Project: Predictive Analytics Capstone:
Combining Predictive Techniques

Task 1: Determine Store Formats for Existing Stores

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

The optimal number of store format is 3,

I arrived at this number after performing the cluster analysis and going through the K-Means cluster analysis report,

From the below table we can see that 3 clusters have the highest median values in AR indices and CH indices

Report

K-Means Cluster Assessment Report

Summary Statistics

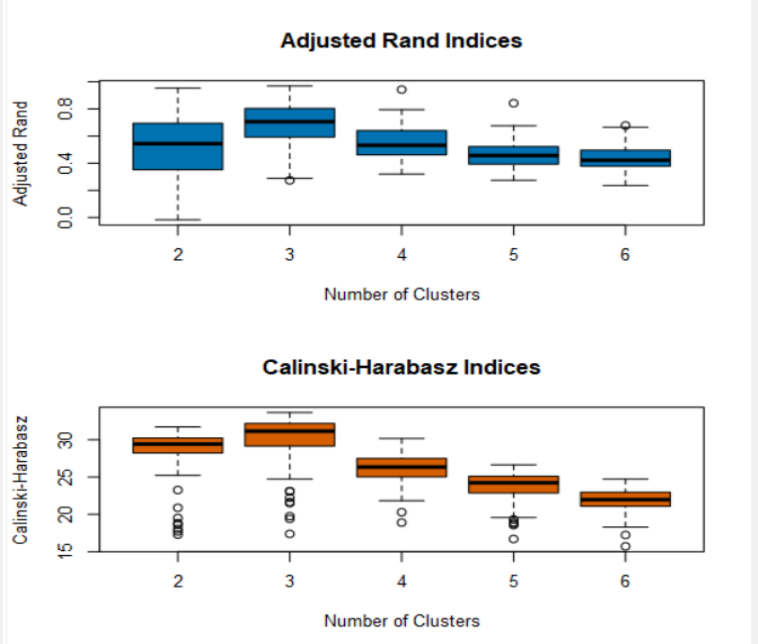
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.27351	0.31976	0.274316	0.235718
1st Quartile	0.35943	0.594017	0.46406	0.39294	0.377774
Median	0.544023	0.705326	0.53195	0.456588	0.421798
Mean	0.524263	0.69161	0.548167	0.470346	0.435429
3rd Quartile	0.694147	0.800179	0.635682	0.520656	0.493589
Maximum	0.952939	0.969034	0.942222	0.841981	0.677532

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.281	17.38103	18.89398	16.69676	15.71092
1st Quartile	28.22121	29.21236	25.03471	22.86498	21.10249
Median	29.4157	31.14178	26.33467	24.22188	21.96958
Mean	28.56936	30.07118	26.18037	23.72205	21.92474
3rd Quartile	30.21867	32.17467	27.4999	25.09459	22.95561
Maximum	31.71569	33.63781	30.1583	26.63063	24.72038

Plots



2. **How many stores fall into each store format?**
   The stores number of stores in each format is as below

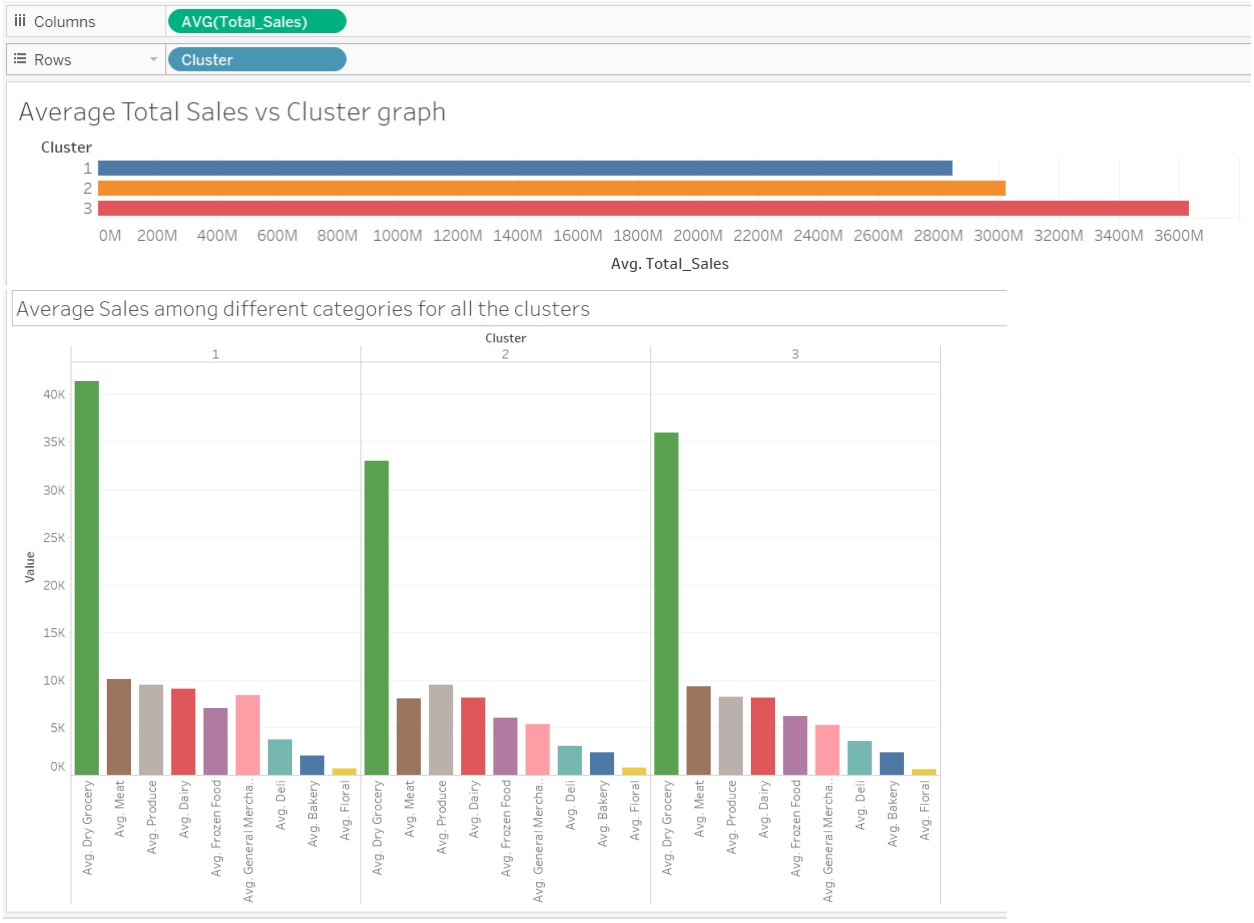
Cluster 1: 23
Cluster 2: 29
Cluster 3: 33

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

   The cluster differs from each other based on the average total sales and also has different average sales for dry grocery items.

   The below visualization gives the better understanding of the clustering model

https://public.tableau.com/profile/rakshit.kumar.satish#!/vizhome/Project8_Task1_Sheet1/Sheet1



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.**

The tableau visualization has been saved as a public file and is present in the below path

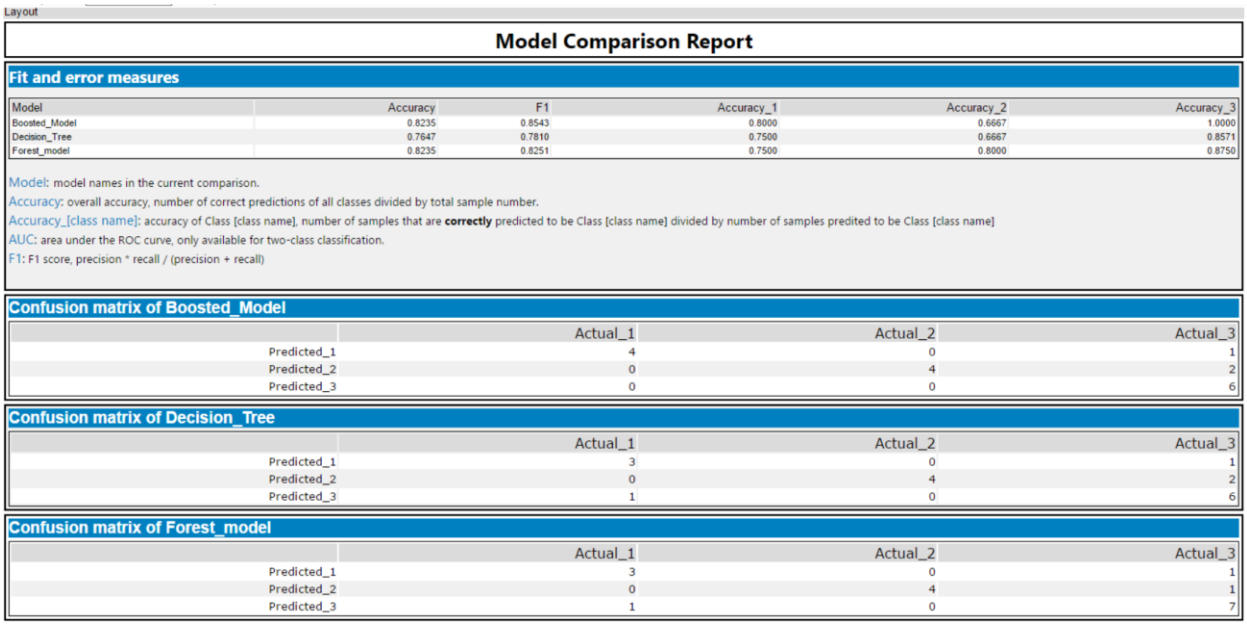
https://public.tableau.com/profile/rakshit.kumar.satish#!/vizhome/Project8_task1_Q4/Sheet1

