Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

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

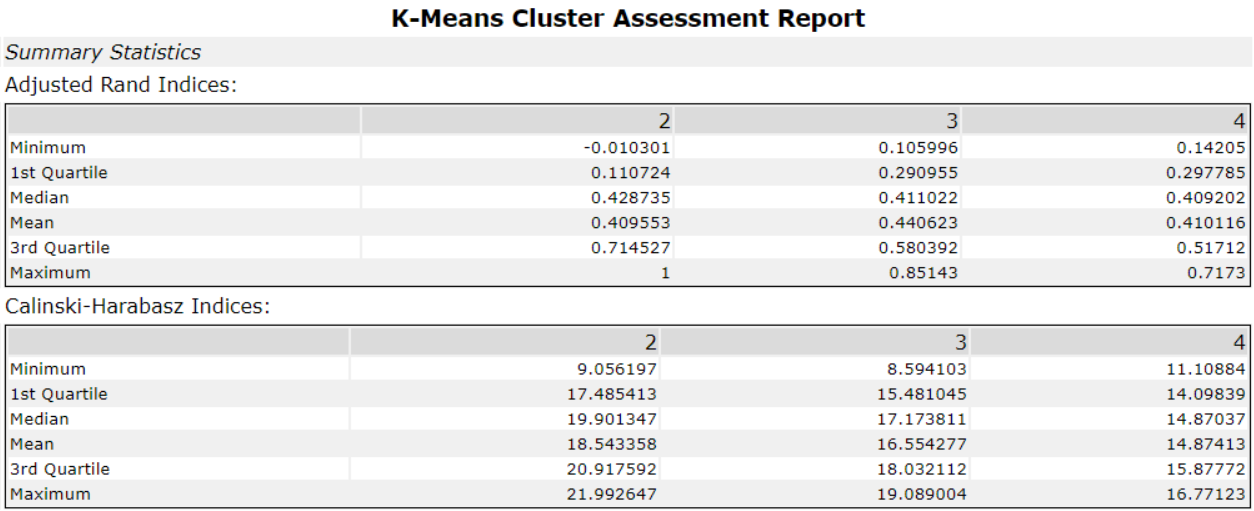


Fig: K-Means Cluster Assessment Report

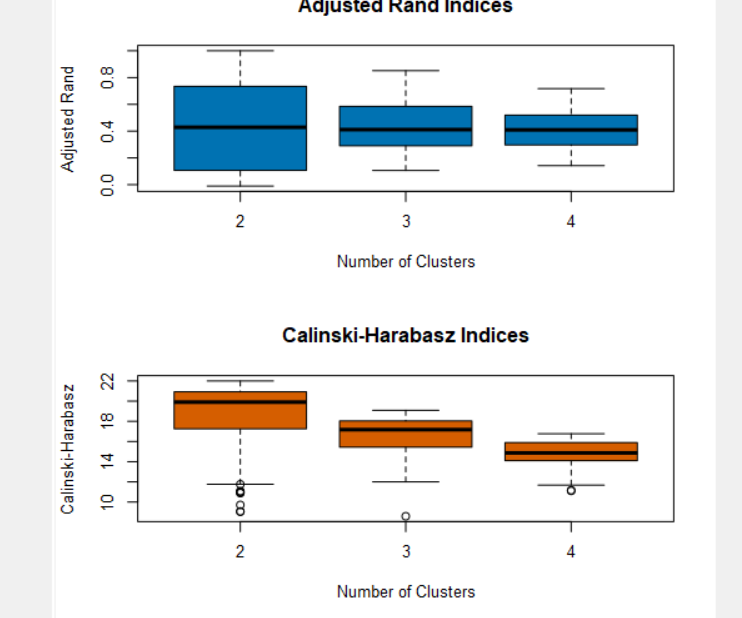


Fig: Adjusted Rand Indices and Calinski-Harabasz Indices

Based on the K-means report, Adjusted Rand and Calinski-Harabasz indices below, the optimal number of store formats is **3** because for 3 clusters CH and AR indices have high median value and are compact.

