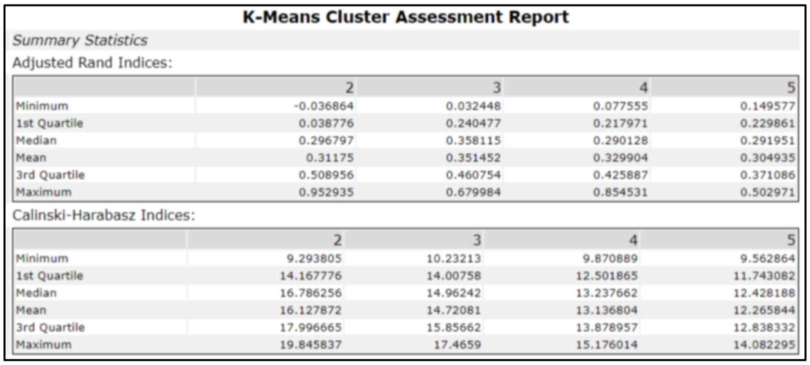
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   After cleaning and preparing the data from the store sales data file and store information file, I used the K-centroids diagnostics tool to output the following report below, which concluded 3 as the optimal number of clusters by using the K-mean clustering method it showed that 3 clusters have the highest median and mean.



K-Means Cluster Assessment Report

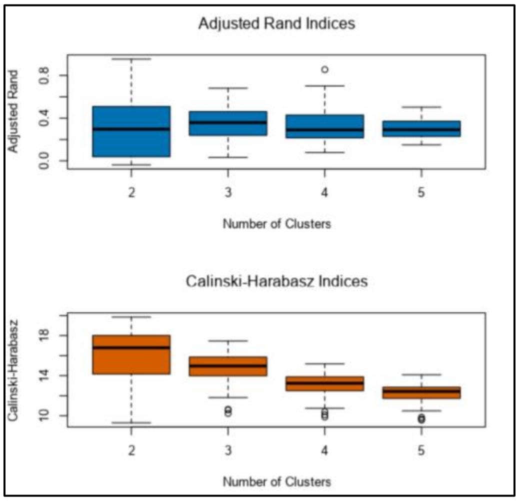
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5
Minimum	-0.036864	0.032448	0.077555	0.149577
1st Quartile	0.038776	0.240477	0.217971	0.229861
Median	0.296797	0.358115	0.290128	0.291951
Mean	0.31175	0.351452	0.329904	0.304935
3rd Quartile	0.508956	0.460754	0.425887	0.371086
Maximum	0.952935	0.679984	0.854531	0.502971

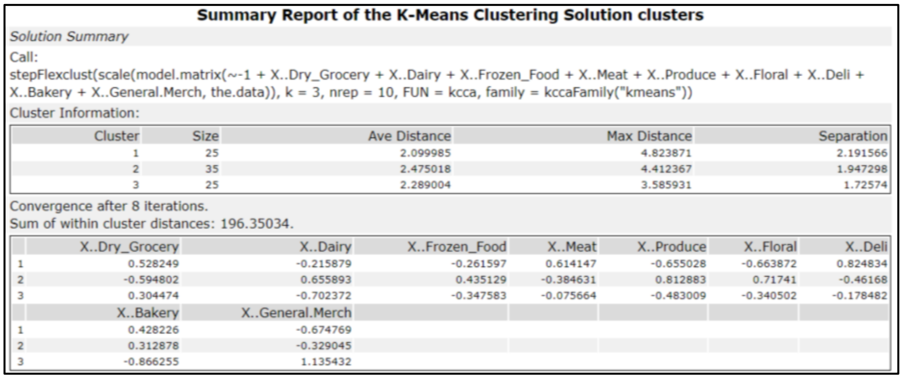
Calinski-Harabasz Indices:

	2	3	4	5
Minimum	9.293805	10.23213	9.870889	9.562864
1st Quartile	14.167776	14.00758	12.501865	11.743082
Median	16.786256	14.96242	13.237662	12.428188
Mean	16.127872	14.72081	13.136804	12.265844
3rd Quartile	17.996665	15.85662	13.878957	12.838332
Maximum	19.845837	17.4659	15.176014	14.082295



2. How many stores fall into each store format?

   My clustering concluded that first cluster has 25 stores, second cluster has 35 stores, and the third cluster has 25 stores.



