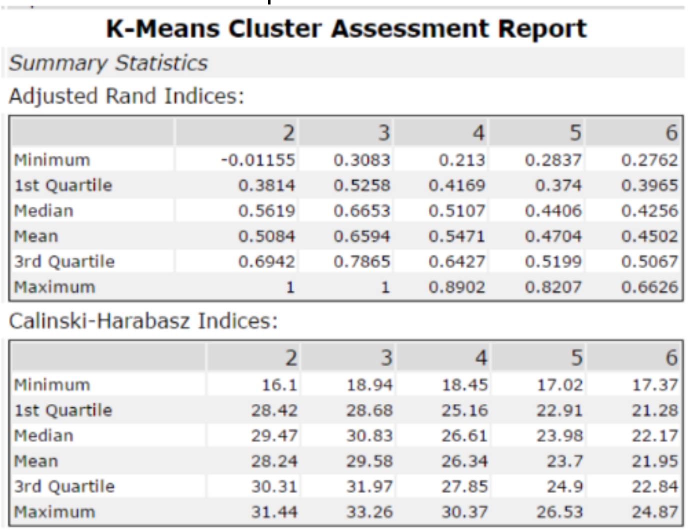
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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



K-Means Cluster Assessment Report

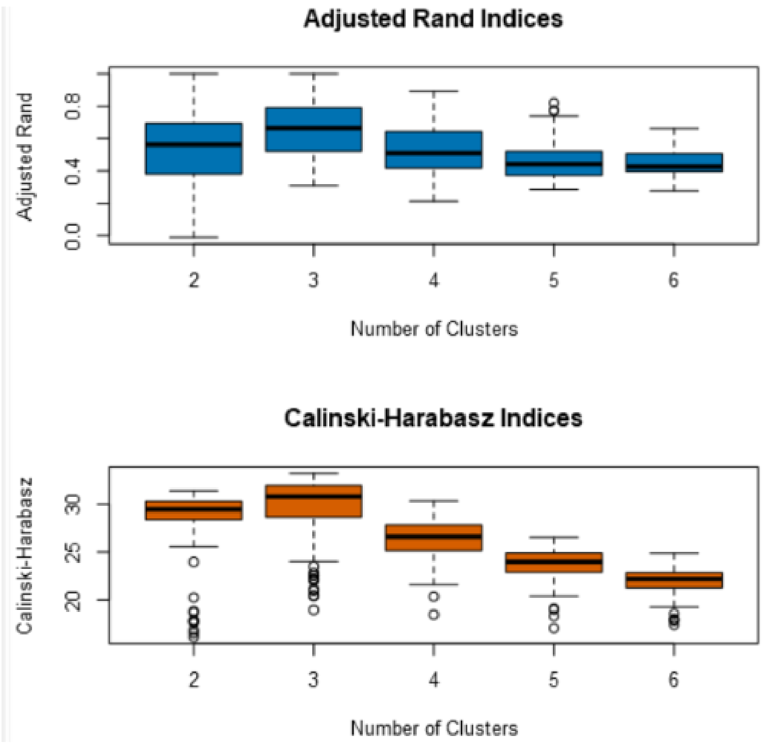
Summary Statistics

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

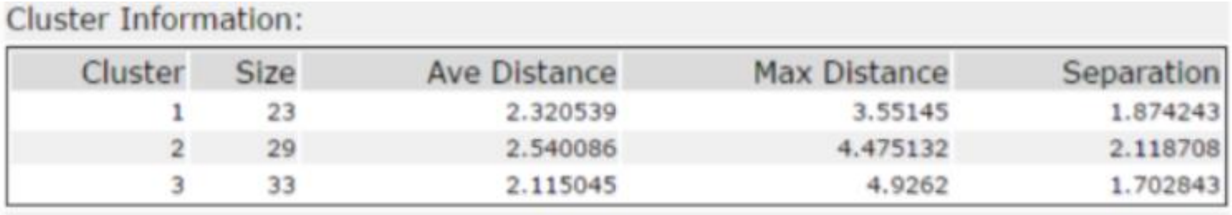
Calinski-Harabasz Indices:

	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



optimal number of store formats is **3** when both the indices registered the highest median value.

2. How many stores fall into each store format?
   Cluster 1= 23 stores, cluster 2 = 29 stores & cluster 3 = 33 stores.

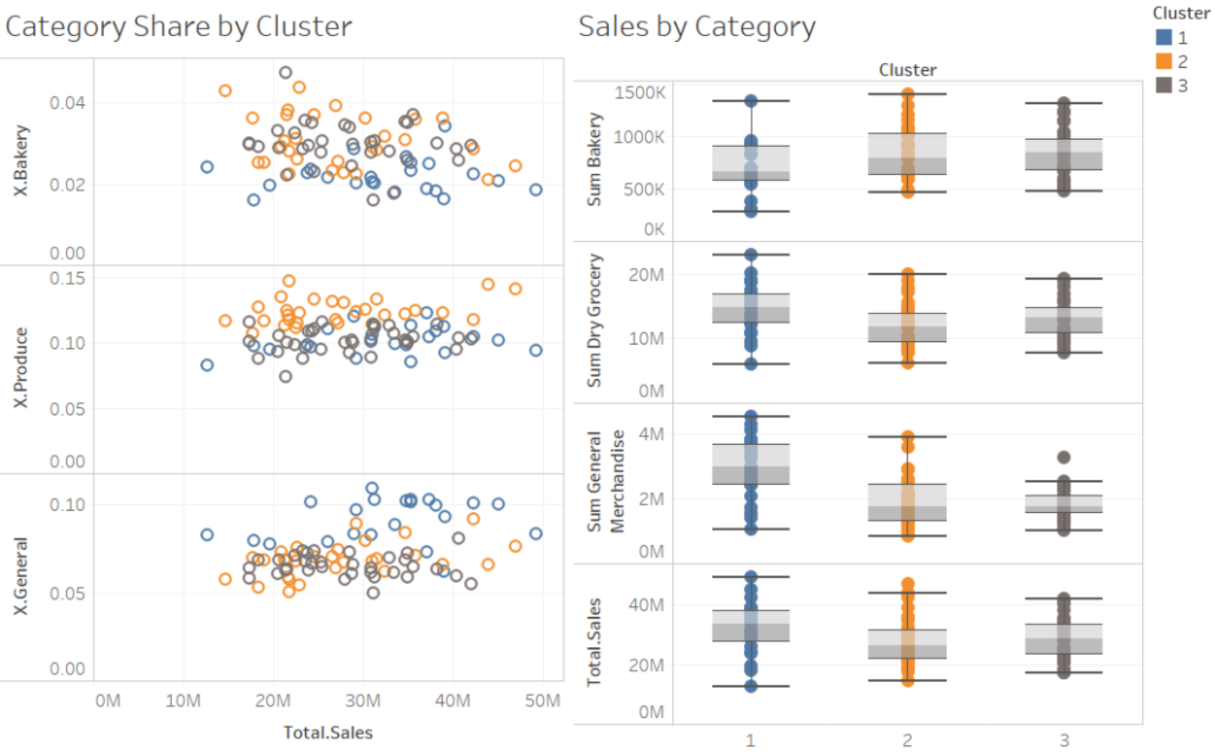


Cluster Information:

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.

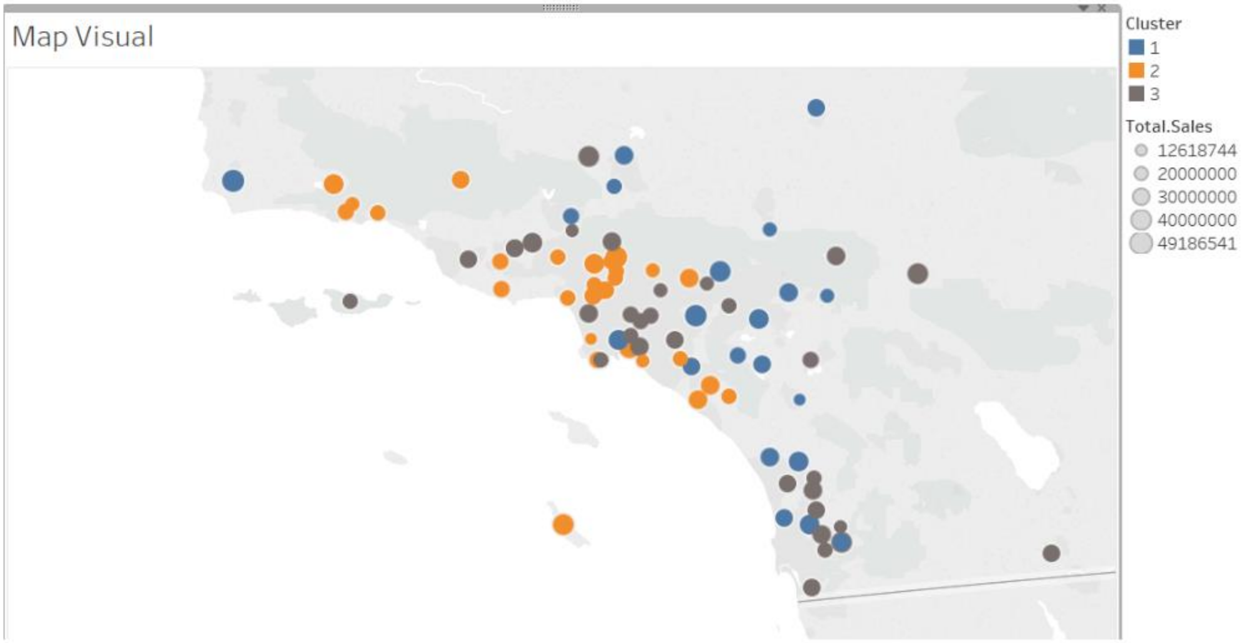


Category Share by Cluster

Sales by Category

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.

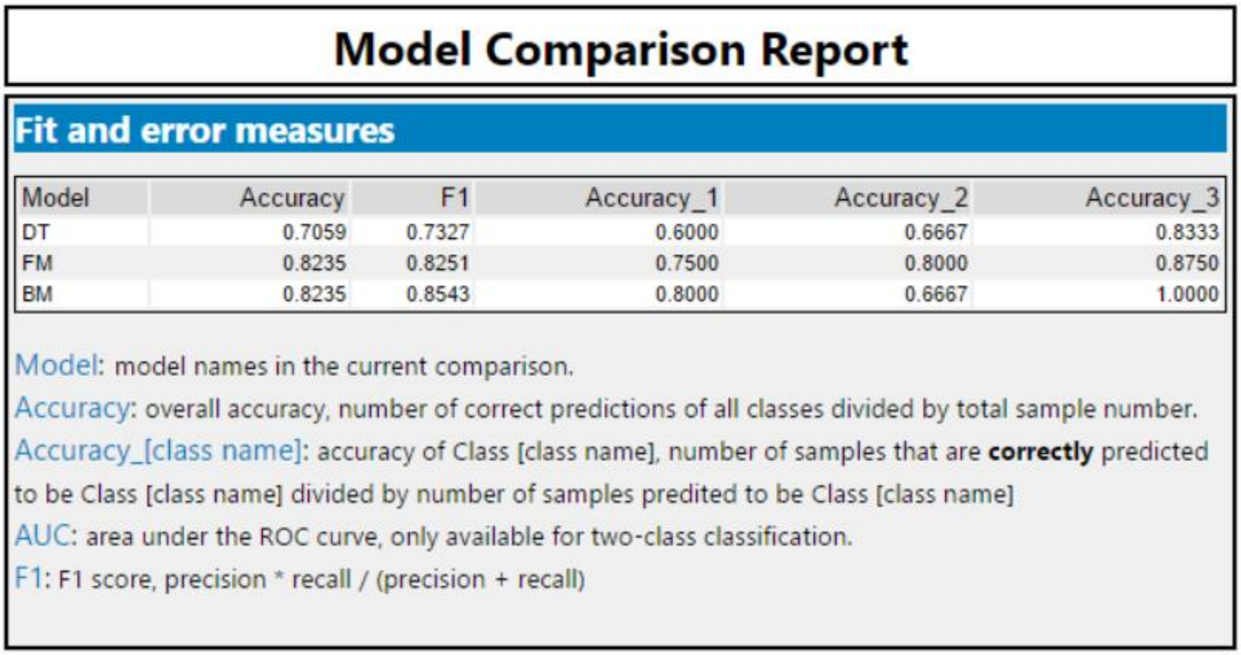
Tableau - https://public.tableau.com/profile/vivek.patel2802#!/vizhome/Task1_15542288322810/MapVisual?publish=yes



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.)

   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

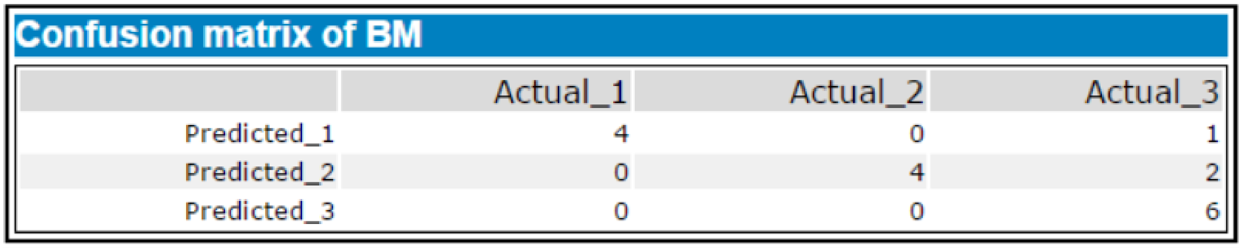
Model: model names in the current comparison.

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]

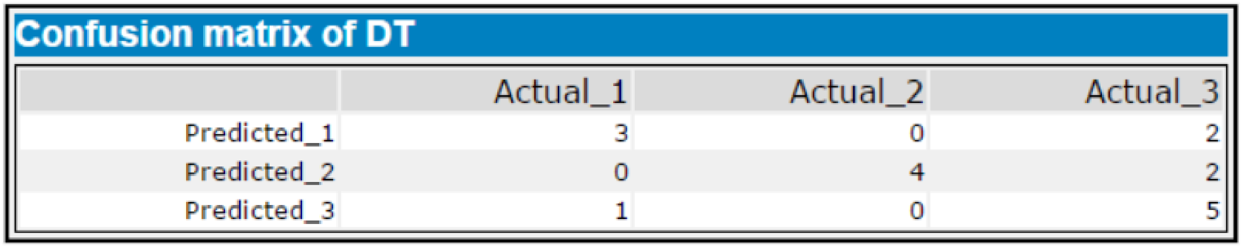
AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)



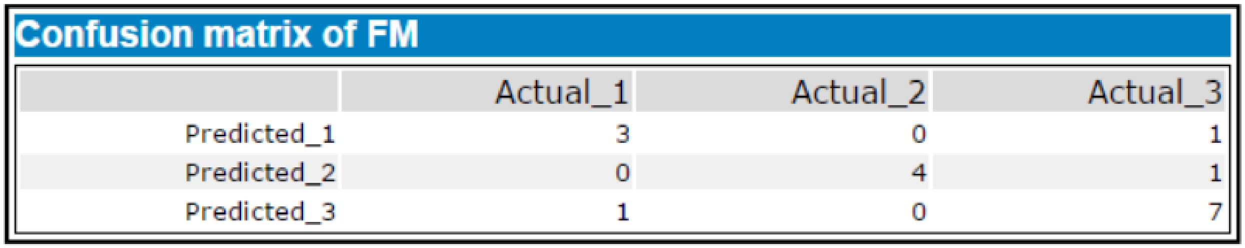
Confusion matrix of BM

	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.

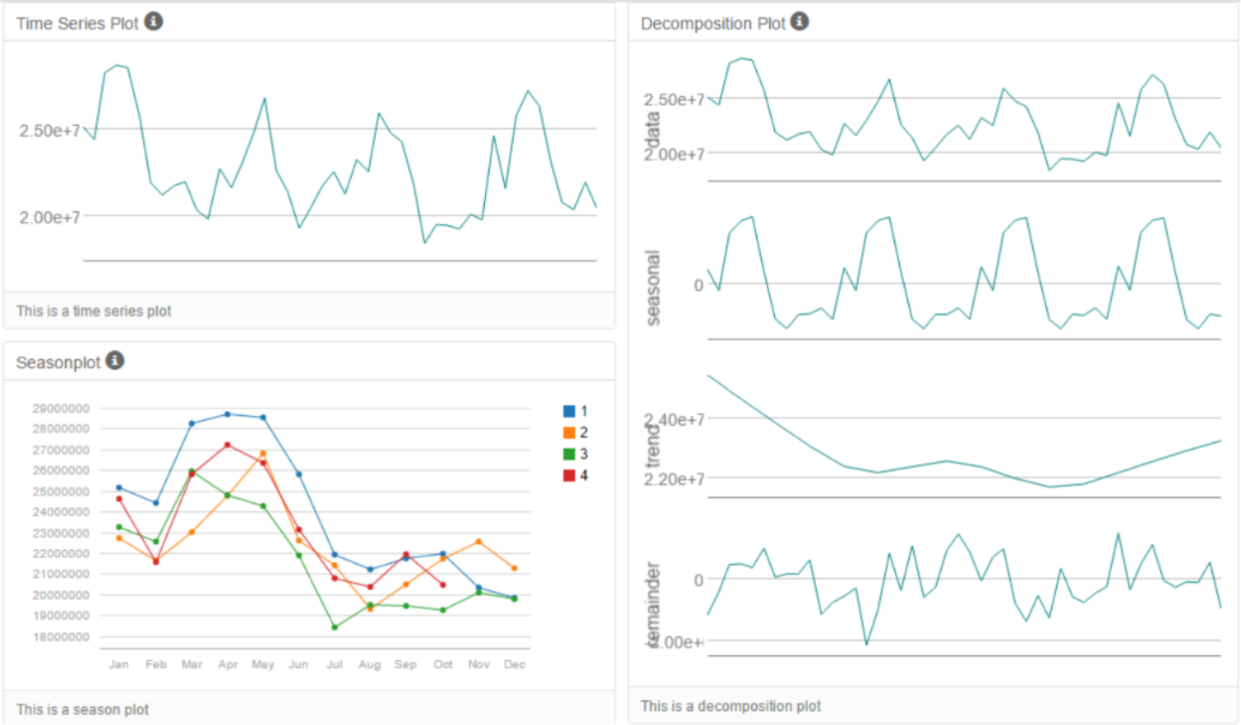
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?

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.


Time Series Plot

This is a time series plot

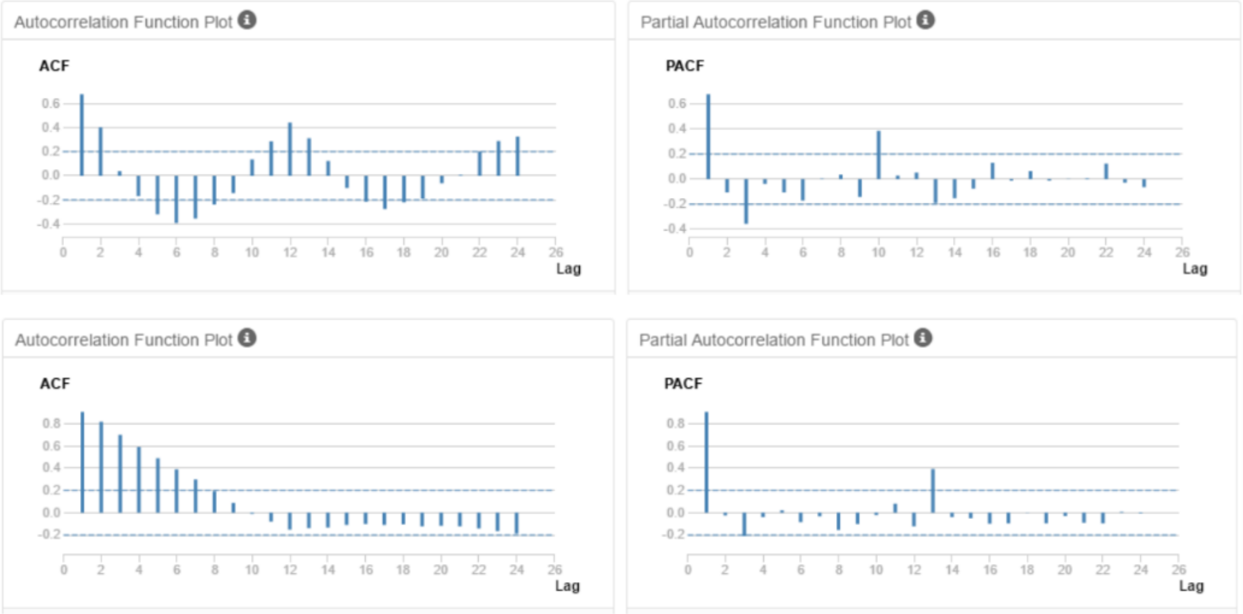
Seasonplot

This is a season plot

Decomposition Plot

This is a decomposition plot

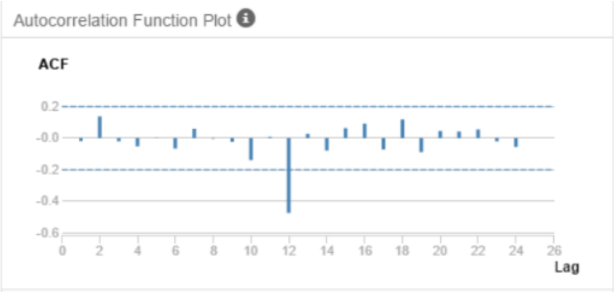
ARIMA(0,1,2)(0,1,0) is used as seasonal difference and seasonal first difference were performed. There is a lag-2.

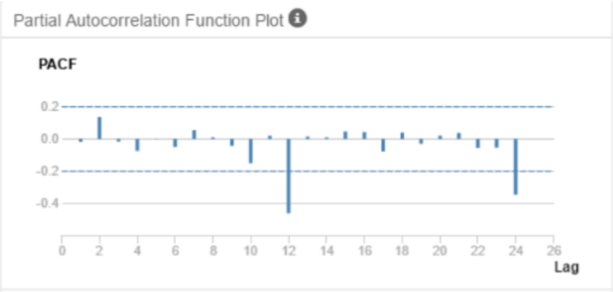

Autocorrelation Function Plot

Partial Autocorrelation Function Plot

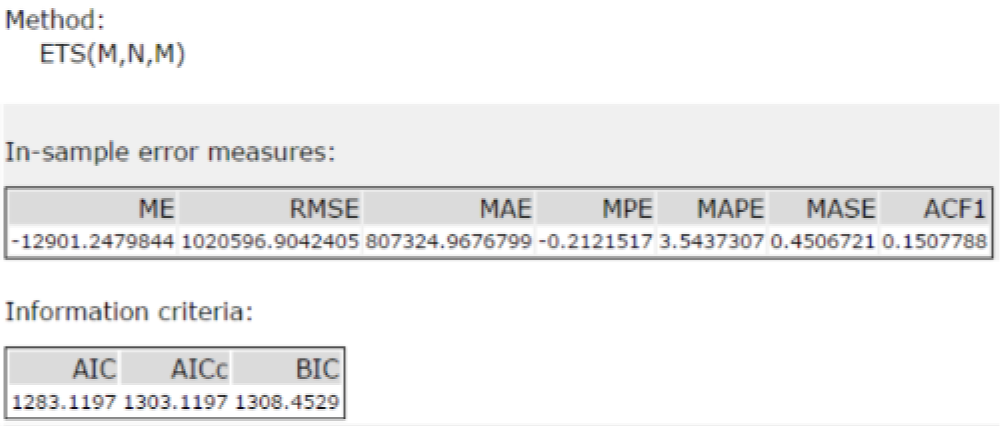
Autocorrelation Function Plot

Partial Autocorrelation Function Plot





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**.



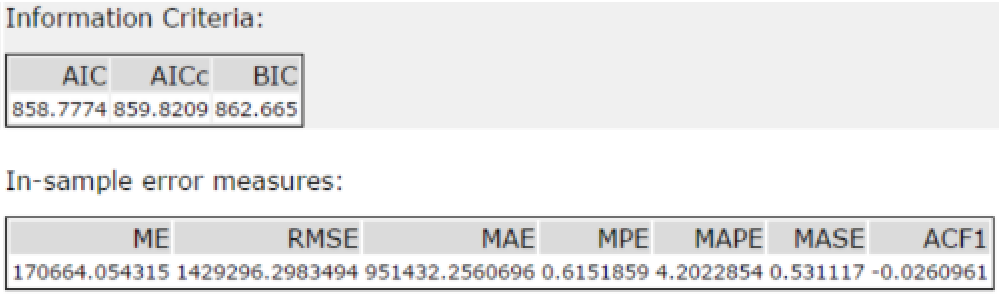
Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529



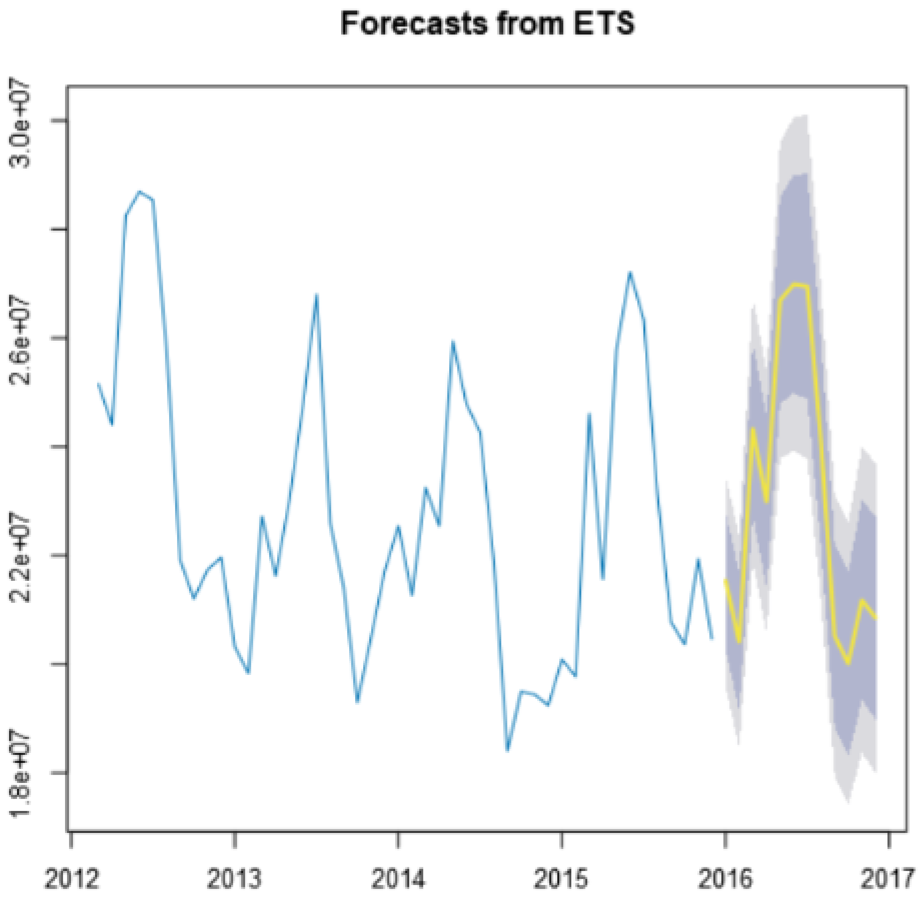
Information Criteria:

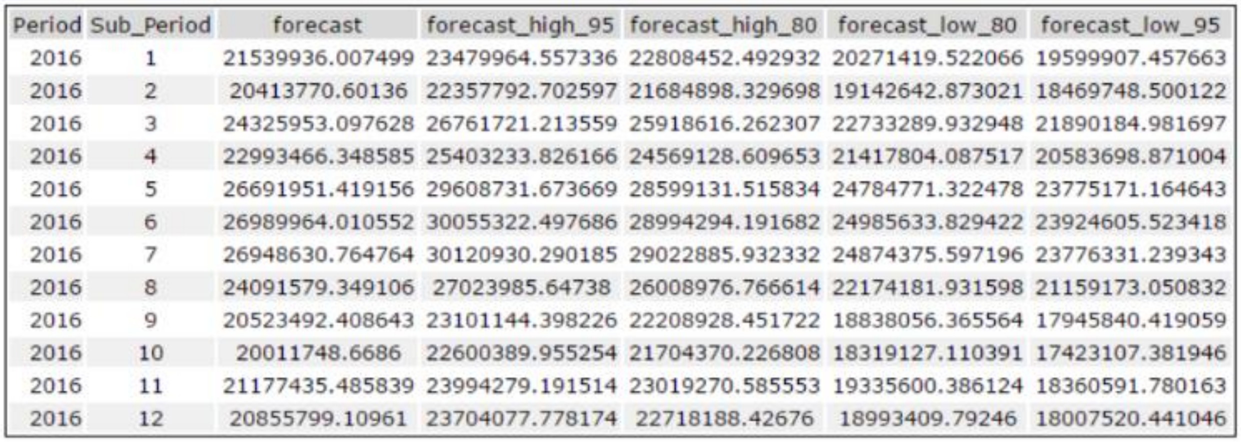
AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961

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

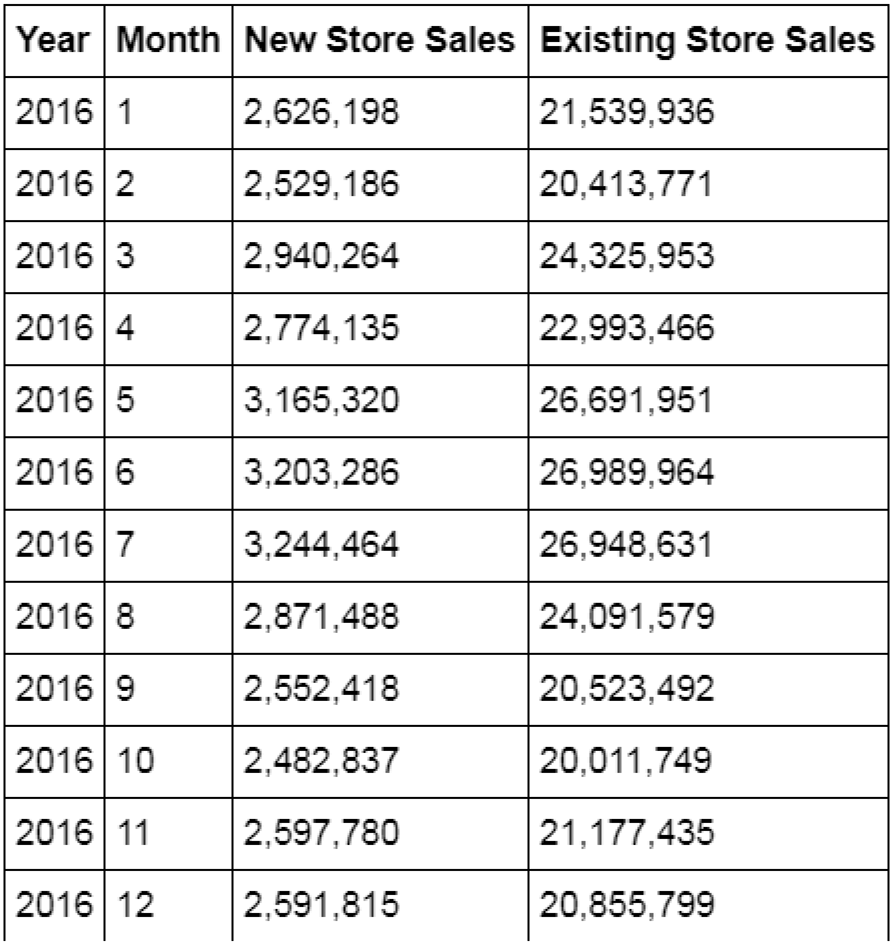




Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

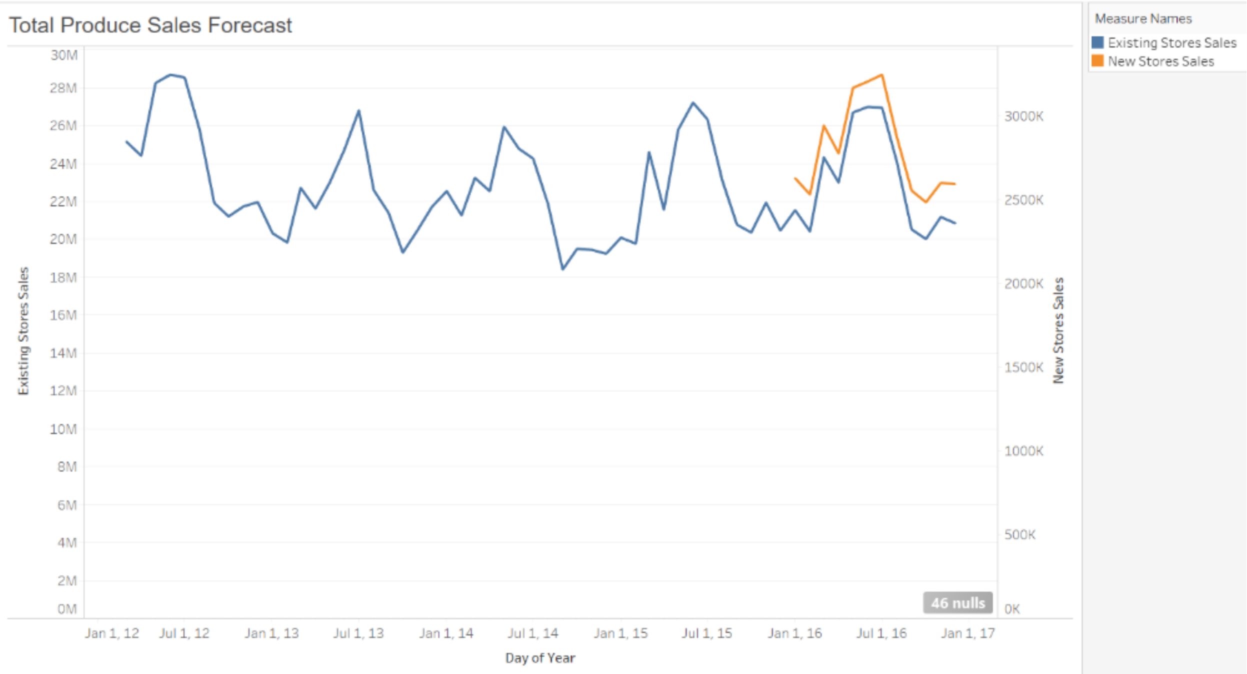
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

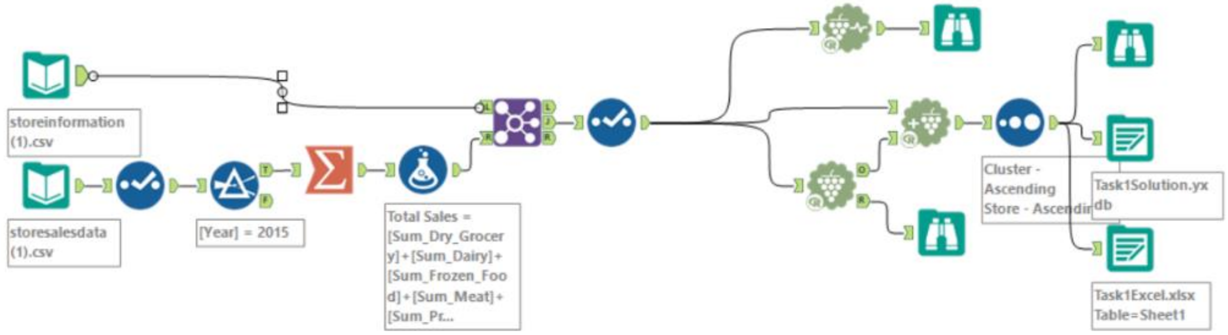
Table below shows the forecast sales for existing stores and new stores. New store sales is obtained by using ETS(M,N,M) analysis with all the 3 individual cluster to obtain the average sales per store. The average sales value (x3 cluster 1, x6 cluster 2, x1 cluster 3) are added up produce New Store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



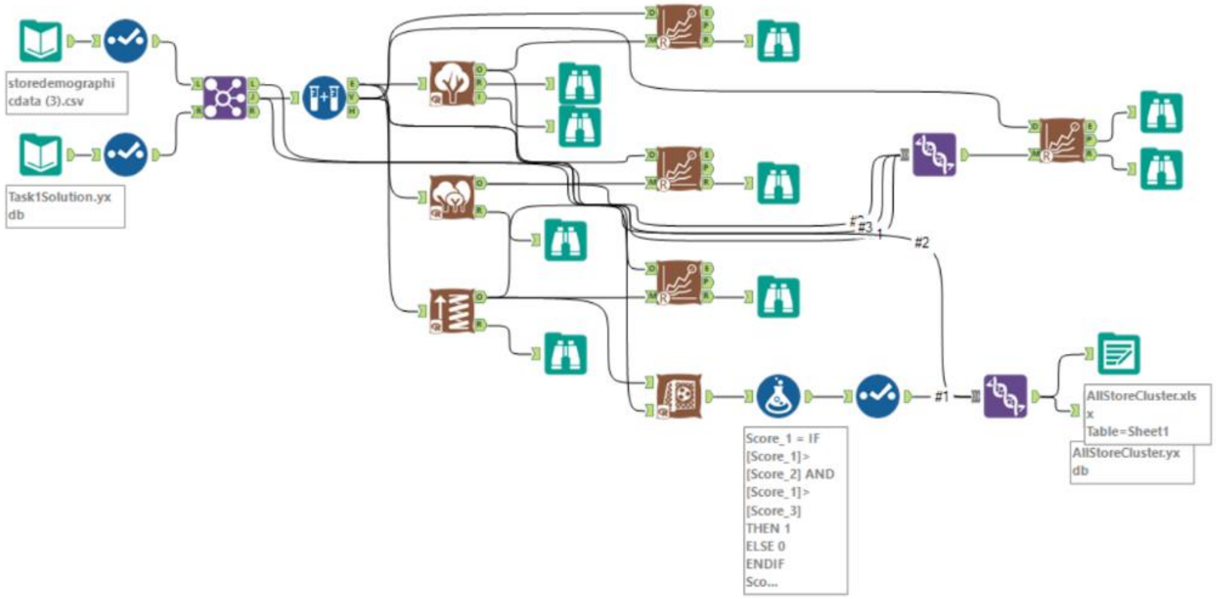
Year	Month	New Store Sales	Existing Store Sales
2016	1	2,626,198	21,539,936
2016	2	2,529,186	20,413,771
2016	3	2,940,264	24,325,953
2016	4	2,774,135	22,993,466
2016	5	3,165,320	26,691,951
2016	6	3,203,286	26,989,964
2016	7	3,244,464	26,948,631
2016	8	2,871,488	24,091,579
2016	9	2,552,418	20,523,492
2016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

Tableau - https://public.tableau.com/profile/vivek.patel2802#!/vizhome/Task3_15542284980890/TotalProduceSalesForecast?publish=yes

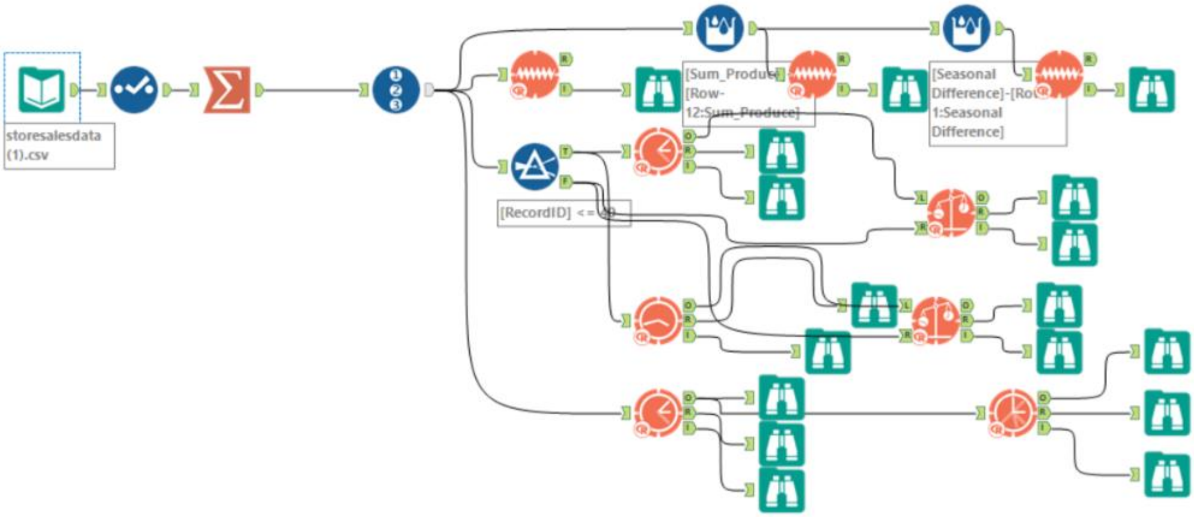

Total Produce Sales Forecast
Measure Names
🟦 Existing Stores Sales
🟧 New Stores Sales
30M
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
Existing Stores Sales
3000K
2500K
2000K
1500K
1000K
500K
0K
New Stores Sales
46 nulls
Jan 1, 12
Jul 1, 12
Jan 1, 13
Jul 1, 13
Jan 1, 14
Jul 1, 14
Jan 1, 15
Jul 1, 15
Jan 1, 16
Jul 1, 16
Jan 1, 17
Day of Year


storeinformation (1).csv
storesalesdata (1).csv
[Year] = 2015
Total Sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...
Cluster - Ascending
Store - Ascendir
Task1Solution.yxdb
Task1Excel.xlsx Table=Sheet1

Workflow 1: Workflow for Task 1


storedemographicdata (3).csv
Task1Solution.yxdb
Score_1 = IF [Score_1]> [Score_2] AND [Score_1]> [Score_3] THEN 1 ELSE 0 ENDIF Sco...
#1
#2
#3
AllStoreCluster.xlsx Table=Sheet1
AllStoreCluster.yxdb

Workflow 2: Workflow for Task 2


storesalesdata (1).csv
[Sum_Produce] [Row-12:Sum_Produce]
[Seasonal Difference]-[Row-1:Seasonal Difference]
[RecordID] <= 43

Workflow 3: Workflow for Task 3