Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms "formats" and "segments" will be used interchangeably throughout this project.

You've been asked to:

- Determine the optimal number of store formats based on sales data.
    - Sum sales data by StoreID and Year
    - Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
    - Use only 2015 sales data.
    - Use a K-means clustering model.
- Segment the 85 current stores into the different store formats.
- Use the StoreSalesData.csv and StoreInformation.csv files.

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

Based on the K-means report, Adjusted Rand and Calinski-Harabasz indices below, the optimal number of store formats is **3** when both the indices registered the highest median value.

K-Means Cluster Assessment Report

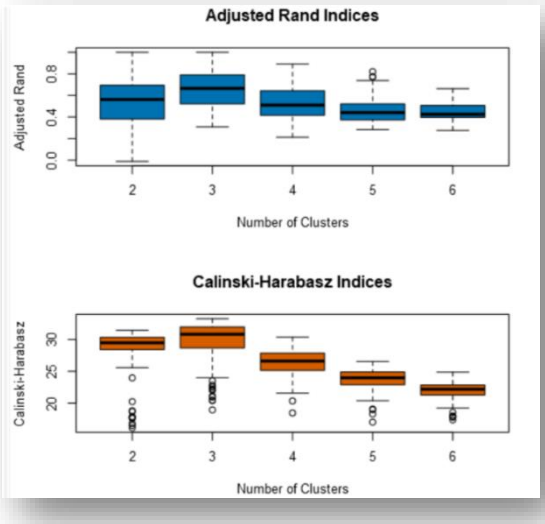
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.01155	0.3083	0.213	0.2837	0.2762
1st Quartile	0.3814	0.5258	0.4169	0.374	0.3965
Median	0.5619	0.6653	0.5107	0.4406	0.4256
Mean	0.5084	0.6594	0.5471	0.4704	0.4502
3rd Quartile	0.6942	0.7865	0.6427	0.5199	0.5067
Maximum	1	1	0.8902	0.8207	0.6626

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	16.1	18.94	18.45	17.02	17.37
1st Quartile	28.42	28.68	25.16	22.91	21.28
Median	29.47	30.83	26.61	23.98	22.17
Mean	28.24	29.58	26.34	23.7	21.95
3rd Quartile	30.31	31.97	27.85	24.9	22.84
Maximum	31.44	33.26	30.37	26.53	24.87



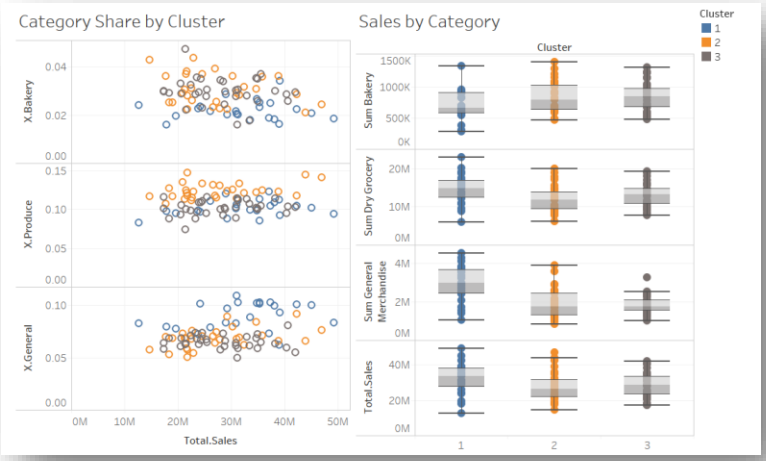
2. How many stores fall into each store format?

   Cluster 1 has 23 stores, cluster 2 has 29 stores while cluster 3 has 33 stores.

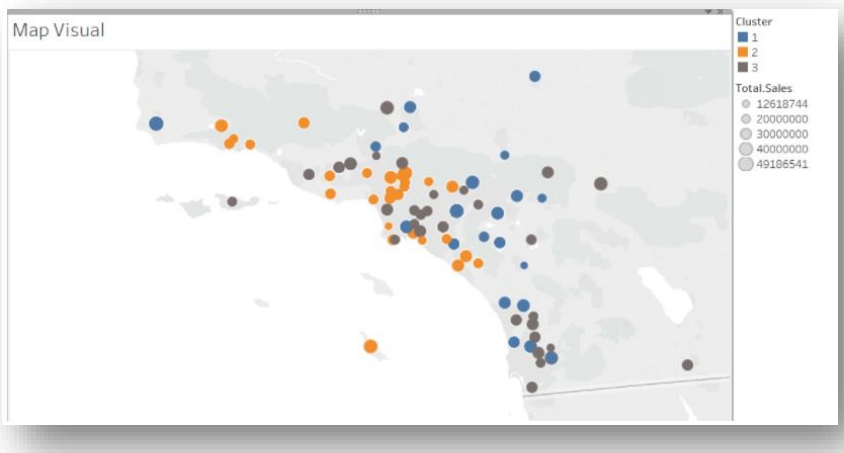
Cluster Information:

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?
   Cluster 1 stores sold more General Merchandise in terms of percentage while Cluster 2 stores sold more Produce.
   Cluster 1 stores have highest medial total sales when compared to the other 2. Its range of total sales and most of other categorical sales are also the largest. Cluster 3 stores are the most similar in terms of sales due to more compact range.



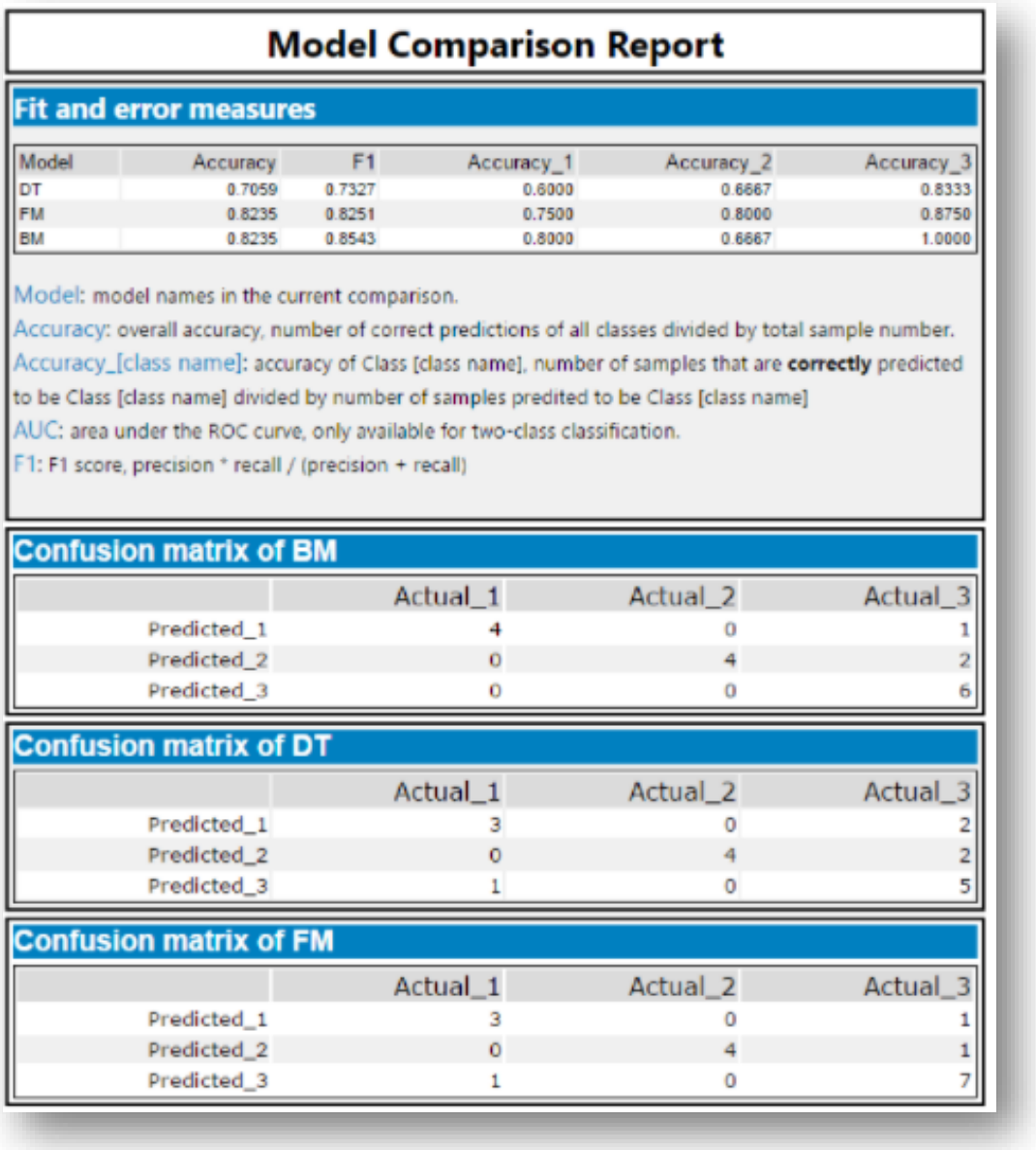
4. Please provide a Tableau visualization that shows the location of the stores, uses color to show cluster, and size to show total sales.



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.)

   The model comparison report below shows comparison matrix of Decision Tree, Forest Model and Boosted Model.
   Boosted Model is chosen despite having same accuracy as Forest Model due to higher F1 value.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7327	0.6000	0.6667	0.8333
FM	0.8235	0.8251	0.7500	0.8000	0.8750
BM	0.8235	0.8543	0.8000	0.6667	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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of DT

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

Confusion matrix of FM

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

StoreNumber	Segment
S0086	1
S0087	2
S0088	3
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?

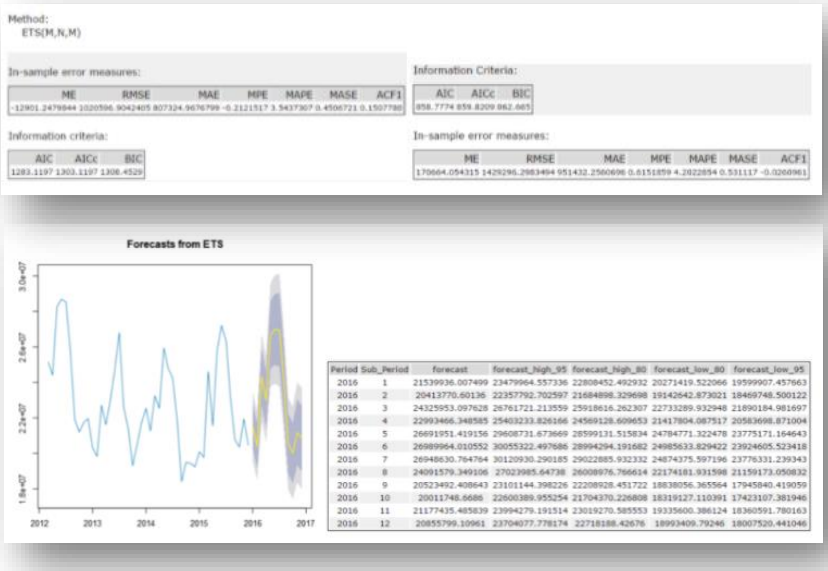
ETS(M,N,M) with no dampening is used for ETS model.

The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and nothing should be applied. Its error is irregular and should be applied multiplicatively.

ARIMA(0,1,2)(0,1,0) is used as seasonal difference and seasonal first difference were performed. There is a lag-2.

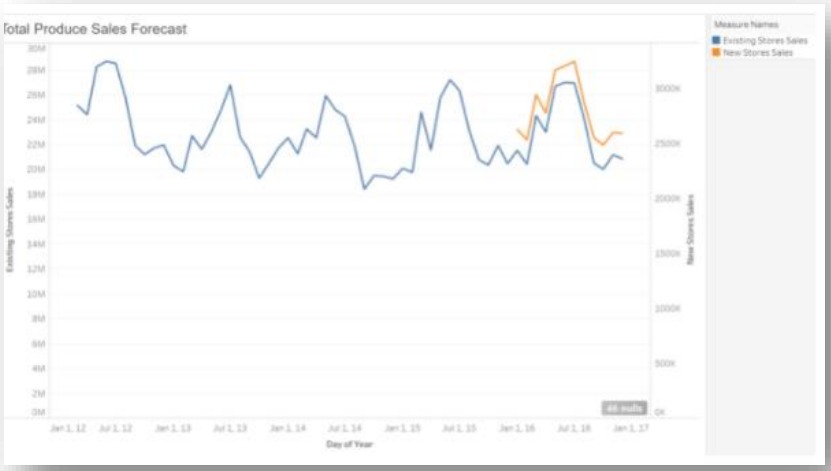
ETS model's accuracy is higher when compared to ARIMA model. A holdout sample of 6 months data is used. Its RMSE of **1,020,597** is lower than ARIMA's **1,429,296** while its MASE is **0.45** compared to ARIMA's **0.53**. ETS also has a higher AIC at **1,283** while ARIMA's AIC is **859**.

The graph and table below shows actual and forecast value with 80% & 95% confidence level interval.



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.

Year	Month	NewStore Sales	ExistingStoreSales
2016	1	2,626,198	21,539,936
2016	2	2,529,186	20,413,771
2016	3	2,940,264	24,325,953
2016	4	2,774,135	22,993,466
2016	5	3,165,320	26,691,951
2016	6	3,203,286	26,989,964
2016	7	3,244,464	26,948,631
2016	8	2,871,488	24,091,579
2016	9	2,552,418	20,523,492
2016	10	2,482,837	20,011,749
2016	11	2,597,780	21,177,435
2016	12	2,591,815	20,855,799



The chart above shows the historical and forecast sales for existing stores and new stores over the period from Mar-12 to Dec-16.

Alteryx flow (Task1,Task2,Task3)