Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

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

A K-Centroid analysis using K-means clustering is being used to determine the optimal number of stores. K value ranges from 3 to 6. Ultimately after the analysis, K=3 is selected as it has the best mean in both the indices as is evident in the below report. Note k=2 is not selected as it would mean more than 40 stores per cluster and k=1 is not possible.

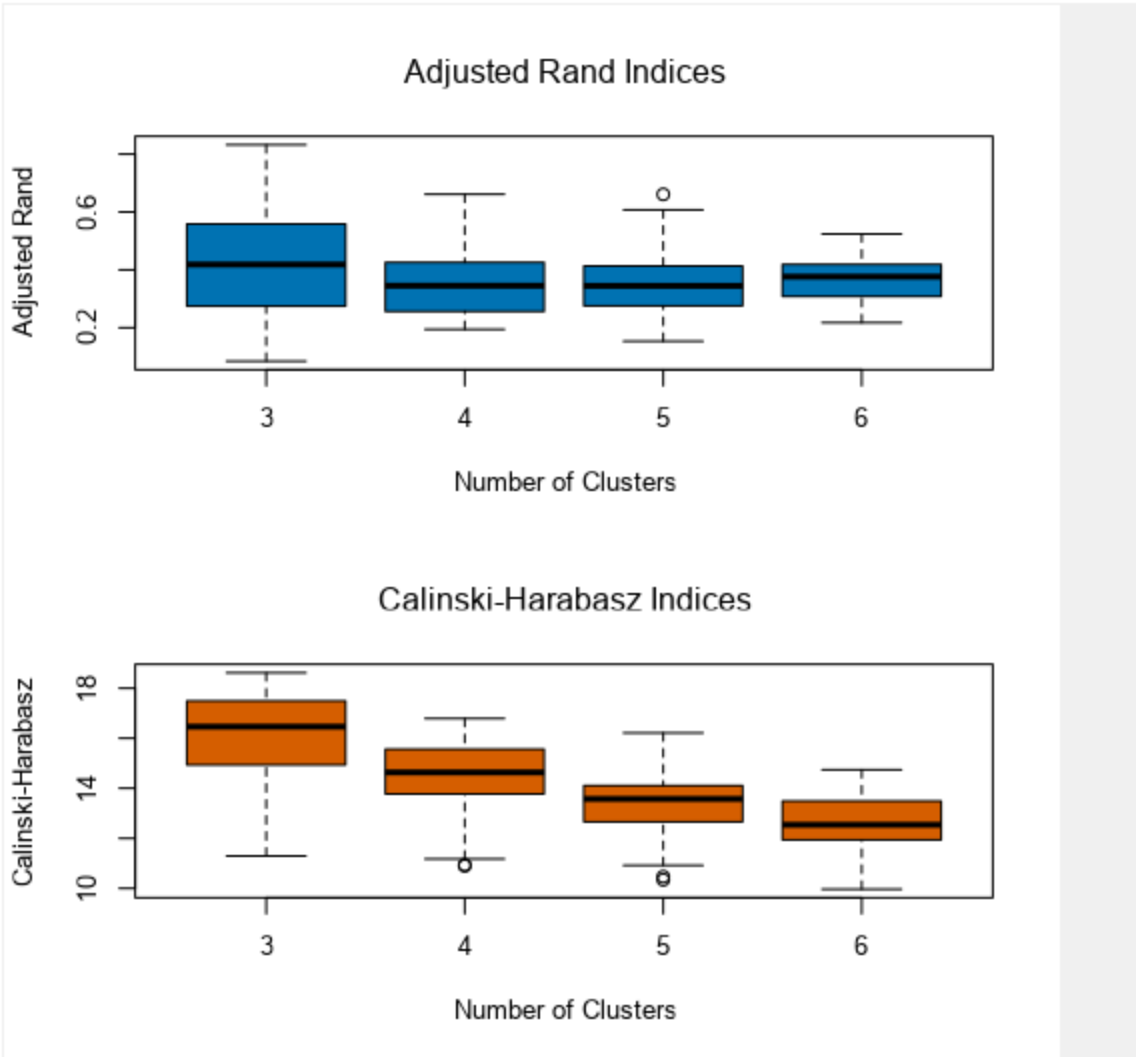
Record	Report
1	**K-Means Cluster Assessment Report**
2	*Summary Statistics*
3	Adjusted Rand Indices:

	3	4	5	6
Minimum	0.084659	0.194793	0.153402	0.217594
1st Quartile	0.281183	0.258162	0.279272	0.309885
Median	0.418371	0.345391	0.344932	0.376624
Mean	0.432245	0.366253	0.349889	0.372618
3rd Quartile	0.555496	0.422207	0.409252	0.418408
Maximum	0.832182	0.661275	0.661121	0.524212

5 Calinski-Harabasz Indices:

	3	4	5	6
Minimum	11.29684	10.90095	10.34414	9.971049
1st Quartile	14.98526	13.796	12.67167	11.947254
Median	16.46413	14.63847	13.57379	12.542131
Mean	16.09654	14.4821	13.45941	12.626029
3rd Quartile	17.46618	15.55479	14.11102	13.473271
Maximum	18.6146	16.78912	16.21386	14.740767

Plots



2. How many stores fall into each store format?

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

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

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

Following is the report of the K means Clustering.

Report

Summary Report of the K-Means Clustering Solution X

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + pct_Dry_Grocery + pct_Diry + 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	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

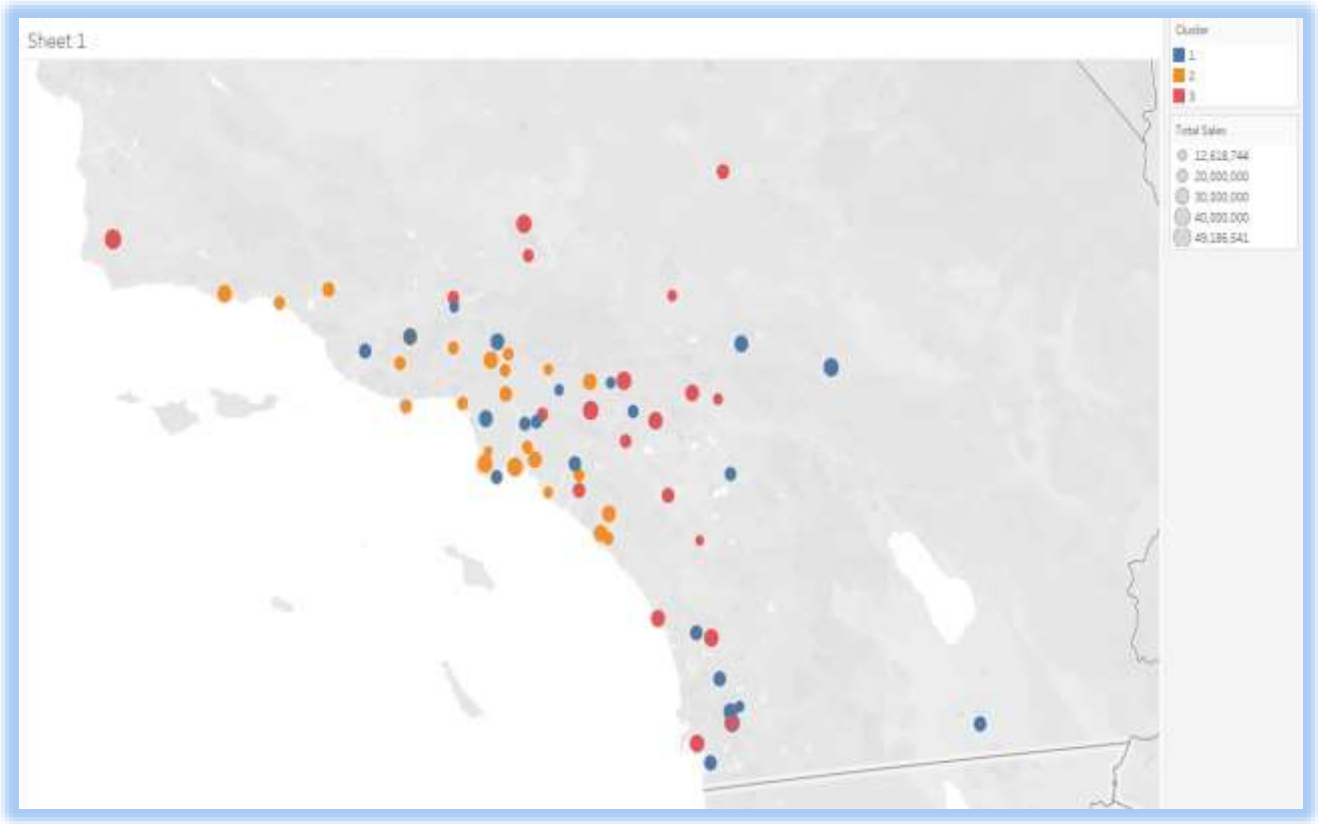
Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	pct_Dry_Grocery	pct_Diry	pct_Frozen_Food	pct_Meat	pct_Produce	pct_Floral	pct_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

	pct_Bakery	pct_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

Cluster 3 contains more negative values than cluster 1 and cluster 2, besides, cluster 3 also contains the most positive value for General merchandise. This indicates the variations in the 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.



Clustering - Dinesh | Tableau Public

https://public.tableau.com/profile/dinesh7244#!/vizhome/Clustering_16150690780430/Sheet1?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.)

The following methods were used: Decision Tree, Forest and Boosted models. The Training set was 80% and the validation data was 20%.
Following is the result of the model comparison:

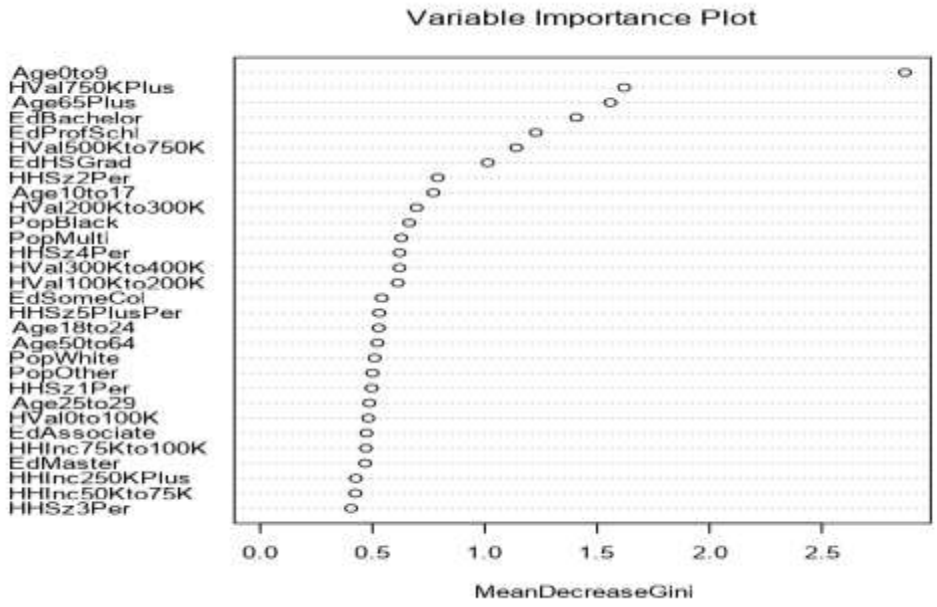
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
boosted	0.9706	0.9730	1.0000	0.9667	0.9524
forest	1.0000	1.0000	1.0000	1.0000	1.0000
Decision_Tree	0.8088	0.7845	0.7059	0.9333	0.7143

It can be concluded from the above that the Forest model is the best fit.

From the forest model, we can conclude that following are the three most important variables.
Age0to9 ,Hval750KPlus,Age65Plus



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

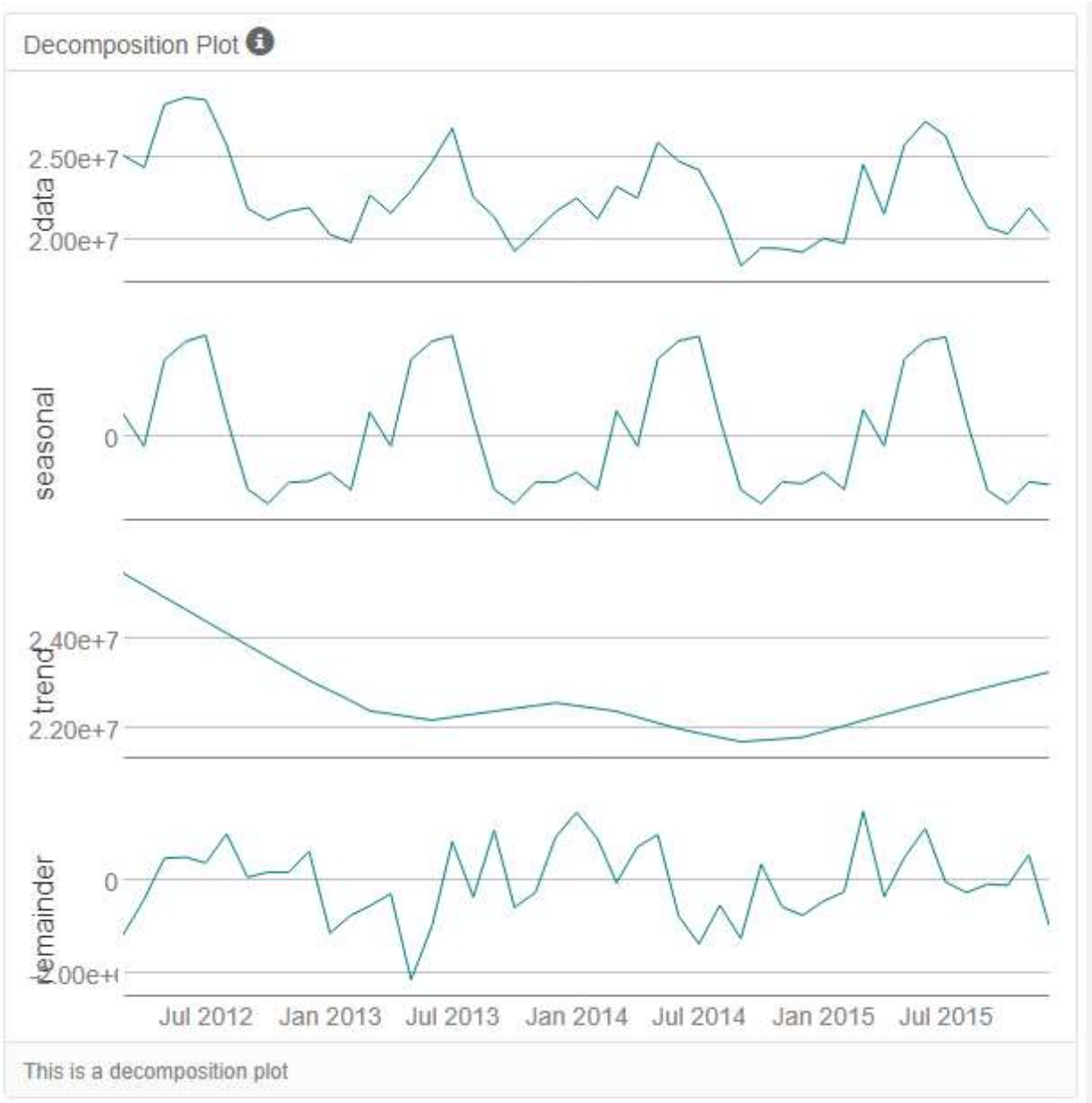
Store Number	Segment
S0086	1
S0087	2
S0088	1
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

Store	cluster
S0086	1
S0087	2
S0088	1
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?

The following can be concluded using the below decomposition plot:



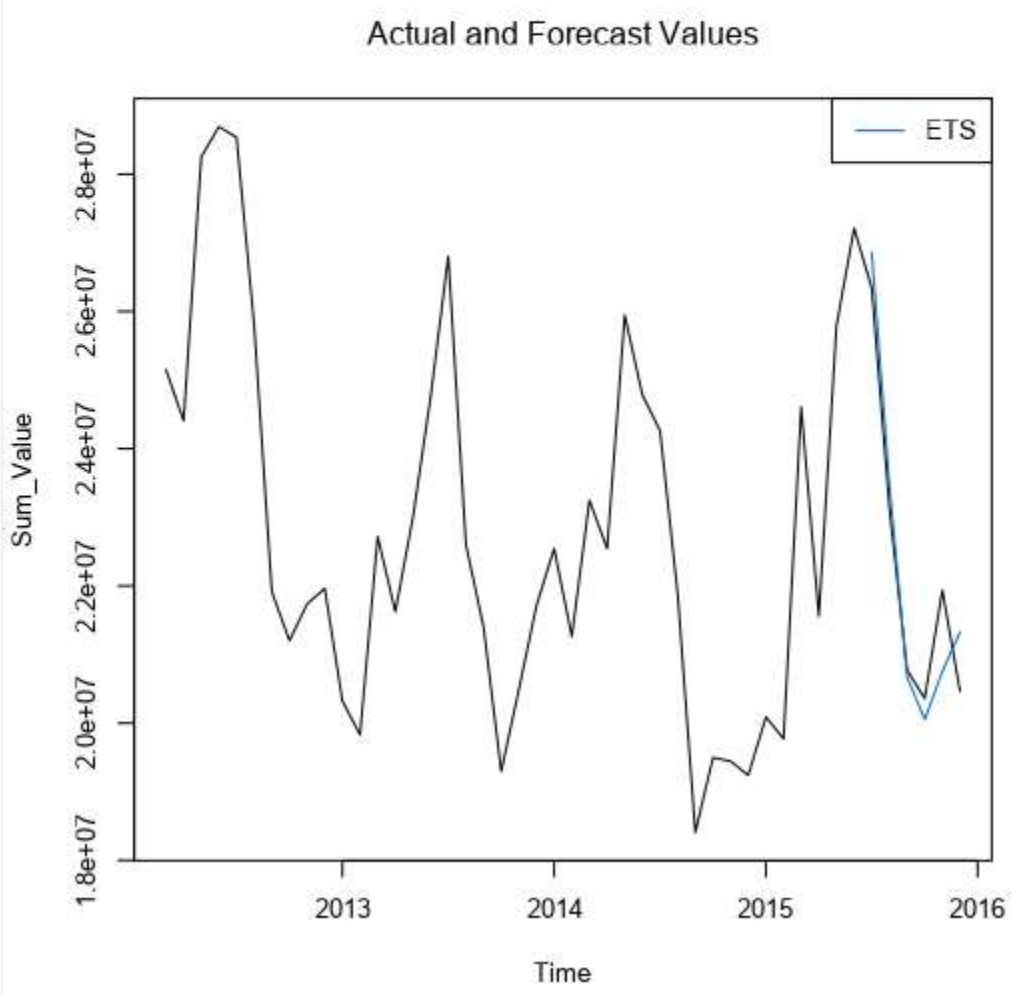
a) Seasonal component is available and must be applied multiplicatively.
b)Trend component is non-linear and trends upward towards the end and should not be applied.
c) Reminder component is decreasing hence applied multiplicatively.

For ARIMA we can use a (0,1,1) (0,1,1)12 since there is a seasonal component in the time series.

Record	Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
1	ETS	-21,581.1252	663,707.1529	553,511.4848	-0.0437	2.5135	0.3257	[Null]
2	Arima	-604,232.3185	1,050,239.1965	928,412.0306	-2.6156	4.0942	0.5463	[Null]

Accuracy Measures:

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

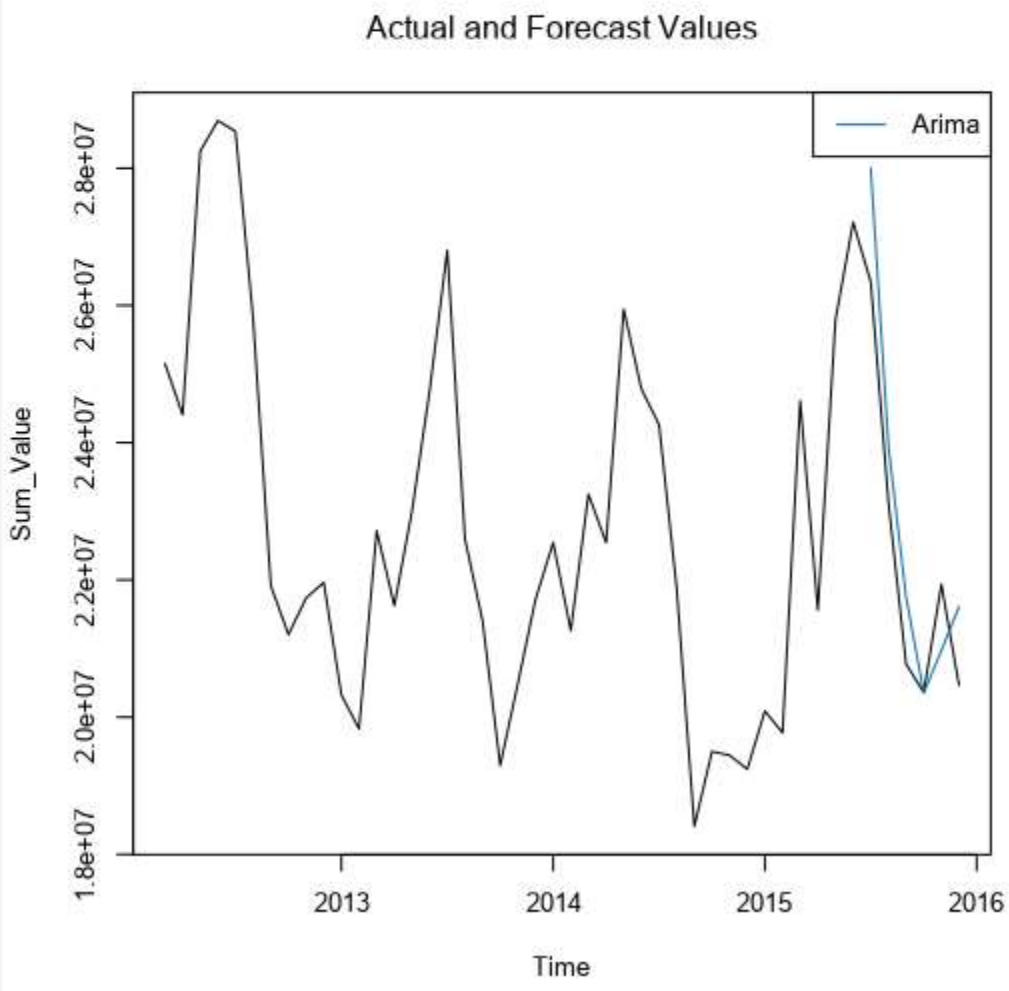


3

Accuracy Measures:

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

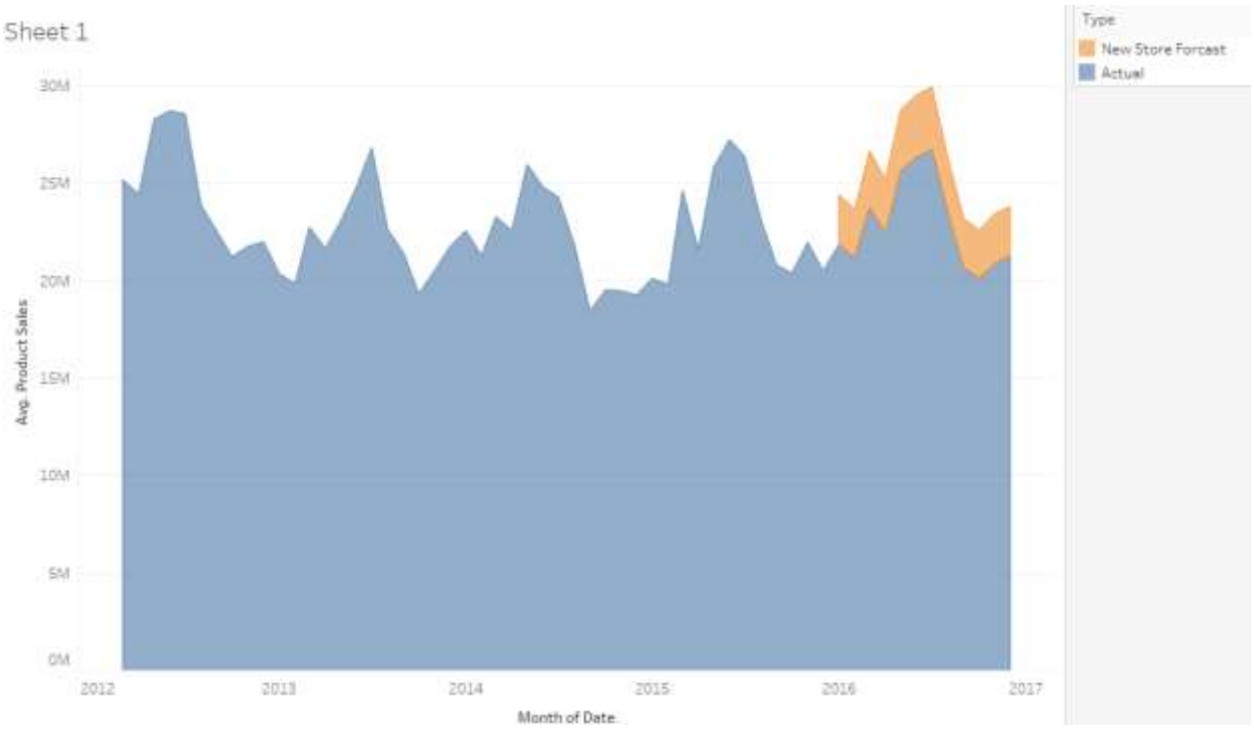
4



It can be observed from the above graphs the ETS model has higher accuracy than ARIMA model. The ME, RMSE and MAE values of the ETS model are lower than the ARIMA model. Also ETS model has a better fit on the graph.

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.

Date	Existing Stores Forcast	New Store Forcast	Combined
1/1/2016	21829060.03	2563357.91	24392417.94
2/1/2016	21146329.63	2483924.73	23630254.36
3/1/2016	23735686.94	2910944.15	26646631.08
4/1/2016	22409515.28	2764881.87	25174397.15
5/1/2016	25621828.73	3141305.87	28763134.59
6/1/2016	26307858.04	3195054.20	29502912.24
7/1/2016	26705092.56	3212390.95	29917483.51
8/1/2016	23440761.33	2852385.77	26293147.10
9/1/2016	20640047.32	2521697.19	23161744.51
10/1/2016	20086270.46	2466750.89	22553021.36
11/1/2016	20858119.96	2557744.59	23415864.55
12/1/2016	21255190.24	2530510.81	23785701.05



Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.