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?

Ans: The optimum number of store formats is **3.** I arrived this number by using K-Centroids Diagnostic Tool and K-Means method which produces the K-Means Cluster Assessment report. Based on the report, I arrived the above conclusion of 3 store formats.

Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.017586	0.160572	0.141156	0.194525	0.099825	0.219315	0.24855
1st Quartile	0.352613	0.33363	0.327799	0.301981	0.329073	0.316515	0.316829
Median	0.508815	0.482501	0.377323	0.381081	0.377999	0.359416	0.388908
Mean	0.496373	0.467738	0.404383	0.386771	0.390182	0.380264	0.378821
3rd Quartile	0.694097	0.574995	0.474807	0.465295	0.460644	0.437733	0.414644
Maximum	0.952939	0.792638	0.874682	0.62036	0.615218	0.6271	0.55165

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	10.47509	10.31461	12.07536	10.19514	9.606981	9.577281	9.271036
1st Quartile	18.7784	15.92972	14.2579	12.9405	12.218793	11.409148	11.140336
Median	20.10162	16.91185	15.11582	13.63886	12.778061	11.973964	11.642876
Mean	19.0993	16.64721	14.89573	13.63096	12.781839	12.137698	11.628092
3rd Quartile	20.87407	17.77524	15.74766	14.3606	13.559392	12.82982	12.227619
Maximum	22.41555	18.90096	16.93911	16.10526	15.308616	14.460895	14.074849

Fig 1: K-Means Cluster Assessment Report.

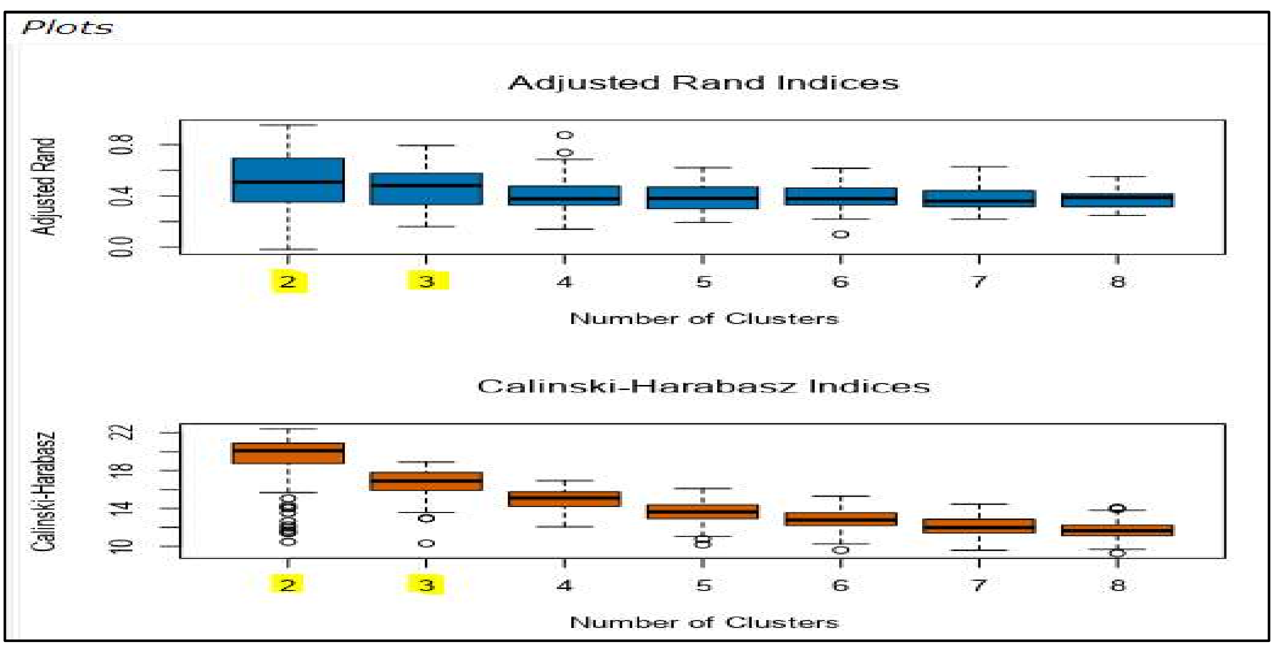


Fig 2: Adjusted Rand Indices & Calinski-Harabaz Indices Plots.

In the above K-Means Cluster Assessment report, Clusters 2 and 3 shows high median values in both Adjusted Rand Indices plot. But, the Calinskii-Harabasz Indices shows many outliers for cluster 2. Hence, we choose the Number of Clusters i.e. the **number of store formats as 3**.

2. How many stores fall into each store format?

Ans: To find the Number of Stores by each format, I used K-Centroid Cluster Analysis Tool, which produces the following report.

Report

Summary Report of the K-Means Clustering Solution K_Means

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Pct_Dry_Grocery + Pct_Dairy + Pct_Frozen_Food + Pct_Meat + Pct_Produce + Pct_Floral + Pct_Deli + Pct_Bakery + Pct_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823872	2.191565
2	35	2.475018	4.412367	1.947297
3	25	2.289004	3.585931	1.72574

Convergence after 8 iterations.
Sum of within cluster distances: 196.35035.

	Pct_Dry_Grocery	Pct_Dairy	Pct_Frozen_Food	Pct_Meat	Pct_Produce	Pct_Floral	Pct_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-0.663872	0.824834
2	-0.594802	0.655893	0.435128	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702371	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	Pct_Bakery	Pct_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Fig 3: Summary of K-Means Cluster Analysis.

Cluster 1 has **25** stores, Cluster 2 has **35** stores and Cluster 3 has **25** stores.

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

Ans: Cluster 3 has maximum average distance 2.289004 when compared to Cluster 1 (average distance: 2.099985) and Cluster 2 (average distance: 2.475018) which shows Cluster 1 is more compact and stable than the other 2 clusters.

Cluster 1 has maximum distance of 4.823872 compared with cluster 2 (Max distance: 4.412367) and Cluster 3 (Max distance: 3.585931). Also, Cluster 1 has the Separation of 2.191565, which is higher than the other two, separated more from the other 2 clusters.

The Stores that fall under Cluster 1 need an increase in Total sales as compared to the stores in other clusters.

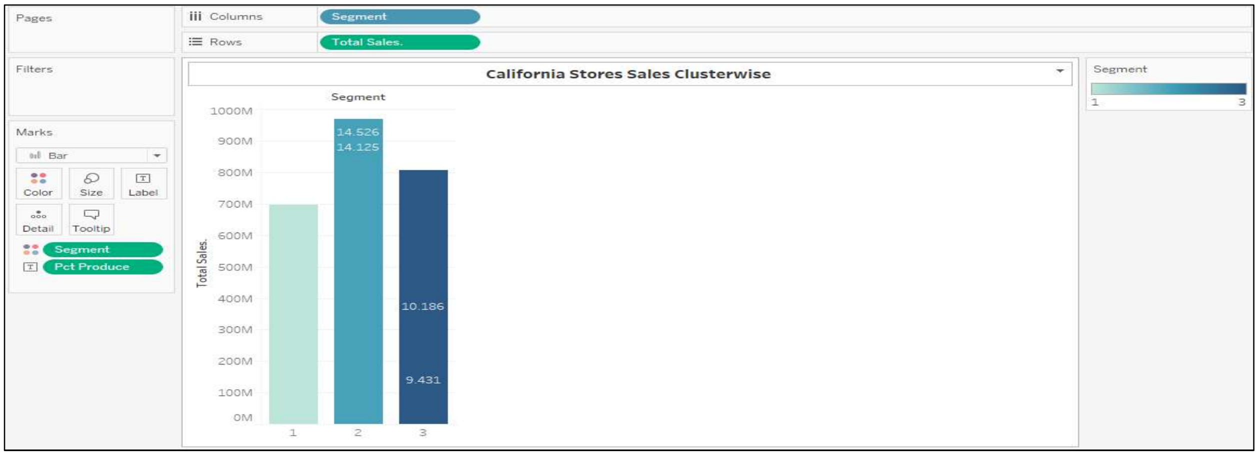


Fig 4: California Stores Total Sales – Cluster wise.

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.

Ans: The California Grocery Stores Clusters are visualized in Tableau. The following is the sheet generated.

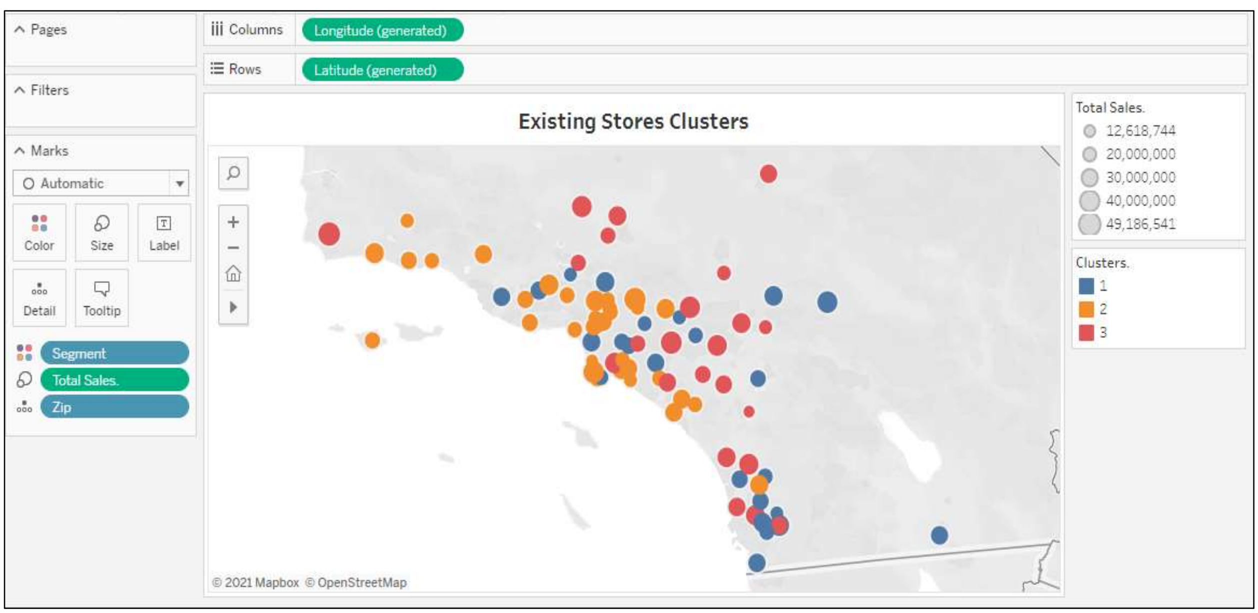


Fig 5: Tableau Visualization of Existing Stores in California as Clusters.

Tableau Visualization Public Link:

https://public.tableau.com/app/profile/umadevi7726/viz/Task1ExistingClustersUmaDevi/Sheet1

The Alteryx Workflow for Task 1 is as follows:

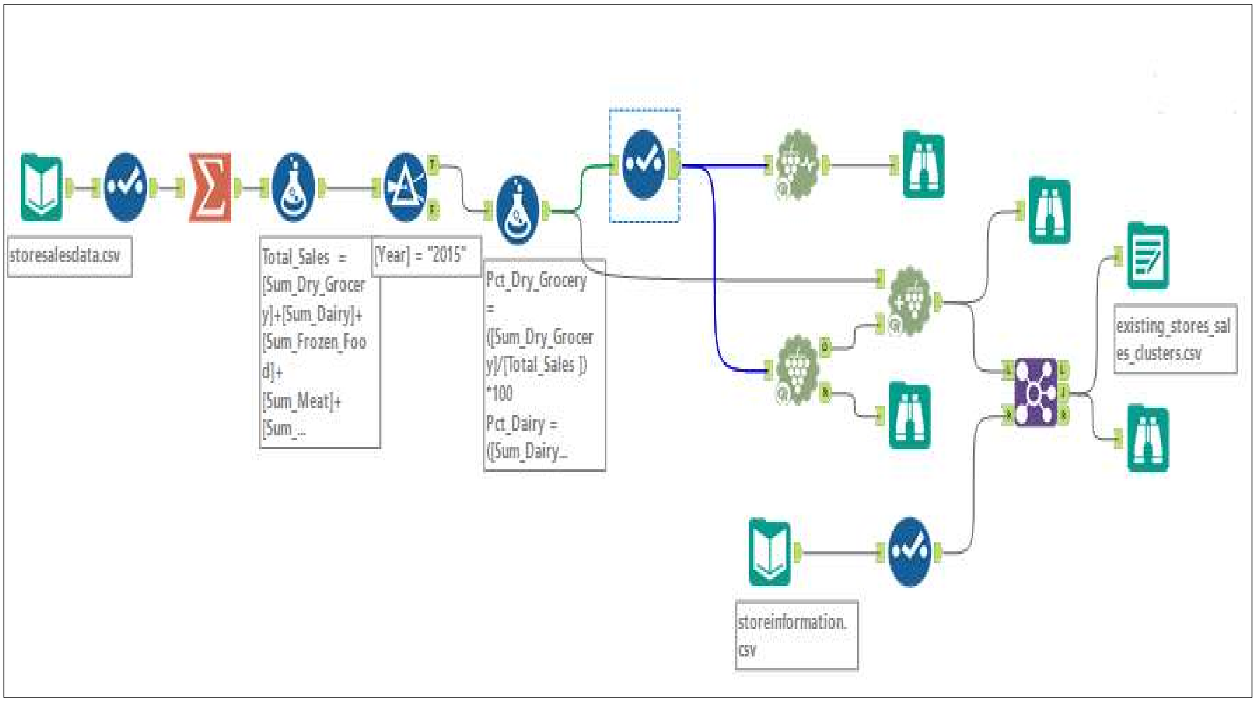


Fig 6: Alteryx Workflow: Existing Stores in California – Clusters.

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.)

Ans: Initially I used 3 models: Decision Tree Model, Forest Model and Boosted Model to predict which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.

I compared the accuracies of all the above 3 models using the Model Comparison Tool. The output of the Model Comparison Tool is shown below. From the report I chose **Boosted Model,** because it has the **highest F1 Accuracy score of 0.8333** over the other two models, **highest average accuracy rate of 0.7467** than the other 2.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7083	0.6250	1.0000	0.5000
FM	0.7059	0.7500	0.5000	1.0000	0.7500
BM	0.7647	0.8333	0.5000	1.0000	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] 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 BM

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	0
Predicted_2	2	5	0
Predicted_3	2	0	4

Confusion matrix of DT

	Actual_1	Actual_2	Actual_3
Predicted_1	5	0	2
Predicted_2	2	5	0
Predicted_3	1	0	2

Confusion matrix of FM

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	2	5	0
Predicted_3	2	0	3

Fig 7: Model Comparison Report

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

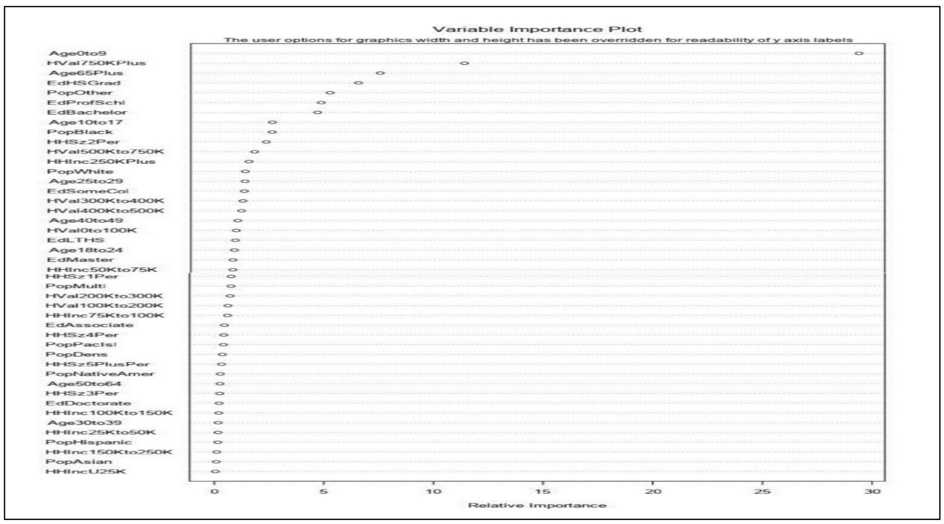
Ans: Below is the Boosted Model Report.

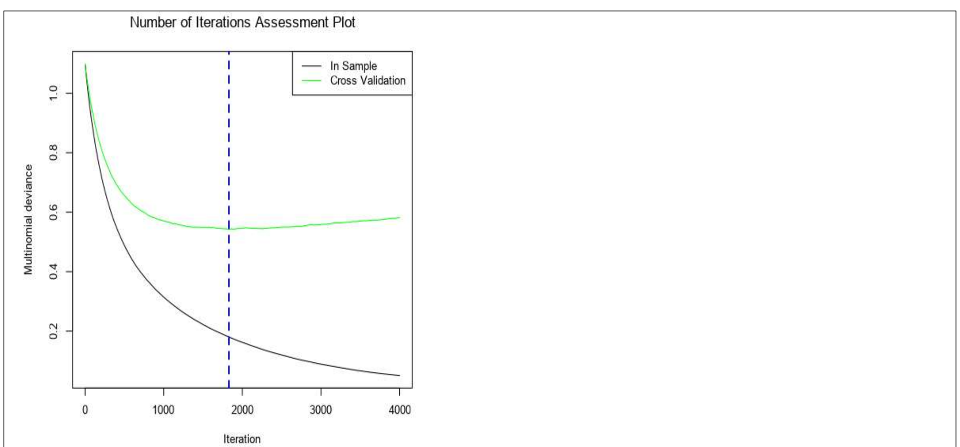
Report

Report for Boosted Model BM

Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1829





The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specfied assessment criteria (cross validation, the use of a test sample, or out-of-bag prediction).

Fig 8: Boosted Model Reports.

From the above report, the important variables are: Age0to9, H750KPlus, Age65Plus & EdHSGrad.

3. 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	3
S0092	2
S0093	3
S0094	2
S0095	2

The Alteryx Workflow for Task 2 is given below:

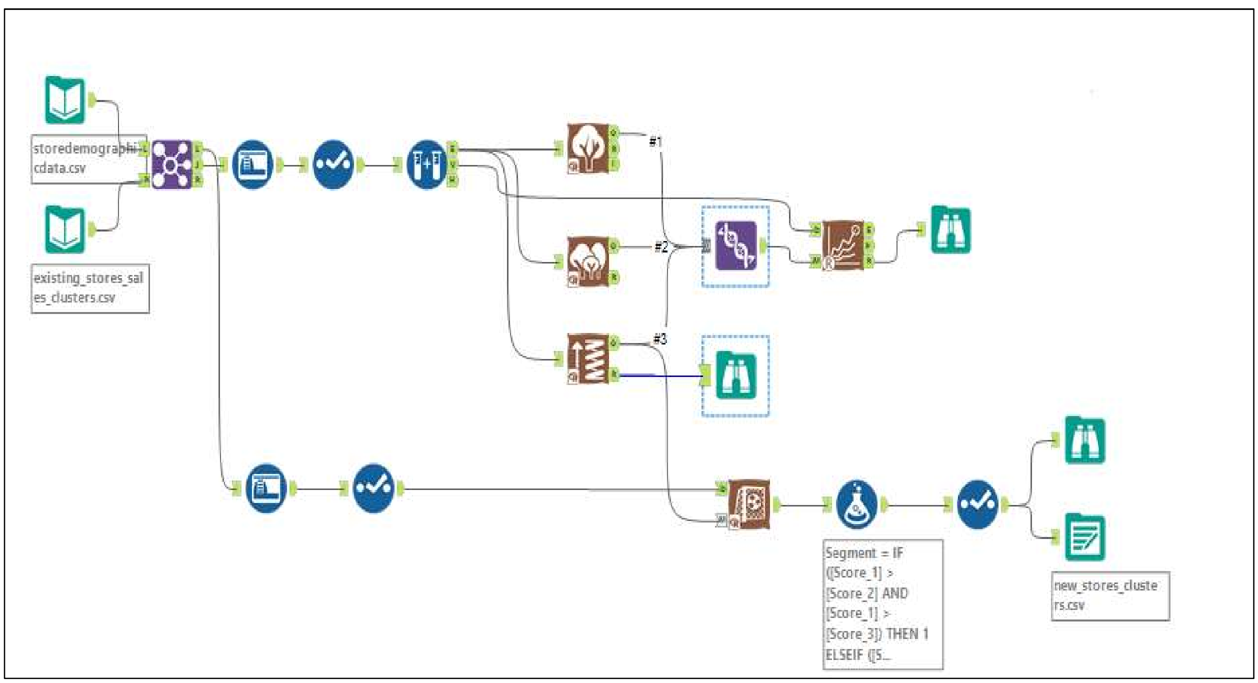


Fig 9: Alteryx Workflow: New Stores – Clusters.

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?

Ans: We have to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. We are going to use a 6 month holdout sample for the TS Compare Tool because we do not have that much data so using a 12 month holdout would remove too much of the data.

Steps followed:

1. Load storesalesdata.csv file on to the canvas, using Select Tool changed the datatypes because since it is a csv file, all the columns were string by default.
2. Then using Summarize Tool, group by Year then group by Month and calculated Sum_Produce.
3. Using TS Plot Tool, a Time Series Decomposition plot was generated for Sum_Produce with Target Field Frequency as Monthly, Year the series starts as 2012 and the quarter of series starts as 1 as settings. Below is the output generated by TS Compare Tool.

