**Predictive Analytics Capstone**

**Task 1: Determine store formats for existing stores.**

**The Business Scenario: Store Format**

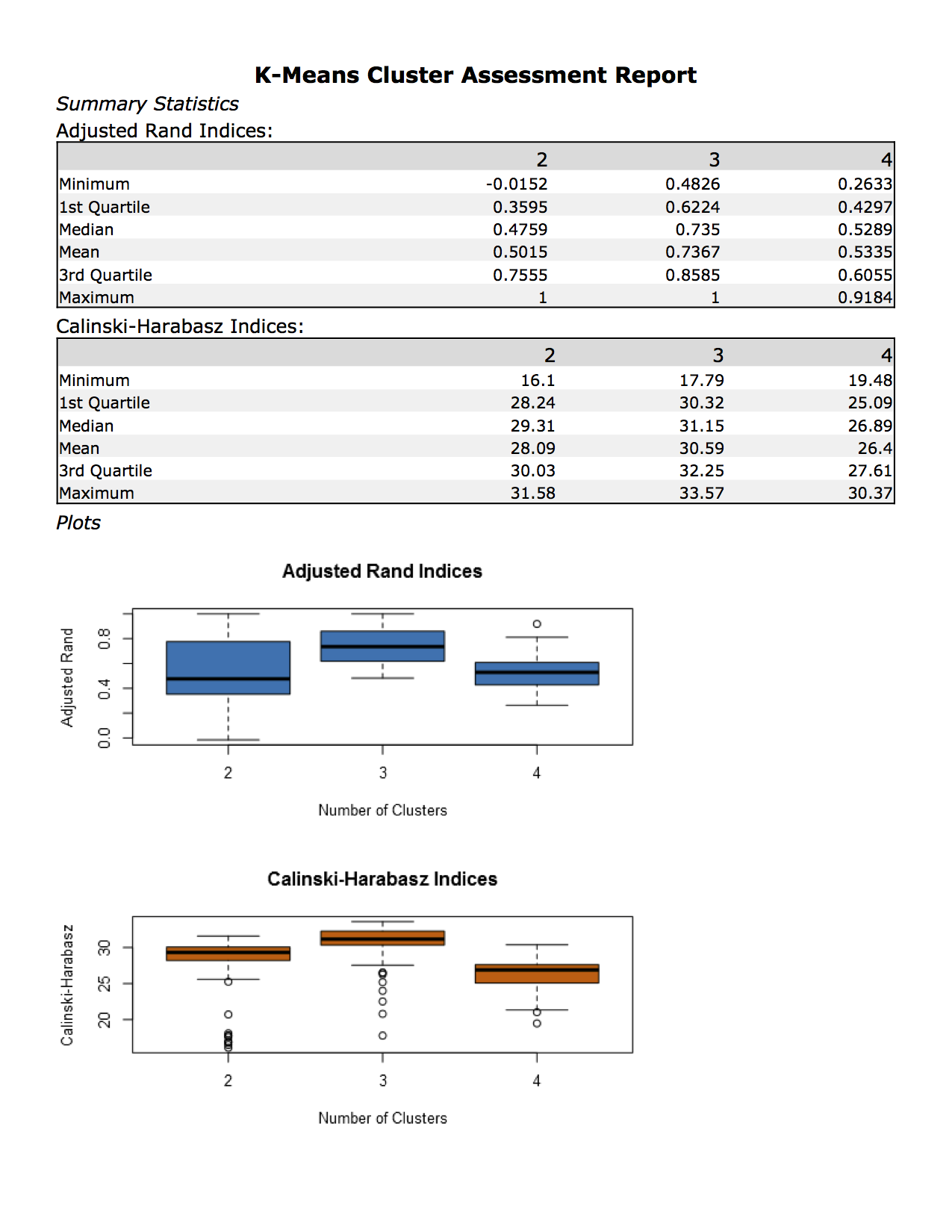
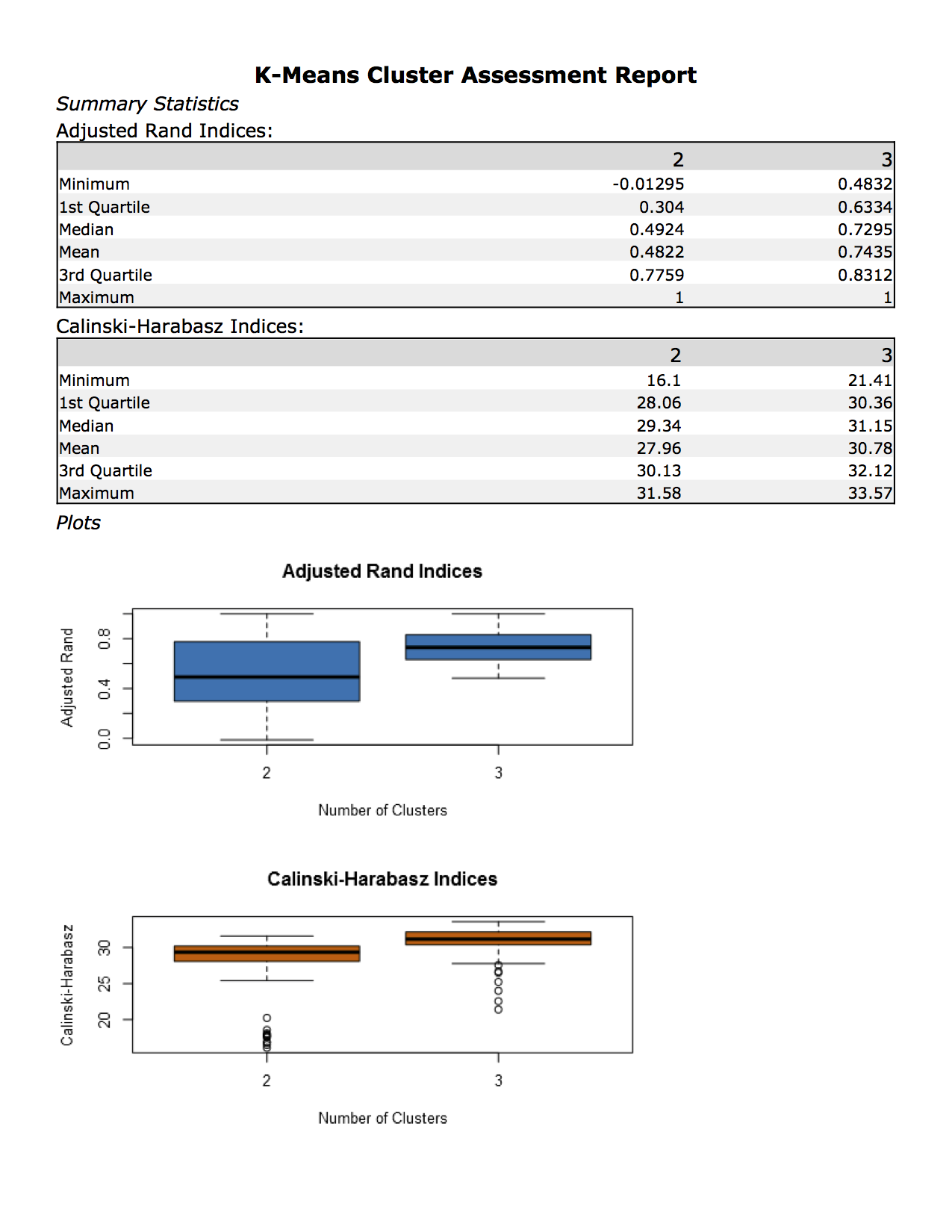
The company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products.

