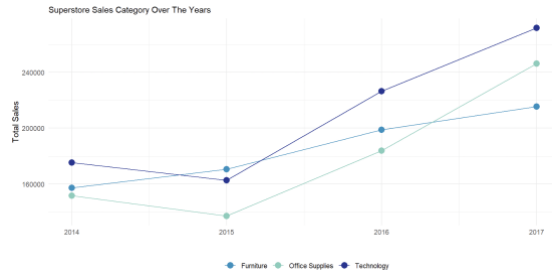
Time Series Final Project: Executive Summary
Superstore Dataset
Group 5: Aniket W, Simran C, Vijay T, Yamini R

Introduction:

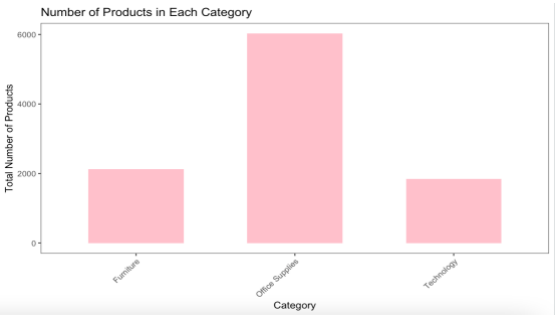
The aim of the project to forecast the sales for furniture segment of a superstore in the Unites States (USA) by utilizing time series analysis and forecasting techniques enabling the Superstore to make informed business decisions and optimize inventory management.



Description about the data:

The dataset offers a comprehensive overview of a Superstore in the USA. It encompasses various product categories, including furniture, office supplies, and technology, which are further categorized into sub-categories. The dataset includes detailed information about customer orders, shipment details, and sales performance.

The data is structured to provide insights at different levels of granularity, starting from the national level (USA) and further breaking down into states and cities. This allows for analysis of sales patterns and customer behavior across different regions.



Key attributes in the dataset include the order details (such as order IDs, order dates, and quantities), shipment information (such as shipping modes and dates), and customer-related data (including customer IDs and names). Additionally, the dataset provides profitability information through columns representing sales, quantity, discounts, and profit metrics.

Overall, this dataset serves as a valuable resource for conducting exploratory data analysis, understanding customer behavior, and informing strategic decision-making for the Superstore.

Data Preprocessing:

Data Filtering: We focused on the Furniture segment of the Superstore dataset, narrowing down our analysis to this specific product category.

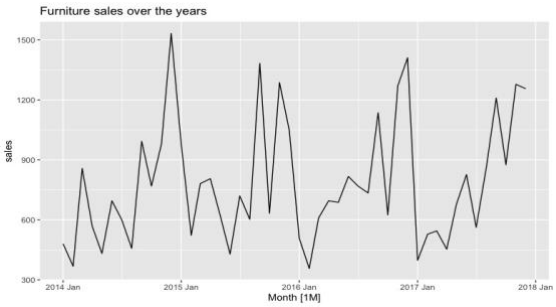
Date Conversion: We transformed the format of the Order Date column from a specific date to month and year values. This allowed us to aggregate and analyze the data at a higher level of granularity.

Handling Null Values: We carefully examined the dataset for any missing or null values.

Train-Test Split: To build a reliable forecasting model, we split the pre-processed dataset into a training and test set (90:10).

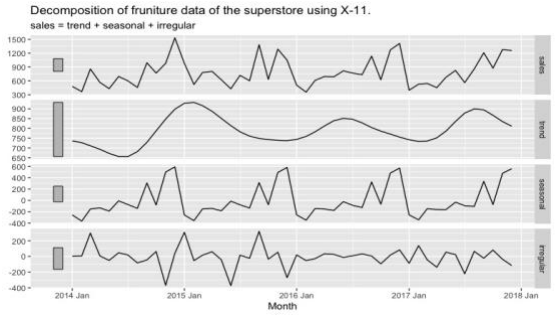
Understanding the Time Series data:

The auto plot function was used to visualize the patterns in the data. From the plot, it is apparent that the data exhibits seasonality.

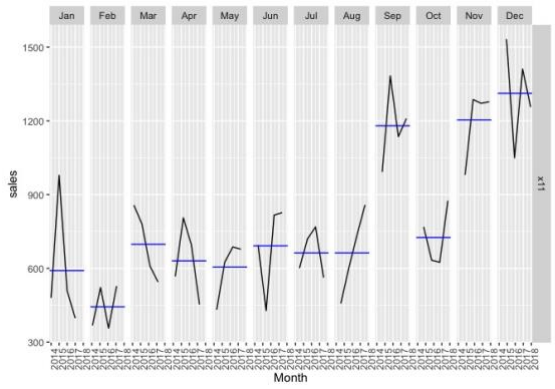


Determining the trend of the data is challenging based solely on the auto plot and thus to gain a deeper understanding of the time series pattern, the data was decomposed using the X_13ARIMA_seats method. The decomposition process involved separating the original time series into its constituent components, namely trend, seasonality, and residual. By decomposing the data, better insights were gained in understanding the individual components and their contributions to the overall pattern.

