733. -1453. -1453. -1429.
## 2 arima510210 0.00313    581. -1145. -1145. -1113.
## 3 arima110210 0.00459    502.  -997.  -997.  -981.
```

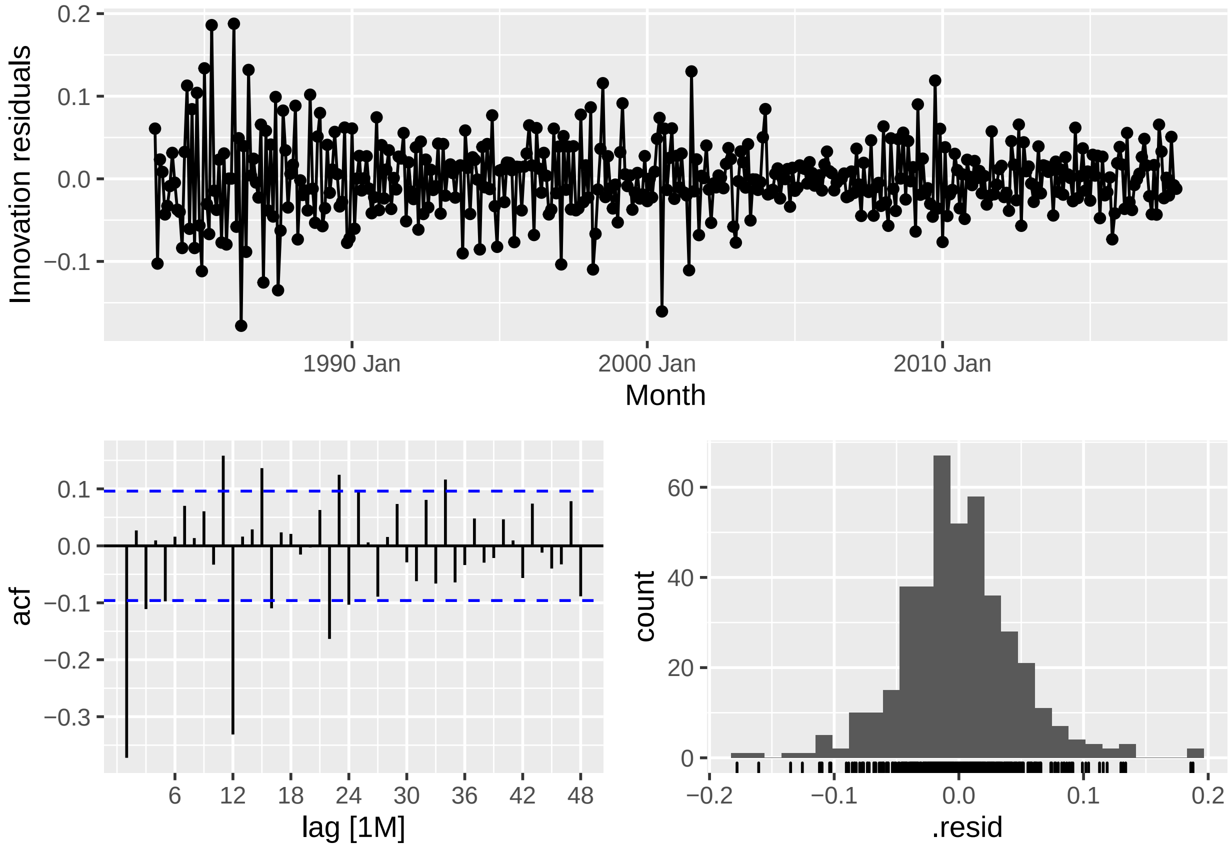
If we look at the AICc values of each of these ARIMA models, shown above, we can see that the ARIMA(p=0,d=1,q=3)(P=0,D=1,Q=2)[12] model fits our transformed data the best as it has the lowest AICc value. Following this, we have to decide the best ETS models. Since there is no real trend component or seasonality to this modified data, we can focus on the errors. Since there are only multiplicative or additive errors we can test the following 2 models: ETS(M,N,N) and ETS(A,N,N)

```
## Series: first_diff
## Model: ETS(M,N,N)
##   Smoothing parameters:
##     alpha = 0.1894753
##
##   Initial states:
##           l
##  0.02457382
##
##   sigma^2:  1870.09
##
##      AIC     AICc      BIC
## 1094.485 1094.543 1106.577
```

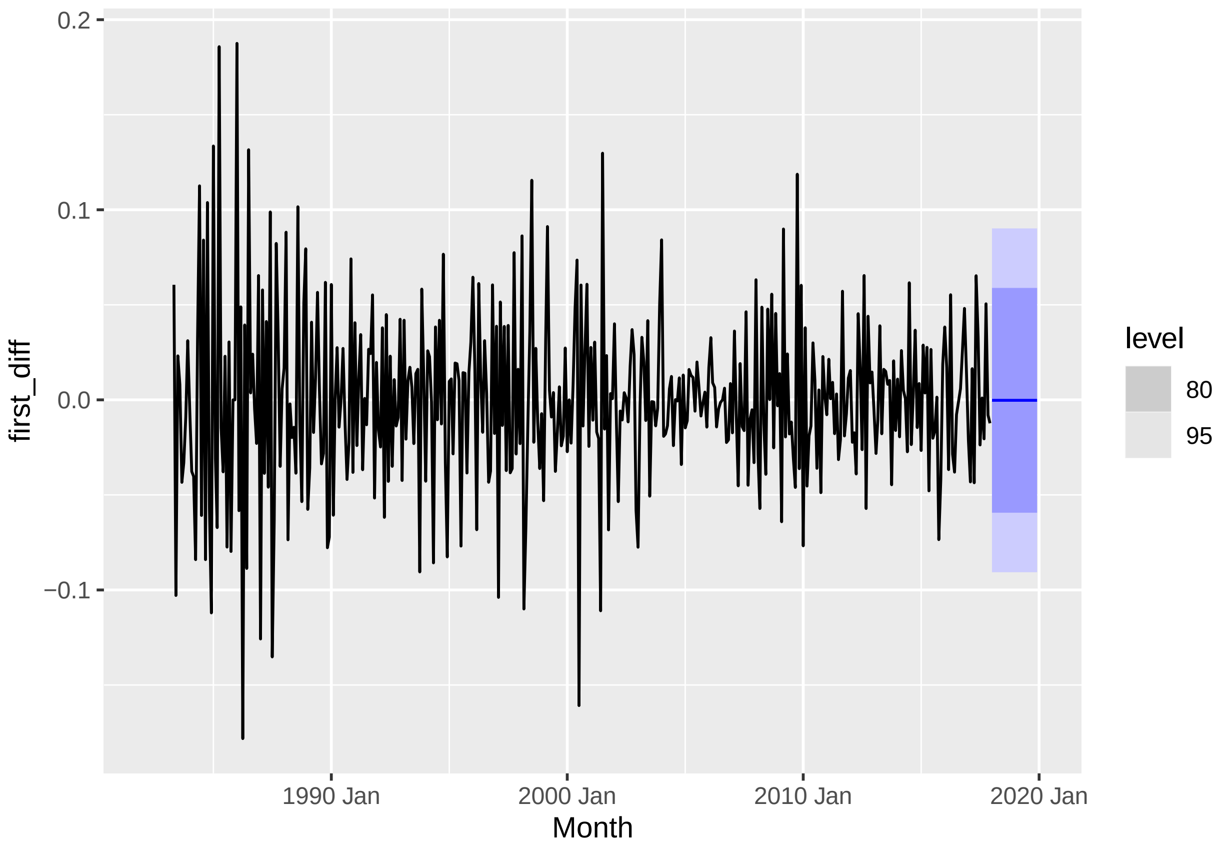
```
## Series: first_diff
## Model: ETS(A,N,N)
##   Smoothing parameters:
##     alpha = 0.0001000002
##
##   Initial states:
##             l
##  -0.000237666
##
##   sigma^2:  0.0021
##
##       AIC      AICc       BIC
## -46.60786 -46.54961 -34.51581
```

We can see above that the AICc for the ETS(A,N,N) model is much lower at -46.55 versus 1094.54 for the ETS(M,N,N) model. This indicates, that by looking at AICc alone the ETS(A,N,N) model is good for this data. Using the ARIMA(p=0,d=1,q=3)(P=0,D=1,Q=2)[12] model and ETS(A,N,N) model we can forecast future values and check for accuracy.

