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?
   3 clusters.
   The CHI and ARI indices are as follows:

Summary Statistics

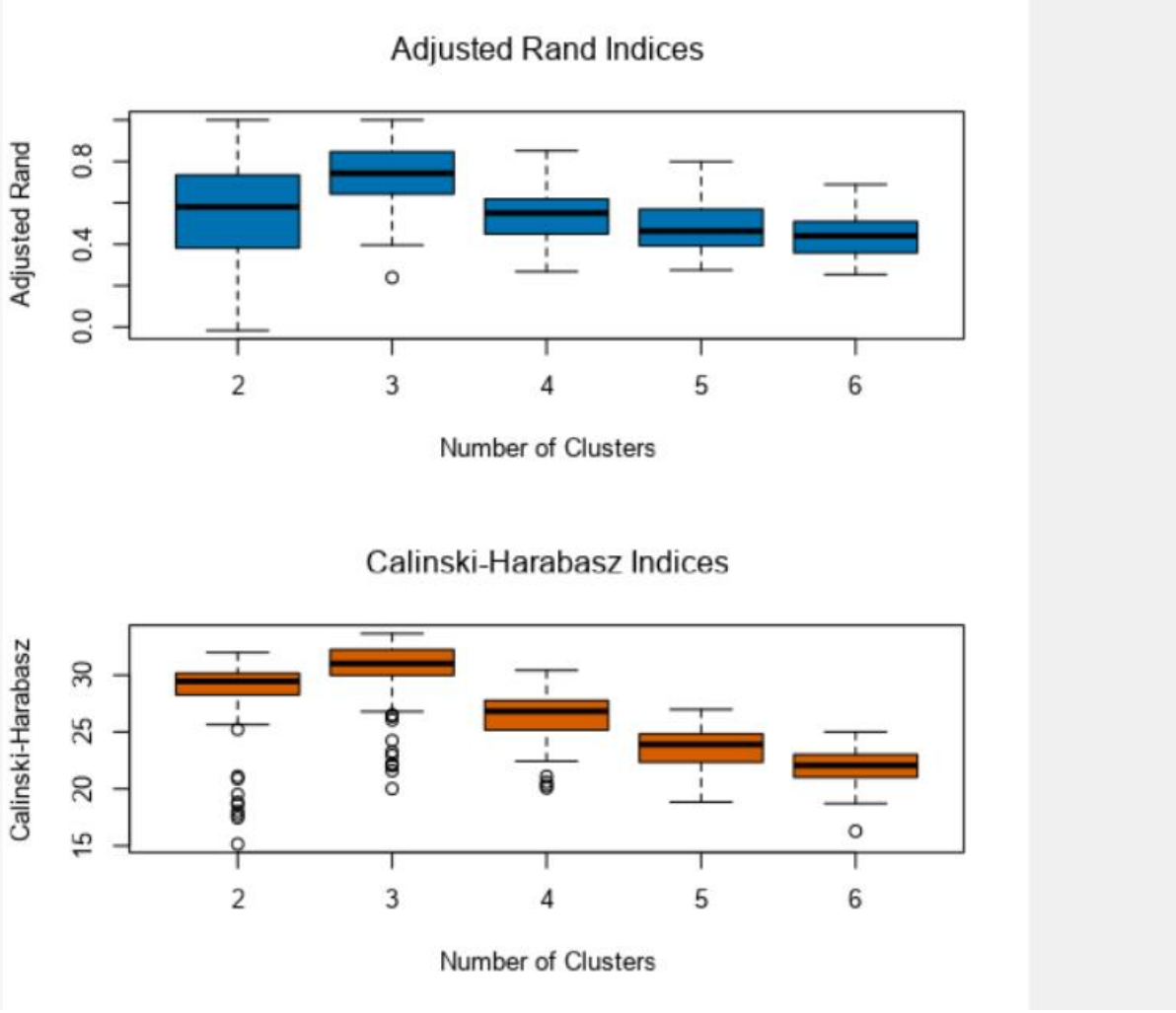
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.238908	0.26746	0.275161	0.254075
1st Quartile	0.389138	0.643526	0.451546	0.393179	0.361002
Median	0.579832	0.742946	0.550094	0.46327	0.440569
Mean	0.538248	0.716946	0.539436	0.480527	0.444128
3rd Quartile	0.734477	0.841627	0.618537	0.564177	0.507959
Maximum	1	1	0.851619	0.798934	0.689104

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.14927	20.01657	20.07469	18.84105	16.28411
1st Quartile	28.27367	30.07272	25.16346	22.35521	21.04521
Median	29.4511	31.00382	26.81884	23.89722	22.0471
Mean	28.40735	30.28555	26.35179	23.56802	21.93001
3rd Quartile	30.16162	32.23534	27.76016	24.82346	22.99673
Maximum	31.9781	33.63781	30.41396	26.97019	25.00769

   The inidices are high for a 3 cluster model. Also, the median is the highest when it comes to 3 cluster models.
   The boxplots are as follows:



Hence, I am using 3 clusters as the optimal number of clusters

2. How many stores fall into each store format?
   Cluster 1 has 23 stores
   Cluster 2 has 29 stores
   Cluster 3 has 33 stores

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Cluster 1 has the least number of stores, then cluster 2 and cluster 3 has the highest.

Summary Report of the K-Means Clustering Solution clustering_model

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Sum_Dry_Grocery + Sum_Dairy + Sum_Frozen_Food + Sum_Meat + Sum_Produce + Sum_Floral + Sum_Deli + Sum_Bakery + Sum_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

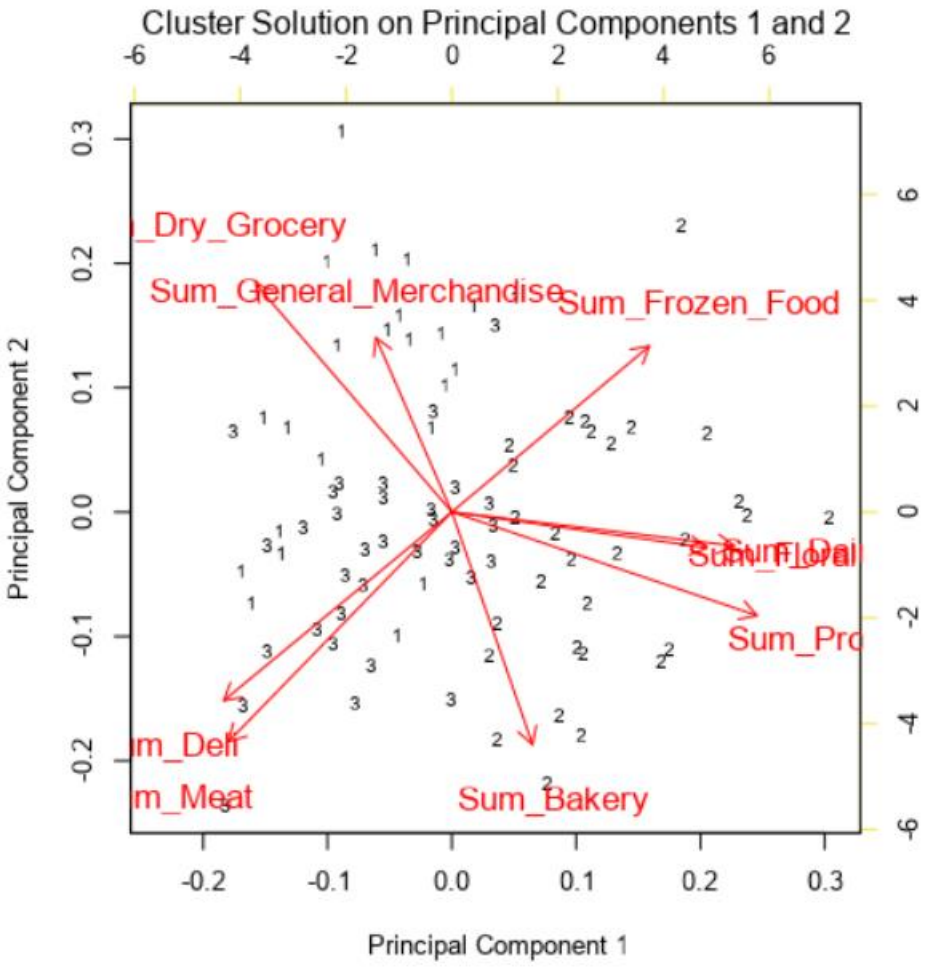
Cluster Information:

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

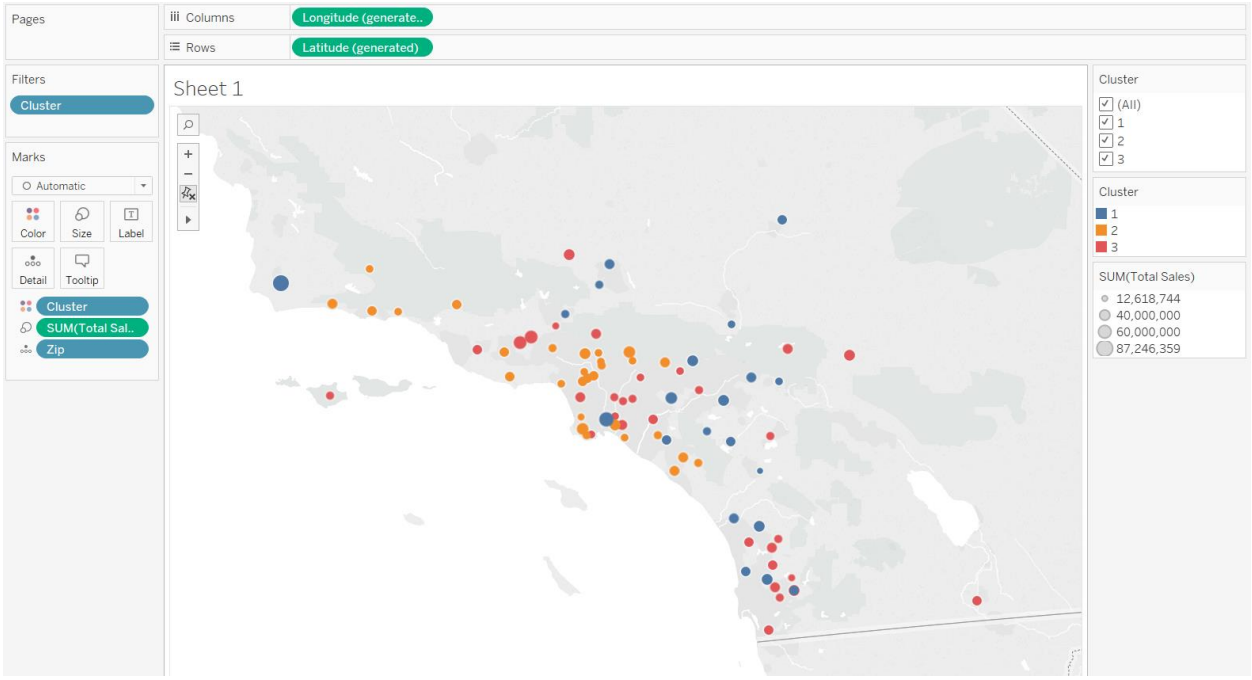
Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Sum_Dry_Grocery	Sum_Dairy	Sum_Frozen_Food	Sum_Meat	Sum_Produce	Sum_Floral	Sum_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

	Sum_Bakery	Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389



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.



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 compared boosted, random forest and decision trees and I found that boosted model had the highest accuracy and f1 score. It also had good PPV and NPV values as compared to the rest

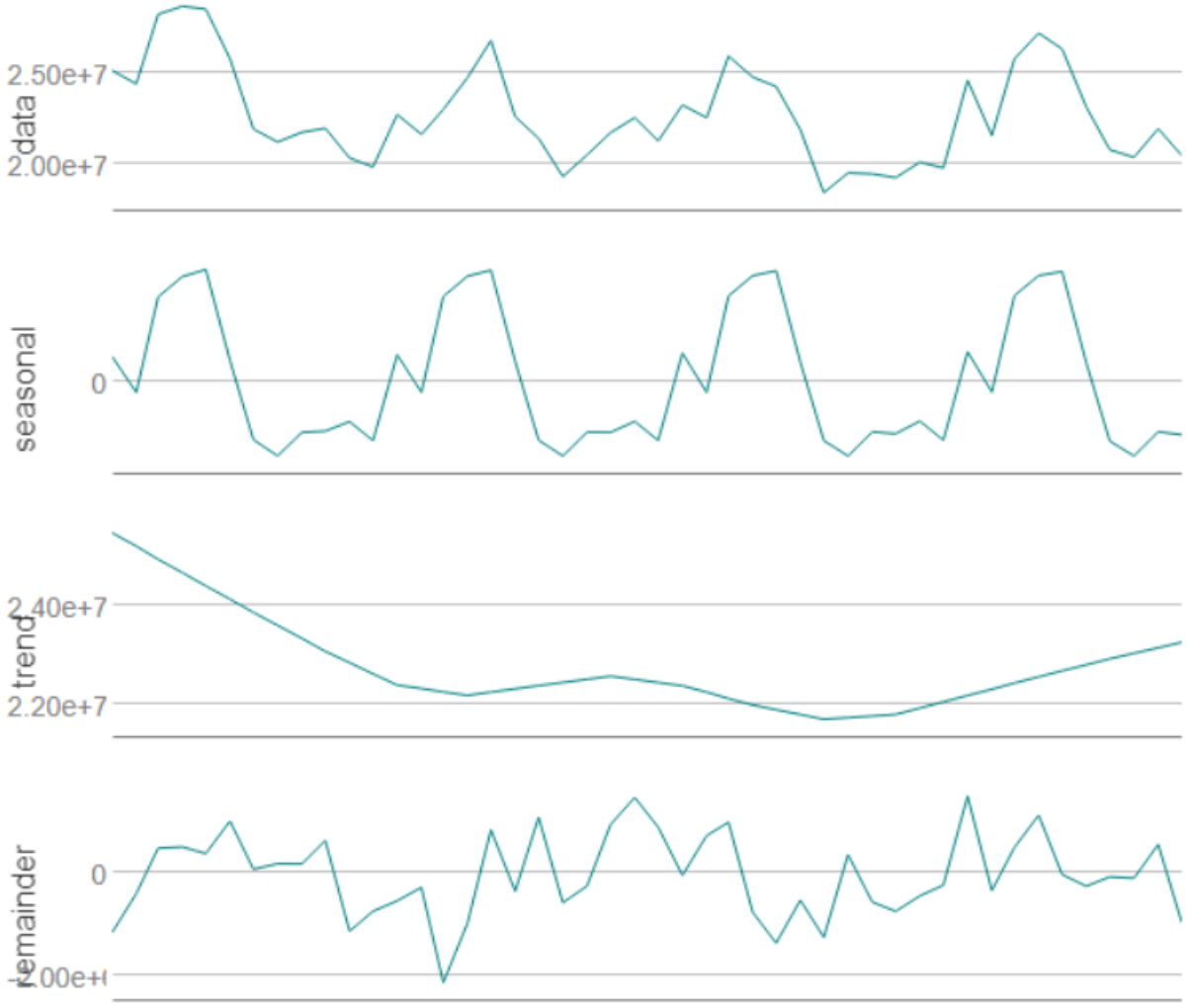
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

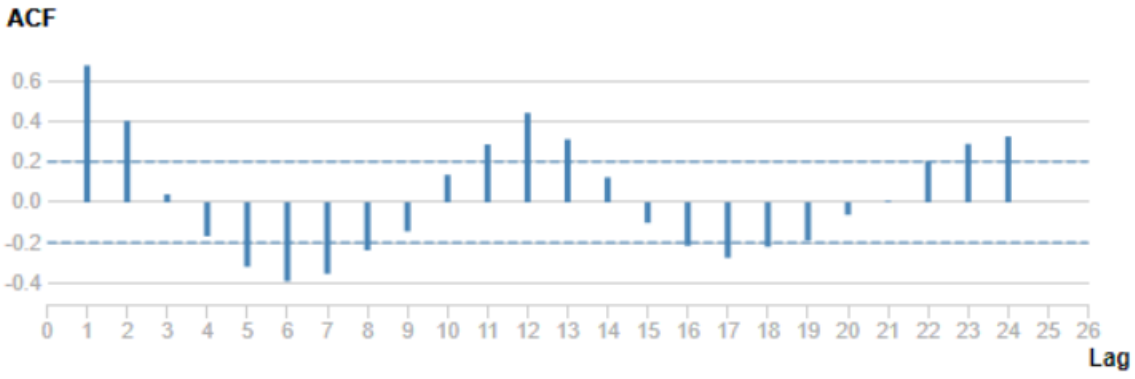
Task 3: Predicting Produce Sales

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?
For ETS I used an MNM model. For ARIMA, I used ARIMA(1,0,0)(1,1,0)[12]
Using the TS Plot tool:

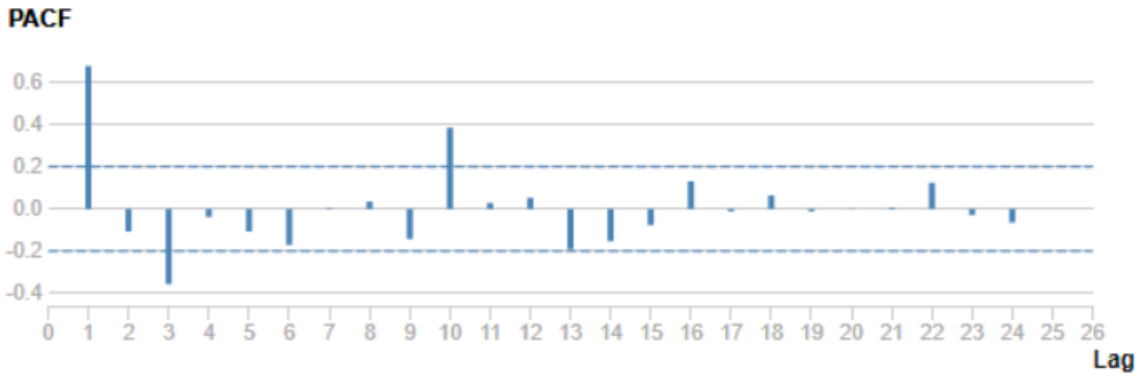
Decomposition Plot



Autocorrelation Function Plot



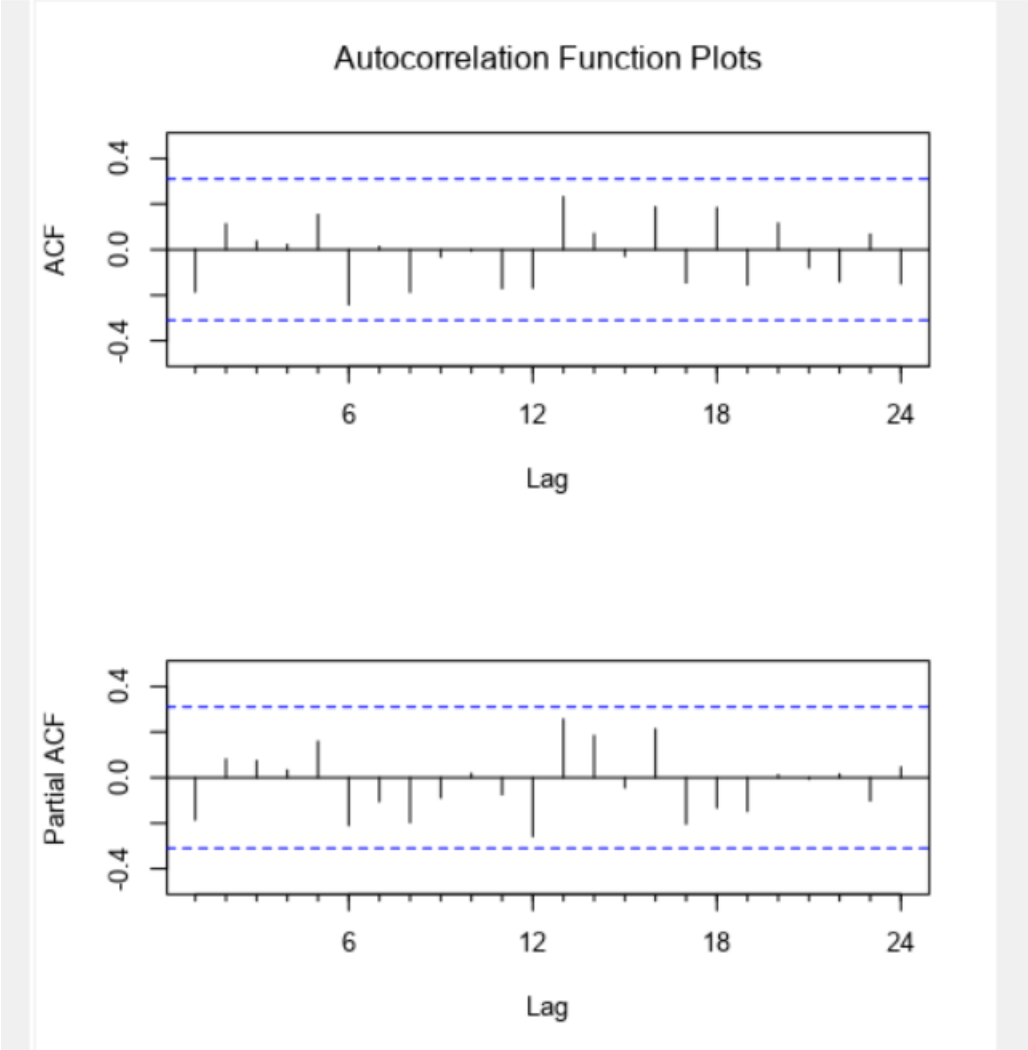
Partial Autocorrelation Function Plot



With respect to the decomposition plot, you can see multiplicative seasonality and error with almost no trend. So I am using a MNM ETS model.

Record	Report
1	**Summary of ARIMA Model Arima**
2	Method: ARIMA(1,0,0)(1,1,0)[12]
3	Call: auto.arima(Sum_Produce)
4	Coefficients: ar1 sar1 Value 0.79852 -0.700441 Std Err 0.126448 0.140181
5	sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224
6	Information Criteria: AIC AICc BIC 880.4445 881.4445 884.4411
7	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
8	Ljung-Box test of the model residuals: Chi-squared = 15.0973, df = 12, p-value = 0.23616

Plots

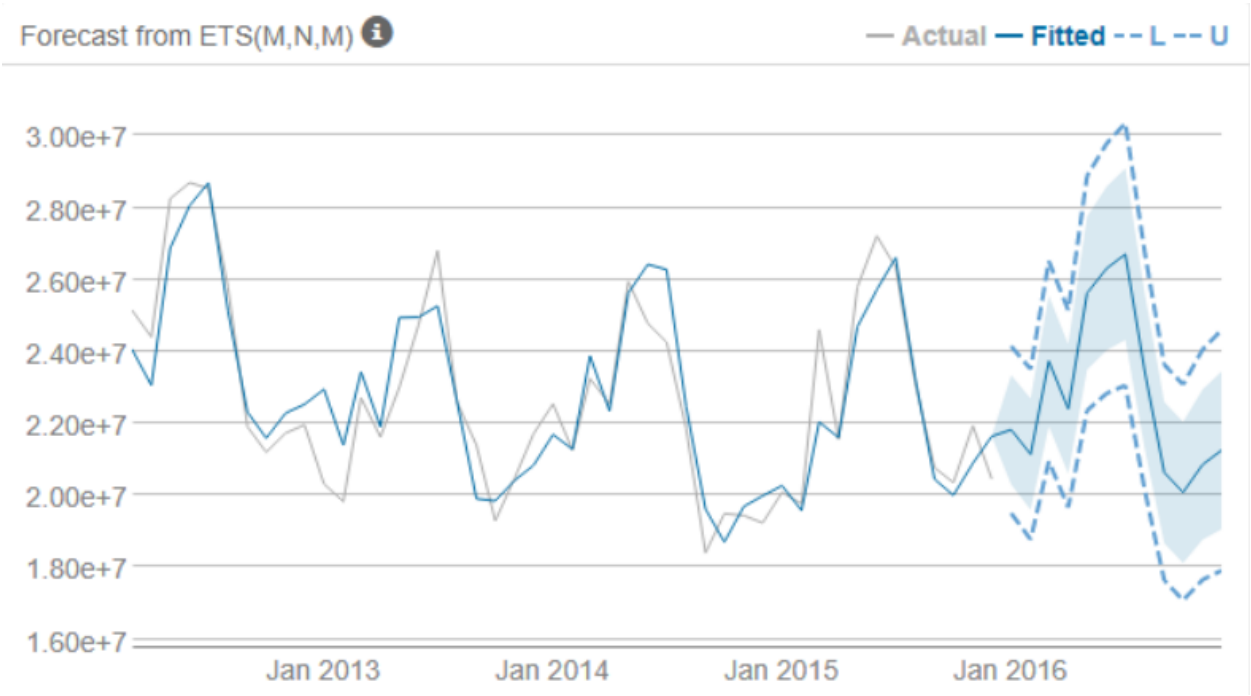

Autocorrelation Function Plots
ACF
0.4
0.0
-0.4
6
12
18
24
Lag
Partial ACF
0.4
0.0
-0.4
6
12
18
24
Lag

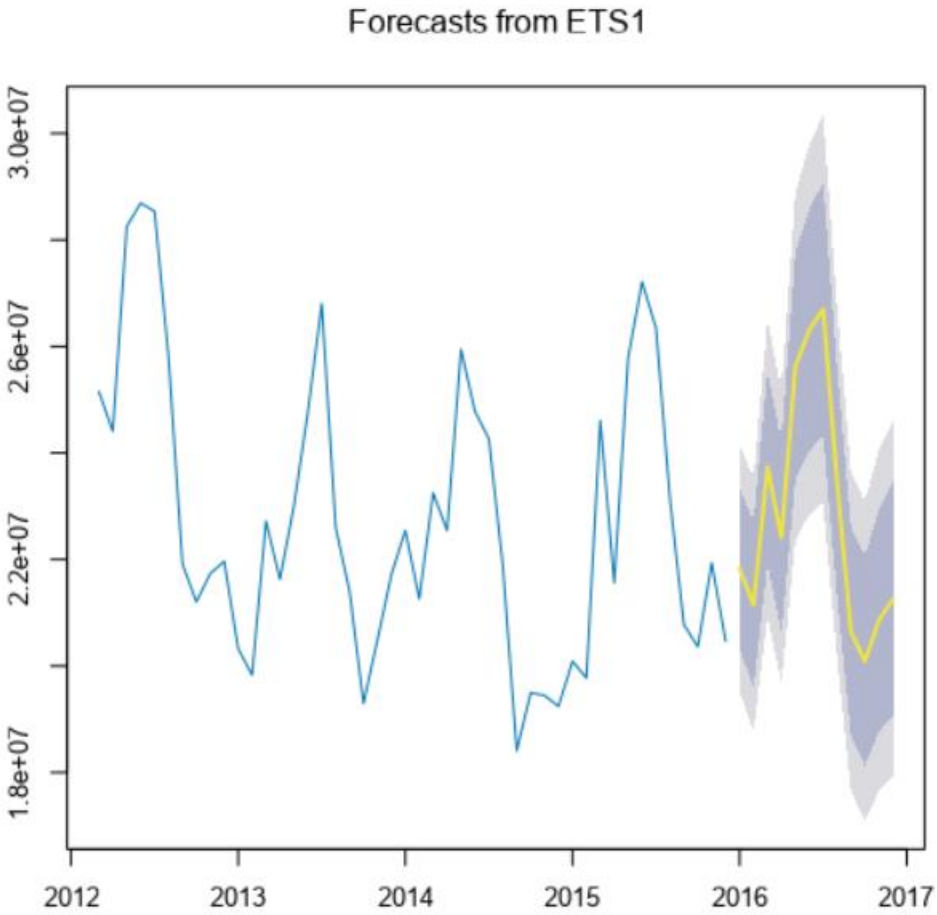

Actual vs. Forecast Values
Actual
Arima
2.80e+7
2.60e+7
2.40e+7
2.20e+7
2.00e+7
1.80e+7

Yu

You can see the actual v predicted values for the ETS model is better than the arima model. I chose ETS model to forecast sales.

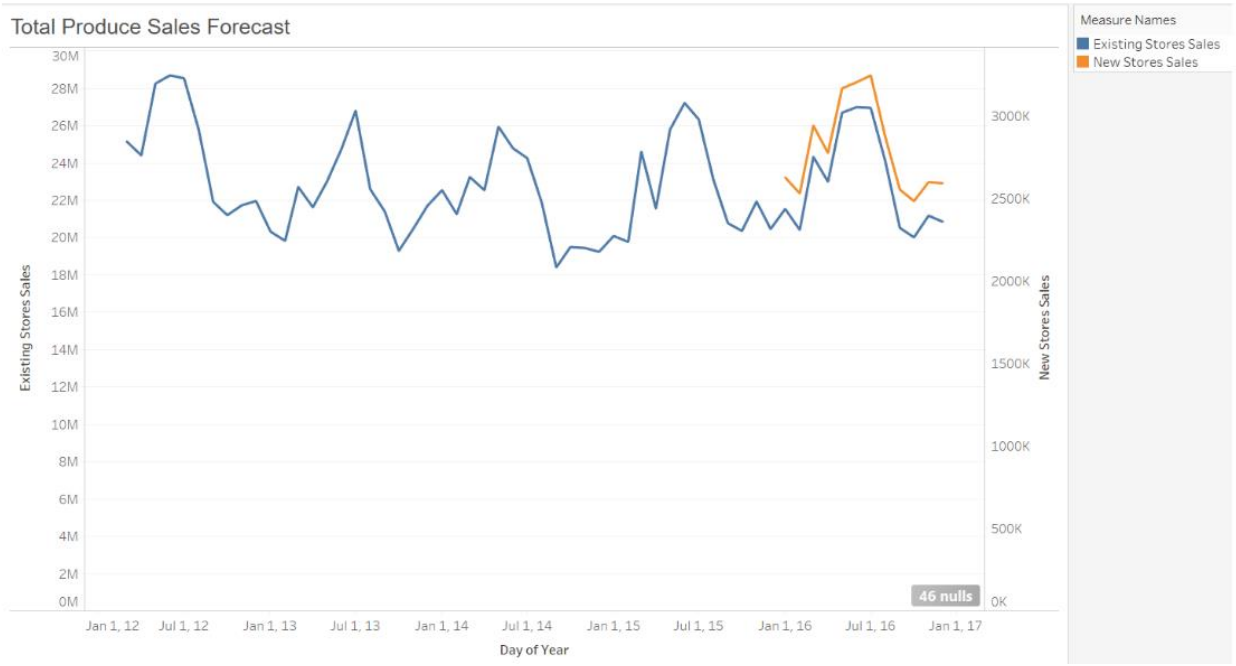
Forecast from ETS model





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.

Record	Year	Month	Forecast_Integer	New_Stores_Sales
1	2016	1	21829060	2603262
2	2016	2	21146330	2508878
3	2016	3	23735687	2989458
4	2016	4	22409515	2849287
5	2016	5	25621829	3224711
6	2016	6	26307858	3269623
7	2016	7	26705093	3288334
8	2016	8	23440761	2937302
9	2016	9	20640047	2606592
10	2016	10	20086270	2536270
11	2016	11	20858120	2631293
12	2016	12	21255190	2586562



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.