Time Series Final Project: Executive Summary
Superstore Dataset
Group 5: Aniket W, Simran C, Vijay T, Yamini R



The data does not exhibit a noticeable trend over time. However, 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.



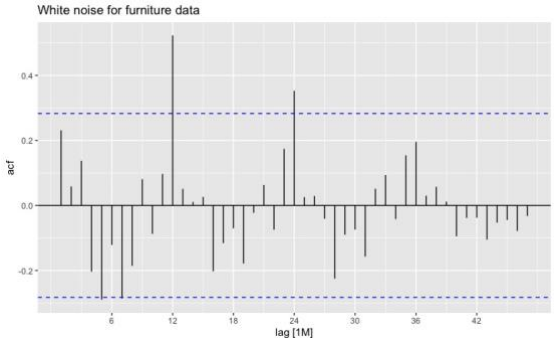
The features function was additionally used to validate the strengths in the data. The feat_stl function was further used to validate the features related to the trend, seasonality, and other characteristics of the data.

```
+   features(sales, feat_stl)
# A tibble: 1 × 9
  trend_strength seasonal_strength_year seasonal_peak_year seasonal_trough_year spikiness linearity curvature stl_e_acf1 stl_e_acf10
           <dbl>                  <dbl>              <dbl>               <dbl>      <dbl>     <dbl>     <dbl>      <dbl>       <dbl>
1          0.103                  0.804                  0                   2     231312.      173.     -56.7    -0.0229       0.238
# … with abbreviated variable names ¹linearity, ²curvature, ³stl_e_acf1, ⁴stl_e_acf10
```

In the output provided, the trend strength with a value of 0.103 suggests that there is a relatively weak trend component, and the data shows a gradual and consistent increase or decrease over time, but the magnitude of the trend is not very strong. The seasonal strength is with a value of 0.804 indicates a strong seasonal pattern in the data. This suggests that the time series exhibits a pronounced and regular seasonal fluctuation. The high value indicates that the seasonal component has a significant influence on the overall pattern of the data.

Verifying white noise in the data:

Using the autocorrelation (ACF) function to measures the correlation between observations at different lags, which will further provide insights if there is any temporal dependencies or patterns in the data.

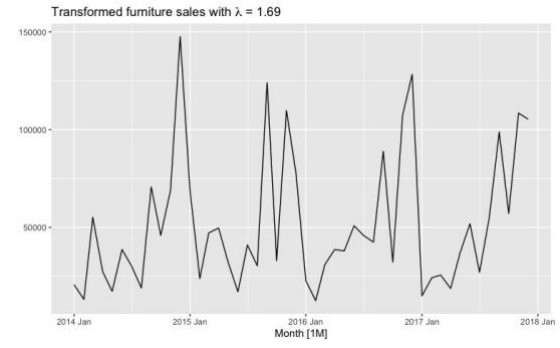


From the above plot we can see that there is significance of the lag coefficients at 12 and 24 further validates a strong seasonal pattern in the data. These specific lags show a notable relationship with the target variable, indicating that the values at these time points are influential in predicting future values. Furthermore, the negative values observed at lags 5 and 7 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 ACF coefficients lie outside the blue line indicates that the data does not exhibit white noise characteristics.

Transforming the Time Series data:

The Box-Cox transformation method was used to stabilize the variance of a time series. The lambda value derived from using the Box-cox method is 1.69, the value indicates a logarithmic transformation, which will help stabilize the variance of a time series with a positive skew.

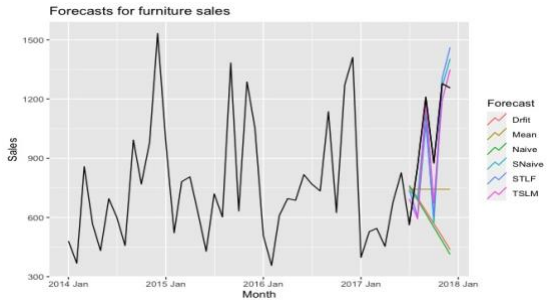


Time Series Final Project: Executive Summary
Superstore Dataset
Group 5: Aniket W, Simran C, Vijay T, Yamini R

Fitting the Models and Forecasting:

1. Basic Models

Post understanding the trends and patterns in the data, transforming the data, several basic forecasting models were fitted to the data. These models include Mean, Naïve, SNaive, Drift, STLF and TSLM. Fitting different models to the data allows for comparison and evaluation of their performance in terms of accuracy and ability to capture the underlying patterns and dynamics of the time series.

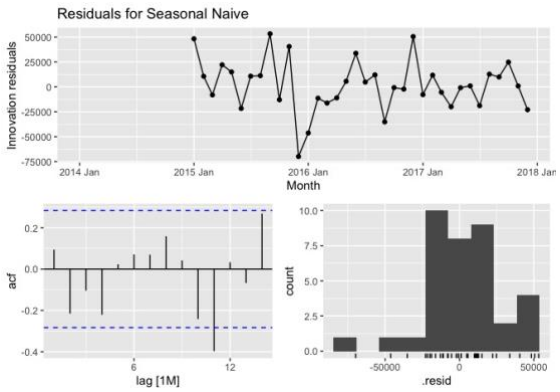


Looking at the forecast auto plot, TSLM and SNaive are the models which fit well to the actual data. To validate the same, accuracy function was used to compare between their forecasted values and the actual data.

```
# 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
>
```

The results of the accuracy assessment for different models (Drift, Mean, Naive, SNaive, STLF, and TSLM) on the test data are shown in the above table. TSLM model has the lowest MAPE, however SNaive model has the lowest values for RMSE, MAE, MAPE, MASE, and RMSSE, suggesting that it performs the best among the considered models.

Evaluating residuals of the SNaive model using ljung_box test and to examine the presence of white noise in the residuals of the models. The Pvalue for SNaive is greater than 0.05 and thus we accept the null hypothesis and conclude that there is white noise in the residuals.



The forecasting plot, accuracy test, examining residuals suggest that SNaive is better forecasting performance compared to the other models.

2. Exponential Smoothening:

Exponential smoothing another popular forecasting method was used to predict future values in time series data. Exponential smoothening model is helpful in handling seasonal data because it can capture both the trend and seasonal patterns in the data. This model gives more importance to recent data and less weight to older observations and adapt to changes in the data over time. The techniques used was auto, ANN, Additive, Holts method and Damped Holts method.

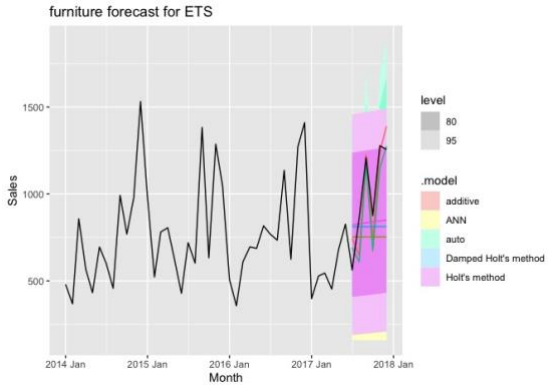
```
> report(furni_fit_ES)
# A tibble: 5 × 9
  .model                sigma2 log_lik   AIC  AICc
  <chr>                  <dbl>   <dbl> <dbl> <dbl>
1 auto                  0.0591   -287.  604.  622.
2 ANN                  91834.    -317.  641.  642.
3 additive             35480.    -288.  611.  636.
4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

The above table provides a summary of the fitted exponential smoothing models for the furniture data. Lower values for AIC, AICc, BIC indicate better model fit and accuracy.
From the reported results, the "auto" model seems to have relatively lower AICc and MSE values compared to the other models, suggesting that it provides a good balance between model fit and complexity for the furniture data.

```
> 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
```

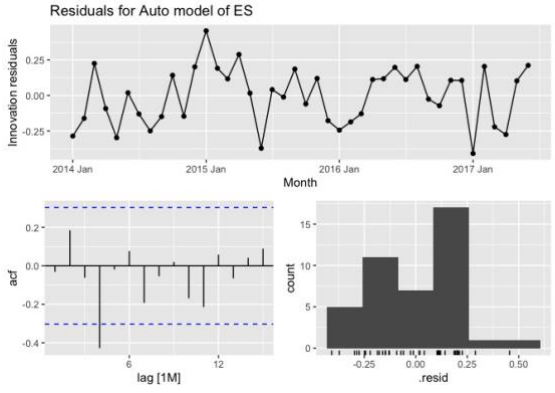
Time Series Final Project: Executive Summary
Superstore Dataset
Group 5: Aniket W, Simran C, Vijay T, Yamini R

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 low MAPE, MPE, MAE and RMSE and we can conclude that its ability to capture the underlying patterns and dynamics of the data.



Validating the results by looking at residuals. The Pvalue for auto model is greater than 0.05 and thus we accept the null hypothesis and conclude that there is white noise in the residuals.

```
> furni_LJ_ES
# A tibble: 5 × 3
  .model                lb_stat lb_pvalue
  <chr>                   <dbl>     <dbl>
1 additive                12.7   0.388
2 ANN                     24.3   0.0184
3 auto                    15.9   0.197
4 Damped Holt's method    26.2   0.00991
5 Holt's method           22.4   0.0329
```



3. ARIMA:(AutoRegressive Integrated Moving Average)

ARIMA uses the autoregression (AR), differencing (I), and moving average (MA) to capture the temporal dependencies and patterns in the data. This helps to capture a wide range of time series patterns and make predictions based on historical observations.

Stationarity: Stationarity is an important property in ARIMA, when a time series is not stationary, it may exhibit trends, seasonality, or other patterns that can lead to inaccurate forecasts.

```
> 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 fail to reject the null hypothesis of stationarity. This suggests that there is not enough evidence to conclude that the time series is non-stationary. Further validating the same using Augmented Dickey-Fuller (ADF) test.

```
> furni_ts1_train %>%
+   features(sales, unitroot_ndiffs)
# A tibble: 1 × 1
  ndiffs
   <int>
1      0
```

The estimated number of differences (ndiffs) is 0 which means that the original series is already stationary and does not require any differencing to achieve stationarity.

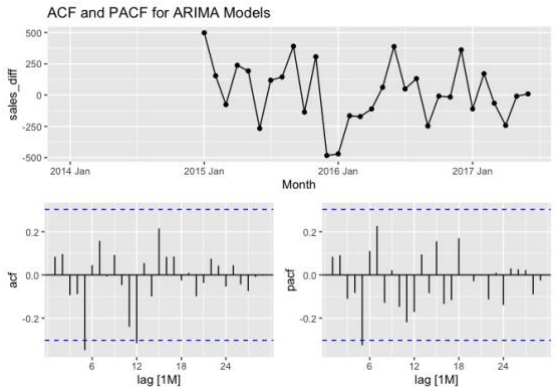
```
> furni_ts1_train %>%
+   features(sales, unitroot_nsdiffs)
# A tibble: 1 × 1
  nsdiffs
    <int>
1       1
```

Seasonal differencing is performed to remove the seasonal pattern or variation in the data. The estimated number of seasonal differences (nsdiffs) is 1 and thus the original series requires one seasonal difference to achieve stationarity in the time series.

Selecting ARIMA Model:

Selection of appropriate values for p, d, and q is important in ARIMA models. Analyzing the autocorrelation and partial autocorrelation functions of the time series data for selecting the accurate model.

Time Series Final Project: Executive Summary
Superstore Dataset
Group 5: Aniket W, Simran C, Vijay T, Yamini R



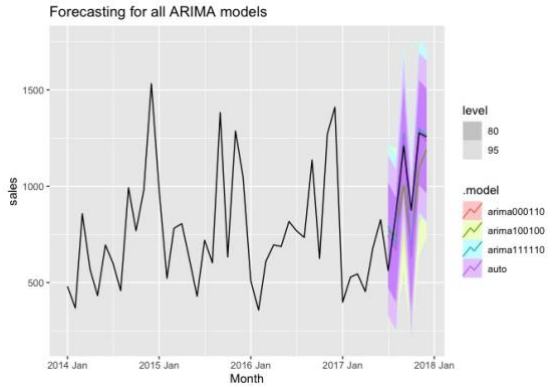
Post analyzing the spikes in ACF and PACF, following model was fitted to the data.

```
> 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.
```

Based on the given information, the ARIMA (1,1,1)(1,1,0)[12] model is the best fit among the models, as it has lower AIC, AICc, and BIC values and includes significant roots in both the AR and MA components.

```
> 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 Ljung-Box test results indicate that the residuals of the ARIMA (1,1,1) (1,1,0)[12] model do not exhibit significant autocorrelation up to a lag of 12. This suggests that the model adequately captures the temporal dependence in the data.



Conclusion:

Based on the analysis and comparison between Exponential Smoothing (ES) and ARIMA models for the furniture sales time series data, the ARIMA (1,1,1)(1,1,0)[12] model emerges as the most suitable choice.

Although Exponential Smoothing models, particularly the "auto" model, show good performance in capturing seasonality and adapting to changes in the data, the ARIMA model provides greater flexibility and handles a wider range of time series patterns. 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. Additionally, the ARIMA (1,1,1)(1,1,0)[12] model demonstrates a lower AIC, AICc, and BIC, indicating a better fit to the data. The Ljung-Box test also suggests that the ARIMA model adequately captures temporal dependence in the residuals.