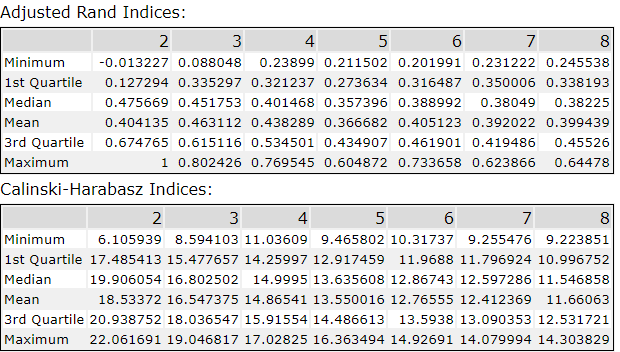
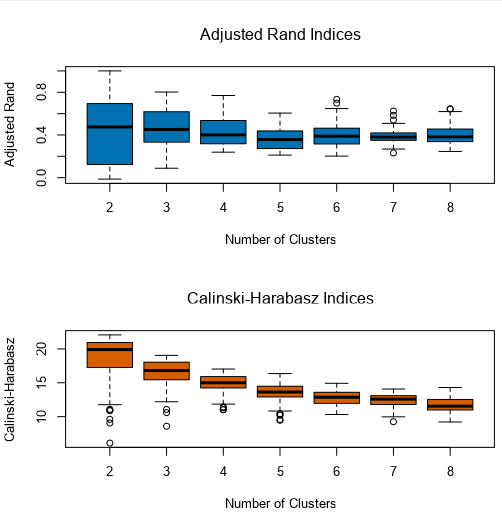
Project: Predictive Analytics Capstone

## Task 1: Determine Store Formats for Existing Stores

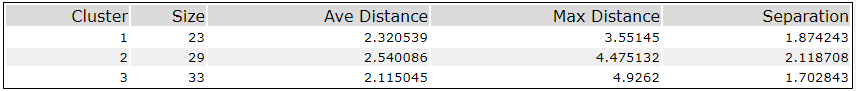
1. What is the optimal number of store formats? How did you arrive at that number?





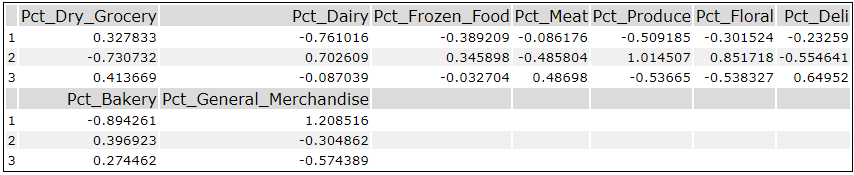
The optimal number of store formats is 3. By running a K-Means clustering model, we are able to see the spread and median of the Adjusted Rand Indices and the Calinski-Harabasz Indicies. With this information, we are able to see that having 2 or 3 clusters give us the highest median, however we chose to use 3 clusters because the variance is more compact.

1. How many stores fall into each store format?



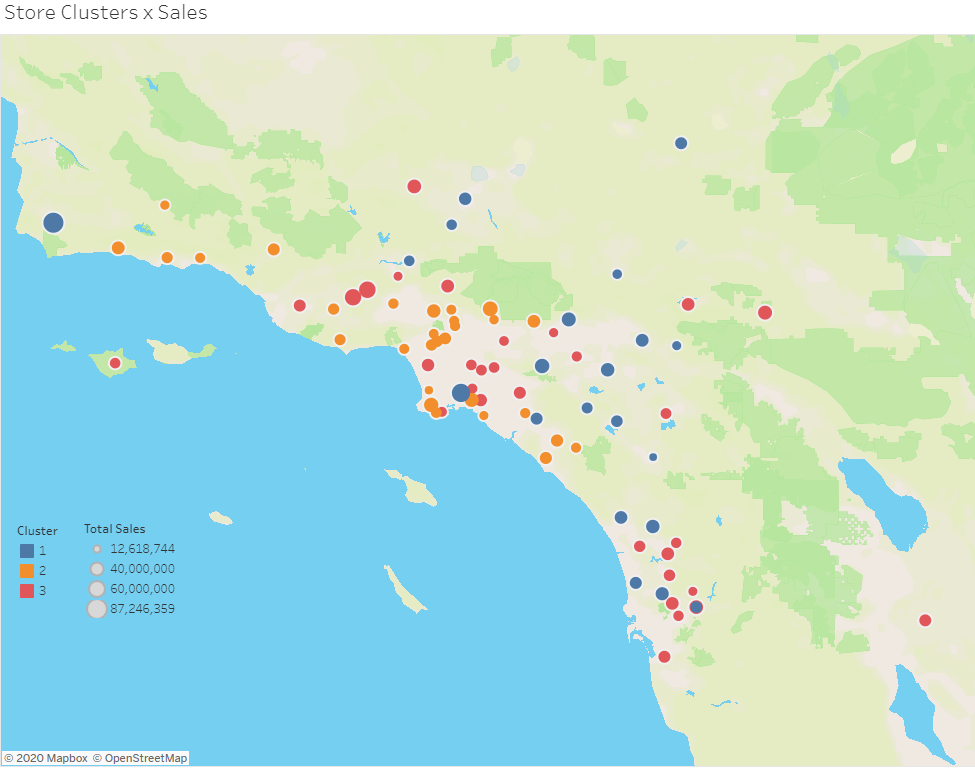
In cluster 1, there are 23 stores, cluster 2 has 29, and cluster 3 has 33.

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



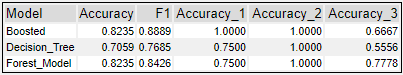
Cluster 1 seems to have a higher amount of General Merchandise sales, Cluster 2 has a higher amount of Dairy, Frozen Food, Produce, and Floral sales, and Cluster 3 has more Meat and Deli sales.

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.



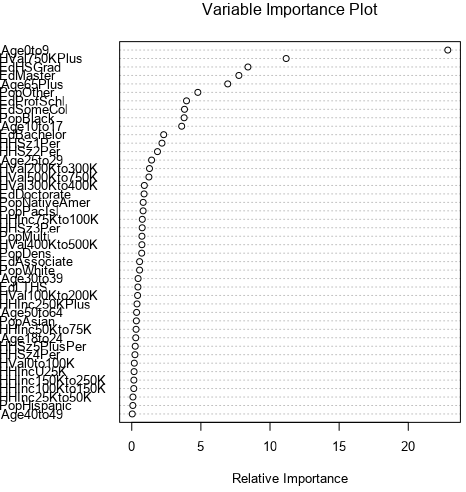
## Task 2: Formats for New Stores

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.)



After comparing the three models, the model that was best for predicting the new store formats was the Boosted Model. The Boosted Model had the highest accuracy and the highest F1 score.

1. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.



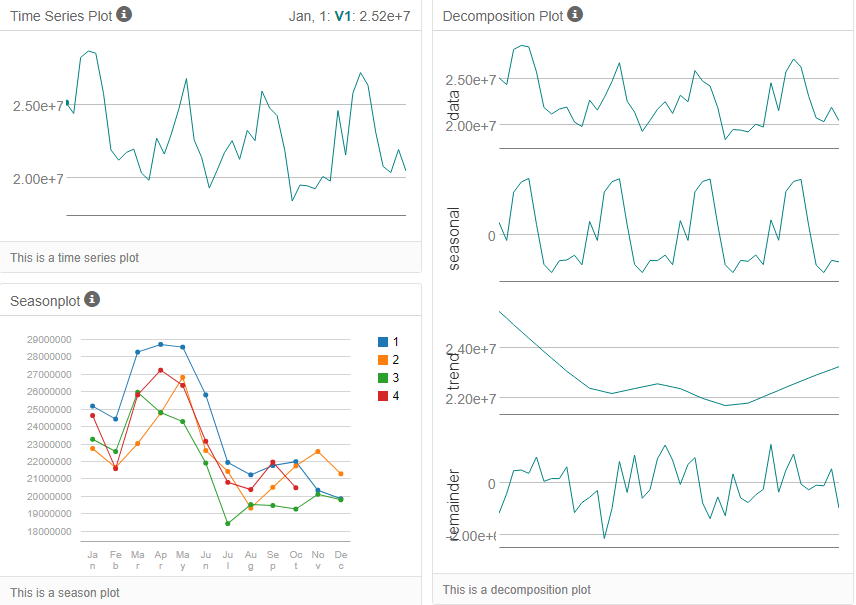
The most important variables are Age0to9 HVal750KPlus and EdHSGrad

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 | 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?



The model we are using for this forecast is ETS (M,N,M). Looking at the decomposition plot, we can see that seasonality and error are both multiplicative, and that there is no trend.

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.

|  |  |  |
| --- | --- | --- |
| **Month** | **New Stores** | **Existing Stores** |
| Jan-16 | 2,588,357 | 21,829,060 |
| Feb-16 | 2,498,567 | 21,146,330 |
| Mar-16 | 2,919,067 | 23,735,687 |
| Apr-16 | 2,797,280 | 22,409,515 |
| May-16 | 3,163,765 | 25,621,829 |
| Jun-16 | 3,202,813 | 26,307,858 |
| Jul-16 | 3,228,212 | 26,705,093 |
| Aug-16 | 2,868,915 | 23,440,761 |
| Sept-16 | 2,538,372 | 20,640,047 |
| Oct-16 | 2,485,732 | 20,086,270 |
| Nov-16 | 2,583,448 | 20,858,120 |
| Dec-16 | 2,562,182 | 21,255,190 |

