Retail Project

Shu To Yeung30159326

19/05/2021

Introduction

The purpose of this report is producing forecast of turnover in retail industry. It would be focusing on the turnover in Furniture, floor coverings, houseware and textile goods retailing industry at Victoria. First it would provide a background information about the statistical properties of the data. Then, we would fit some ETS and ARIMA models. We would choose one model from each method, using it to do some forecasts. Finally, we would compare the forecasts between the 2 models we have chosen to see which one gave us the better forecasts.

Loading in neccessary libraries

```
library(fpp3)
```

```
## -- Attaching packages -------------------------------------------- fpp3 0.4.0 --
```

```
## v tibble      3.1.2      v tsibble     1.0.1
## v dplyr       1.0.6      v tsibbledata 0.3.0
## v tidyr       1.1.3      v feasts      0.2.1
## v lubridate   1.7.10     v fable       0.3.0
## v ggplot2     3.3.3
```

```
## -- Conflicts ------------------------------------------------- fpp3_conflicts --
## x lubridate::date()    masks base::date()
## x dplyr::filter()      masks stats::filter()
## x tsibble::intersect() masks base::intersect()
## x tsibble::interval()  masks lubridate::interval()
## x dplyr::lag()         masks stats::lag()
## x tsibble::setdiff()   masks base::setdiff()
## x tsibble::union()     masks base::union()
```

```
library(readabs)
```

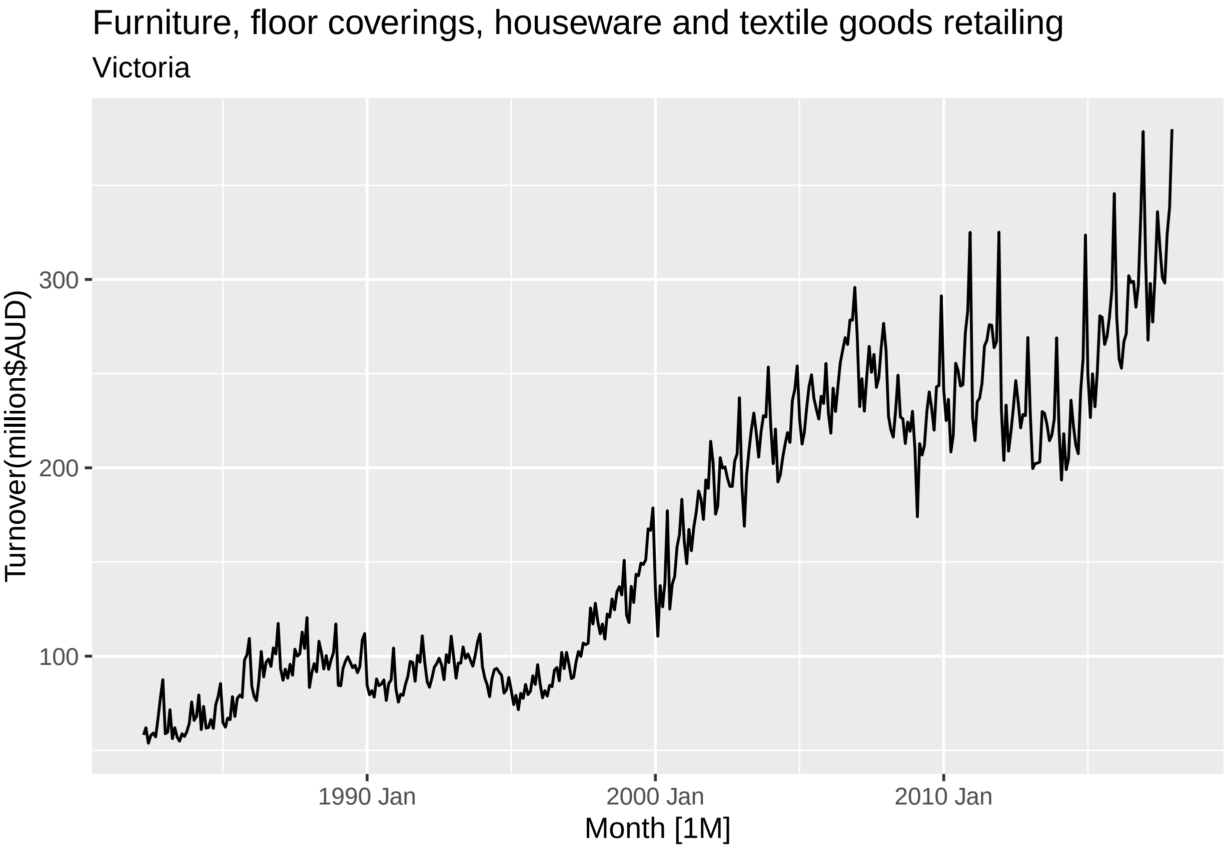
```
## Environment variable 'R_READABS_PATH' is unset. Downloaded files will be saved in a temporary directo
## You can set 'R_READABS_PATH' at any time. To set it for the rest of this session, use
##  Sys.setenv(R_READABS_PATH = <path>)
```

Reading in the data

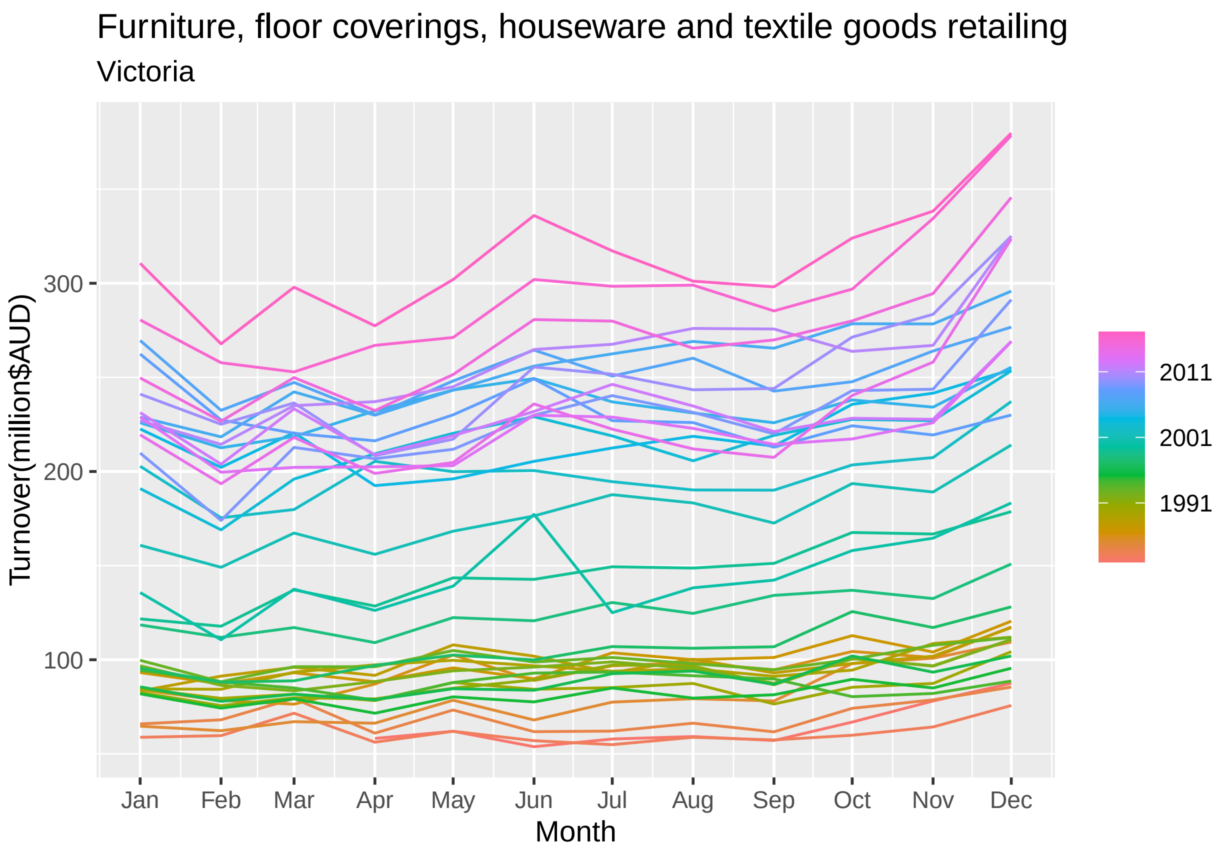
```
## Selecting by Turnover
```

```
## # A tsibble: 5 x 5 [1M]
## # Key:       State, Industry [1]
##   State   Industry                                 `Series ID`    Month Turnover
##   <chr>   <chr>                                    <chr>          <mth>    <dbl>
## 1 Victor~ Furniture, floor coverings, houseware a~ A3349413L   2015 Dec     346.
## 2 Victor~ Furniture, floor coverings, houseware a~ A3349413L   2016 Dec     378.
## 3 Victor~ Furniture, floor coverings, houseware a~ A3349413L   2017 Jun     336
## 4 Victor~ Furniture, floor coverings, houseware a~ A3349413L   2017 Nov     338.
## 5 Victor~ Furniture, floor coverings, houseware a~ A3349413L   2017 Dec     380.
```

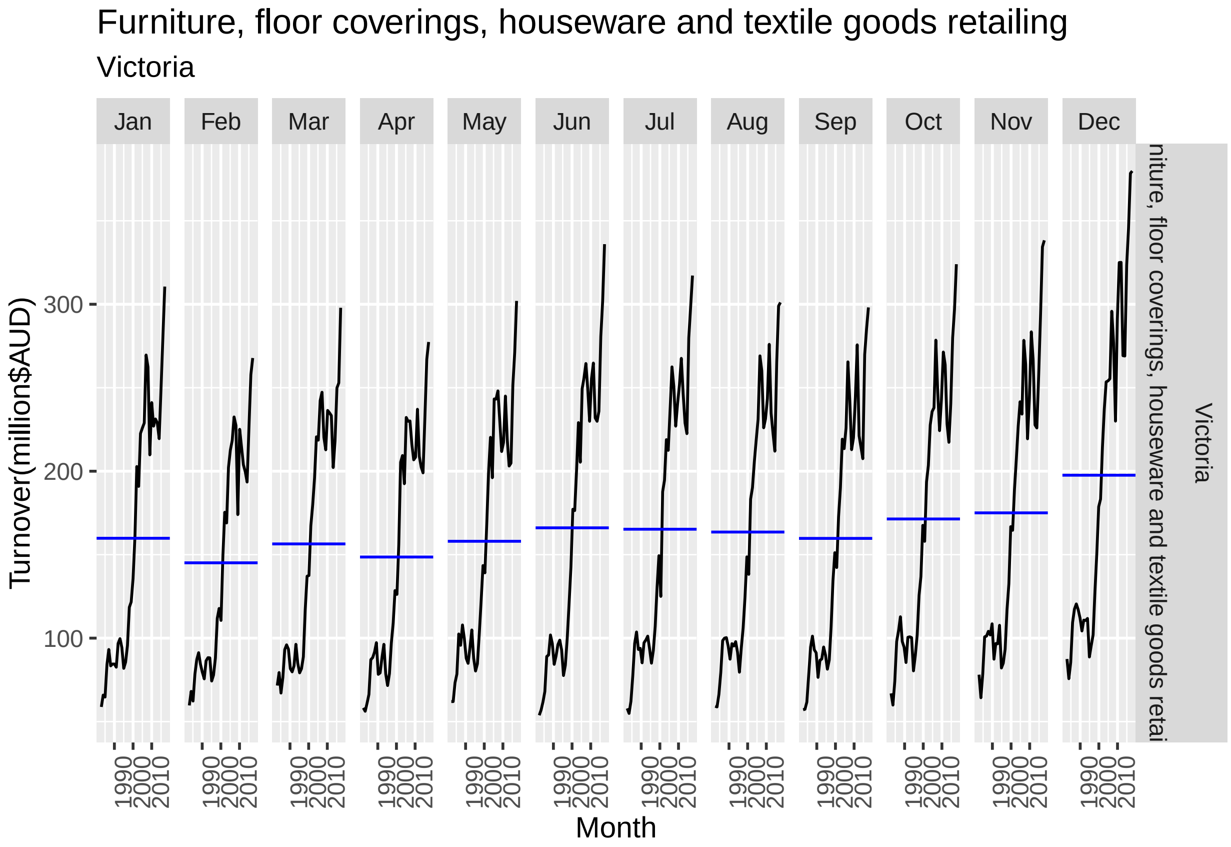
Dicussion of statistical features



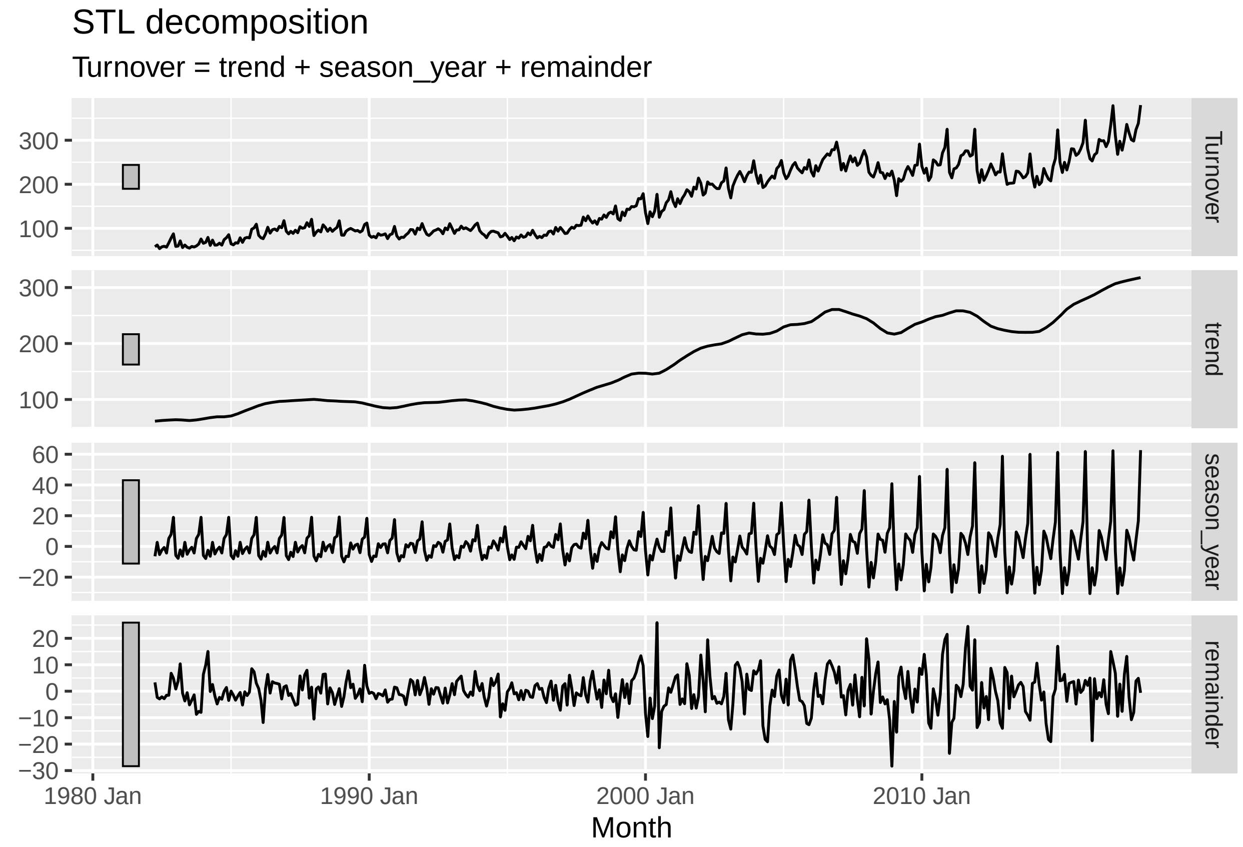
Generally we can see a postitve trend, the variance of the data increase over time. Therefore, it doesn't have a constant mean and constant variance.



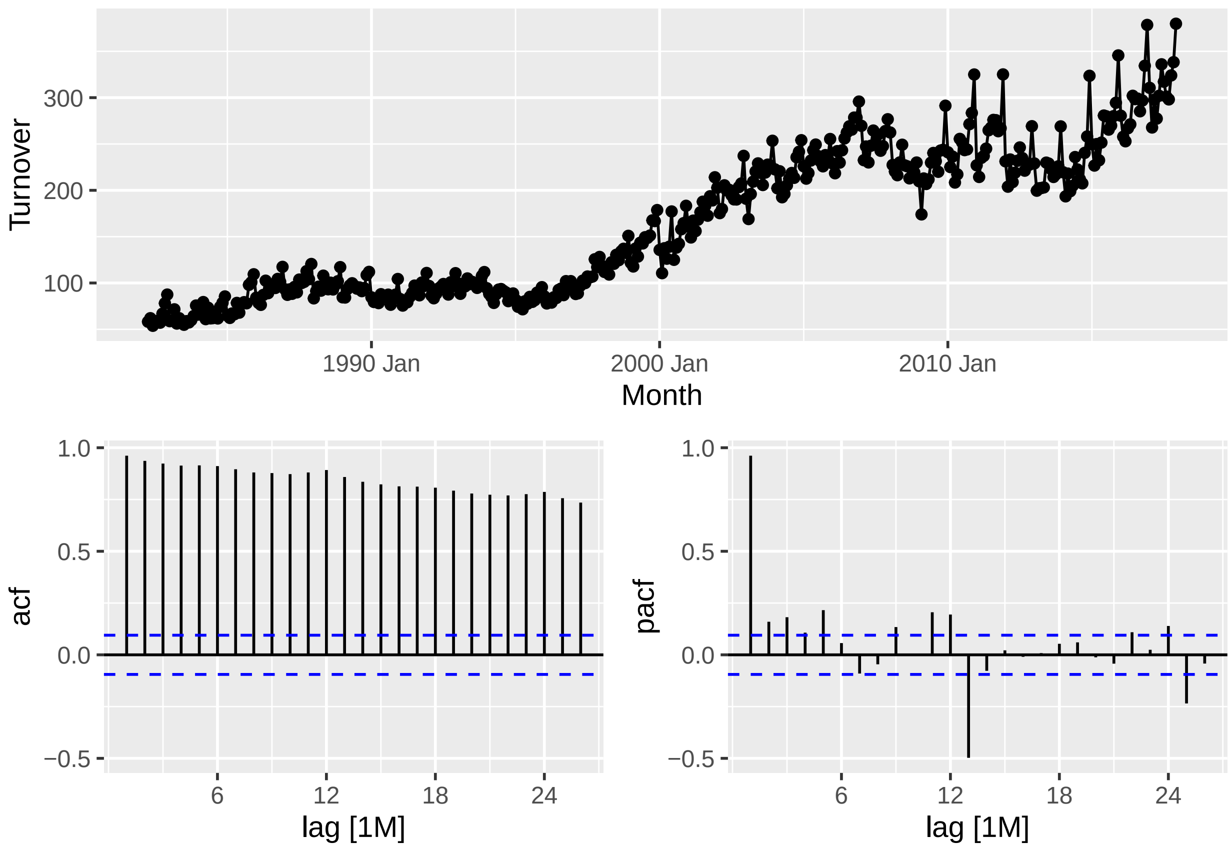
From the gg_season plot, the turnover is the highest in December. There is a second peak at June in recent years. We can see in the middle of the plot, there is a large spike at June in around 2000.



From the gg_subseries plot, We can see that the pattern of each month is quite similar. But the turnover is obviously higher in December. The turnover is lower at February. Maybe is because there are only 28 days in February most the times.

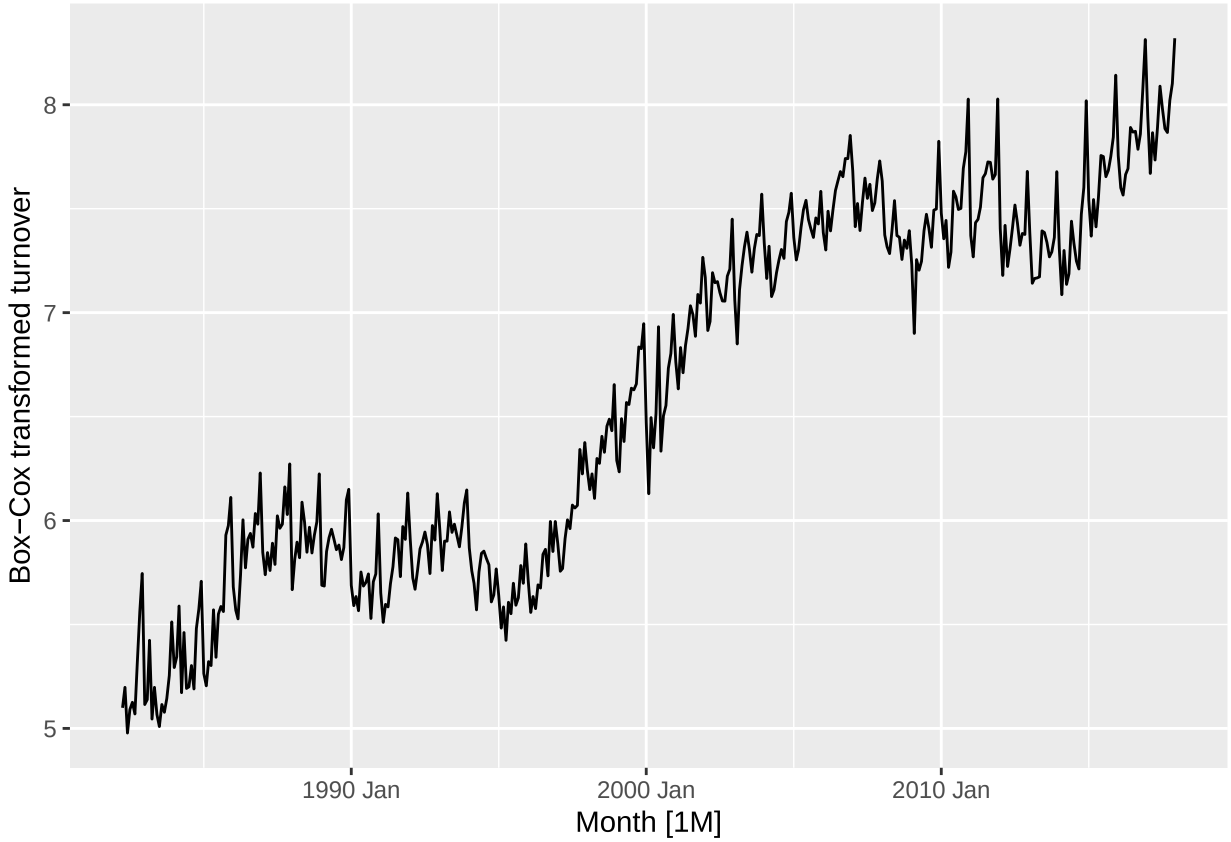


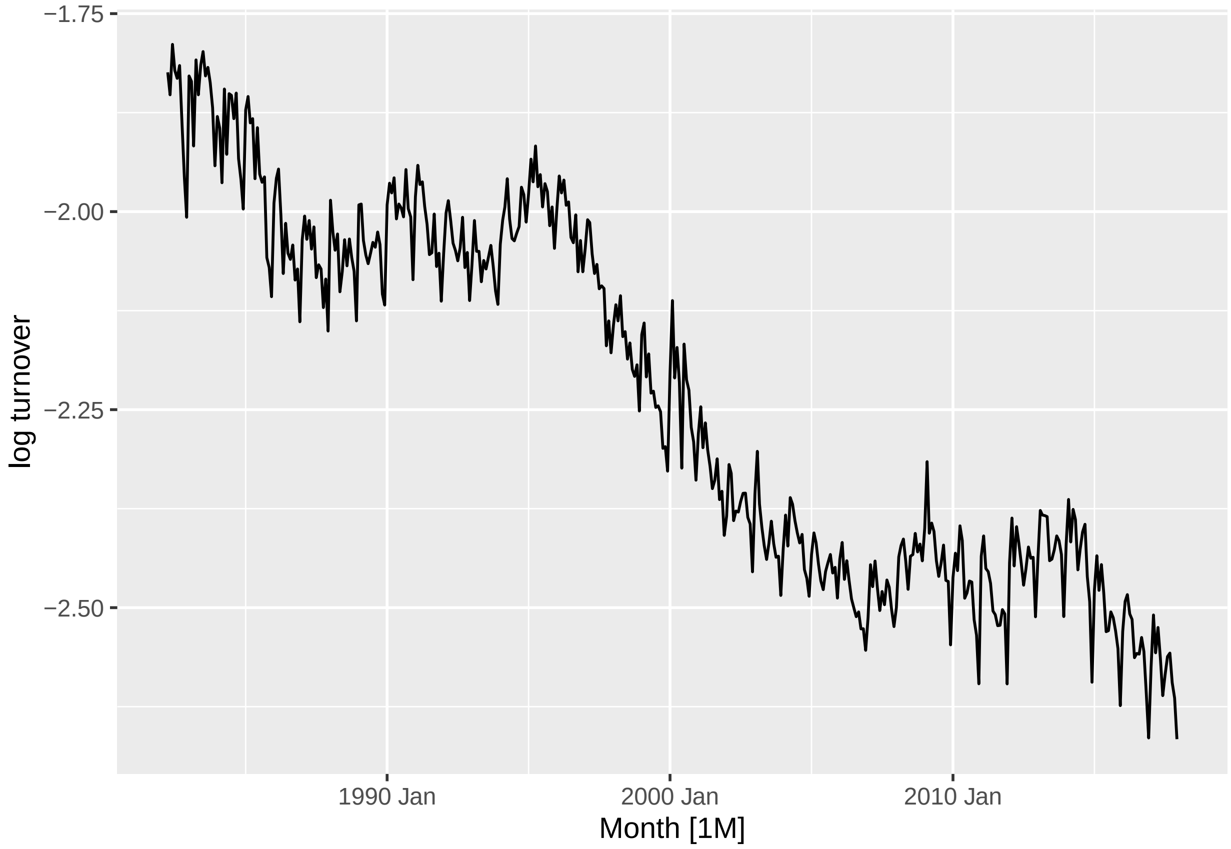
As STL decomposition can only apply in addititve decomposition, we can only see the trend. But for season and remainder part, we can see the variation is increasing as time proceesed. Therefore, we can't get many information from it.



From the ACF part in the tsdisplay plot, we can also see that our data has a trend and seasonality. There is a scalloped shape in the acf plot due to the seasonality. Also, the slow decrease acf as lags increase is due to the trend.

Explanation of transformations and differencing used.



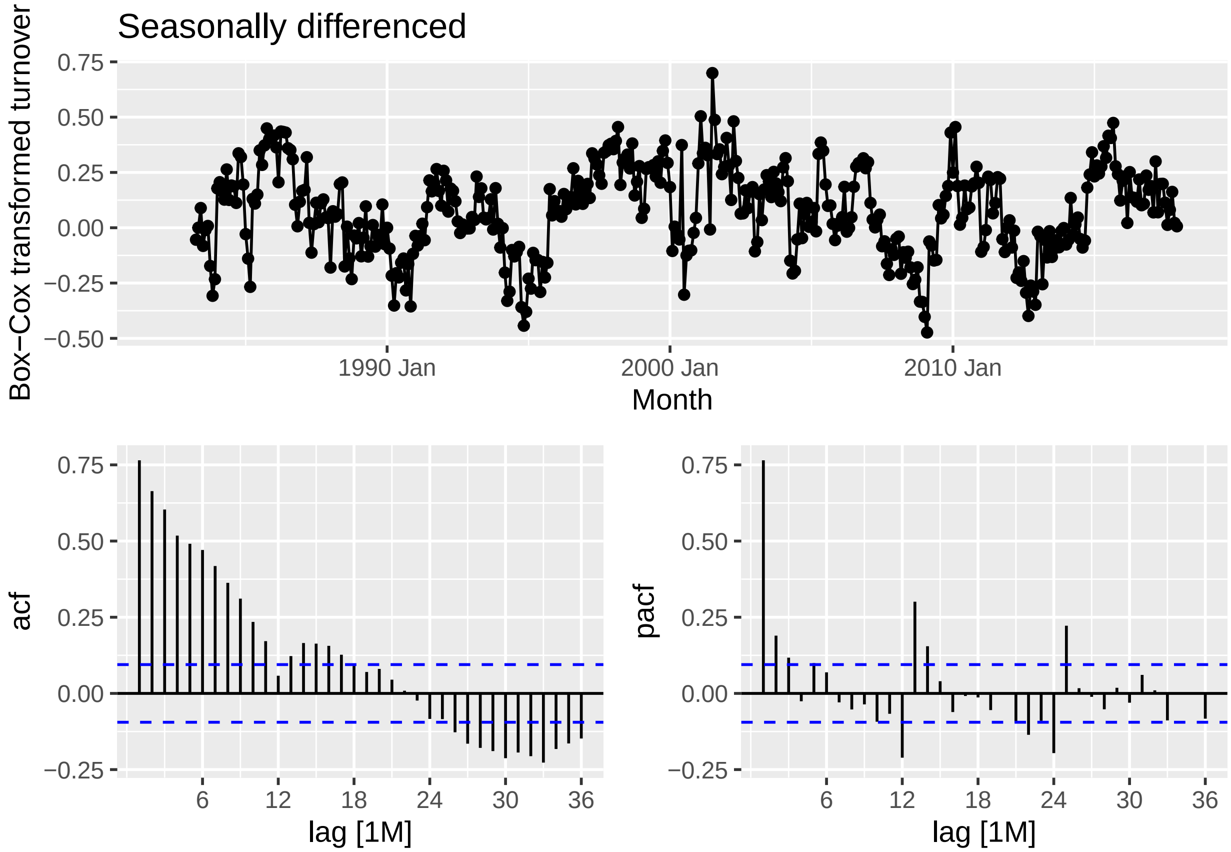


First, given the autoplot from the above part, we can see that the data is multiplicative decomposition as the turnover increased, variance has increased too. After that, we fit a log transformation to the data. We see it actually balance the variation of at the start and the end. To get a more accurate lambda value in box-cox transformations, we apply the guerrero method and get the lambda value of 0.1 which is closed to 0, the log transformations. Therefore, 0.1 lambda value provides a similar autoplot as the log transformations. However, in terms of accuracy, we would choose to use lambda value provided by the box-cox transformation.

```
## # A tibble: 1 x 5
##   State   Industry                                      nsdiffs kpss_stat kpss_pvalue
##   <chr>   <chr>                                           <int>     <dbl>       <dbl>
## 1 Victor~ Furniture, floor coverings, houseware a~            1      6.73        0.01
```

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

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



Second, from the unitroot test, we should apply 1 seasonal difference to our data to remove the seasonality as the kpss p value is smaller than 0.05. We can reject the null and conclude that our data is not stationary. From the ACF plot, the lags are slowly decaying.

```
## # A tibble: 1 x 5
##   State   Industry                                        ndiffs kpss_stat kpss_pvalue
##   <chr>   <chr>                                            <int>     <dbl>       <dbl>
## 1 Victor~ Furniture, floor coverings, houseware an~            0     0.161         0.1
```