**2. How many stores fall into each store format?**

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

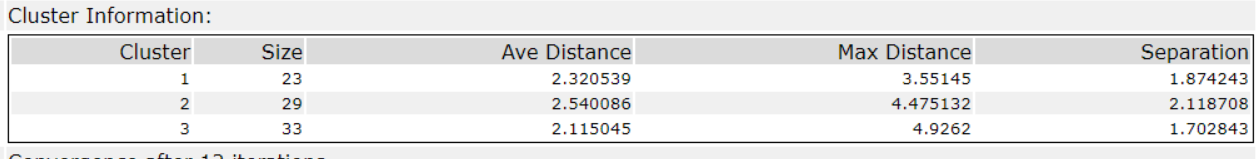


Fig: Cluster Information

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

Cluster 1 stores have the highest total sales. Cluster 2 stores sold more Produce. Cluster 1 stores sold more General Merchandise.

Cluster 1 stores have the highest median total sales when compared to the other 2. Cluster 3 stores are the most similar in terms of sales due to their compact range.

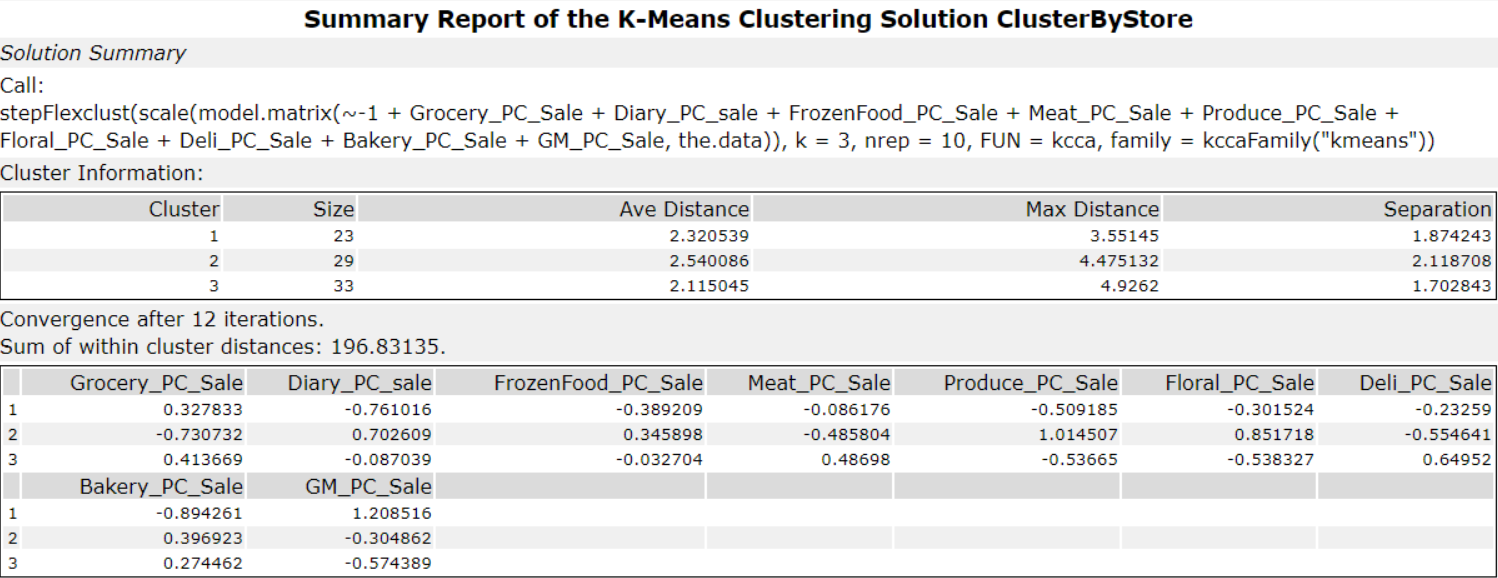


Fig: K-Means Clustering Summary

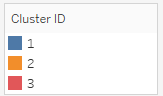
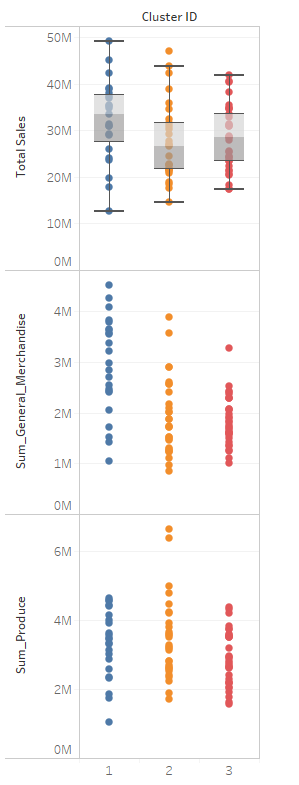


Figure 4: Tableau Visualization

**4. 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.**

<https://public.tableau.com/profile/doibajit.medhi#!/vizhome/StoreByClustersSales/StoreBycluster?publish=yes>



## Task 2: Formats for New Stores

**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 comparison matrix of Decision Tree, Forest Model, and Boosted Model.

**Boosted Model** is chosen despite having the same accuracy as Forest Model due to higher F1 value.

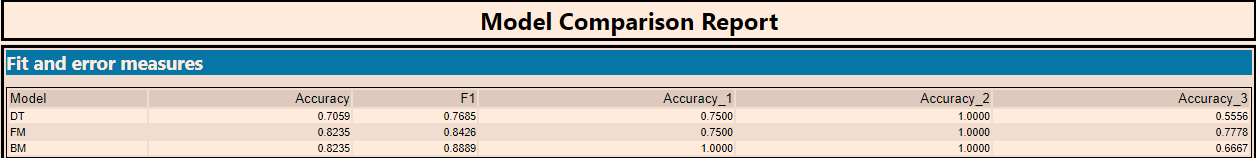


Fig 6: Model Comparison Report

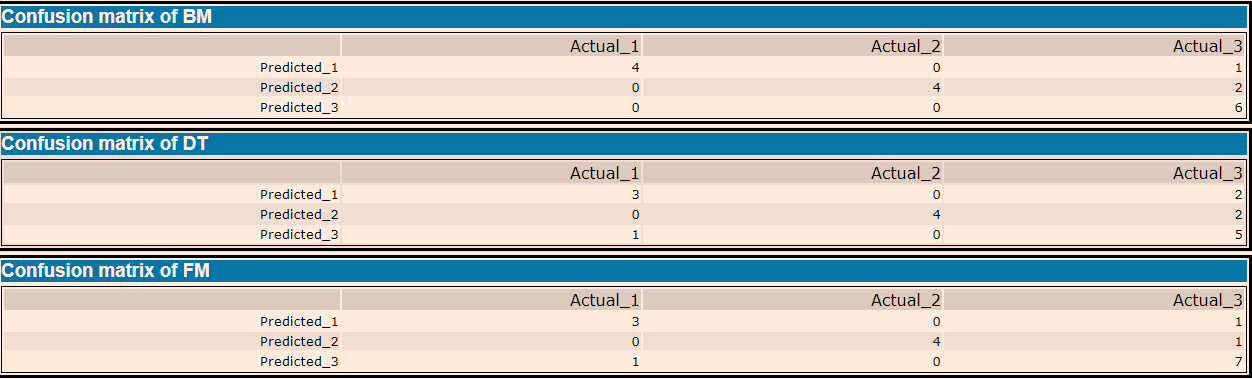


Fig: Confusion Matrices for all models.

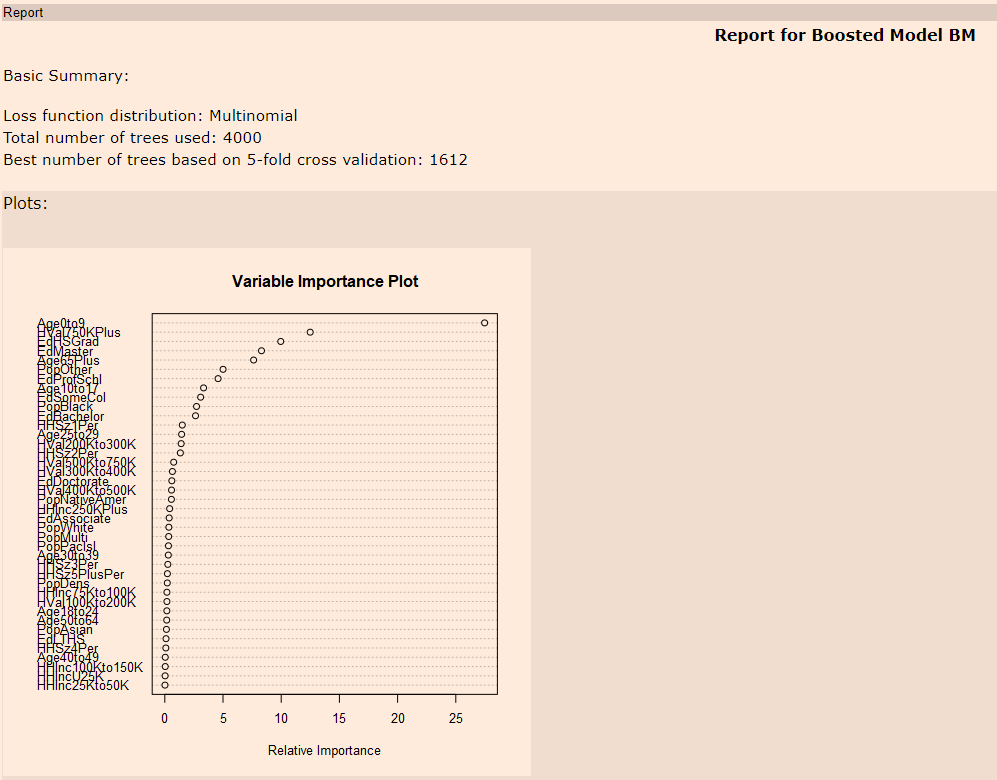


Fig: Variance Importance Plot

Ave0to9, HVal750KPlus, and EdHSGrad are the three most important variables.

**3. What format does 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 |

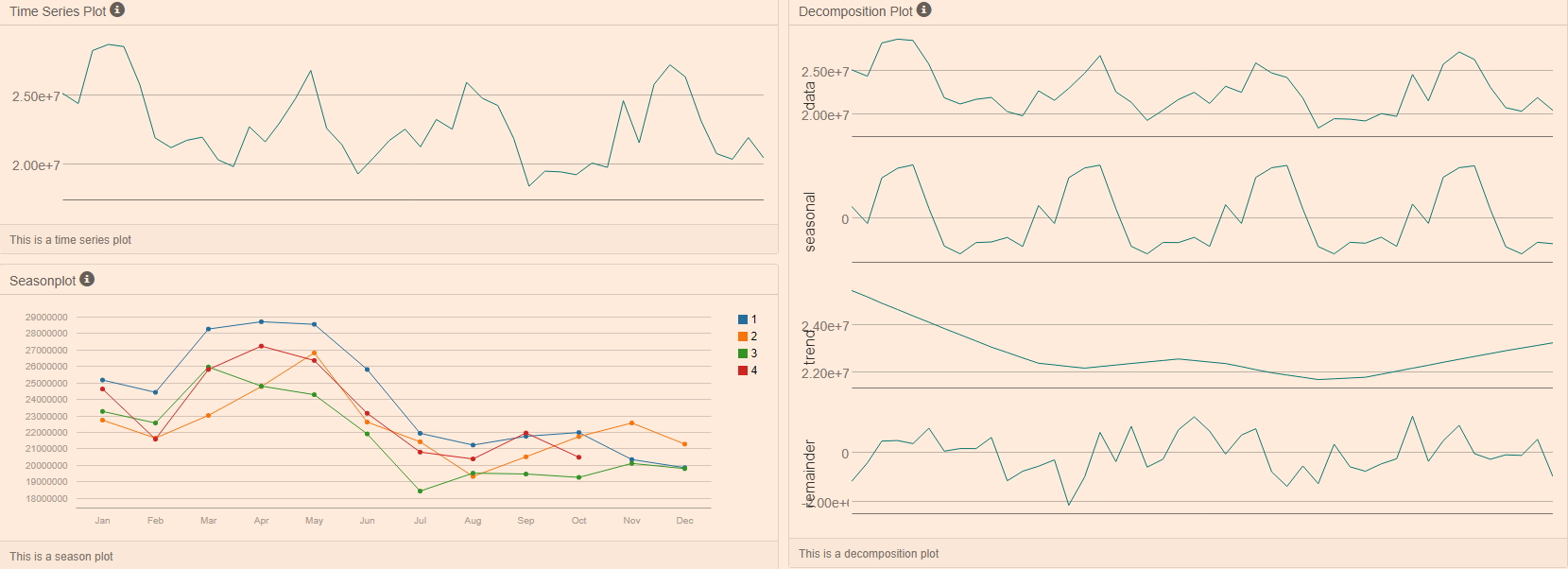
## Fig: Store Number and Segment

## Task 3: Predicting Produce Sales

**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** is used for ETS model.

The seasonality shows increasing peaks and valley trend and should be applied multiplicatively. The trend is not clear therefore nothing should be applied. The error is irregular and should be applied multiplicatively.



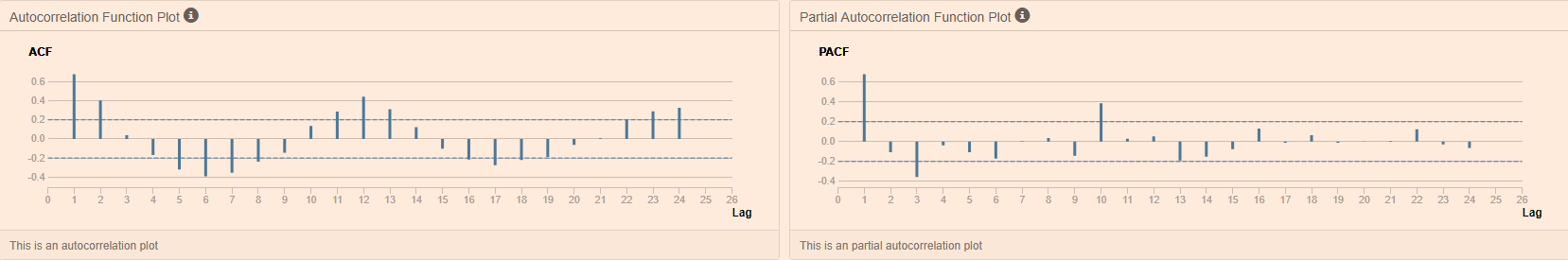


Fig: TS Plot

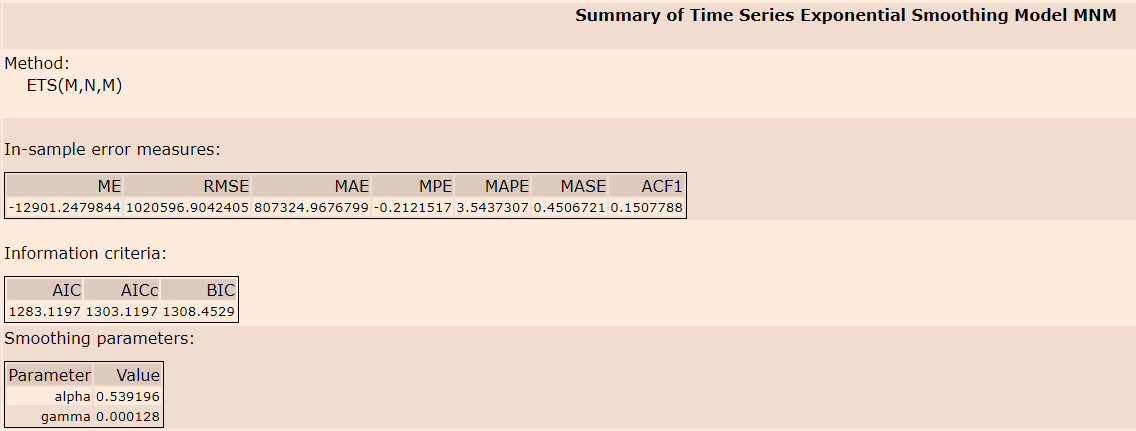


Fig: ETS Model summary

For the ARIMA model, the model is set to calculate options automatically. ARIMA (1,0,0)(1,1,0)[12] is run.

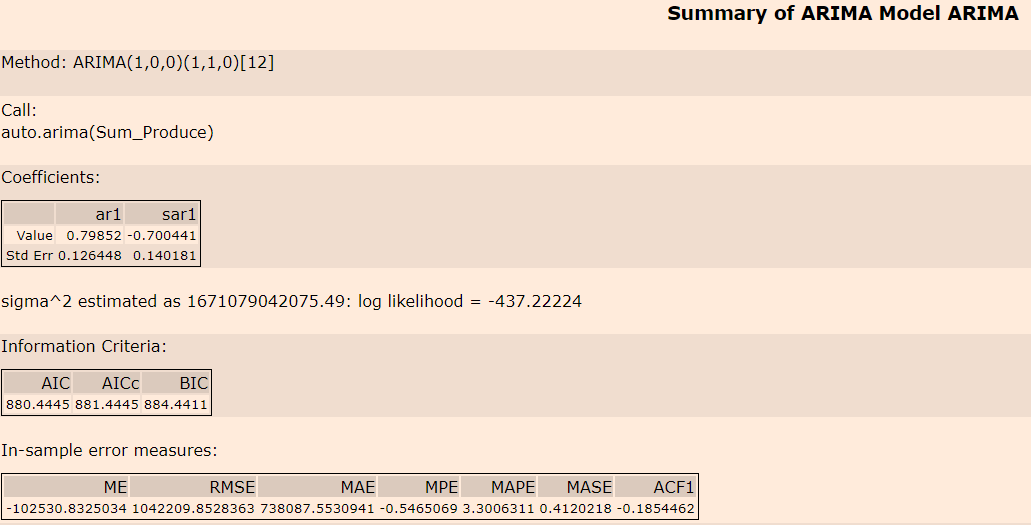


Fig: ARIMA Model Summary

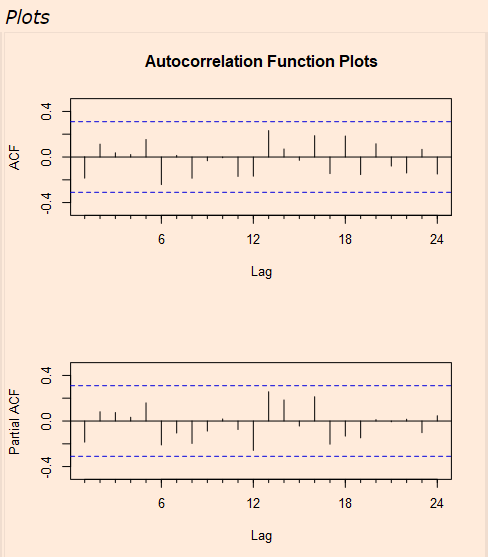


Fig: ARIMA model ACF and PCF plots

ETS model’s accuracy is higher when compared to the ARIMA model. A holdout sample of 6 months of data is used. Its RMSE of 1020596 is lower than ARIMA’s 1042209. ETS model also has a higher AIC in 1283 while ARIMA’s AIC is 880.

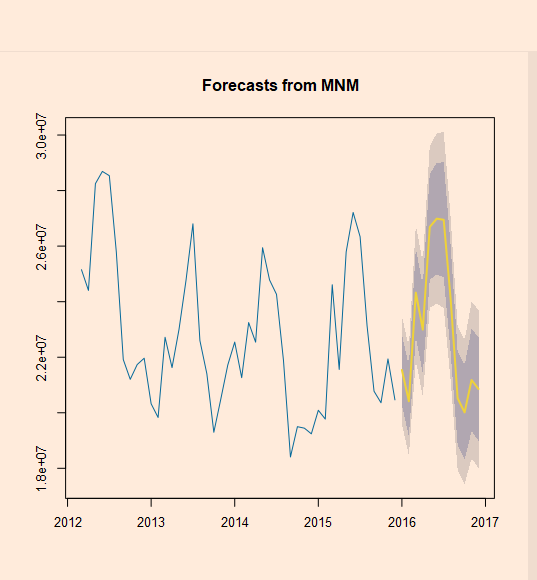


Fig: 12 Period Forecast from ETS MNM

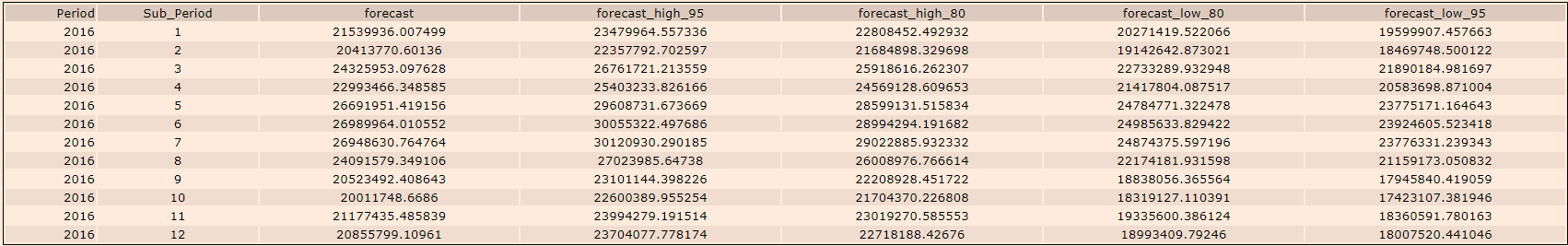


Fig: Forecast with confidence intervals

**2. 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.**

<https://public.tableau.com/profile/doibajit.medhi#!/vizhome/TableauForecastforProduceSales/Task3-Forecast?publish=yes>



Fig: Produce Forecast table for 2016

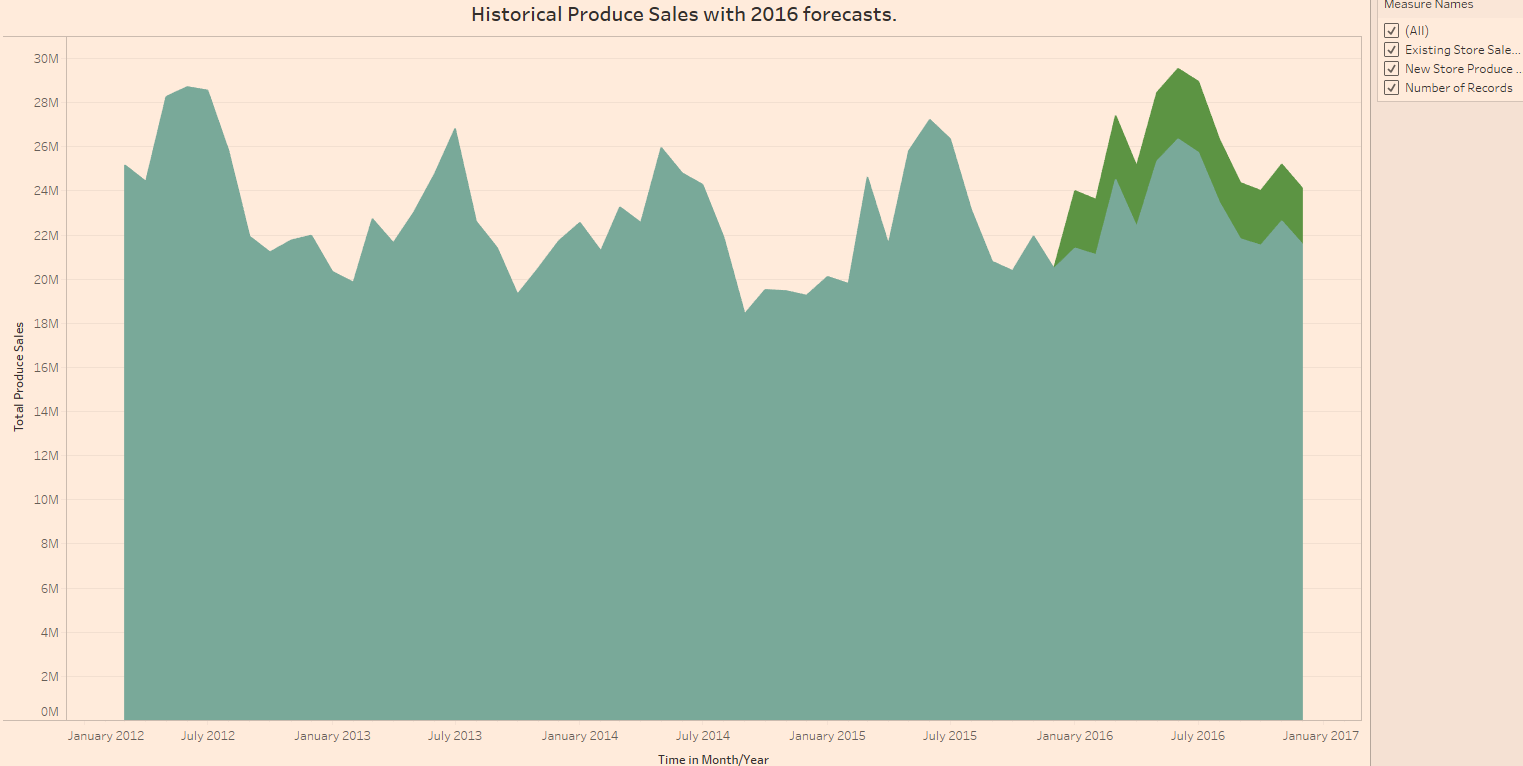


Fig: Tableau visualization for historical and forecast sales for existing stores and new stores

Alteryx Workflows:

