# 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 StoreSalesData.csv file contains sales by product category for all existing stores from 2012 through to 2015. I prepared the data to reflect the percentage sales per category per store so as to obtain the right metric for clustering. I then applied the K-Centroid Diagnostics tool using the K-Means clustering method to determine the optimal number of clusters. The Summary Statistics, showing the Adjusted Rand Indices and the Calinski-Harabasz Indices, are as shown below;

**K-Means Cluster Assessment Report**

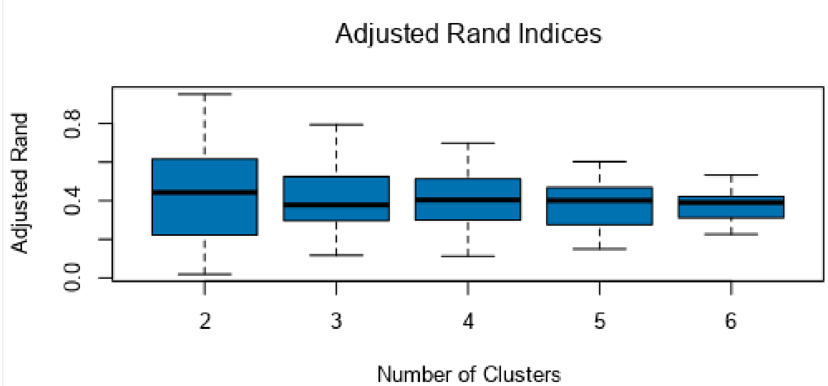
*Summary Statistics*

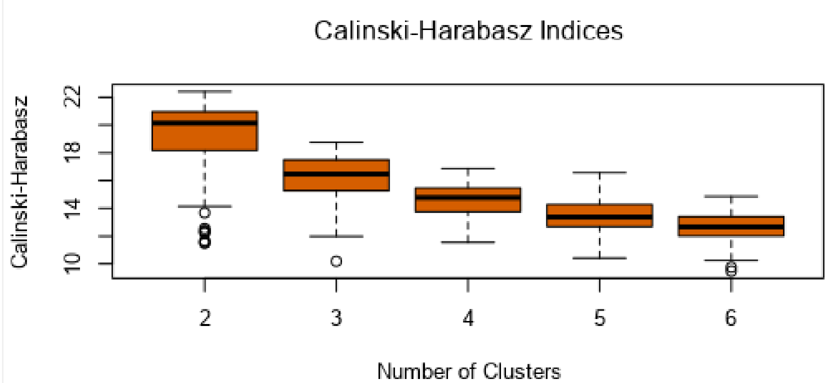
Adjusted Rand Indices:

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | 0.020011 | 0.116722 | 0.113112 | 0.150418 | 0.226486 |
| 1st Quartile | 0.225885 | 0.297259 | 0.300394 | 0.278121 | 0.31428 |
| Median | 0.443086 | 0.377155 | 0.403137 | 0.400518 | 0.390769 |
| Mean | 0.430858 | 0.421041 | 0.403641 | 0.3825 | 0.377712 |
| 3rd Quartile | 0.607523 | 0.525492 | 0.511782 | 0.468717 | 0.421245 |
| Maximum | 0.952115 | 0.794667 | 0.698784 | 0.602951 | 0.532821 |

Calinski-Harabasz Indices:

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | 11.49694 | 10.18405 | 11.55369 | 10.41516 | 9.47976 |
| 1st Quartile | 18.25951 | 15.28341 | 13.77306 | 12.69596 | 12.01426 |
| Median | 20.14522 | 16.47868 | 14.77552 | 13.36888 | 12.66809 |
| Mean | 19.09552 | 16.27797 | 14.57626 | 13.43979 | 12.64158 |
| 3rd Quartile | 20.94642 | 17.45689 | 15.46508 | 14.25764 | 13.39232 |
| Maximum | 22.41555 | 18.75042 | 16.86351 | 16.57168 | 14.8625 |



From the indices above, cluster number 2 could be a good choice because it has higher medians on both the Adjusted Rand(AR) Index and the Calinski-Harabasz(CH) Index. However, the spread of the interquartile range is quite high and it seems loose compared to other clusters.

Comparing other clusters, cluster number 3 seems to be a good choice. This is so because, it has higher mean and median on the CH index and higher mean on the AR index with slightly lower median. The spread over the interquartile range also shows that the distribution is fairly compact. Since the aim of this diagnosis is to determine the cluster with high level of stability, distinctness and compactness, I will be using 3 cluster to build my cluster model.