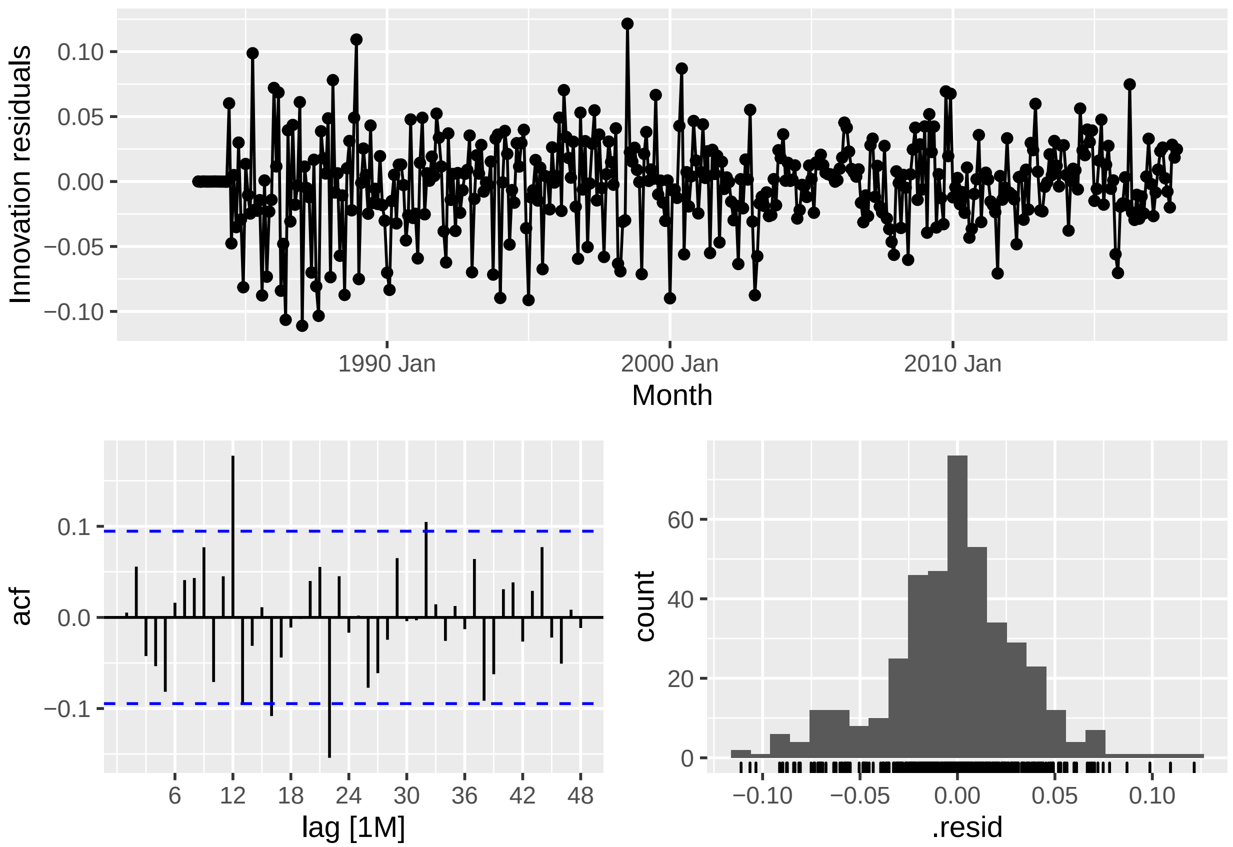
We can see below that the ETS(A,N,N) has a lot of significant lags indicating it is not white noise and the Ljung-Box test also shows a very insignificant p-value, very close to 0.



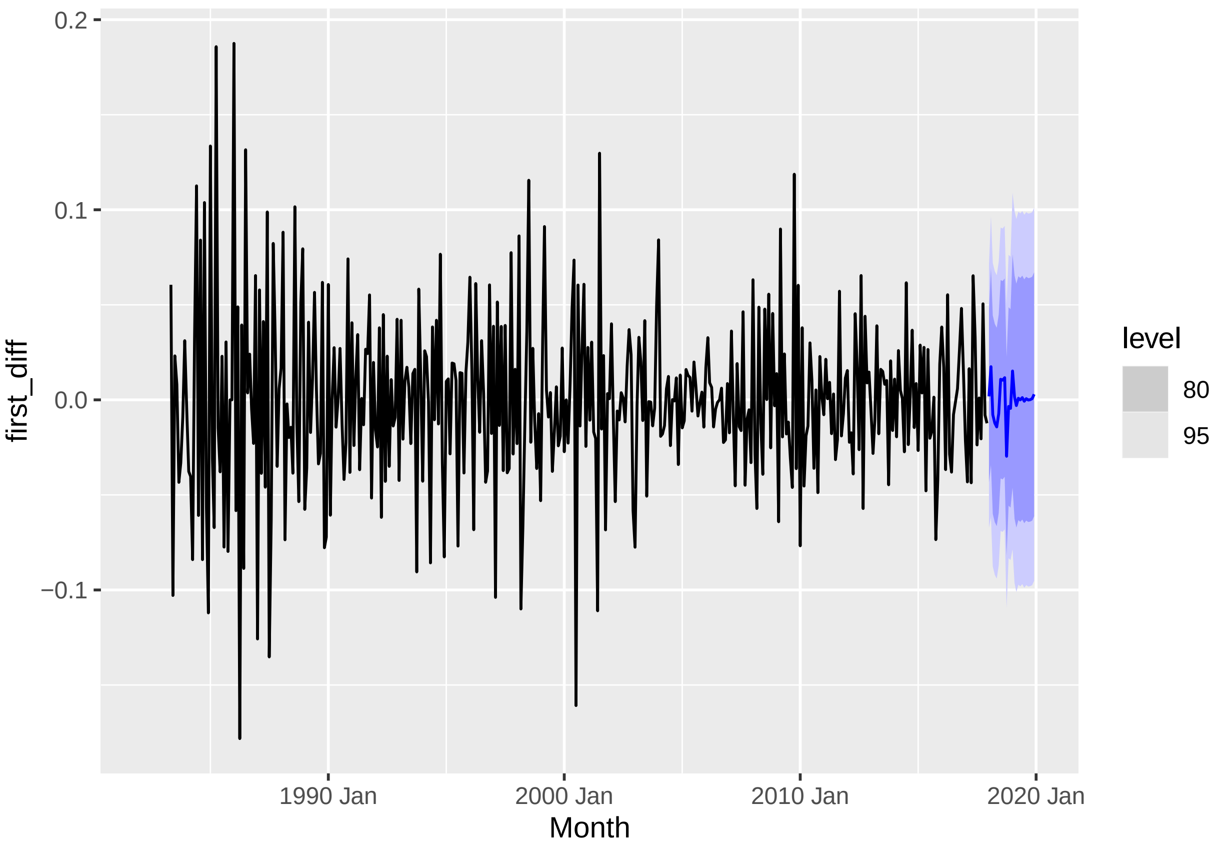
```
## # A tibble: 1 x 5
##   State                        Industry                  .model lb_stat lb_pvalue
##   <chr>                        <chr>                     <chr>    <dbl>     <dbl>
## 1 Australian Capital Territory Footwear and other pers~ "(ETS~    169.         0
```



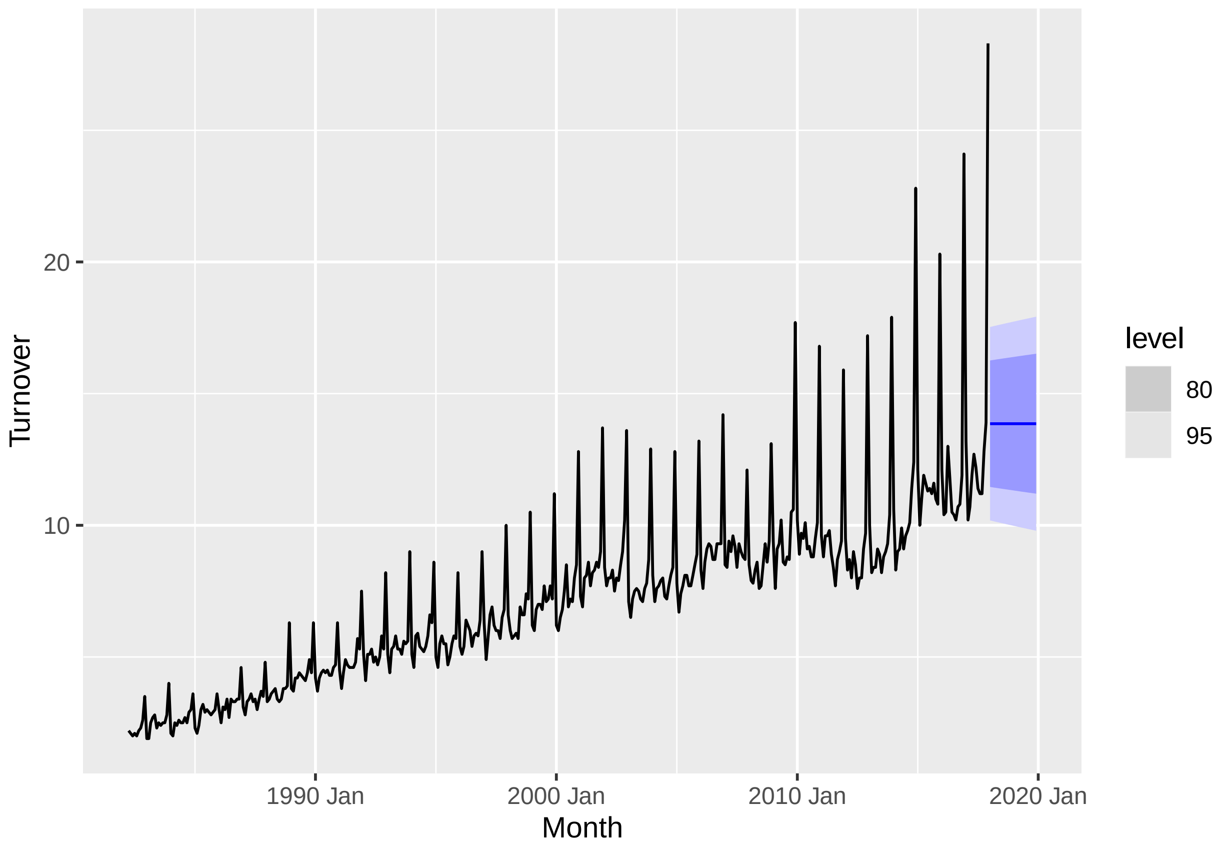
We can see below that the ARIMA(p=0,d=1,q=3)(P=0,D=1,Q=2)[12] model on the other had has fewer significant lags indicating it is closer to white noise and the Ljung-Box test also shows a low p-value, which is still higher than the ETS(A,N,N) model had. The forecasts given by this model also look much better, indicating the ARIMA model is the one to use for the full data set.



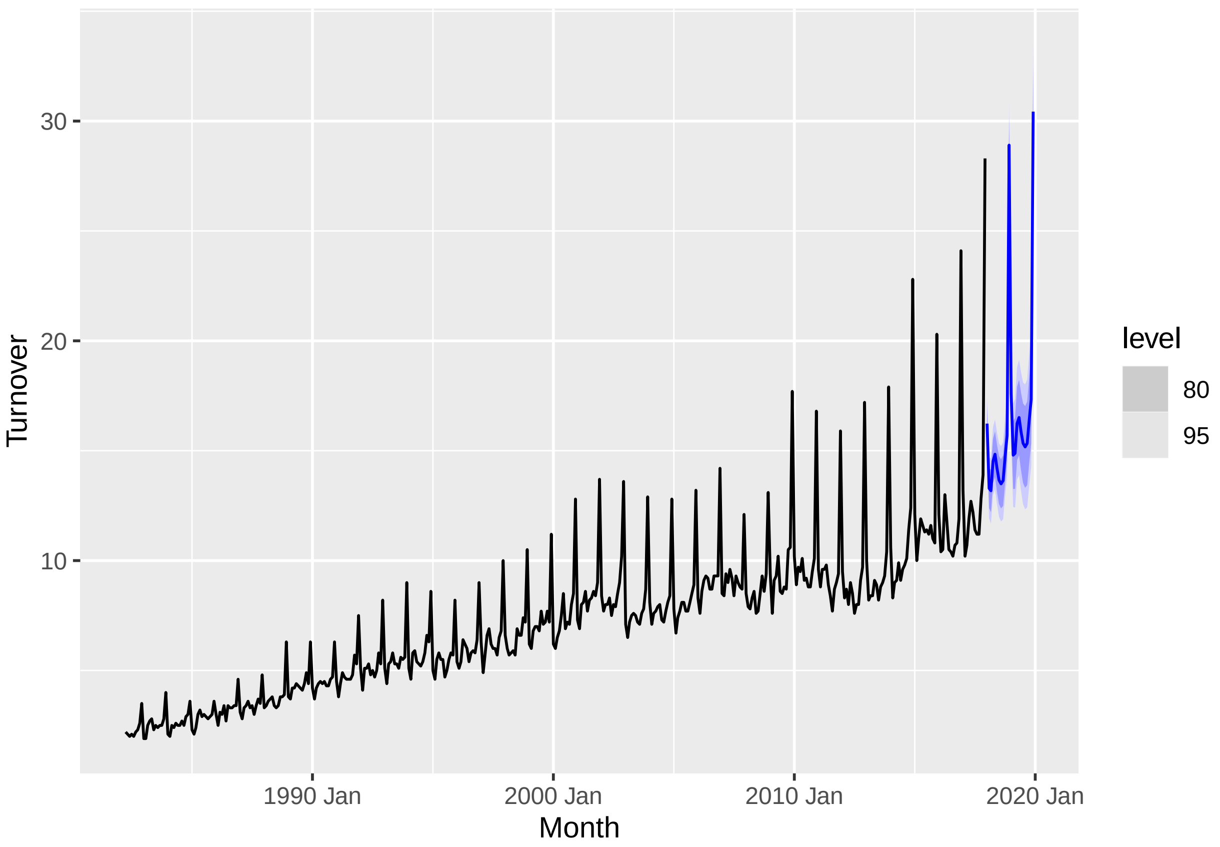
```
## # A tibble: 1 x 5
##   State                        Industry                 .model lb_stat lb_pvalue
##   <chr>                        <chr>                    <chr>    <dbl>     <dbl>
## 1 Australian Capital Territory Footwear and other pers~ arima~    50.7  0.000175
```

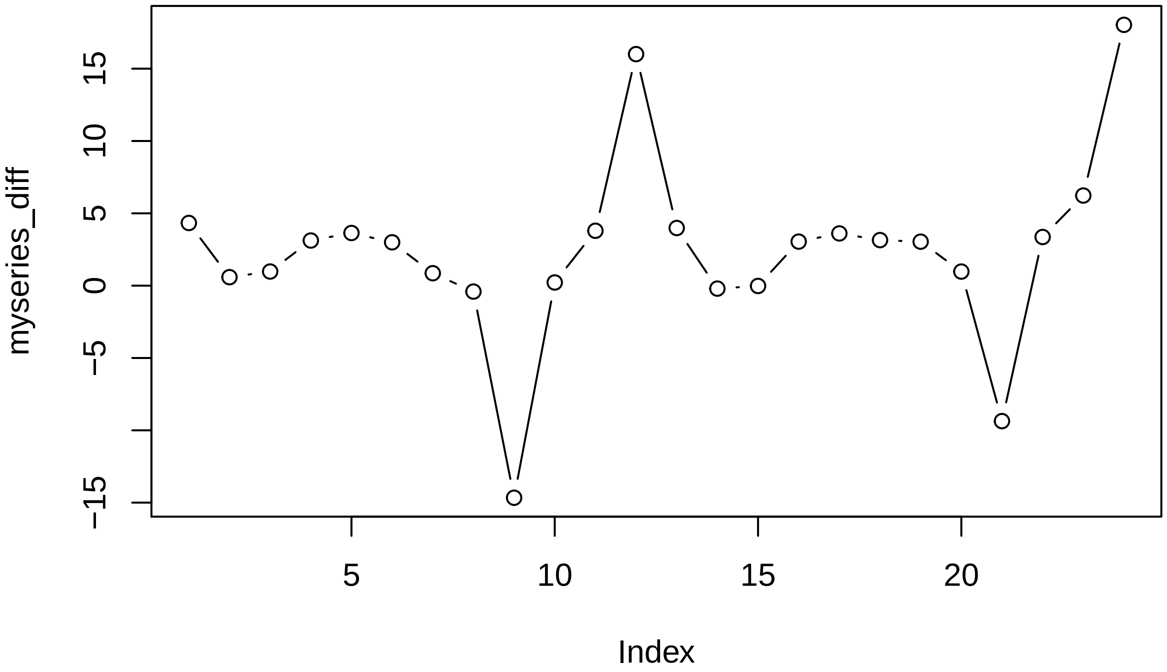


Using our full data set, we can forecast future values of Australia retail turnover, as shown below. The ETS(A,N,N) model does not provide a very good forecast for the seasonal data set we have as can be seen below.



The ARIMA(p=0,d=1,q=3)(P=0,D=1,Q=2)[12] model on the other hand provides a very good looking forecast


Turnover
30
20
10
1990 Jan
2000 Jan
2010 Jan
2020 Jan
Month
level
80
95



From the actual ABS data we can see above that this model tended to overestimate when it came to the end of year sales especially.