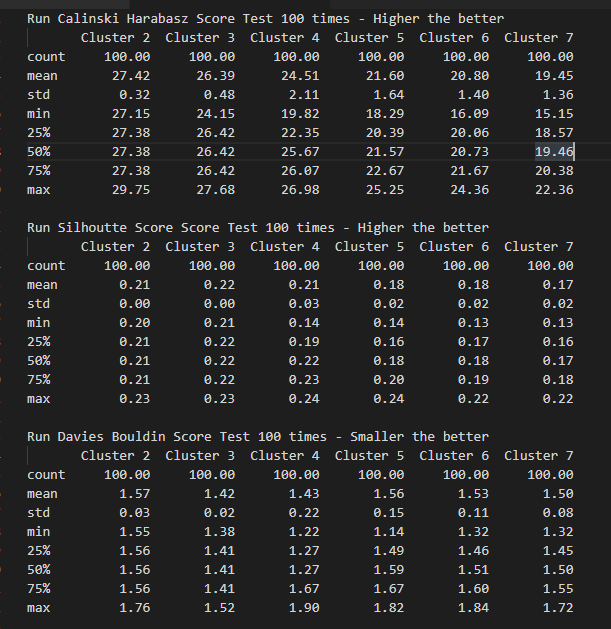
Project: Predictive Analytics Capstone

## 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 three. The the test used to select optimal number of store format.

* Davies Bouldin Score Test show 3 is optimal store format number
* Silhoutte Score Test show 3 is optimal store format number
* Calinski Harabasz Score Test show 2 is optimal store format number.
* Overall I chose 3 as the optimal number of store formats

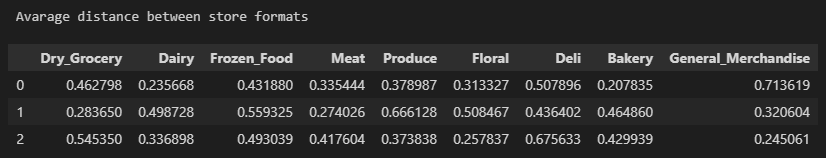


1. How many stores fall into each store format?

* Store Format 0 has 22 stores
* Store Format 1 has 32 stores
* Store Format 2 has 31 stores

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

* Store format 1 has small average distance compared to other store formats

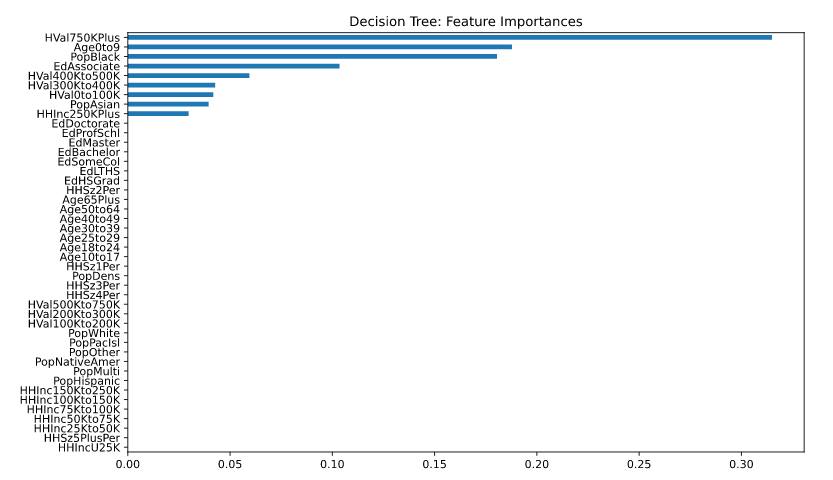


1. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

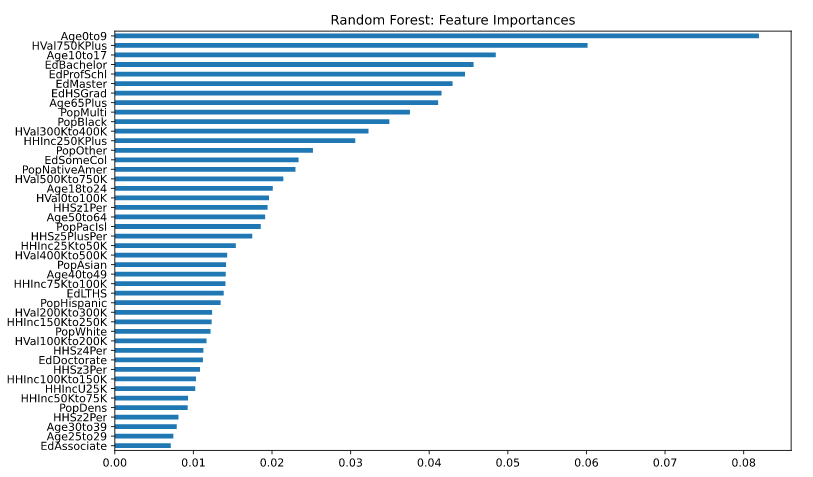
## Task 2: Formats for New Stores

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.)