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.)**

   This is a non-binary classification model and we need to predict the best store format for the new stores and below table gives the model comparison report for the different types of models.

   The boosted model is used to predict the best store format, the accuracy for the boosted model is 0.8235 and the F1 value is 0.8543 which is better than the decision tree and the forest models



Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Decision_Tree	0.7647	0.7810	0.7500	0.6667	0.8571
Forest_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)

Confusion matrix of Boosted_Model

	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

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	2
Predicted_3	1	0	6

Confusion matrix of Forest_model

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

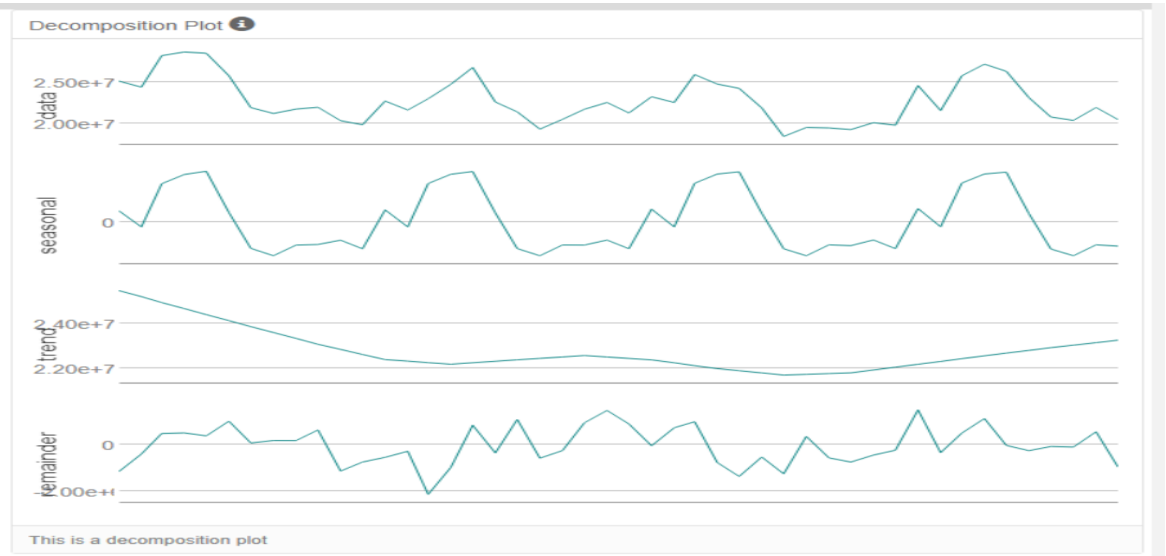
2. **What format do each of the 10 new stores fall into? Please fill in the table below.**

Store ID	Cluster Name
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? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?**

The TS plot for the data given in the current data set is shown below,



we can see that the error varies over the time so the error needs to be applied multiplicatively,

There is no clear trend which means the T = none.

The seasonal component varies slightly over the time and should be applied multiplicatively.

The model to be chosen for predicting the time series is ETS(M,N,M) model.

The ACF and PACF plots when we do not consider any lag is given below, we can observe that the data are highly corelated and cannot be used as such and needs to be differenced,

Image 1: Original ACF and PACF Plots with no differencing

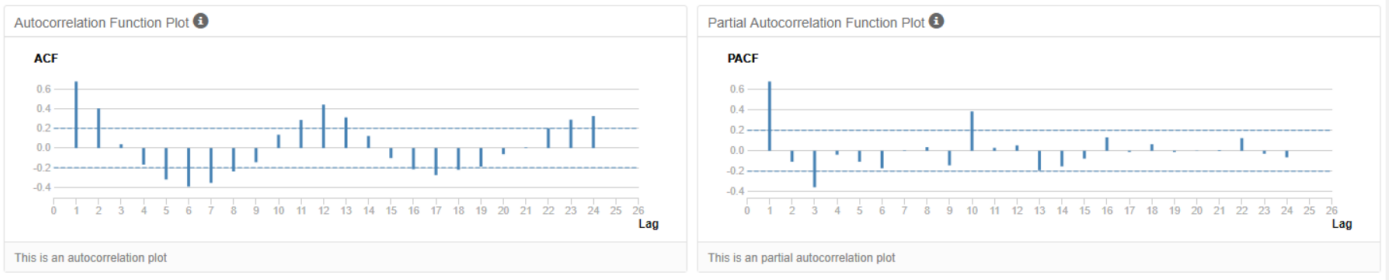


Image 2: ACF and PACF Plots with differencing

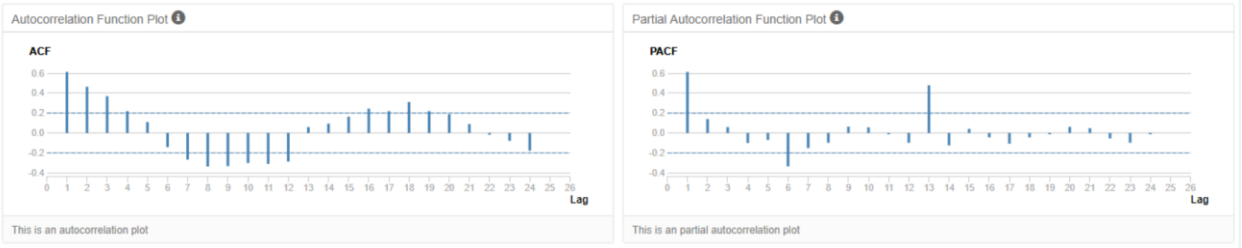
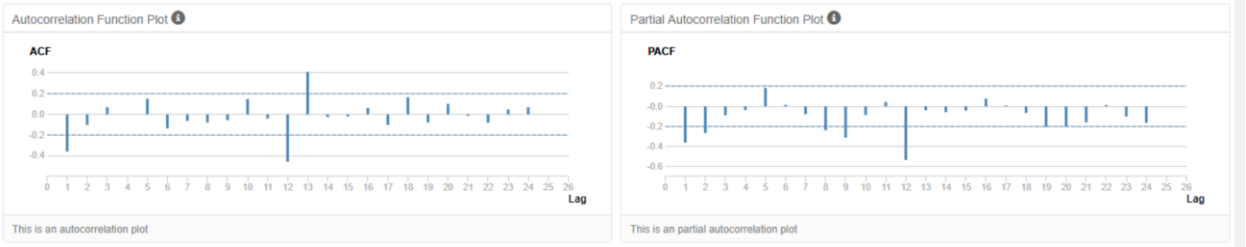


Image 3: ACF and PACF Plots after seasonal first differencing



We know that the time series given to us is seasonal and we have to remove the seasonality to have a better understanding of the ARIMA model.

We first do the seasonal differencing to remove the seasonality and then check the ACF and PACF plots, after the seasonal differencing we still see the correlation in the lags and the lags are significant.

Since we have a significant correlation, we have to make the series more stationary.

When we do the seasonal first differencing, all the significant lags are smoothened and this can be seen in the ACF and PACF plots, so there is no need for further differencing.

We have D(1) and d(1) since we have removed the seasonality in the first differencing,

The ACF and PACF plots show that there is a strong negative correlation at lag 1, this suggests that the MA(2) for the non-seasonal component.

When we look at the difference ACF-PACF plots the lags at 12 and 24 are not very significant, so there is no need for the seasonal component.

ARIMA (p,d,q) (P,D,Q)m

Non-Seasonal (p,q,d) is (0,1,2) and Seasonal (P,D,Q) is (0,1,0) , m is 12.

The model comparison between the ETS(M,N,M) and ARIMA(0,1,2)(0,1,0)[12] is given below

Summary of Time Series Exponential Smoothing Model ETS_Model

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

The comparison of the ETS and ARIMA model is shown below

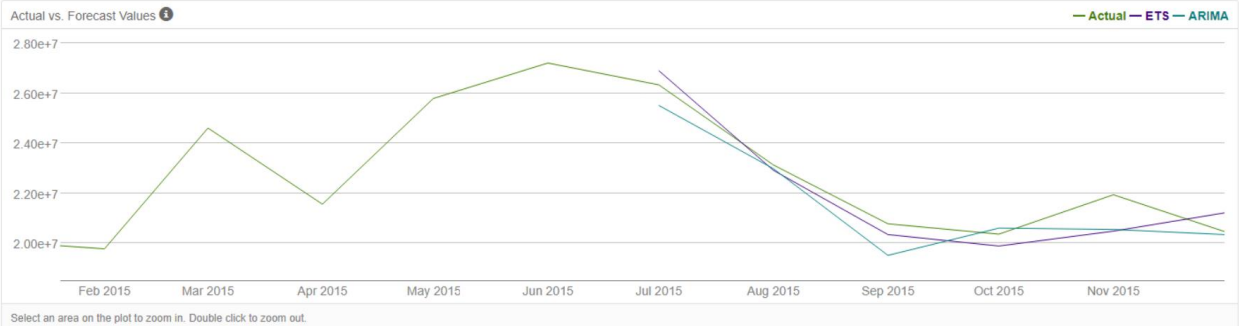
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909	NA

From the above table we can observe that the ETS model's acccuracy is higher than the ARIMA, the ME and RMSE of the ETS is lower when compared to that of the ARIMA.

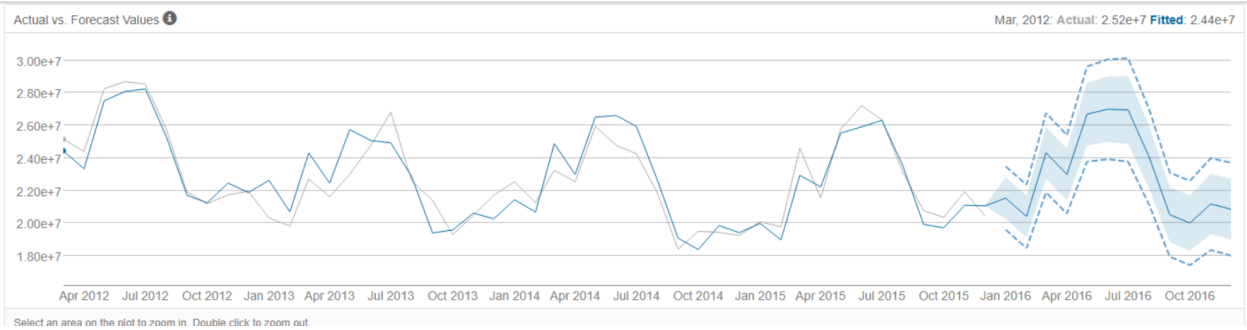
When the holdout sample of 6 is used to check the accuracy of the prediction the ETS model forecast better fits into the existing timeseries and hence we use ETS(M,N,M) for the forecasting.

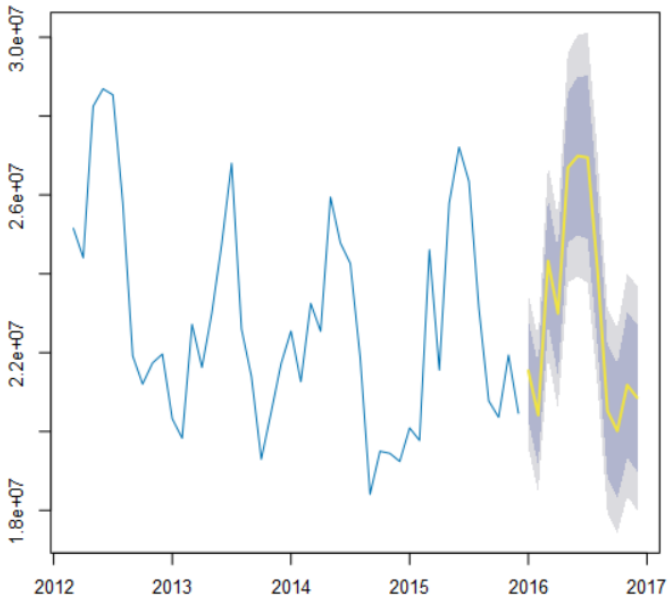
Comparison of the ETS and ARIMA forecasting against the holdout sample



The forecasting for the existing stores using the ETS(M,N,M) model is shown below

Period	Sub_Period	ETS_Forecast	ETS_Forecast_high_95	ETS_Forecast_high_80	ETS_Forecast_low_80	ETS_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

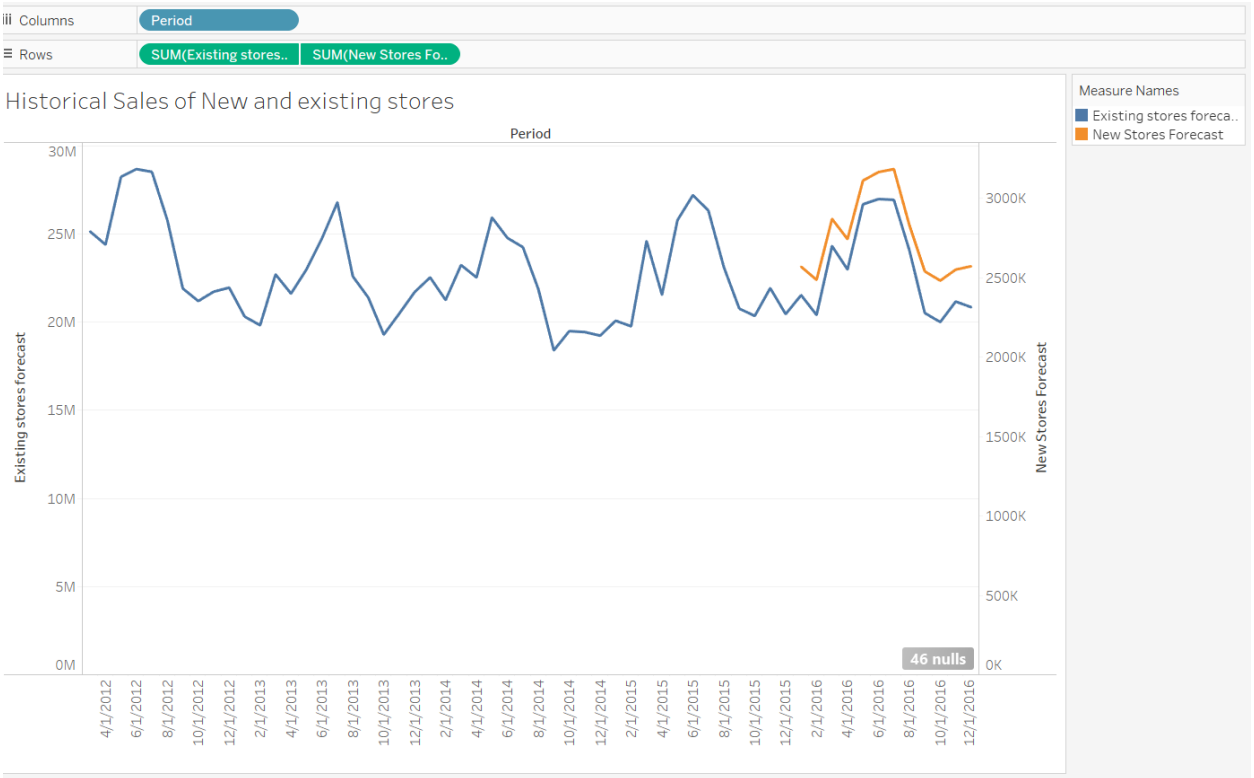

Actual vs. Forecast Values
Mar, 2012: Actual: 2.52e+7 Fitted: 2.44e+7
3.00e+7
2.80e+7
2.60e+7
2.40e+7
2.20e+7
2.00e+7
1.80e+7
Apr 2012
Jul 2012
Oct 2012
Jan 2013
Apr 2013
Jul 2013
Oct 2013
Jan 2014
Apr 2014
Jul 2014
Oct 2014
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Select an area on the plot to zoom in. Double click to zoom out


3.0e+07
2.6e+07
2.2e+07
1.8e+07
2012
2013
2014
2015
2016
2017

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.**

The table with the forecasts for the existing stores is given below.

Period	Existing stores forecast	New Stores Forecast
1/2016	21539936.01	2568379.761
2/2016	20413770.6	2485351.648
3/2016	24325953.1	2869220.974
4/2016	22993466.35	2742082.95
5/2016	26691951.42	3110168.799
6/2016	26989964.01	3163811.727
7/2016	26948630.76	3182833.176
8/2016	24091579.35	2830752.473
9/2016	20523492.41	2539033.656
10/2016	20011748.67	2480782.282
11/2016	21177435.49	2550045.743
12/2016	20855799.11	2571383.76



https://public.tableau.com/profile/rakshit.kumar.satish#!/vizhome/Project8_Task3/Sheet1