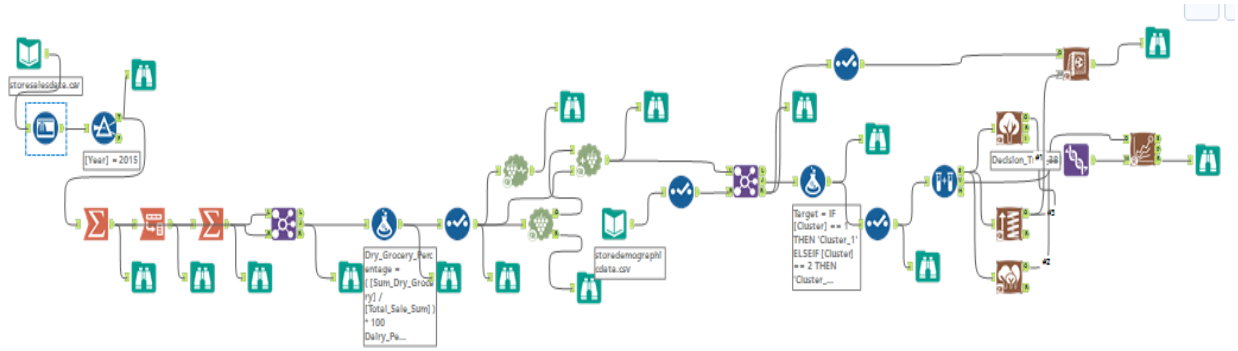
Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Alteryx Workflow



Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?
   For best number of store formats I used k-centroids analysis was done using k-means clustering method for k=3,4,5,6.
   Optimal number of store formats are 3.

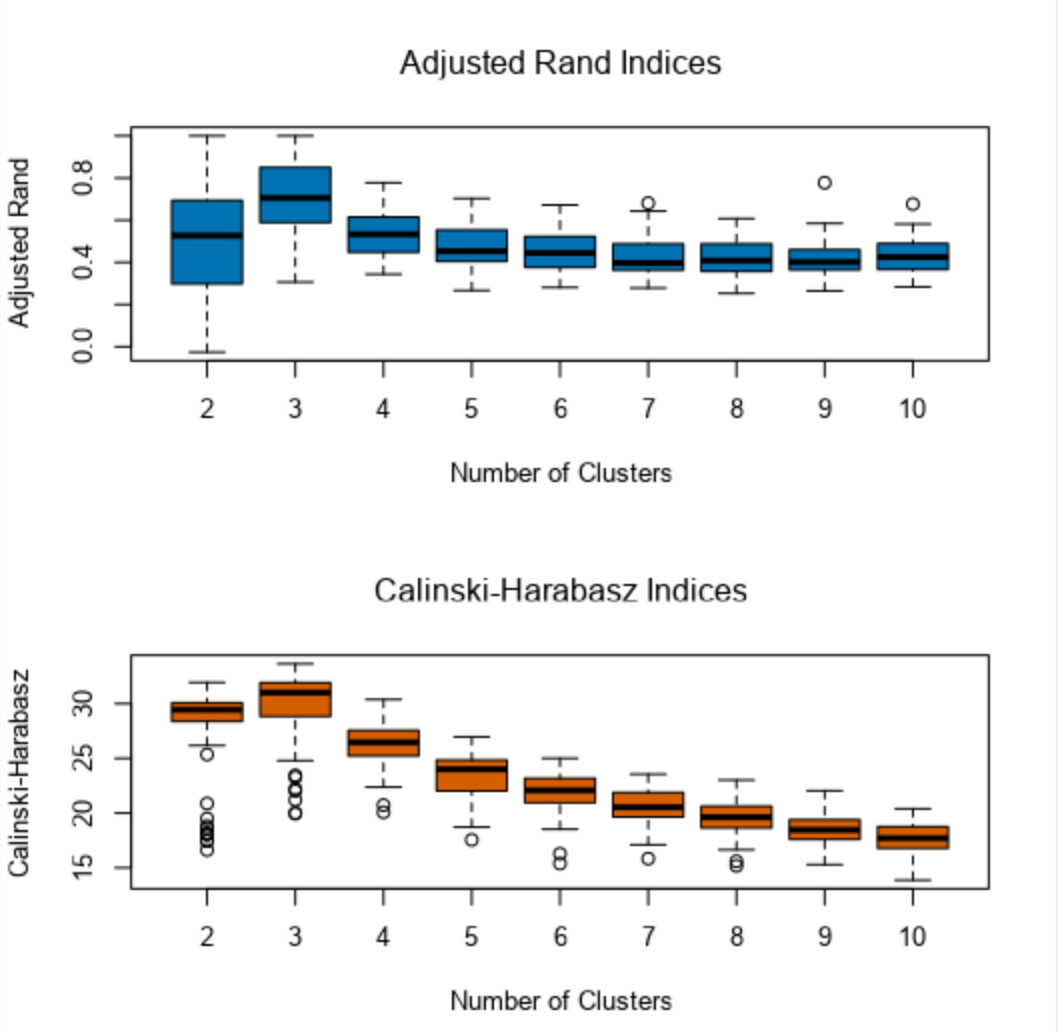
2. How many stores fall into each store format?

   Cluster 1 -> 23 Stores
   Cluster 2 ->29 Stores
   Cluster 3 -> 33 Stores

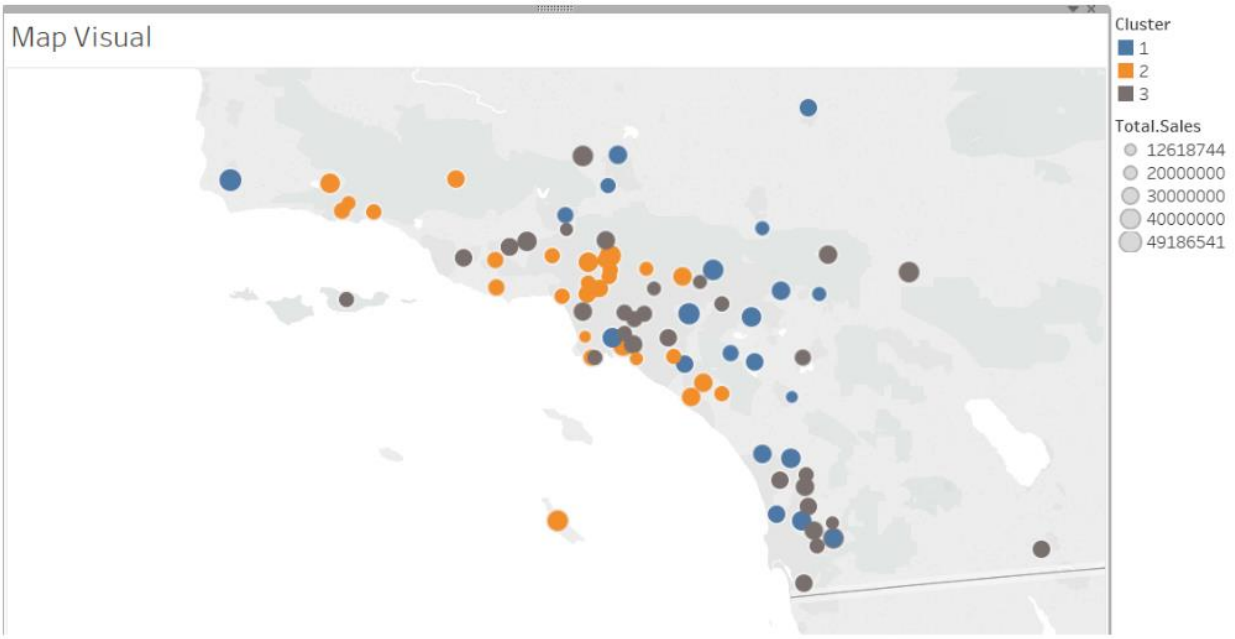
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   Based on the results above, I can see cluster 1 is most positive for percentage of general merchandise sales vs cluster 3 which is the most negative. This would indicate that these two clusters are the most different in terms of the variable for percentage of general merchandise sales.

   AR and CH value for each of the cluster is compared, in which Median of 3 cluster is high. AR value 0.6936 and CH Value 30.7.



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.)

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750

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 Boosted_Model

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest_Model

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

2. Constructed and compared the three models Decision Tree, Boosted Model and Forest

Model and the Model comparison report is shown above. Based on the F1 score, I decided to use Boosted Model.

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

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

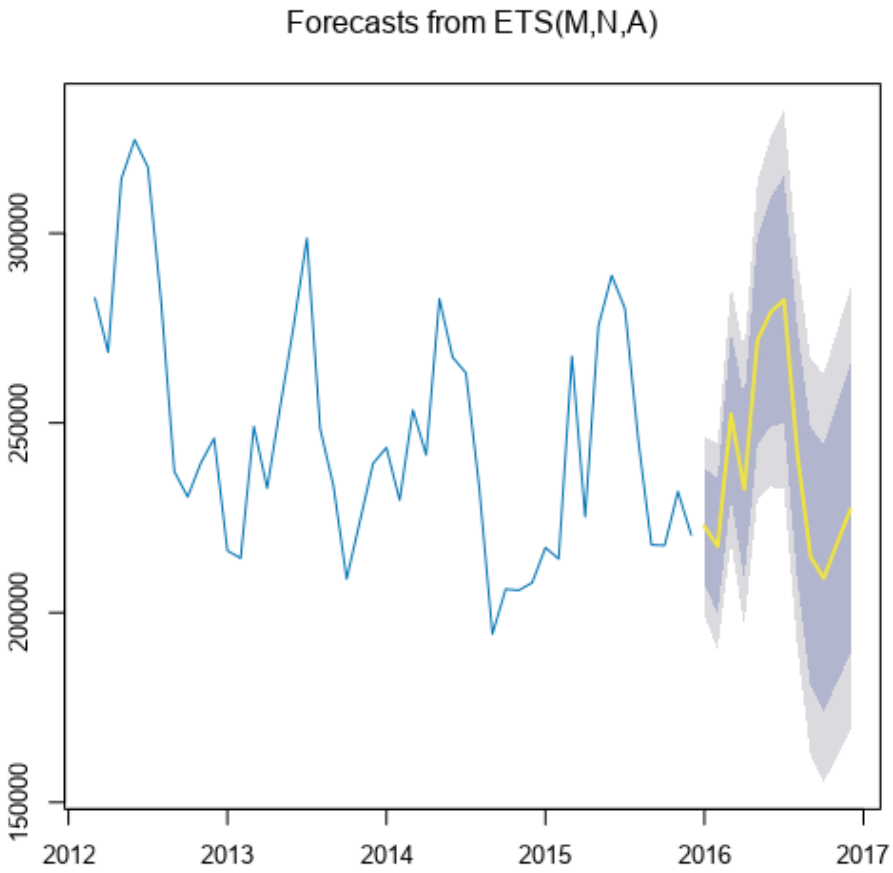
Task 3: Predicting Produce Sales

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?

To forecast sales for existing stores, Decomposition plot with the Produce data was generated with the TS-Plot tool.

The sales fluctuate in similar intervals, thereby indicating the presence of Seasonality. It can also be observed that the Sales is also growing, hence it would be Multiplicative. So, based on the above observations, the ETS Model would be ETS(M,N,M).

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.

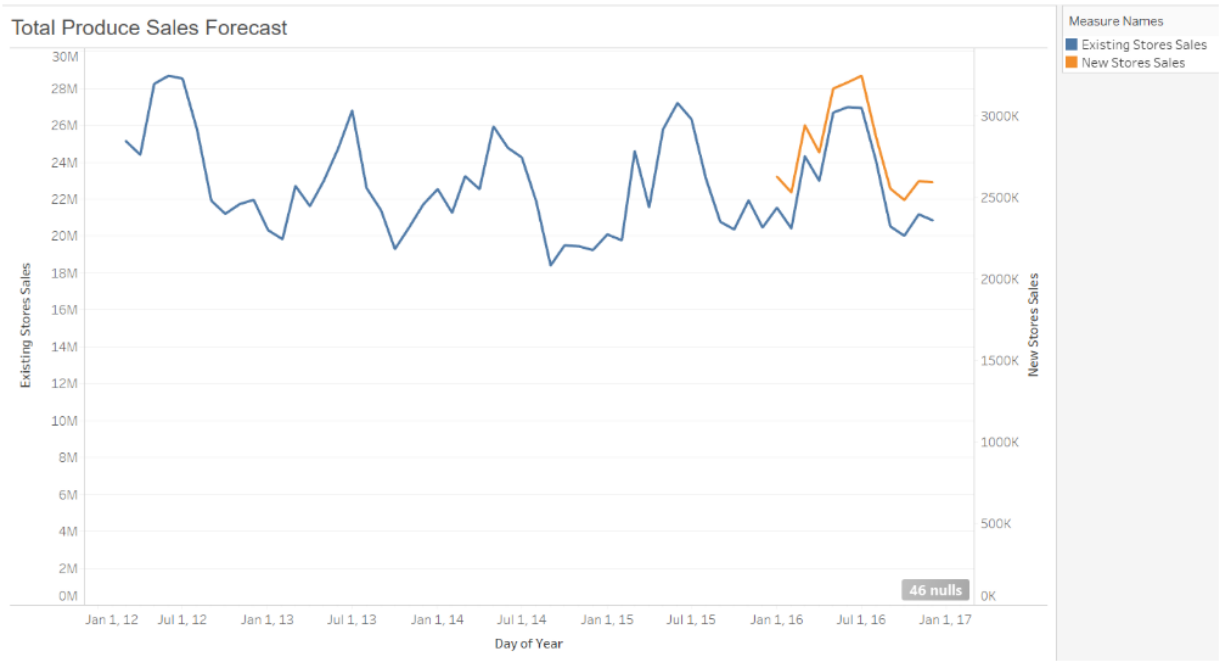


The Forecast Plot shows the historic data in black and the expected value in blue. The orange in the plot shows the 90% confidence interval, and the yellow shows the 95% confidence interval.

Table below shows the forecast sales for existing stores and new stores. New store sales is obtained by using **ETS(M,N,M)** analysis with all the 3 individual cluster to obtain the average sales per store. The average sales value (x3 cluster 1, x6 cluster 2, x1 cluster 3) are added up produce New Store Sales.

Year	Month	New Store Sales	Existing Store Sales
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

Year	Month	New Store Sales	Existing Store Sales
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



This chart show historical and forecast sale for each store and new store.

Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.