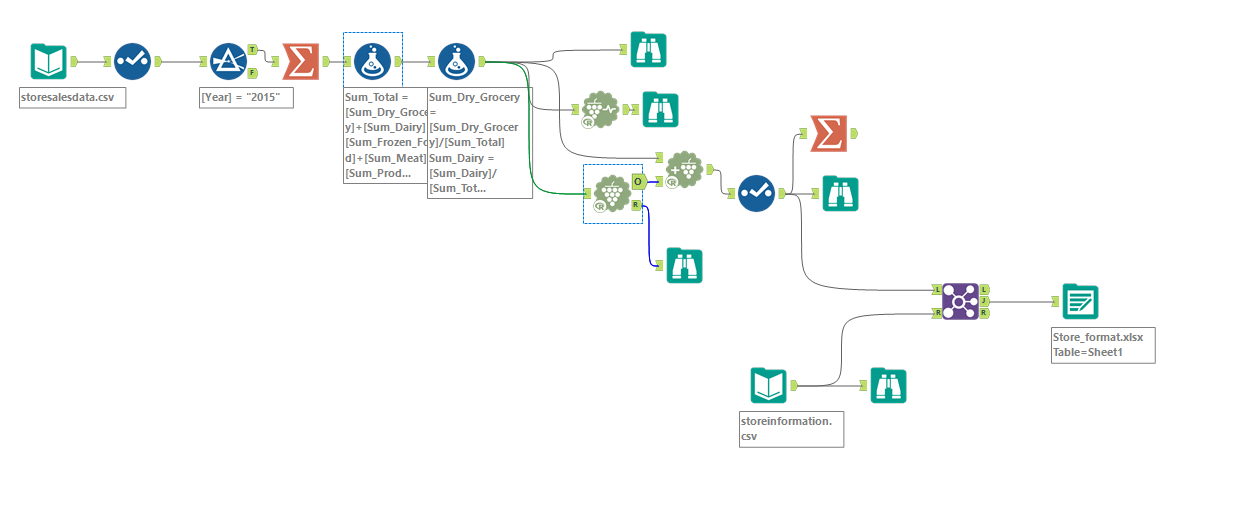
Project: Predictive Analytics Capstone

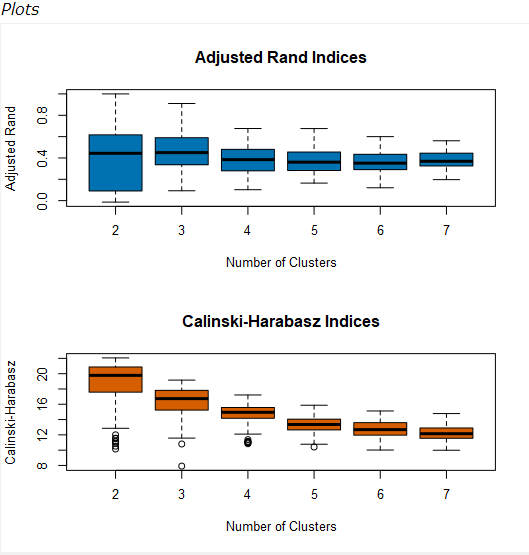
## Task 1: Determine Store Formats for Existing Stores

Company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all store use the same store format to sell all their products, which begin to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. I am asked to provide analytical support to make decisions about store formats and inventory planning.



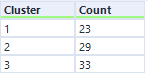
1. What is the optimal number of store formats? How did you arrive at that number?

I used only 2015 sales data and applied K-means clustering model. Based on K-means cluster assessment report, I picked the final optimal number of store formats to be 3 as it has highest rand in Adjusted Rand Indices chart and highest rand in Calinski\_harabasz Indices chart as well.



1. How many stores fall into each store format?

Cluster 1 has 23 stores, cluster 2 has 29 stores and cluster 3 has 33 stores.

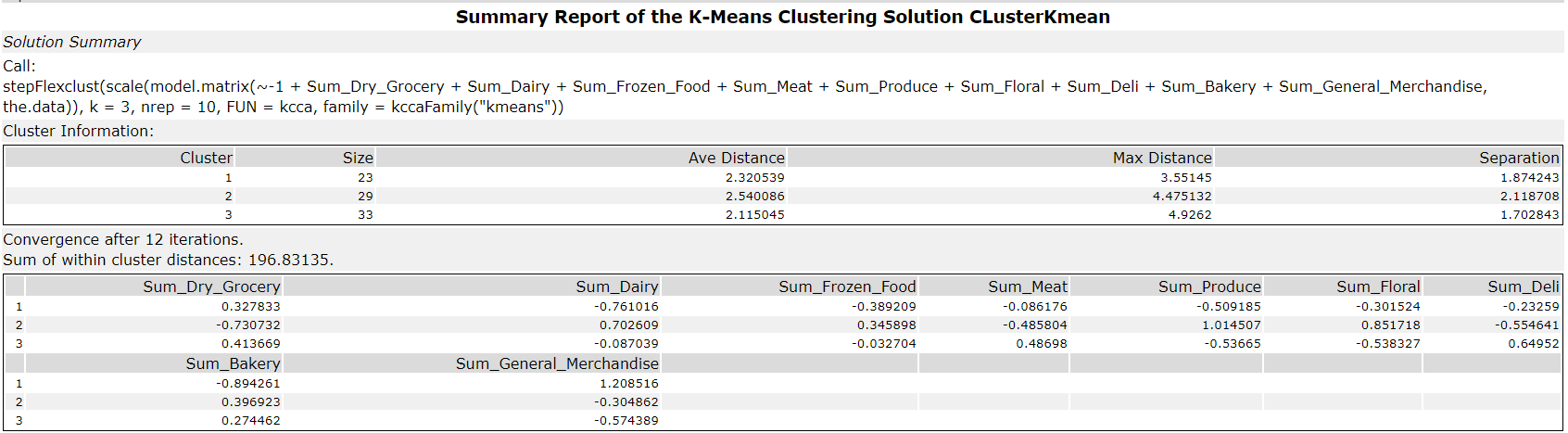


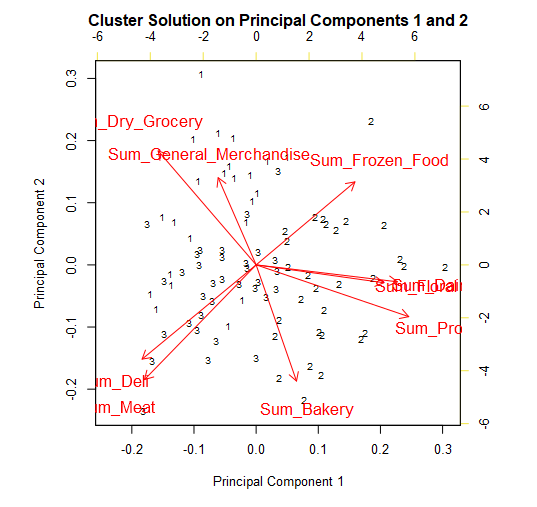
1. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Cluster 1 has more general merchandise sales and less Dairy sales

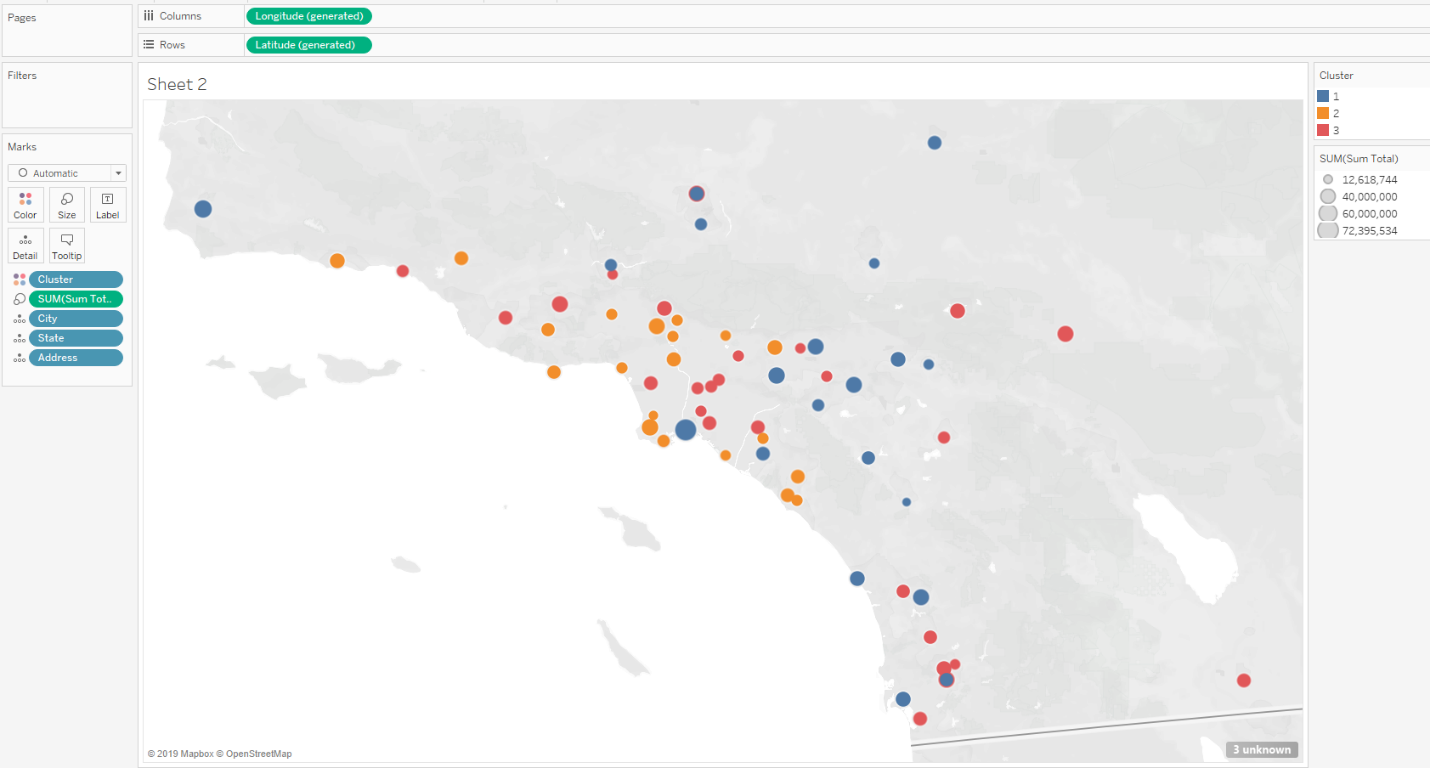
Cluster 2 has more dairy, frozen food, produce, floral sales and less dry grocery, meat sales.

Cluster 3 has more meat and deli sales and less general merchandise sales.





1. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

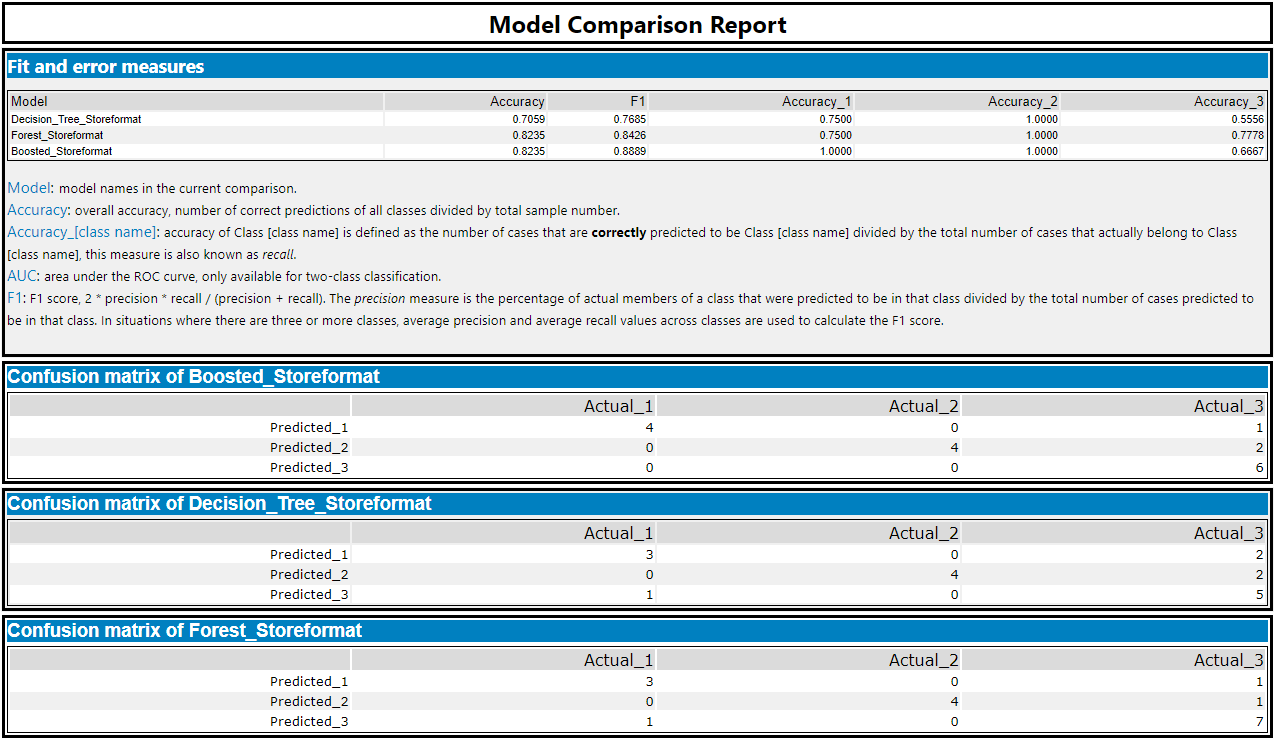


## Task 2: Formats for New Stores

The grocery store chain has 10 new stores opening up at the beginning of the year. I am asked to determine which store format each of the new stores should. Since I don’t have sales data for these new stores, we need to determine the format based on demographic data.

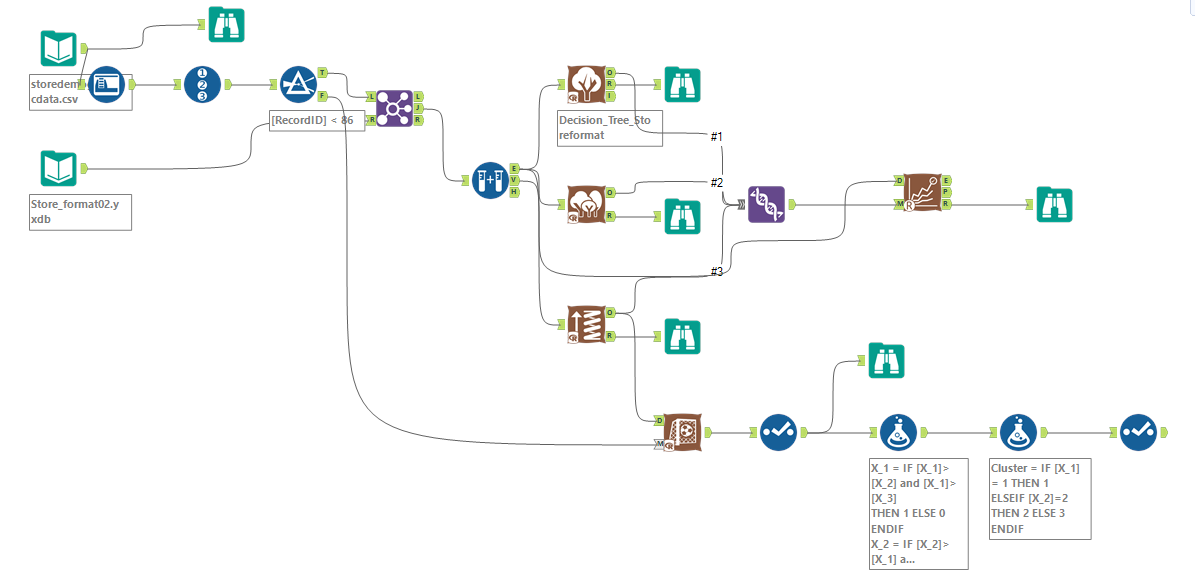
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

The model comparison report below shows the comparison matrix of Decision Tree, Forest Model and Boosted Model. From the report below, we can see that Boosted Model has the highest accuracy and F1 score. Thus, Boosted model is the best model to apply for the prediction.



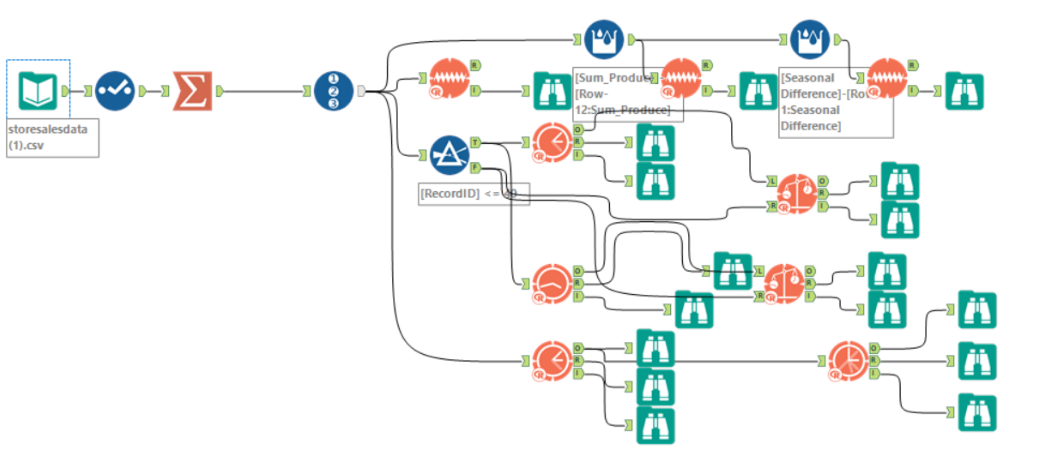
1. What format do each of the 10 new stores fall into? Please fill in the table below.

|  |  |
| --- | --- |
| Store Number | Segment |
| S0086 | 3 |
| S0087 | 2 |
| S0088 | 1 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 1 |
| S0092 | 2 |
| S0093 | 1 |
| S0094 | 2 |
| S0095 | 2 |



## Task 3: Predicting Produce Sales

I am asked to predict the accurate monthly sale forecast.

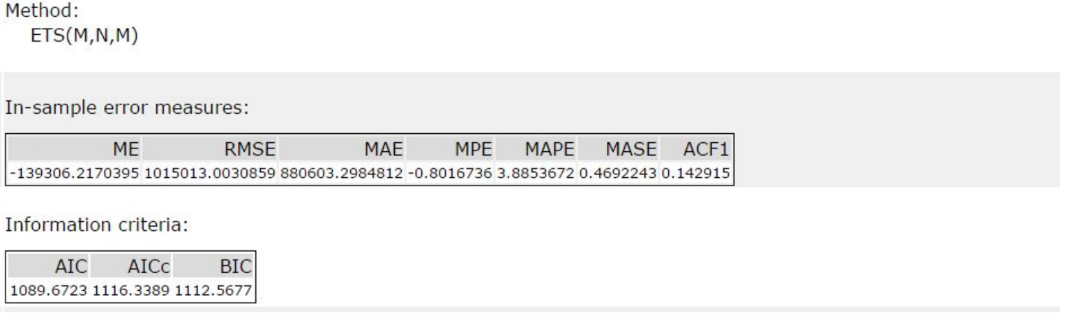


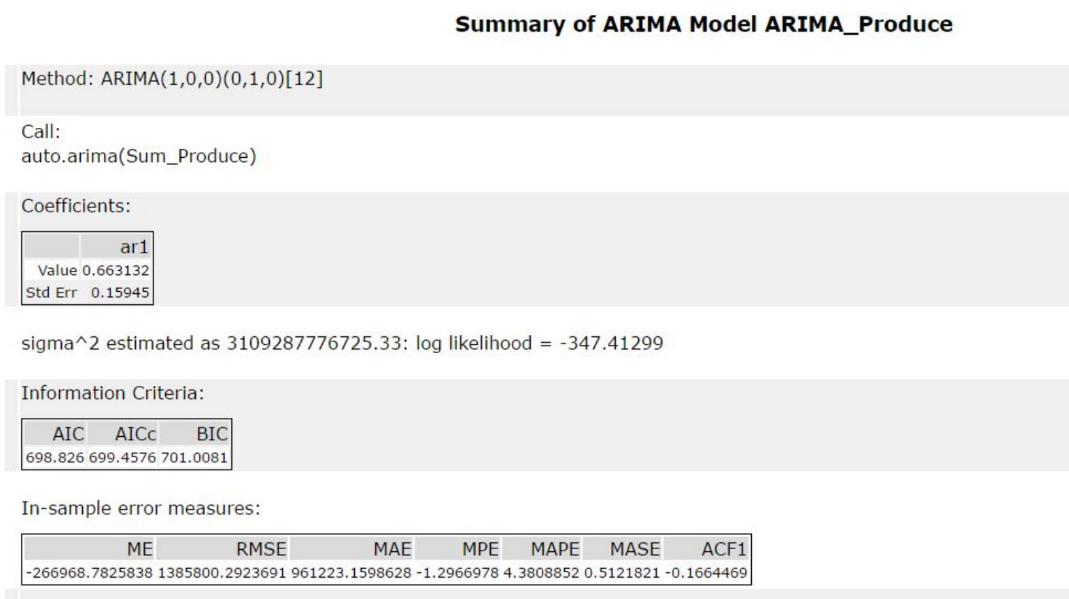
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

ETS(M,N,M) with no dampening should be used for ETS model.

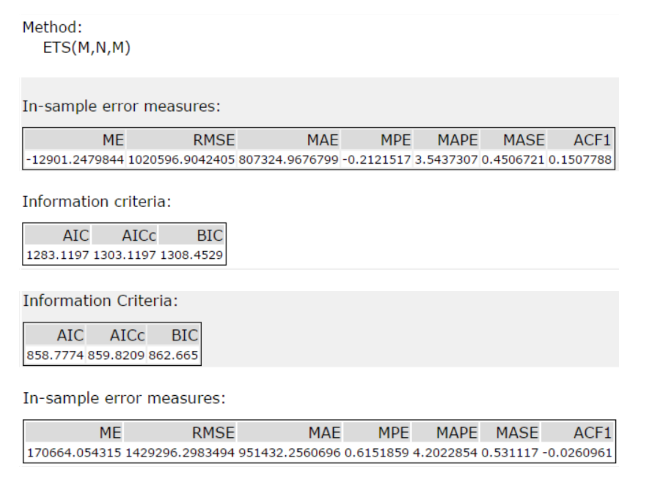
The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and nothing should be applied. The error chart is irregular and should be applied multiplicatively as well.

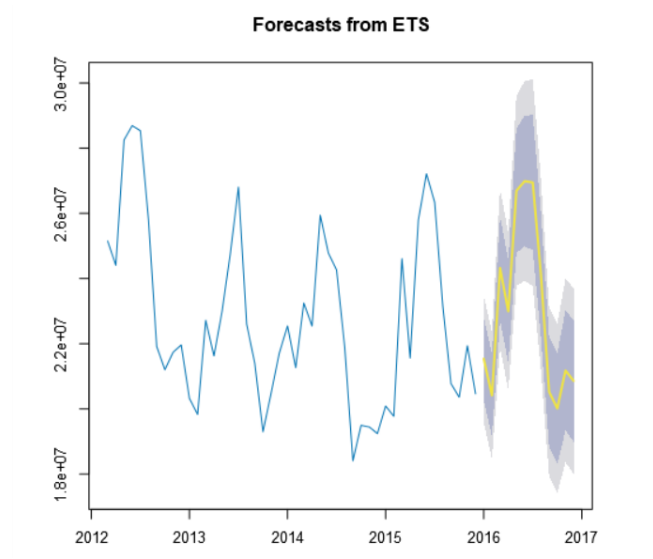
ARIMA(0,1,2)(0,1,0) is set to calculate elements automatically.

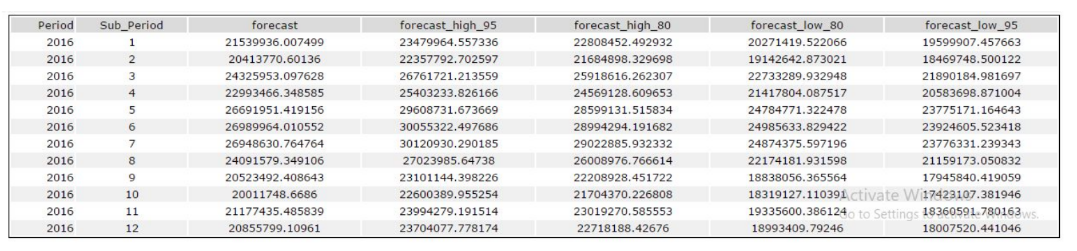




ETS model’s accuracy is higher when compared to ARIMA model. A holdout sample of 6 months data is used. Its RMSE of 1,020,597 is lower than ARIMA’s 1,429,296 while its MASE is 0.45 compared to ARIMA’s 0.53. ETS also has a higher AIC at 1,283 while ARIMA’s AIC is 859.







1. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Please see below forecast table for existing and new stores.



