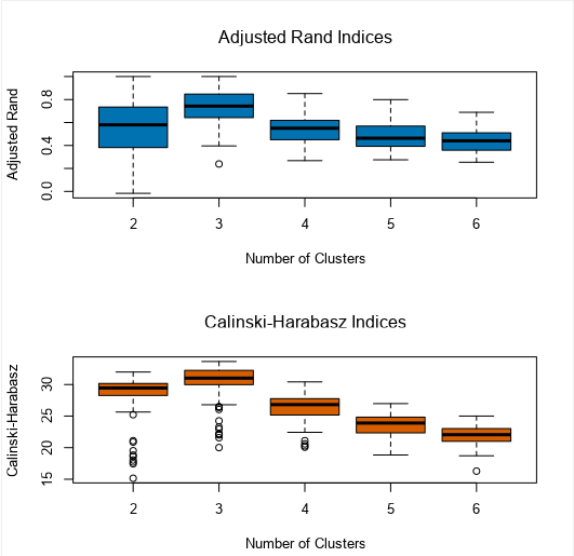
# Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

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

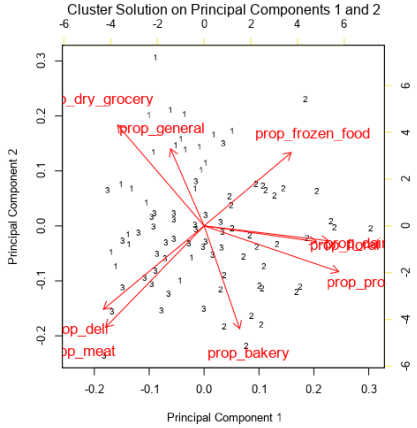



*Based on the adjusted rand indices and Calinski-Harabasz indices, optimal of 3 cluster give the highest median than 2, 4, 5, 6 clusters.*

2. How many stores fall into each store format?
*Cluster 1: 23 stores. Cluster 2: 29 stores. Cluster 3: 33 stores*

| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 23 | 2.320539 | 3.55145 | 1.874243 |
| 2 | 29 | 2.540086 | 4.475132 | 2.118708 |
| 3 | 33 | 2.115045 | 4.9262 | 1.702843 |

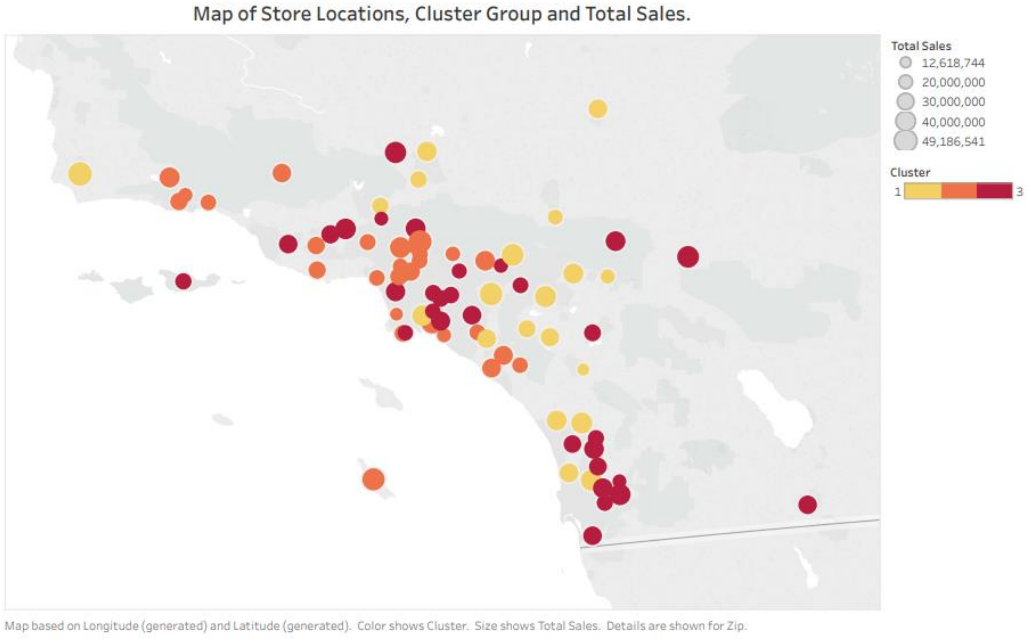
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?




*Based on principal component analysis in 2D plane for three different clusters, cluster 1 is more located with high sales proportion of dry_grocery and general_merchandise. Cluster 3 is more located with high sales proportional of deli and meat. Cluster 2 has the*

*remaining other categories.*

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.




