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?
   3 even though at first by looking at AR and CH Indices we may think 2 is the optimal number.
   However, if we look closely, we can see in both indices that the compactness is better with 3 clusters as both min and max and IQ range is shorter with 3 clusters and the high median values are still close.
   So, 3 is the optimal number of clusters

K-Means Cluster Assessment Report

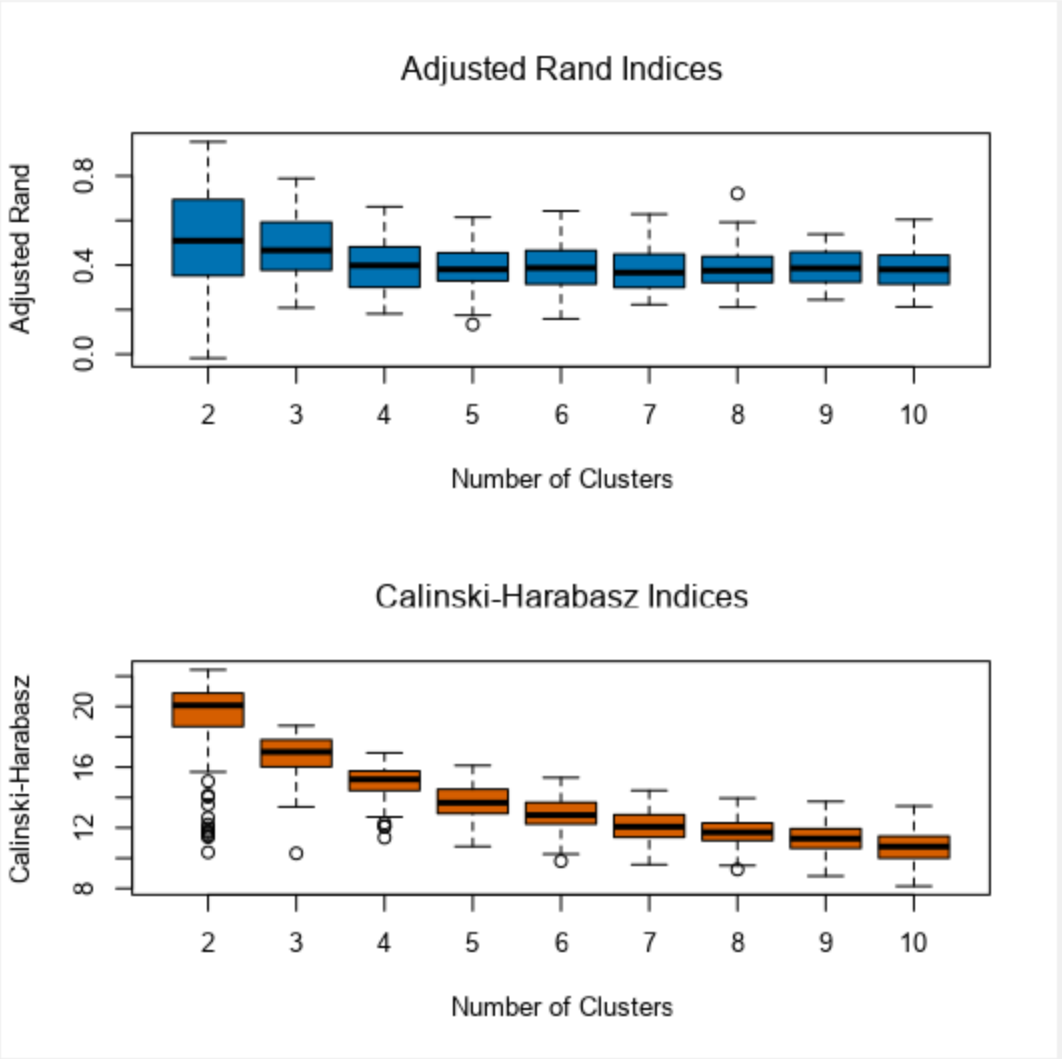
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.017586	0.208197	0.181585	0.133772	0.158757	0.222502	0.21093
1st Quartile	0.352613	0.377392	0.302314	0.331809	0.314419	0.299658	0.322749
Median	0.509257	0.466169	0.398104	0.380556	0.387434	0.366279	0.375409
Mean	0.494056	0.479493	0.404888	0.388834	0.39306	0.381404	0.384298
3rd Quartile	0.693746	0.58771	0.481097	0.454895	0.46369	0.447859	0.436717
Maximum	0.952939	0.788895	0.661744	0.614672	0.64242	0.62851	0.720498
	9	10					
Minimum	0.244439	0.212783					
1st Quartile	0.325103	0.315087					
Median	0.386151	0.380127					
Mean	0.390303	0.379638					
3rd Quartile	0.457811	0.442954					
Maximum	0.538277	0.604545					

