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?

Cluster Information:

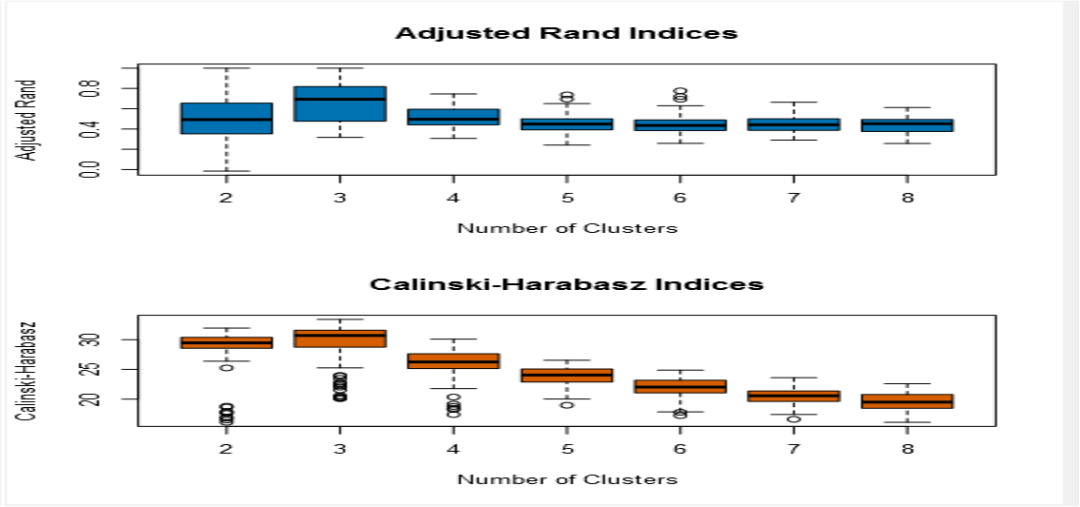
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.0152	0.3171	0.3072	0.2412	0.2586	0.2903	0.2568
1st Quartile	0.352	0.4819	0.4431	0.3943	0.3896	0.3877	0.377
Median	0.4926	0.6936	0.4964	0.4487	0.4348	0.4417	0.4526
Mean	0.484	0.6575	0.5125	0.4623	0.4532	0.4498	0.4411
3rd Quartile	0.655	0.816	0.5913	0.4982	0.489	0.4997	0.491
Maximum	1	1	0.7458	0.7366	0.7762	0.6637	0.6118

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	20.09	17.41	18.98	17.24	16.61	16.11
1st Quartile	28.61	28.76	25.16	22.91	21.05	19.61	18.46
Median	29.47	30.7	26.25	24.05	22.02	20.56	19.5
Mean	28.41	29.47	25.99	23.88	21.96	20.48	19.62
3rd Quartile	30.39	31.58	27.62	25.06	23.14	21.35	20.77
Maximum	31.95	33.41	30.09	26.53	24.87	23.6	22.59



From above Adjusted Rand Indices and Calinski-Harabasz Indices, Mean and Median values of 2, 3, 4 clusters are high Compare to other cluster values (5, 6, 7, 8)

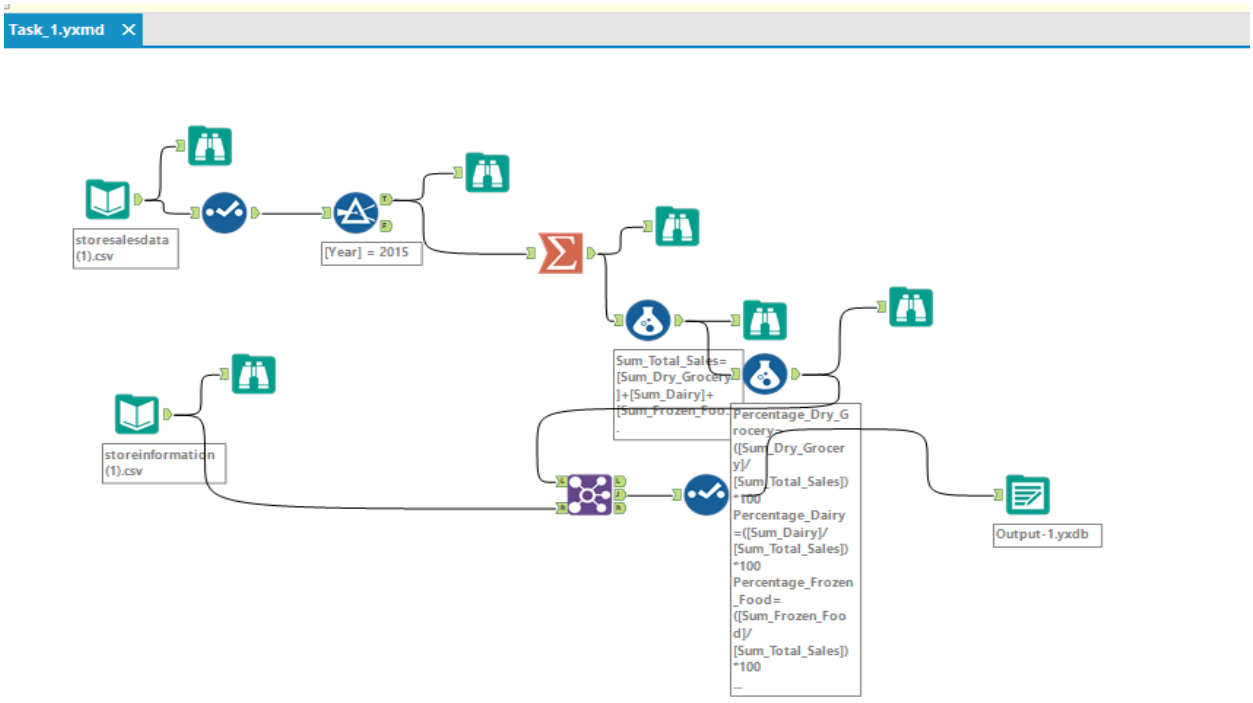
From the above justification the optimal numbers of Clusters are 3 and Stores format numbers which is falls under each cluster is shown below.