Summary Report of the K-Means Clustering Solution clusters

Solution Summary

Call:

```
stepFlexclust(scale(model.matrix(~-1 + X..Dry_Grocery + X..Dairy + X..Frozen_Food + X..Meat + X..Produce + X..Floral + X..Deli +
X..Bakery + X..General.Merch, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))
```

Cluster Information:

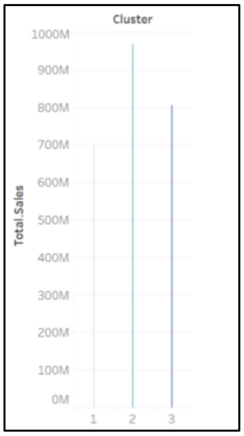
Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

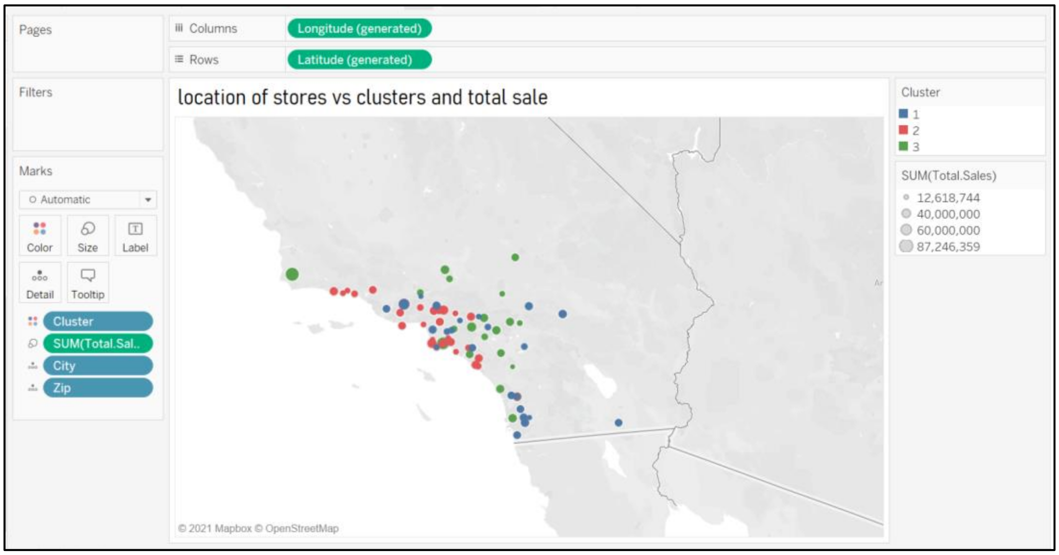
	X..Dry_Grocery	X..Dairy	X..Frozen_Food	X..Meat	X..Produce	X..Floral	X..Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482
	X..Bakery	X..General.Merch					
1	0.428226	-0.674769					
2	0.312878	-0.329045					
3	-0.866255	1.135432					

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

One of the differences is that the 2nd cluster seems to have higher sales than the other clusters



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.


Columns: Longitude (generated)

Rows: Latitude (generated)

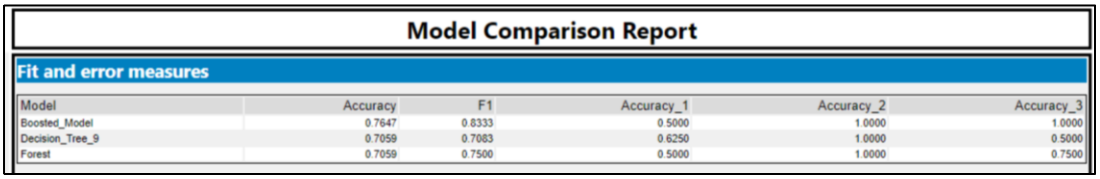
Marks: Automatic; Cluster, SUM(Total.Sal.., City, Zip

location of stores vs clusters and total sale

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.)