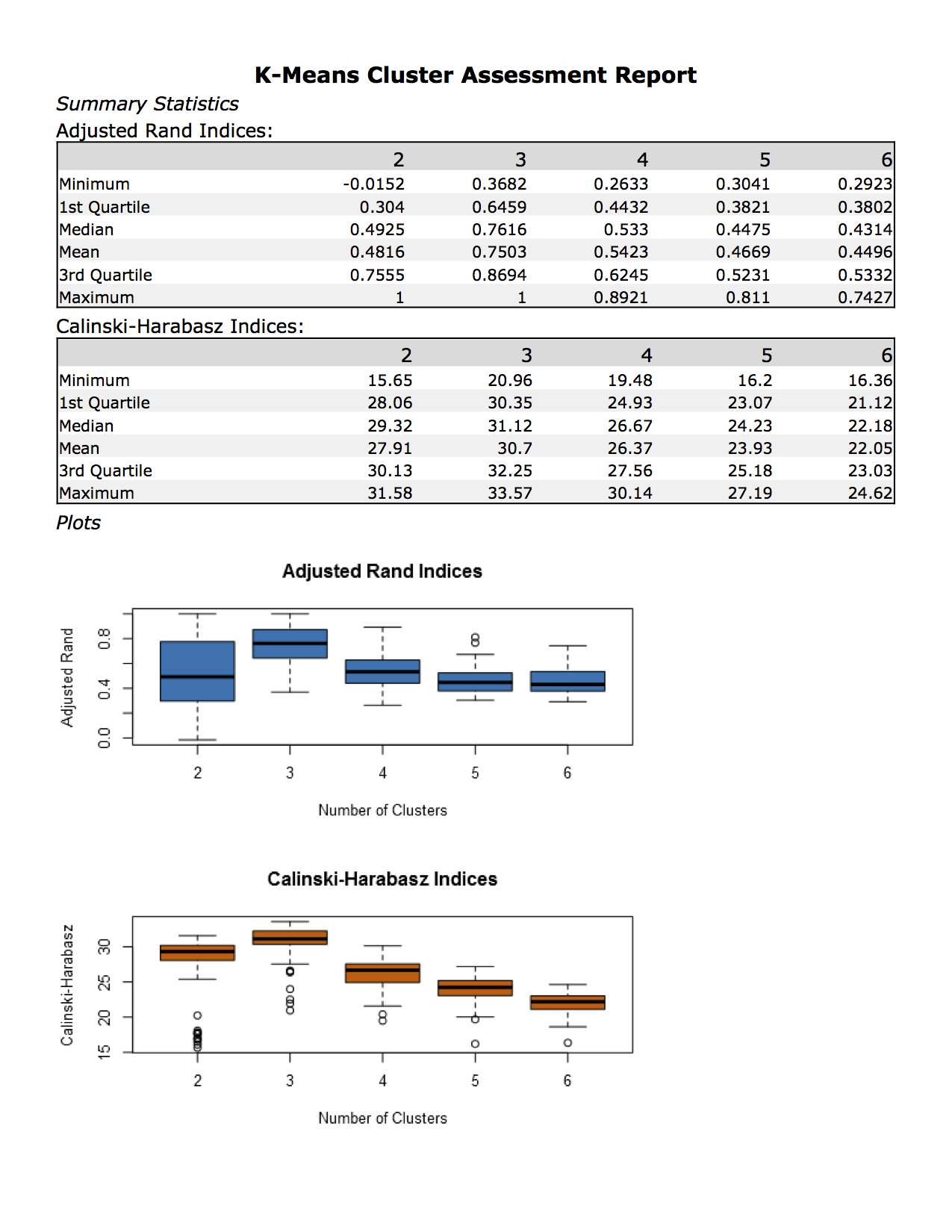
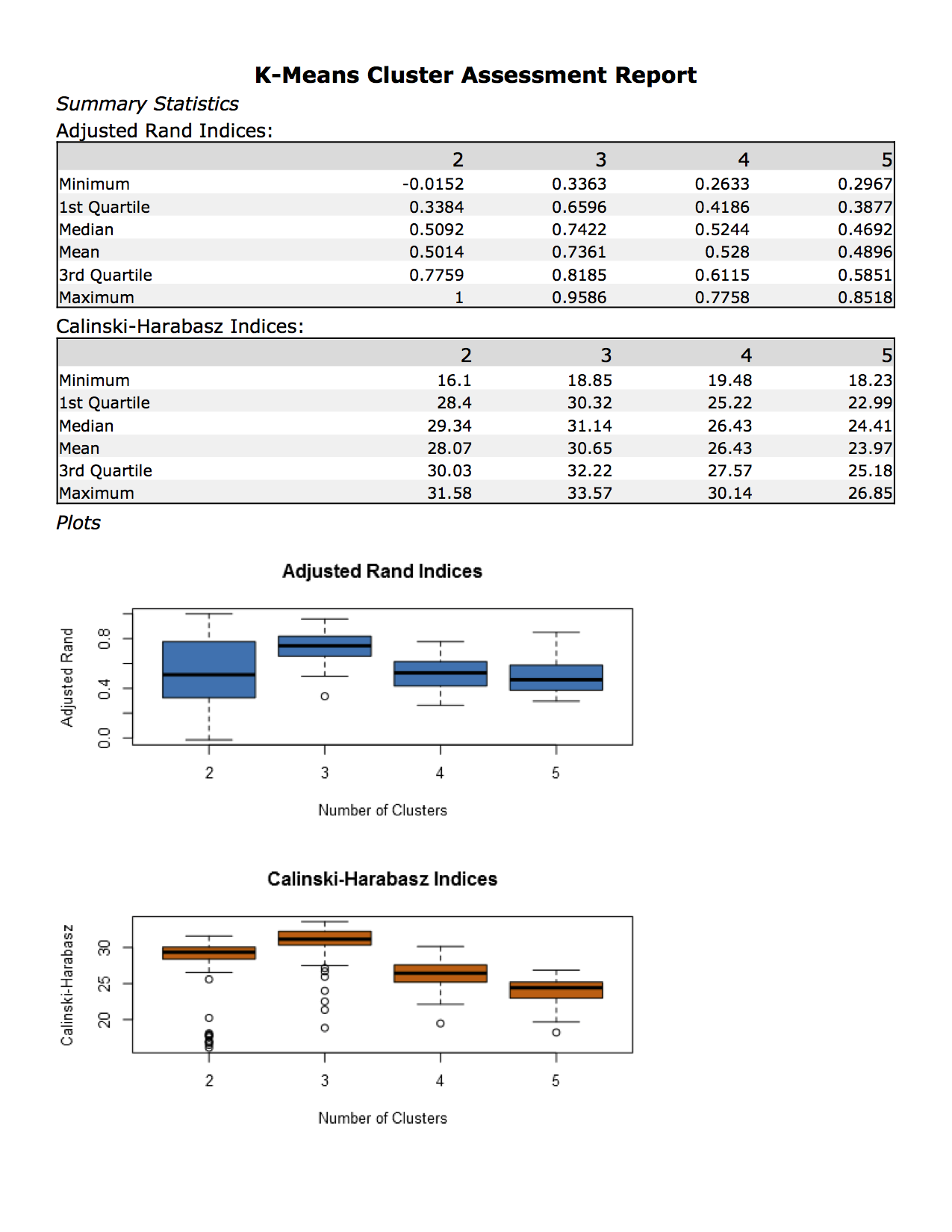
Up until now, the company has treated all stores the same, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others.

**What is the optimal number of store formats?**

In order to determine the best number of store formats to use. A k-centroids analysis was done using k-means clustering method for k=3,4,5,6.

Below are the results:





From the above diagnostics, cluster 3 seems to be the best cluster and therefore I will use cluster 3 as my base to compare the number of k terms. Comparing cluster 3, between all the differing k terms, it looks like using k=3 for my analysis offers the best results with cluster 3 having the tightest range and highest mean when k=3.

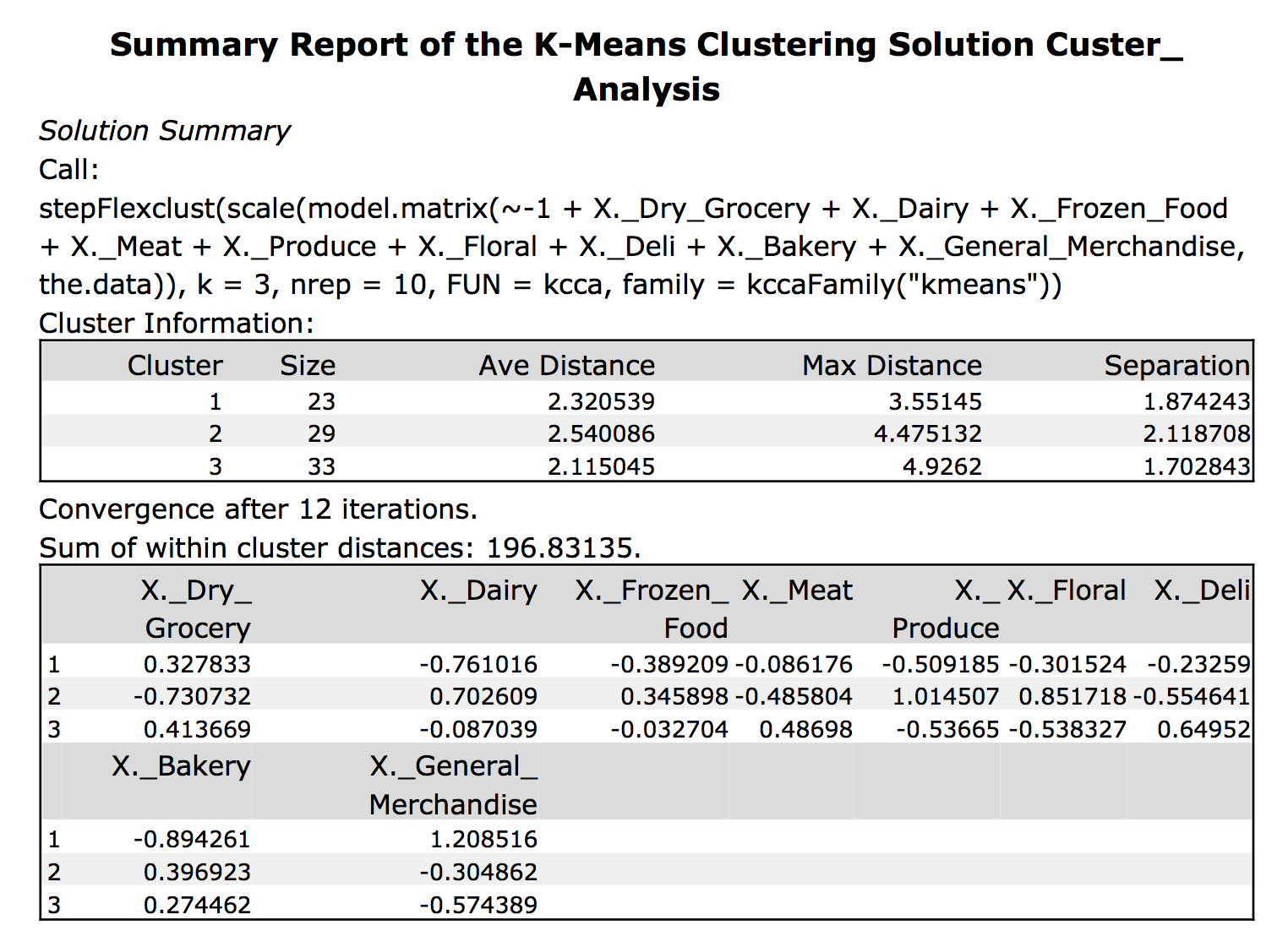
The optimal number of store formats in my opinion is 3.