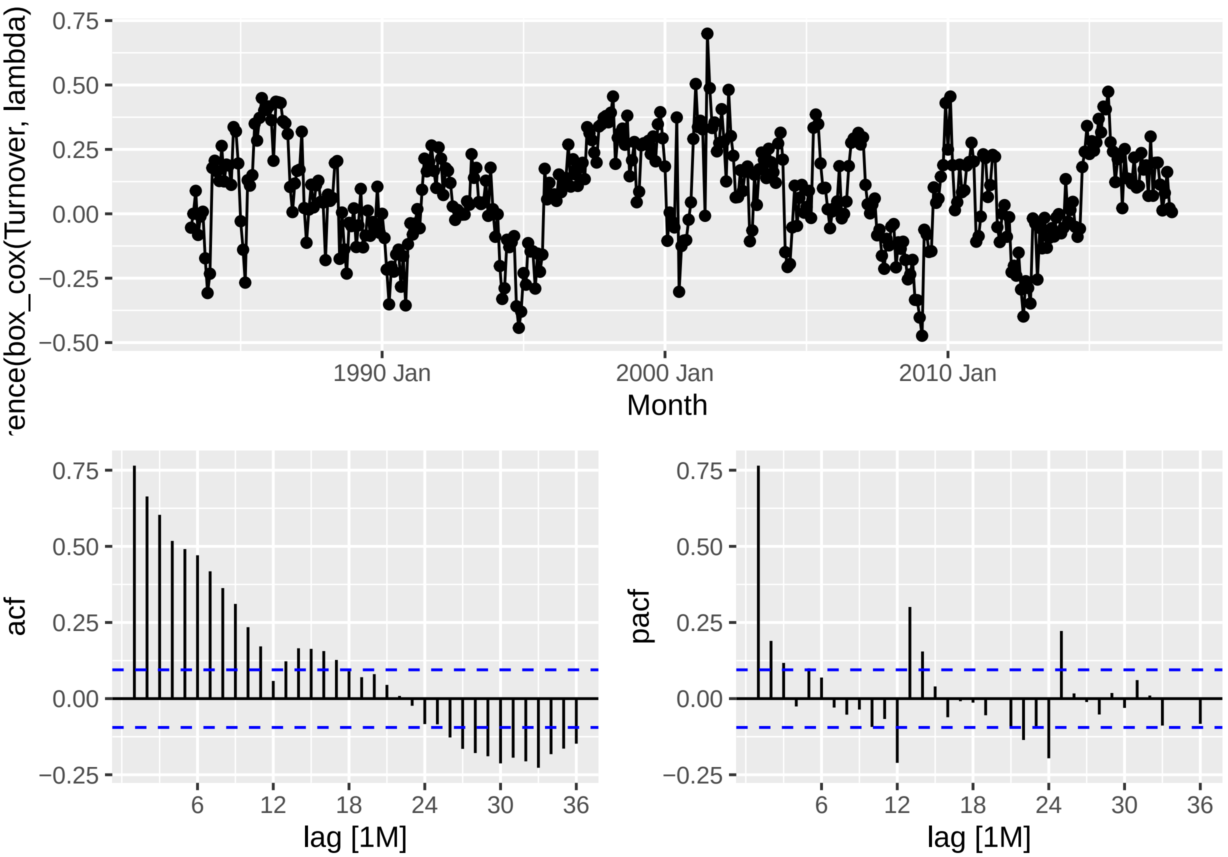
Third, we do a unitroot test to test do we need a further ordinary difference. From the kpss p value, it is larger than 0.05 (0.01>0.05). So, we can conclude that we can accept the null that our data is stationary now and we don't need to do an ordinary difference.

A short list of appropriate ARIMA model

Look at acf and pcf which one is a simpler model to start with

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

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



```
## $y
## [1] "Box-Cox transformed turnover"
##
## $title
## [1] "Seasonally differenced"
##
## attr(,"class")
## [1] "labels"
```

box_cox(Turnover, lambda)~ ARIMA(p=2,d=0,q=0)(P=2,D=1,Q = 0)[12] # lag2 has dropped a lot but still quite significant box_cox(Turnover, lambda)~ ARIMA(p=3,d=0,q=0)(P=2,D=1,Q = 0)[12] # lag3 is barely significant

From the second part of analysis, the test suggest we should use a seasonal difference to our data. Therefore, for all ARIMA models we have included in the list have a seasonal difference too. From the acf plot, as it is still slowly decaying, it is hard to choose a lag to apply to our model. From the pacf plot, the 1st lag is highly significant and 2nd lag had dropped a lot but is also significant. Therefore, p=2 could be a good start. Instead, the 3rd lag is barely significant, we should also include in our list to see how the model perform. Likewise, for the seasonal component, P=2, should be included, last seasonal lag is 2nd one, and we have a monthly data, so lag 24 is a significant lag.

A short list of appropriate ETS model

Turnover ~ error("M")+trend("A")+season("M"))

Turnover ~ error("M")+trend("Ad")+season("M"))

For the ETS model, from the original autoplot of turnover, we can see that the variation of each year is becoming bigger. Therefore, we should consider to use a multiplicative seasonal component to our data. As we have multiplicative seasonality, we should use multiplicative errors. Additionally, ETS need not worry about the transformation of the reponse variable as it can take care of it. For the trend componenet, it is obvious that our data has a trend, so we can try to use "A" or "Ad" component to see which one can produce a better forecast.

For both ARIMA and ETS model, an automatic model chosen by the R function ETS() and ARIMA() has been added our list. We would like to see which one would be our best models.

```
## # A tibble: 6 x 13
##   State  Industry   .model  sigma2 log_lik   AIC  AICc   BIC   MSE  AMSE     MAE
##   <chr>  <chr>      <chr>    <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 Victo~ Furniture~ madm   0.00407  -2088. 4213. 4214. 4285.  94.8  128.  0.0474
## 2 Victo~ Furniture~ mam    0.00405  -2089. 4211. 4213. 4279.  94.7  126.  0.0476
## 3 Victo~ Furniture~ auto_~ 0.00405  -2089. 4211. 4213. 4279.  94.7  126.  0.0476
## 4 Victo~ Furniture~ arima~ 0.0127     301. -590. -590. -566.  NA     NA  NA
## 5 Victo~ Furniture~ arima~ 0.0122     308. -602. -601. -574.  NA     NA  NA
## 6 Victo~ Furniture~ auto_~ 0.0110     326. -636. -636. -604.  NA     NA  NA
## # ... with 2 more variables: ar_roots <list>, ma_roots <list>
```

```
## # 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_ets Test   17.5  21.4  18.8  5.55  6.04  1.21  1.05 0.483
## 2 madm     Test   15.7  20.5  17.5  4.89  5.58  1.13  1.01 0.533
## 3 mam      Test   17.5  21.4  18.8  5.55  6.04  1.21  1.05 0.483
```

```
## # 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 arima200210c Test  18.2   21.3 19.3   5.89  6.31 1.25  1.05  0.323
## 2 arima300210c Test  13.7   17.1 15.3   4.41  5.01 0.988 0.840 0.220
## 3 auto_arima   Test   7.30  12.3  9.42  2.22  3.04 0.608 0.604 0.205
```

When comparing with ETS model, the mam model is producing a slightly lower AIC value than madm model. When comparing with ARIMA model, the auto selected model has the lowest AIC which is 1 + ARIMA(1,0,2)(2,1,1)[12].

Based on the test accuracy, for ETS model, the madm model is having a lower RMSE, RMSSE, MAE and MASE. For arima model, the auto selected model is having the lowest RMSE, RMSSE, MASE and MAE.

As the AIC value of madm is just slightly lower than the mam model, and the test accuracy for madm is lower than mam quite a lot. In addition, we expect that the trend would be a bit flattern than before. Therefore, we would choose madm as the most appropriate model.

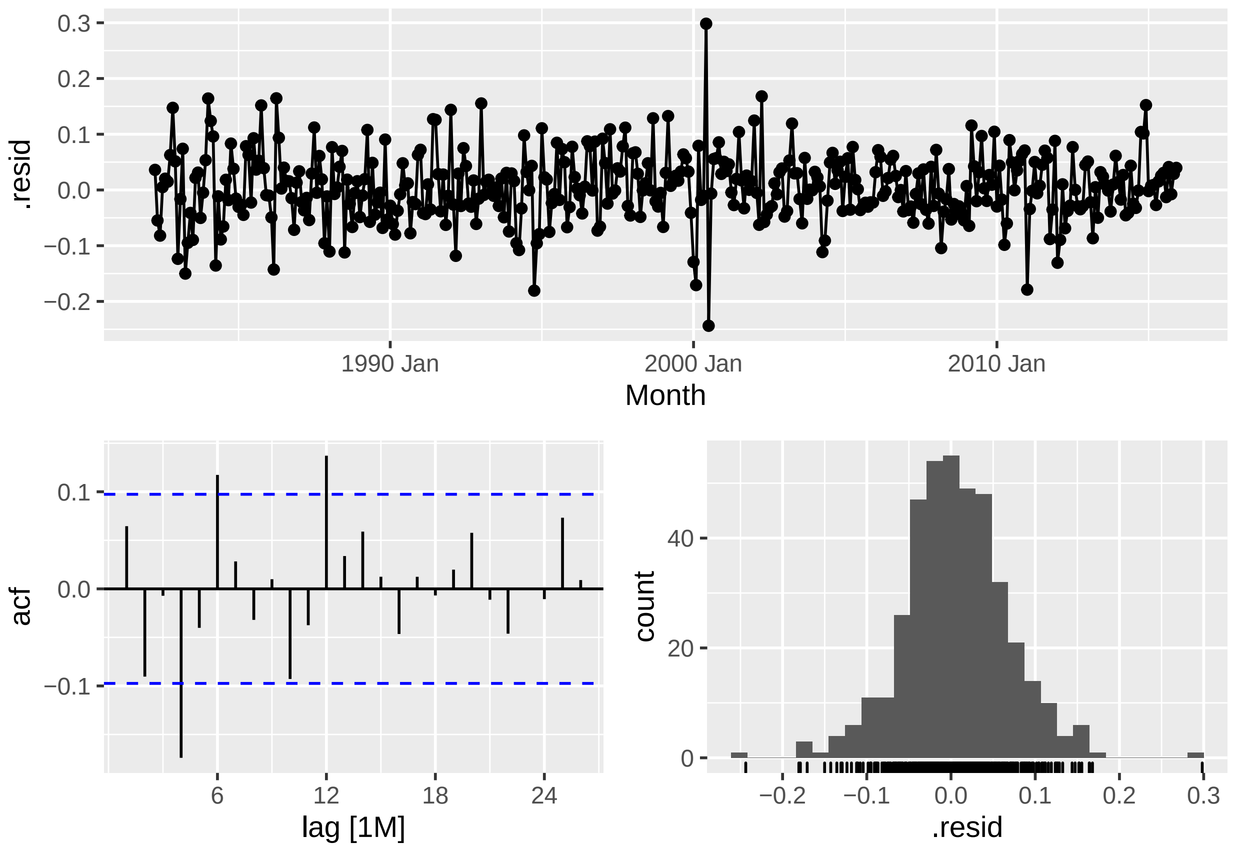
For ARIMA model, we would choose, the auto selected model,1 + ARIMA(1,0,2)(2,1,1)[12], as it has the lowest AIC value and lowest RMSE among our list.