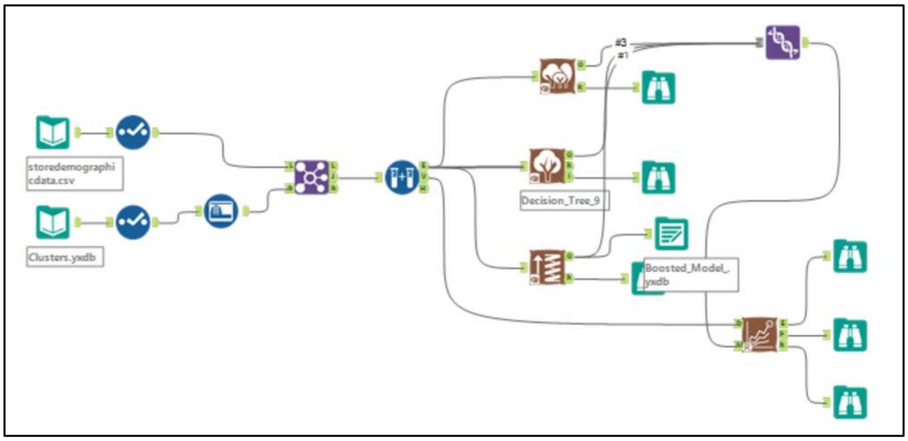
After generating a model comparison report, I have chosen the boosted model as it showed the highest accuracy results



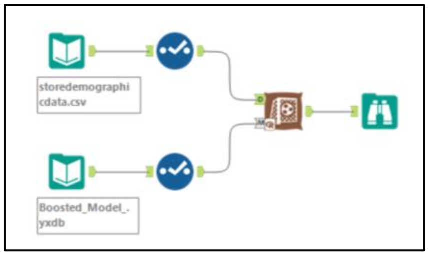
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted_Model	0.7647	0.8333	0.5000	1.0000	1.0000
Decision_Tree_9	0.7059	0.7083	0.6250	1.0000	0.5000
Forest	0.7059	0.7500	0.5000	1.0000	0.7500

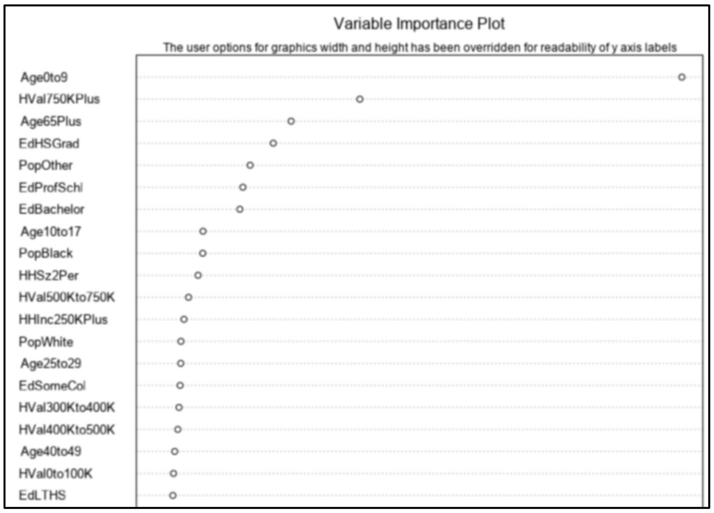


2. What format do each of the 10 new stores fall into? Please fill in the table below.



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

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

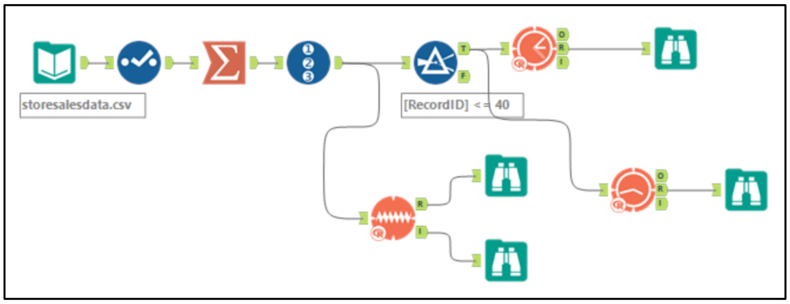


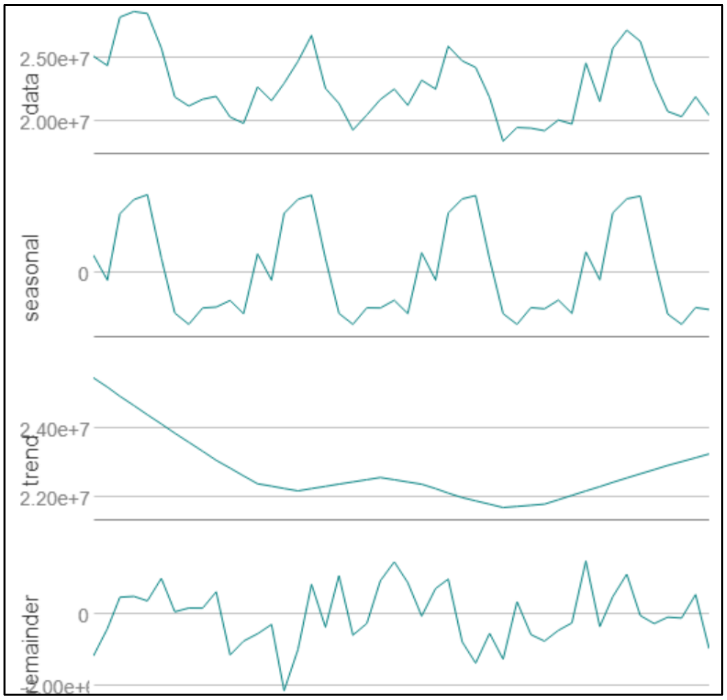
Age0to9, HVal750KPlus, and Age65Plus are the most important variables.

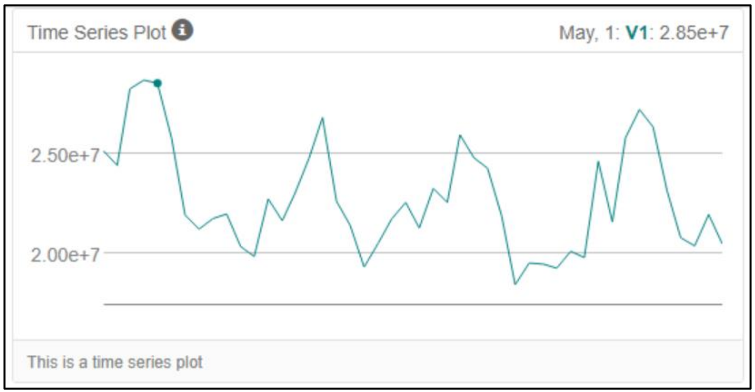
Task 3: Predicting Produce Sales

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?

After comparing the ETS and ARIMA model, the ETS (M,N,M) model showed best results of the forcast and lower error values. As observed in the plot below, there is no trend and seasonal is multiplicative and error is multiplicative.. Thus, by comparing, the **ETS preformed** better than the ARIMA model. In-sample error also showed better results on ETS unlike the ARIMA model interms of accurecy.

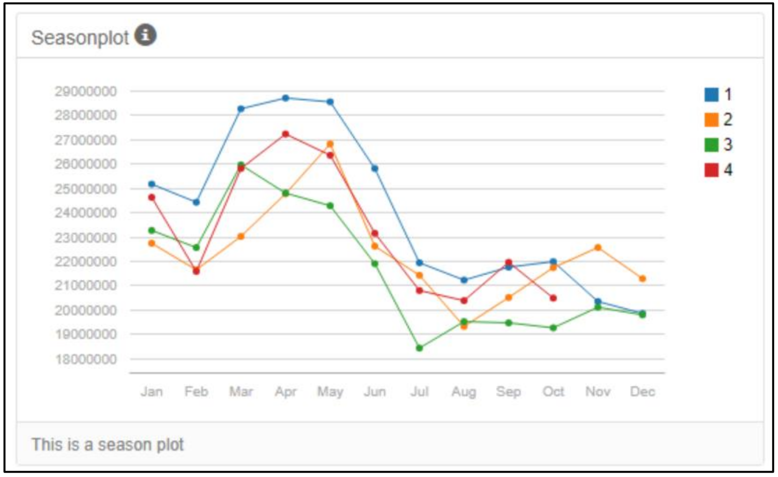





Time Series Plot ℹ

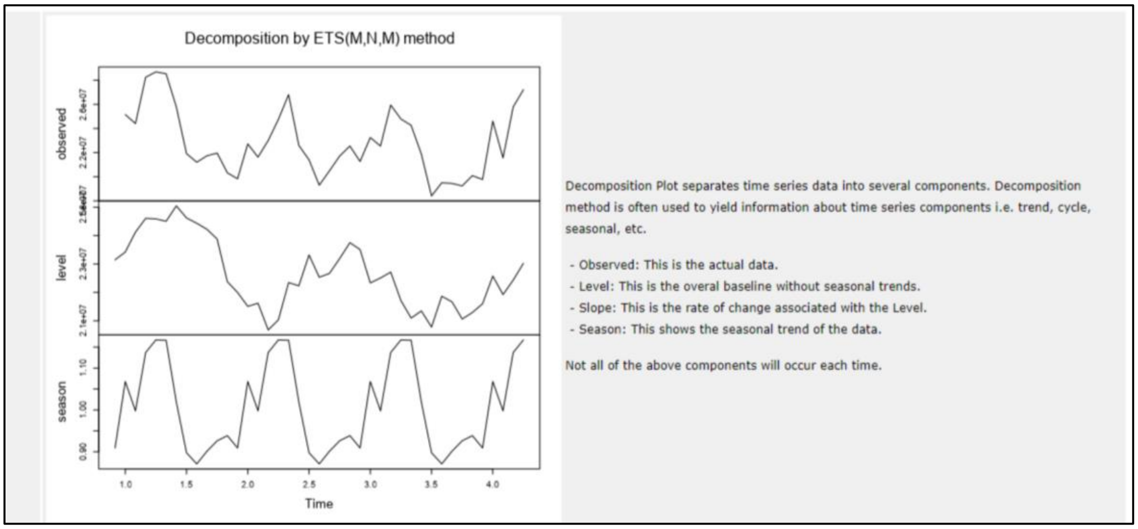
May, 1: **V1**: 2.85e+7

This is a time series plot


Seasonplot ℹ

This is a season plot

ETS (M,N,M)decomposition plot

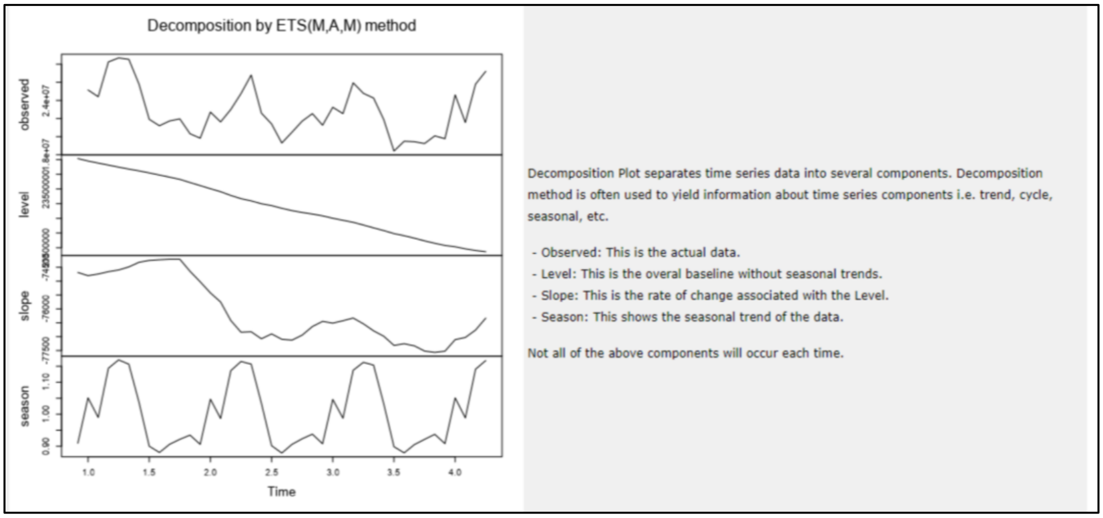


Decomposition Plot separates time series data into several components. Decomposition method is often used to yield information about time series components i.e. trend, cycle, seasonal, etc.

- Observed: This is the actual data.
- Level: This is the overal baseline without seasonal trends.
- Slope: This is the rate of change associated with the Level.
- Season: This shows the seasonal trend of the data.

Not all of the above components will occur each time.

ETS (M,A,M)decomposition plot

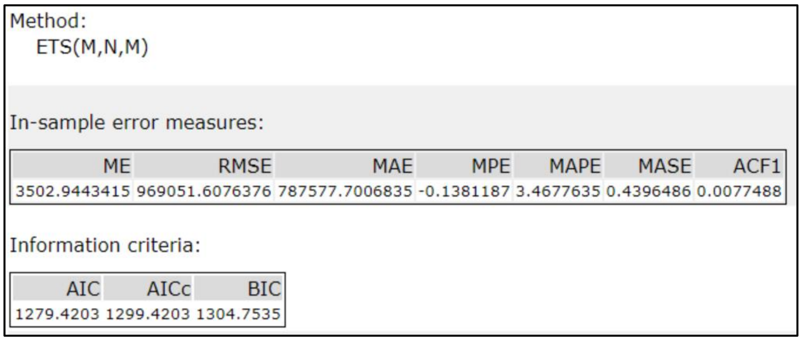


Decomposition Plot separates time series data into several components. Decomposition method is often used to yield information about time series components i.e. trend, cycle, seasonal, etc.

- Observed: This is the actual data.
- Level: This is the overal baseline without seasonal trends.
- Slope: This is the rate of change associated with the Level.
- Season: This shows the seasonal trend of the data.

Not all of the above components will occur each time.

ETS



Method:
ETS(M,N,M)

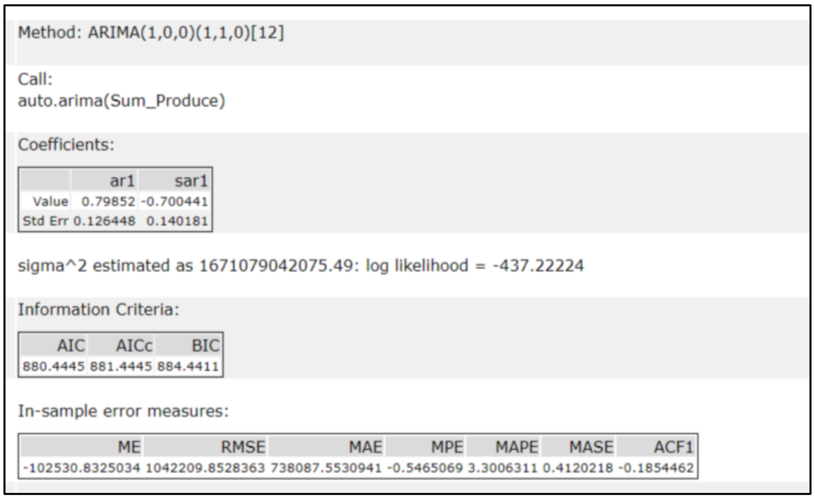
In-sample error measures:

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

ARIMA



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

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

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

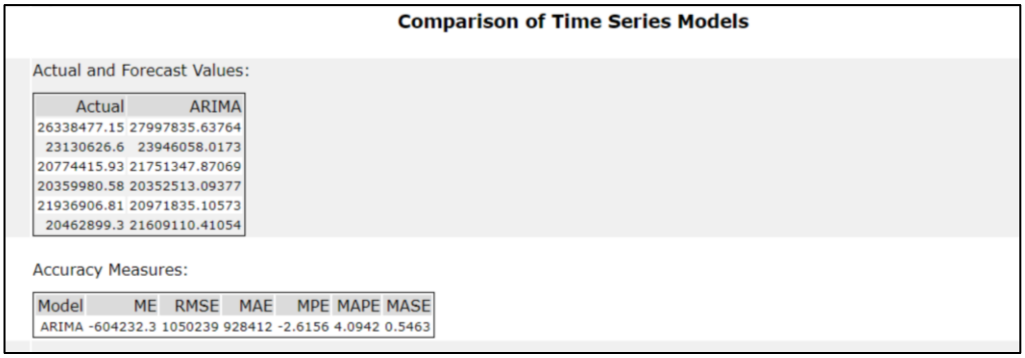
Information Criteria:

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

accuracy measures/forecast error measurements against the holdout sample; ETS showed way better results than ARIMA



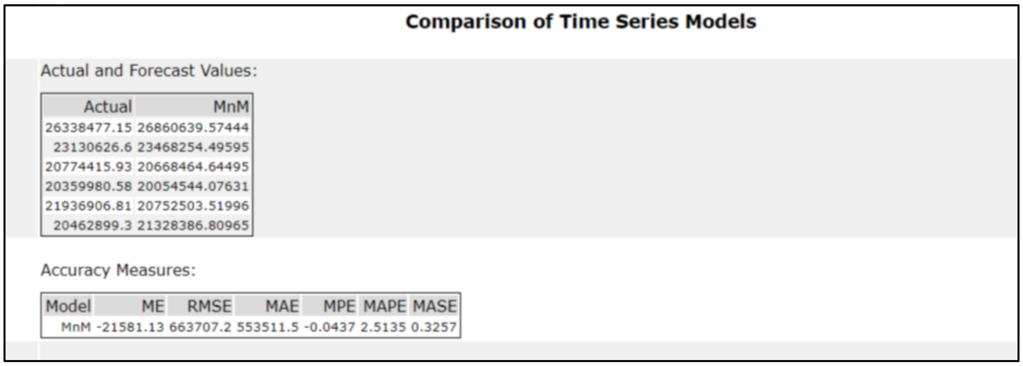
Comparison of Time Series Models

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



Comparison of Time Series Models

Actual and Forecast Values:

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

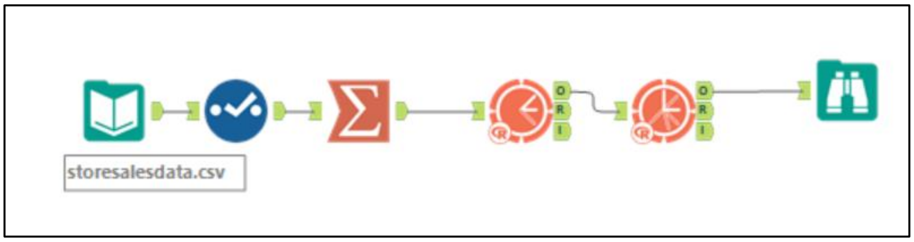
Accuracy Measures:

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

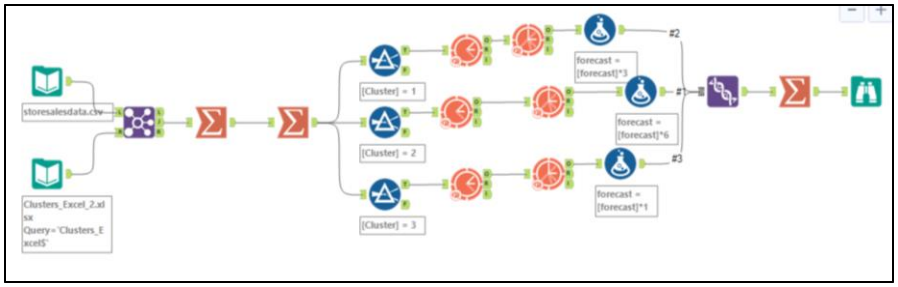
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Month	Existing Stores	New Stores
Jan-16	20,814,202.36	2491319
Feb-16	20,101,180.14	2408385
Mar-16	22,829,934.27	2833157

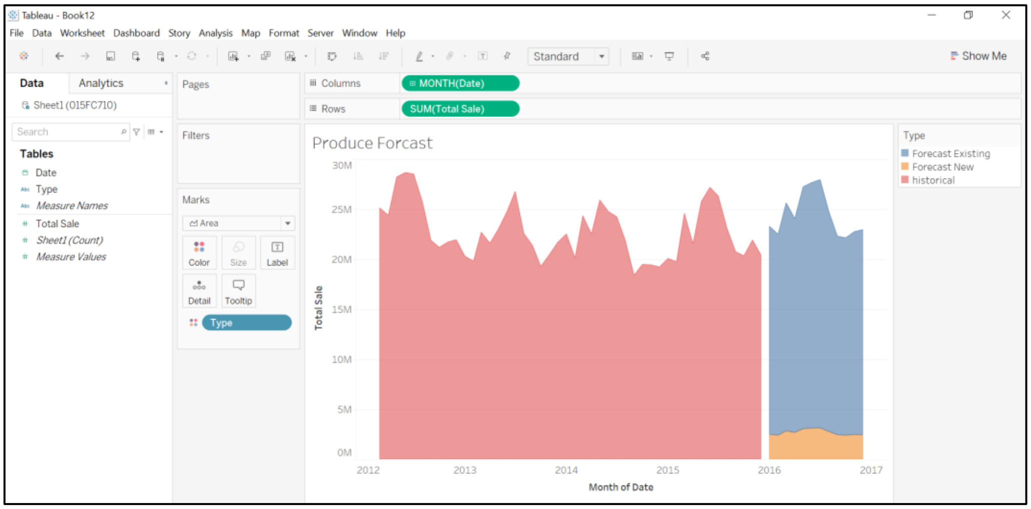
Apr-16	21,396,217.72	2679433
May-16	24,202,378.04	3054886
Jun-16	24,580,208.32	3106152
Jul-16	24,846,391.97	3132699
Aug-16	22,035,840.79	2776154
Sep-16	19,871,327.62	2451566
Oct-16	19,751,047.19	2401772
Nov-16	20,298,711.96	2477302
Dec-16	20,518,134.12	2452170



Existing stores workflow



New stores workflow

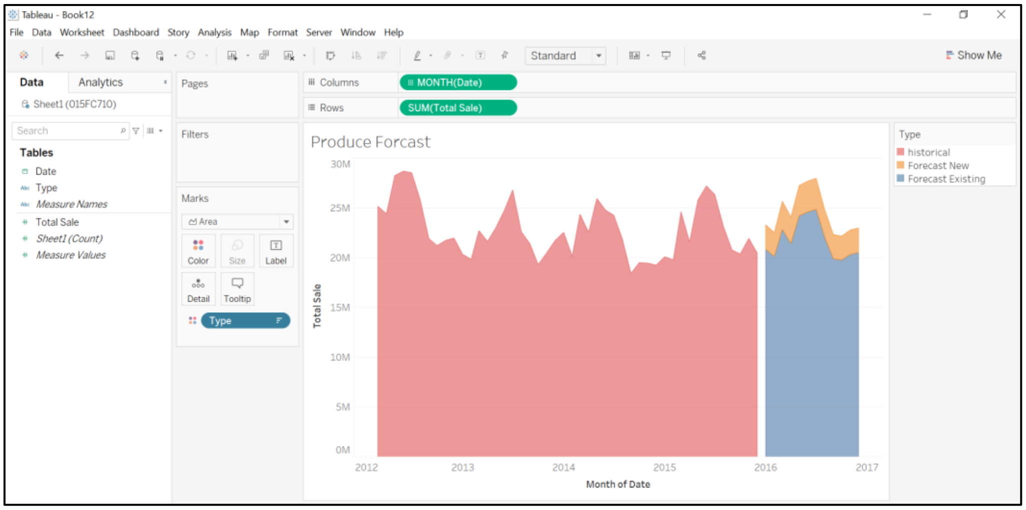


Columns: MONTH(Date)

Rows: SUM(Total Sale)

Marks: Area

Type



Columns MONTH(Date)

Rows SUM(Total Sale)