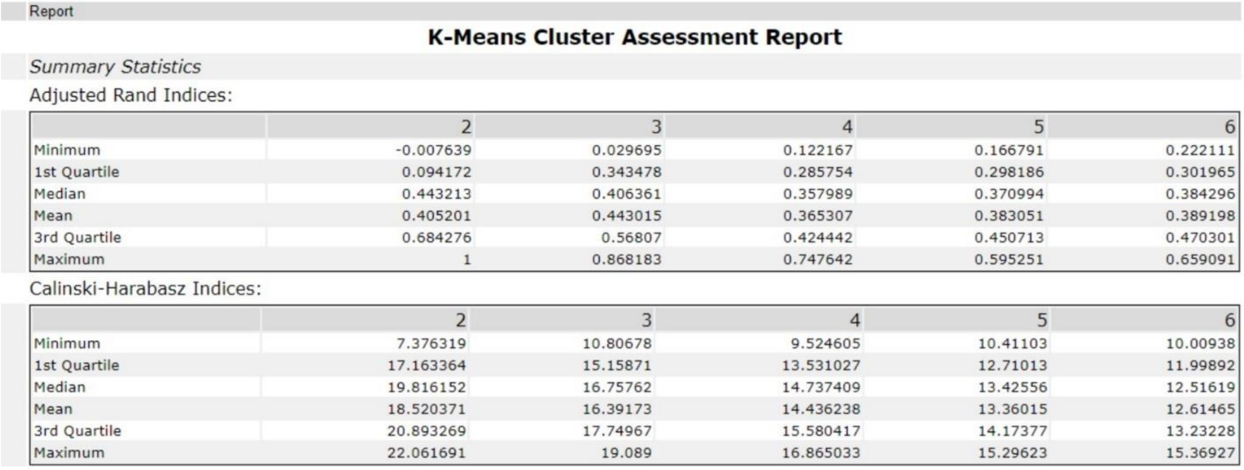
Project: Predictive Analytics Capstone

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 K-means report, Adjusted Rand and Calinski-Harabasz, the optimal number of store formats is 3. See below – both indices have high median values at 3 and the spread of the iterations is minimized.



Report

K-Means Cluster Assessment Report

Summary Statistics

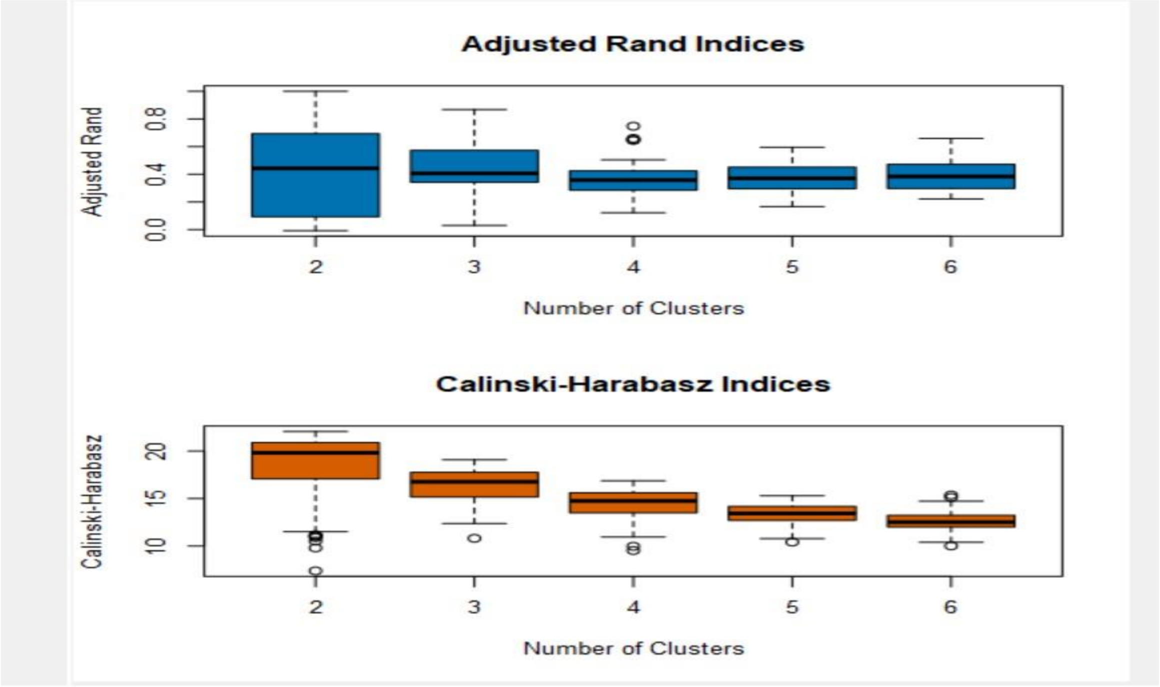
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.007639	0.029695	0.122167	0.166791	0.222111
1st Quartile	0.094172	0.343478	0.285754	0.298186	0.301965
Median	0.443213	0.406361	0.357989	0.370994	0.384296
Mean	0.405201	0.443015	0.365307	0.383051	0.389198
3rd Quartile	0.684276	0.56807	0.424442	0.450713	0.470301
Maximum	1	0.868183	0.747642	0.595251	0.659091

Calinski-Harabasz Indices:

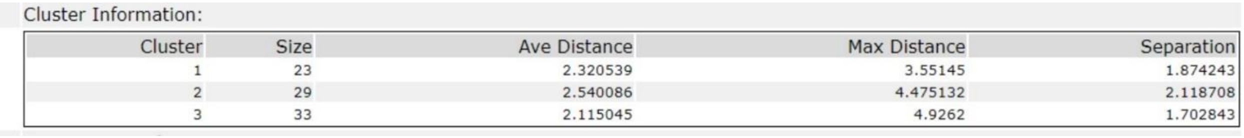
	2	3	4	5	6
Minimum	7.376319	10.80678	9.524605	10.41103	10.00938
1st Quartile	17.163364	15.15871	13.531027	12.71013	11.99892
Median	19.816152	16.75762	14.737409	13.42556	12.51619
Mean	18.520371	16.39173	14.436238	13.36015	12.61465
3rd Quartile	20.893269	17.74967	15.580417	14.17377	13.23228
Maximum	22.061691	19.089	16.865033	15.29623	15.36927

Plots



2. How many stores fall into each store format?

   Cluster 1 has 23 stores, Cluster 2 has 29 stores and Cluster 3 has 33 stores. See below.



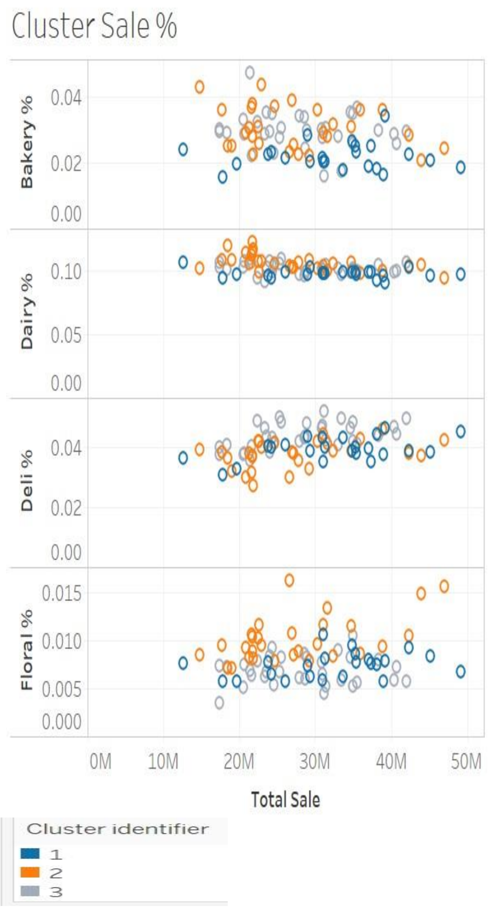
Cluster Information:

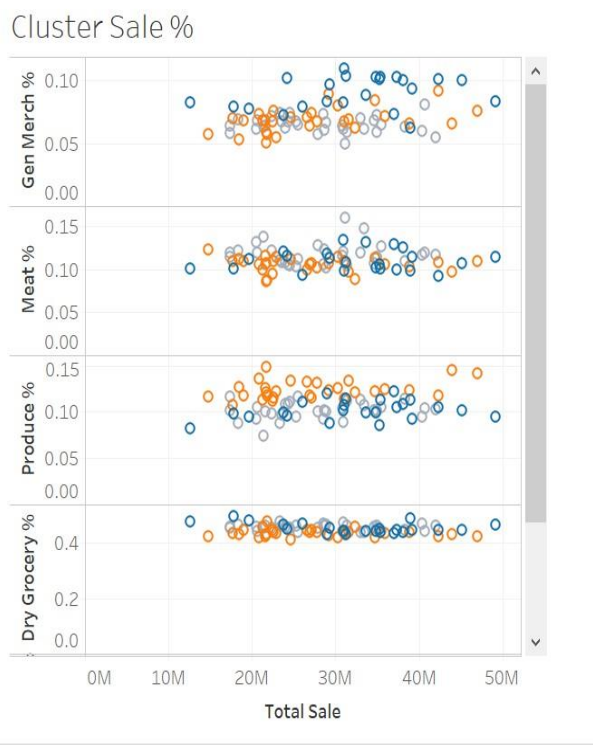
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

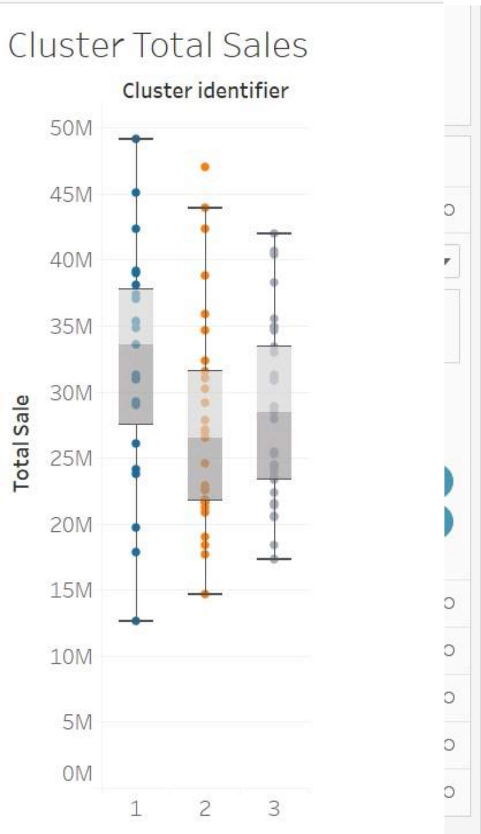
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   Percentage of floral and produce sales are highest in cluster 2.
   Percentage of general merchandise sales are highest in Cluster 1.

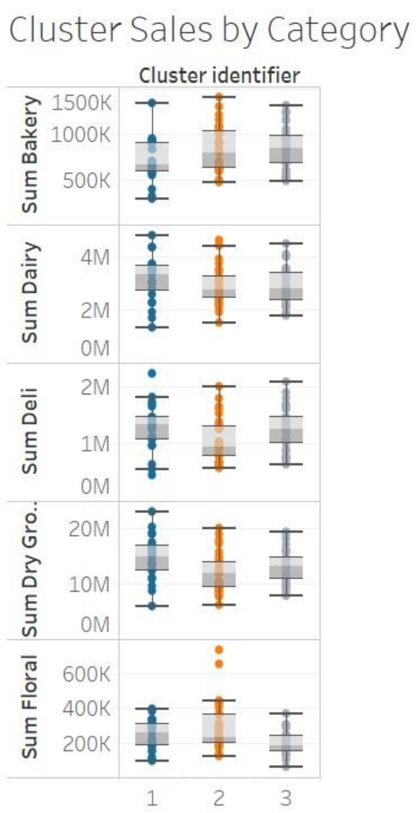
   Cluster 1 has the highest median and range of total sales. Cluster 3 is the most compact of the three store clusters. See below and next page.



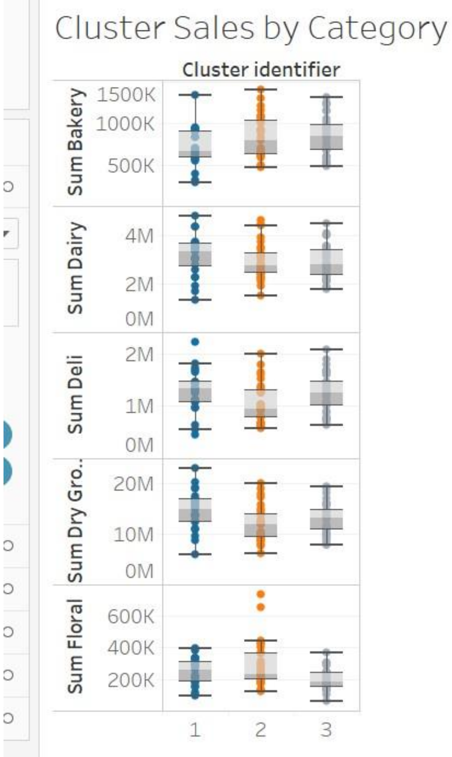




Cluster Total Sales

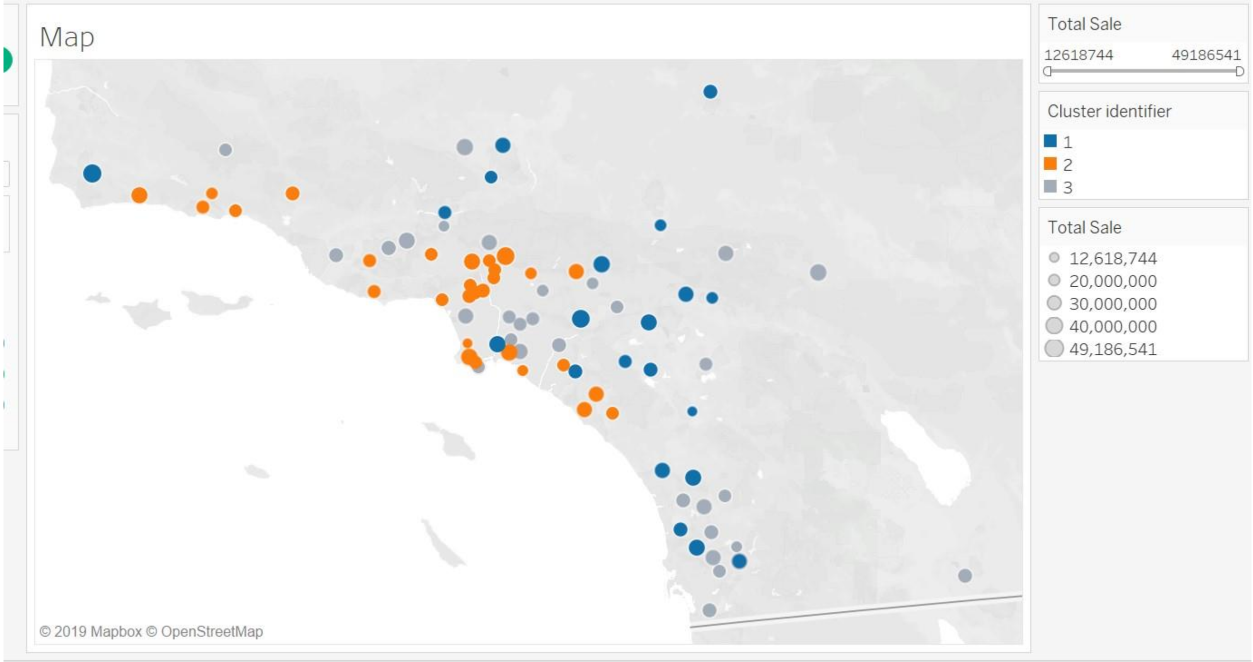


Cluster Sales by Category



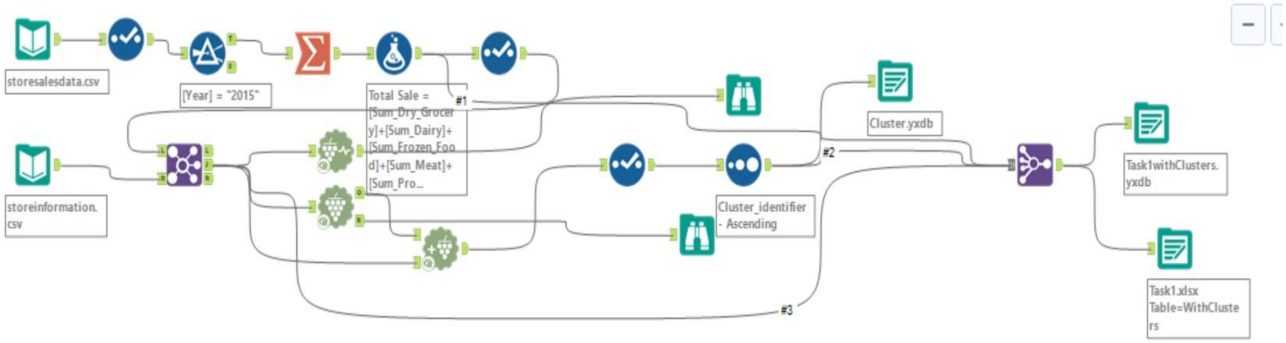
Cluster Sales by Category

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.



https://public.tableau.com/profile/farhana.hasan#!/vizhome/ClusterMap_15655815631400/Map

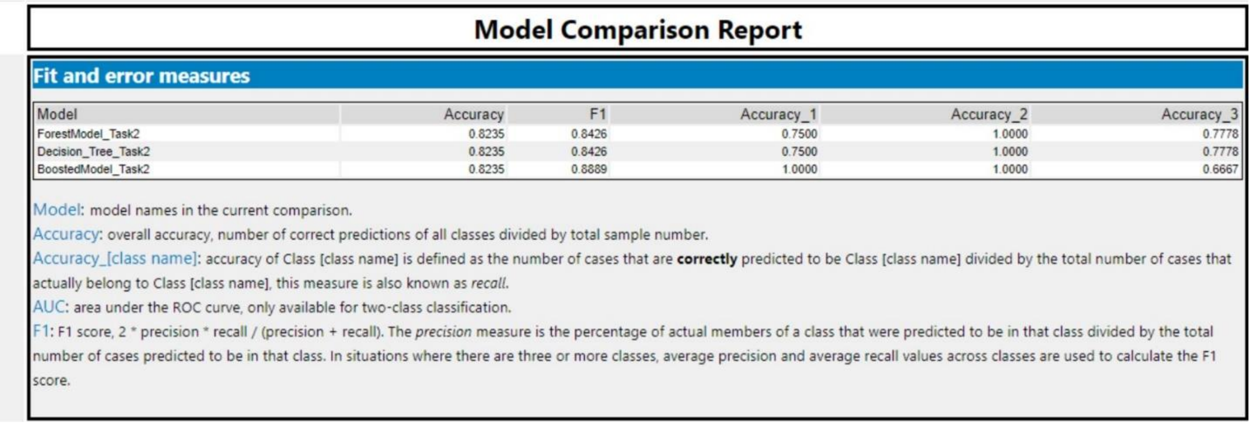
Alteryx Workflow for Task1


storesalesdata.csv
[Year] = "2015"
Total Sale = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pro...
#1
storeinformation.csv
Cluster.yxdb
#2
Cluster_identifier - Ascending
Task1withClusters.yxdb
Task1.xlsx Table=WithClusters
#3

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 chose the Boosted Model. It has a high F1 score compared to other models. It also accurately segments clusters 1 & 2 - 100% of the time and cluster 3 - 66.67% of the time. See below.

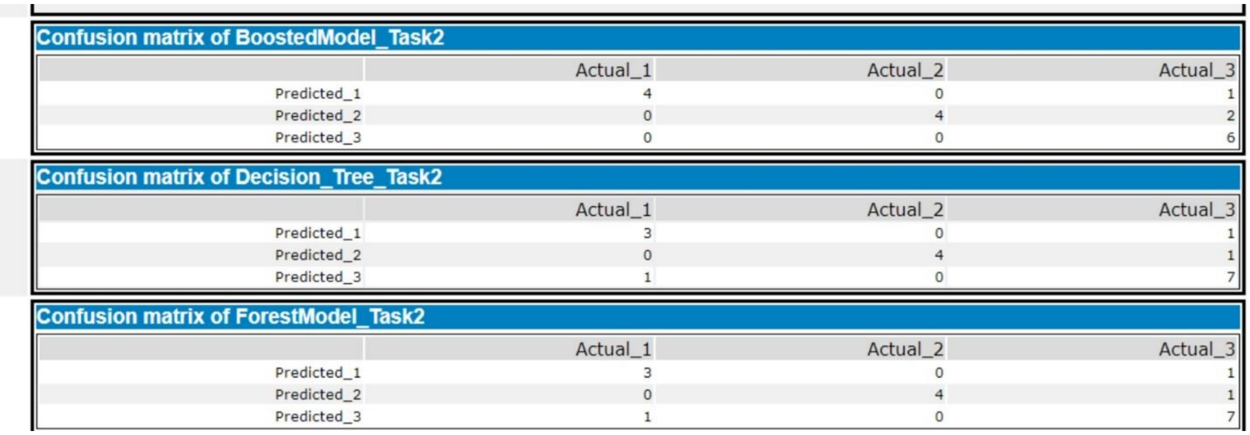


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
ForestModel_Task2	0.8235	0.8426	0.7500	1.0000	0.7778
Decision_Tree_Task2	0.8235	0.8426	0.7500	1.0000	0.7778
BoostedModel_Task2	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 BoostedModel_Task2

	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_Task2

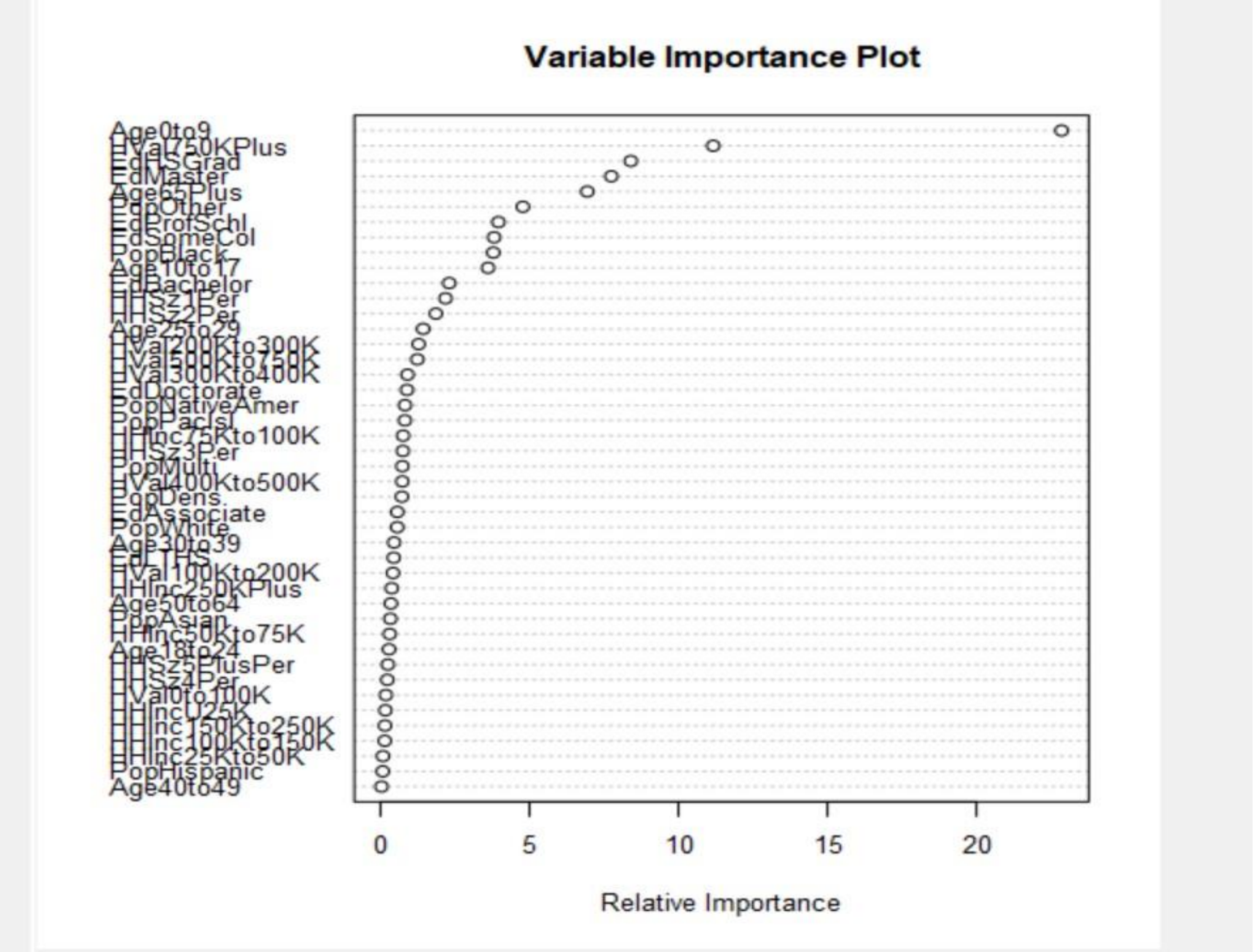
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

Confusion matrix of ForestModel_Task2

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

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

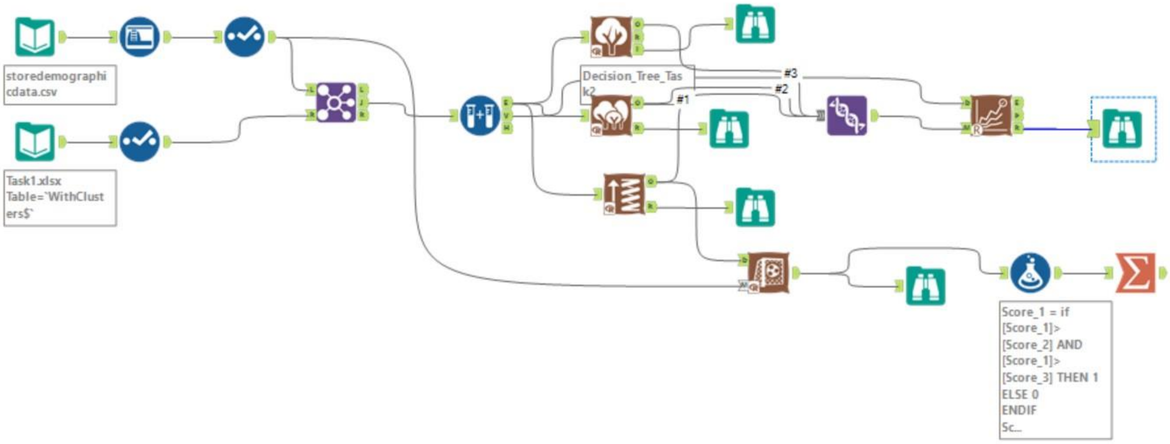
   The three most important variables are Age0to9, HVal750KPlus and EdHSGrad. See next page.



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

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

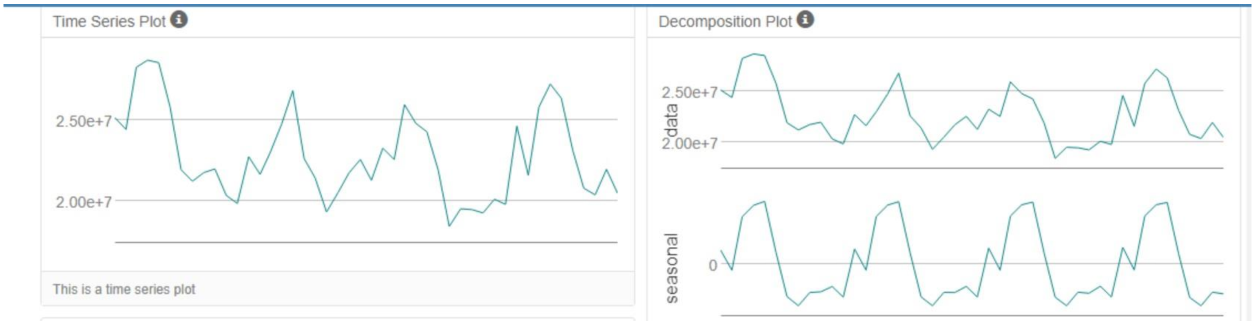
Alteryx workflow for Task 2

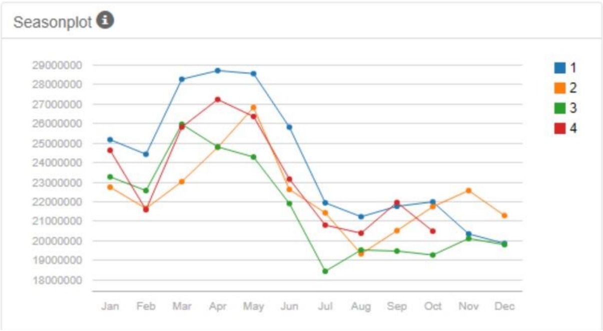


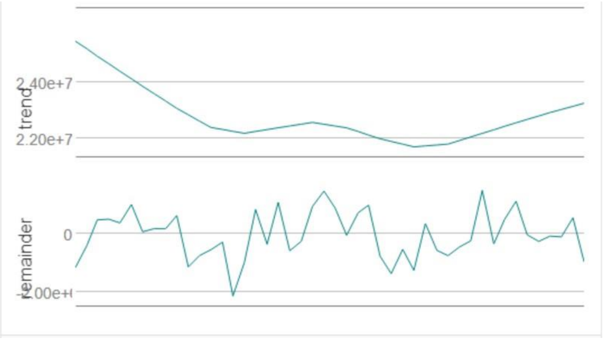
Task 3: Predicting Produce Sales

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?

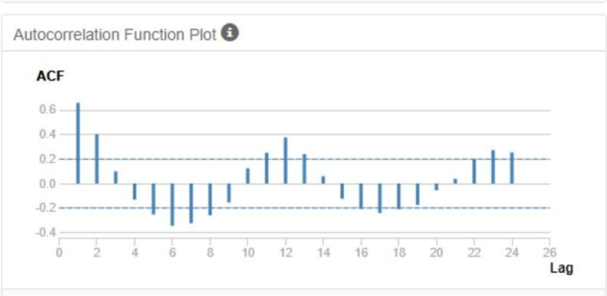
ETS (MNM) without dampening is used for the ETS model because of its lower AIC. The error does not have constant variance over time, so it should be applied multiplicatively. The trend line changes direction, so we should select none for the ETS tool trend line. Finally, the seasonality is growing over time so it should be applied multiplicatively. I also used auto configuration to confirm these findings. See below.

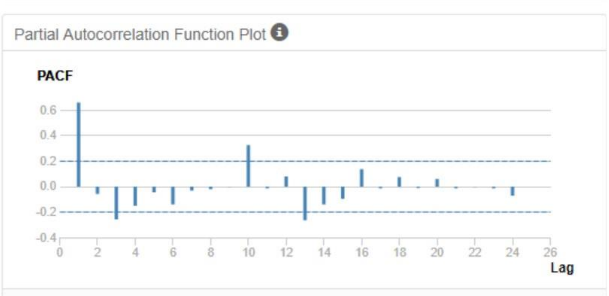


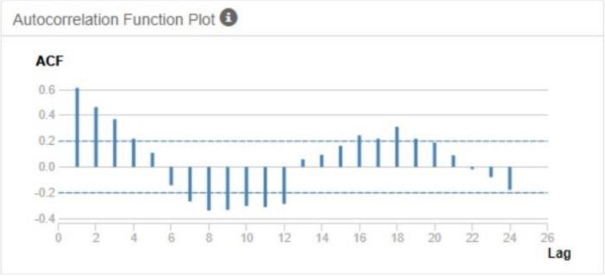


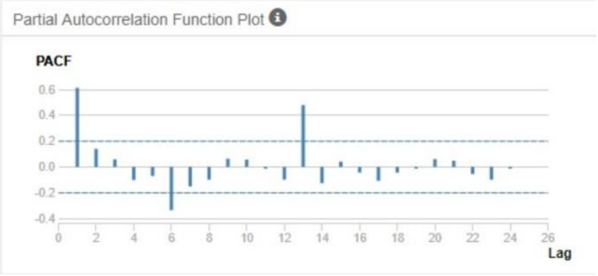


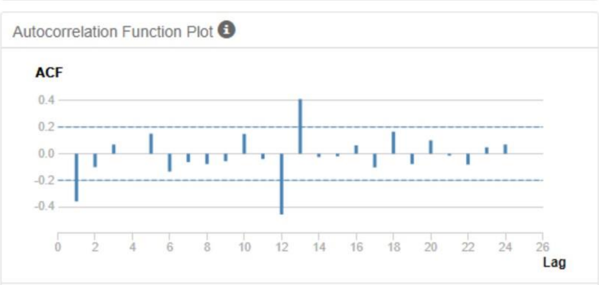
For ARIMA consider the ACF and PACF plots below. There is a lag at 2. ARIMA (0,1,2)(0,1,0) is used and seasonal difference and seasonal first difference are performed.

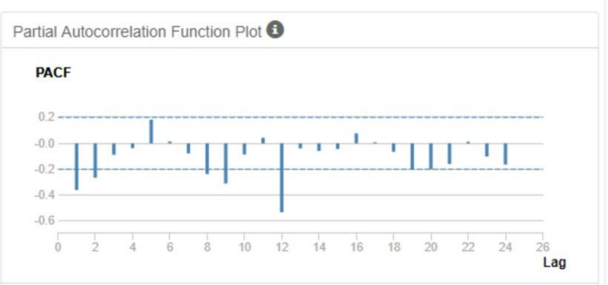




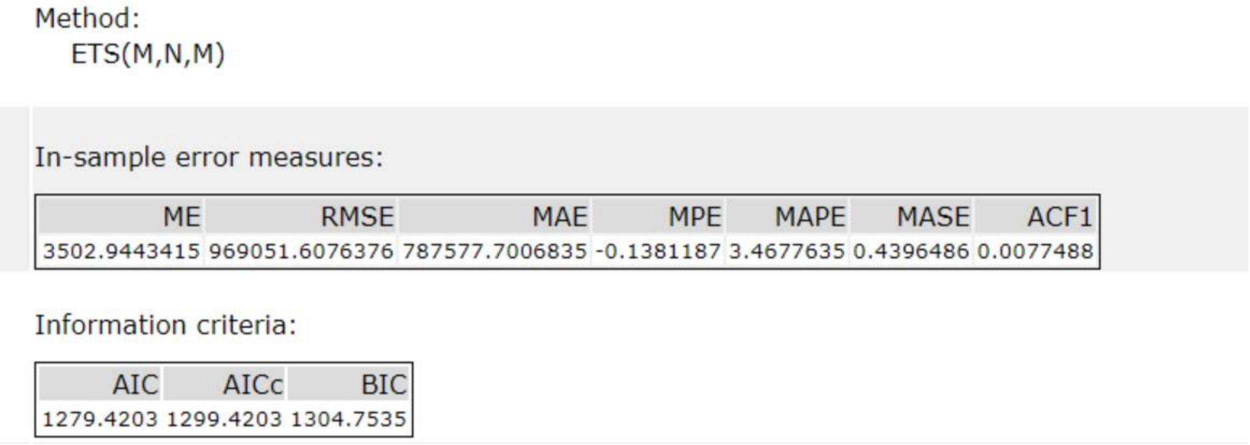








The ETS model is chosen. ETS RMSE (standard deviation from mean) of 969,052 is lower than the ARIMA RMSE of 1,429,296. ETS model MASE is 0.44 and the ARIMA model MASE is 0.53. Moreover, ARIMA MAPE is 4.2% as compared to the ETS MAPE of 3.47%. Finally, the AIC of the ARIMA model is 858 and the ETS is 1,279. See below.



Method:

ETS(M,N,M)

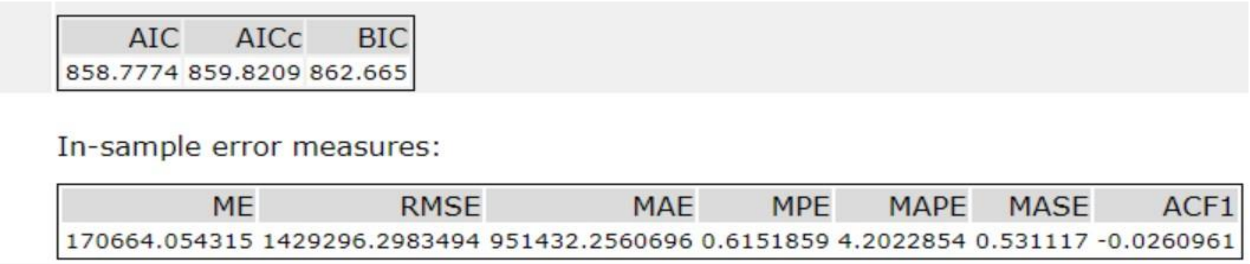
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

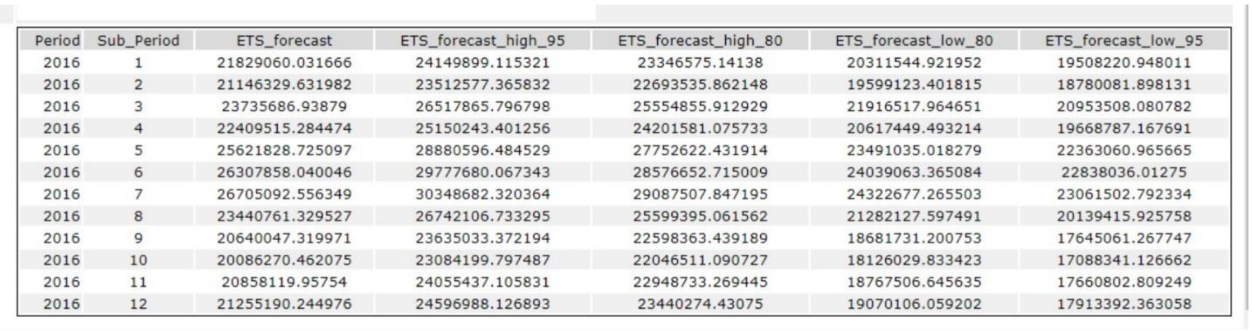
ARIMA Model (0,1,2)(0,1,0)



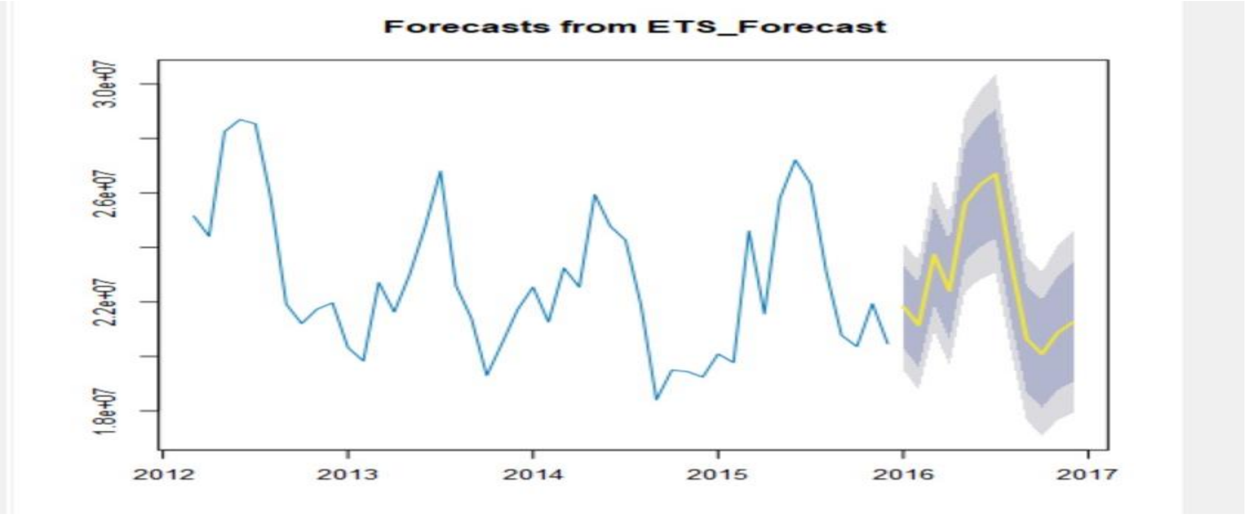
AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

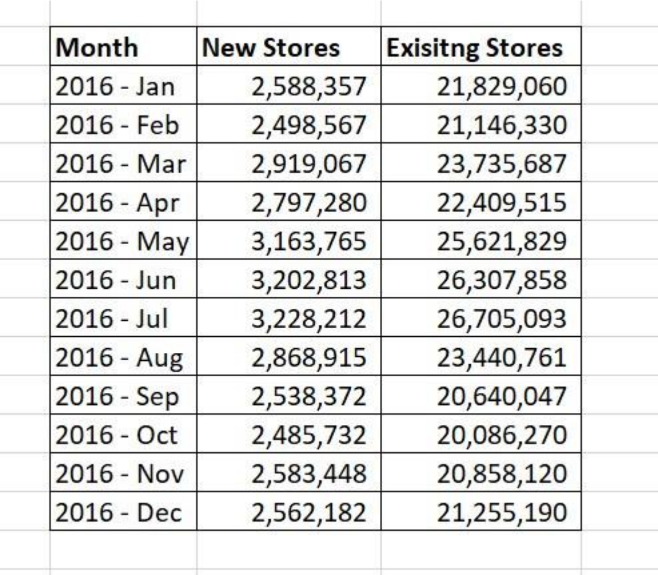
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961



Period	Sub_Period	ETS_forecast	ETS_forecast_high_95	ETS_forecast_high_80	ETS_forecast_low_80	ETS_forecast_low_95
2016	1	21829060.031666	24149899.115321	23346575.14138	20311544.921952	19508220.948011
2016	2	21146329.631982	23512577.365832	22693535.862148	19599123.401815	18780081.898131
2016	3	23735686.93879	26517865.796798	25554855.912929	21916517.964651	20953508.080782
2016	4	22409515.284474	25150243.401256	24201581.075733	20617449.493214	19668787.167691
2016	5	25621828.725097	28880596.484529	27752622.431914	23491035.018279	22363060.965665
2016	6	26307858.040046	29777680.067343	28576652.715009	24039063.365084	22838036.01275
2016	7	26705092.556349	30348682.320364	29087507.847195	24322677.265503	23061502.792334
2016	8	23440761.329527	26742106.733295	25599395.061562	21282127.597491	20139415.925758
2016	9	20640047.319971	23635033.372194	22598363.439189	18681731.200753	17645061.267747
2016	10	20086270.462075	23084199.797487	22046511.090727	18126029.833423	17088341.126662
2016	11	20858119.95754	24055437.105831	22948733.269445	18767506.645635	17660802.809249
2016	12	21255190.244976	24596988.126893	23440274.43075	19070106.059202	17913392.363058

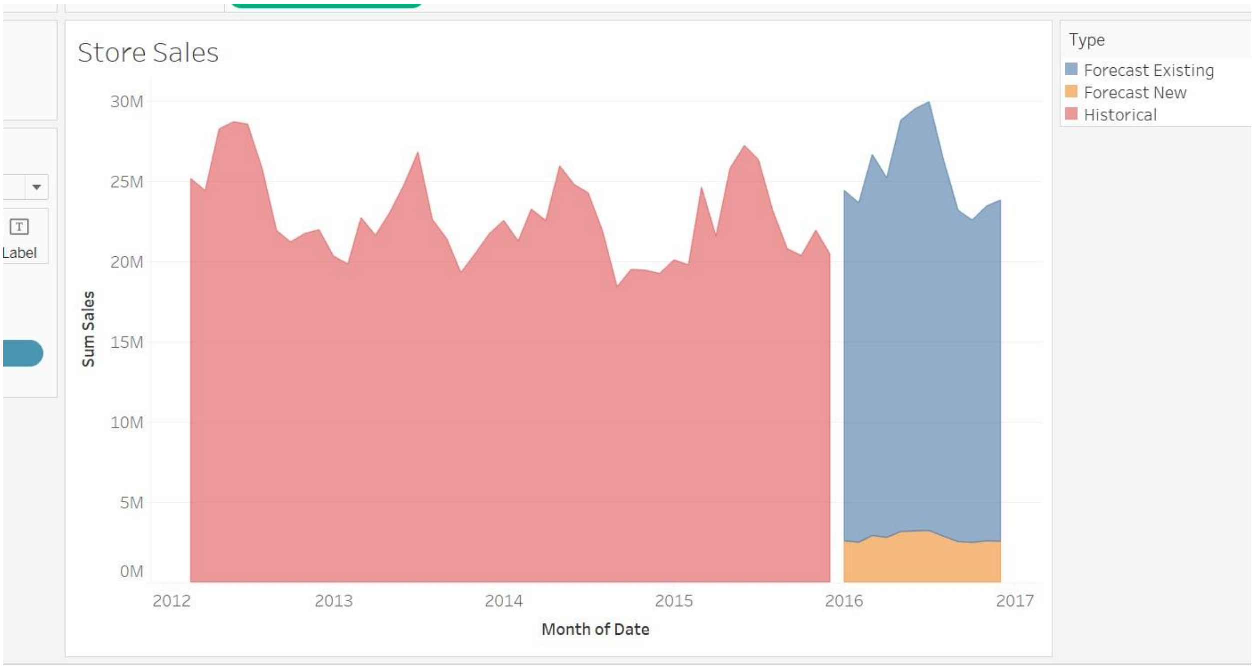


2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.



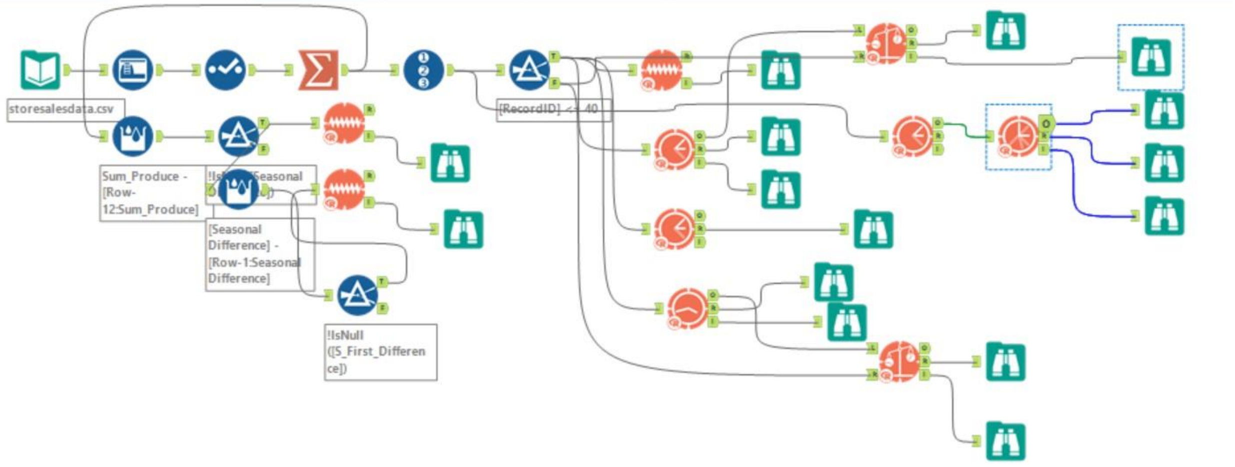
Month	New Stores	Exisitng Stores
2016 - Jan	2,588,357	21,829,060
2016 - Feb	2,498,567	21,146,330
2016 - Mar	2,919,067	23,735,687
2016 - Apr	2,797,280	22,409,515
2016 - May	3,163,765	25,621,829
2016 - Jun	3,202,813	26,307,858
2016 - Jul	3,228,212	26,705,093
2016 - Aug	2,868,915	23,440,761
2016 - Sep	2,538,372	20,640,047
2016 - Oct	2,485,732	20,086,270
2016 - Nov	2,583,448	20,858,120
2016 - Dec	2,562,182	21,255,190

Tableau Visualization of historical, existing and new store forecasts

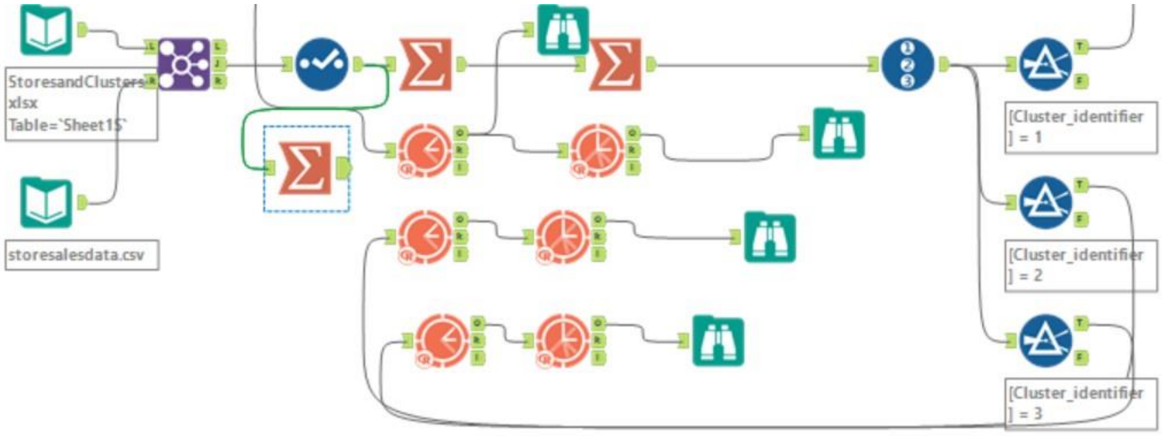


https://public.tableau.com/profile/farhana.hasan#!/vizhome/ETSForecast/StoreSales?publish=yes

Task 3 Workflow – 1


storesalesdata.csv
Sum_Produce - [Row-12:Sum_Produce]
!Is [illegible] Seasonal [illegible])
[Seasonal Difference] - [Row-1:Seasonal Difference]
!IsNull ([S_First_Difference])
[RecordID] <= 40

Task 3 Workflow 2


StoresandClusters.xlsx Table=`Sheet1$`
storesalesdata.csv
[Cluster_identifier] = 1
[Cluster_identifier] = 2
[Cluster_identifier] = 3