The Optimal number of store formats is 3, Under 3 formats stores are divided into

Cluster 1=23
Cluster 2=29
Cluster 3=33

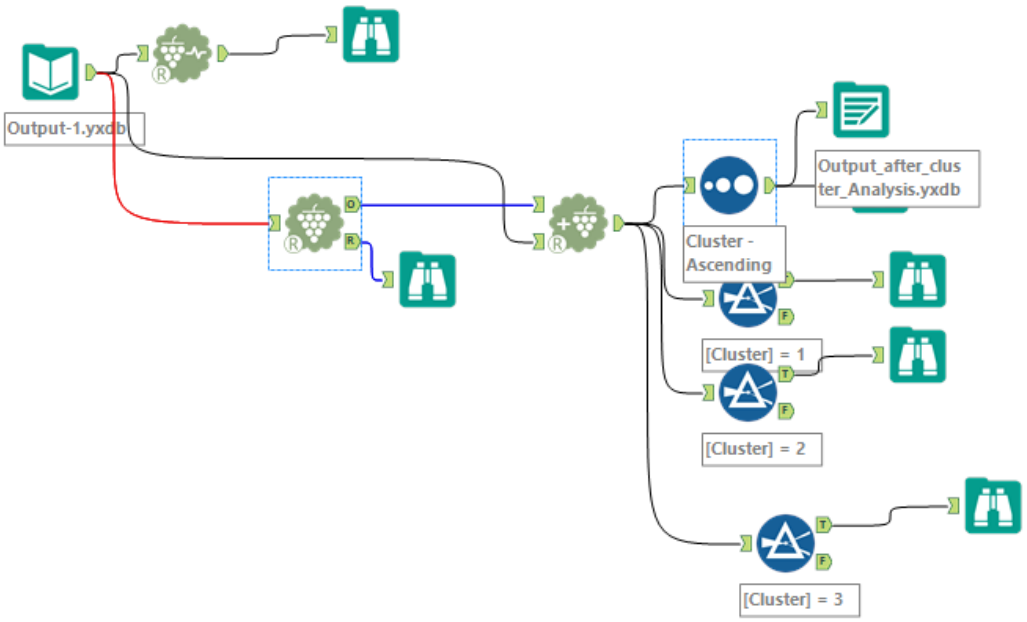
Above stores numbers under each cluster has calculated by adding all the sales of each category (Dry Grocery, Dairy, Frozen Food, Meat, Produce, Floral, Deli, and Bakery) for each store only for 2015, and then we have to calculate total sales by adding all category sales. Then we have to calculate Percentages of each category by dividing sum total sales of each category by Total sales of all categories.

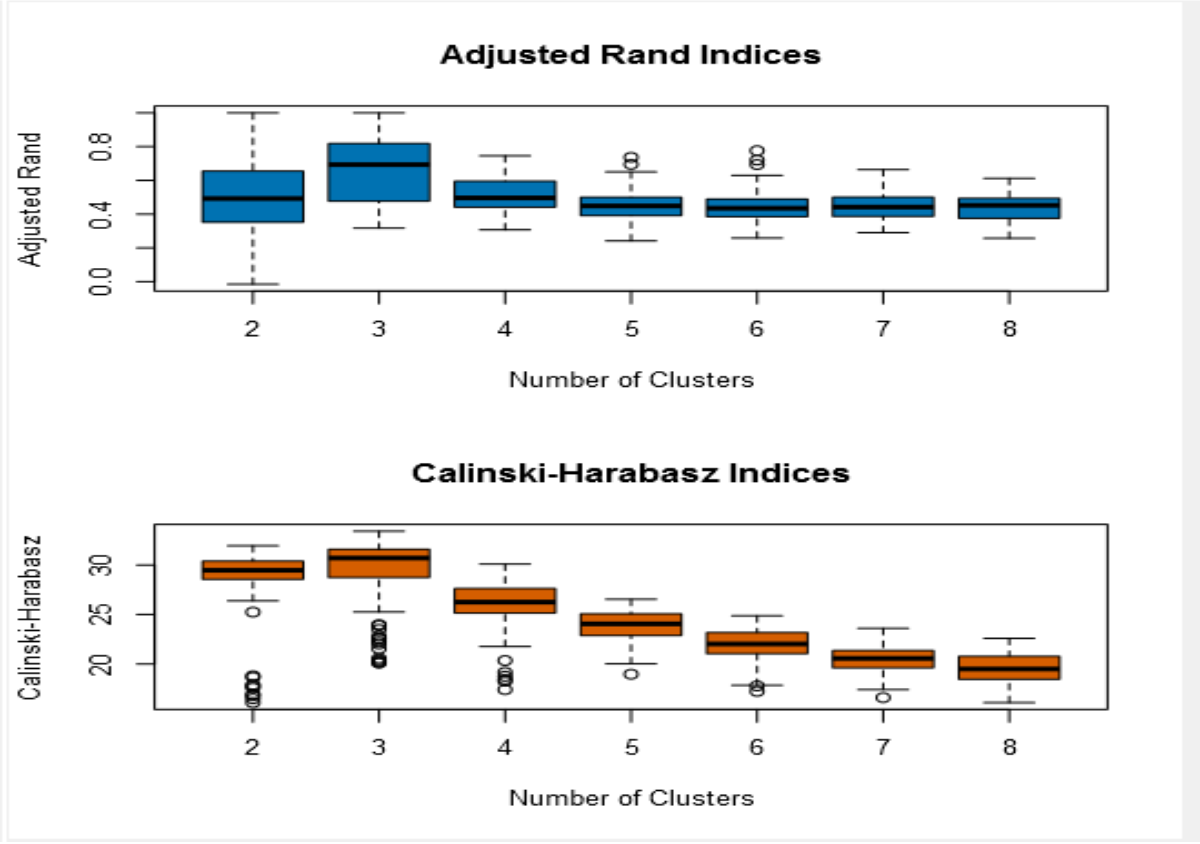
Total sales Calculated using Alteryx has shown below.



Then Using K-Centroids cluster analysis we have to calculate how many stores comes under each cluster here 3 cluster and K_Means clustering method is used as per the project requirement.

Store formats calculated using K-Centroids Cluster Analysis using Alteryx shown below.





Mean and Medians values of K Means clustering method is high compare to other clustering method (K-median and neural)

2. How many stores fall into each store format?

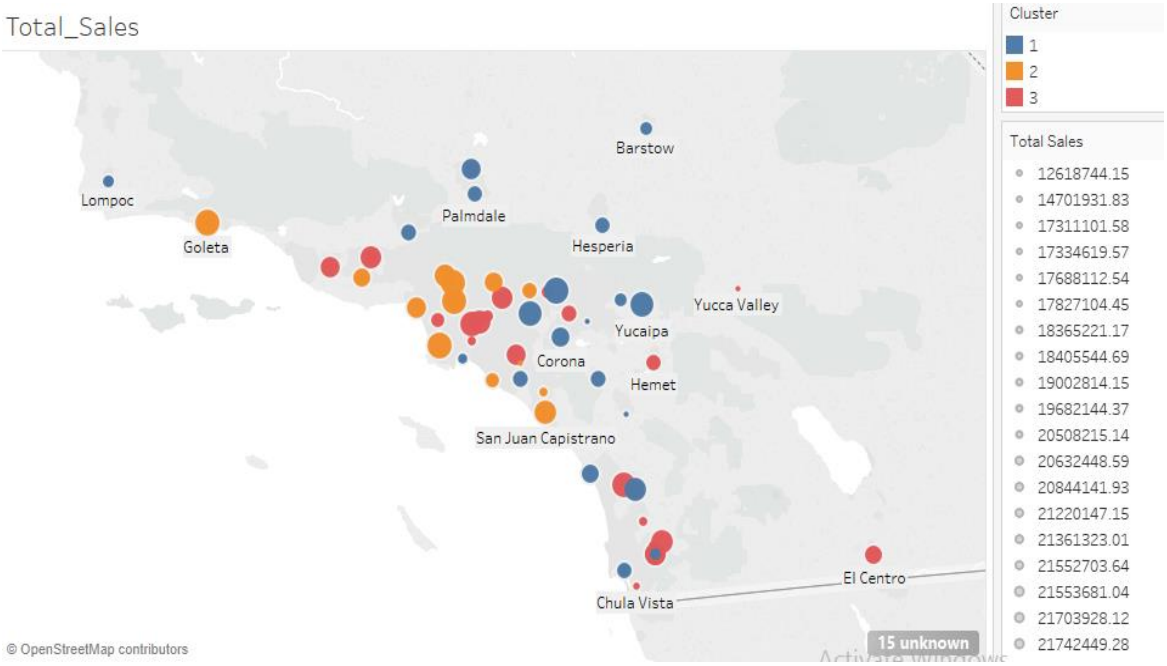
   Cluster 1=23
   Cluster 2=29
   Cluster 3=33

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

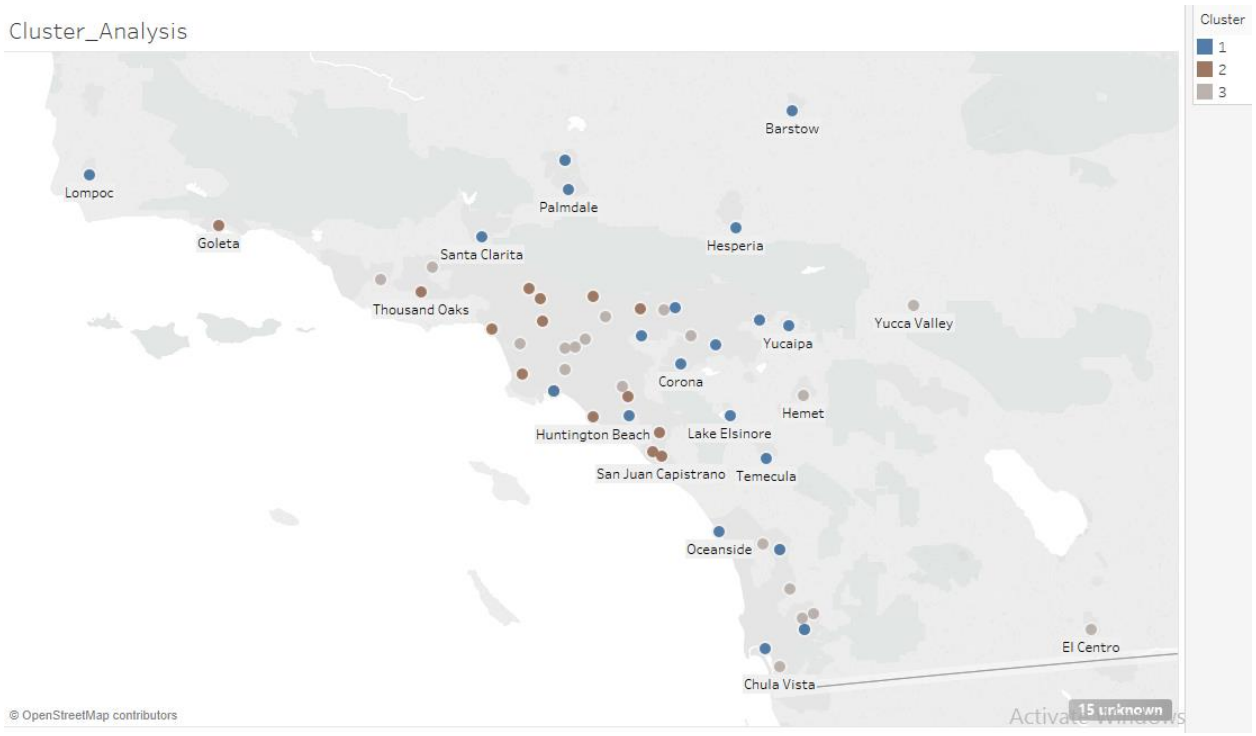
   All three clusters vary in their Total sales. Stores those have higher sales are grouped into cluster 3, Stores have moderate sales grouped into cluster 2 and stores have less sales are grouped into cluster1.

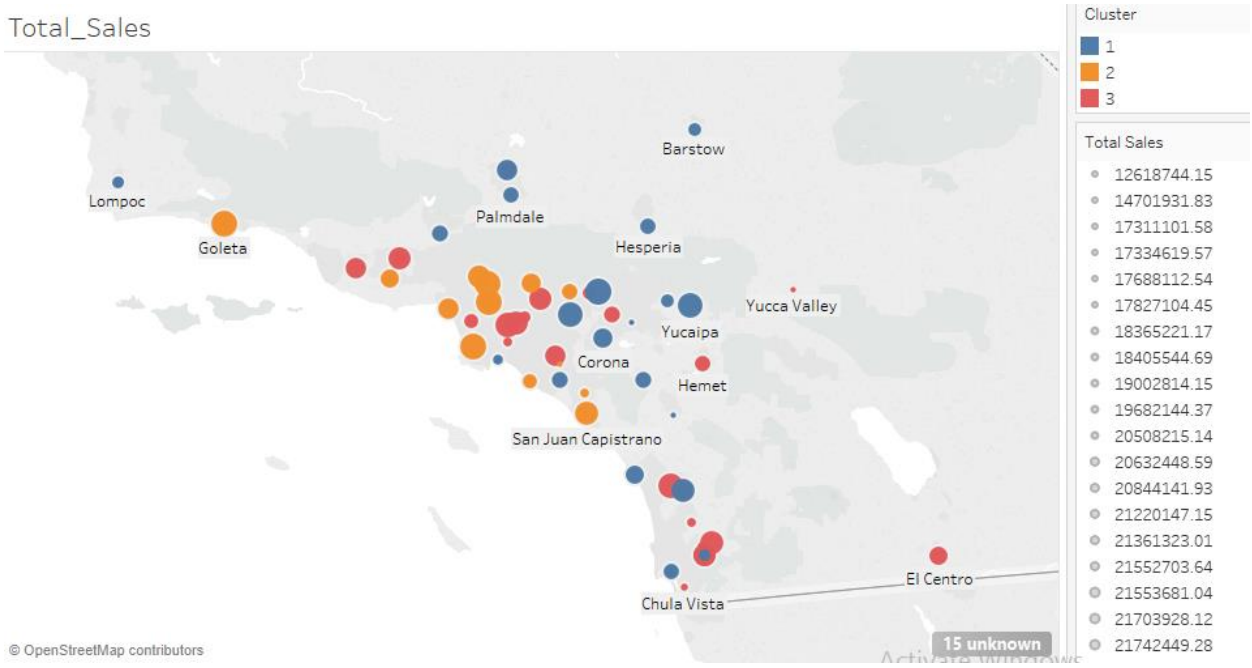
   So, we can say there are 23 stores have less sales, 29 have moderate sales and 33 Stores are having higher sales.

   To justify all three clusters are vary in Total sales, Visualization between cluster and Total-sales have shown below.

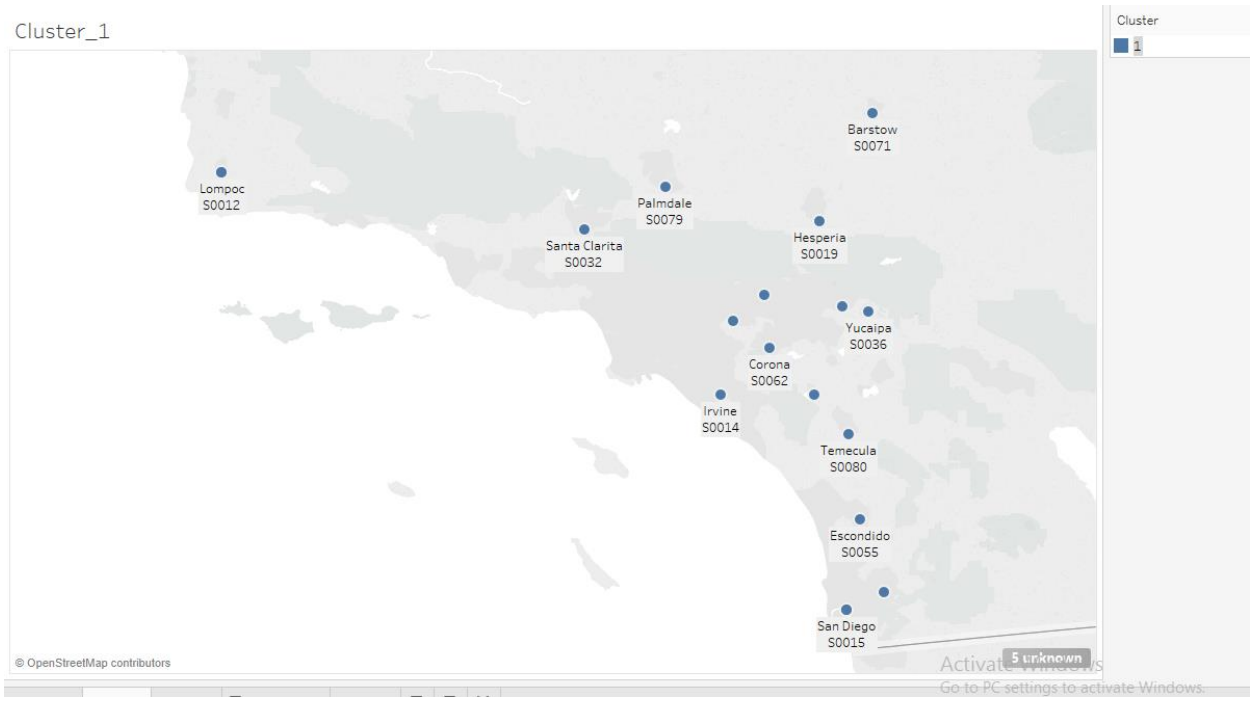


4. Please provide Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.

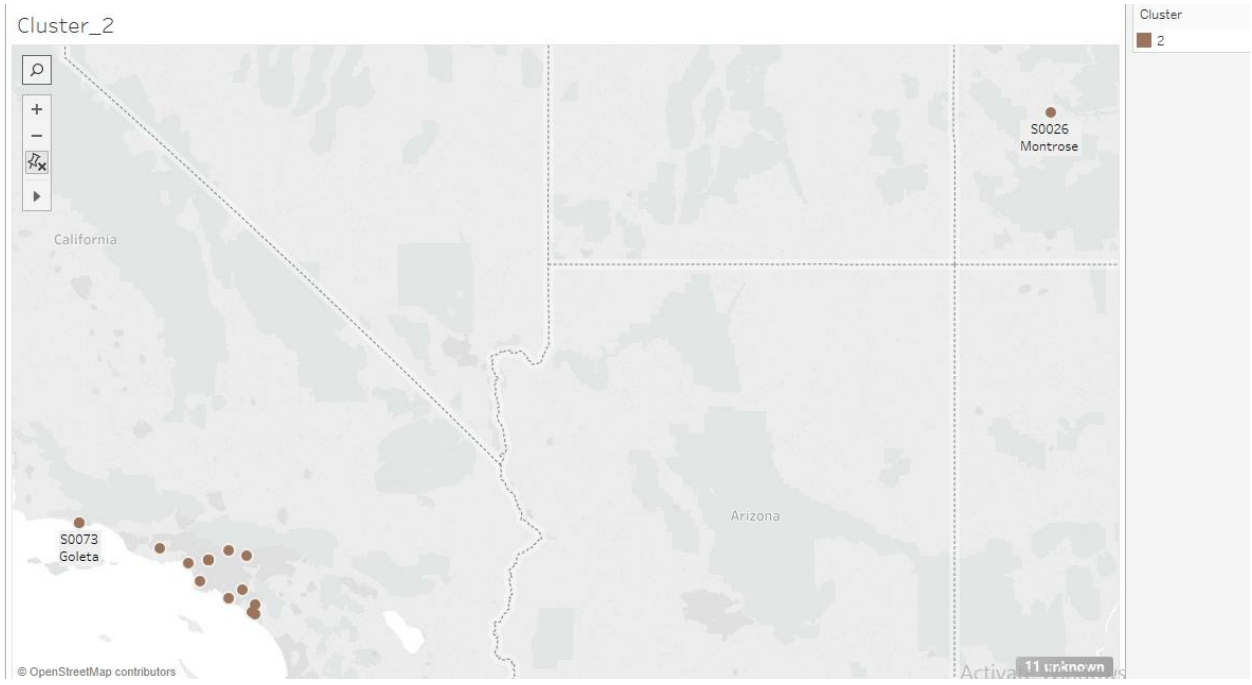




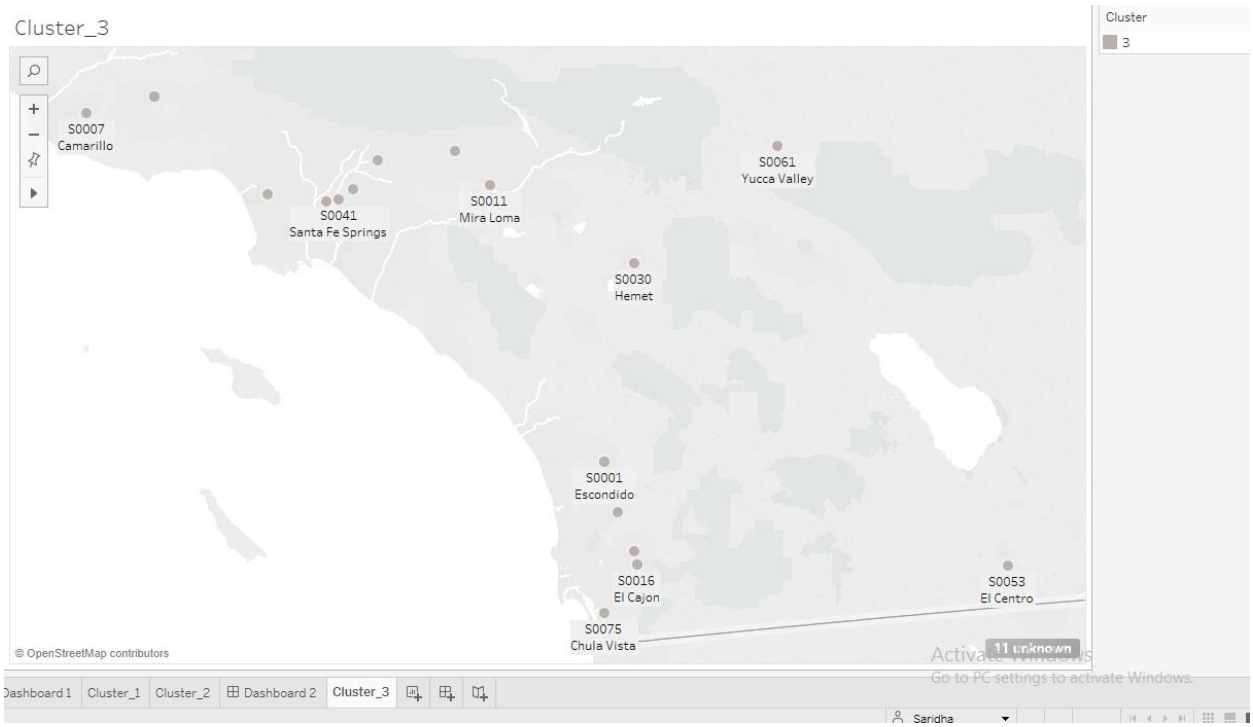
https://public.tableau.com/profile/saridha#!/vizhome/VisualizationforTotalSales/Total_Sales?publish=yes



https://public.tableau.com/profile/saridha#!/vizhome/VisualizationforTotalSales/Cluster_1?publish=yes



https://public.tableau.com/profile/saridha#!/vizhome/VisualizationforTotalSales/Cluster_2?publish=yes



https://public.tableau.com/profile/saridha#!/vizhome/VisualizationforTotalSales/Cluster_3?publish=yes

Task 2: Formats for New Stores

1) What methodology did you use to predict the best store format for the new stores?
I used Boosted Model to predict the best store format for the new stores.

Why did you choose that methodology?

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
BT_Model	0.8235	0.8543	0.8000	0.6667	1.0000
DT_Model	0.7059	0.7327	0.6000	0.6667	0.8333
FT_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.

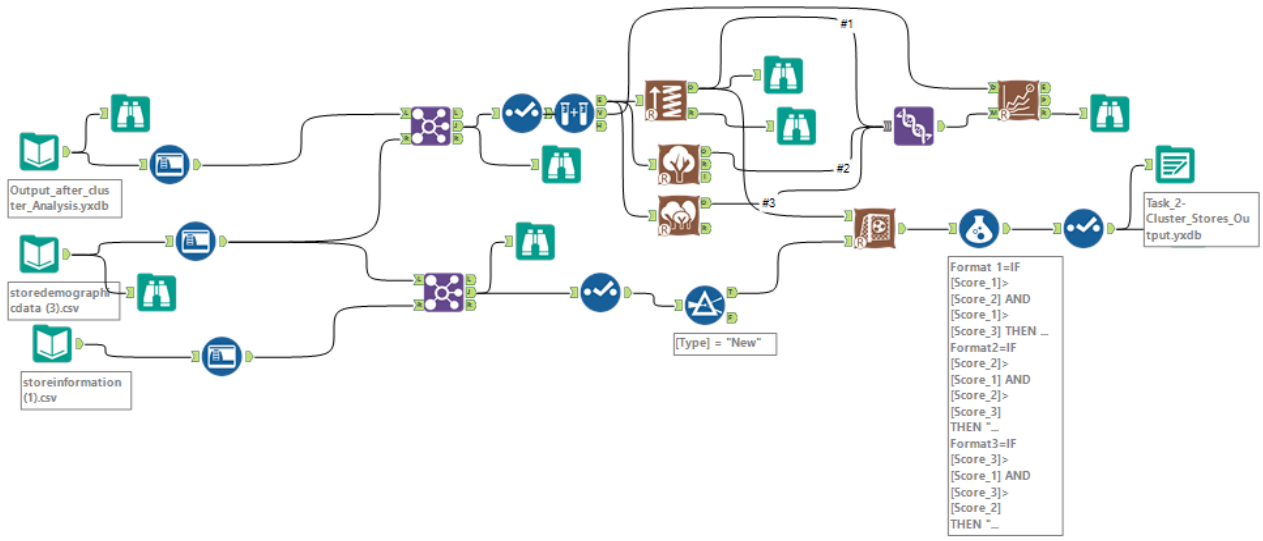
F1: F1 score, precision * recall / (precision + recall)

In this project three Classification models (Decision tree Model, Forest Model and Boosted Model) are used to predict store best format for new stores by setting validation 20% and random seed to 3 as per the project requirement. When we compare overall Accuracy of three models, Boosted Model Accuracy is high so, but Forest Model is same Accuracy as of Boosted

Model, so we have to take Model which have high values in Overall Accuracy and F-Score(test Accuracy) value.

Model	Overall Accuracy	F-Score
Decision Tree Model:	70%	73%
Forest Model	80%	82%
Boosted Model	82%	85%

From the above table, Boosted Model has high Overall Accuracy and F-Score value, chosen as the best Model to predict the best store format for the new stores



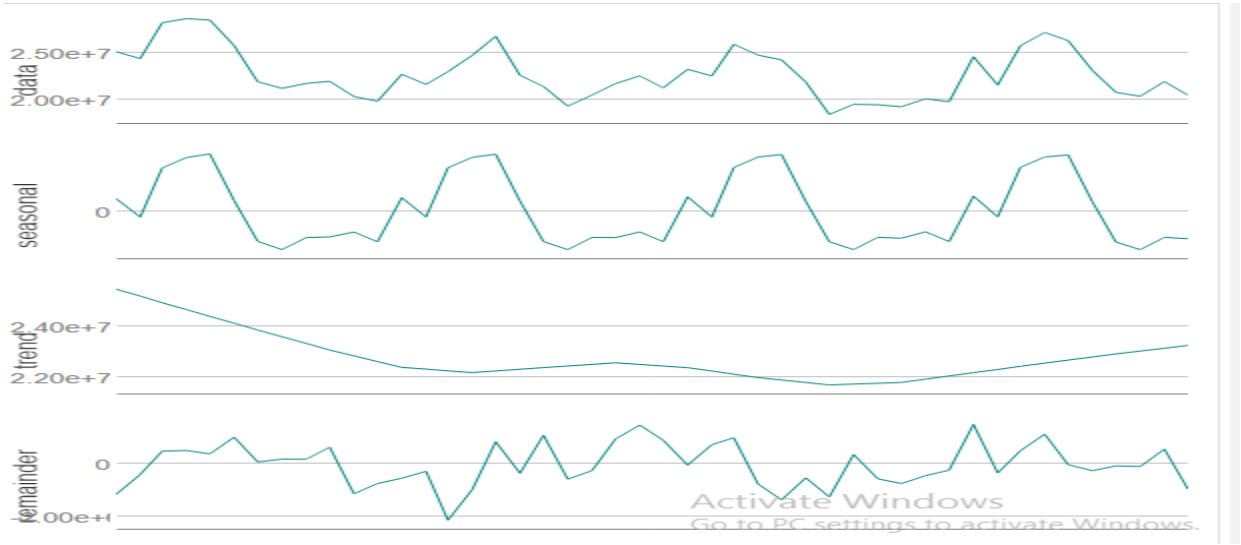
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

Task 3: Predicting Produce Sales

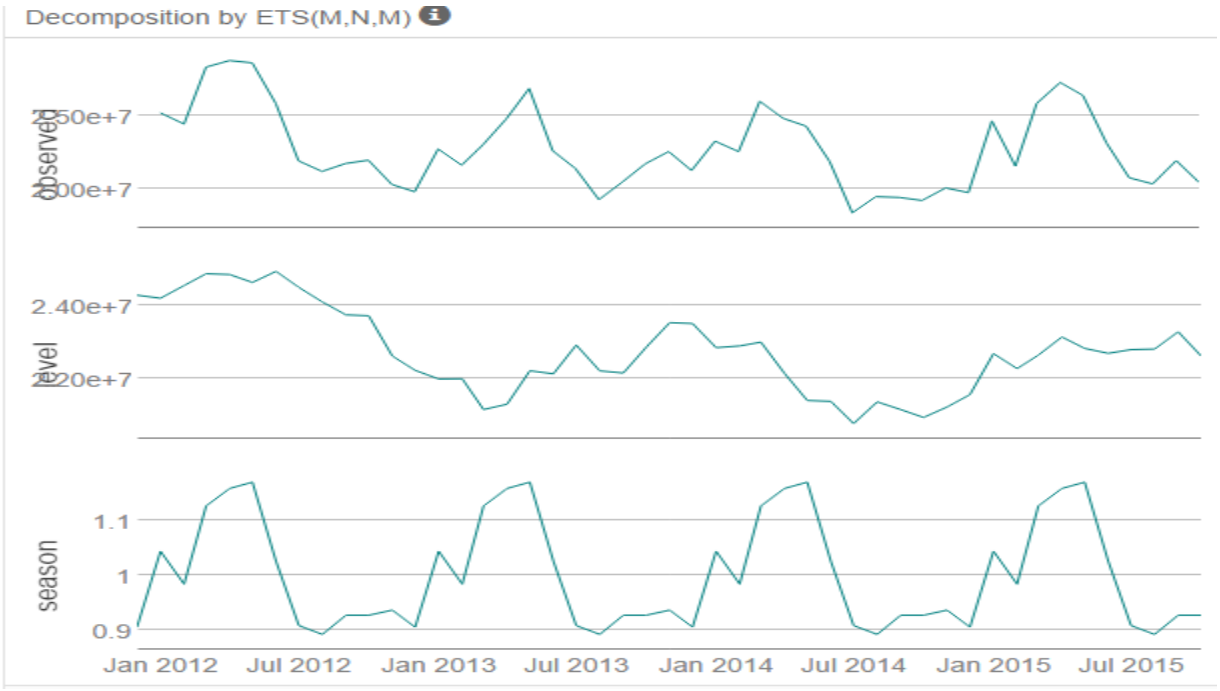
1. What type of ETS or ARIMA model did you use for each forecast?

Decomposition Plot:



Above shown is Time Series plot. By analyzing time series plot Error component is decreasing and increasing (M), Trend component is missing (N), and the seasonal component decreases slightly every year (M). Decomposition plot of ETS Model with (M, N, M) and in sample error measures and AIC values are shown below.

EST Model:



Method:
ETS(M,N,M)

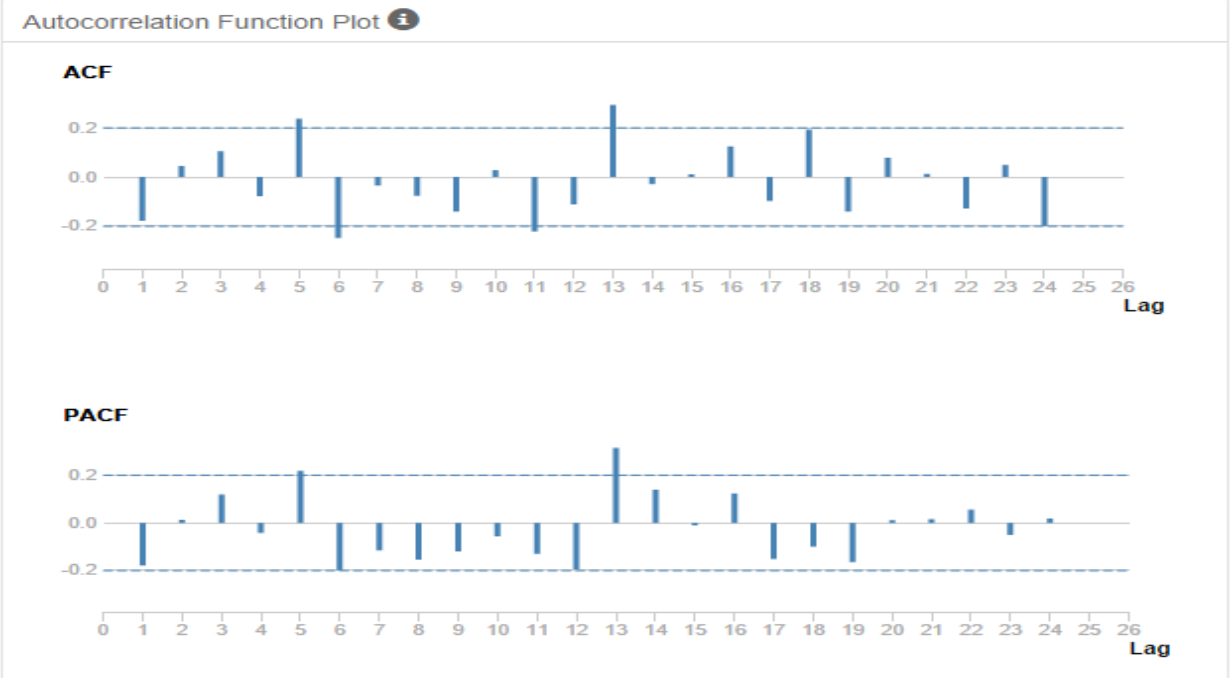
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-65107.9582069	917633.0913823	716577.8296631	-0.3978271	3.1667307	0.4031789	0.0308656

Information criteria:

AIC	AICc	BIC
1466.4904	1480.0387	1492.0913

ARIMA Model:
ACF and PACF plot with ARIMA Model (0,1,1) (0,1,0)12 has shown below.



Information Criteria:

AIC	AICc	BIC
1064.4413	1065.2413	1069.0204

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-154940.1852383	1093684.5811167	795865.729091	-0.7790733	3.560782	0.4477899	-0.1781792

When we compare both AIC, ETS Model AIC is higher than ARIMA Model AIC. Next, mean error, the average difference between Forecast and Actual vales is high in the ETS Model compare to ARIMA Model.

Mean percentage error i.e average percentage difference between actual and forecast value is high in the ETS Model than ARIMA Model.

EST Model after hold out Sample (Not included 2015):

Comparison of Time Series Models

Actual and Forecast Values:

Actual	MAM
20088529.29	19093401.03177
19772333.34	18946411.45986
24608406.71	18799421.88795
21559729.45	18652432.31604
25792074.59	18505442.74413
27212464.15	18358453.17222
26338477.15	18211463.60032
23130626.6	18064474.02841
20774415.93	17917484.4565
20359980.58	17770494.88459
21936906.81	17623505.31268
20462899.3	17476515.74077

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
MAM	4384779	5078248	4384779	18.3746	18.3746	2.8054	NA

ARIMA Model after hold out Sample (Not included 2015):

Report

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ARIMA_AR
20088529.29	20747633.03174
19772333.34	19465587.77174
24608406.71	21450343.27174
21559729.45	20745162.57174
25792074.59	24146221.40174
27212464.15	22985353.08174
26338477.15	22466292.24174
23130626.6	20083163.51174
20774415.93	16610438.23174
20359980.58	17700746.60174
21936906.81	17647927.82174
20462899.3	17443559.40174

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA_AR	2545368	2999243	2655219	11.0071	11.5539	1.6988	NA

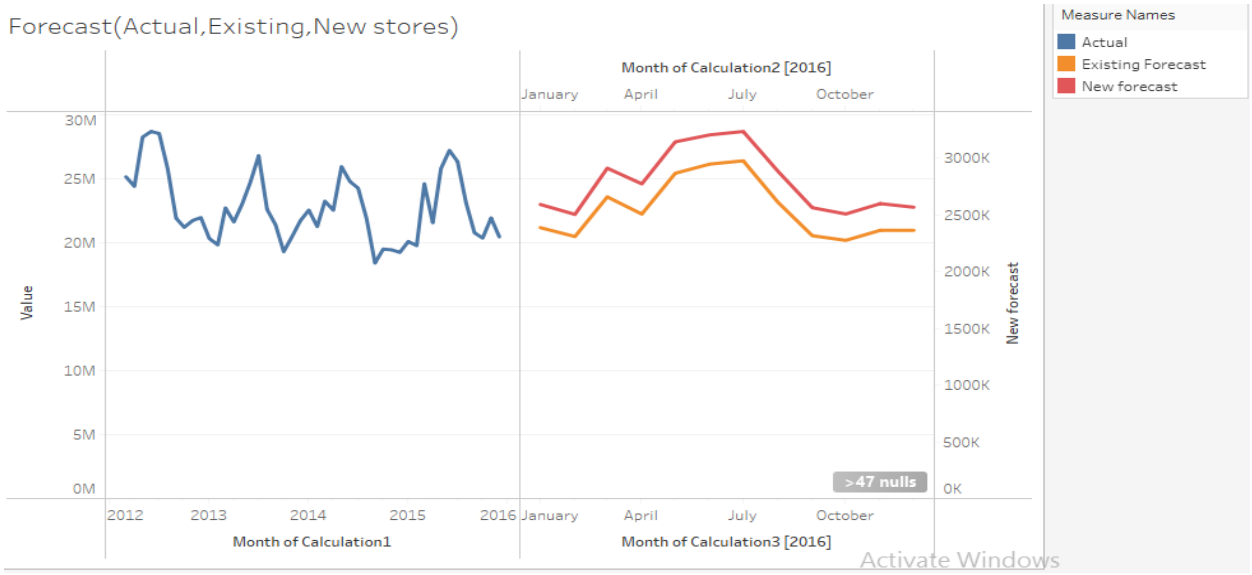
When we compare Mean Error (ME), Mean Percentage Error(MPE),Root Mean Squared Error (RMSE),Mean Absolute Error(MAE) of ETS Model all are have high values then ARIMA Model.

So from above justification ETS is the best model for each forecast.

A table with the correct 12 month forecasts for existing and new stores

Month	New Stores	Existing Stores
Jan-16	2590566.585695	21174989.40366
Feb-16	2503135.097223	20479354.577584
Mar-16	2910154.079513	23580340.680403
Apr-16	2772193.191798	22236546.234698
May-16	3142262.475899	25427255.457065
Jun-16	3203694.414631	26143967.404045
Jul-16	3233436.116192	26399993.267031
Aug-16	2882618.003153	23172393.880014
Sep-16	2562088.683447	20544268.638819
Oct-16	2506670.539657	20182471.085708
Nov-16	2598150.832184	20966876.352466
Dec-16	2566314.03563	20965097.001692

Visualization of Actual, Existing and New stores Forecast:



https://public.tableau.com/profile/saridha#!/vizhome/ForecastActualExistingNewstores/ForecastActualstoresExistingstoresNewstores?publish=yes