2. How many stores fall into each store format?
   - I applied the K-Centroid Cluster Analysis tool to the prepared store sales data using the K-means clustering method and 3 as the number of clusters. The result of the cluster solution is a shown below;

**Summary Report of the K-Means Clustering Solution Clusters**

*Solution Summary*

Call:
stepFlexclust(scale(model.matrix(~-1 + Perc_Dry_Grocery_sales + Perc_Dairy_Sales + Perc_Frozen_Food_Sales + Perc_Meat_Sales + Perc_Produce_Sales + Perc_Floral_Sales + Perc_Deli_Sales + Perc_Bakery_Sales + Perc_GenMerch_Sales, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 25 | 2.099985 | 4.823871 | 2.191566 |
| 2 | 35 | 2.475018 | 4.412367 | 1.947298 |
| 3 | 25 | 2.289004 | 3.585931 | 1.72574 |

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

| | Perc_Dry_Grocery_sales | Perc_Dairy_Sales | Perc_Frozen_Food_Sales | Perc_Meat_Sales | Perc_Produce_Sales | Perc_Floral_Sales | Perc_Deli_Sales |
|---|---|---|---|---|---|---|---|
| 1 | 0.528249 | -0.215879 | -0.261597 | 0.614147 | -0.655028 | -0.663872 | 0.824834 |
| 2 | -0.594802 | 0.655893 | 0.435129 | -0.384631 | 0.812883 | 0.71741 | -0.46168 |
| 3 | 0.304474 | -0.702372 | -0.347583 | -0.075664 | -0.483009 | -0.340502 | -0.178482 |

| | Perc_Bakery_Sales | Perc_GenMerch_Sales |
|---|---|---|
| 1 | 0.428226 | -0.674769 |
| 2 | 0.312878 | -0.329045 |
| 3 | -0.866255 | 1.135432 |

The cluster solution above shows that, out of the 85 existing stores;
25 stores fall into Cluster 1
35 stores fall into Cluster 2
25 stores fall into Cluster 3

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   - Based on the result I obtained from the clustering solution below, I observed that some items in stores belonging so some clusters sell more than others.

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

| | Perc_Dry_Grocery_sales | Perc_Dairy_Sales | Perc_Frozen_Food_Sales | Perc_Meat_Sales | Perc_Produce_Sales | Perc_Floral_Sales | Perc_Deli_Sales |
|---|---|---|---|---|---|---|---|
| 1 | 0.528249 | -0.215879 | -0.261597 | 0.614147 | -0.655028 | -0.663872 | 0.824834 |
| 2 | -0.594802 | 0.655893 | 0.435129 | -0.384631 | 0.812883 | 0.71741 | -0.46168 |
| 3 | 0.304474 | -0.702372 | -0.347583 | -0.075664 | -0.483009 | -0.340502 | -0.178482 |

| | Perc_Bakery_Sales | Perc_GenMerch_Sales |
|---|---|---|
| 1 | 0.428226 | -0.674769 |
| 2 | 0.312878 | -0.329045 |
| 3 | -0.866255 | 1.135432 |

For example, I can see that the item categories of Dairy, Frozen_Food, Produce and Floral all have the highest positive values in cluster 2 compared to the other two clusters. This is an indication that the items sell more in cluster 2.

Also, the item categories of Dry_Grocery, Meat, Deli and bakery tend to sell more in cluster 1 compared to the other two clusters, while the GenMerch category sells more in cluster 3.

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.
    - Provided below is an image showing the location of stores and the clusters they fall into;

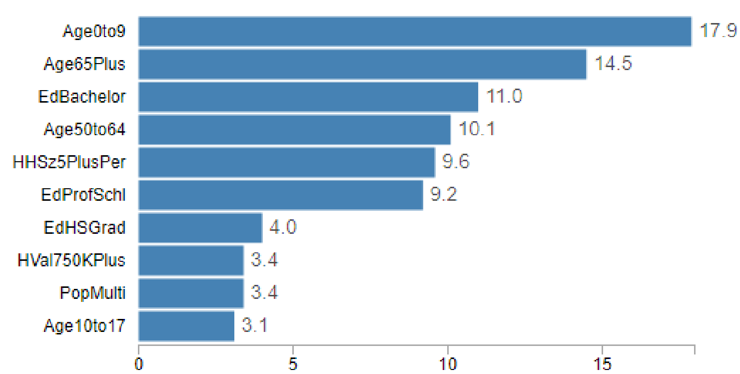
# Task 2: Formats for New Stores

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.)
    - 10 new stores were set to be opened at the beginning of the year 2016 and there is need to predict which store format each new store would fall into based on the demographic data surrounding each store. Since the target variable to be predicted is categorical (clusters) with more than two outcomes, a non-binary classification model is most suitable.

      In order to achieve this, I set aside 20% of the existing data for the purpose of validating the classification models and created a Decision Tree model, a Forest Tree model and a Boosted model. The results of the classification models are as shown below;

**Decision Tree:**

Variable Importance




Confusion Matrix

| Actual \ Predicted | 1 | 2 | 3 | Sum | Accuracy |
|---|---|---|---|---|---|
| 1 | 15 | 0 | 2 | 17 | 88% |
| 2 | 3 | 26 | 1 | 30 | 87% |
| 3 | 6 | 0 | 15 | 21 | 71% |
| Sum | 24 | 26 | 18 | 68 | 82% |