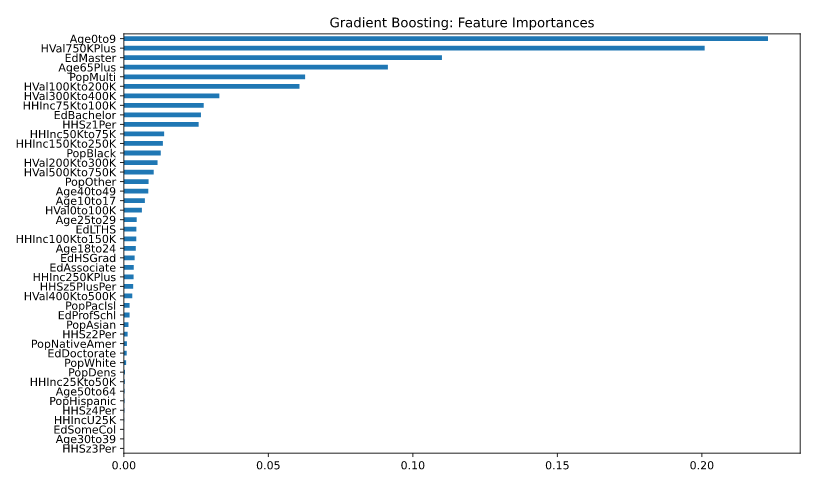
* I tested Decision Tree, Forest and Boosted models to find best model to segment stores
* Decision Tree Model top three important variables are HVal750KPlus, Age0to9, and PopBlack.



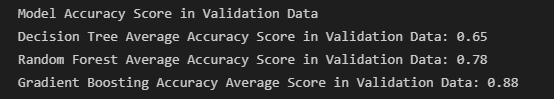
* Forest Model top three important variables are Age0to9, HVal750KPlus, and Age10to17.



* Boosted Model top three important variables are Age0to9, HVal750KPlus, and EdMaster.



* I use average accuracy score on validation data to choose the best model.



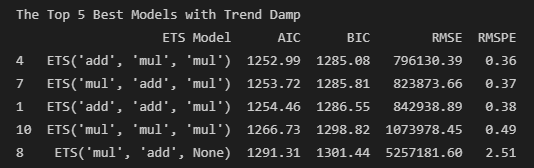
* I will use booted model due to higher score of 88% accuracy in validation data.

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

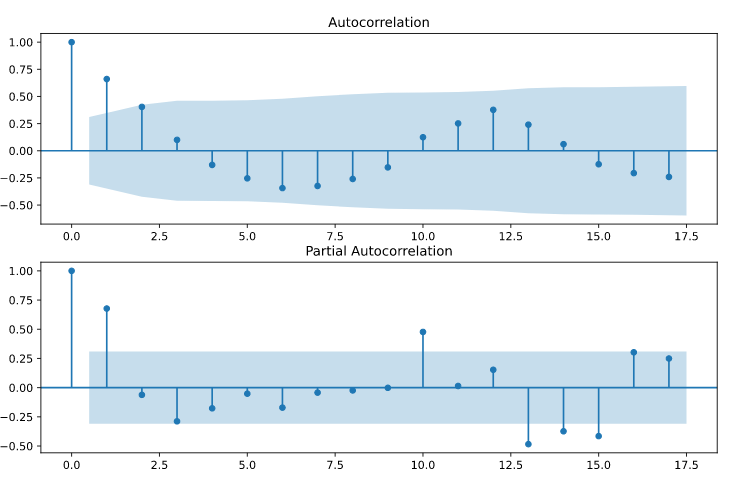


## Task 3: Predicting Produce Sales

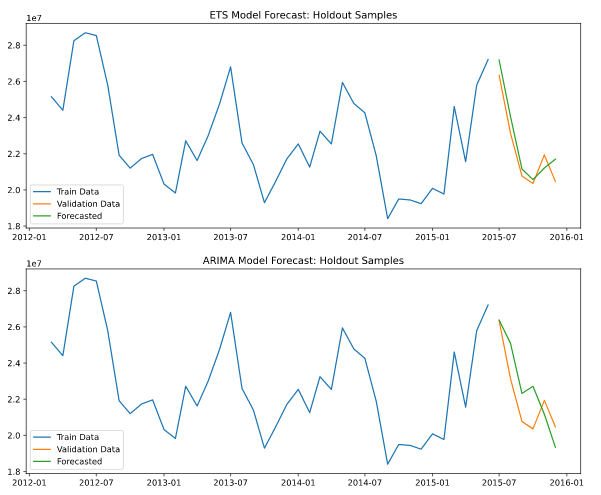
* + - 1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?
* I use ETS(A,M,M) due to lower error AIC and BIC



* I use ARIMA(11,0,0) as acf and pacf plot suggested



* Holdout test error (RMSE) are 796,130 for ETS and 1,506,450 for ARIMA
* I will chose ETS Model to forecast 2016 sales due to lower error in holdout sample
* Forecast plot



* + - 1. 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.