## 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.)
   *I will use three different classification models, forest model, deicison tree and boosted tree method to compare the performance of three models.*

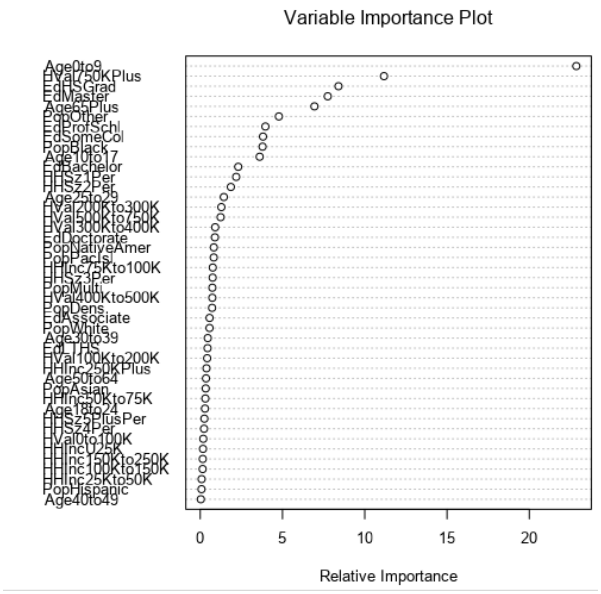
**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| Decision_Tree_7 | 0.7059 | 0.7685 | 0.7500 | 1.0000 | 0.5556 |
| forest_model | 0.8235 | 0.8426 | 0.7500 | 1.0000 | 0.7778 |
| boosted_model | 0.8235 | 0.8889 | 1.0000 | 1.0000 | 0.6667 |

   *As seen in above comparison, boosted model has highest accuracy along with highest F1 score.*

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.
   *Three most important variables in making classification is:* ***Aveoto9, Hval750kplus, EdHSGrad*** *as seen from below classification model feature importance.*

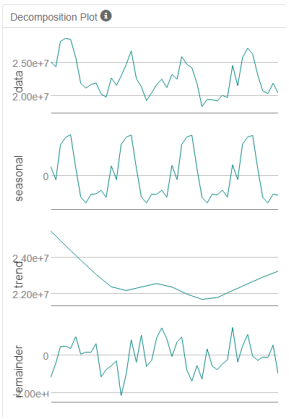



3. What format do each of the 10 new stores fall into? Please fill in the table below.

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



In ETS model, we will use ETS(M, N, M) with no damping model, which is due to (1) error terms are increasing, so use multiplicatively. (2) Trend terms seams no trend (3) season terms slightly increase, so use multiplicatively.

Method:
ETS(M,N,M)

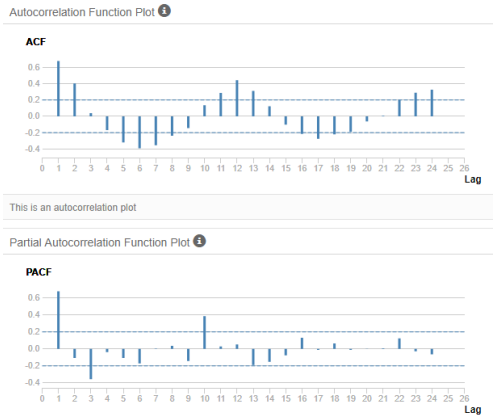
In-sample error measures:

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| 3502.9443415 | 969051.6076376 | 787577.7006835 | -0.1381187 | 3.4677635 | 0.4396486 | 0.0077488 |

Information criteria:

| AIC | AICc | BIC |
|---|---|---|
| 1279.4203 | 1299.4203 | 1304.7535 |

For ARIMA model, when we directly look at timeseries plot of produce sales, we can see that there is peak at 1 and 12, which means that we must use seasonal ARIMA model at 12.




We directly compute the first seasonal difference data and use an ARIMA with seasonal model to fit and validate the model. Software automatically choose ARIMA(1,1,0)(1,1,0)[12] as the best model. And we can also see that after remove seasoning, the ACF and PACF everything is within the range.

**Summary of ARIMA Model arima**

Method: ARIMA(1,1,0)(1,1,0)[12]

Call:
auto.arima(Sum_Produce, d = 1, D = 1, max.p = 1, max.q = 1, max.P = 1, max.Q = 1, ic = "aicc", allowdrift = TRUE)

Coefficients:

| | ar1 | sar1 |
|---|---|---|
| Value | -0.298721 | -0.730228 |
| Std Err | 0.17963 | 0.124086 |