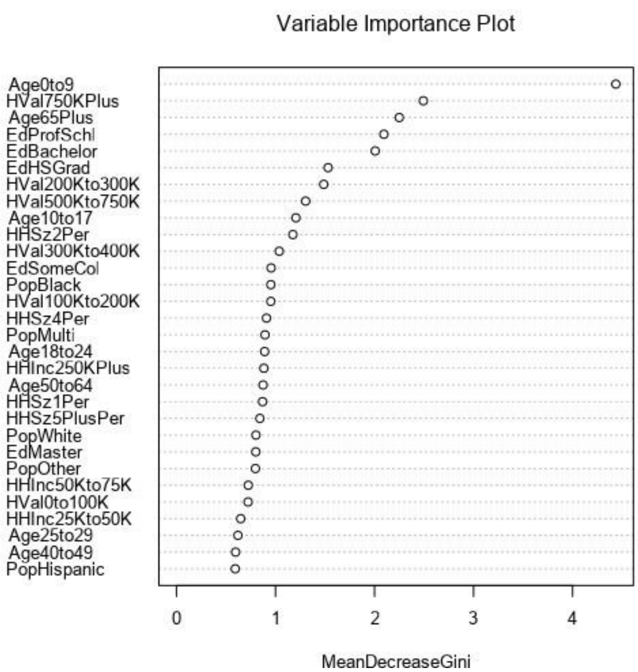
From the Decision Tree report above, the variable importance plot shows that **Age0to9, Age65Plus and EdBachelor** are the three most important variables used in the Decision tree model. Also, the confusion matrix indicates that 82% of the variables were classified correctly overall. I also observed that 88% of cluster 1, 87% of cluster 2 and 71% of cluster 3 were predicted correctly. This model seem to be fairly strong but will have to be compared with other models to see how well it performs against the validation sample.

**Forest Tree:**

| Record | Report |
|---|---|
| 1 | *Basic Summary* |
| 2 | Call:<br>randomForest(formula = Cluster ~ Age0to9 + Age10to17 + Age18to24 + Age25to29 + Age30to39 + Age40to49 + Age50to64 + Age65Plus + EdLTHS + EdHSGrad + EdSomeCol + EdAssociate + EdBachelor + EdMaster + EdProfSchl + EdDoctorate + HHSz1Per + HHSz2Per + HHSz3Per + HHSz4Per + HHSz5PlusPer + HHIncU25K + HHInc25Kto50K + HHInc50Kto75K + HHInc75Kto100K + HHInc100Kto150K + HHInc150Kto250K + HHInc250KPlus + PopAsian + PopBlack + PopHispanic + PopMulti + PopNativeAmer + PopOther + PopPacIsl + PopWhite + HVal0to100K + HVal100Kto200K + HVal200Kto300K + HVal300Kto400K + HVal400Kto500K + HVal500Kto750K + HVal750KPlus + PopDens, data = the.data, ntree = 500, replace = TRUE) |
| 3 | Type of forest: classification<br>Number of trees: 500<br>Number of variables tried at each split: 6 |
| 4 | OOB estimate of the error rate: 17.6% |
| 5 | Confusion Matrix: |
| 6 | |

| | Classification Error | 1 | 2 | 3 |
|---|---|---|---|---|
| 1 | 0.294 | 12 | 2 | 3 |
| 2 | 0.067 | 1 | 28 | 1 |
| 3 | 0.238 | 3 | 2 | 16 |

From the Forest Tree report summary above, I can see that the Out of the bag error is 17.6% which is quite on the high side. Also, the Classification errors for clusters 1 (29.4%) and 3 (23.8%) are significantly high compared to that of cluster 2 (6.7%).



Also, from the Variable importance plot above, **Age0to9, HVal750kPlus and Age65Plus** appear to be the most important predictor variables used in the model.

## Boosted Model:

1

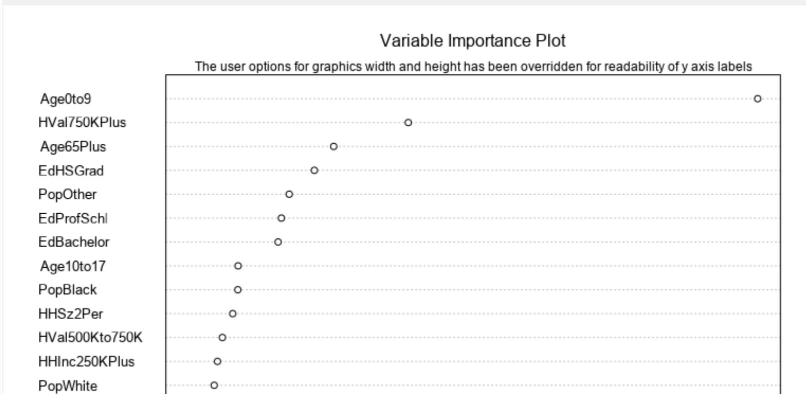
**Report for Boosted Model Boosted_Model**

Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1829

2

Plots:




From the variable importance plot of the boosted model, the three most important variables used in the model are Age0to9, HVal750KPlus and Age65Plus. The number of trees used in the iterative process is 4000 which is quite a lot and gives room for a lot more accuracy in prediction.

- In order to identify the best classification model to use for predicting the format the 10 new stores fall into, I carried out a comparison between the Decision Tree model, Forest Tree model and the Boosted model against the validation sample using the Model Comparison tool. The result of the comparison is as shown below;

**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| Forest_Model | 0.7059 | 0.7500 | 0.5000 | 1.0000 | 0.7500 |
| Decision_Tree | 0.7059 | 0.7083 | 0.6250 | 1.0000 | 0.5000 |
| Boosted_Model | 0.7647 | 0.8333 | 0.5000 | 1.0000 | 1.0000 |

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

**Confusion matrix of Boosted_Model**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 4 | 0 | 0 |
| Predicted_2 | 2 | 5 | 0 |
| Predicted_3 | 2 | 0 | 4 |

**Confusion matrix of Decision_Tree**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 5 | 0 | 2 |
| Predicted_2 | 2 | 5 | 0 |
| Predicted_3 | 1 | 0 | 2 |

**Confusion matrix of Forest_Model**

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 4 | 0 | 1 |
| Predicted_2 | 2 | 5 | 0 |
| Predicted_3 | 2 | 0 | 3 |

From the fit and error measures in the model comparison report, I observed that the Boosted model has the highest overall accuracy of 76.47% compared to the Decision Tree model and the Forest Model which both have overall accuracy of 70.69%. Inspecting the confusion matrices of all three models also shows that the Boosted model did a better job in predicting the three formats.

- Based on the analyses and results above, I chose to use the **Boosted Model** method to predict the best formats the 10 new stores will fall into.

2. What format do each of the 10 new stores fall into? Please fill in the table below.
    - I used the Boosted model to score the data of the new stores and provided below is a table showing the format each of the 10 new stores fall into.

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

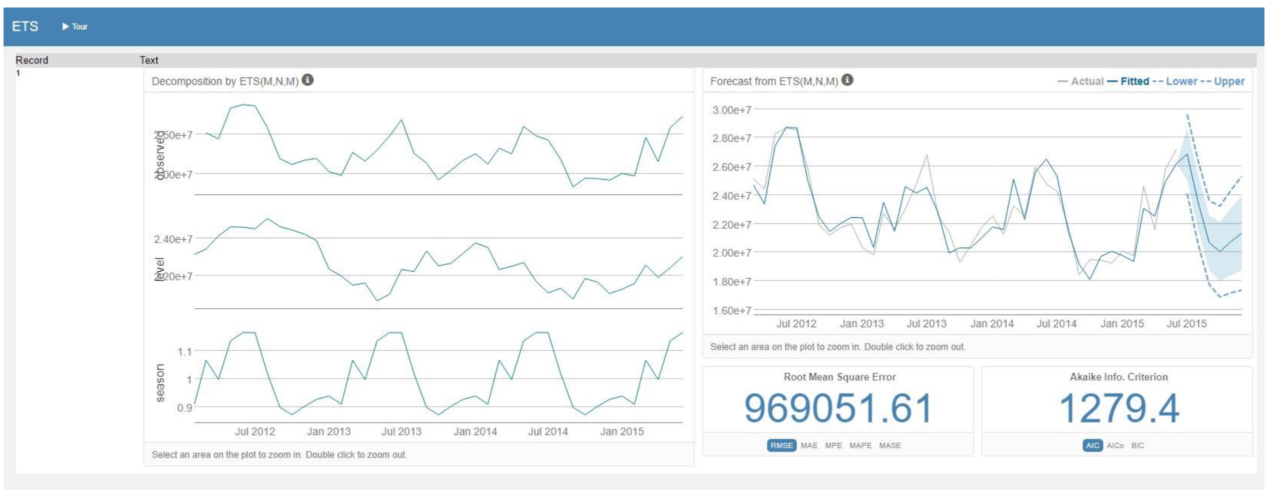
# Task 3: Predicting Produce Sales

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?

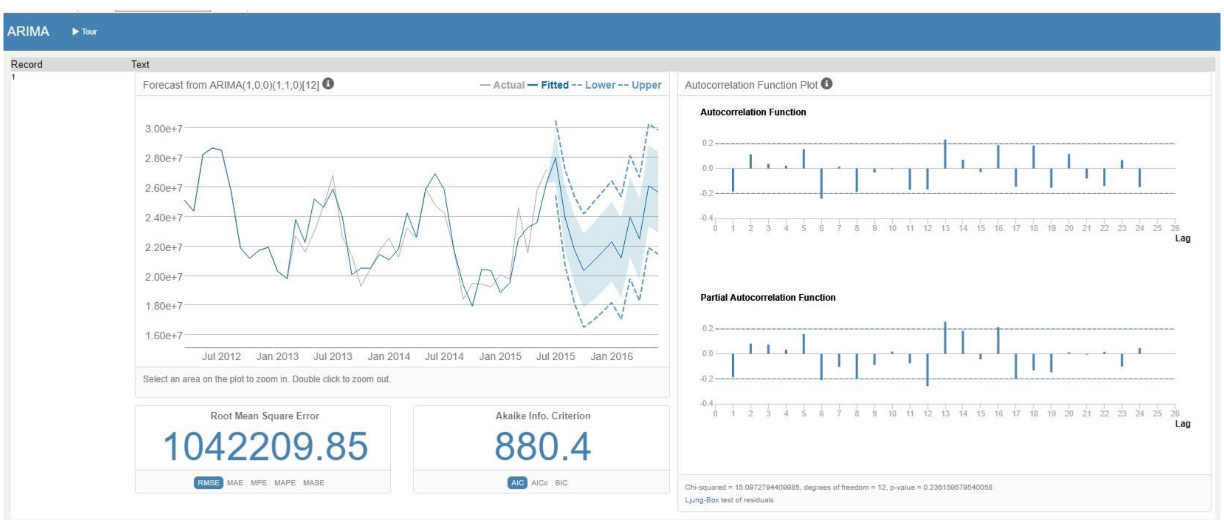
- This section is about predicting the produce sales for both the existing and new stores for the year 2016. For this purpose a time series forecasting will be needed.

- In order to achieve this for the existing stores, I aggregated the existing sales data on produce sales and set aside 6months hold out sample for the purpose of model validation. I created both an ETS and ARIMA models, as shown below, using the Auto settings to enable me achieve the best results for both models.

**ETS Model**




**ARIMA Model**



From the above model outcomes, The ETS model used for the forecast is ETS(M,N,M) while the ARIMA model used is ARIMA(1,0,0)(1,1,0)[12]

- I and further compared the results against the hold out sample using the TS Compare tool. The result of the comparison is as shown below;

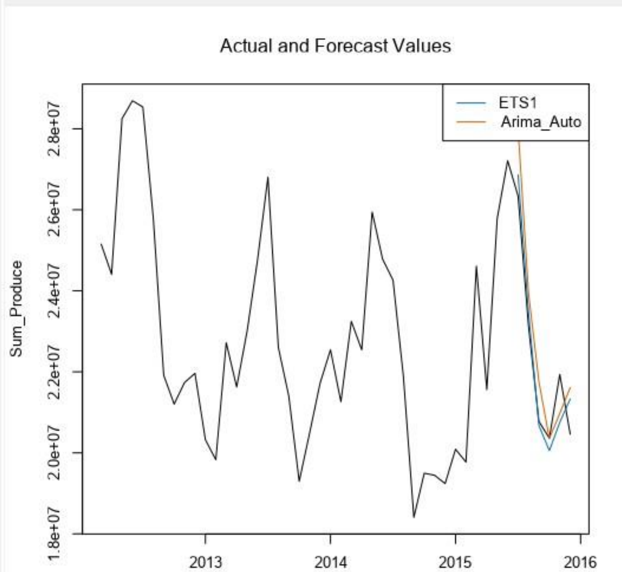
**Comparison of Time Series Models**

Actual and Forecast Values:

| Actual | ETS1 | Arima_Auto |
|---|---|---|
| 26338477.15 | 26860639.57444 | 27997835.63764 |
| 23130626.6 | 23468254.49595 | 23946058.0173 |
| 20774415.93 | 20668464.64495 | 21751347.87069 |
| 20359980.58 | 20054544.07631 | 20352513.09377 |
| 21936906.81 | 20752503.51996 | 20971835.10573 |
| 20462899.3 | 21328386.80965 | 21609110.41054 |

Accuracy Measures:

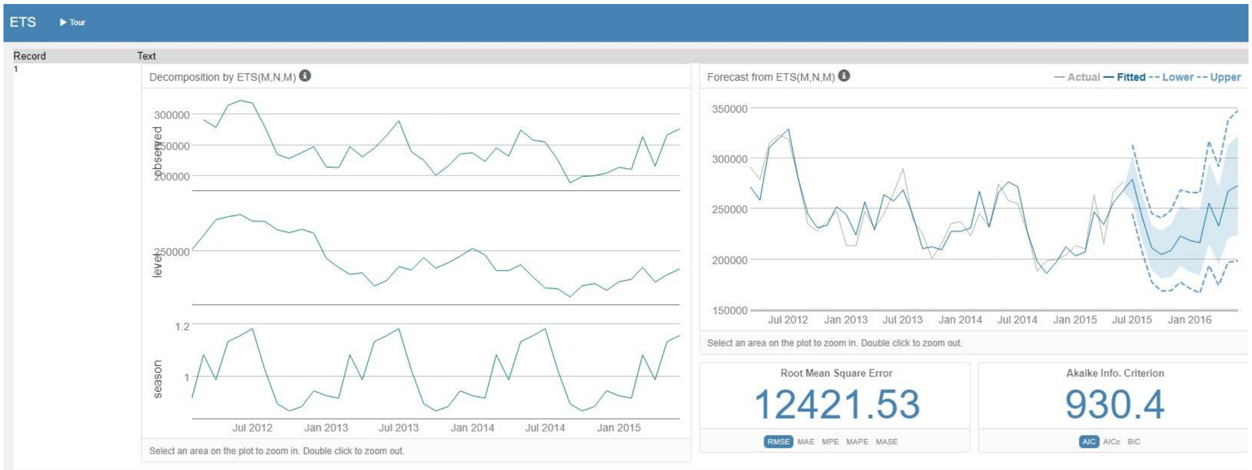
| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS1 | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |
| Arima_Auto | -604232.29 | 1050239.2 | 928412 | -2.6156 | 4.0942 | 0.5463 |




Observing the Actual and Forecasted values of the TS compare result above, the ETS model tends to have more values closer to the actual values compared to that of the ARIMA model. Also, on the Accuracy Measures, The RMSE, MAE, MAPE and the MASE for the ETS model are significantly lower compared to that of the ARIMA model. Overall, I think the ETS model performed better than the ARIMA model and thus, I went ahead using the ETS model to carry out the produce sales forecast for the existing stores.

- In order to achieve the forecasts for the new stores, I aggregated the existing sales data down to the average produce sales per cluster and set aside 6months hold out sample in each cluster for the purpose of model validation. I created both an ETS and ARIMA models for each cluster as shown below;

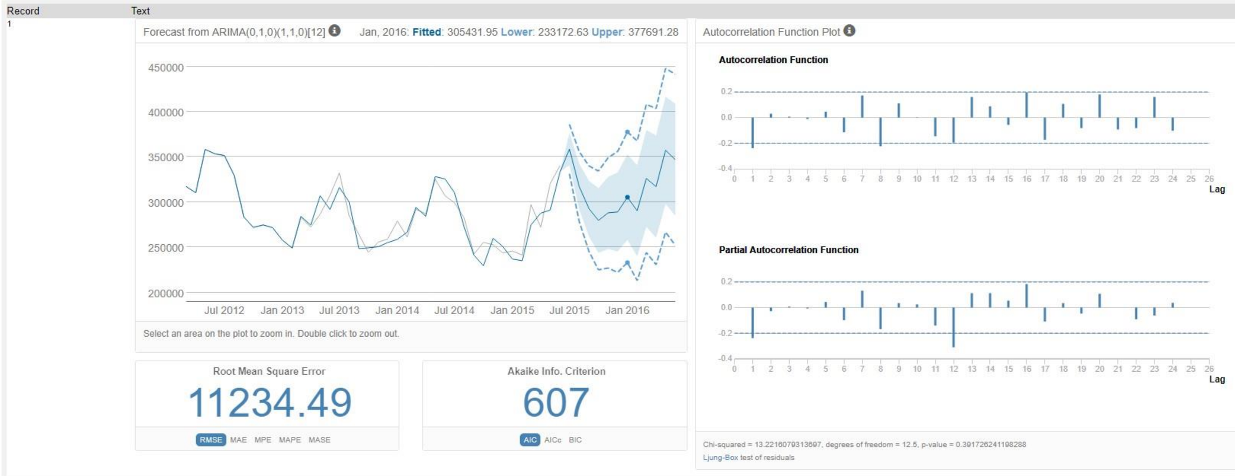
## Cluster 1 ETS



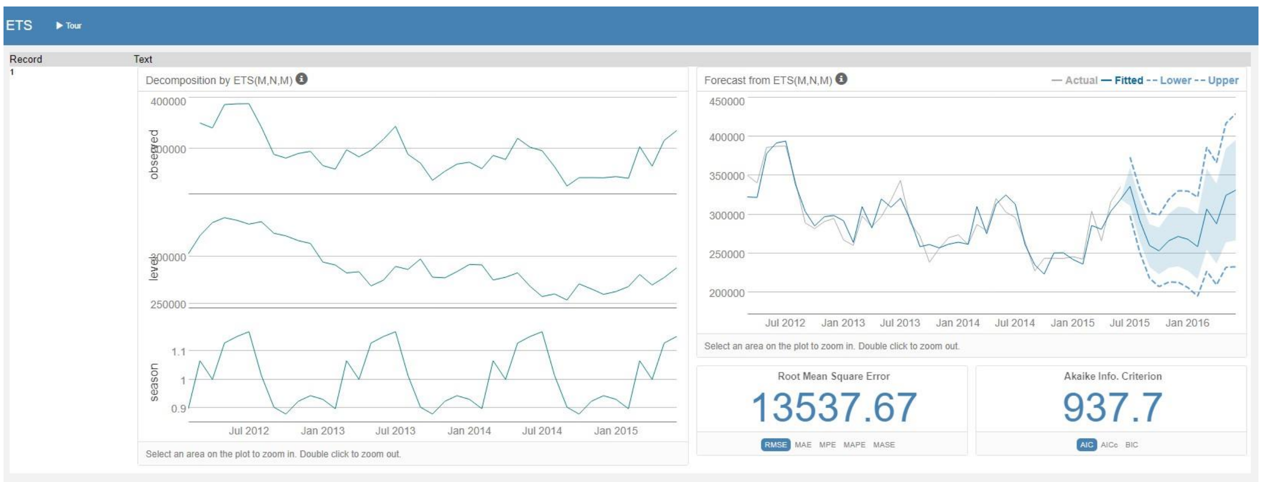

## Cluster 1 ARIMA


## Cluster 2 ETS

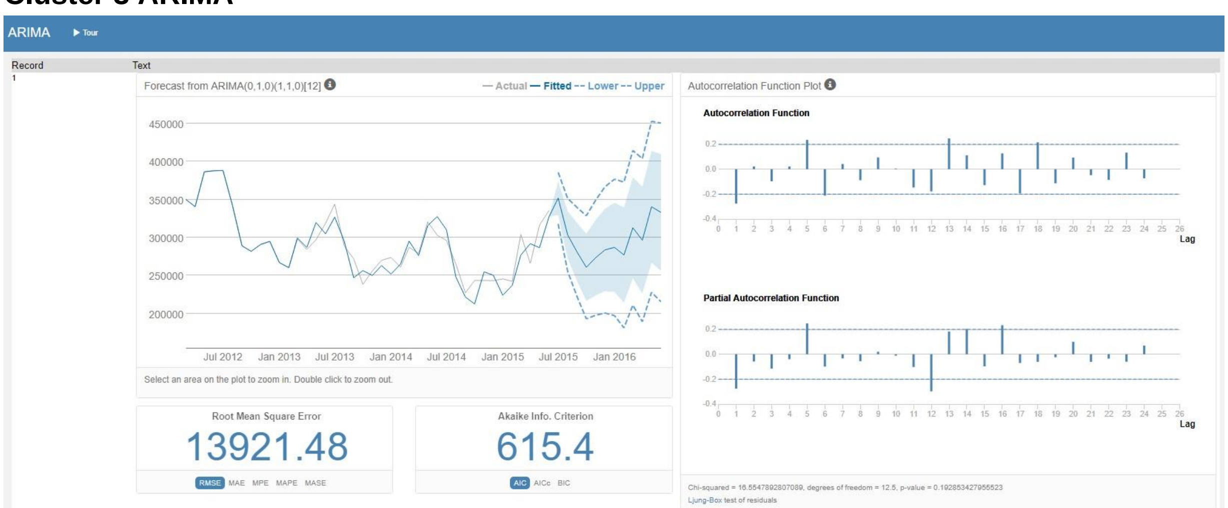
## Cluster 2 ARIMA




## Cluster 3 ETS




## Cluster 3 ARIMA



From the model outcomes above, the ETS model used for Clusters 1, 2 and 3 is ETS(M,N,M). Also, the ARIMA model used for Cluster 1 is ARIMA(0,1,1)(1,1,0)[12] while that used for Clusters 2 and 3 is ARIMA(0,1,0)(1,1,0)[12].

- I further compared the results for each pair of models against their respective hold out samples to ascertain the best forecast model to use on the individual clusters. Provided below are the results from the model comparison for each cluster;

**Cluster 1**

Report

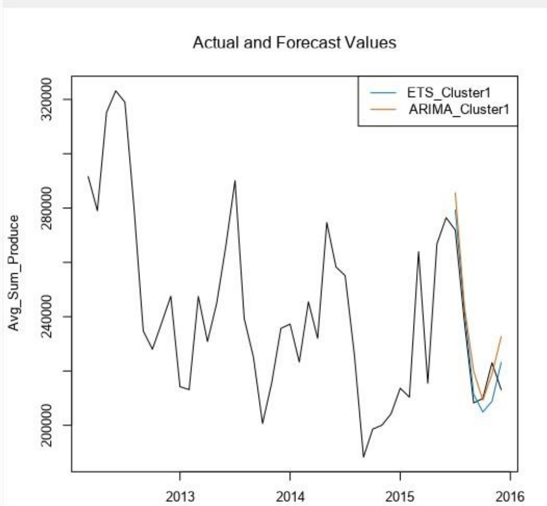
**Comparison of Time Series Models**

Actual and Forecast Values:

| Actual | ETS_Cluster1 | ARIMA_Cluster1 |
|---|---|---|
| 271898.4004 | 279291.66428 | 285596.01211 |
| 237838.3324 | 242091.31355 | 243480.10628 |
| 208225.7536 | 211400.94675 | 219907.46401 |
| 209886.6188 | 204880.21959 | 209258.81682 |
| 223050.8948 | 208847.36008 | 218884.14069 |
| 213114.9824 | 223096.06049 | 232596.29491 |

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS_Cluster1 | -932.0971 | 8258.898 | 7335.408 | -0.3271 | 3.2448 | 0.3522 |
| ARIMA_Cluster1 | -7617.9754 | 11204.122 | 9216.161 | -3.3323 | 4.0547 | 0.4426 |




From the Cluster 1 model comparison report, the Accuracy measures as well as the Actual and Forecasted values suggests that the ETS model performs better than the ARIMA model. More of the ETS forecasted values are closer to the actual values and the accuracy measures also looks better than that of the ARIMA model.

