Time Series : Superstore dataset

Time Series final project

Group 5:

- Aniket W
- Simran C
- Vijay T
- Yamini R

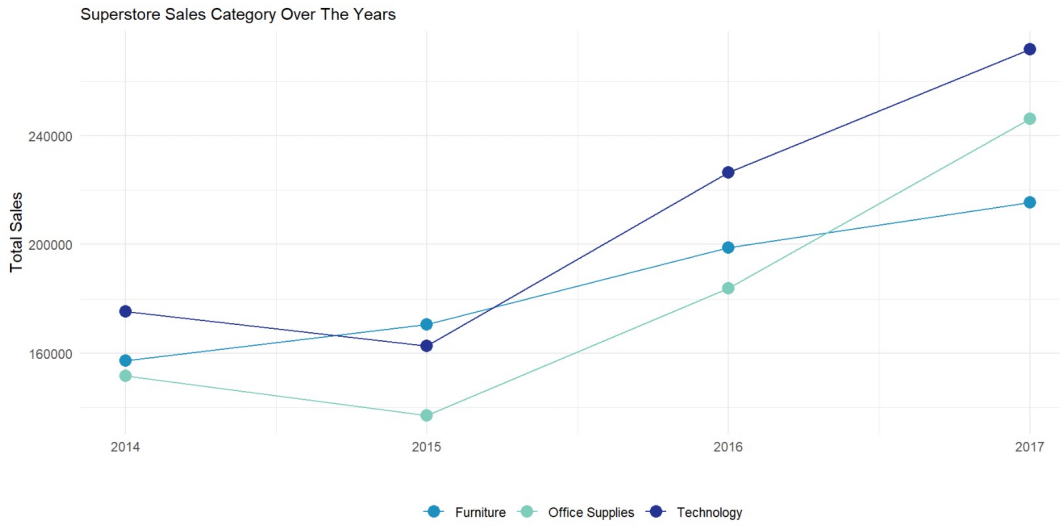
Overview

- Description of data
- Time series methodology application
- Fitting the models
- Evaluating the model
- Forecasting
- Conclusion.

Describing the Dataset

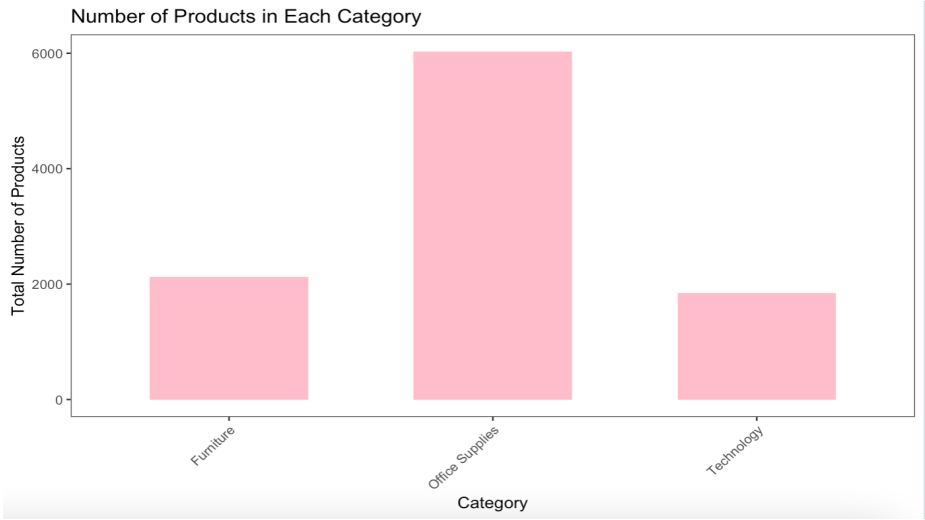
About the dataset

- Data taken from Kaggle :https://www.kaggle.com/code/abiodunonadeji/superstore-eda-and-visualization-with-r
- Retail dataset of a superstore in United states of 4 years (2014 : 2017)
- The dataset has 3 types of product category : Office Supplies, Furniture and Technology

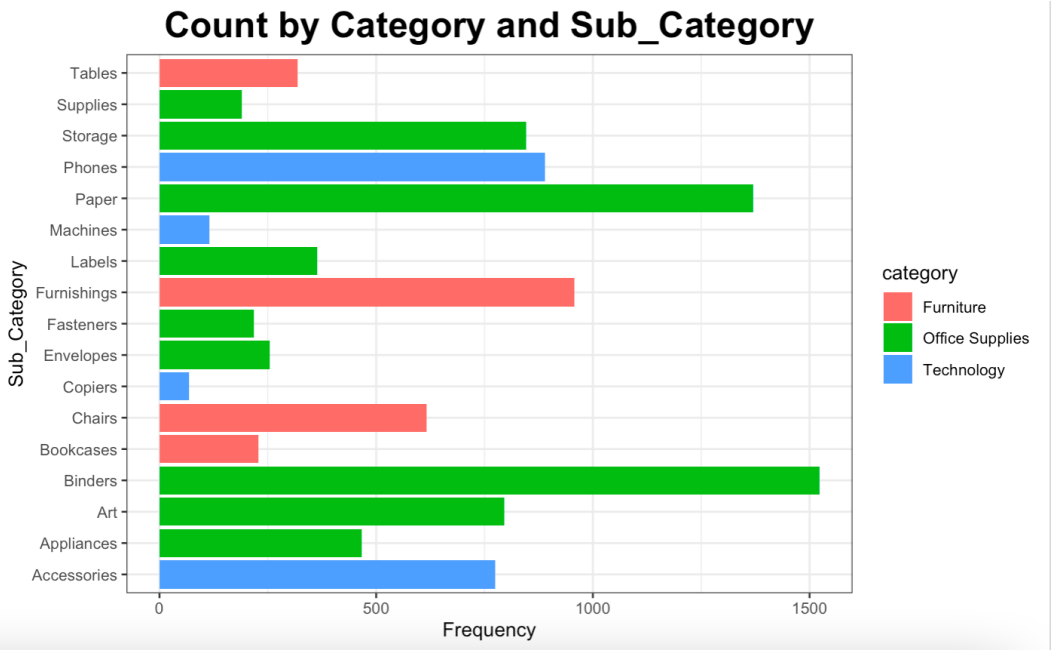


Data Preprocessing

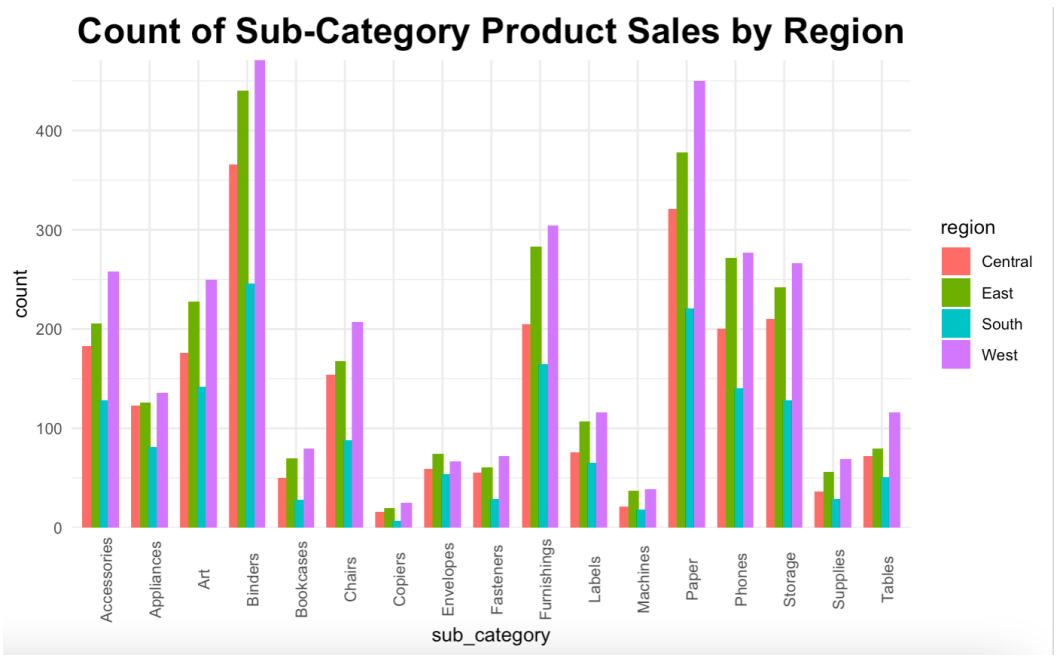
- Filtered Furniture to process further.
- Converted order date into months and year
- Checked the data for missing values.
- Furniture category is used for further analysis and forecasting the sales.
- Data was split into train and test (90:10)



Visualisation of the data

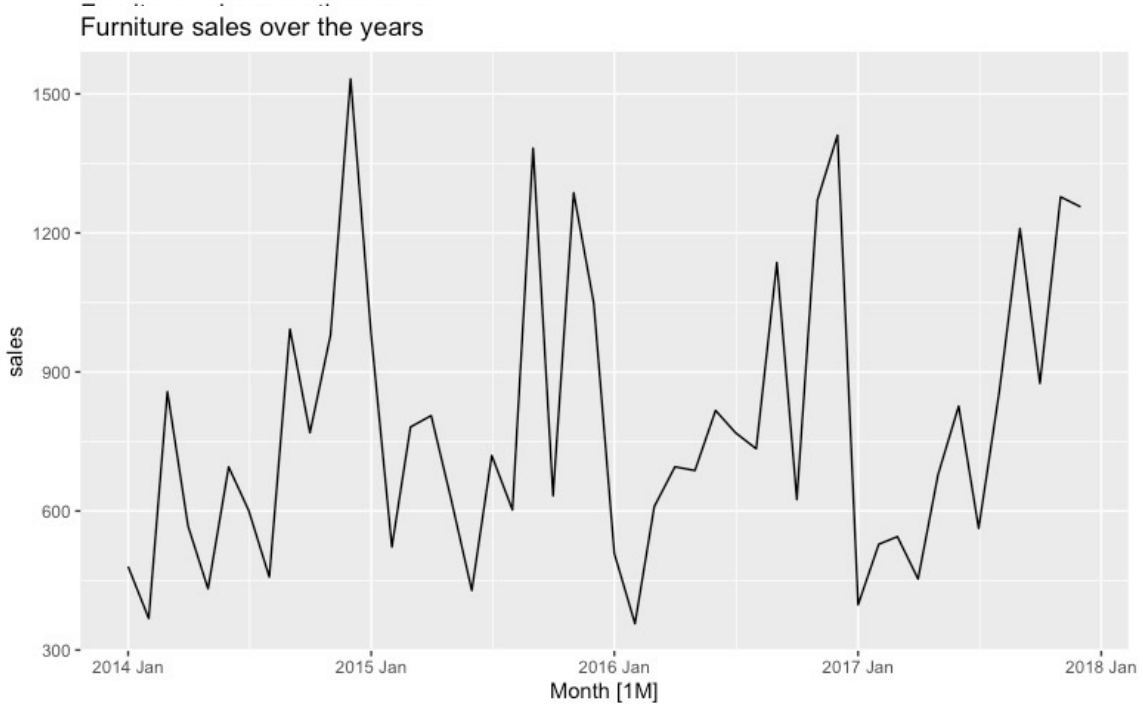


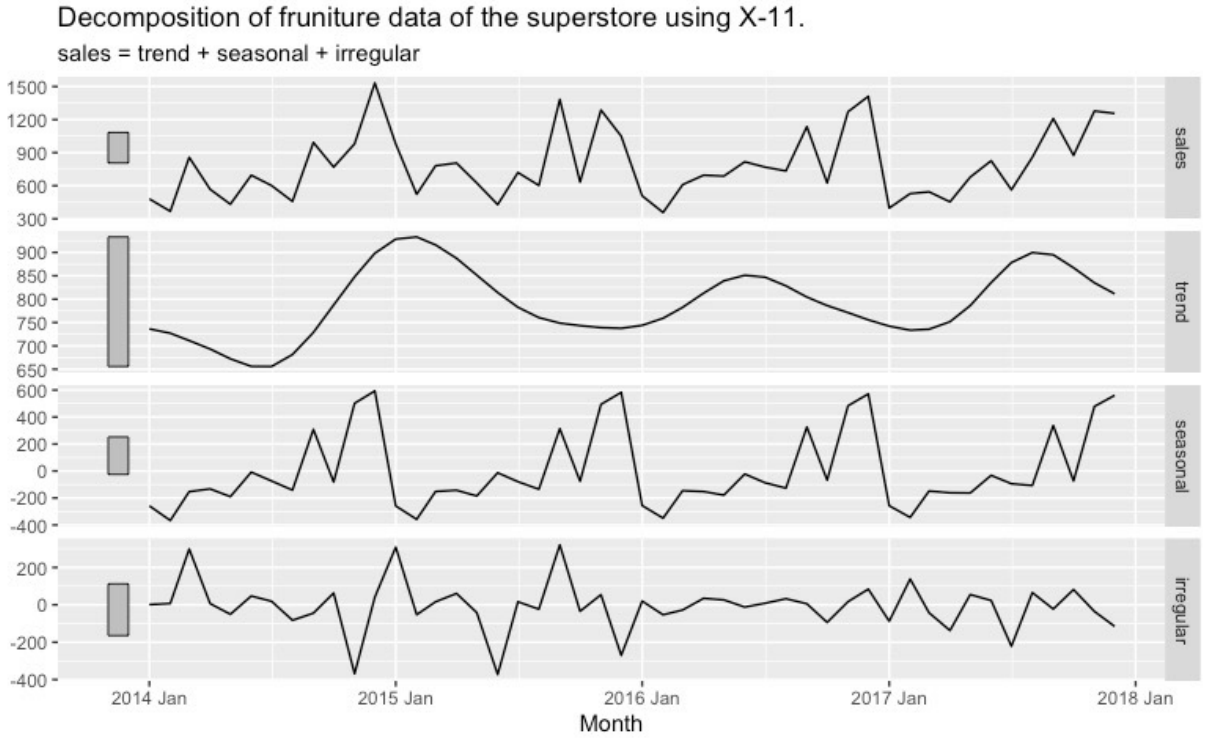
Furnishing has max sales in Furniture, Papers and binders in office supplies and Phones in technology.



People residing in Western part of United State tends to order more from superstore.

Time series Methodology: Understanding the data



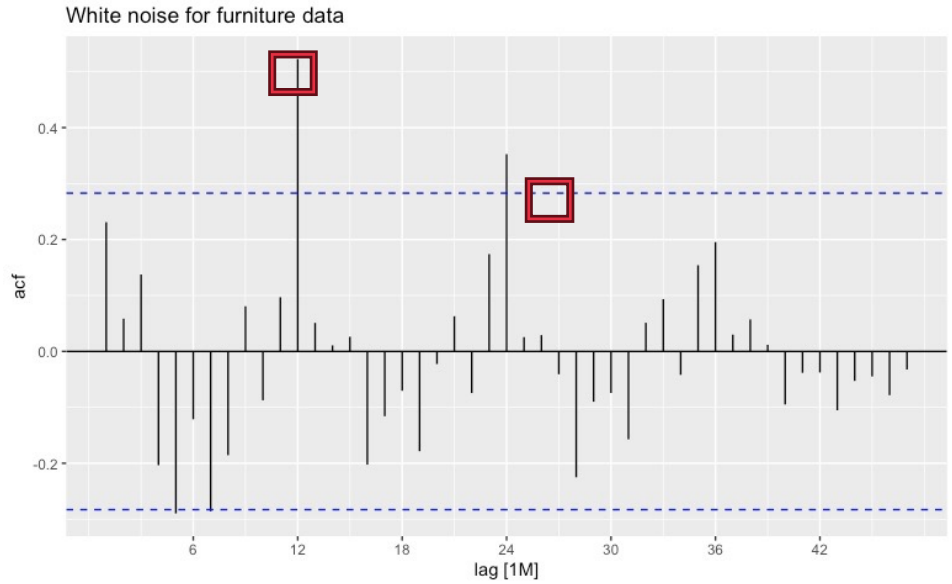


- **Trend:** The data does not exhibit a noticeable trend over time.
- **Seasonality:** A strong seasonal pattern is evident. The sales fluctuate in a consistent pattern, with distinct upward and downward trends.

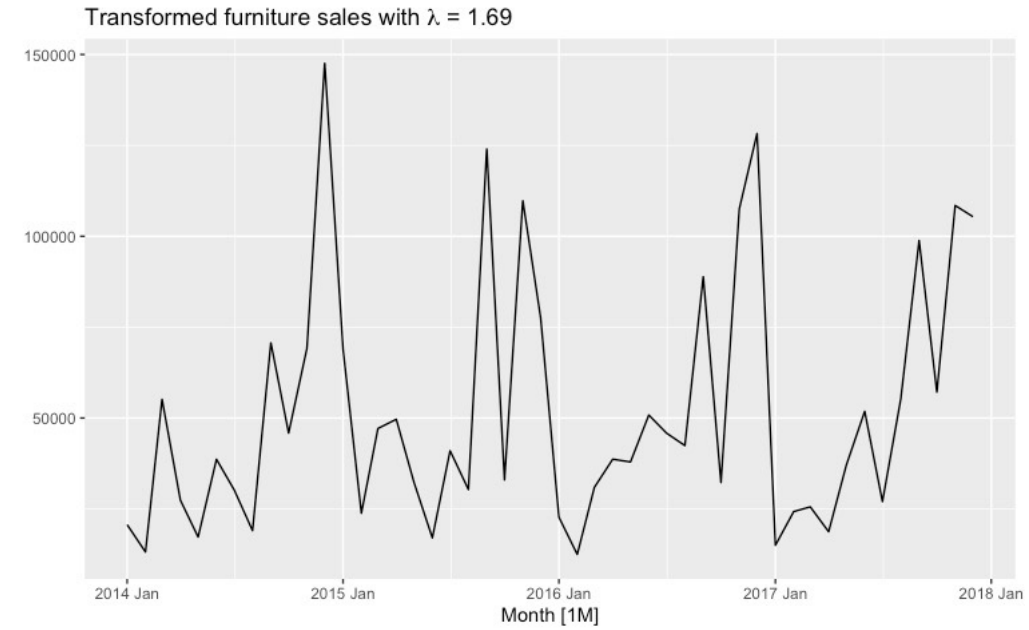
- Sales tend to be lower at the beginning of each year, particularly in February, while they tend to be higher towards the end of the year.

- The time series can be characterized as additive, as the seasonal pattern appears to be consistent across time.

Transformation of the data

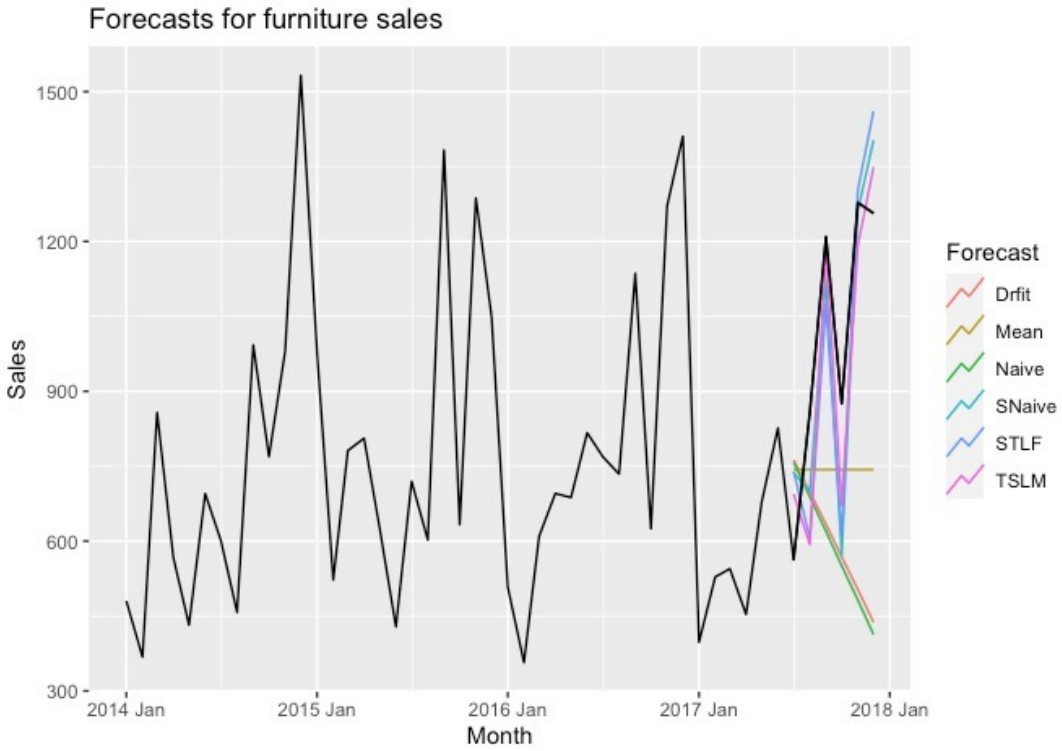


- The significance of the lag coefficients at 12 and 24 suggests a strong seasonal pattern in the data.
- These specific lags show a notable relationship with the target variable, indicating that the values at these time points are influential in predicting future values.
- Lag at 5 and 7 have negative values which indicate a consistent pattern of decreased sales at these intervals.
- These lags display a stronger negative correlation compared to other lags, suggesting a specific recurring pattern in the data.
- Additionally, the fact that 4 out of 48 autocorrelation coefficients lie outside the blue line indicates that the data does not exhibit white noise characteristics.



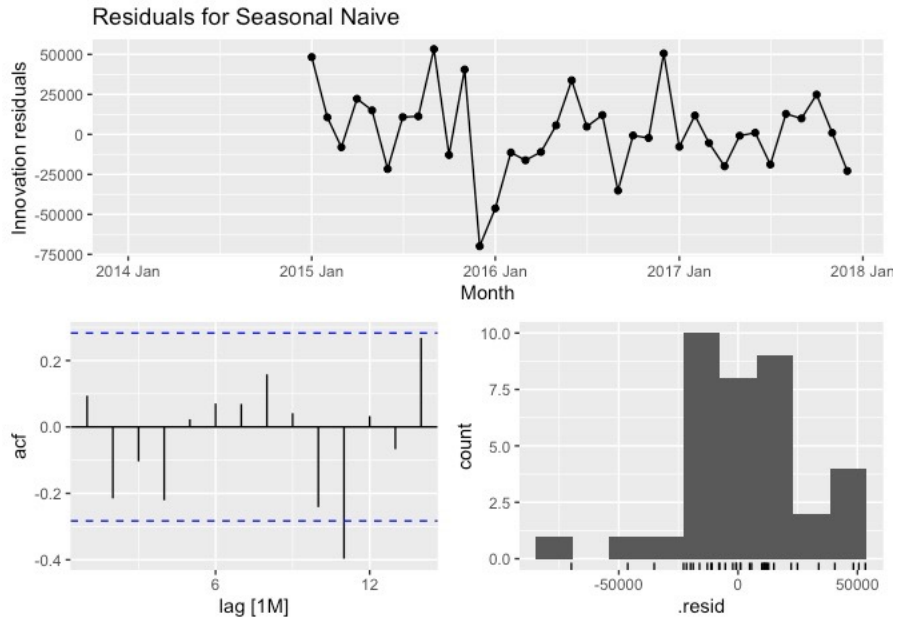
- Performed the Box-Cox transformation to stabilize the variance of a time series.
- A lambda value of 1 helps stabilize the variance of a time series with a positive skew.

Fitting and forecasting: Basic Models



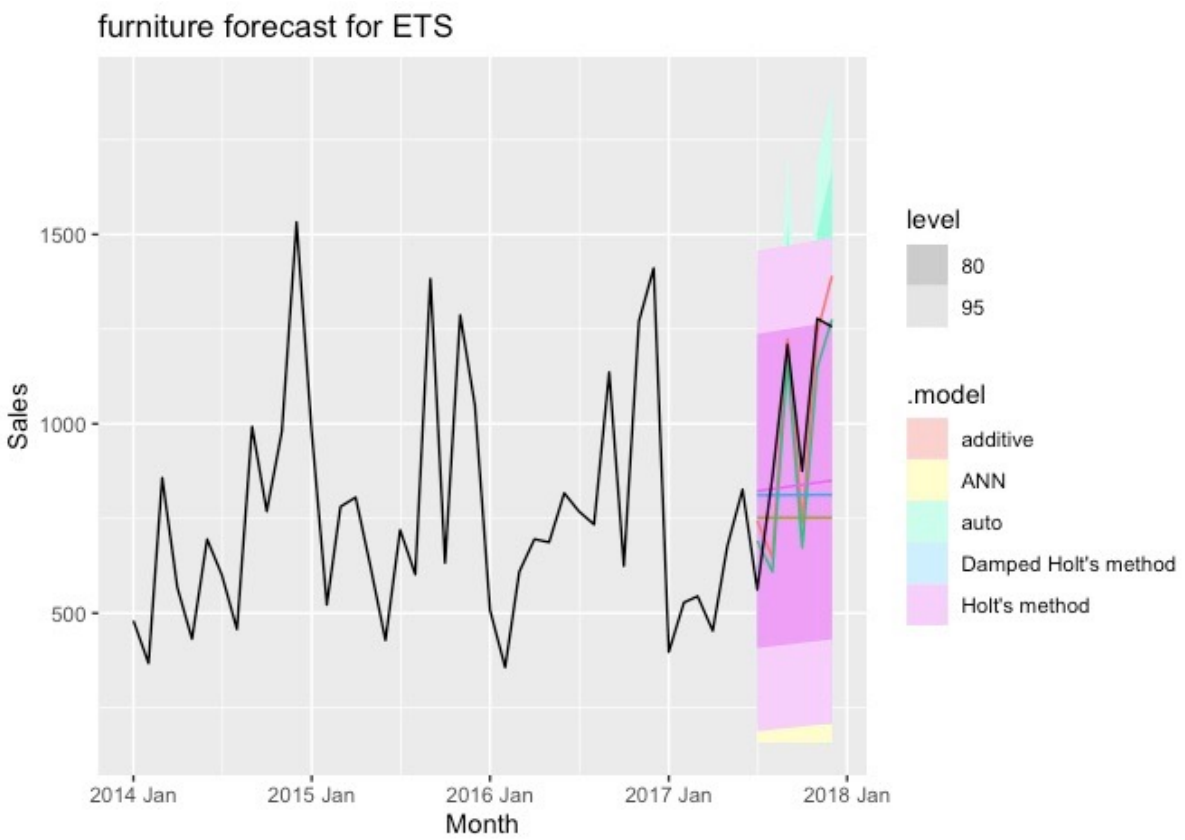

```
# A tibble: 6 × 10
  .model .type    ME  RMSE   MAE   MPE  MAPE
  <chr>  <chr> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Drfit  Test   406.  542.  473. 31.9   43.8
2 Mean   Test   263.  372.  324. 19.6   30.3
3 Naive  Test   421.  558.  486. 33.4   45.0
4 SNaive Test    40.8 172.  148.  3.18  17.4
5 STLF   Test    41.3 195.  177.  3.56  20.2
6 TSLM   Test    59.1 156.  134.  5.36  15.6
>
```

```
> furni_LJ_bx
# A tibble: 6 × 3
  .model lb_stat lb_pvalue
  <chr>    <dbl>      <dbl>
1 Drfit     31.3   0.00178
2 Mean      24.3   0.0184
3 Naive     31.3   0.00176
4 SNaive    15.7   0.206
5 STLF      24.1   0.0200
6 TSLM      18.1   0.113
```



- Performed ljung_box test to examine the presence of white noise in the residuals of the models.
- The Pvalue for SNaive, TSLM is > 0.05, which confirms that there is white noise in the residuals.
- Looking at accuracy table : MAPE is lowest for TSLM, however SNaive model has the lowest values for RMSE, MAE, MAPE, MASE, and RMSSE,
- The accuracy results suggest **Snaive is better forecasting performance compared to the other models.**

Fitting and forecasting: Exponential smoothening




```
> report(furni_fit_ES)
# A tibble: 5 × 9
  .model                 sigma2 log_lik   AIC  AICc
  <chr>                   <dbl>   <dbl> <dbl> <dbl>
1 auto                   0.0591   -287.  604.  622.
2 ANN                   91834.    -317.  641.  642.
3 additive              35480.    -288.  611.  636.
4 Holt's method        104744.    -319.  648.  650.
5 Damped Holt's method  95453.    -317.  643.  645.
```

```
> furni_forecast_ES
# A tibble: 5 × 10
  .model                 .type     ME  RMSE   MAE    MPE  MAPE
  <chr>                  <chr>  <dbl> <dbl> <dbl>  <dbl> <dbl>
1 additive               Test    9.28  139.  119. -0.147  14.6
2 ANN                    Test  254.    365.  317. 18.6    29.9
3 auto                   Test   79.3   152.  129.  6.90   15.0
4 Damped Holt's method   Test  195.    327.  278. 12.2    26.9
5 Holt's method          Test  170.    307.  257.  9.78   25.2
```

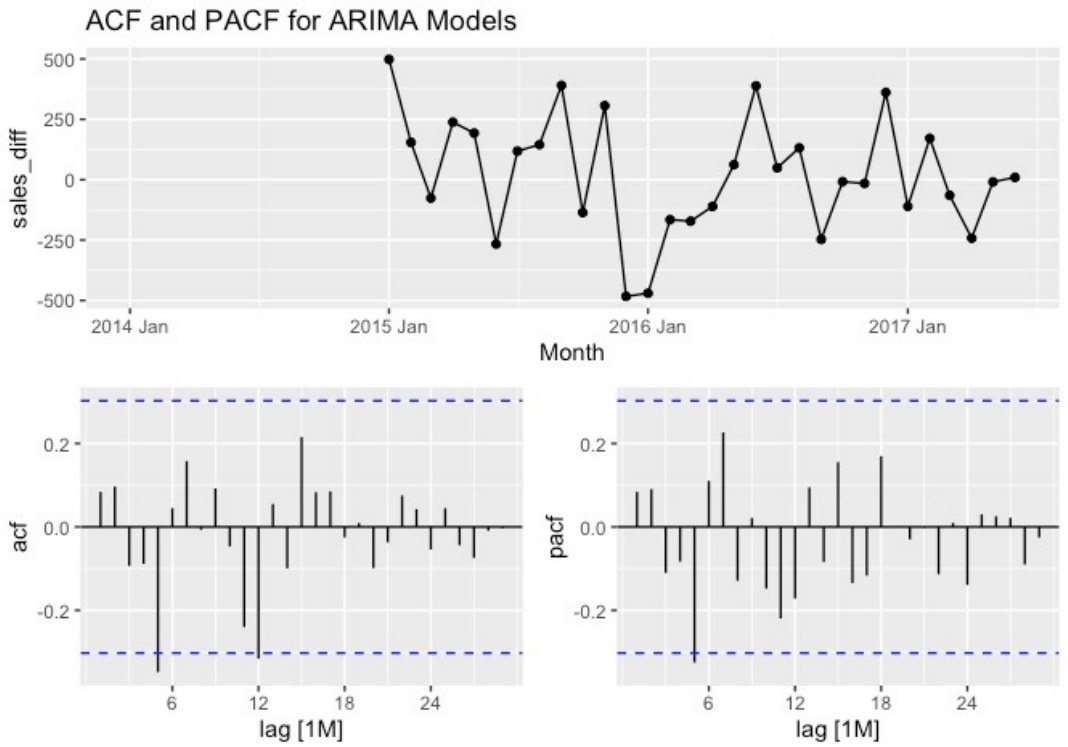
- 5 models of Exponential Smoothing (ES) were fitted to the data.
- The AICc is lowest for AUTO however the MAPE is lowest for Additive model.
- Although the Additive model had the lowest MAPE, the **AUTO model's overall performance was determined to be superior** based on its AICc value and its ability to capture the underlying patterns and dynamics of the data.

Fitting the models – ARIMA (1/3)

Check if data is stationary:

```
> furni_ts1_train %>% features(sales, unitroot_kpss)
# A tibble: 1 × 2
  kpss_stat kpss_pvalue
      <dbl>       <dbl>
1    0.0877         0.1
```

The p value is 0.1 which is greater than 0.05 and thus we accept the null hypothesis. **The data is stationary.**



Checking for differencing:

```
> #checking the order of differencing needed
> furni_ts1_train %>%
+   features(sales, unitroot_ndiffs)
# A tibble: 1 × 1
  ndiffs
   <int>
1      0
> #seasonal
> furni_ts1_train %>%
+   features(sales, unitroot_nsdiffs)
# A tibble: 1 × 1
  nsdiffs
    <int>
1       1
```

← Performing 1 seasonal difference on the data

Fitting the models – ARIMA (2/3)

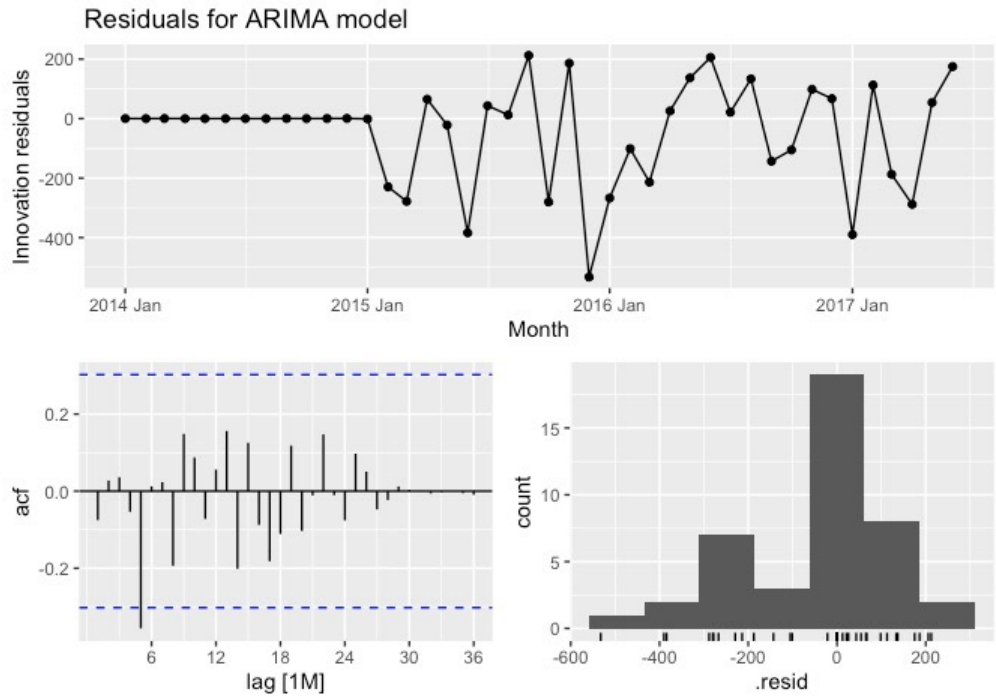
Following model for ARIMA was fit

```
> furni_ar_fit
# A mable: 1 x 4
                      arima100100               arima111110               arima000110                      auto
                          <model>                   <model>                   <model>                   <model>
1 <ARIMA(1,0,0)(1,0,0)[12] w/ mean> <ARIMA(1,1,1)(1,1,0)[12]> <ARIMA(0,0,0)(1,1,0)[12]> <ARIMA(0,0,0)(1,1,0)[12]>
```

```
> glance(furni_ar_fit)
# A tibble: 4 × 8
  .model      sigma2 log_lik   AIC  AICc   BIC
  <chr>        <dbl>   <dbl> <dbl> <dbl> <dbl>
1 arima100100 52482.    -290.  588.  589.  595.
2 arima111110 49829.    -200.  407.  409.  413.
3 arima000110 45097.    -204.  413.  413.  416.
4 auto        45097.    -204.  413.  413.  416.
```

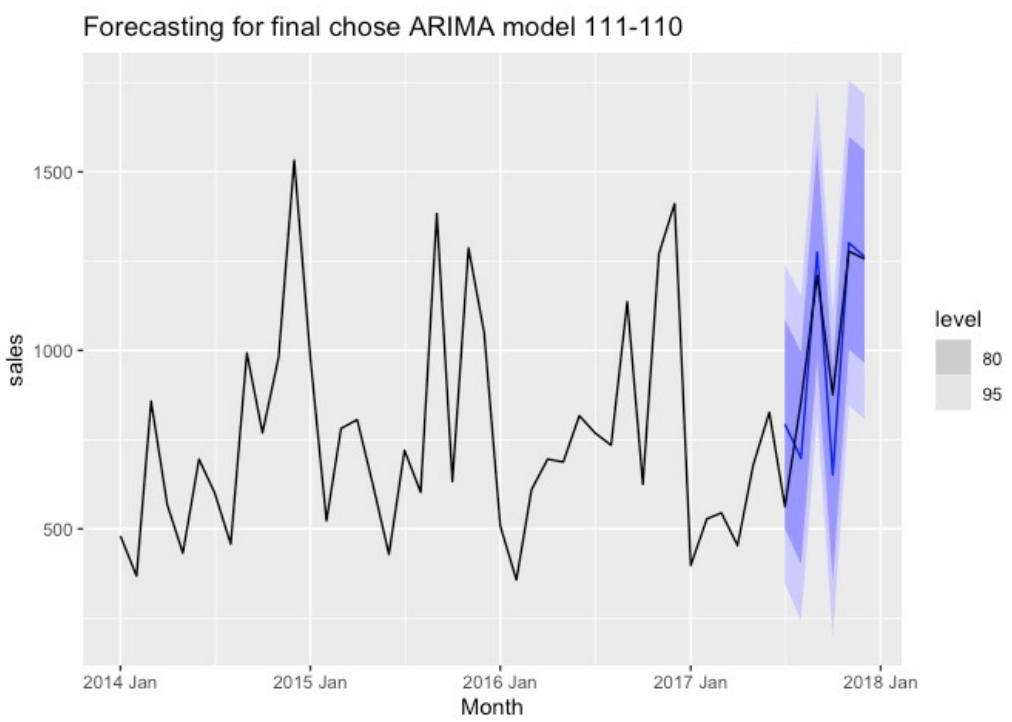
AICc is lowest for ARIMA 111,110

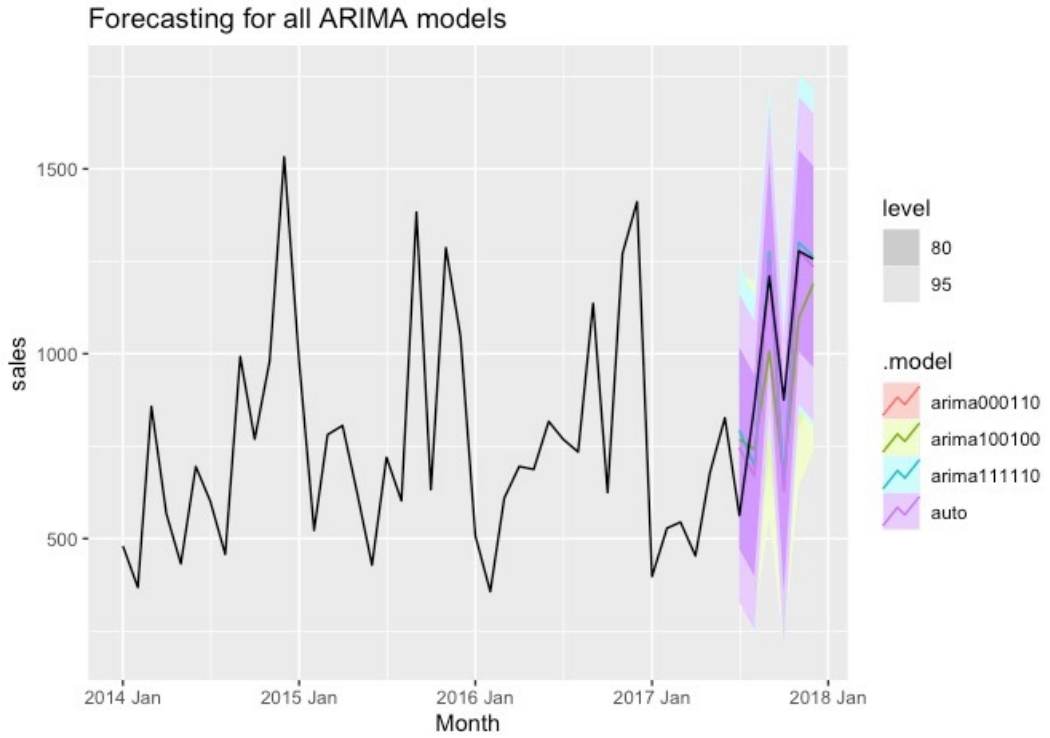
```
> augment(arima_final) %>%
+   features(.innov, ljung_box, lag = 12, dof=4)
# A tibble: 1 × 3
  .model      lb_stat lb_pvalue
  <chr>         <dbl>     <dbl>
1 arima111110    11.1     0.197
```



The Pvalue for the same is > 0.05, which confirms that there is white noise in the residuals.

Forecasting using ARIMA (3/3)

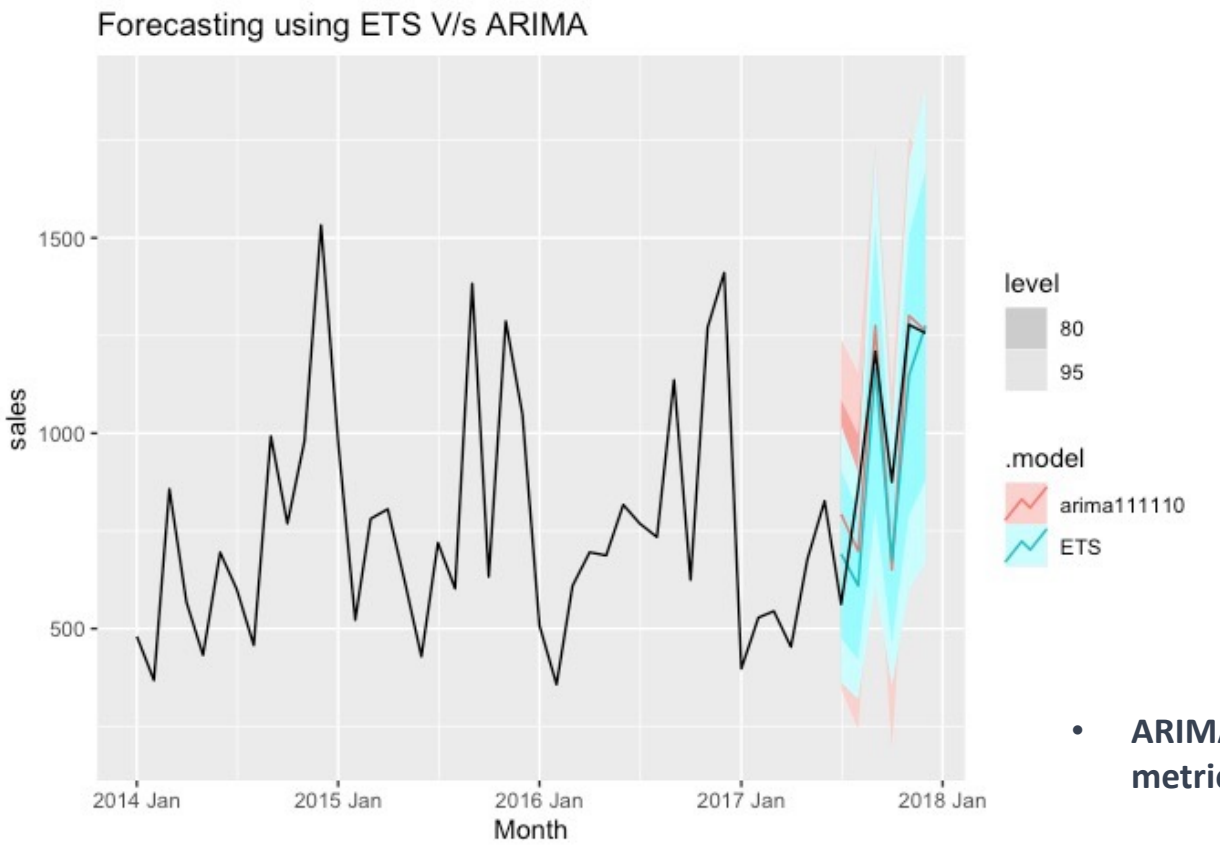




- The ARIMA (1,1,1)(1,1,0) model having the lowest MAPE indicates that it has the smallest average percentage error in its forecasts, making it the most accurate model among the ones considered.

- Taking both the AICc and MAPE into account, it can be concluded that the **ARIMA (1,1,1)(1,1,0) model is the preferred choice for forecasting the data.**

Conclusion: ETS V/s ARIMA




```
> glance(final_check)
# A tibble: 2 × 11
  .model          sigma2 log_lik   AIC  AICc   BIC
  <chr>            <dbl>   <dbl> <dbl> <dbl> <dbl>
1 arima111110 49829.       -200.  407.  409.  413.
2 ETS             0.0591   -287.  604.  622.  630.
```

```
> final_check_fc
# A tibble: 2 × 10
  .model       .type     ME  RMSE   MAE     MPE  MAPE
  <chr>        <chr>  <dbl> <dbl> <dbl>   <dbl> <dbl>
1 arima111110 Test    9.87  149.  118. -0.733   15.5
2 ETS         Test   79.3   152.  129.  6.90    15.0
```

- **ARIMA (1,1,1)(1,1,0) model exhibits the lowest values for both the AICc and MAPE metrics.**
- The ARIMA model, with its incorporation of autoregressive, moving average, and differencing components, proves effective in capturing temporal dependencies and accounting for both trend and seasonal variations.

Thank You