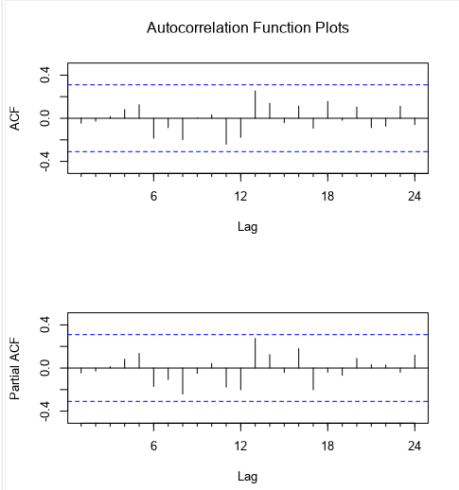
sigma^2 estimated as 1620740758948.9: log likelihood = -421.42529

Information Criteria:

| AIC | AICc | BIC |
|---|---|---|
| 848.8506 | 849.8941 | 852.7381 |

In-sample error measures:

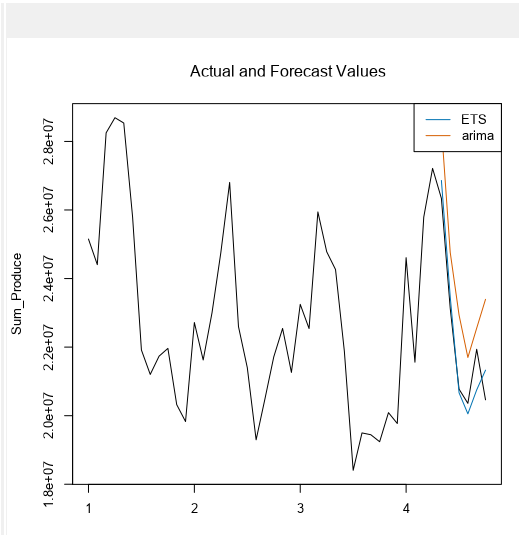
| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| 182768.8209824 | 1006460.617383 | 673969.1592174 | 0.7918625 | 2.9723252 | 0.376229 | -0.0463213 |



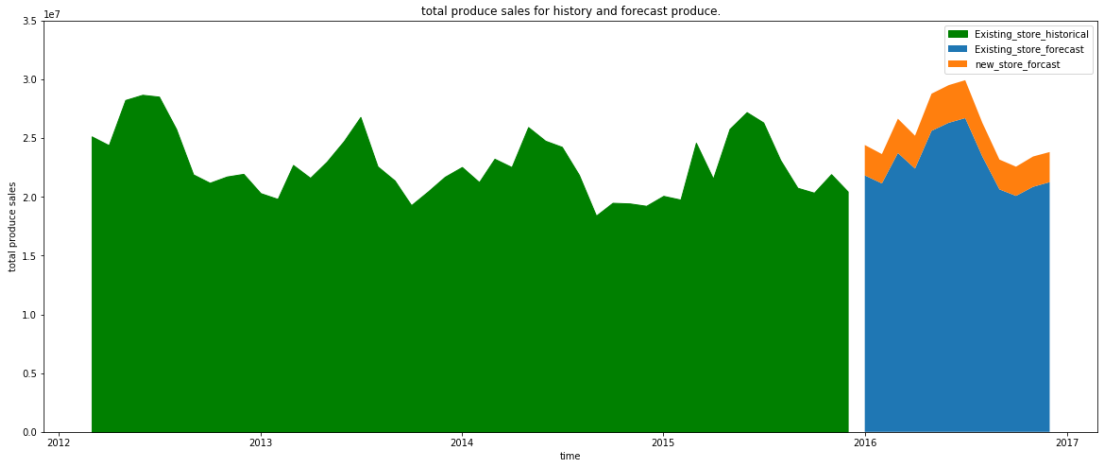

By looking at RMSE, it looks like ETS(M,N,M) give the lowest RMSE, thus I will use ETS model to make prediction.

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |
| arima | -1795372.98 | 1935635.6 | 1795373 | -8.1855 | 8.1855 | 1.0564 |




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.




| Date | Existing stores prediction | New store prediction |
|---|---|---|
| 01/2016 | 24,829,060.03 | 2,588,356.55 |
| 02/2016 | 23,146,329.63 | 2,498,567.17 |
| 03/2016 | 26,735,686.94 | 2,919,067.02 |
| 04/2016 | 26,409,515.28 | 2,797,280.08 |
| 05/2016 | 27,621,828.72 | 3,163,764.85 |
| 06/2016 | 24,307,858.04 | 3,202,813.28 |
| 07/2016 | 20,705,092.55 | 3,228,212.24 |
| 08/2016 | 20,440,761.32 | 2,868,914.81 |
| 09/2016 | 21,640,047.31 | 2,538,372.26 |
| 10/2016 | 20,086,270.46 | 2,485,732.28 |
| 11/2016 | 21,858,119.95 | 2,583,447.59 |
| 12/2016 | 20,255,190.24 | 2,562,181.69 |

Because existing stores have 85 stores, while new store is only 10 stores, thus there is a big difference between forecast total sales per month.

Workflow attachment: