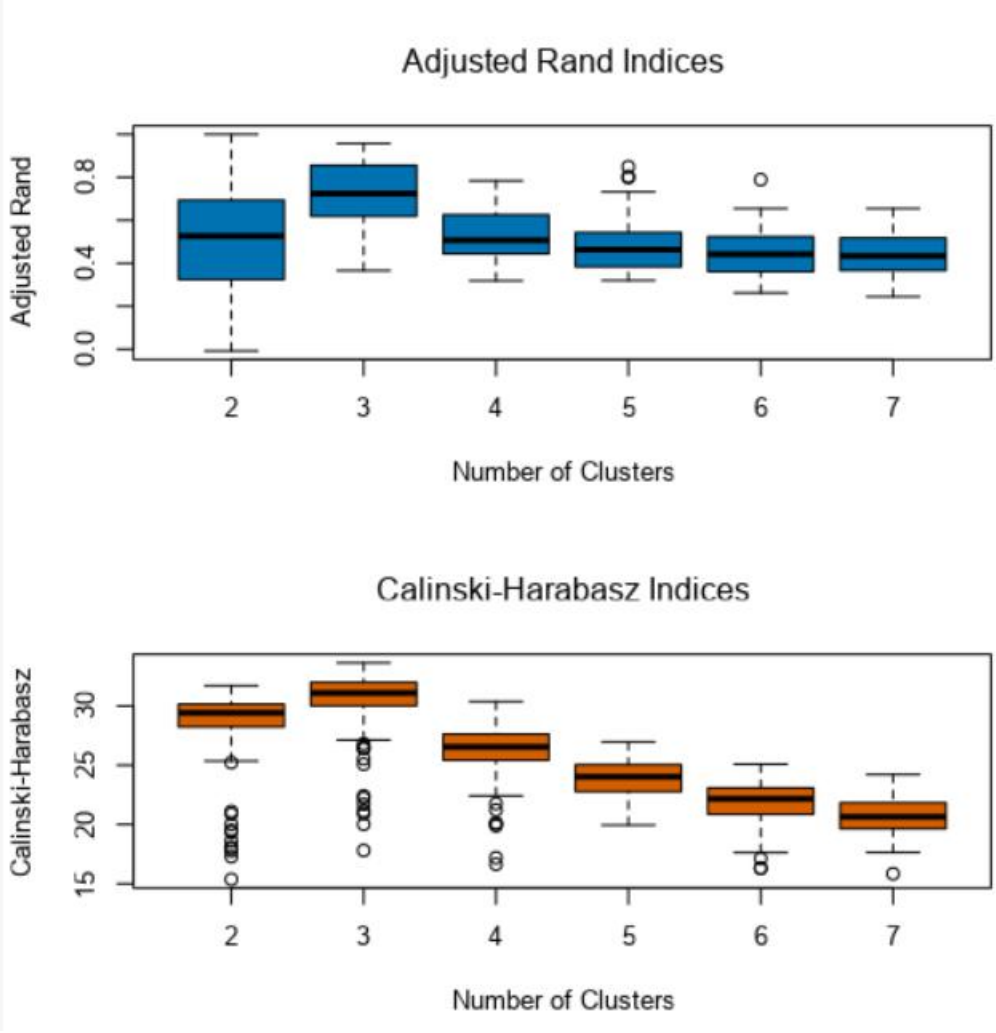
Project: Predictive Analytics Capstone

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 below for K=2-7, the optimal number of store formats is 3. This is because in both indices, 3 has the highest median and a relatively tight spread.



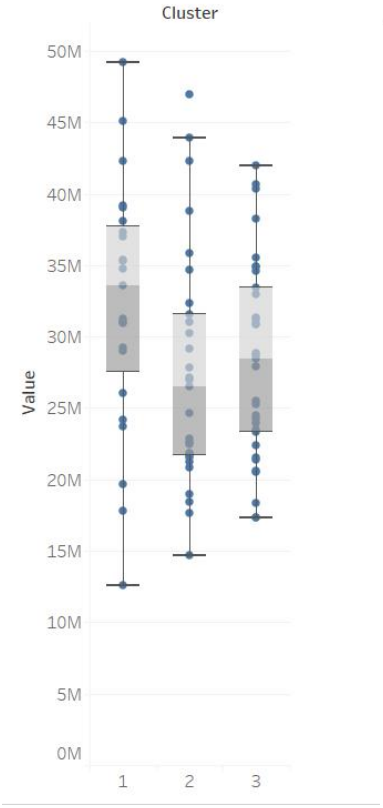
2. How many stores fall into each store format?

Record	Cluster	Count
1	1	23
2	2	29
3	3	33

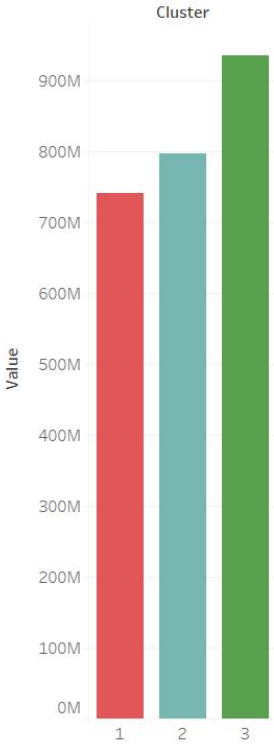
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Cluster 1 is more variable with higher median sales on average, while cluster 3 is less variable with higher total sales due to a larger proportion. Cluster 2 has the lowest median sales.

box and whisker

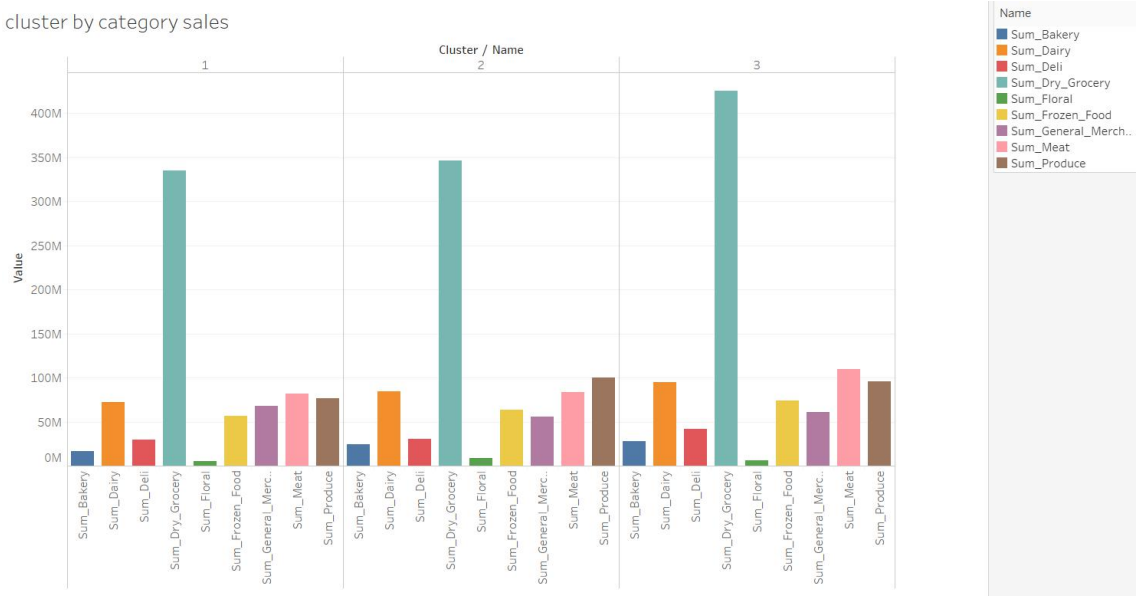


cluster by sales

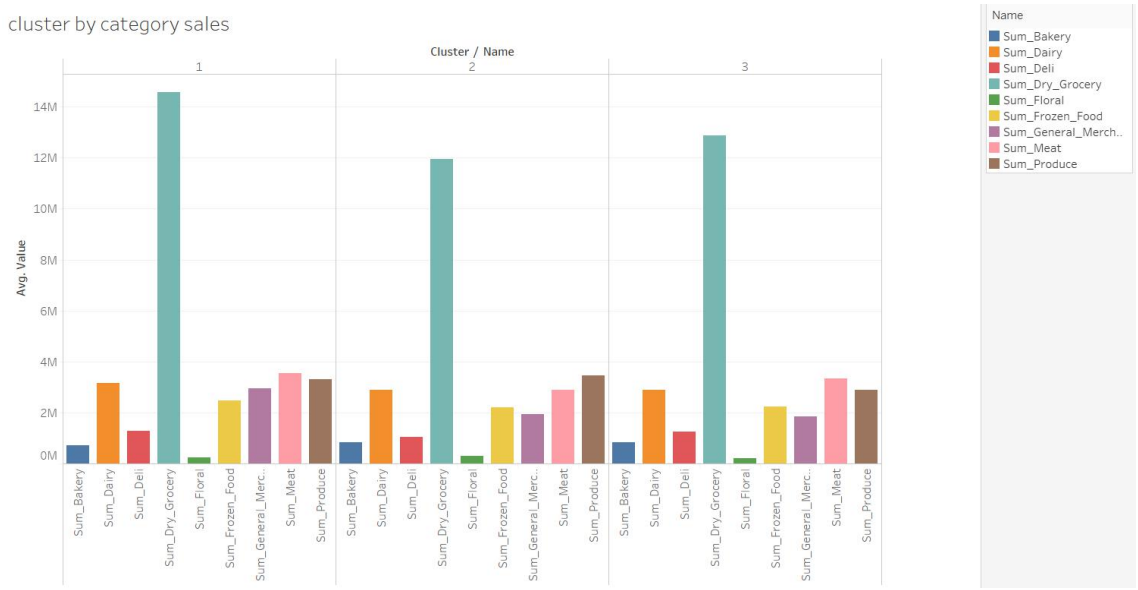


Dry Grocery sales in total are much higher in cluster 3 and Produce sales in total are higher in cluster 2.

cluster by category sales



When looking at average sales by cluster, cluster 1 has the highest for most of the grocery items.



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.

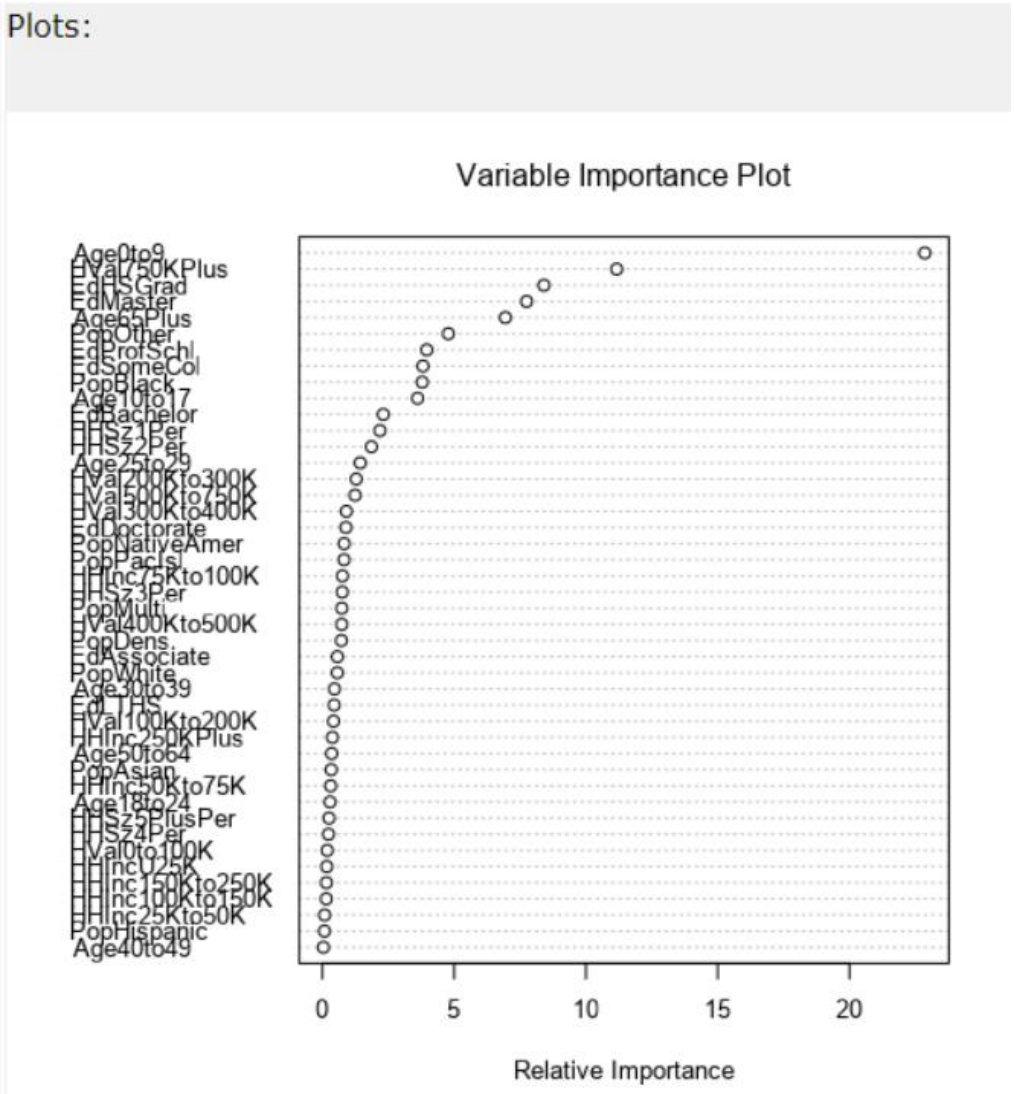
Task 2: Formats for New Stores

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 compared the performance of a Decision Tree, Decision Forest and Boosted model with a 20% holdout sample. Based on the results of model comparison, I used the boosted model to predict the format of each new store as it had the best error measures when measured on the holdout sample.

Based on the feature importance plot from my boosted model, the three most important features are Age0to9, HVal750KPlus and EdHSGrad.

Plots:



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DecisionTree	0.7647	0.8056	0.7500	1.0000	0.6667
Forest	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted	0.8235	0.8889	1.0000	1.0000	0.6667

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted

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

Confusion matrix of DecisionTree

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	1
Predicted_3	1	0	6

Confusion matrix of Forest

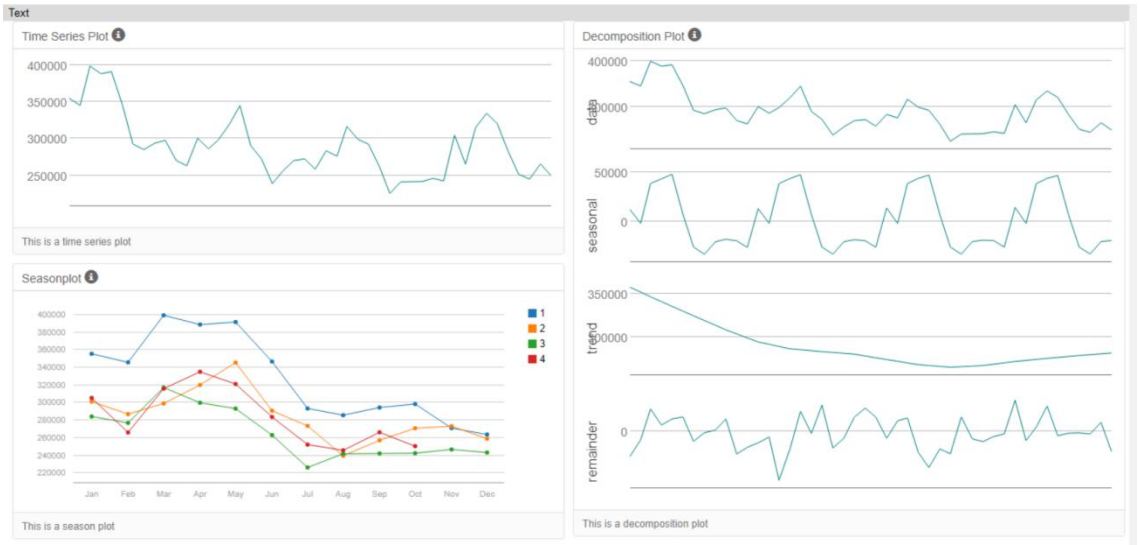
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

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

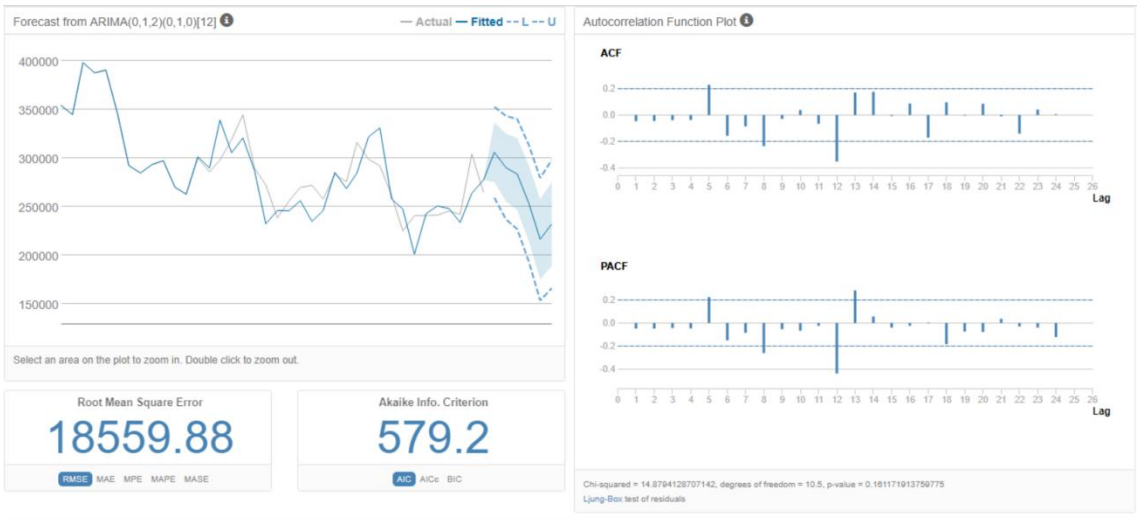
Task 3: Predicting Produce Sales

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?



I have used the TS Plot above for just Cluster 1 in my decisions below. I have also used a holdout sample of 6 months of data to test the models against.

ARIMA: A seasonal difference and seasonal first difference were performed to make the data static. There is a lg 2 present too so an ARIMA(0,1,2)(0,1,0)[6]



ETS: After an initial fall the seasonality shows an increasing trend, the trend is unclear and the error is also irregular. Therefore an ETS(M, N, M) model will be used.

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-1182.5271886	13896.4447271	11592.9229025	-0.5578903	3.992998	0.3608805	-0.0203925

Information criteria:

AIC	AICc	BIC
891.87	913.6881	916.4337

Information Criteria:

AIC	AICc	BIC
579.1989	580.3417	582.8555

In-sample error measures:

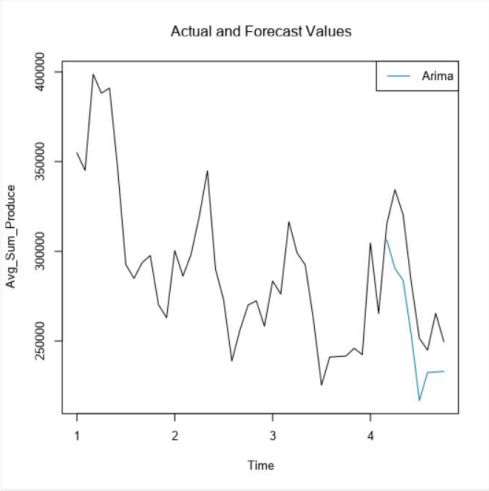
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3071.4096869	18559.8835676	11886.5029218	1.0560319	4.2744103	0.3700194	-0.0473466

Based on the comparison of the models above, the ETS model outperforms the ARIMA model. The ETS model has a lower RMSE (13896 vs 18560), a lower MASE (0.36 vs 0.37) and a higher AIC (892 vs 579). This suggests I should use the ETS(M,N,M) model for forecasting. However, first I will check both models' performance against the 6 month holdout sample.

On the holdout sample, the ETS model also outperformed the ARIMA model. RMSE is lower (14687 vs 29361) and MASE is much lower (0.5792 vs 1.2681). This validates my choice that using he ETS(M, N, M) model will be better for my forecast.

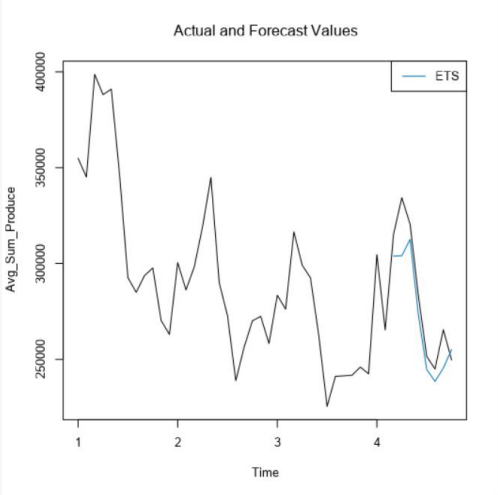
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
Arima	26855.85	29361.19	26855.85	9.4343	9.4343	1.2681



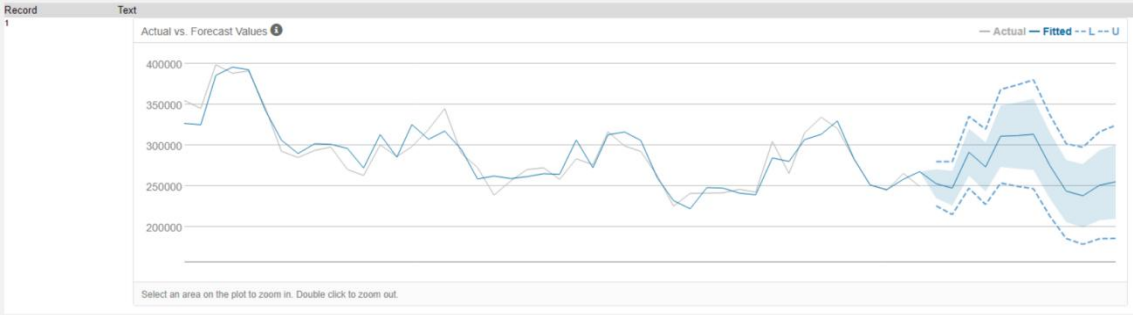
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	10907.47	14686.09	12266.82	3.6654	4.2098	0.5792

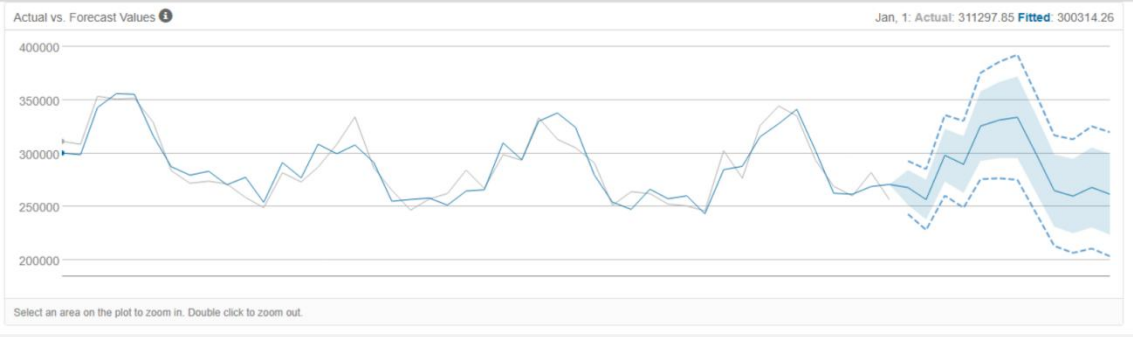


2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

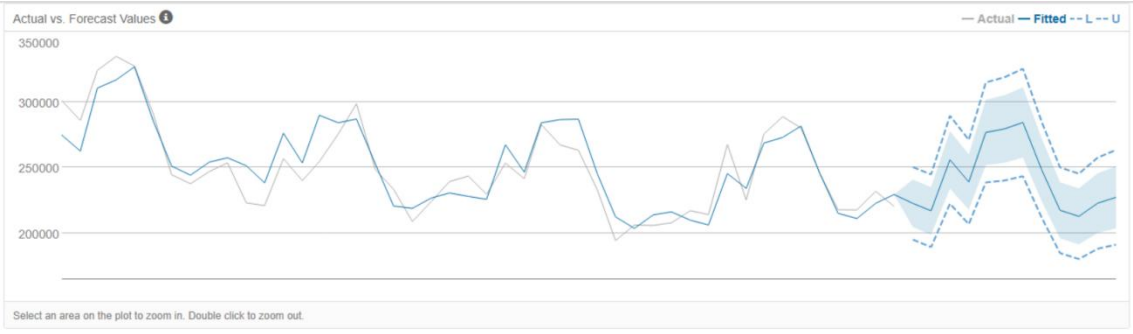
Cluster 1 Forecast:



Cluster 2 Forecast



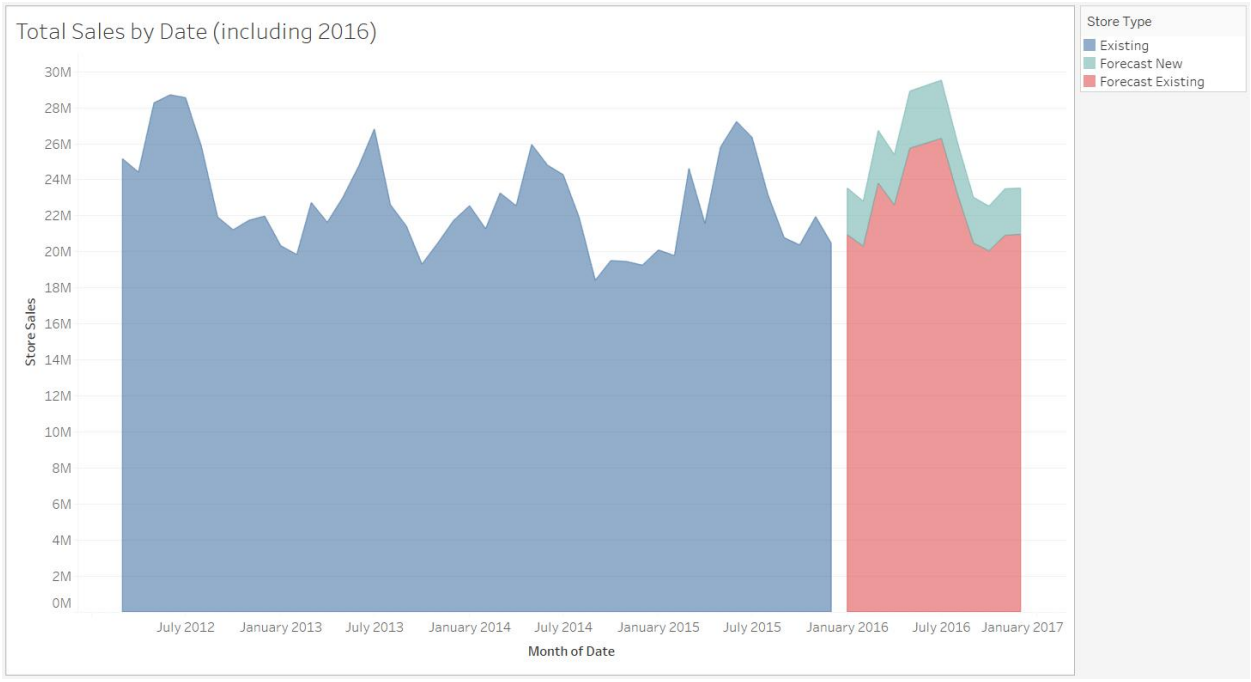
Cluster 3 Forecast:



Below is the complete table with forecast produce sales per month for 2016

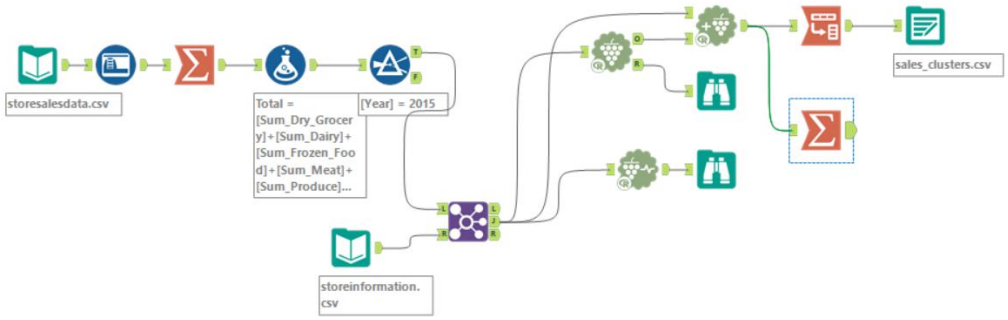
Year	Month	New Stores	Existing Stores
2016	1	$2,588,356.56	$20,933,918.66
2016	2	$2,498,567.17	$20,292,892.36
2016	3	$2,919,067.02	$23,796,371.52
2016	4	$2,797,280.08	$22,577,227.91
2016	5	$3,163,764.86	$25,731,590.76
2016	6	$3,202,813.29	$26,004,247.14
2016	7	$3,228,212.24	$26,278,306.57
2016	8	$2,868,914.81	$23,191,853.31
2016	9	$2,538,372.27	$20,466,390.85
2016	10	$2,485,732.28	$20,032,492.03
2016	11	$2,583,447.59	$20,897,760.51
2016	12	$2,562,181.70	$20,955,902.15

Below is a visualisation of the store sales data forecast to the end of 2016.

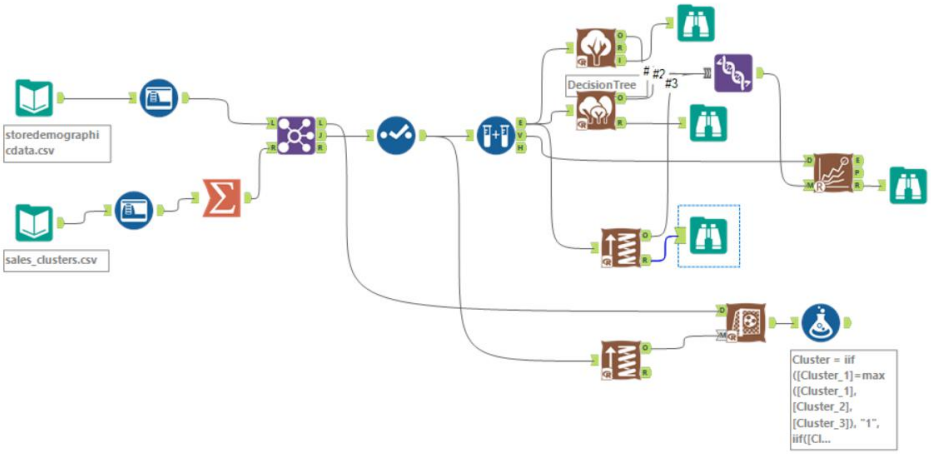


Alteryx Workflows:

Task 1:



Task 2:



Task 3: