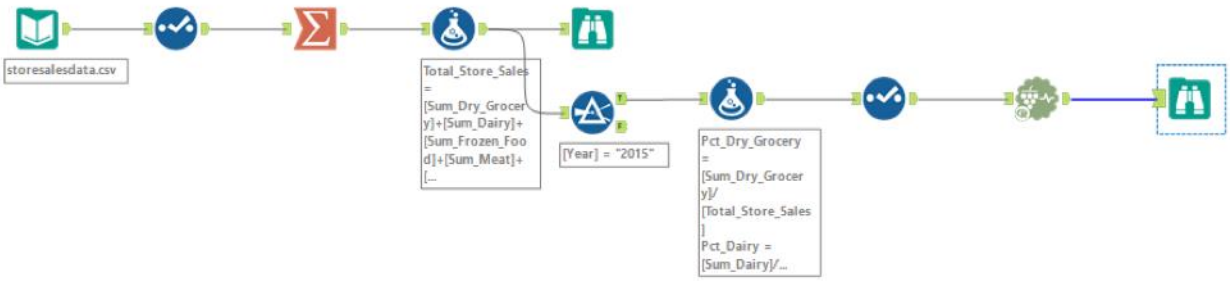
Project: Predictive Analytics Capstone

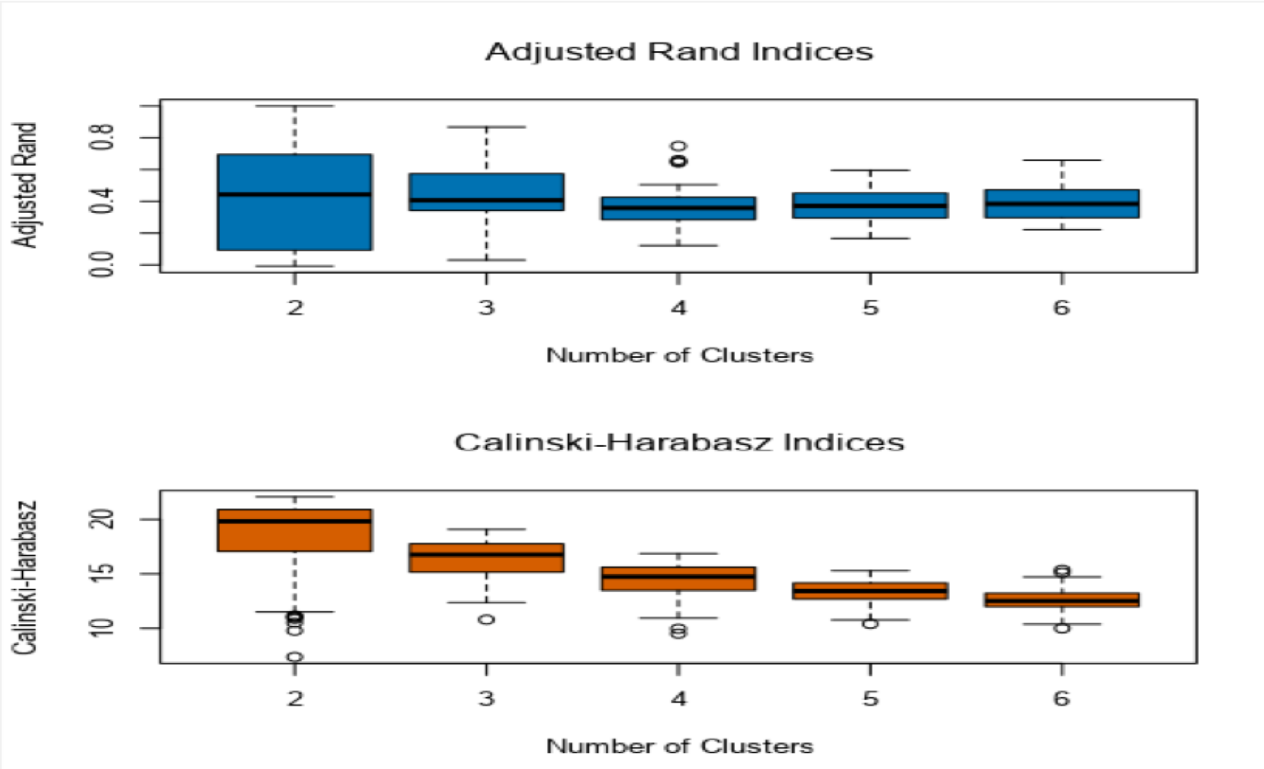
Project: Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. You've been asked to provide analytical support to make decisions about store formats and inventory planning.

Task 1: Determine Store Formats for Existing Stores

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



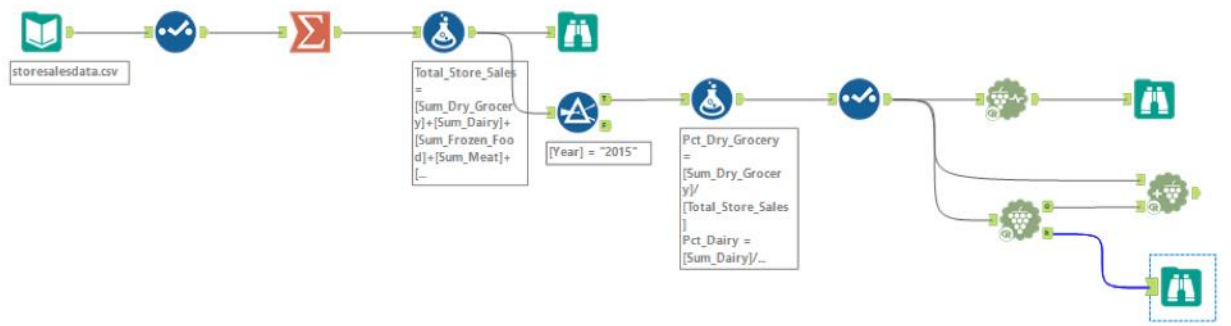
Plots



The optimal store format of 3 is selected using the Adjusted Rand and Calinski-Harabasz Indices.

The value of 3 was selected as the box and whiskers are very compact and the median is relatively high.

2. How many stores fall into each store format?



Report

Summary Report of the K-Means Clustering Solution Store_Cluster

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	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

Using the above model, we obtain the values given 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?

	Pct_Dry_Grocery	Pct_Dairy	Pct_Frozen_Food	Pct_Meat	Pct_Produce	Pct_Floral	Pct_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Pct_Bakery	Pct_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Given the findings listed under the summary report of percentage of sales for each category:

- Cluster 1 sells more General Merchandize
- Cluster 2 sells more Produce, Dairy and Floral
- Cluster 3 sells more Deli and Meat products.

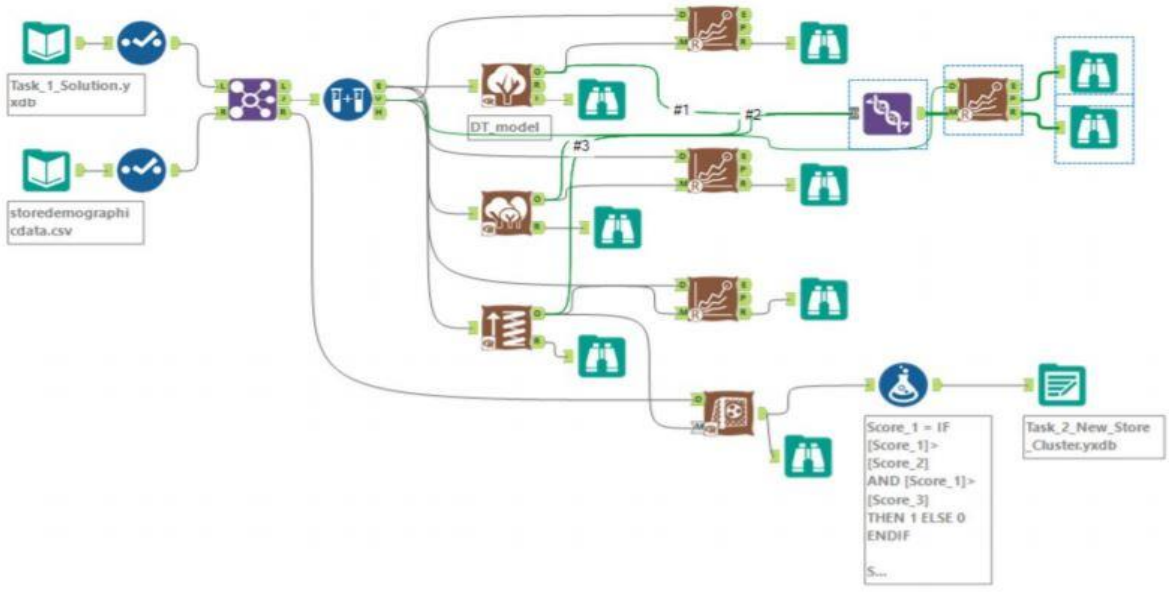
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the

location of the stores, uses color to show cluster, and size to show total sales.

URL:
https://public.tableau.com/profile/mihir.shinde7903#!/vizhome/StoreLocationbySalesandCluster/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.)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_model	0.8235	0.8426	0.7500	1.0000	0.7778
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	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 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 DT_model

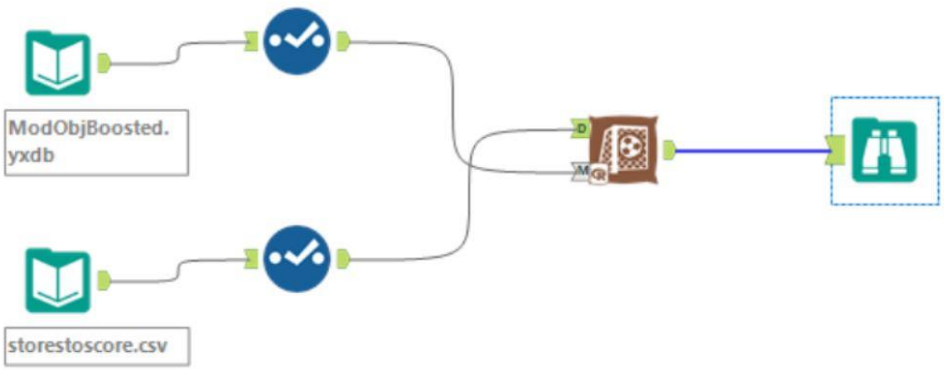
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

As the accuracy metrics for all three models is same at 0.8235, we choose the Boosted model for its higher F1 score of 0.8889. This higher F1 score denotes a higher Recall and precision value than the other models.

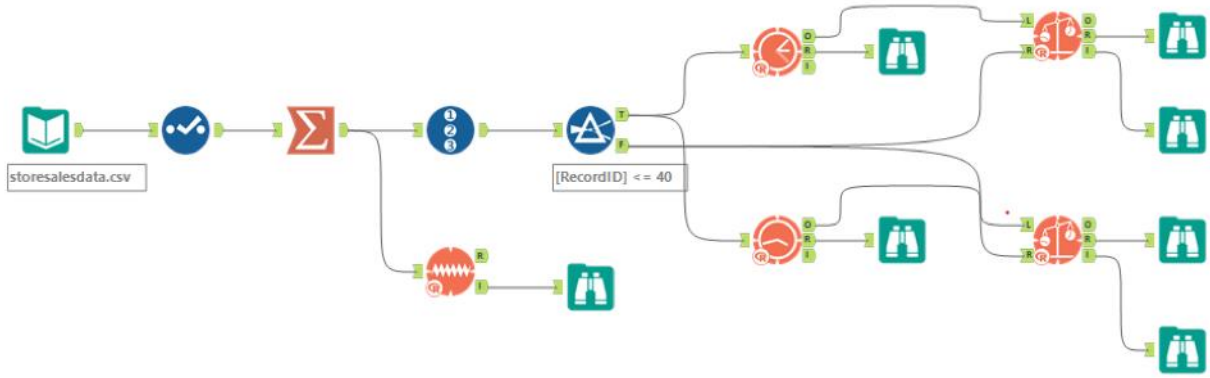
2. What format do each of the 10 new stores fall into? Please fill in the table below.



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

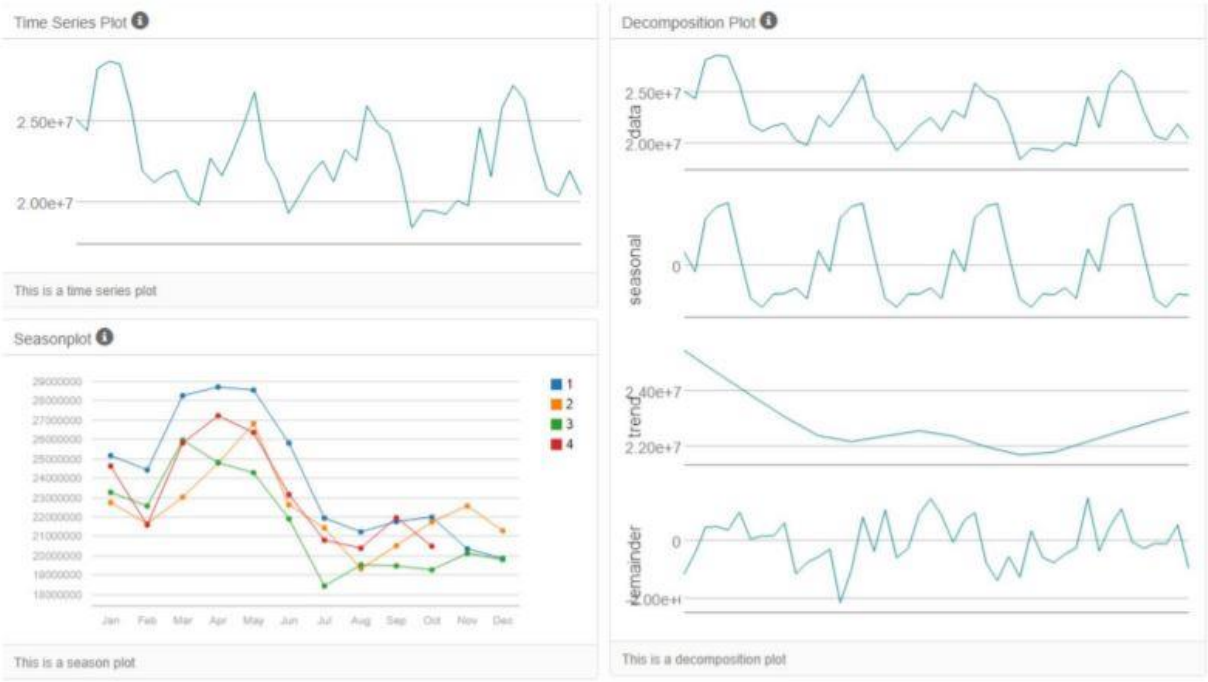
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?



Using the storesalesdata to summarize the sales by year and month. From the 46 records separating 6 records as a holdout sample and 40 records to training the models.

- Error is randomly distributed around the mean (m) – Applying Multiplicatively
- Trend is not clear (n) - Neutral
- Seasonality is visible (m) – Applying Multiplicatively

ETS Model : (m,n,m)



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822

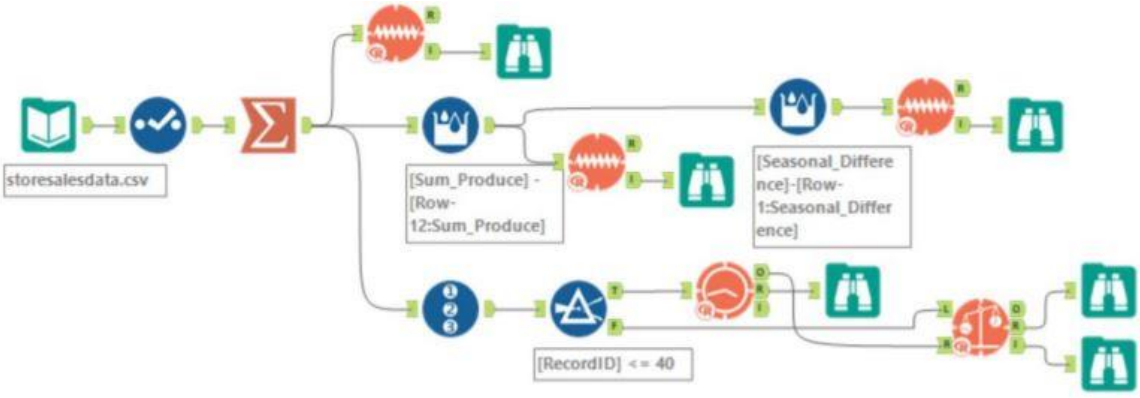
Using the TS compare tool on last six months of holdout sample, we observe the MASE score is well below 1.0 value which is a good sign.

ARIMA Model

From the plot below it seems the data is not stationarized around the y-axis.

The plot below looks better after first seasonal differencing

After looking at the ACF and PACF plots we find the model terms to be:
(p,d,q)(P,D,Q)m = (0,1,2)(0,1,0)12

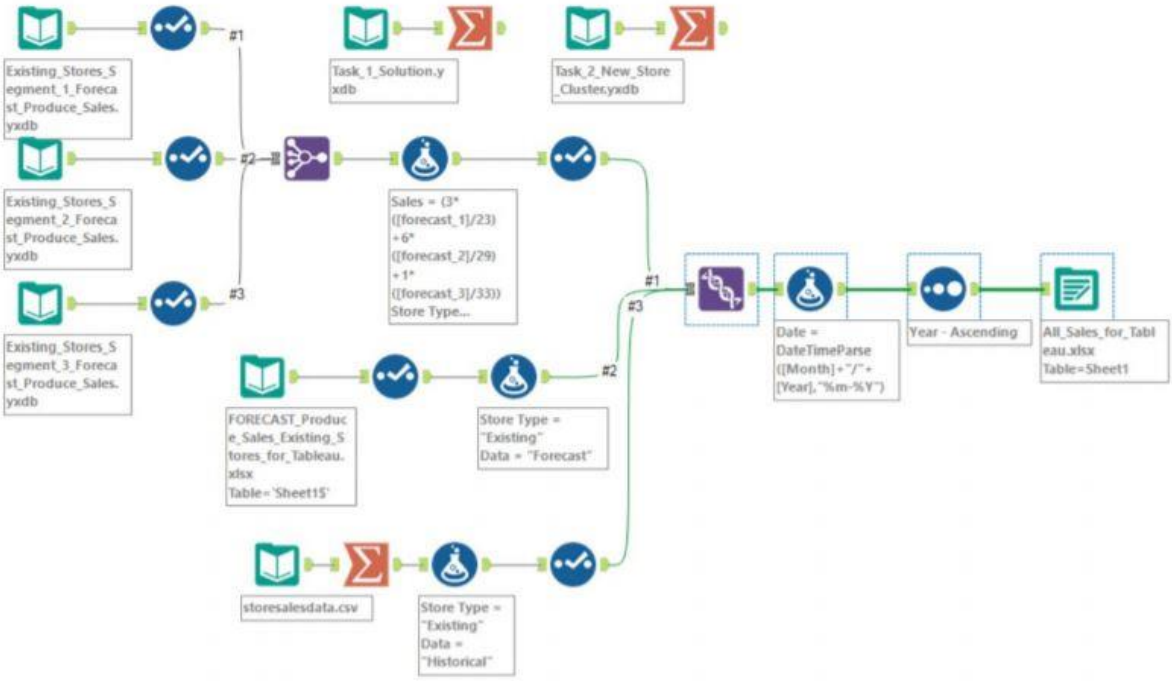


Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_0_1_2__0_1_0_12	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

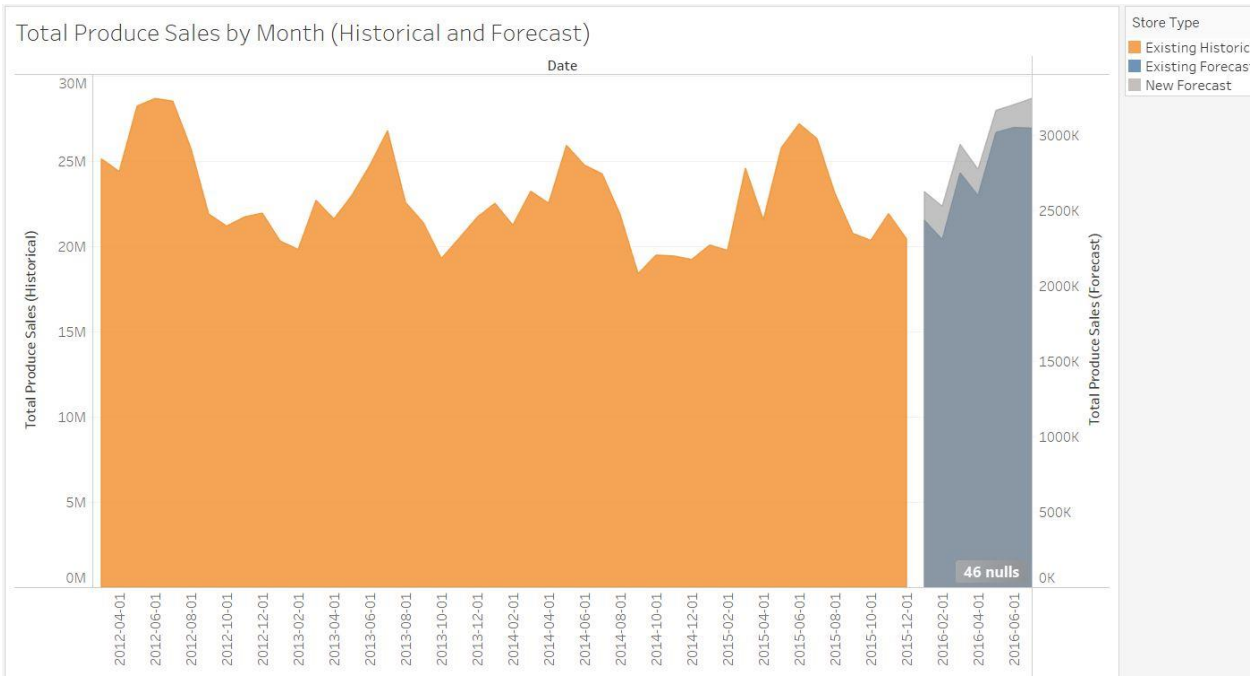
The MASE score of 0.3909 on ARIMA model is more than the 0.3822 found on the ETS model, indicates the ETS model is more accurate. The low RSME value on the ETS model indicates a forecast with a narrow range of possible values. Therefore, choosing ETS for forecasting.

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.



Combining all historic sales and total forecasted sales for new stores

Date	Existing Stores Forecast	New Stores Forecast
Jan-16	$2,15,39,936	$26,26,198
Feb-16	$2,04,13,771	$25,29,186
Mar-16	$2,43,25,953	$29,40,264
Apr-16	$2,29,93,466	$27,74,135
May-16	$2,66,91,951	$31,65,320
Jun-16	$2,69,89,964	$32,03,286
Jul-16	$2,69,48,631	$32,44,464
Aug-16	$2,40,91,579	$28,71,488
Sep-16	$2,05,23,492	$25,52,418
Oct-16	$2,00,11,749	$24,82,837
Nov-16	$2,11,77,435	$25,97,780
Dec-16	$2,08,55,799	$25,91,815



URL: https://public.tableau.com/profile/mihir.shinde7903#!/vizhome/Forecast-TotalProduceSalesbyMonth/TotalProduceSalesbyMonthHistoricalandForecast?publish=yes