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	10.38298	10.31461	11.34984	10.77356	9.80353	9.577281	9.253901
1st Quartile	18.69647	16.03968	14.46704	12.9405	12.24542	11.378557	11.166056
Median	20.07012	17.00754	15.19152	13.65142	12.83476	12.07357	11.697797
Mean	19.08577	16.73685	14.98778	13.68998	12.83426	12.156743	11.681178
3rd Quartile	20.87407	17.78773	15.74729	14.53404	13.67175	12.859807	12.311206
Maximum	22.41555	18.73715	16.93911	16.10526	15.30862	14.460893	13.955665
	9	10					
Minimum	8.822973	8.153824					
1st Quartile	10.648806	10.002731					
Median	11.287124	10.760594					
Mean	11.359959	10.745482					
3rd Quartile	11.937564	11.429852					
Maximum	13.731897	13.433832					



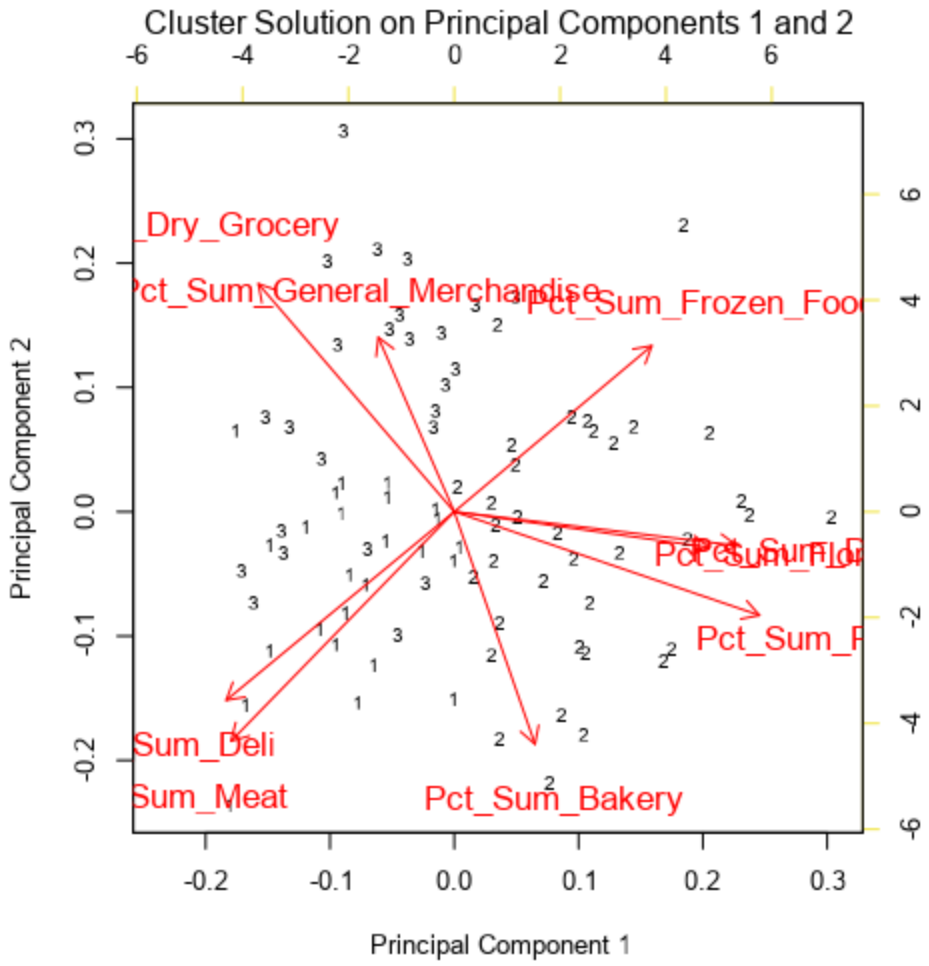
2. How many stores fall into each store format?

   Cluster 1: 25 stores
   Cluster 2: 35 stores
   Cluster 3: 25 stores

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

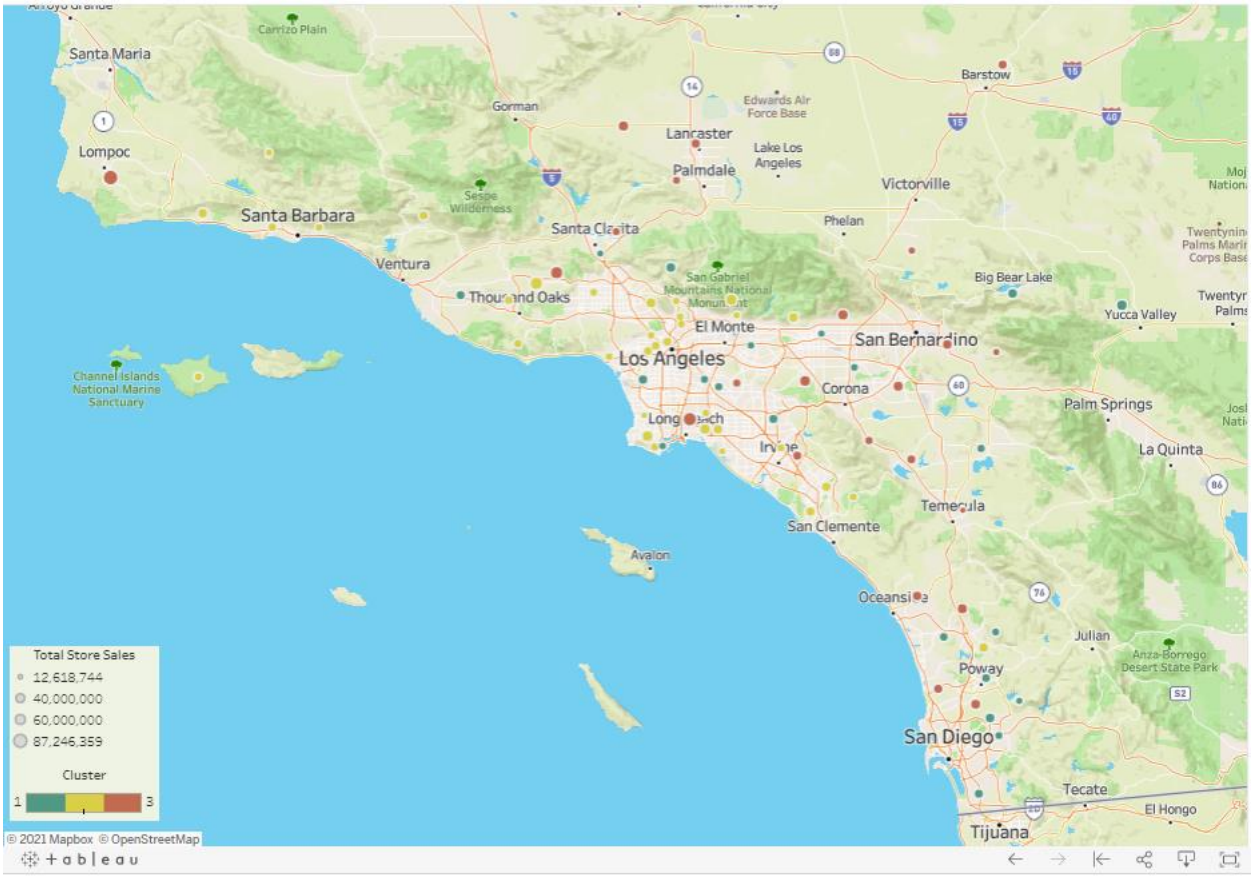
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   One possible interpretation could be that cluster 1 sells more products in the categories “Meat” and “Deli”, cluster 2 sells more products in the categories “Produce” and “Floral” and cluster 3 sells more in the categories “General Merchandise” and perhaps “Dry Grocery”



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.

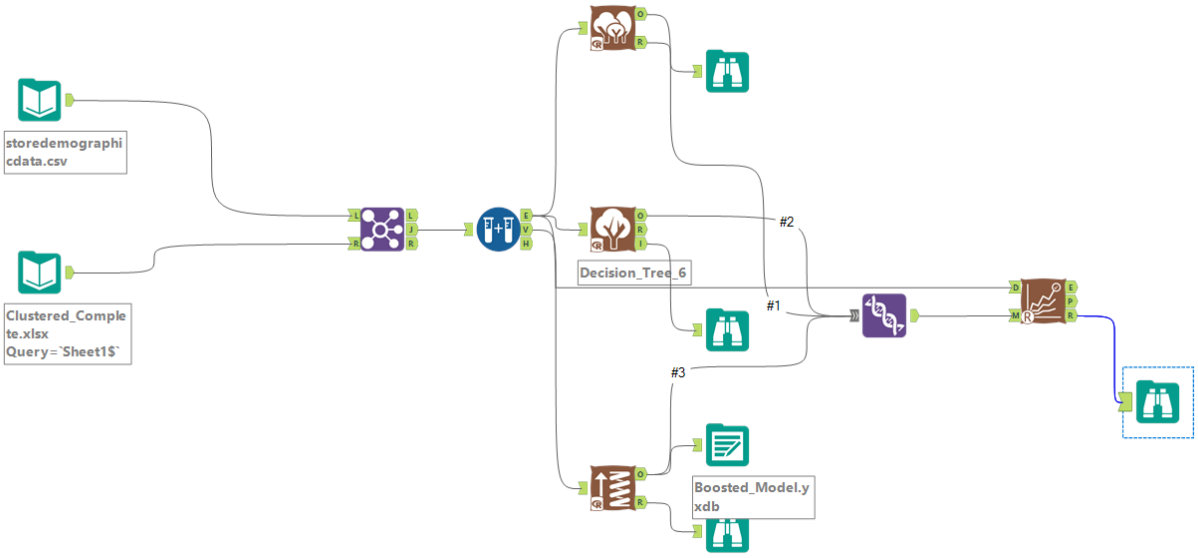
   https://public.tableau.com/app/profile/antonios.fledos/viz/UdacityPr_/Dashboard4?publish=yes



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 used Boosted Model after comparing it with Forest model and Decision Tree with the Model Comparison Tool.



Boosted Model showed the higher Accuracy and F1 score. It also had the most True Positives out of all the models.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest	0.6471	0.7083	0.3750	1.0000	0.7500
Decision_Tree_6	0.6471	0.6667	0.5000	1.0000	0.5000
Boosted	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 Boosted

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

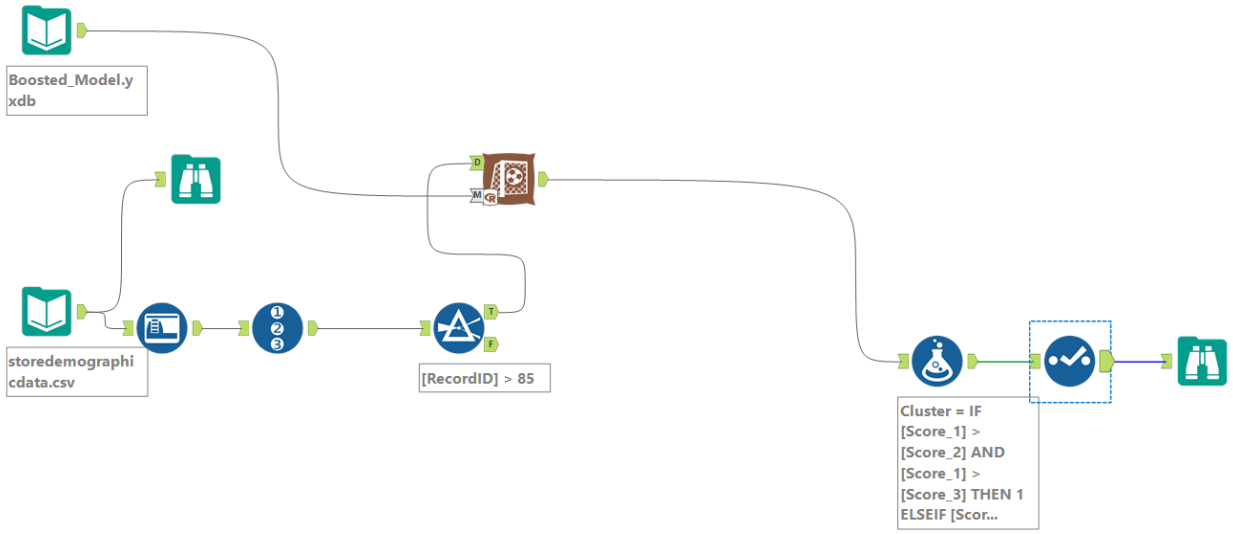
Confusion matrix of Decision_Tree_6

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

Confusion matrix of Forest

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

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



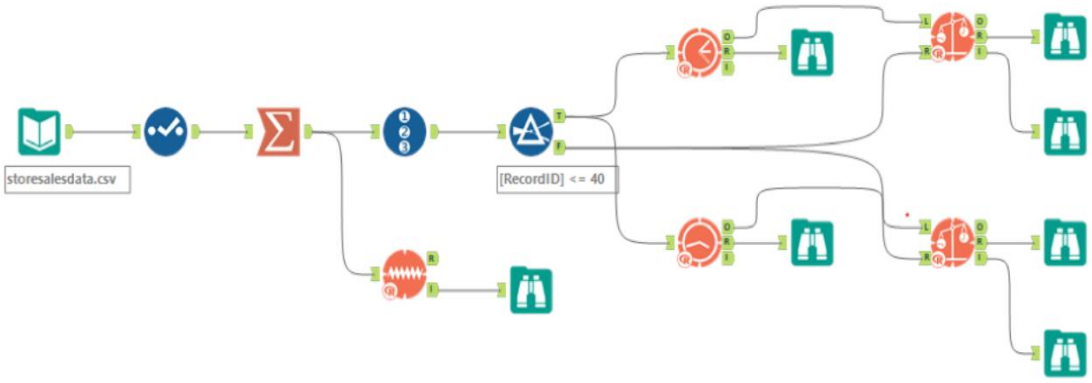
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

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?

I used ETS(M,N,M).

I did it after trying both ETS and ARIMA models as follows.



I used a summary tool summarizing the Produce sales grouped by “Year” and “Month”, I then added a Record ID tool and filtered my dataset. I kept the first 40 observations for the models’ training and the last 6 as validation/holdout set. I added a TS Compare tool to obtain the forecast error measurements against the holdout set for each model and then trained ETS and ARIMA models.

From the TS Compare Tool I looked at the Decomposition Plot, where we can see that the seasonality is multiplicative, the trend is none and the error is multiplicative as well.

So we used ETS(M,N,M).

I finally compared the results of the models against the holdout samples and the ETS model performs better than the ARIMA.

Report

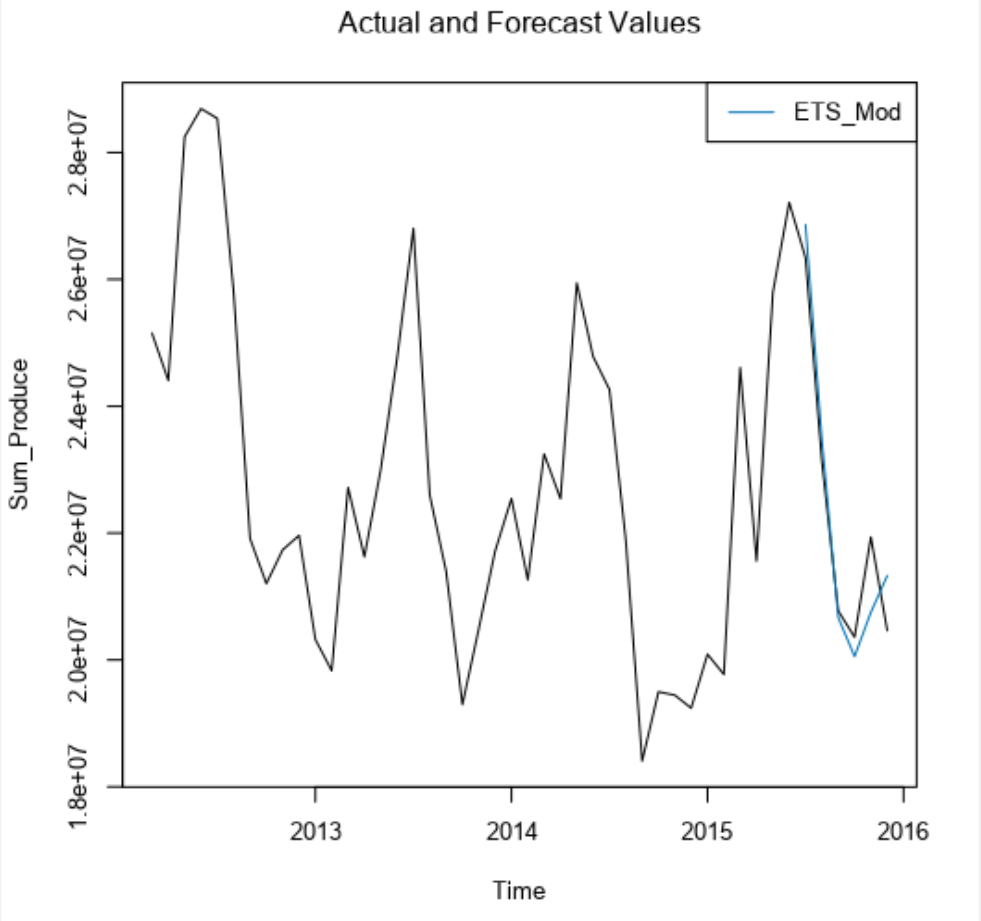
Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS_Mod
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_Mod	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257



Report

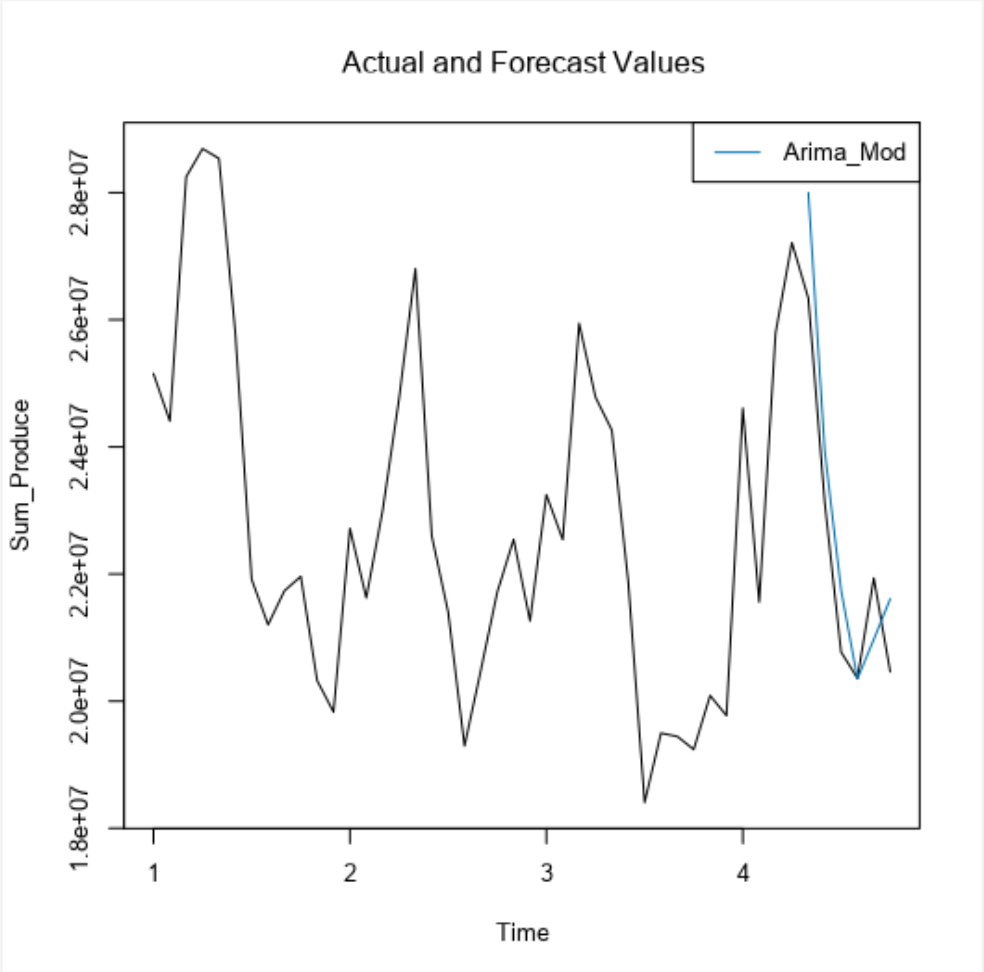
Comparison of Time Series Models

Actual and Forecast Values:

Actual	Arima_Mod
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_Mod	-604232.3	1050239	928412	-2.6156	4.0942	0.5463



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	Existing Stores
Jan-16	2563358	21829060
Feb-16	2483925	21146330
Mar-16	2910944	23735687
Apr-16	2764882	22409515
May-16	3141306	25621829
Jun-16	3195054	26307858
Jul-16	3212391	26705093
Aug-16	2852386	23440761
Sep-16	2521697	20640047
Oct-16	2466751	20086270
Nov-16	2557745	20858120
Dec-16	2530511	21255190