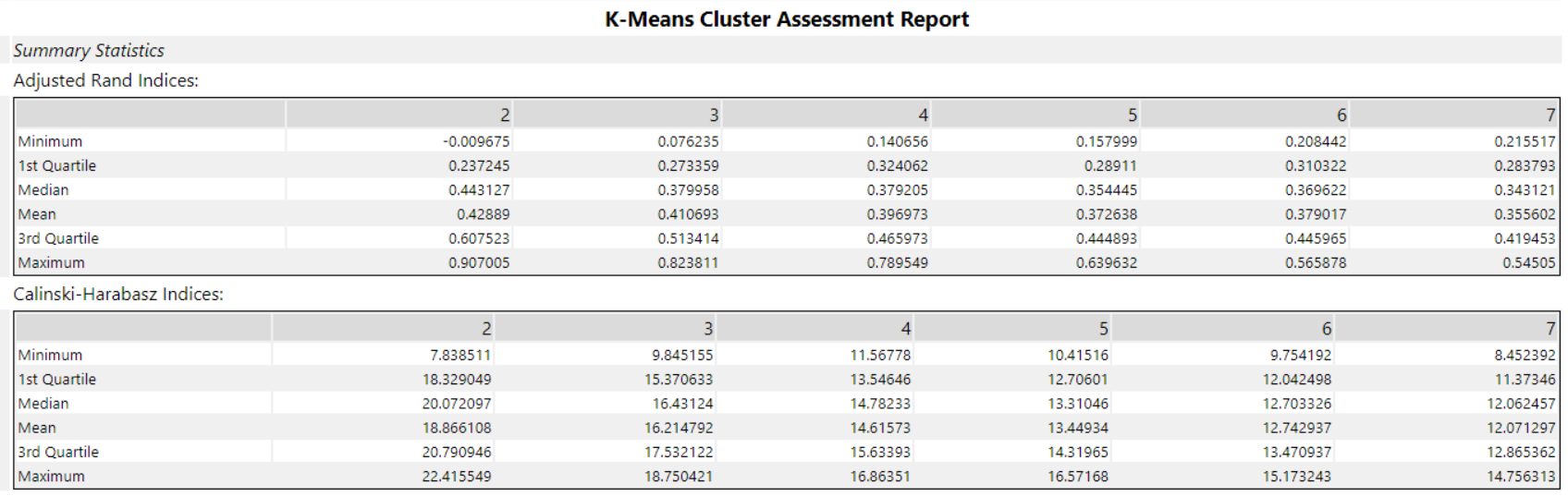
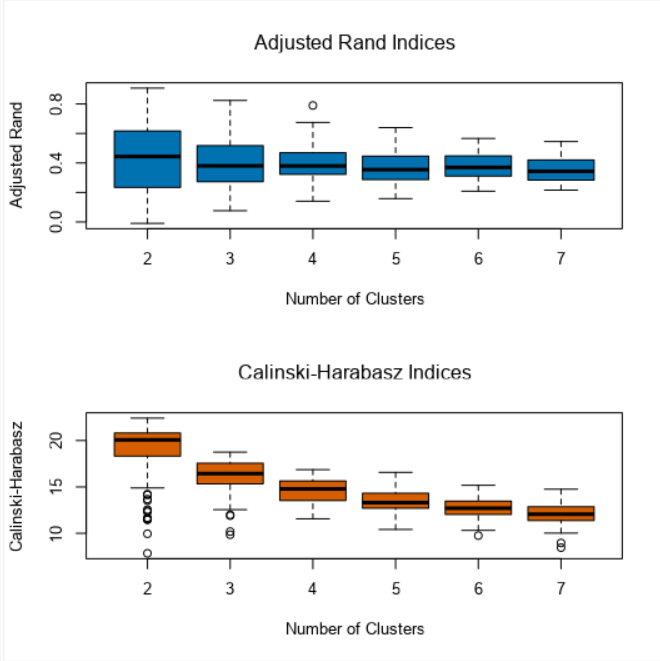
**Project: Predictive Analytics Capstone**