**Cluster 2**

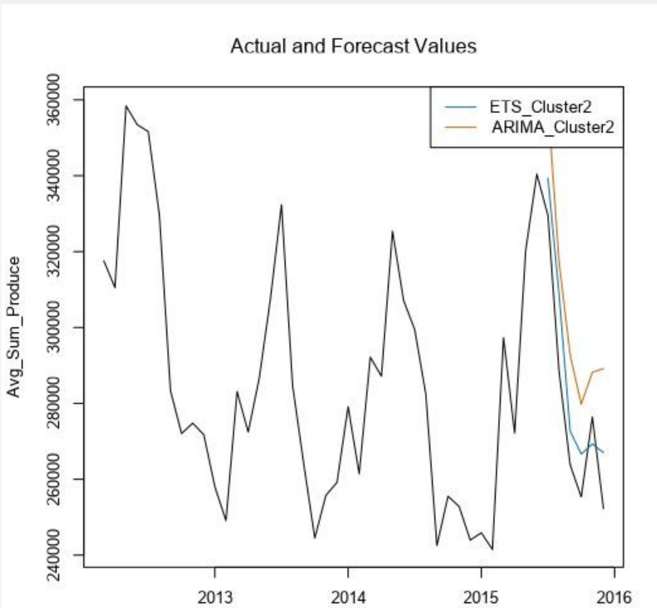
**Comparison of Time Series Models**

Actual and Forecast Values:

| Actual | ETS_Cluster2 | ARIMA_Cluster2 |
|---|---|---|
| 329532.84 | 339311.08596 | 358715.04895 |
| 288438.062857 | 308870.08425 | 317273.59819 |
| 263815.184857 | 272707.54569 | 292769.59114 |
| 255322.114286 | 266633.13046 | 279746.05225 |
| 276422.870857 | 269300.83168 | 288197.69145 |
| 252259.076571 | 267015.52284 | 289137.30542 |

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS_Cluster2 | -9674.675 | 12835.96 | 12048.69 | -3.5208 | 4.3797 | 0.6253 |
| ARIMA_Cluster2 | -26674.856 | 27738.74 | 26674.86 | -9.7121 | 9.7121 | 1.3843 |




From the Cluster 2 model comparison report, the Accuracy measures as well as the Actual and Forecasted values suggests that the ETS model performs better than the ARIMA model. All of the ETS forecasted values are closer to the actual values and the accuracy measures also looks better than that of the ARIMA model.

**Cluster 3**

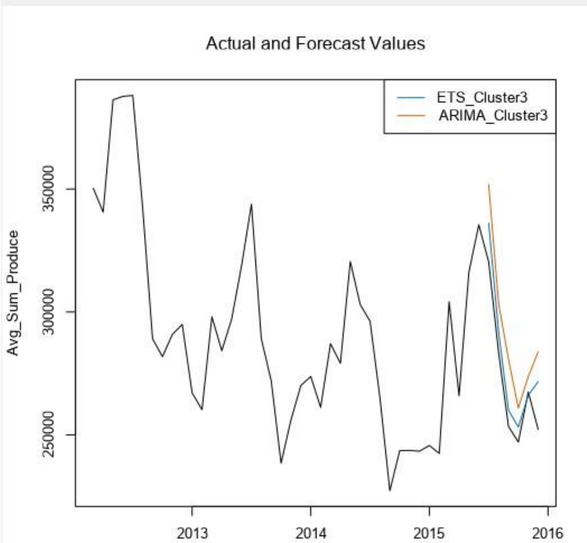
**Comparison of Time Series Models**

Actual and Forecast Values:

| Actual | ETS_Cluster3 | ARIMA_Cluster3 |
|---|---|---|
| 320294.7096 | 336067.37118 | 351738.42781 |
| 283573.4436 | 291977.43192 | 303468.10067 |
| 253409.6248 | 260078.9091 | 280744.48293 |
| 247061.6444 | 253132.04902 | 260888.75028 |
| 267433.3584 | 266062.5522 | 273687.57195 |
| 252238.2824 | 271662.57877 | 283746.12312 |

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS_Cluster3 | -9161.638 | 11401.14 | 9618.574 | -3.3608 | 3.5317 | 0.4488 |
| ARIMA_Cluster3 | -21710.399 | 23645.81 | 21710.399 | -8.0077 | 8.0077 | 1.013 |




From the Cluster 3 model comparison report, the Accuracy measures as well as the Actual and Forecasted values suggests that the ETS model performs better than the ARIMA model. All of the ETS forecasted values are closer to the actual values and the accuracy measures also looks better than that of the ARIMA model.

- Based on the above analyses, I chose to use the ETS(M,N,M) to forecast the produce sales for both the existing and new stores for the year of 2016.

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.

- Provided below is a table showing the forecasted store sum produce for both the existing stores and new stores;

| Month | Forecasted_Existing | Forecasted_New |
|---|---|---|
| 1 | 21829060.03 | 2563357.91 |
| 2 | 21146329.63 | 2483924.73 |
| 3 | 23735686.94 | 2910944.14 |
| 4 | 22409515.28 | 2764881.87 |
| 5 | 25621828.73 | 3141305.87 |
| 6 | 26307858.04 | 3195054.20 |
| 7 | 26705092.56 | 3212390.95 |
| 8 | 23440761.33 | 2852385.77 |
| 9 | 20640047.32 | 2521697.19 |
| 10 | 20086270.46 | 2466750.89 |
| 11 | 20858119.96 | 2557744.59 |
| 12 | 21255190.24 | 2530510.81 |

- Also provided below is a tableau visualization showing the historical store sum produce as well as the forecasted store sum produce for the existing stores and new stores;