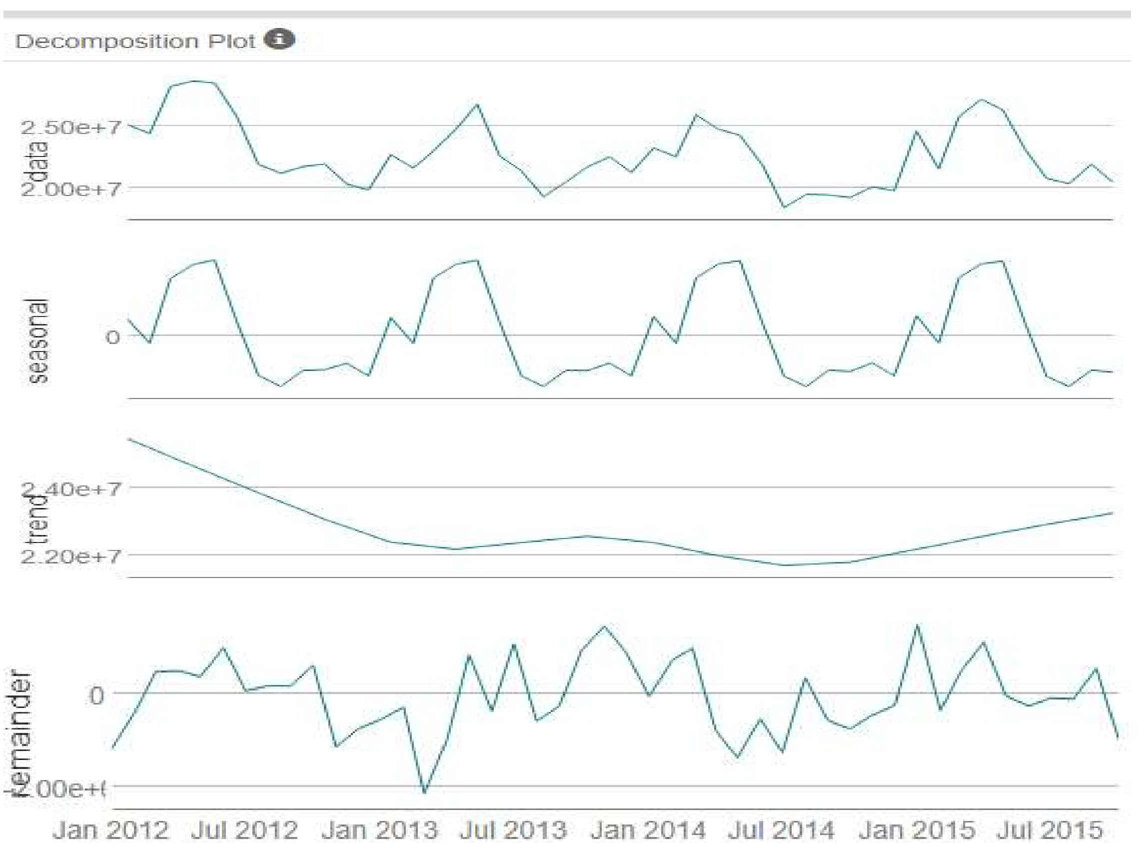


Fig 10: Decomposition Plot

4. The Decomposition plot clearly depicts that there is **NO Trend, the Seasonality is Multiplicative and the Error is also Multiplicative. So, I chose ETS (M, N, M) model** to forecast sales.

5. From the 46 records created by Summarize Tool, first 40 records were filtered and the remaining 6 records were holdout samples.

6. ETS model and ARIMA model were trained with these 40 records, and validated with the 6 holdout samples. Their output were compared. For both the models, the series starting period is 2012, quarter of series start as 3, number of periods to include in the forecast is 12.

Summary of Time Series Exponential Smoothing Model ETS

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

Smoothing parameters:

Parameter	Value
alpha	0.674884
gamma	0.000203

Actual and Forecast Values:

Actual	ETS
26338477.15	26860639.57444
23130626.6	23468254.49595
20774415.93	20668464.64495
20359980.58	20054544.07631
21936906.81	20752503.51996
20462899.3	21328386.80965

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

Fig 11: The output by ETS model.

Report

Summary of ARIMA Model ARIMA

Method: ARIMA(1,0,0)(1,1,0)[12]

Call:
Arima(Sum_Produce, order = c(1, 0, 0), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616

Actual and Forecast Values:

Actual	ARIMA
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

Fig 12: The output by ARIMA model.

7. From the above two reports it is clear that, **the ETS model performance is better than ARIMA in terms of Accuracy.**

- The **RMSE** value of **ETS** model is **1,042,209** whereas the RMSE value for **ARIMA** model is **1,050,239.**
- The **MASE** value of **ETS** model is **0.412** whereas the MASE value for **ARIMA** model is **0.546**.
- The **AIC** value of **ETS** model is **880** whereas the AIC value for ARIMA model is **1279**.

Hence, I chose the ETS Model for forecasting the Sales Value for the Existing stores & NEW stores.

The Alteryx workflow for the above steps is given below:

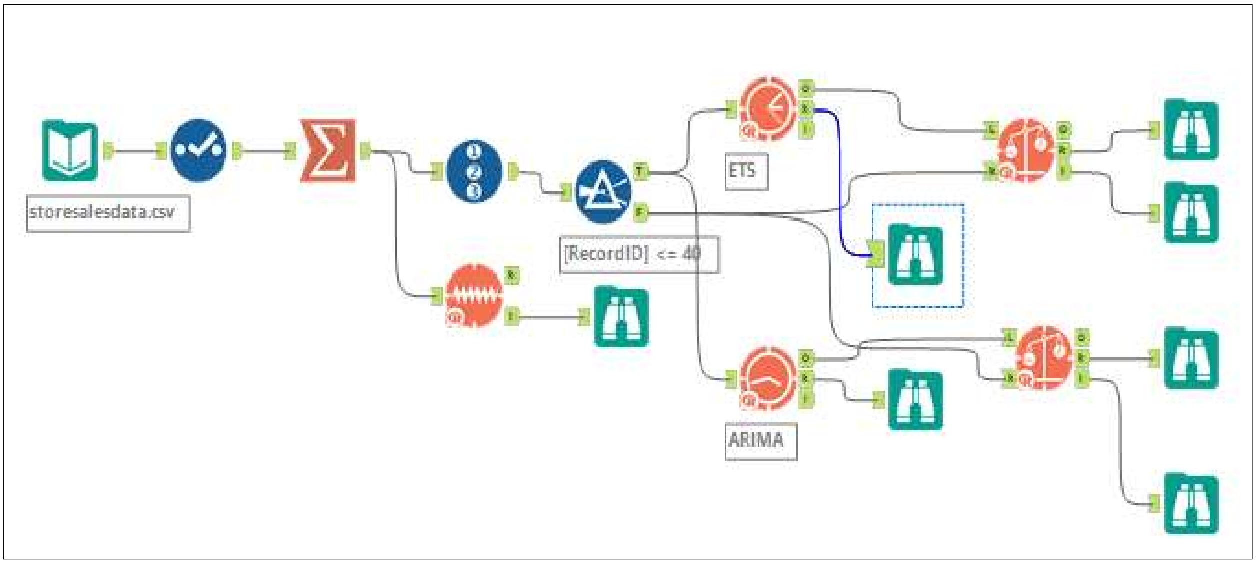
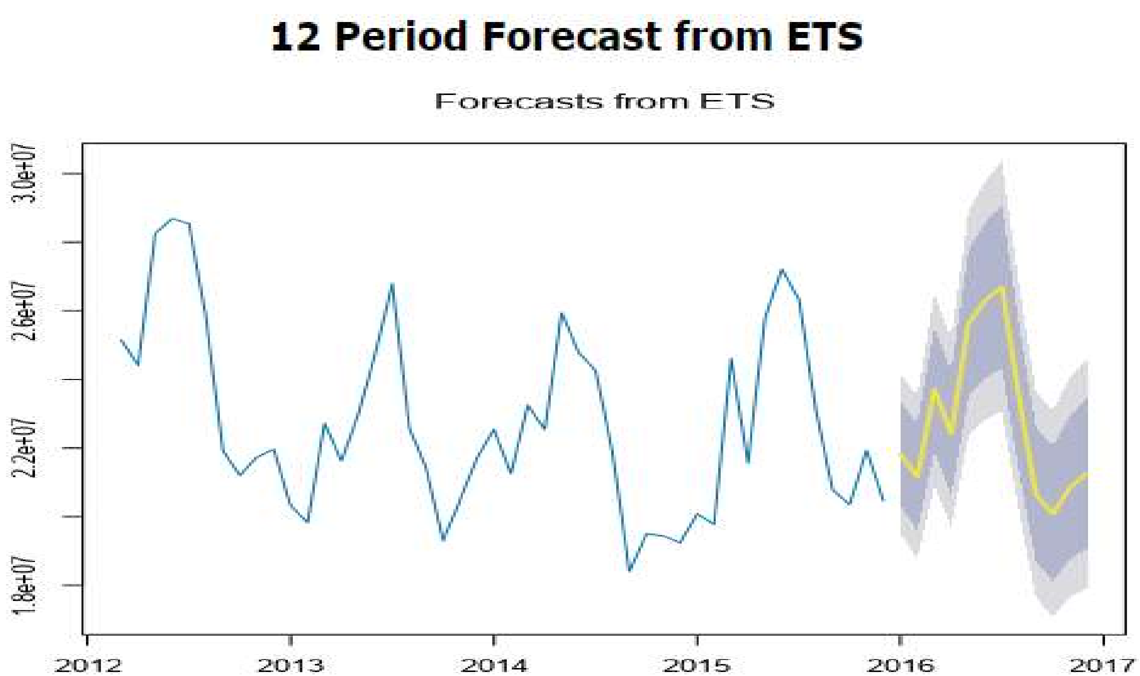


Fig 13: Alteryx Workflow for comparing performances of ETS & ARIMA models.

8. **The ETS model is used to forecast the Sales Value for the next 12 periods for the Existing Stores.** The output and the Alteryx workflow for the same is given below.



Period	Sub_Period	TS_Forecast	TS_Forecast_high_95	TS_Forecast_high_80	TS_Forecast_low_80	TS_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

Fig 14: ETS Model Forecast Report for Existing Stores.

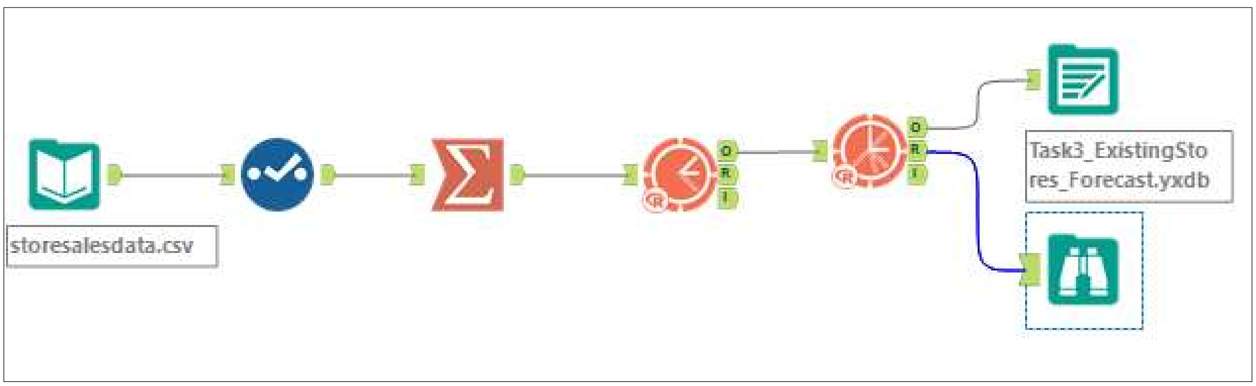


Fig 15: Alteryx Workflow for ETS Model Forecast for the Existing Stores.

9. Now we are going to predict the Sales Value for the New stores Cluster wise. For that, the **storessalesdata.csv** and **existing_stores_sales_clusters.csv** files are bringing on to the canvas using the Input Data Tool and joined together by Store field using Join Tool.

10. Using a Summarize Tool, grouping of data is done first by Store followed by Segment, Year, and Month. Then Sum_Produce column was produced by summing Produce column. Another Summarize Tool is used to group data by Segment, Year, Month and Avg_Sum_Produce was calculated.

11. We know that there are 3 Clusters (Segments). So 3 Filter Tools are used to separate the data into Cluster wise. Then 3 ETS Model tools are used to generate model for each segment of stores, and then 3 TS Forecast Tools are used to forecast the Sales Value for the next 12 periods for the New stores. The output is given below.

Record	Period	Sub_Period	Sum_TS_Forecast
1	2,016	1	2,563,357.910041
2	2,016	2	2,483,924.727562
3	2,016	3	2,910,944.145687
4	2,016	4	2,764,881.869697
5	2,016	5	3,141,305.867305
6	2,016	6	3,195,054.203804
7	2,016	7	3,212,390.95409
8	2,016	8	2,852,385.769198
9	2,016	9	2,521,697.18679
10	2,016	10	2,466,750.893696
11	2,016	11	2,557,744.587714
12	2,016	12	2,530,510.805133

Fig 16: TS_Forecast for the next 12 periods for the New Stores.

12. Now the first TS_Forecast is multiplied by 1 since 1 store at segment 1, the second TS_Forecast multiplied by 6 since 6 stores at segment 2 and the third one by 3 since 3 stores under segment 3.

13. All the 3 outputs from the 3 Formula Tools are joined using Union Tool, the using Summarize Tool, then grouped by Year (Period), Month (Sub_Period), then Sum_TS_Forecast is calculated. The Alteryx workflow for forecasting New stores Sales value is given below.

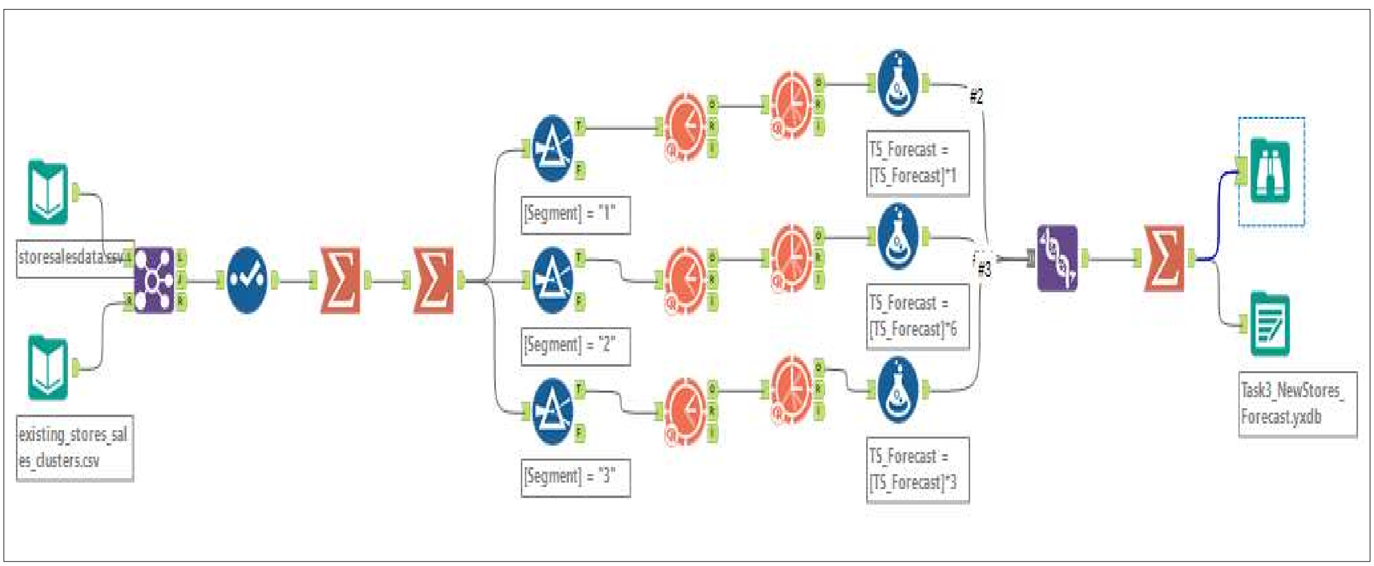


Fig 17: Alteryx Workflow to Forecast Sales Value for New stores for the next 12 periods.

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.

S.No	**Month**	**Existing Stores**	**New Stores**
1.	JAN 2016	21,829,060	2,563,358
2.	FEB 2016	21,146,330	2,483,925
3.	MAR 2016	23,735,687	2,910,944
4.	APR 2016	22,409,515	2,764,882
5.	MAY 2016	25,621,829	3,141,306
6.	JUNE 2016	26,307,858	3,195,054
7.	JULY 2016	26,705,093	3,212,391
8.	AUG 2016	23,440,761	2,852,386
9.	SEP 2016	20,640,047	2,521,697
10.	OCT 2016	20,086,270	2,466,751
11.	NOV 2016	20,858,120	2,557,745
12.	DEC 2016	21,255,190	2,530,511

Finally I did the Visualization of these forecasts that include historical data, existing stores forecasts and the new stores forecasts in **TABLEAU**.

Here is the link of my Tableau Public Visualizations:

https://public.tableau.com/app/profile/umadevi7726/viz/AllForecastsFinalUmaDevi/Sheet1

https://public.tableau.com/authoring/FinalDashboardTask3UmaDevi/Dashboard1#1

Tableau Visualization of Historical data, Existing and New Stores Forecasts.

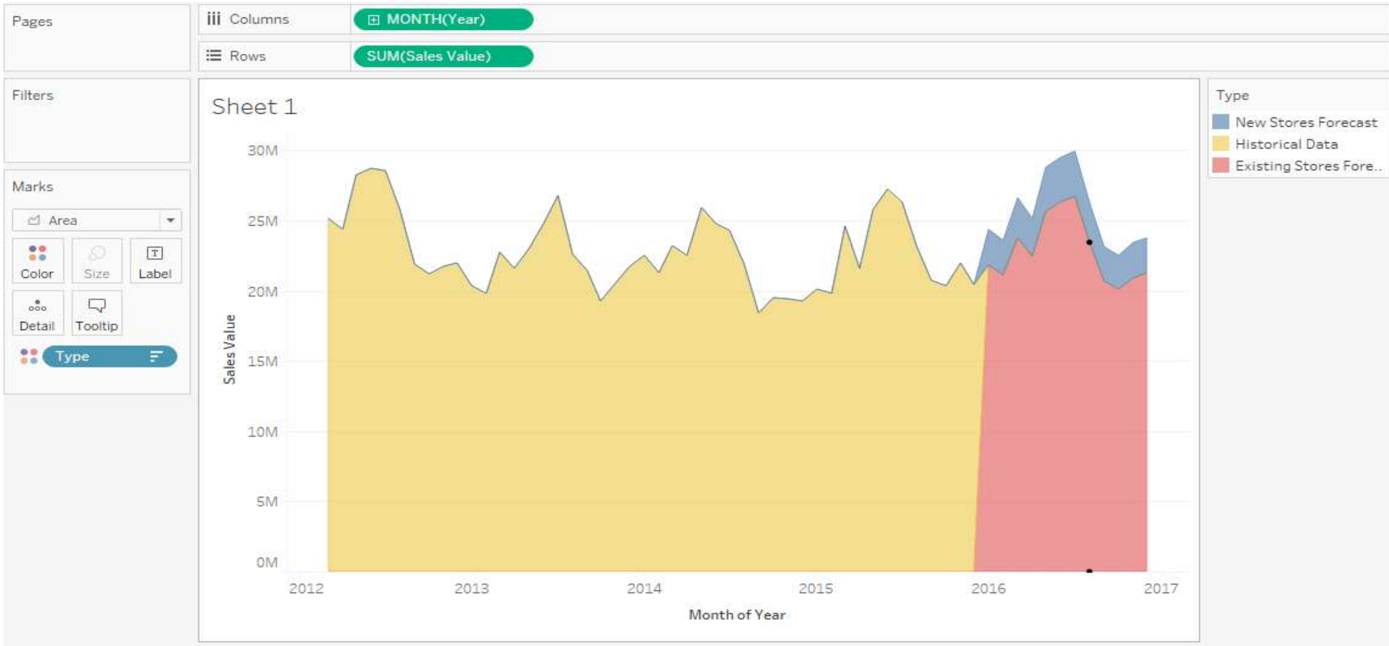
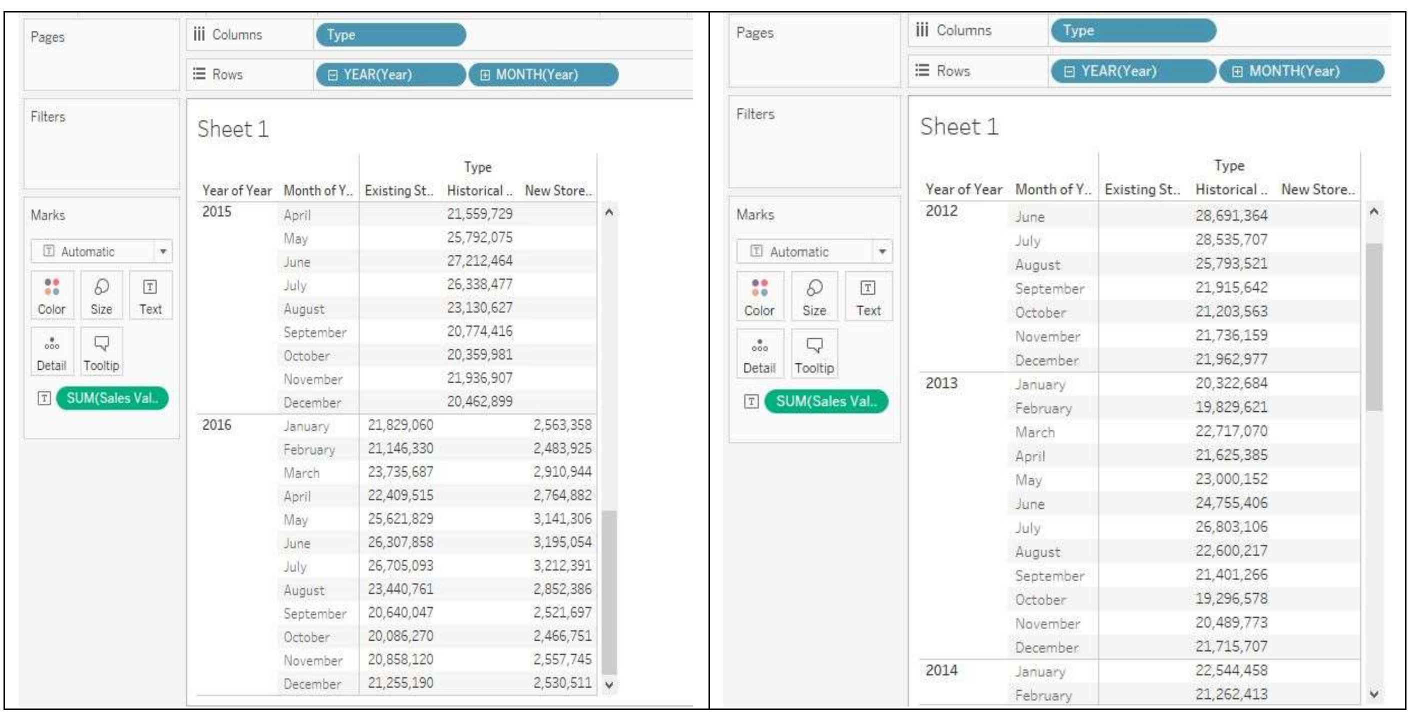


Fig 18: Tableau Visualization - Forecast for Historical data, existing Stores & New Stores.



Year of Year	Month of Y..	Existing St..	Historical ..	New Store..
2015	April		21,559,729	
	May		25,792,075	
	June		27,212,464	
	July		26,338,477	
	August		23,130,627	
	September		20,774,416	
	October		20,359,981	
	November		21,936,907	
	December		20,462,899	
2016	January	21,829,060		2,563,358
	February	21,146,330		2,483,925
	March	23,735,687		2,910,944
	April	22,409,515		2,764,882
	May	25,621,829		3,141,306
	June	26,307,858		3,195,054
	July	26,705,093		3,212,391
	August	23,440,761		2,852,386
	September	20,640,047		2,521,697
	October	20,086,270		2,466,751
	November	20,858,120		2,557,745
	December	21,255,190		2,530,511

Year of Year	Month of Y..	Existing St..	Historical ..	New Store..
2012	June		28,691,364	
	July		28,535,707	
	August		25,793,521	
	September		21,915,642	
	October		21,203,563	
	November		21,736,159	
	December		21,962,977	
2013	January		20,322,684	
	February		19,829,621	
	March		22,717,070	
	April		21,625,385	
	May		23,000,152	
	June		24,755,406	
	July		26,803,106	
	August		22,600,217	
	September		21,401,266	
	October		19,296,578	
	November		20,489,773	
	December		21,715,707	
2014	January		22,544,458	
	February		21,262,413	

Fig 19: Forecast of Sales of Historical data, Existing Stores & New Stores in Table Form.