## Task 1: Determine Store Formats for Existing Stores

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

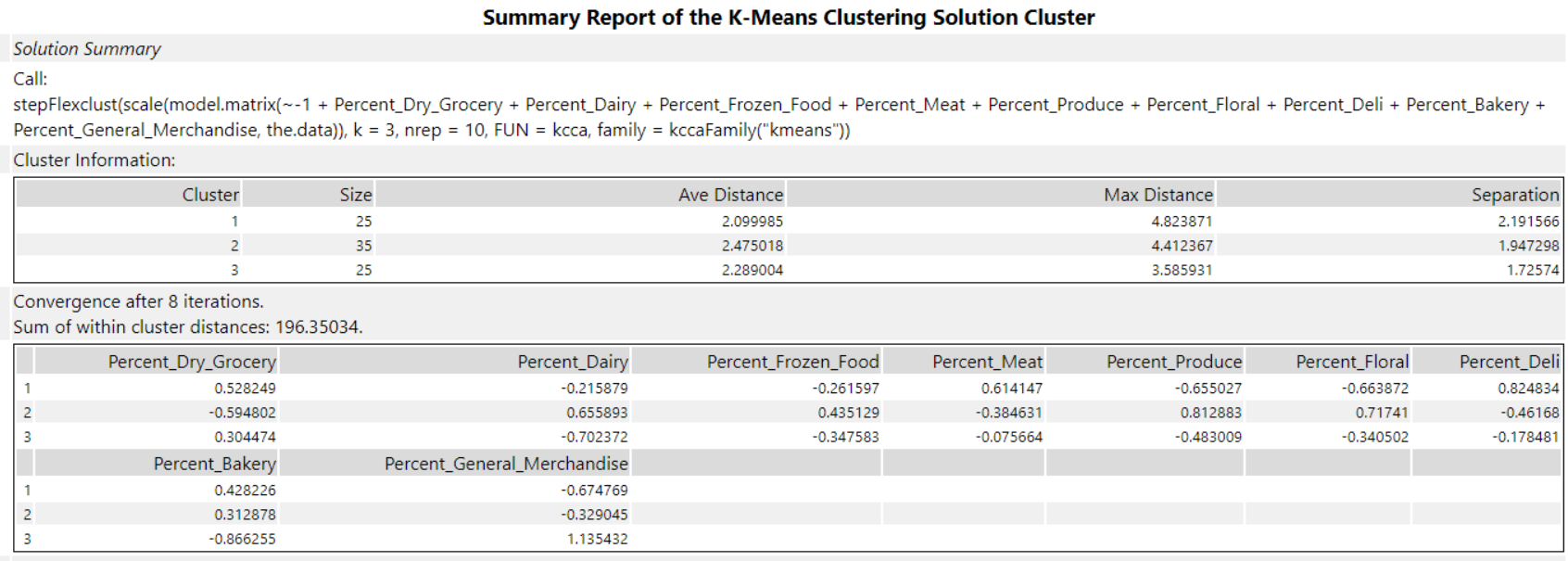
The optimal number of store formats is 3. I arrived this number by using K-Centroids Cluster analysis and K-Centroids Diagnostics Tool with K-Mean Clustering Method. According to K-Mean analysis or below report, both Adjusted Rand Indices and Calinski-Harabasz Indices shows highest mean value at 2 and 3 indicating that the optimal number of stores formats is 3 (but we can also take 2 stores).



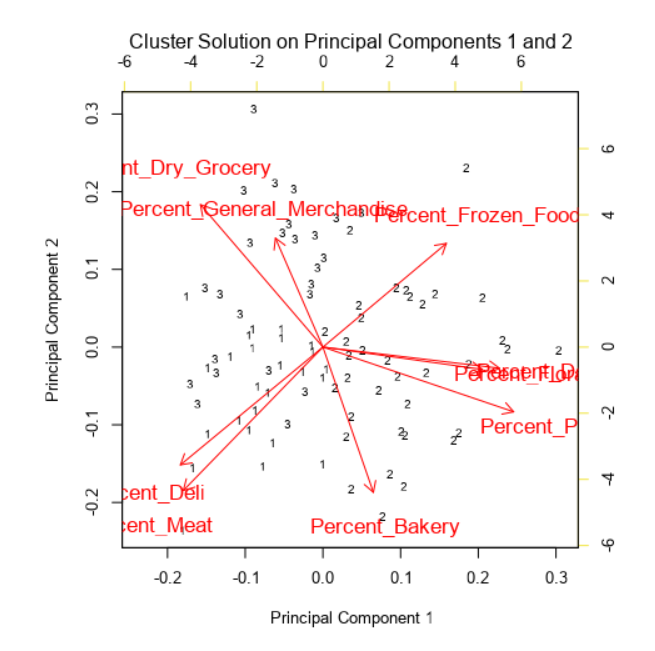


1. **How many stores fall into each store format?**

According to Cluster Information: Cluster 1 has 25 stores; Cluster 2 has 35 stores and Cluster 3 has 25 stores.

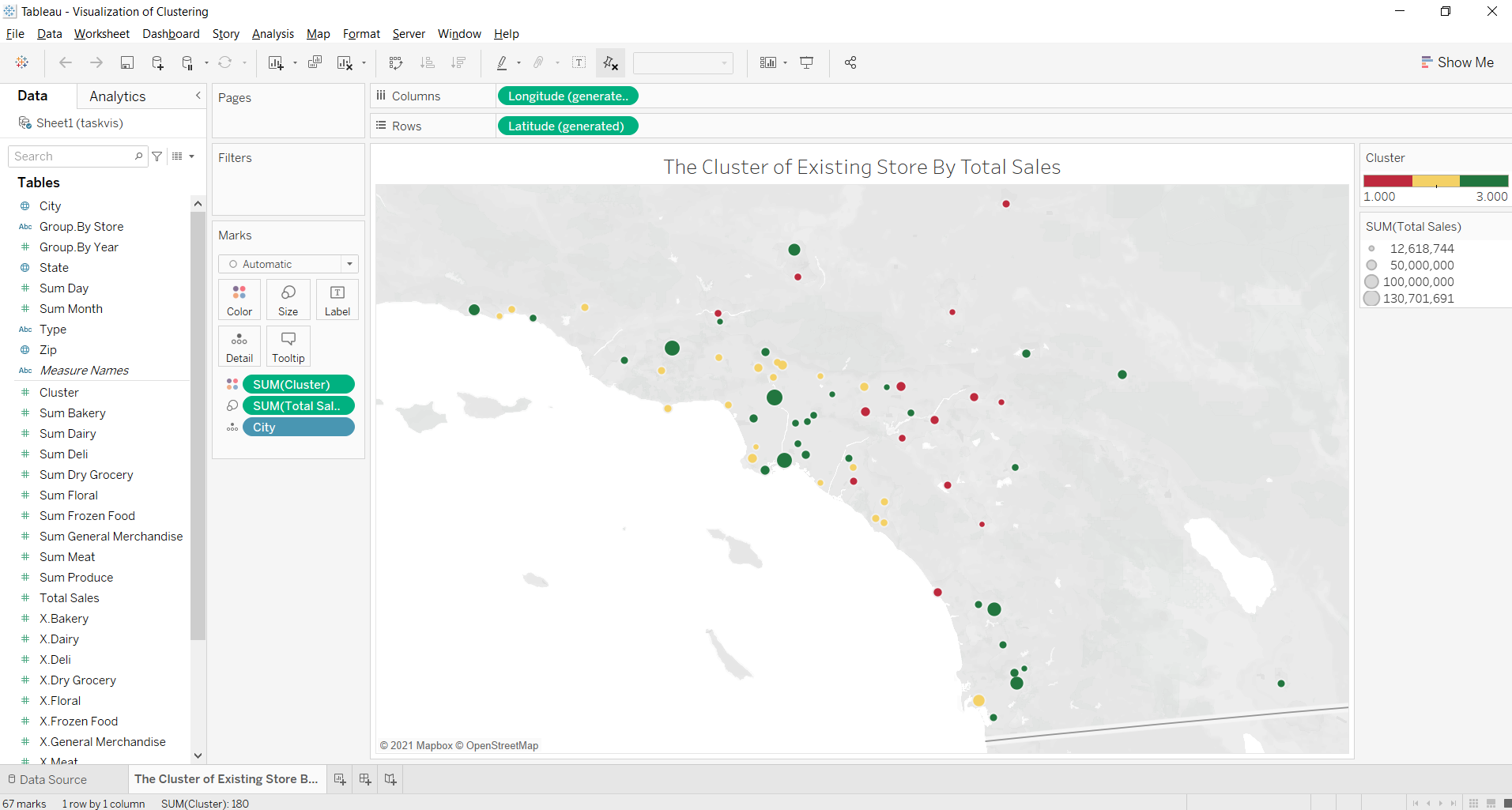


1. **Based on the results of the clustering model, what is one way that the clusters differ from one another?**

****

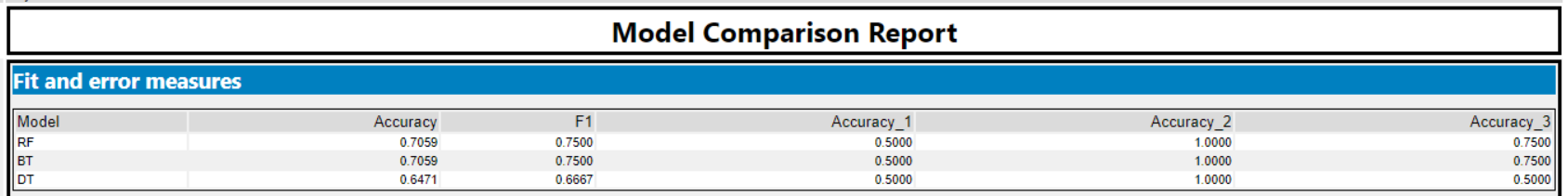
From the report of K-Mean Clustering, we can see the cluster 1 and 3 both of them have same size of 25. Cluster 1 has the highest Max Distance which is 4.82 and have lowest Ave Distance of 2.09 with highest separation of 2.19. Cluster 2 has the largest size of 35 with Highest Ave Distance of 2.47, Max Distance of 4.41 and Separation of 1.94. Similarly, Cluster 3 has Ave Distance of 2.28, Max Distance of 3.58 and Separation of 1.75.

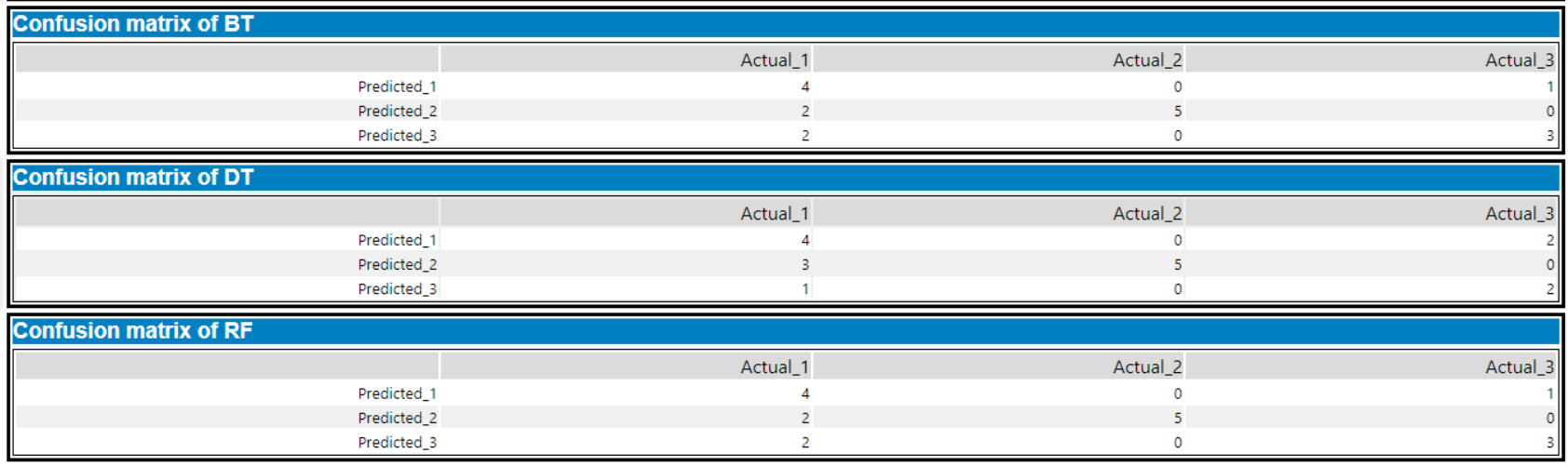
1. **Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.**

****

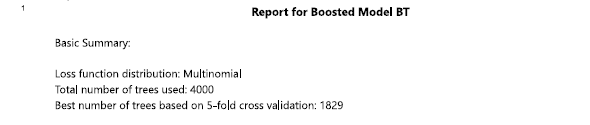
## Task 2: Formats for New Stores

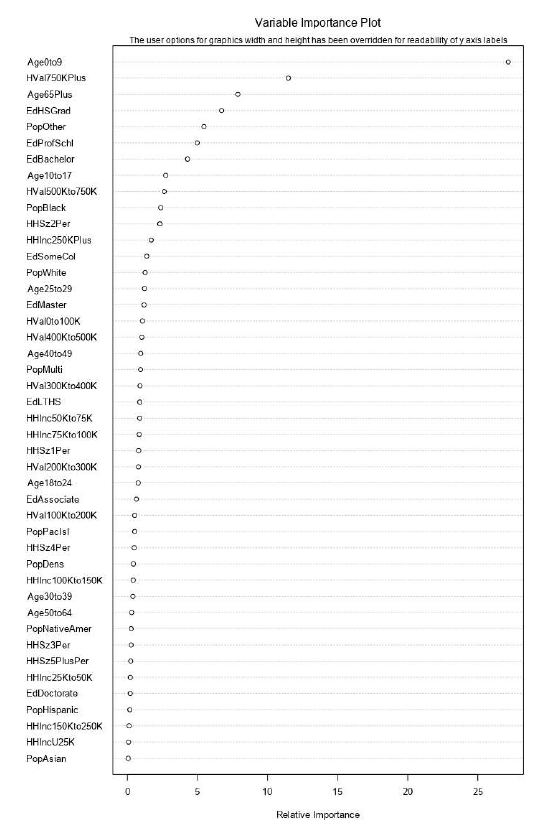
1. **What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)**

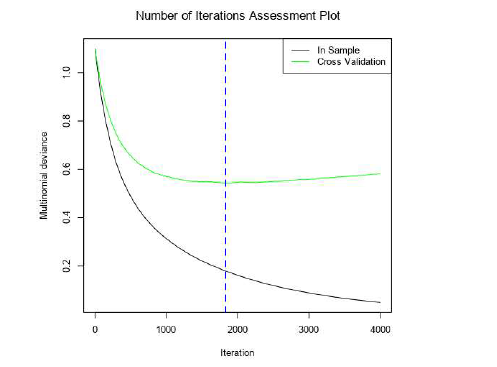
****

****

In any classification problem we will need to set an estimation sample 80 and a validation sample 20 of my data. This helps us compare different classification models to see which better fit the data. Both Random Model and Boosted Model. I going to use Boosted model because dataset is small in number so that we can run the boosted model in short time. The comparison result made me choose boosted model has the best result in Accuracy = 70.59%, F1 = 75% and Accuracy\_1 = 50%.







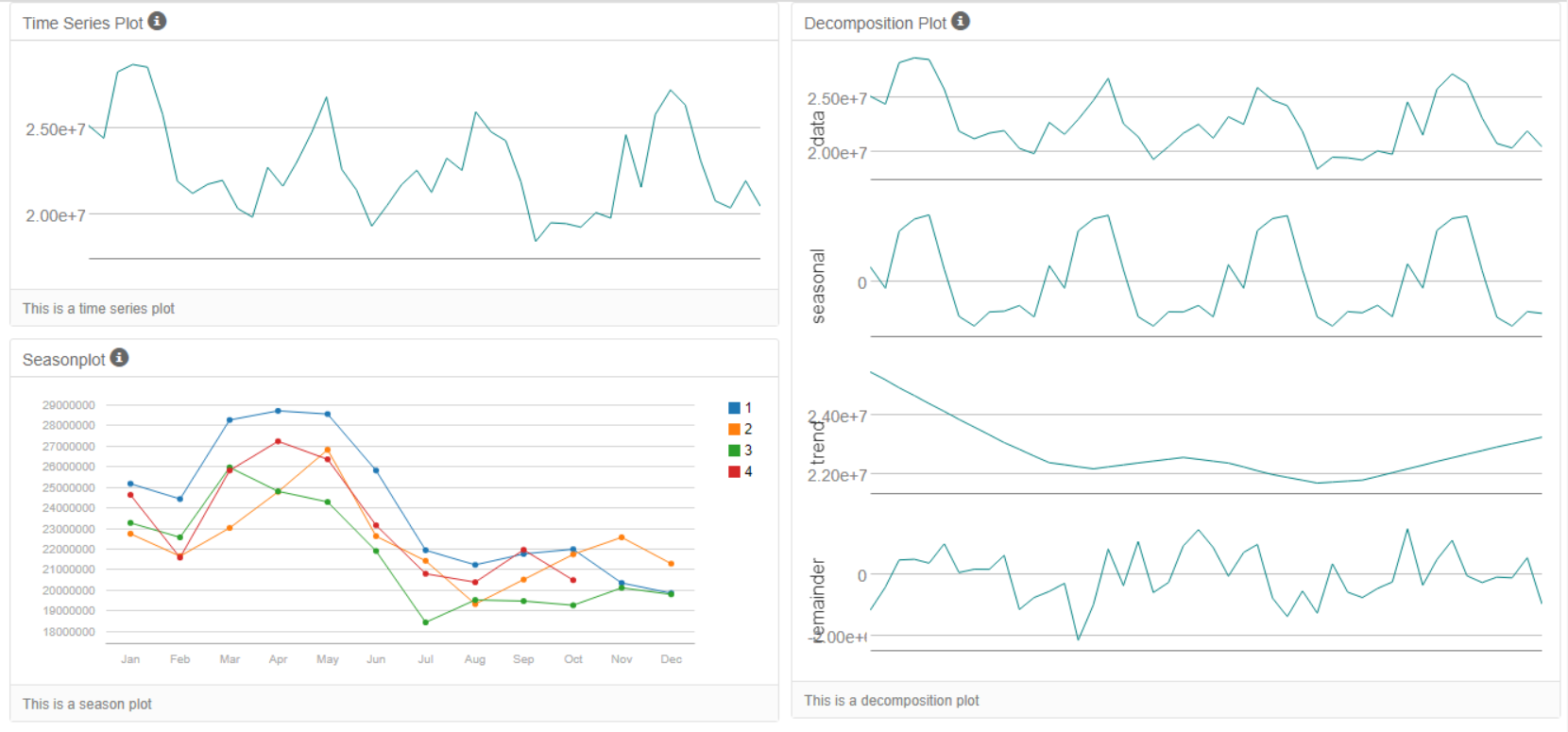
1. **What format do each of the 10 new stores fall into? Please fill in the table below.**

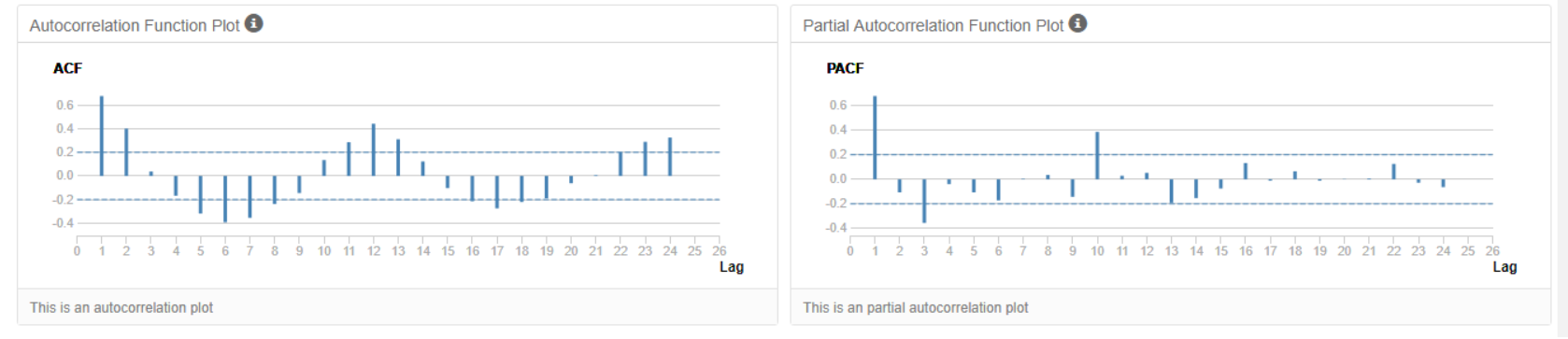
|  |  |
| --- | --- |
| Store Number | Segment |
| S0086 | 1 |
| S0087 | 2 |
| S0088 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 3 |
| S0092 | 2 |
| S0093 | 3 |
| S0094 | 2 |
| S0095 | 2 |

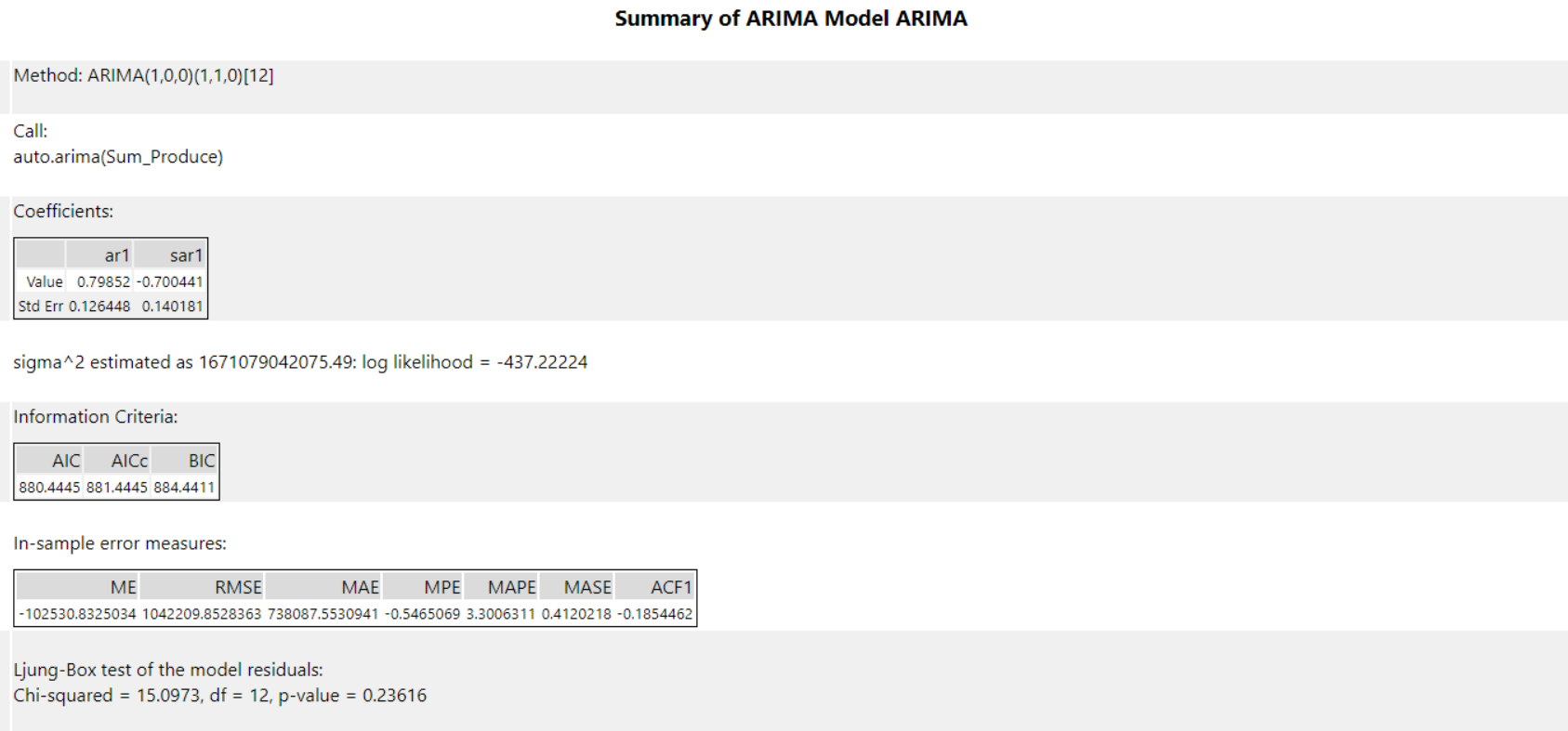
## Task 3: Predicting Produce Sales

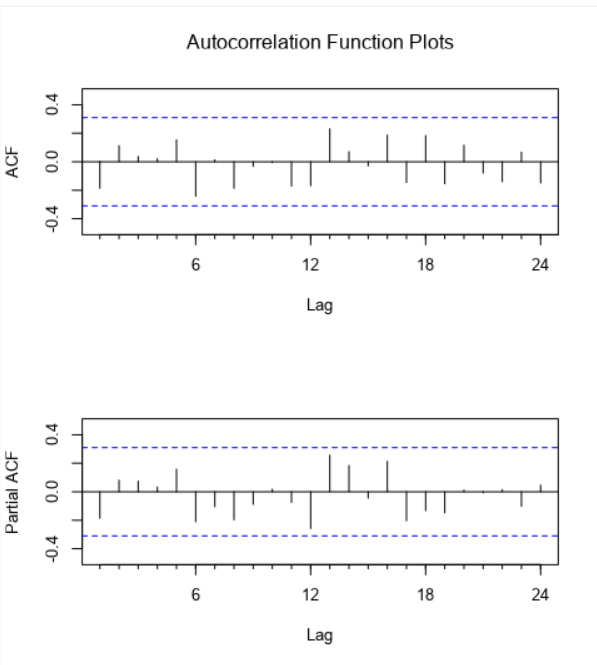
**1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?**

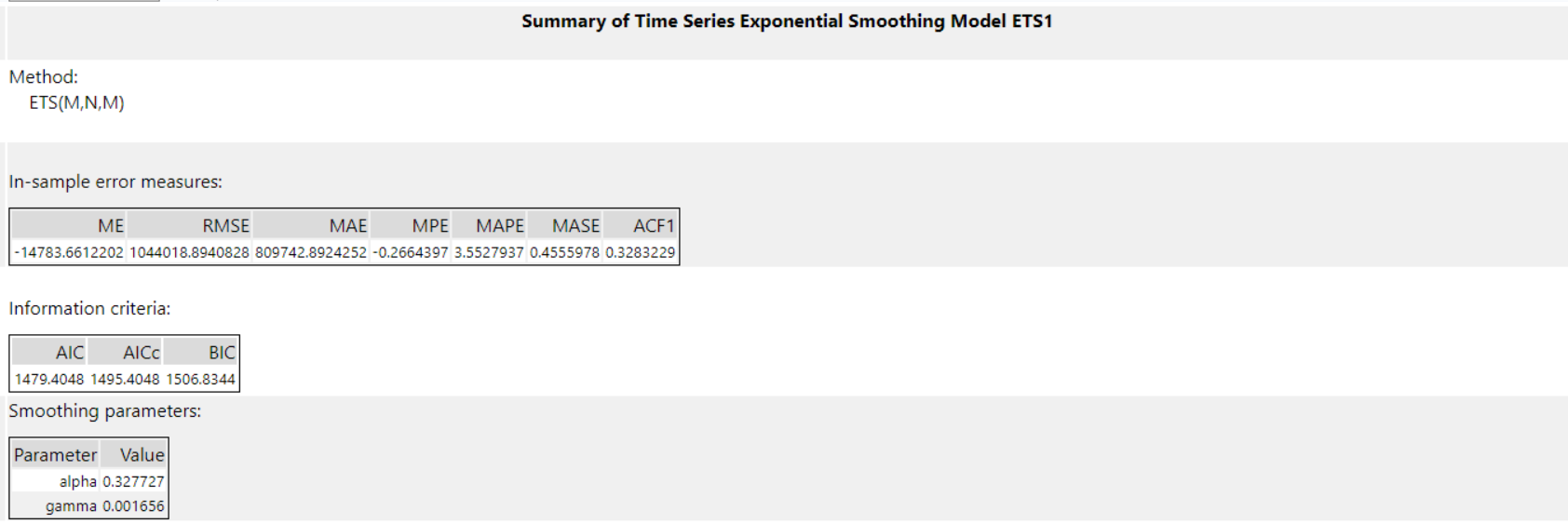
I have used ETS model for forecast. I have come to this decision by comparing between ETS and ARIMA and using TS plot tool. From below Decomposition plot, I have seen the error is multiplicative, the trend is non-exciting and seasonality has an increase trend and multiplicative as the perks change over time. So, I have chosen the ETS model.



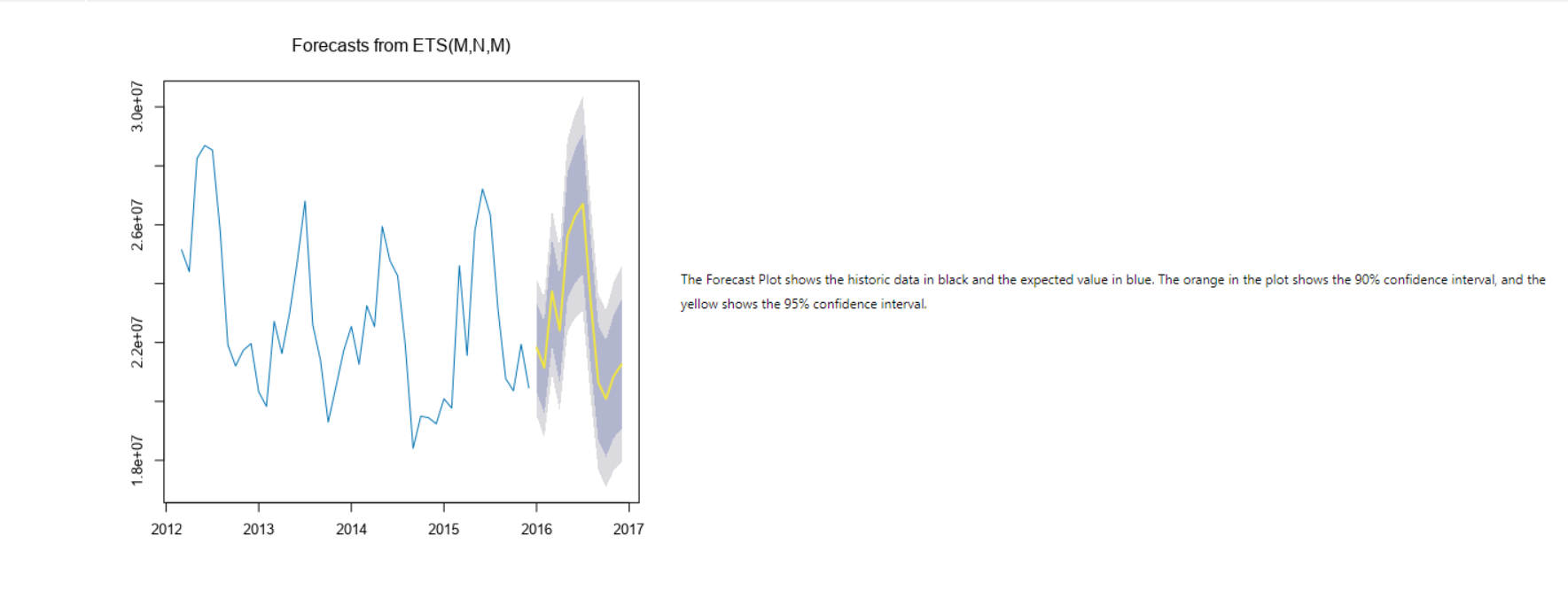


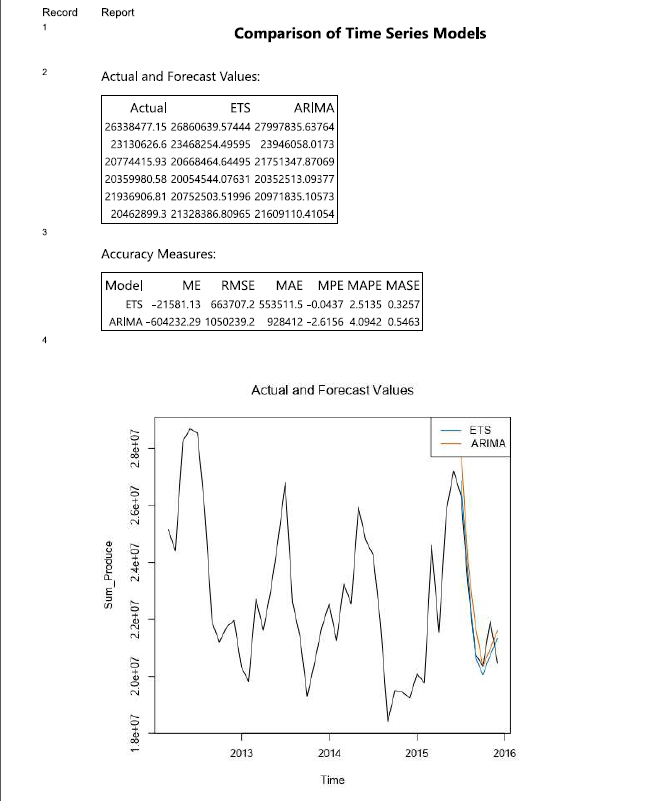












From the Actual vs. Forecast Values for Arima and ETS plots above, I can see the forecast values by the ETS model is most near to the actual values than the forecast values by the Arima model.

**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.**

The forecasted values for produce, monthly in 2016 for new and existing stores, table down shows the historical data together with these forecasts.

