Forecasting Business Turnover with ARIMA model

Changsoo Byun

```
p1 <- data |>
  autoplot(y)

p2 <- data |>
  autoplot(log(y))

p3 <- data |> autoplot(log(y) |> difference(12))

data |>
  mutate(log_turnover = difference(log(y), 12)) |>
  features(log_turnover, unitroot_kpss)
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1      2.38        0.01
```

```
p4 <- data |> autoplot(log(y) |> difference(12) |> difference())


data |>
  mutate(log_turnover2 = difference(log(y), 12) |> difference()) |>
  features(log_turnover2, unitroot_kpss)
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1    0.0117         0.1
```

p1: The original data exhibits a noticeable trend and seasonality, and the variance increases as the level of the series increases. Despite these characteristics, the data is not stationary. Applying transformations and differencing can help address these issues and make the data appear stationary.

p2: Applying a log transformation simplifies the patterns in the historical data by removing known sources of variation and making the overall pattern more consistent. However, the data is still not stationary.

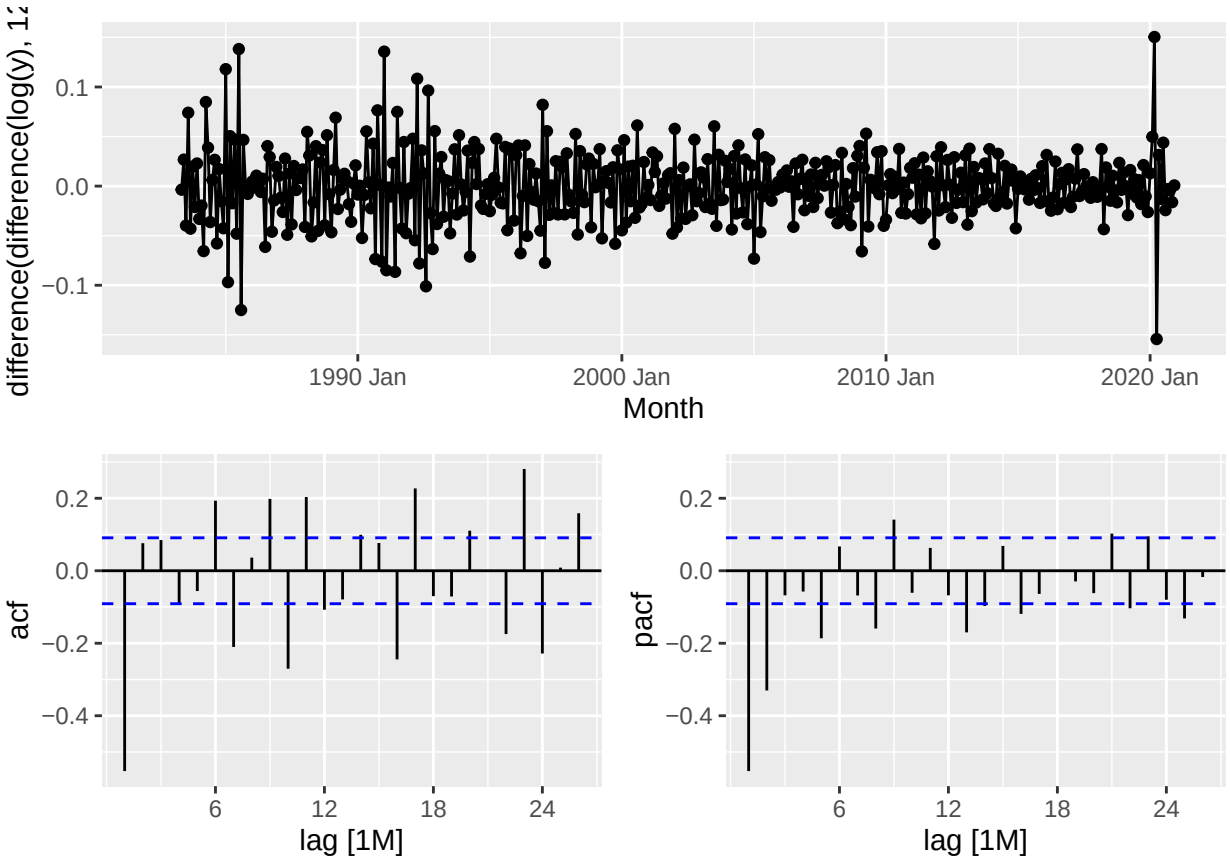
p3: The data in p3 represents the result of seasonal differencing. The KPSS test yields a p-value of 0.01, indicating that the null hypothesis of stationarity is rejected. This suggests that the seasonal differenced data is not stationary and requires further differencing.

p4: The data in p4 represents the seasonal and first-order differenced data. The KPSS test yields a p-value of 0.1, which does not provide sufficient evidence to reject the null hypothesis. Therefore, it can be concluded that the seasonal and first-order differenced data appear to be stationary.

```
data |> gg_tsdisplay(
  log(y) |> difference(12) |>
    difference(),plot_type="partial"
)
```

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

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

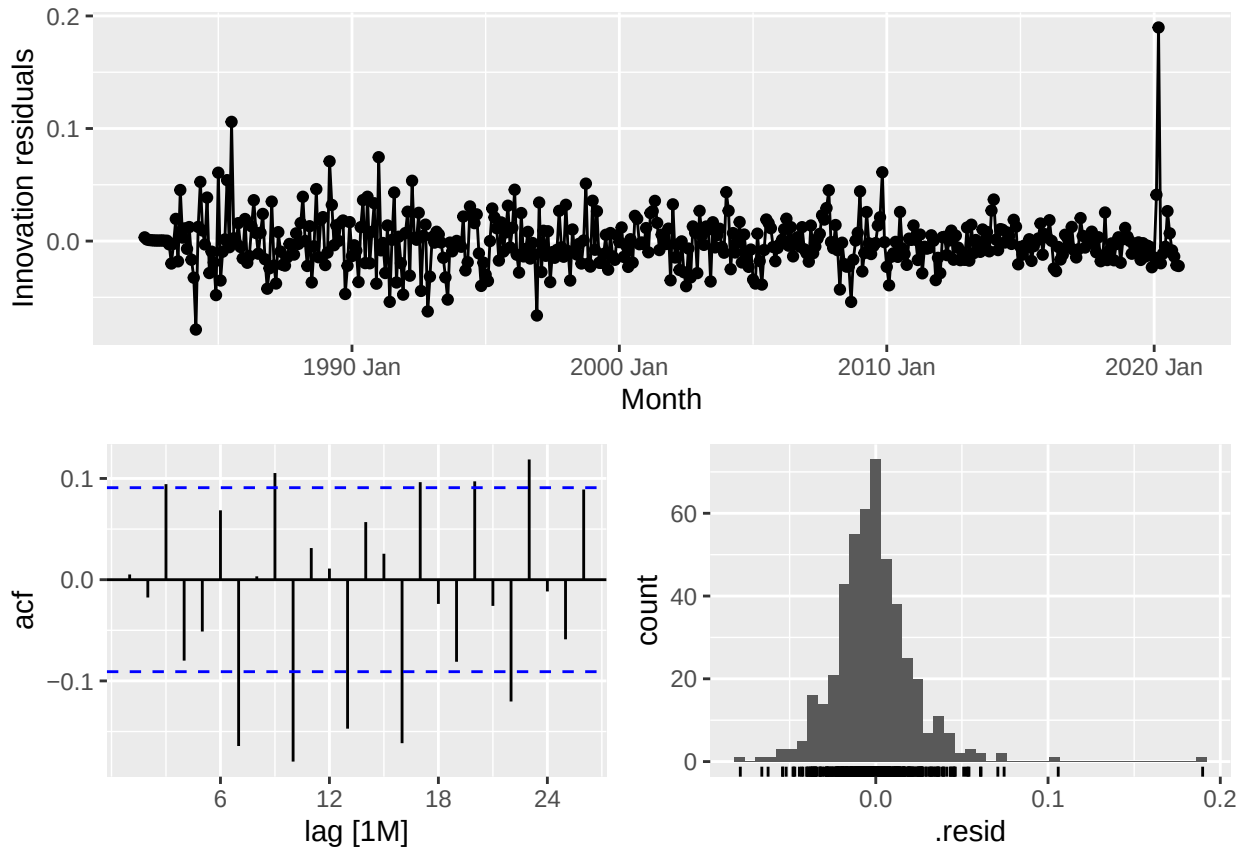



```
fit <- data |>
  model(arima = ARIMA(log(y) ~ pdq(0,1,3) + PDQ(0,1,2)))
report(fit)
```

```
## Series: y
## Model: ARIMA(0,1,3)(0,1,2)[12]
## Transformation: log(y)
##
## Coefficients:
##           ma1     ma2      ma3     sma1     sma2
##       -0.8007  0.1533  -0.0579  -0.5075  -0.3342
## s.e.   0.0498  0.0712   0.0543   0.0501   0.0486
##
## sigma^2 estimated as 0.0005625:  log likelihood=1045.7
## AIC=-2079.39   AICc=-2079.2   BIC=-2054.71
```

For the seasonal component, acf shows MA(2) and pacf shows AR(2). They have the same parameter, MA(2) is chosen to use in this case. For non-seasonal components which beloew 12, there are 3 significant spikes in acf, and 5 significant spikes in pacf.Hence, MA(5) is selected as it is more parsimonious model. d id 1 for both because of both seasonal and first order difference. Thus, (0,1,5)(0,1,2) is used

```
fit |> gg_tsresiduals()
```




```
augment(fit) |>
  features(.innov, ljung_box, lag=24, dof=5)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 arima     96.7  2.08e-12
```

There are several significant spikes crossed the bounds, so the series is not white noise. And ljung box test p value rejects the null hypothesis that the series is white noise. It is better to consider the alternative models there are many significant spikes.

```
fit2 <- data |>
  model(
    arima = ARIMA(log(y) ~ pdq(0,1,3) + PDQ(0,1,2)), #Originally chosen one
    arima2 = ARIMA(log(y) ~ pdq(0,1,3) + PDQ(2,1,0)), #Same number of spikes in seasonal components
    arima3 = ARIMA(log(y) ~ pdq(0,0,1) + PDQ(0,1,2)), #Only seasonal difference
    arima4 = ARIMA(log(y) ~ pdq(0,0,2) + PDQ(0,1,2)), #Only seasonal difference
    )
glance(fit2) |> arrange (AICc)
```

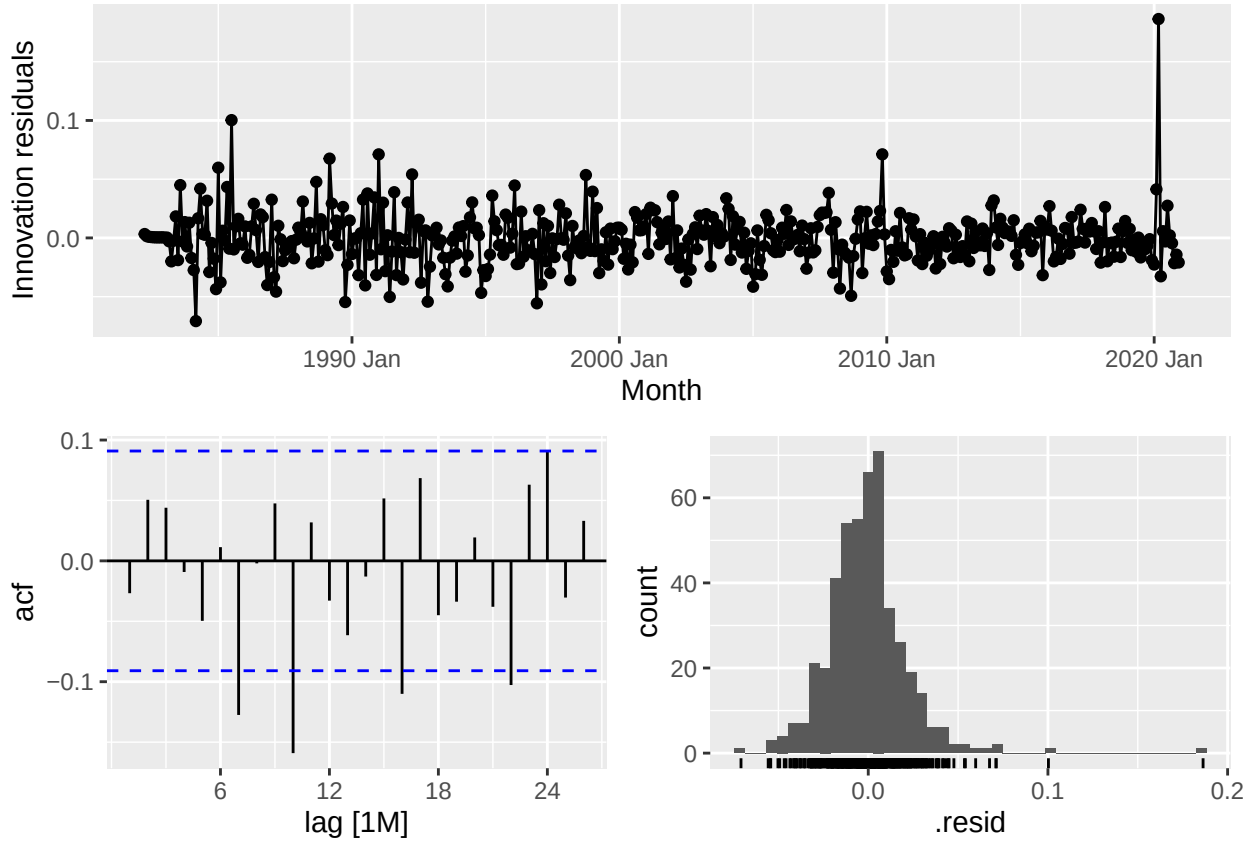
```
## # A tibble: 4 x 8
##   .model   sigma2 log_lik    AIC   AICc    BIC ar_roots   ma_roots
##   <chr>     <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <list>     <list>
## 1 arima  0.000562   1046. -2079. -2079. -2055. <cpl [0]>  <cpl [27]>
## 2 arima2 0.000683   1007. -2003. -2003. -1978. <cpl [24]> <cpl [3]>
## 3 arima4 0.000861    958. -1904. -1904. -1879. <cpl [0]>  <cpl [26]>
## 4 arima3 0.000974    930. -1850. -1849. -1829. <cpl [0]>  <cpl [25]>
```

According to the AICc value, ARIMA(0,1,3)(0,1,2) model is better as it produces the lowest value. It is the originally selected model in the previous question

```
fit3 <- data |>
  model(auto = ARIMA(log(y),
                     stepwise = FALSE,
                     approximation = FALSE)
        )

report(fit3)
```

```
## Series: y
## Model: ARIMA(0,1,1)(2,1,2)[12]
## Transformation: log(y)
##
## Coefficients:
##           ma1    sar1     sar2     sma1    sma2
##       -0.6976  0.8141  -0.5703  -1.3630  0.5734
## s.e.   0.0369  0.0712   0.0559   0.0737  0.0673
##
## sigma^2 estimated as 0.0005112:  log likelihood=1064.71
## AIC=-2117.41   AICc=-2117.23   BIC=-2092.73
```

```
fit3 |> gg_tsresiduals()
```




```
augment(fit3) |>
  features(.innov, ljung_box, lag=24, dof=5)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 auto      50.5  0.000112
```

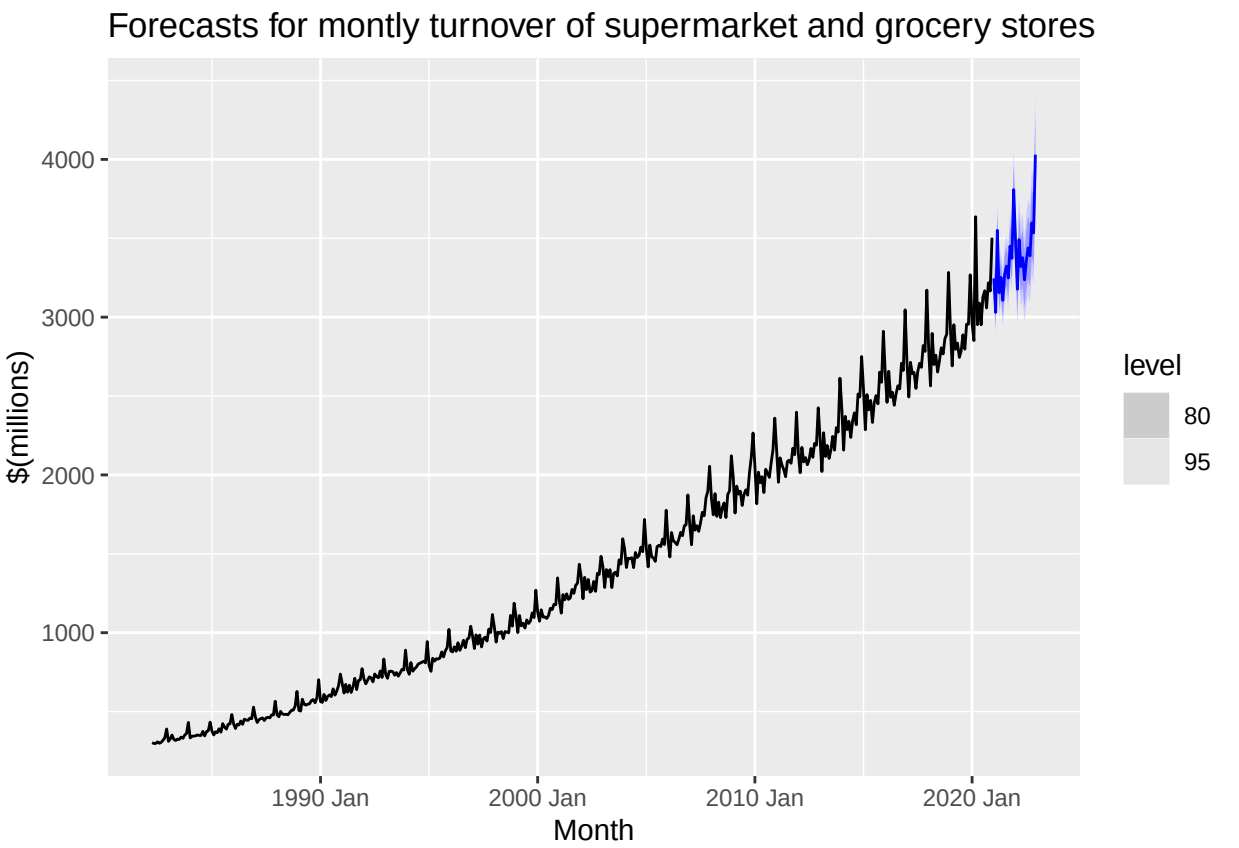
The ARIMA() function uses an ARIMA(0,1,1)(2,1,2) model, which is more complex compared to the ARIMA(0,1,3)(0,1,2) model chosen in Q4. However, the ARIMA() function has a lower AICc value, indicating that it provides a better model according to information criteria. The residuals show clear differences between the two models, ARIMA(0,1,1)(2,1,2) has fewer significant spikes and some of them got closer to the bounds. The Ljung-Box test rejects the null hypothesis that the series is white noise, suggesting that the ARIMA(0,1,1)(2,1,2) model is not performing that well.

```
test <- data |>
  slice(1:441)

test |>
  model(
    arima = ARIMA(log(y) ~ pdq(0,1,3) + PDQ(0,1,2)),
    arima2 = ARIMA(log(y) ~ pdq(0,1,3) + PDQ(2,1,0)),
    arima3 = ARIMA(log(y) ~ pdq(0,0,1) + PDQ(0,1,2)),
    arima4 = ARIMA(log(y) ~ pdq(0,0,2) + PDQ(0,1,2)),
    auto = ARIMA(log(y),
                 stepwise = FALSE,
                 approximation = FALSE)

      ) |>
  forecast(h="2 years") |>
  accuracy(data)|>
  select(.model, RMSE:MAPE)
```

```
## # A tibble: 5 x 5
##   .model  RMSE   MAE    MPE  MAPE
##   <chr>  <dbl> <dbl>  <dbl> <dbl>
## 1 arima   128.  73.5  0.299  2.30
## 2 arima2  142.  85.1  1.60   2.66
## 3 arima3  133. 109.  -2.64   3.54
## 4 arima4  136. 114.  -2.83   3.69
## 5 auto    131.  76.3  0.391  2.40
```

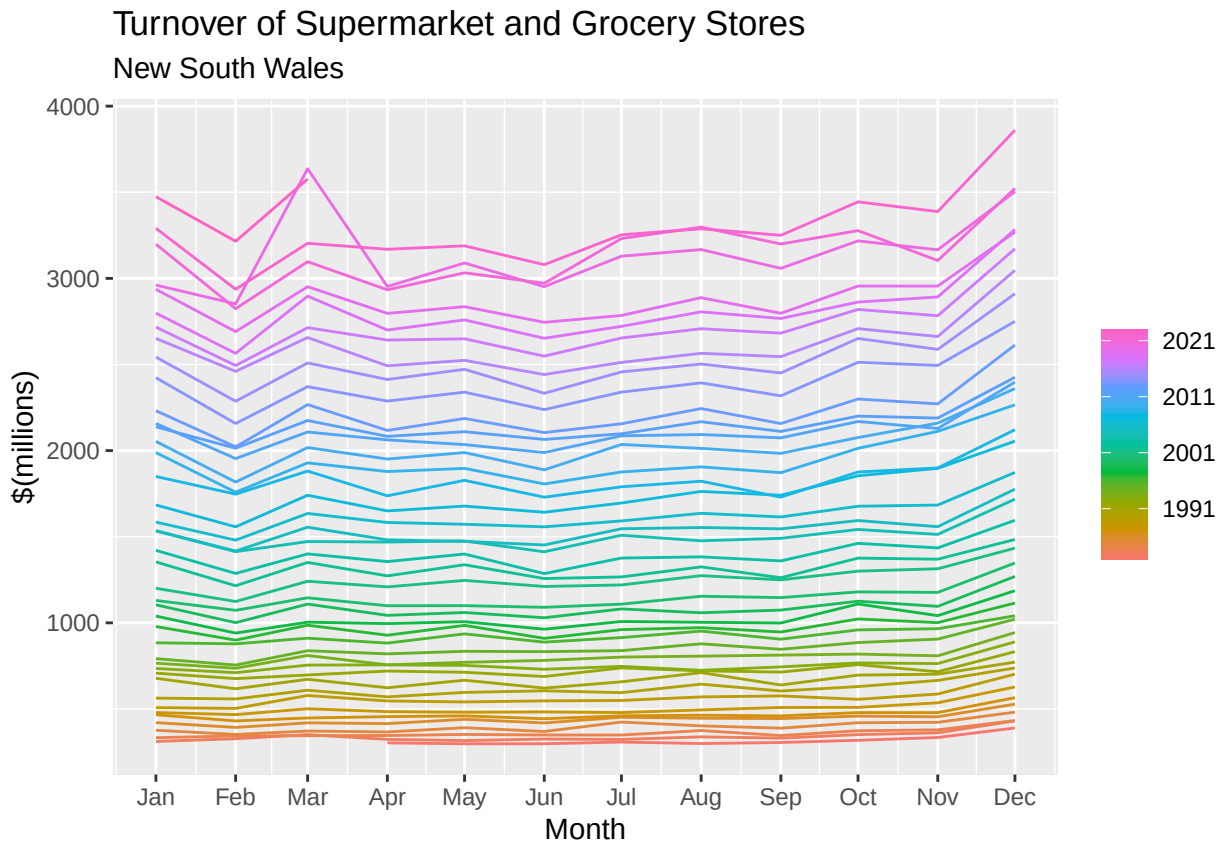
The auto model ARIMA(0,1,1)(2,1,2) will be chosen as the models chosen manually are close to the best model over this test set based on the RMSE values, while the model chosen automatically with ARIMA() is not far behind. It has the lowest AICc and second lowest RMSE.

```
fit3 |>
  forecast(h= "2 years") |>
  autoplot(data)+
  labs(y= "$(millions)",
       title= "Forecasts for montly turnover of supermarket and grocery stores")+
  guides(colour = guide_legend(title = "Forecast"))
```



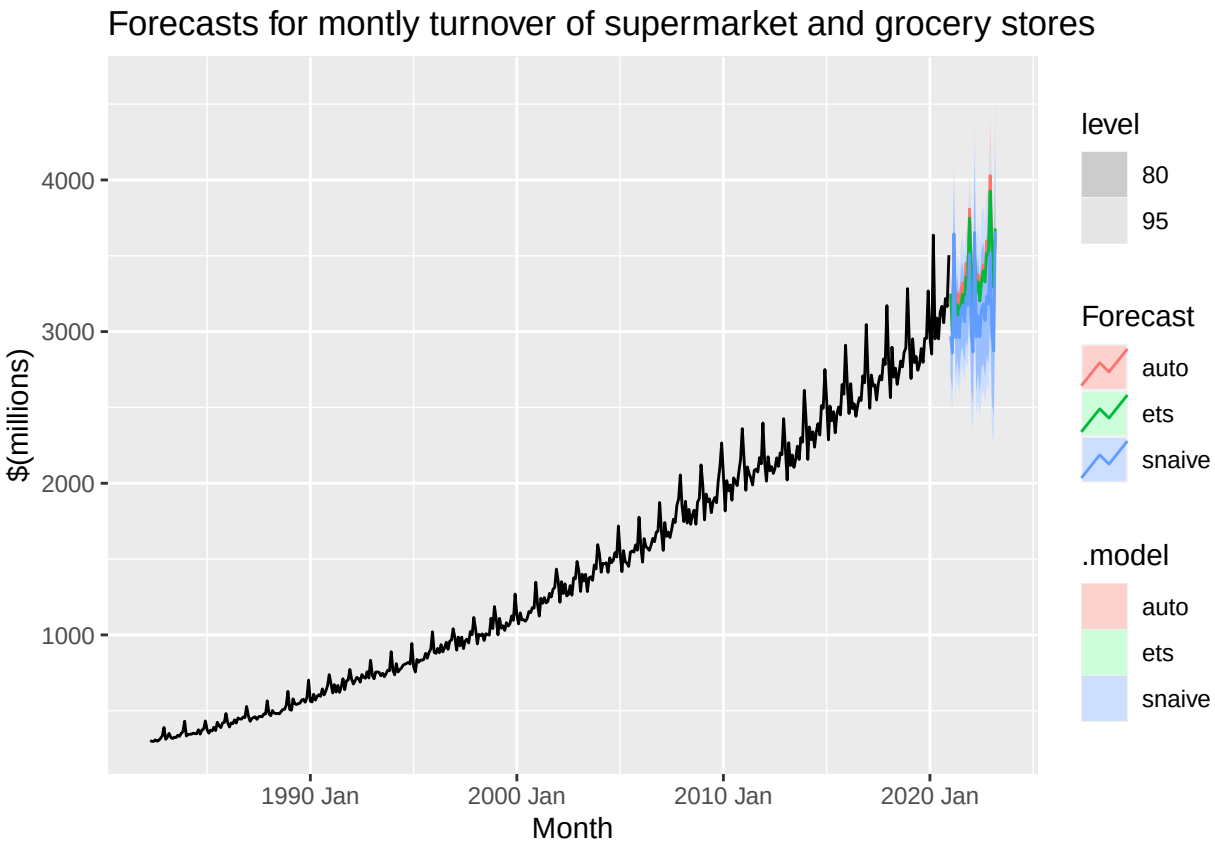
The plot shows the 80% and 95% prediction intervals and the point forecasts for the turnover of supermarket and grocery stores in New South Wales based on ARIMA method.The point forecasts look reasonable, but the intervals are narrow, recalling that only few of autocorrelation was left over in the residuals hence these will affect predictions intervals.

```
np <- newdata |>
  autoplot(y) +
  labs(title="Turnover of Supermarket and Grocery Stores",
       subtitle="New South Wales",
       y="$(millions)")

nsp <- newdata %>% gg_season(y) +
  labs(title="Turnover of Supermarket and Grocery Stores",
       subtitle="New South Wales",
       y="$(millions)")
nsp
```



The turnover rate in the supermarket follows a consistent trend and seasonality in the time plot. However, there was a significant increase in March 2021, likely due to people buying a large quantity of products in preparation for COVID isolation. This anomaly stands out and reflects the impact of external events on customer behavior and supermarket performance.

```
fc <- data |>
  model(
    snaive=SNAIVE(log(y)),
    ets=ETS(log(y)),
    auto = ARIMA(log(y),
                      stepwise = FALSE,
                      approximation = FALSE)

  ) |>
  forecast(h=27)

fc |> autoplot(data)+
  labs(y= "$(millions)",
       title= "Forecasts for montly turnover of supermarket and grocery stores")+
  guides(colour = guide_legend(title = "Forecast"))
```



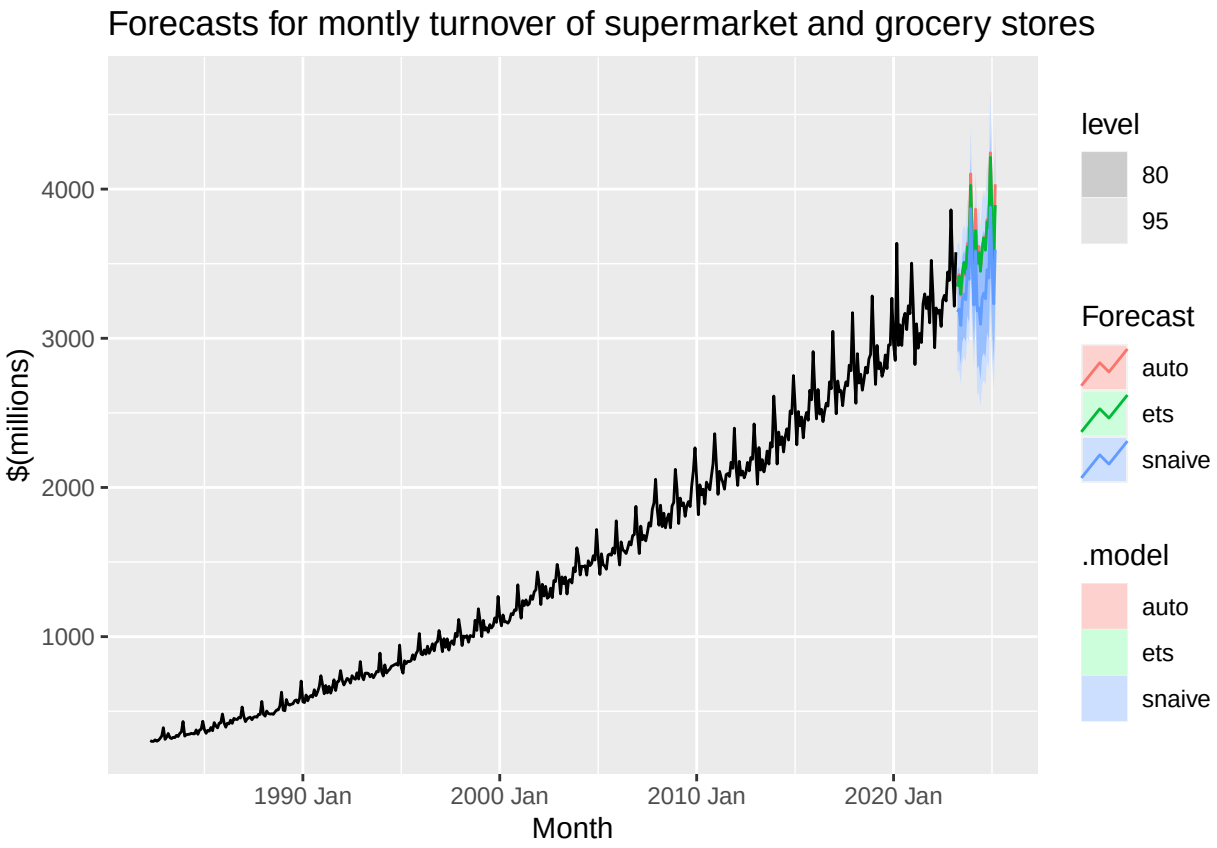
The point forecasts look to be quite similar, but SNAIVE produces lower values and wider forecast interval than other two models.

```
fc |> accuracy(newdata)
```

```
## # A tibble: 3 x 10
##   .model .type     ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 auto   Test  -166.   190.  166. -5.20  5.20  2.21  2.01 0.430
## 2 ets    Test  -113.   142.  123. -3.57  3.87  1.64  1.50 0.515
## 3 snaive Test    79.7  228.  174.  2.31  5.32  2.32  2.42 0.262
```

Baed on the table, ETS model has the lowest RMSE value, indicating this model gives the most accurate forecasts.

```
fc2 <- newdata |>
  model(
    snaive=SNAIVE(log(y)),
    ets=ETS(log(y)),
    auto = ARIMA(log(y),
                      stepwise = FALSE,
                      approximation = FALSE)

  ) |>
  forecast(h=24)

fc2 |> autoplot(newdata)+
  labs(y= "$(millions)",
       title= "Forecasts for montly turnover of supermarket and grocery stores")+
  guides(colour = guide_legend(title = "Forecast"))
```



Based on the generated plot, both the ARIMA and ETS models exhibit reasonable forecast points by capturing the trend and seasonality of the turnover rate, even in the presence of the pandemic's impact. Despite the significant increase in turnover rate in March 2021, the seasonality and trend remain consistent across all three models.

However, the SNAIVE model shows lower forecast values and wider intervals compared to the other two models. Nevertheless, it still manages to capture the underlying seasonality and trend observed in the data. It is worth noting that the turnover rate of supermarkets and grocery stores in NSW generally experiences an increase during March, coinciding with the drastic surge observed in March 2021. This suggests that the impact of COVID on the turnover rate may not be as pronounced in the series.

Overall, despite some variations in forecast values and intervals, all three models successfully capture the seasonality and trend of the turnover rate, with the ARIMA and ETS models demonstrating reasonable forecasts.