Choose one ARIMA model and one ETS model based on this analysis and show parameter estimates, residual diagnostics, forecasts and prediction intervals for both models.

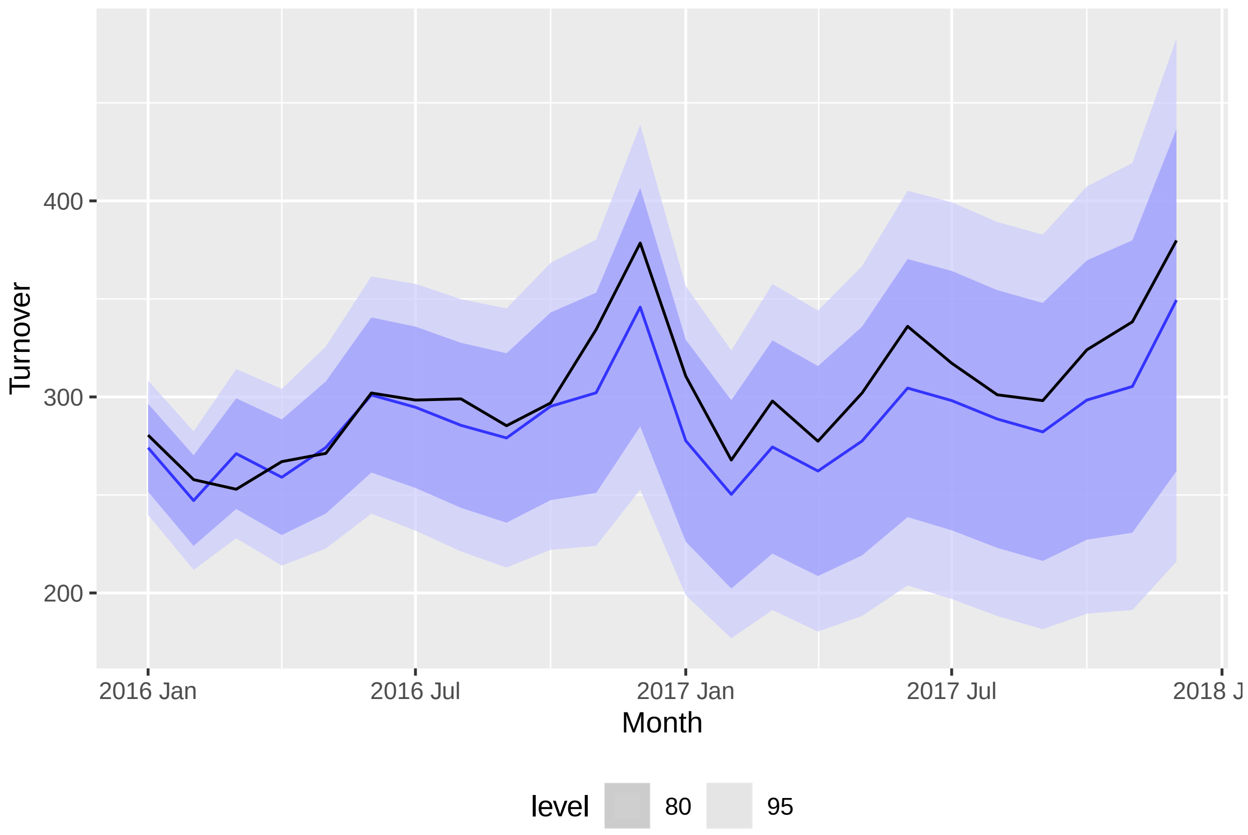
ETS: madm ARIMA: 1 + ARIMA(1,0,2)(2,1,1)[12]

ETS: madm

```
## Series: Turnover
## Model: ETS(M,Ad,M)
##   Smoothing parameters:
##     alpha = 0.5496792
##     beta  = 0.004422253
##     gamma = 0.1249318
##     phi   = 0.9799988
##
##   Initial states:
##         l        b        s1        s2        s3       s4       s5       s6
##  59.55503 0.567321 0.9915552 0.9005886 0.9345242 1.200707 1.069526 1.031098
##         s7        s8        s9       s10      s11       s12
##  0.9363328 0.9844062 0.9881665 0.9682453 1.060349 0.9345016
##
##   sigma^2:  0.0041
##
##      AIC     AICc      BIC
## 4212.669 4214.441 4284.739
```



```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 madm      23.7   0.00130
```

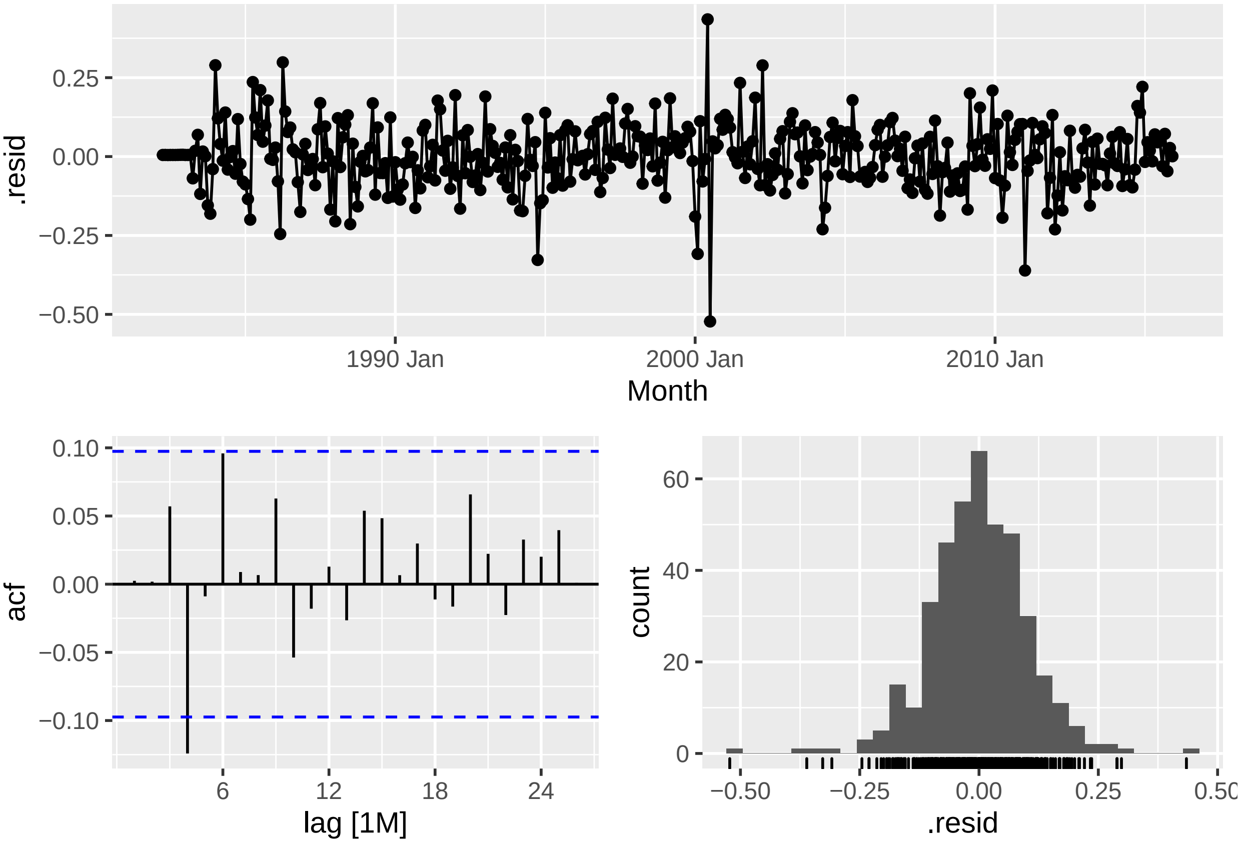


```
## # A tsibble: 24 x 8 [1M]
## # Key:       State, Industry, .model [1]
##    State  Industry   .model      Month     Turnover .mean                  `80%`
##    <chr>  <chr>      <chr>       <mth>       <dist> <dbl>                 <hilo>
##  1 Victo~ Furniture~ "ETS(Tu~ 2018 Jan  N(307, 369)  307. [282.0660, 331.2740]80
##  2 Victo~ Furniture~ "ETS(Tu~ 2018 Feb  N(275, 393)  275. [249.4218, 300.2315]80
##  3 Victo~ Furniture~ "ETS(Tu~ 2018 Mar  N(299, 579)  299. [267.6938, 329.3839]80
##  4 Victo~ Furniture~ "ETS(Tu~ 2018 Apr  N(289, 651)  289. [255.9626, 321.3451]80
##  5 Victo~ Furniture~ "ETS(Tu~ 2018 May  N(305, 850)  305. [267.6791, 342.3960]80
##  6 Victo~ Furniture~ "ETS(Tu~ 2018 Jun N(335, 1175)  335. [291.1000, 378.9601]80
##  7 Victo~ Furniture~ "ETS(Tu~ 2018 Jul N(326, 1256)  326. [280.5879, 371.4208]80
##  8 Victo~ Furniture~ "ETS(Tu~ 2018 Aug N(317, 1321)  317. [269.9859, 363.1276]80
##  9 Victo~ Furniture~ "ETS(Tu~ 2018 Sep N(309, 1389)  309. [261.1982, 356.7256]80
## 10 Victo~ Furniture~ "ETS(Tu~ 2018 Oct N(327, 1709)  327. [274.5058, 380.4741]80
## # ... with 14 more rows, and 1 more variable: 95% <hilo>
```

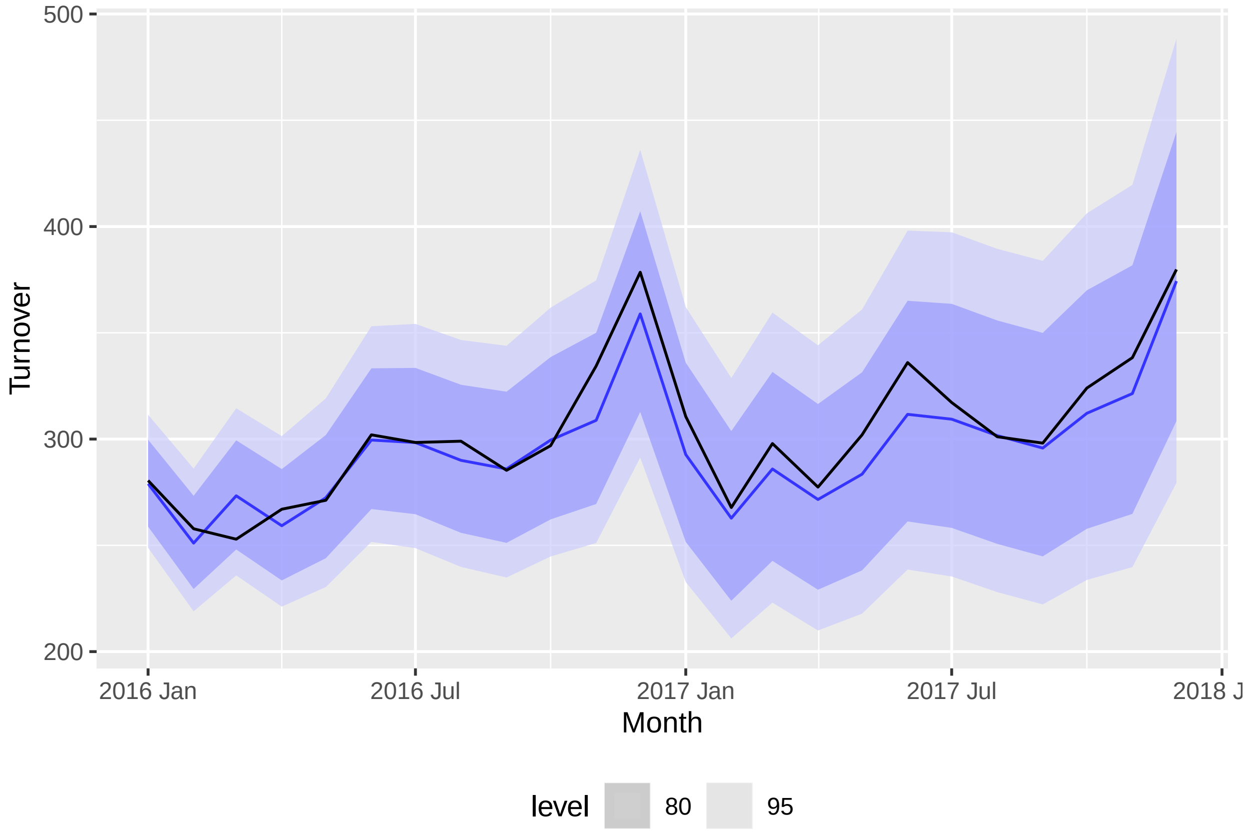
For the estimates of ETS model, we have a very small beta value which is close to 0, so we would expect the trend is quite constant. However, there would have some flexibility changing over time. And we have a large gamma, suggesting that the seasonality is changing over time. The coefficient estimates of phi is very close to one, suggesting not a lot of damping is required. Also, the alpha value is quite large, implying that the level changes quite rapidly.

ARIMA: 1 + ARIMA(1,0,2)(2,1,1)[12]

```
## Series: Turnover
## Model: ARIMA(1,0,2)(2,1,1)[12] w/ drift
## Transformation: box_cox(Turnover, lambda)
##
## Coefficients:
##          ar1      ma1      ma2    sar1     sar2     sma1  constant
##       0.9768  -0.3535  -0.1113  0.1161  -0.0617  -0.7979    0.0017
## s.e.  0.0120   0.0530   0.0562  0.0763   0.0677   0.0624    0.0006
##
## sigma^2 estimated as 0.01102:  log likelihood=326
## AIC=-635.99   AICc=-635.62   BIC=-604.2
```



```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 auto_arima     34.5    0.0230
```



```
## # A tsibble: 24 x 8 [1M]
## # Key:       State, Industry, .model [1]
##    State  Industry .model    Month          Turnover .mean                    `80%`
##    <chr>  <chr>    <chr>     <mth>            <dist> <dbl>                   <hilo>
##  1 Victo~ Furnitu~ "ARIM~ 2016 Jan t(N(7.7, 0.011))  279. [258.8128, 299.7078]80
##  2 Victo~ Furnitu~ "ARIM~ 2016 Feb t(N(7.5, 0.015))  251. [229.4675, 273.3138]80
##  3 Victo~ Furnitu~ "ARIM~ 2016 Mar t(N(7.7, 0.018))  273. [248.0696, 299.4050]80
##  4 Victo~ Furnitu~ "ARIM~ 2016 Apr t(N(7.6, 0.021))  259. [233.5045, 285.8797]80
##  5 Victo~ Furnitu~ "ARIM~ 2016 May t(N(7.7, 0.023))  272. [243.9672, 301.9034]80
##  6 Victo~ Furnitu~ "ARIM~ 2016 Jun t(N(7.9, 0.025))  300. [267.0780, 333.2854]80
##  7 Victo~ Furnitu~ "ARIM~ 2016 Jul t(N(7.9, 0.028))  298. [264.6462, 333.4775]80
##  8 Victo~ Furnitu~ "ARIM~ 2016 Aug  t(N(7.8, 0.03))  290. [255.8983, 325.5544]80
##  9 Victo~ Furnitu~ "ARIM~ 2016 Sep t(N(7.8, 0.032))  286. [251.1789, 322.2894]80
## 10 Victo~ Furnitu~ "ARIM~ 2016 Oct t(N(7.9, 0.034))  300. [262.2109, 338.5619]80
## # ... with 14 more rows, and 1 more variable: 95% <hilo>
```

For the estimates of ARIMA model, we can see that it has a very high ar1 coefficient estimates, it can capture the slowly decaying pattern in the ACF plot.

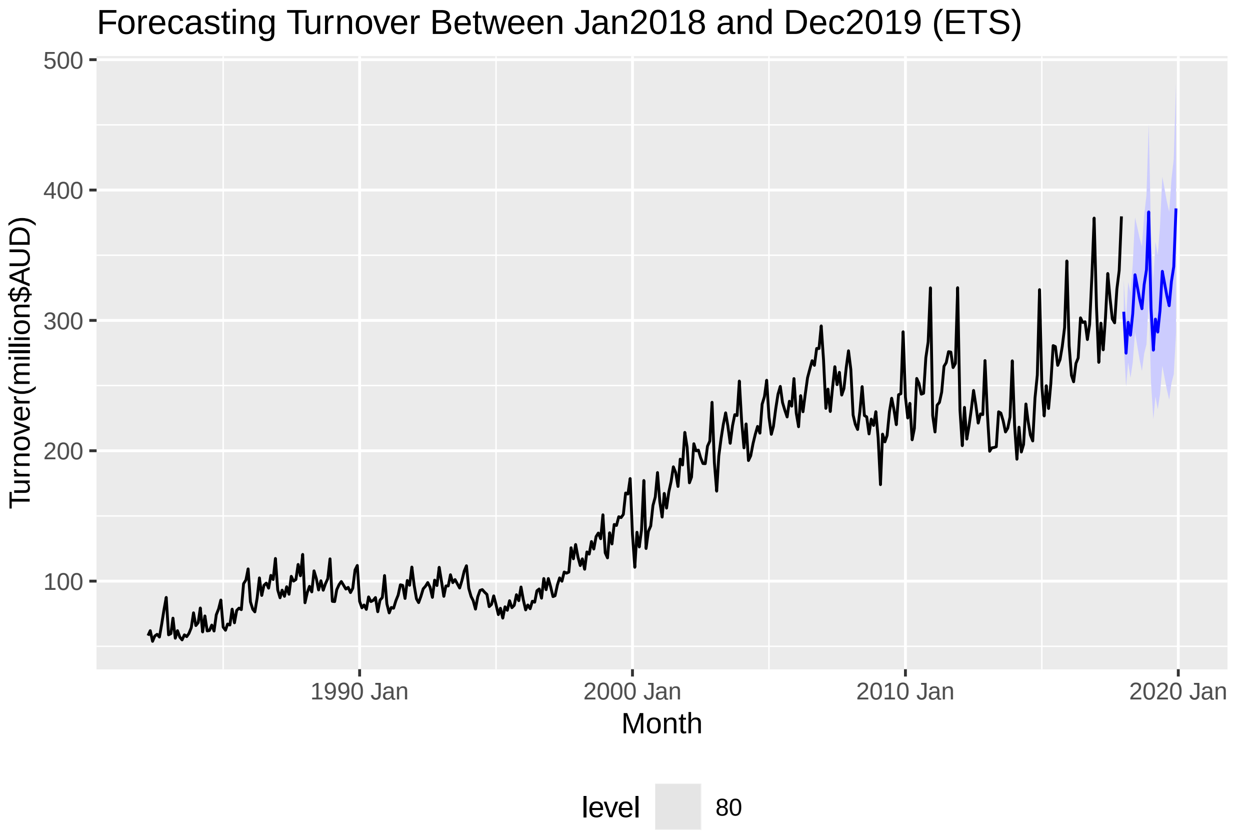
For the reisduals diagnostics, from the ljung-Box test, both models are also having a p-value lower than 0.05, which means we are going to reject the null and conclude that our erros are not white noise. This conclusion is aligned with the ACF plot of the residuals as we have some significant lags in the ACF plot. From ETS model, we can see that there is 4 significant lags. For ARIMA model, there is one significant lag at lag 4 and one lag which is barely significant at lag 6. Both of the models doesn't seems like it has constant variance, the variance is changing ove time.

Comparison of the results from each of your preferred models.

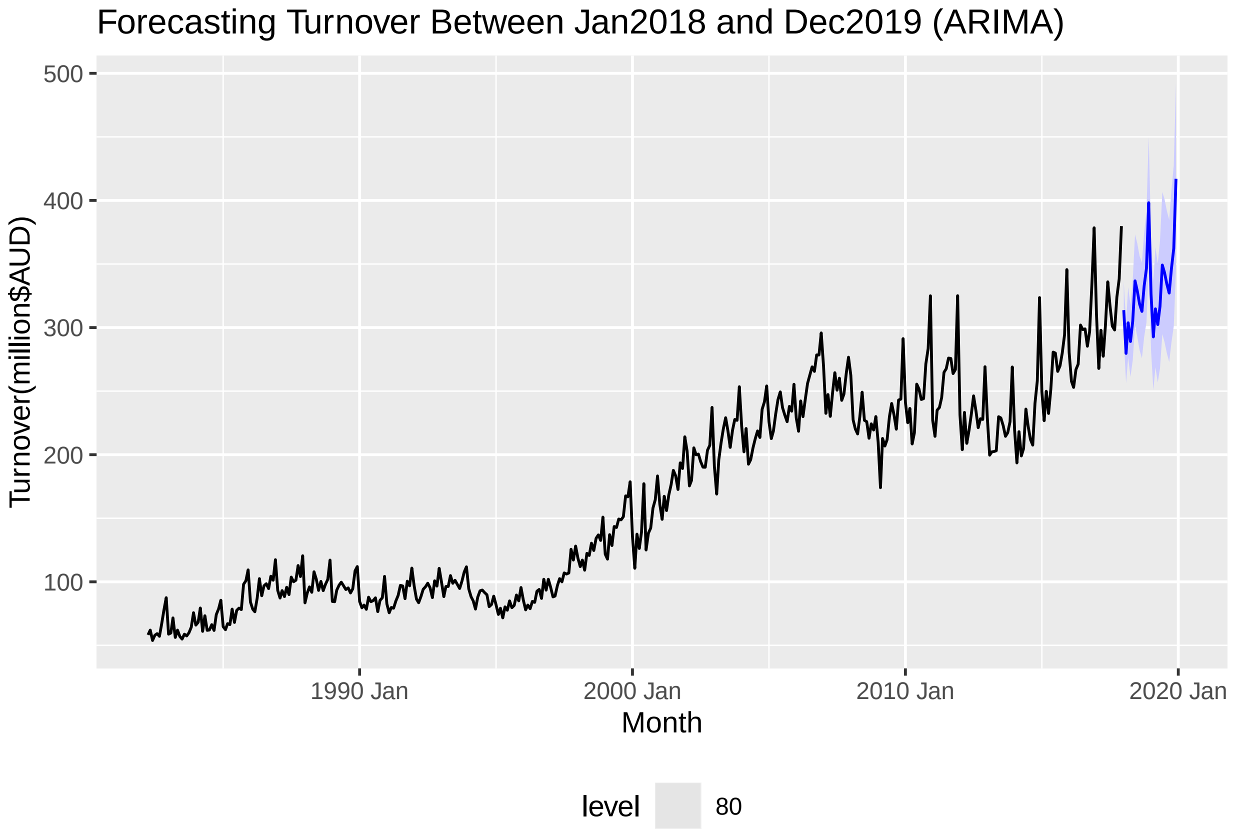
ARIMA model has provide a better point forecast than ETS model comparing with the test set data. ARIMA model is having a narrower prediction interval than ETS model. Which means that we have less unsertainty about the predictions.

However, as our residuals are not white noise, our prediction interval could be baised and we might not truly trust it.

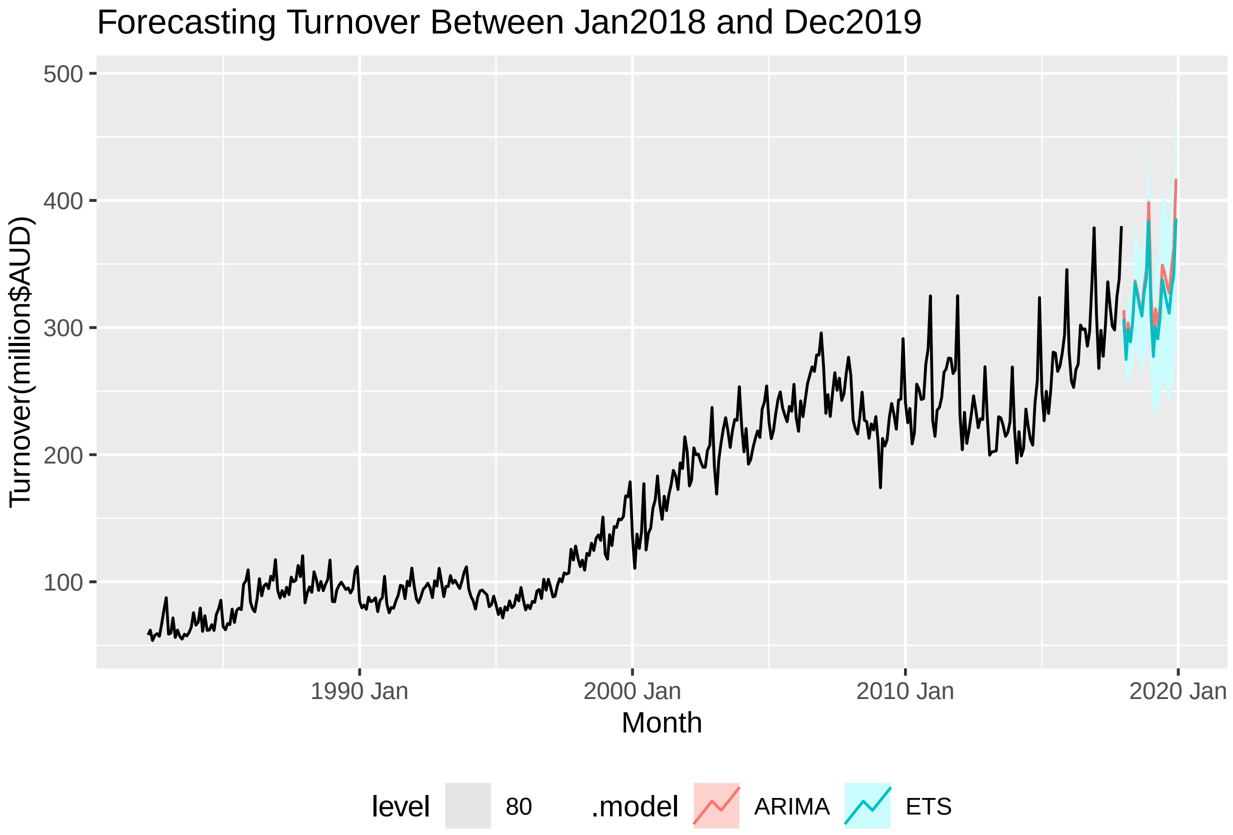
Apply your two chosen models to the full data set and produce out-of-sample point forecasts and 80% prediction intervals for each model for two years past the end of the data provided.



```
## # A tsibble: 24 x 7 [1M]
## # Key:       State, Industry, .model [1]
##    State  Industry     .model    Month     Turnover .mean                   `80%`
##    <chr>  <chr>        <chr>     <mth>       <dist> <dbl>                  <hilo>
##  1 Victo~ Furniture, ~ ETS    2018 Jan  N(307, 369)  307. [282.0660, 331.2740]80
##  2 Victo~ Furniture, ~ ETS    2018 Feb  N(275, 393)  275. [249.4218, 300.2315]80
##  3 Victo~ Furniture, ~ ETS    2018 Mar  N(299, 579)  299. [267.6938, 329.3839]80
##  4 Victo~ Furniture, ~ ETS    2018 Apr  N(289, 651)  289. [255.9626, 321.3451]80
##  5 Victo~ Furniture, ~ ETS    2018 May  N(305, 850)  305. [267.6791, 342.3960]80
##  6 Victo~ Furniture, ~ ETS    2018 Jun N(335, 1175)  335. [291.1000, 378.9601]80
##  7 Victo~ Furniture, ~ ETS    2018 Jul N(326, 1256)  326. [280.5879, 371.4208]80
##  8 Victo~ Furniture, ~ ETS    2018 Aug N(317, 1321)  317. [269.9859, 363.1276]80
##  9 Victo~ Furniture, ~ ETS    2018 Sep N(309, 1389)  309. [261.1982, 356.7256]80
## 10 Victo~ Furniture, ~ ETS    2018 Oct N(327, 1709)  327. [274.5058, 380.4741]80
## # ... with 14 more rows
```

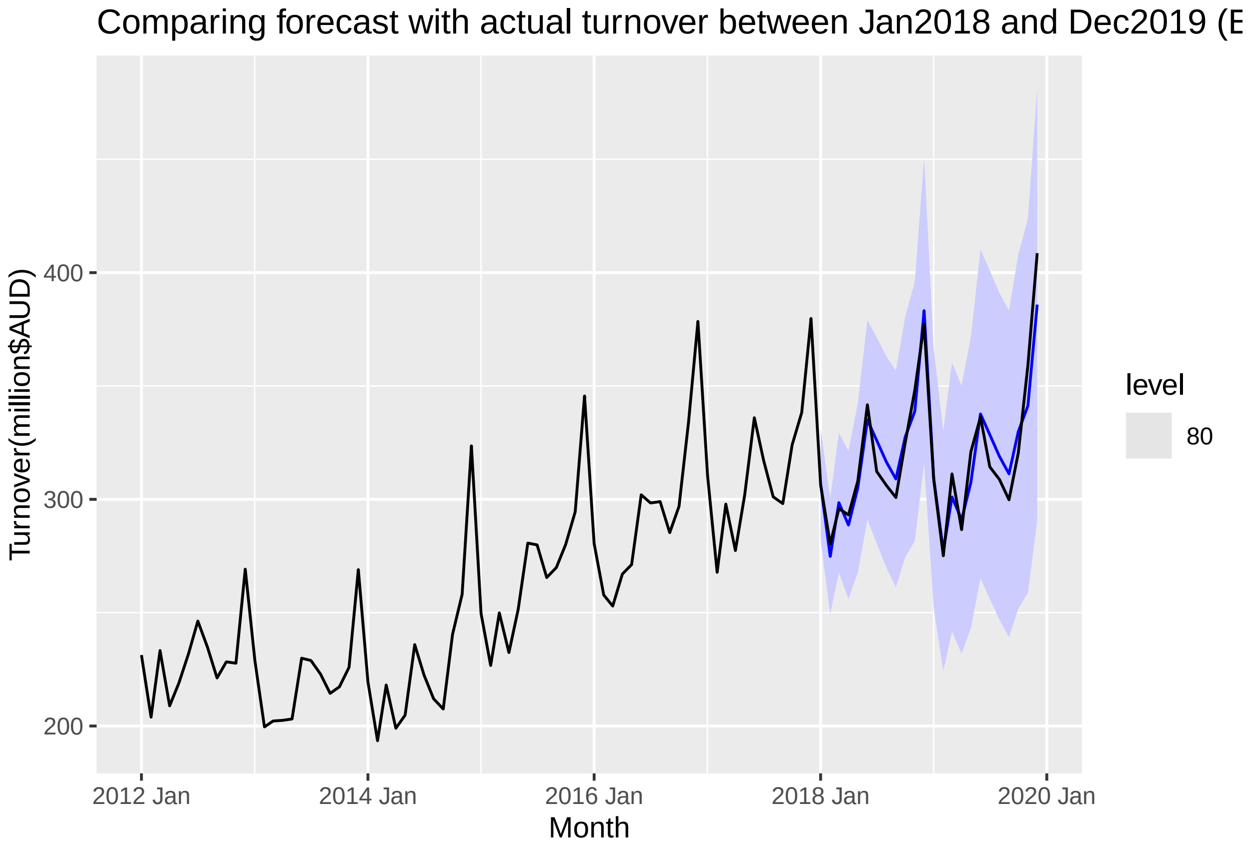


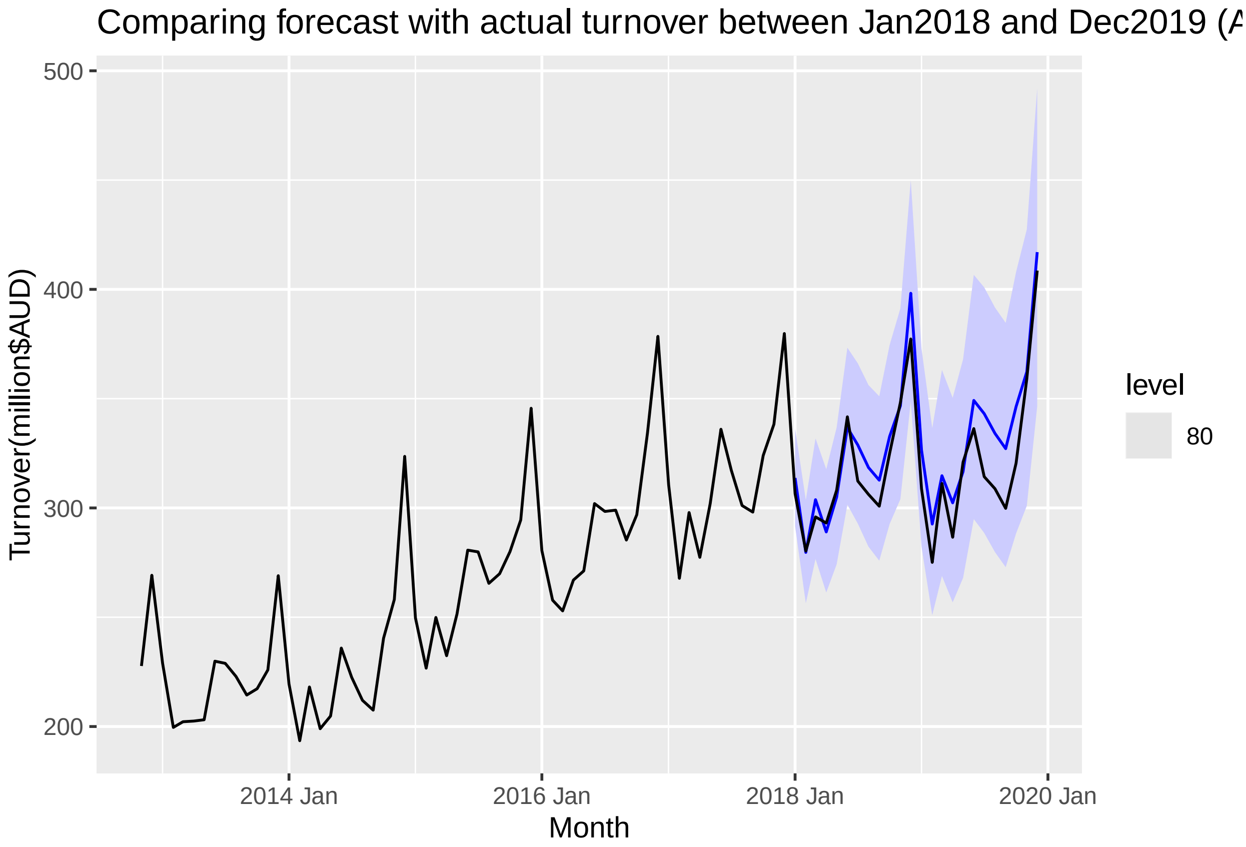
```
## # A tsibble: 24 x 7 [1M]
## # Key:       State, Industry, .model [1]
##    State  Industry .model    Month        Turnover .mean                    ‘80%‘
##    <chr>  <chr>    <chr>     <mth>          <dist> <dbl>                   <hilo>
##  1 Victo~ Furnitu~ ARIMA  2018 Jan   t(N(8, 0.011))  314. [291.6506, 336.3218]80
##  2 Victo~ Furnitu~ ARIMA  2018 Feb t(N(7.7, 0.015))  280. [256.3918, 303.6000]80
##  3 Victo~ Furnitu~ ARIMA  2018 Mar t(N(7.9, 0.017))  304. [276.6536, 331.7871]80
##  4 Victo~ Furnitu~ ARIMA  2018 Apr  t(N(7.8, 0.02))  289. [261.3166, 317.7033]80
##  5 Victo~ Furnitu~ ARIMA  2018 May t(N(7.9, 0.022))  305. [274.2286, 336.8160]80
##  6 Victo~ Furnitu~ ARIMA  2018 Jun t(N(8.1, 0.024))  337. [301.5266, 373.3026]80
##  7 Victo~ Furnitu~ ARIMA  2018 Jul   t(N(8, 0.026))  329. [292.8871, 366.1870]80
##  8 Victo~ Furnitu~ ARIMA  2018 Aug   t(N(8, 0.029))  318. [282.2602, 356.2021]80
##  9 Victo~ Furnitu~ ARIMA  2018 Sep  t(N(7.9, 0.03))  313. [275.8907, 351.0825]80
## 10 Victo~ Furnitu~ ARIMA  2018 Oct t(N(8.1, 0.032))  333. [292.6747, 374.5468]80
## # ... with 14 more rows
```

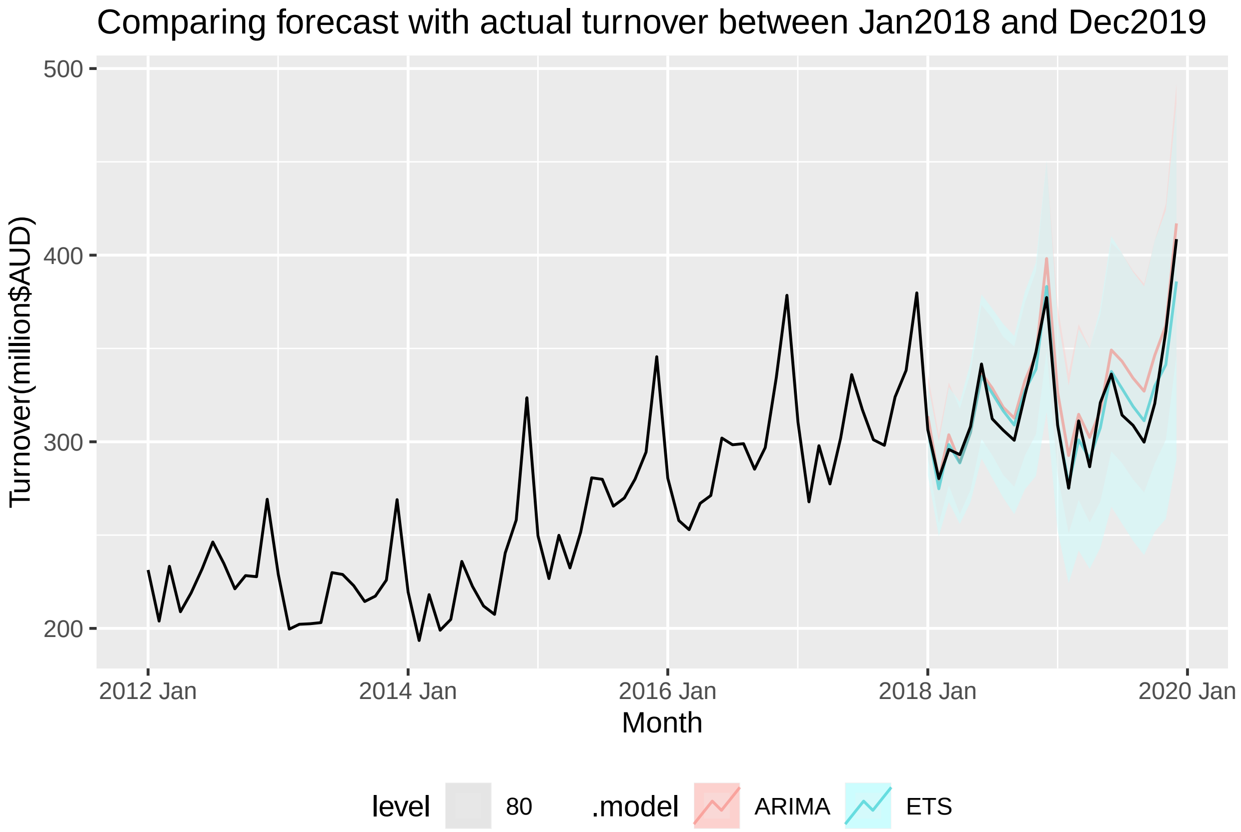


Comparing forecasts with the actual numbers from ABS website

```
## Finding URLs for tables corresponding to ABS series ID

## Attempting to download files from series ID , Retail Trade, Australia

## Downloading https://www.abs.gov.au/statistics/industry/retail-and-wholesale-trade/retail-trade-austra

## Extracting data from downloaded spreadsheets

## Tidying data from imported ABS spreadsheets
```


Comparing forecast with actual turnover between Jan2018 and Dec2019 (E
Turnover(million$AUD)
400
300
200
2012 Jan
2014 Jan
2016 Jan
2018 Jan
2020 Jan
Month
level
80


Comparing forecast with actual turnover between Jan2018 and Dec2019 (A
500
400
300
200
Turnover(million$AUD)
2014 Jan
2016 Jan
2018 Jan
2020 Jan
Month
level
80



```
## # A tibble: 2 x 12
##   .model State Industry .type      ME  RMSE   MAE    MPE  MAPE  MASE RMSSE  ACF1
##   <chr>  <chr> <chr>    <chr>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 ARIMA  Vict~ Furnitu~ Test  -10.5    14.8  12.1  -3.35   3.84 0.761 0.716 0.378
## 2 ETS    Vict~ Furnitu~ Test   -0.161  9.73  7.93 -0.171  2.43 0.500 0.470 0.249
```

By comparing the forecasts to the actual turnover from 2018 Jan to 2019 Dec, ETS model is doing a more accurate point forecast. And ETS model has a better test accuracy than ARIMA model. But both model is doing a pretty good job. And again, the ARIMA model is having a narrower prediction interval than the ETS model.

benefits and limitations of the models for your data.

Regarding on benefitsm for ETS model, we can do one step less as it can take care with non-stationary data. We need not to do any transformation of the reponse variable before modelling the data. For ARIMA model, there is more combinations of models to fit the data. Once we have checked how many MA process or AR process should be included in the model form the ACF and PACF plot, we can start try a mix of it, for example ARIMA(p=2,d=0,q=1),ARIMA(p=1,d=0,q=1). Also, both models can also produce a point forecast and forecast distributions capturing the uncertainty of the prediction. Both models can also produce a fairly close forecast capturing the trend and seasonlity comparing with the actual data.

However, both model can't pick up the correlated errors, our errors are not white noise. We might still have left some information in the residuals. And maybe we should choose other models to fit the data.