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

|  |  |
| --- | --- |
| **Store Format** | **Number of stores** |
| 1 | 23 |
| 2 | 29 |
| 3 | 33 |

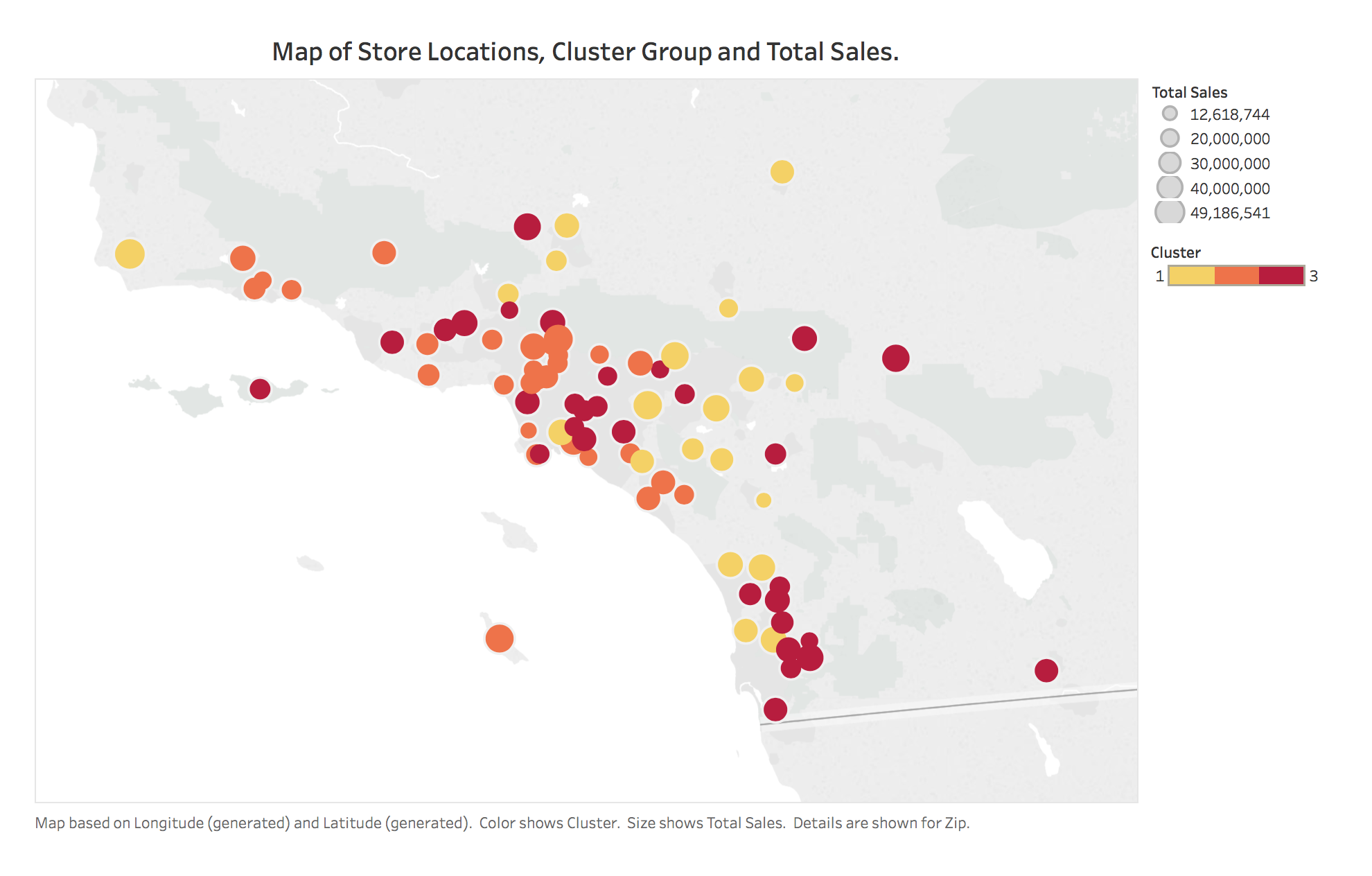
**Based on the results of the clustering model, what is the one way that the clusters differ from one another?**

Below is the summary of the cluster analysis for k=3.



Based on the results above, I can see cluster 1 is most positive for percentage of general merchandise sales vs cluster 3 which is the most negative. This would indicate that these two clusters are the most different in terms of the variable for percentage of general merchandise sales.

**Tableau Visualization of the location of the stores, showing clusters and size due to total sales of each store.**



**Task 2: Formats for New Stores.**

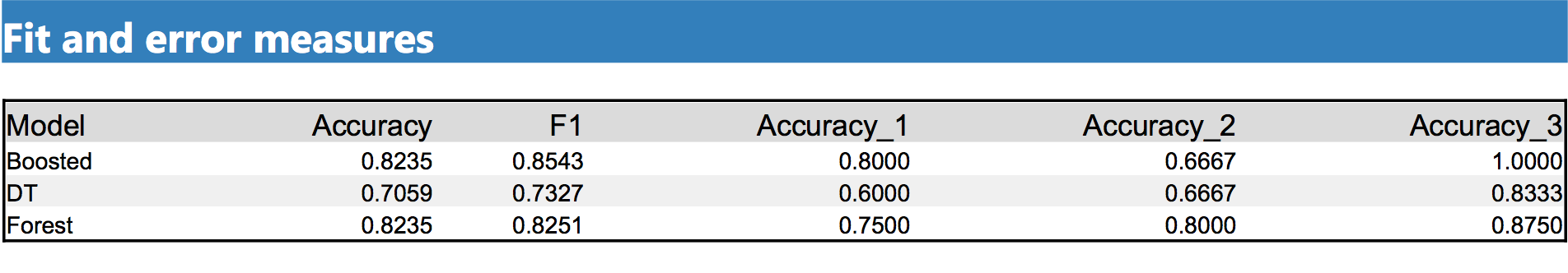
**The Business Scenario: New Stores**

The grocery store chain has 10 new stores opening up at the beginning of the year. The company wants to determine which store format each of the new stores should have. However, we don’t have sales data for these new stores yet, so we’ll have to determine the format using each of the new store’s demographic data.

**What methodology did you use to predict the best store format for the new stores?**

In order to predict the store formats for the new stores, demographic data from StoreDemographicData.csv was used. All the variables were kept as predictor variables and run through a boosted, decision tree and random forest model. An 80/20 split of the data was used for training and validating the models.

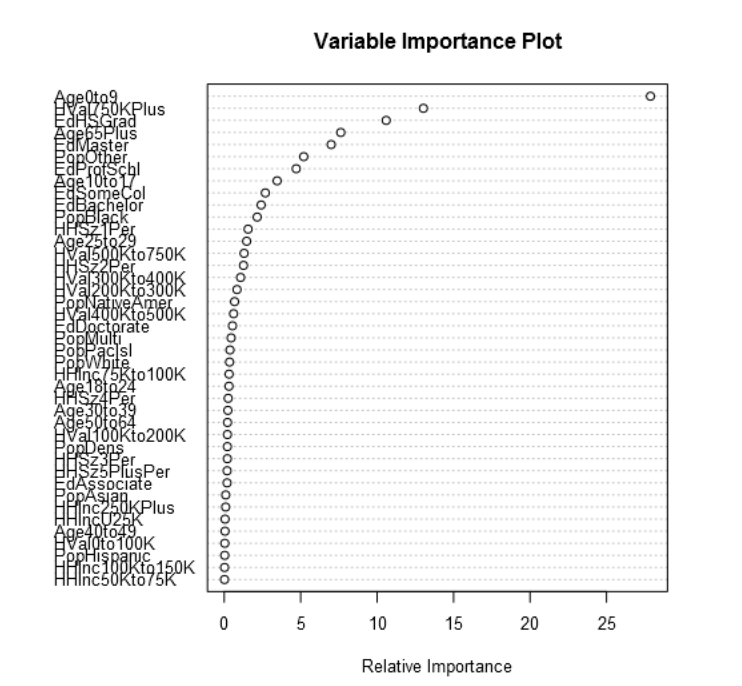
**Below is a comparison of the models:**



Based on the above results, I have decided to use the Boosted, even though the forest model and boosted model both have the same accuracy score, I have used the higher F1 score of the Boosted model as my deciding factor.

**What are the three most important variables that help explain the relationship between demographic indicators and store formats?**

Below is the variable importance plot for the Boosted model chosen for the final prediction.



Based on the above plot, the 3 most important variables for the Boosted model are:

|  |
| --- |
| **Variable** |
| Age0to9 |
| HVal750KPlus |
| EdHSGrad |

**What format do each of the 10 new stores fall into?**

The model gave the new store predictions below:

|  |  |
| --- | --- |
| **Store** | **Store\_Format** |
| S0086 | 3 |
| S0087 | 2 |
| S0088 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 1 |
| S0092 | 2 |
| S0093 | 1 |
| S0094 | 2 |
| S0095 | 2 |

**Task 3: Predicting Product Sales.**

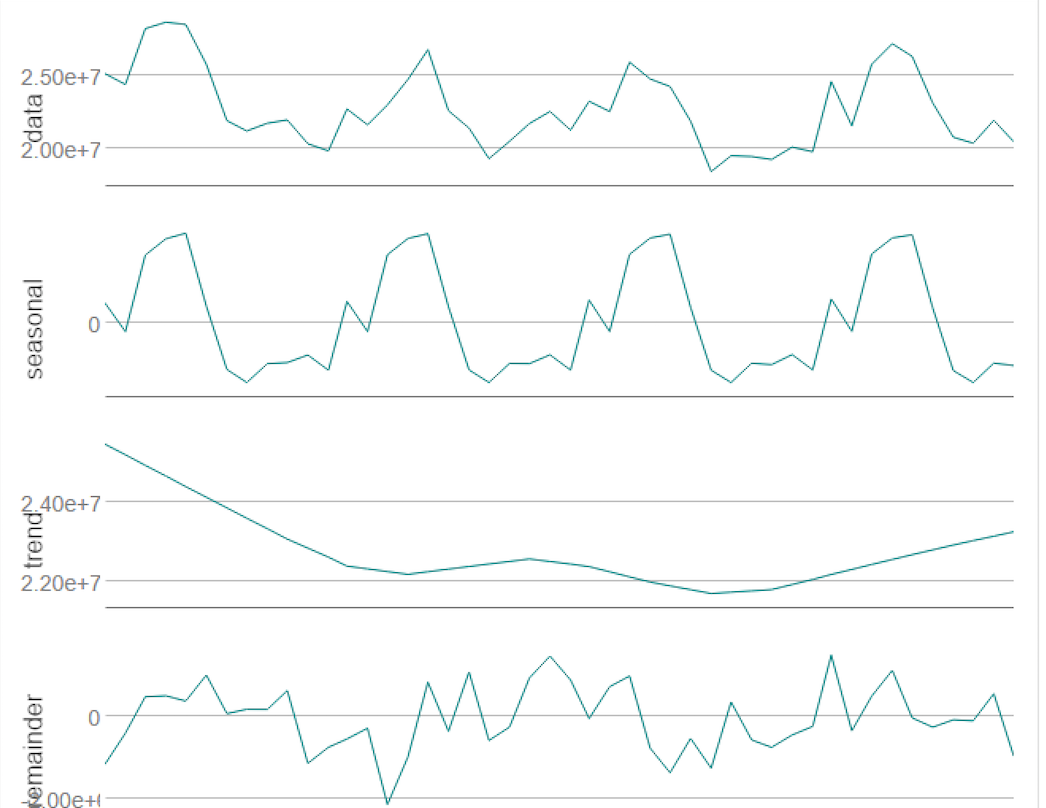
**The Business Scenario: Forecasting**

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast.

**What type of ETS or ARIMA model did you use for each forecast? How did you come to that decision?**

Both ETS and ARIMA models were run for comparison. Analysis of the initial time series decomposition plots below allowed further analysis of model parameters to be established.

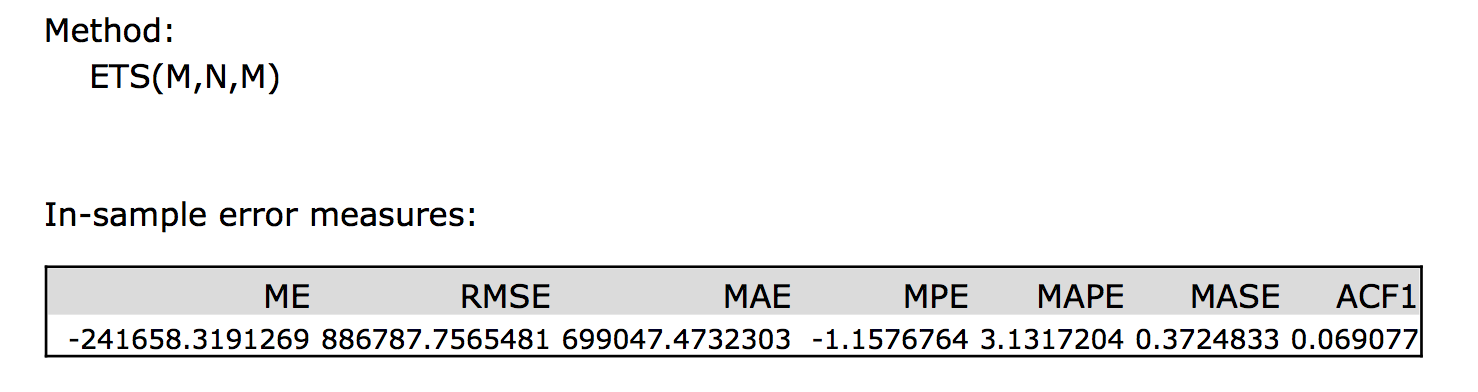
The data used here is sales for produce only per month for all stores aggregated.

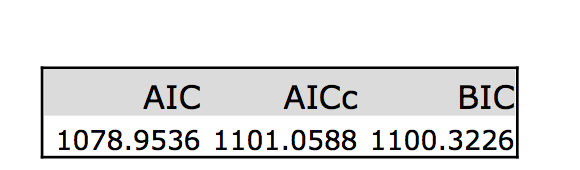


From the above decomposition plots, I can see that the Error element is increasing, Trend element is non-existent and the Seasonal element is also increasing, therefore an ETS(M,N,M) will be used. As for the ARIMA model, I have set the model to calculate the elements automatically.

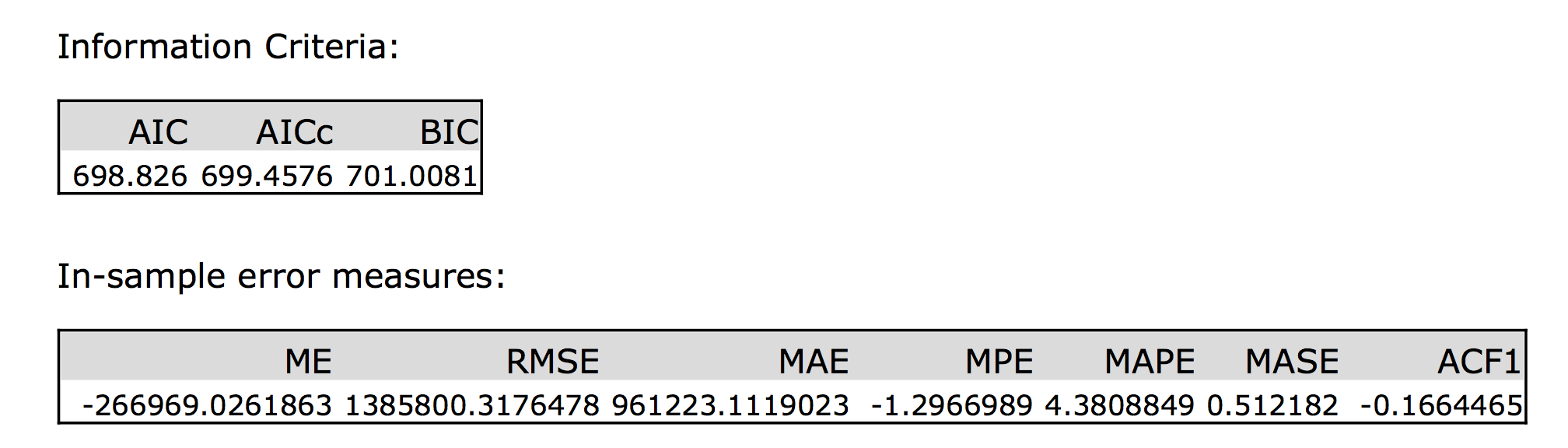
For comparison, a holdout period of 12 periods was used to validate the ETS and ARIMA model.

Below is the ETS(M,N,M) in-sample summary.





The ARIMA(1,0,0)(0,1,0)12 model in-sample summary.



The ETS(M,N,M) will be used for forecasting due to the model having lower error values compared to the ARIMA model.

**Provide a tableau dashboard that includes a table and plot of the 3 monthly forecasts; one for existing, one for new and one for all stores.**

Below is a table of sales forecasts for existing stores, new stores and both new and existing stores combined.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Month** | **Year** | **Existing Store Sales Forecast** | **New Store Sales Forecast** | **Combined Store Sales Forecast** |
| 1 | 2016 | 21,381,830.22 | 2,600,354.85 | 23,982,185.07 |
| 2 | 2016 | 21,081,311.62 | 2,505,198.46 | 23,586,510.07 |
| 3 | 2016 | 24,502,171.96 | 2,889,940.32 | 27,392,112.28 |
| 4 | 2016 | 22,352,993.13 | 2,743,927.30 | 25,096,920.43 |
| 5 | 2016 | 25,331,350.65 | 3,110,813.81 | 28,442,164.46 |
| 6 | 2016 | 26,330,255.79 | 3,191,154.55 | 29,521,410.34 |
| 7 | 2016 | 25,715,514.09 | 3,219,369.78 | 28,934,883.87 |
| 8 | 2016 | 23,458,933.07 | 2,852,751.79 | 26,311,684.87 |
| 9 | 2016 | 21,801,458.48 | 2,543,602.66 | 24,345,061.14 |
| 10 | 2016 | 21,509,922.65 | 2,477,331.44 | 23,987,254.09 |
| 11 | 2016 | 22,619,212.99 | 2,569,169.56 | 25,188,382.55 |
| 12 | 2016 | 21,582,321.09 | 2,535,481.94 | 24,117,803.02 |

Below is a tableau plot of the sales forecasts for existing, new and the combined store sales for produce only.

