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?

Answer:

Optimal number of store formats is 3. I arrived at this number by looking at the Adjusted Rand and Calinksi-Harabasz Indices.

K-Means Cluster Assessment Report

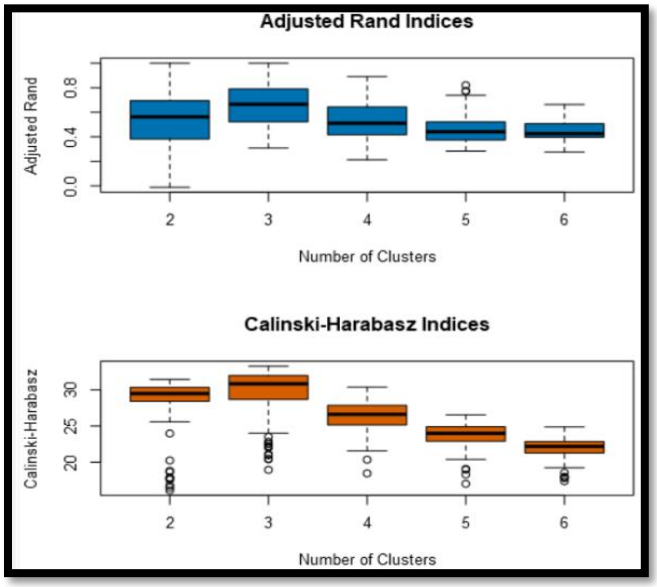
Summary Statistics

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



The Medial value in both the indices is highest at cluster 3. Hence, it helps us decide that 3 clusters are optimal for the business problem.

2. How many stores fall into each store format?

Answer:

23 stores in Cluster 1
29 stores in Cluster 2
33 stores in Cluster 3

Cluster Information:

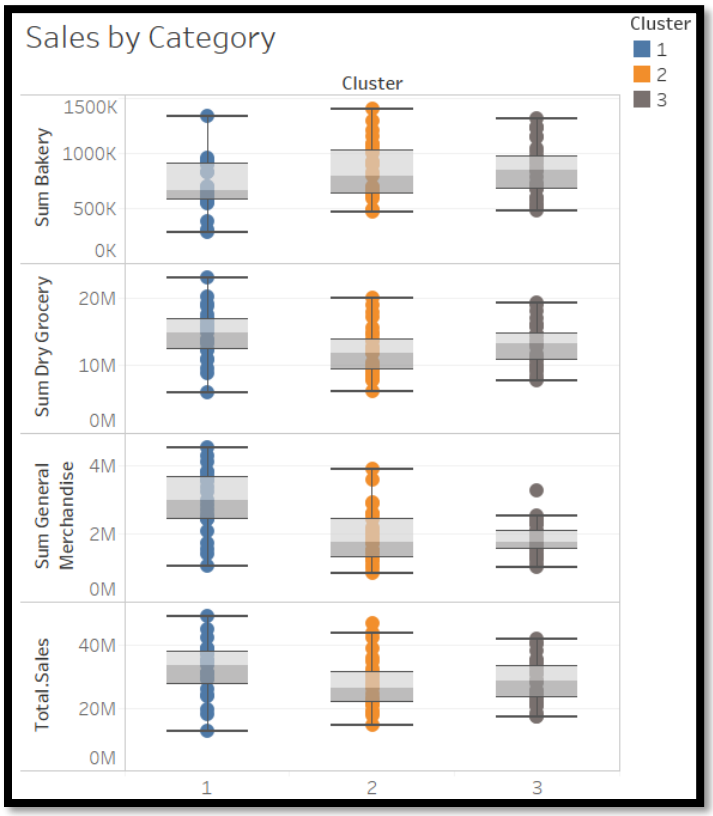
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

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

Answer:

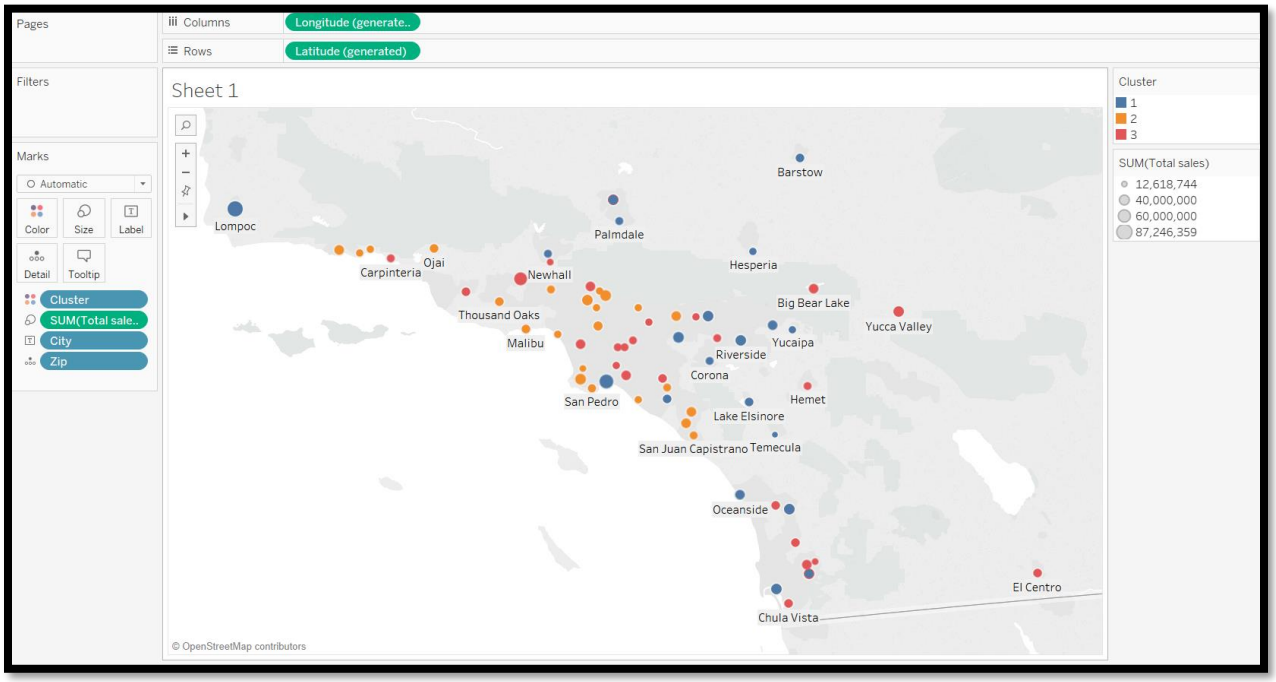
Cluster 1 stores sold more General Merchandise in terms of percentage while Cluster 2 stores sold more Produce.

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 (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

Answer:



https://public.tableau.com/views/Task1_122/GeographicDistributionsofClusters?:embed=y&:display_count=yes&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.)

Answer:

Boosted Model is chosen based on the below model comparison report. It has same accuracy as that of Forest Model but has better 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.

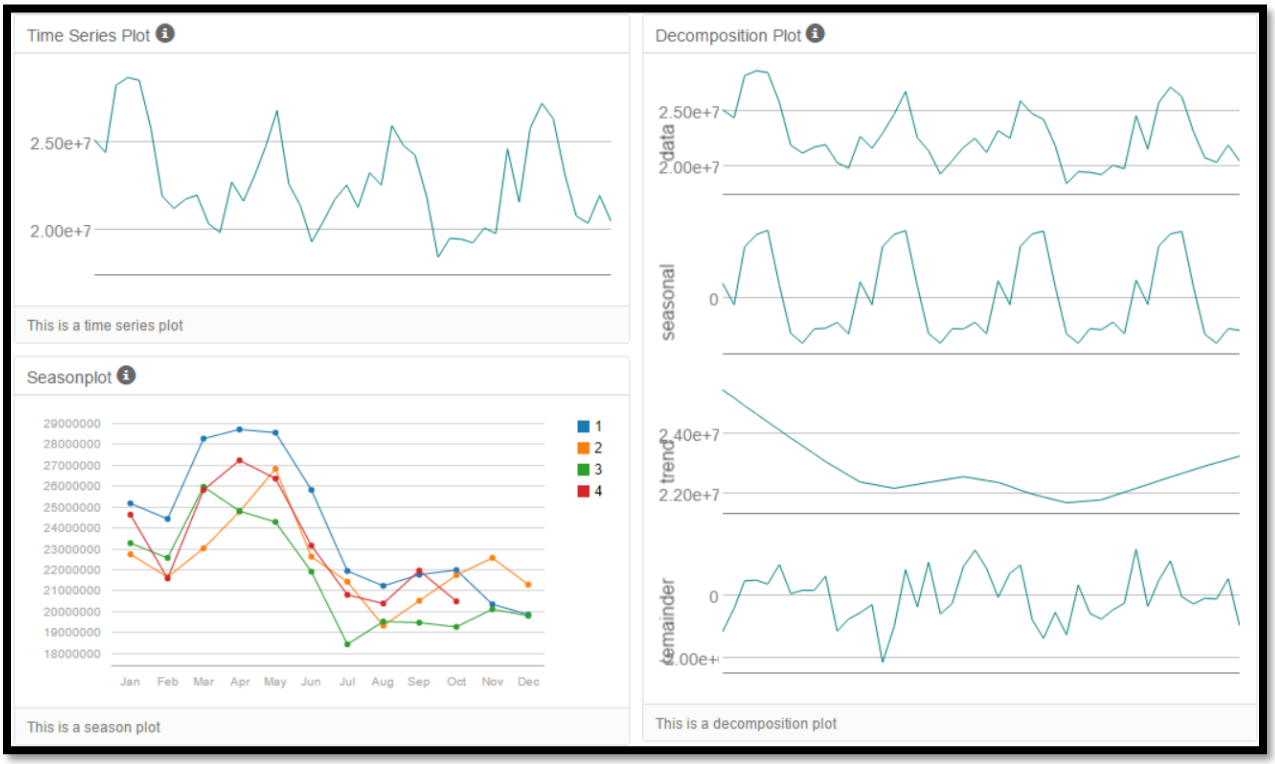
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

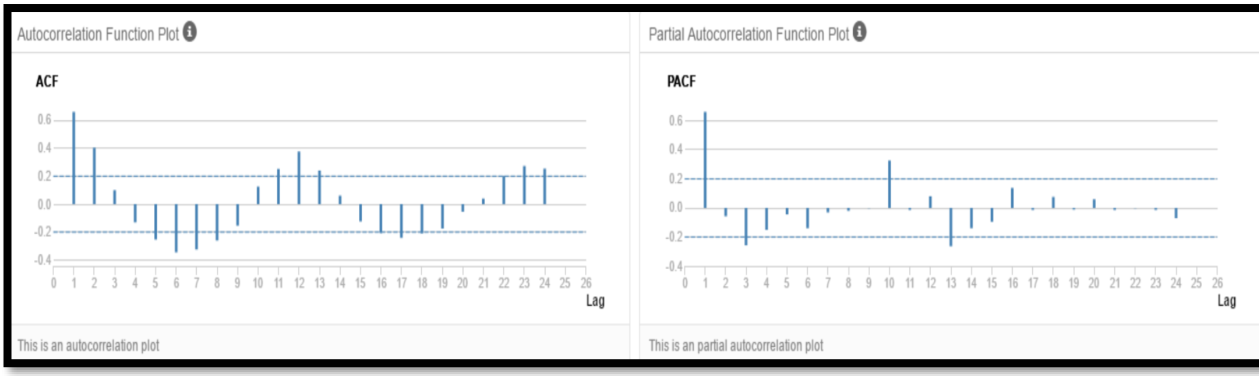
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

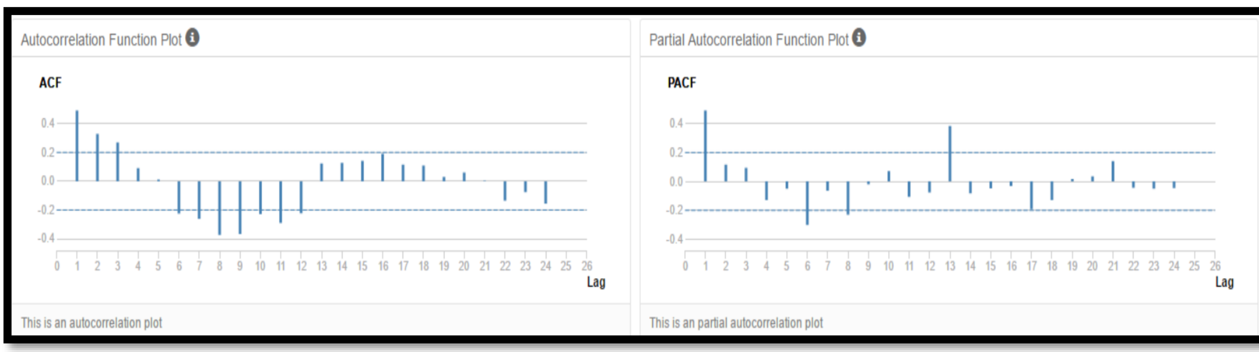
Answer:

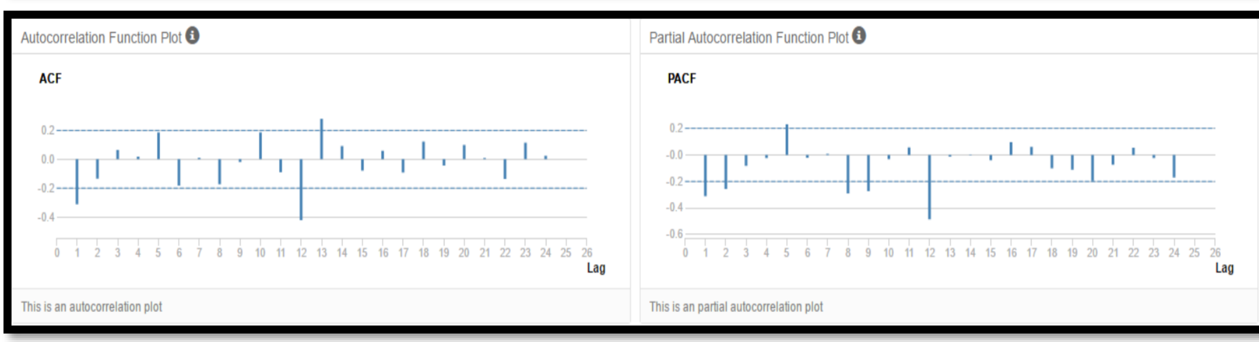
ETS(M,N,M) is used for this problem.
As evident, seasonality increases with time, hence multiplicative.
Trend is not clear so N and Error's magnitude changes with changing seasonality hence, multiplicative.



For ARIMA, ARIMA(0,1,0)(0,1,2)12 is used as seasonal difference and seasonal first difference is used.







In-sample error for ETS:

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2476102	1020596.9028083	807324.9668745	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

In-sample error for ARIMA:

Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.0518584	1429296.2972978	951432.2539369	0.6151859	4.2022854	0.531117	-0.0260961

Accuracy Measures:

Actual and Forecast Values:

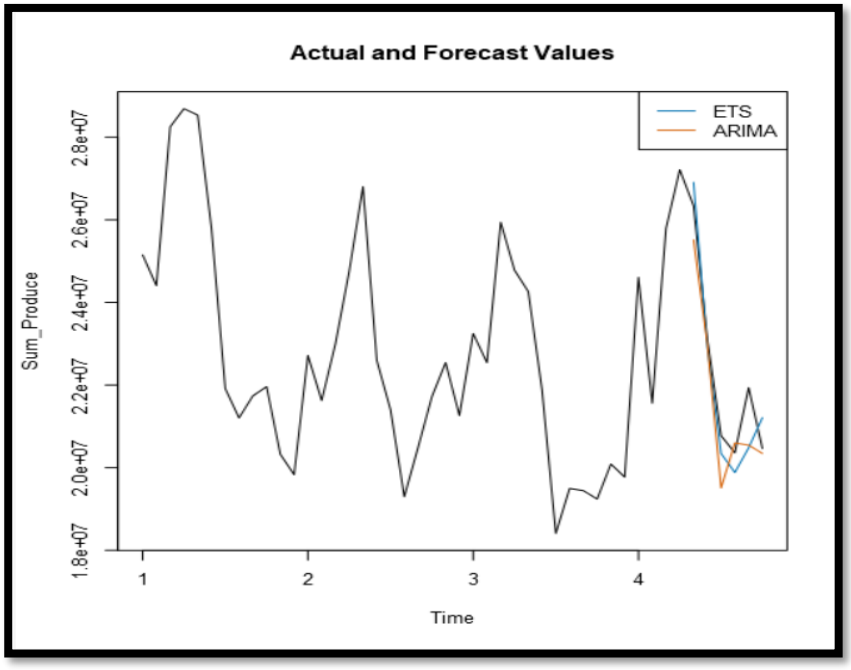
Actual	ETS	ARIMA
26338477.148682	26907095.61646	25515002.56122
23130626.625	22916903.07319	22982398.3927
20774415.934448	20342618.32894	19509673.12135
20359980.583252	19883092.32267	20599981.47462
21936906.821533	20479210.43618	20547162.71748
20462899.330322	21211420.14641	20342794.28431

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA
ARIMA	584382.3	846863.9	664382.6	2.5998	2.9927	0.3909	NA

As we can see, RMSE value of ETS Model is less than that of ARIMA Model which shows less deviation in values. Also, MASE value of ETS is less than that of ARIMA which also makes it a better model since MASE is used to compare difference models (not based on scale).

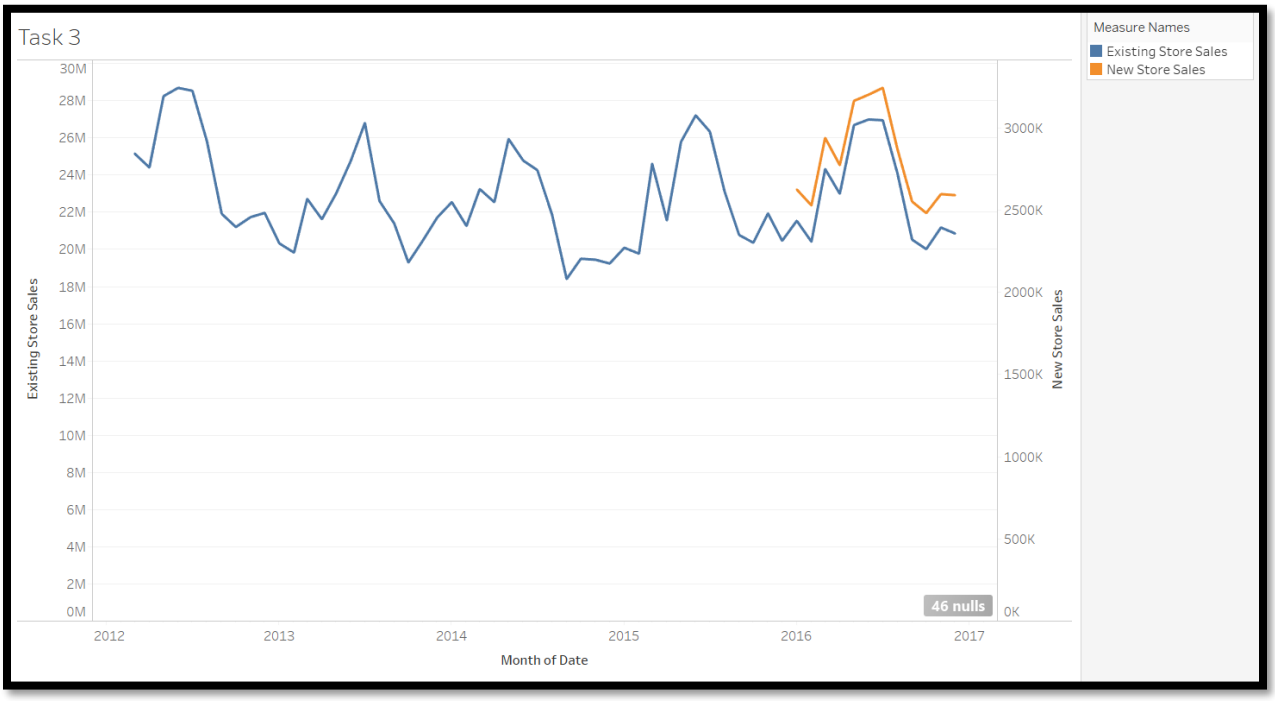
Also, based on graph below it is evident that ETS Model shows better accuracy than ARIMA Model.



3. 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.

Answer:

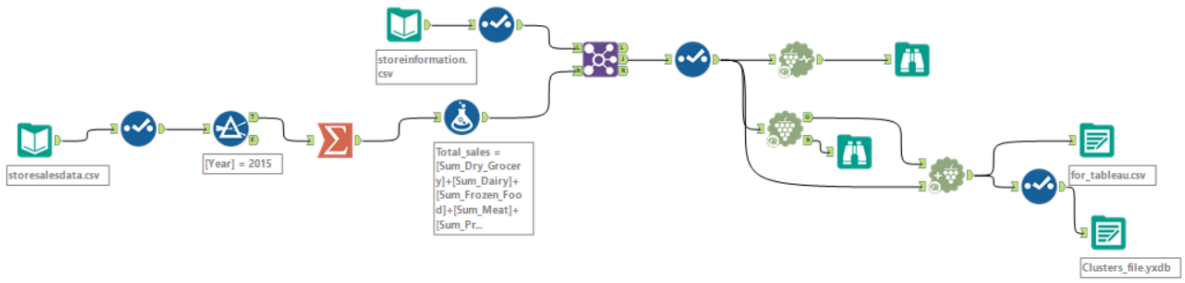
Month	Year	Existing Store Sales	New Store Sales
1	2016	2,15,39,936	26,26,198
2	2016	2,04,13,771	25,29,186
3	2016	2,43,25,953	29,40,264
4	2016	2,29,93,466	27,74,135
5	2016	2,66,91,951	31,65,320
6	2016	2,69,89,964	32,03,286
7	2016	2,69,48,631	32,44,464
8	2016	2,40,91,579	28,71,488
9	2016	2,05,23,492	25,52,418
10	2016	2,00,11,749	24,82,837
11	2016	2,11,77,435	25,97,780
12	2016	2,08,55,799	25,91,815



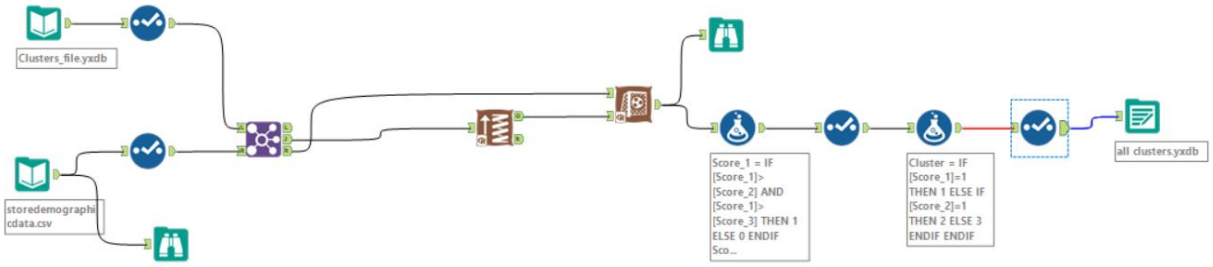
https://public.tableau.com/views/Task3_198/Task3?:embed=y&:display_count=yes&publish=yes

Workflows:

TASK 1 – Assigning clusters to existing stores:



TASK 2 - Assigning clusters to new stores:



TASK 3 - Forecasting: