Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project>

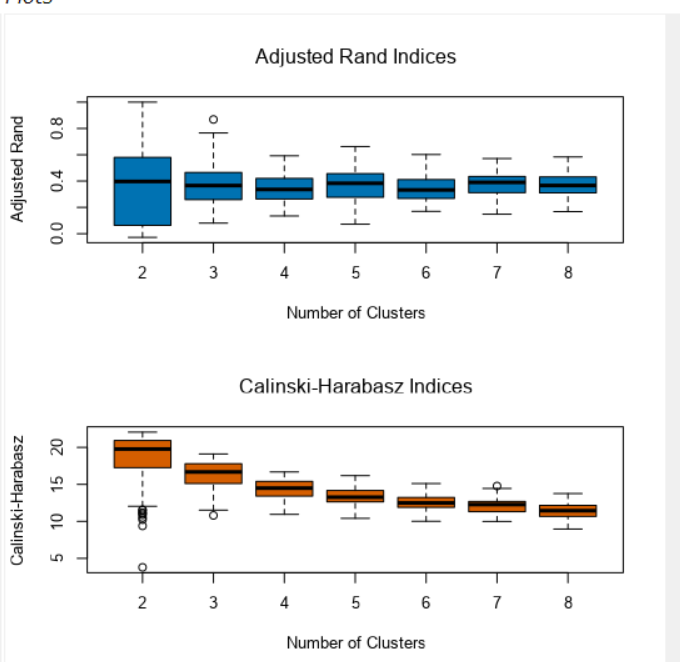
## Task 1: Determine Store Formats for Existing Stores

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

The optimal number of store formats was found to be 3. Based on the 2015 sales data, it was found. The variable percentage of sales by category per store (i.e. each sales category as the percentage of total sales) was standardized by Z-Score to group

Using the K- Centroids Diagnostic tool, the appropriate number of cluster was assessed. K-Means clustering algorithm was used to the method. Measures examined are adjusted Rand Index and Calinski-Harabasz index. The result is shown below.

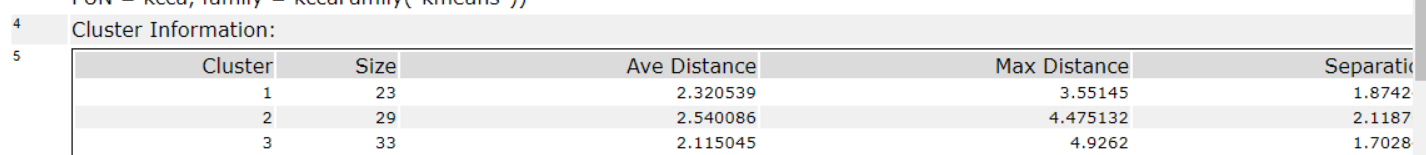
The K-Centroids Diagnostic tool allows an assessment of the appropriate number of clusters. The clustering algorithm selected is K-Means. Two measures examined are the adjusted Rand index and the Calinski-Harabasz index. The graph below shows the ideal number of store formats:



The cluster number is based on each measure correspond to average comparing with the highest median of the solution. Therefore, the ideal storage format would be 3.

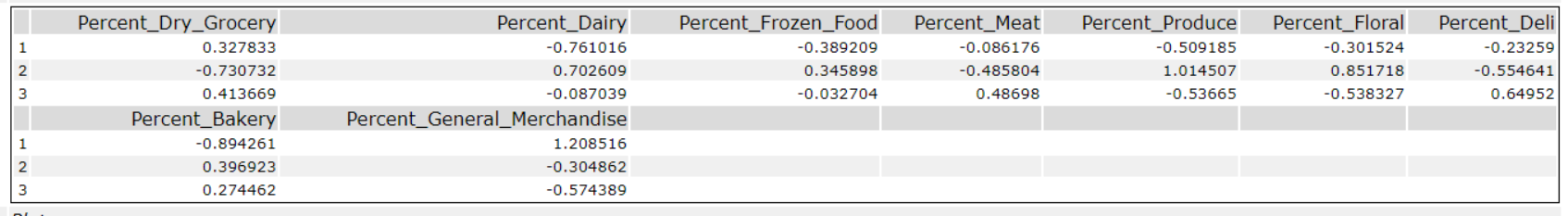
1. How many stores fall into each store format?

K-centered cluster analysis tool was used to find the cluster information. It was found that 23 stores are present in cluster 1, 29 stores in cluster 2 and 33 stores in clulster3.



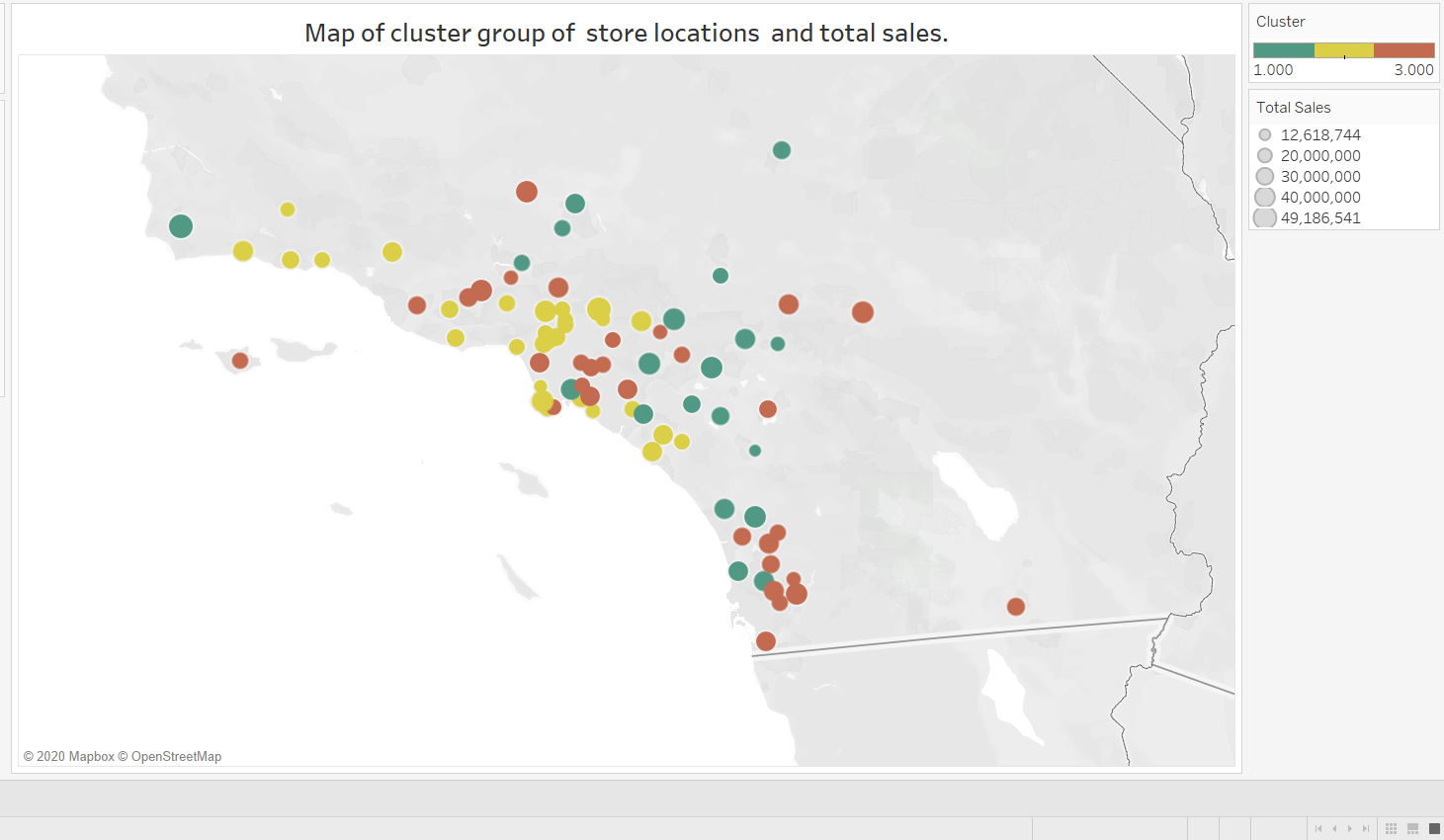
1. Based on the results of the clustering model, what is one way that the clusters differ from one another?

The below is the report genertated by the K-centroid cluster analysis tool.



We can infer that cluster1 stores are generalized by high percent\_general \_merchandise in comparision to cluster 2 and cluster3. Similarly, cluster 2 are generally characterized by high percent\_produce in comparision to cluster1 and cluster3.

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.



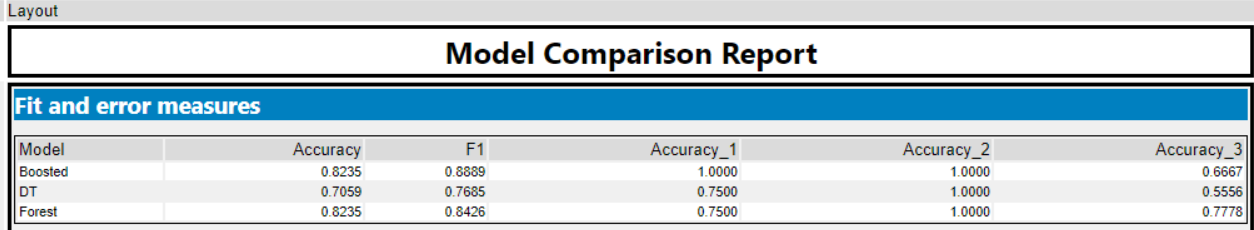
## Task 2: Formats for New Stores

10 new stores of the grocery chain is going to open in the starting of the year. Since, we don’t have any sales data, we will determine the store format using the demographic data present

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.)

To predict the best score format for the new stores, we used the demographic data to run against boosted model, decision tree model and the random forest model. Split of 80-20 was done for training and validation data Then the model comparison tool was used to compare the model.

The result of the model comparison tool is shown below:



Since, the boosted model and the forest model have same accuracy score, I choose to go with the boosted model as the F1 score of the boosted model was higher.

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 | 3 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 1 |
| S0092 | 2 |
| S0093 | 1 |
| S0094 | 2 |
| S0095 | 2 |

## Task 3: Predicting Produce Sales

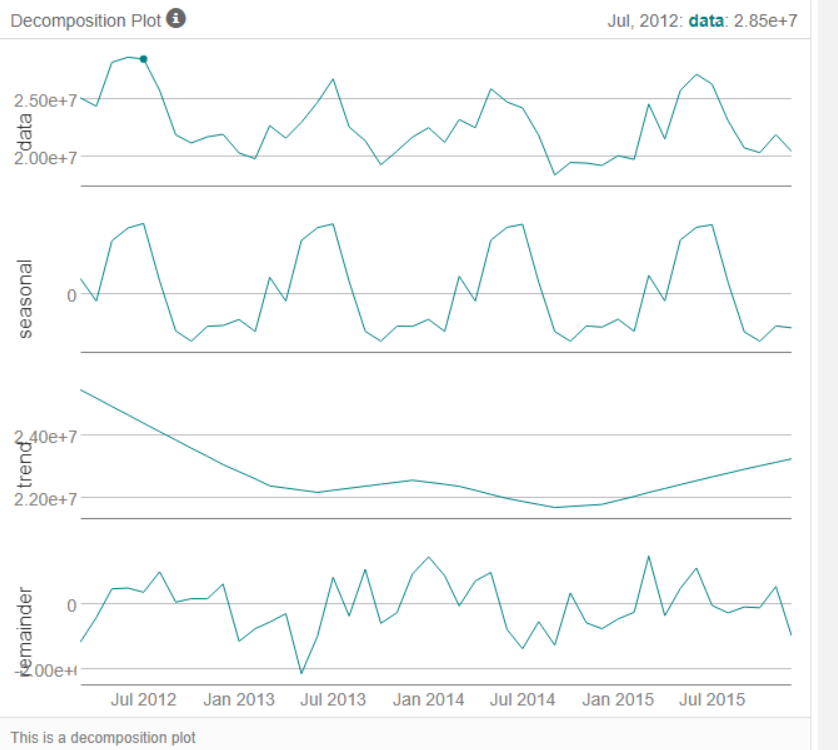
Since, the fresh products have short life span and high cost, we are requested to have an accurate monthy sales forecast

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?

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.

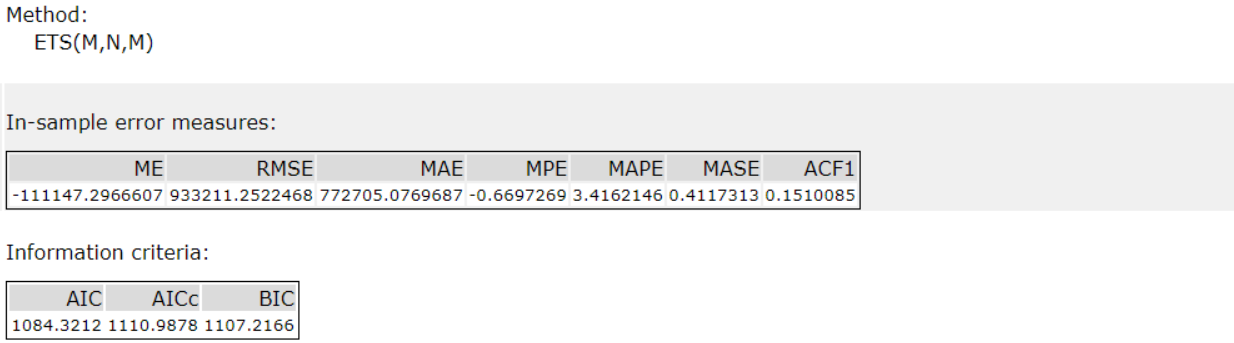
For the forecast, we applied and compare both the ETS and ARIMA model. By analyzing the initial time series decomposition plot which is shown below, we could further find the model parameters.

We are using sales for produce / month for all the stores aggregated.

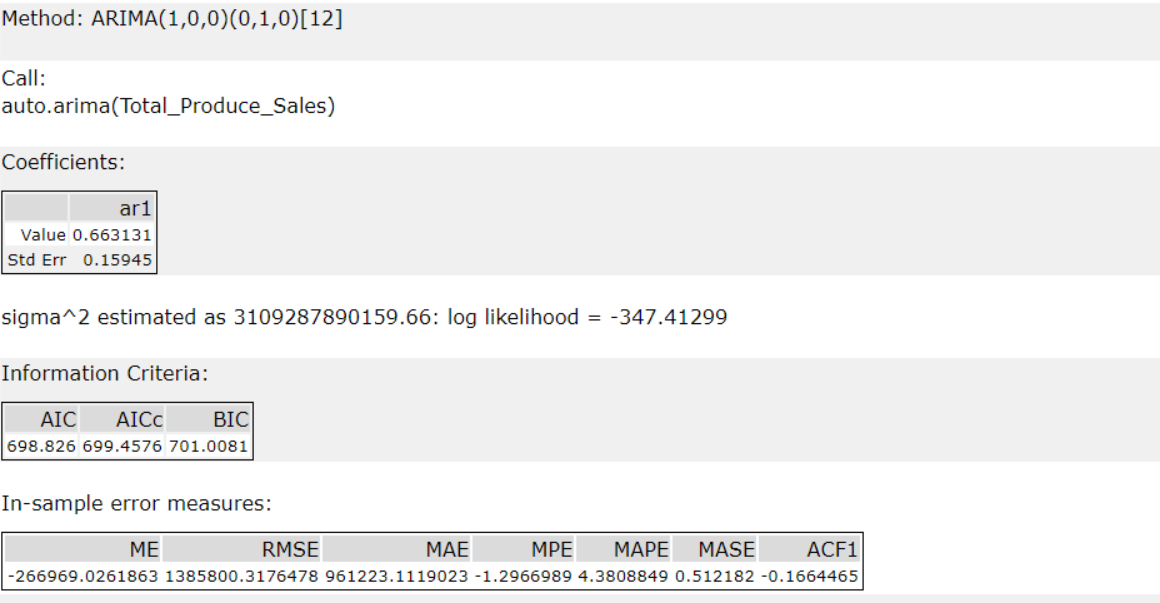


The decomposition plot shows that there is increase in the error element, the trend is non existential and also, the seasonal trend is increasing. Hence, I have used ETS(M,N,M) model. For the ARIMA model, i had set it to auto. Both the model had a holdout period of 12.

For ETS(M,N.M) model:

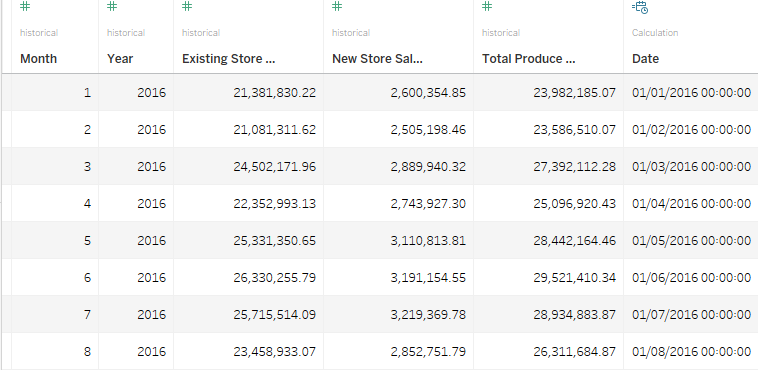


For ARIMA(1,0,0)(0,1,0)12 selected automatically :



We select the ETS(M,N,M) as the forecasting model as the ETS model showed less error valued in comparisons to the ARIMA model

The table of forecast of existing and new stores is given below:



Also, we have made a tabeleau plot for the store sales produce to forecast existing and the new store sales. This is shown below.

