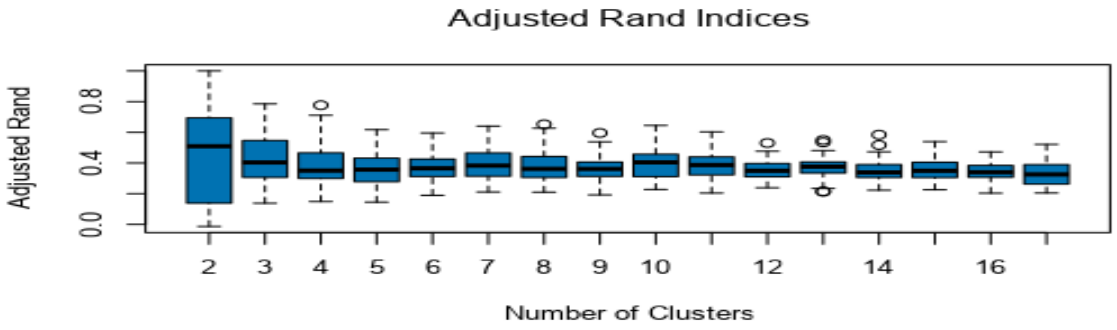
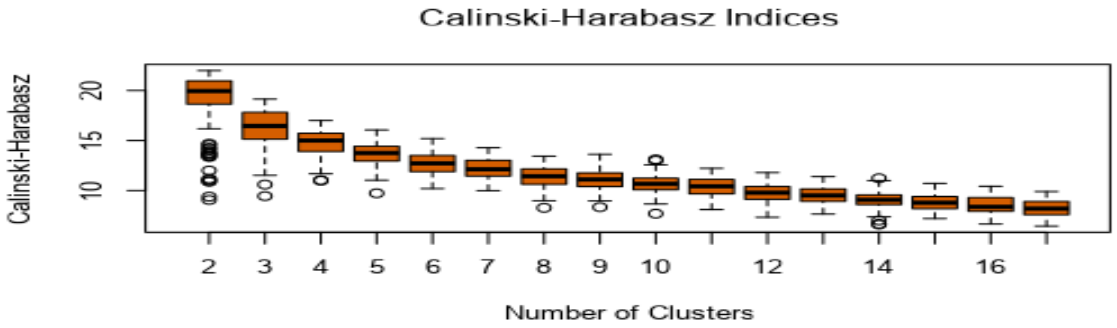
Project: Predictive Analytics Capstone

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?
   I performed the K-means clustering algorithm for 2-17 clusters. I compared the AR and CH indices. According to these indices, the higher the better. According to the box plots below, AR and CH are both highest when the clustering algorithm is set to 2 clusters. However, I notice an abnormally large interquartile range for the Adjusted Rand with two clusters, so I will choose to use three. I also checked the tables to confirm what I visualized.





2. How many stores fall into each store format?
   23 stores fall into cluster 1, 29 stores fall into cluster 2, and 33 stores fall into cluster 3.

Summary Report of the K-Means Clustering Solution KMeans3

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Pct_Sum_Dry_Grocery + Pct_Sum_Dairy + Pct_Sum_Frozen_Food + Pct_Sum_Meat + Pct_Sum_Produce + Pct_Sum_Floral + Pct_Sum_Deli + Pct_Sum_Bakery + Pct_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

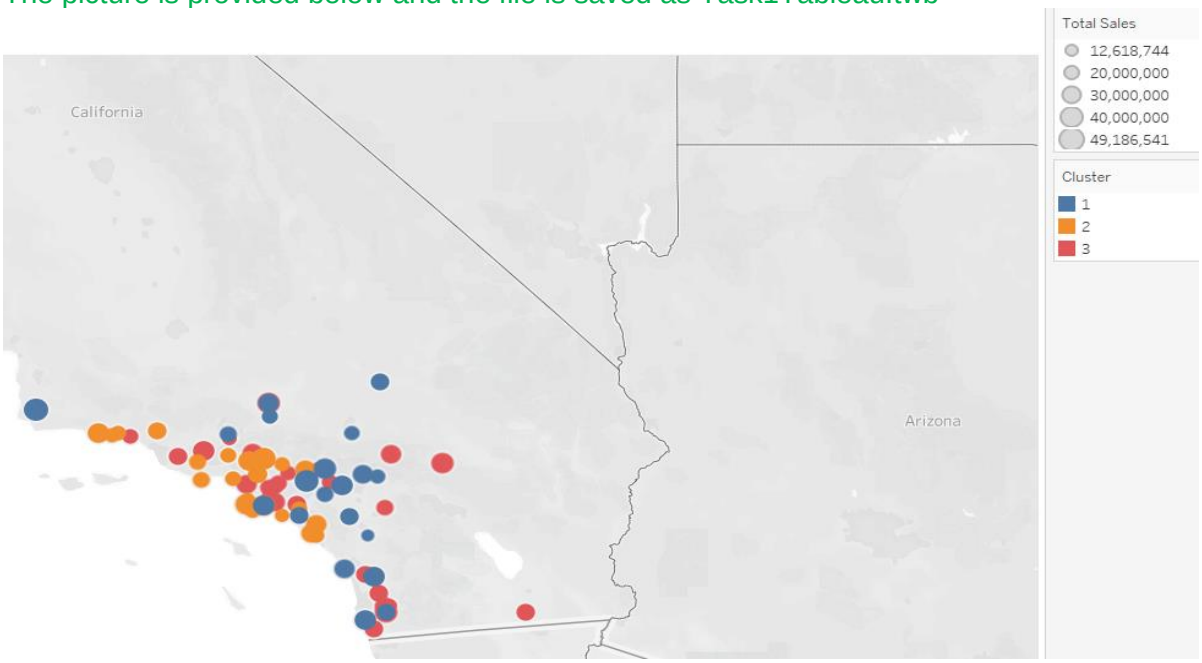
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Looking at the table below, the magnitude of the weights decide which variables have a bigger impact. Cluster 1 has a large magnitude and is different in direction for Dairy, Bakery, and General Merchandise. Cluster 2 has a large magnitude and is different in direction for Grocery, Dairy, Produce, and Floral. Cluster 3 has a large magnitude and is different in direction for Deli.

	Pct_Sum_Dry_Grocery	Pct_Sum_Dairy	Pct_Sum_Frozen_Food	Pct_Sum_Meat	Pct_Sum_Produce	Pct_Sum_Floral	Pct_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

	Pct_Sum_Bakery	Pct_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.
   The picture is provided below and the file is saved as Task1Tableau.twb



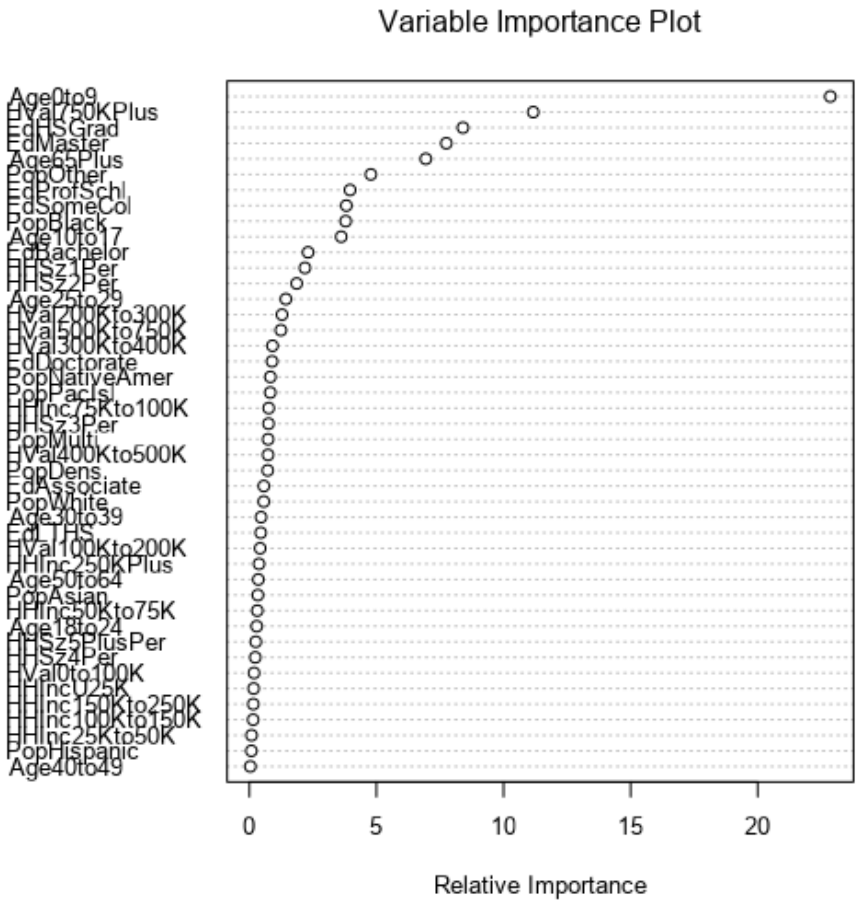
Task 2: Formats for New Stores

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 results from predicting using the Decision Tree, Random Forest, and Boosted Model; the results are listed, respectively. The accuracies are 0.7059, 0.8235, and 0.8235 as shown below.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DecisionTree	0.7059	0.7685	0.7500	1.0000	0.5556
ForestModel	0.8235	0.8426	0.7500	1.0000	0.7778
BoostedModel	0.8235	0.8889	1.0000	1.0000	0.6667

The Boosted Model and the Forest Model Perform similarly when looking at accuracy metrics. However, since the F1 score for the Boosted Model (0.8889) is larger than the F1 score for the Forest Model (0.8426), The Boosted Model will be used.
The Variable Importance plot shows Age0to9, HVal750kPlus, EdHSGrad, EdMaster, and Age65Plus as having the most impact.



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

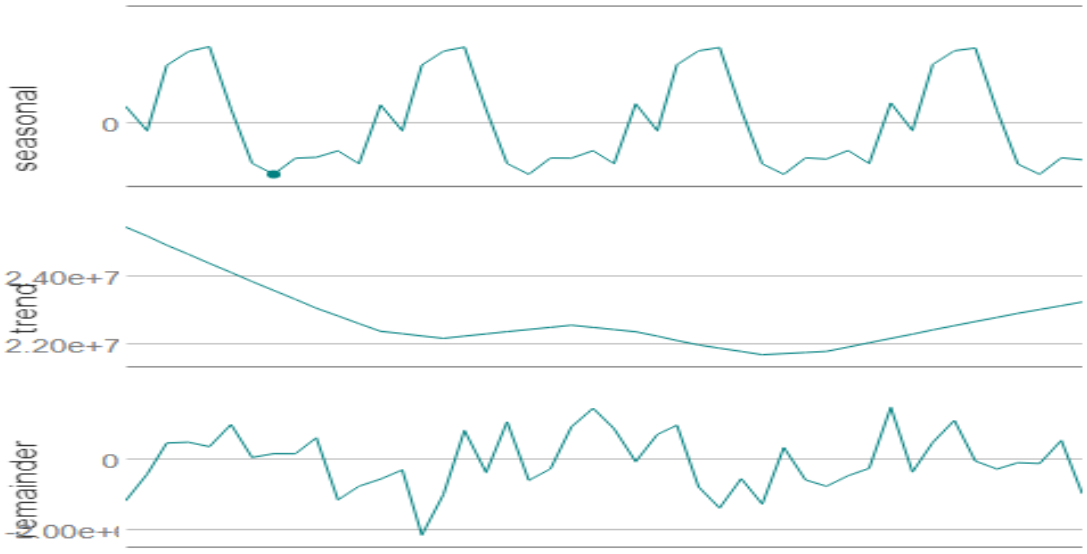
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2

S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

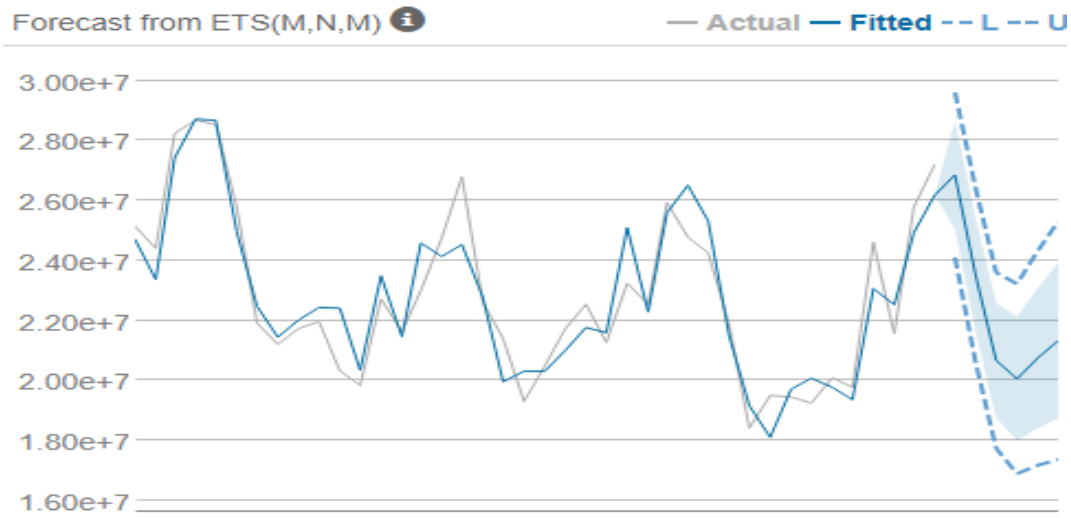
Task 3: Predicting Produce Sales

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?

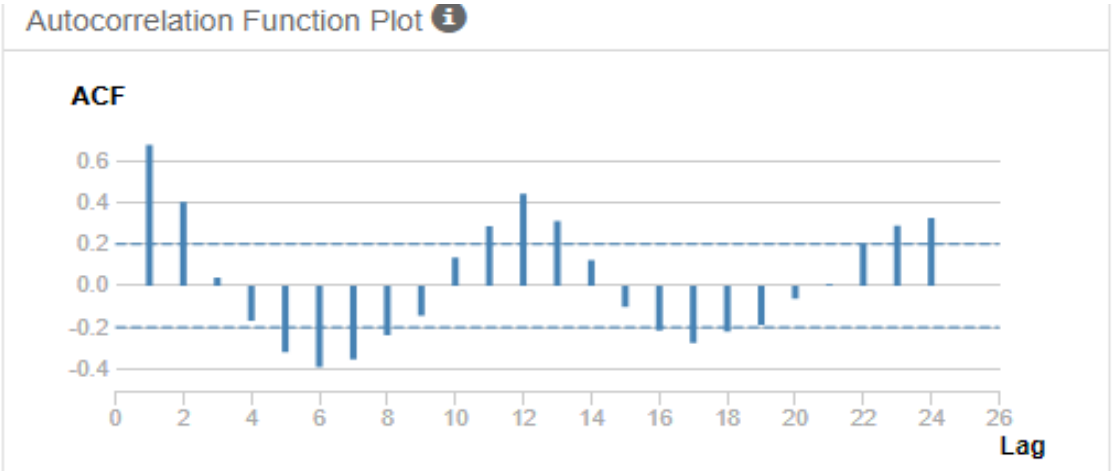
For existing stores, I used both the ETS and ARIMA models to find the best solution. For predicting the aggregate produce for the existing stores, I plotted the Decomposition plots to understand the trend, seasonality and error. Looking at the three plots below, it is apparent that there exists seasonality and the error appears to decrease over time. Since the trend curve slopes upward after a period of time, I will not use that. So, I will have seasonality multiplicatively, trend as none, and remainder multiplicatively giving an ETS(M,N,M).



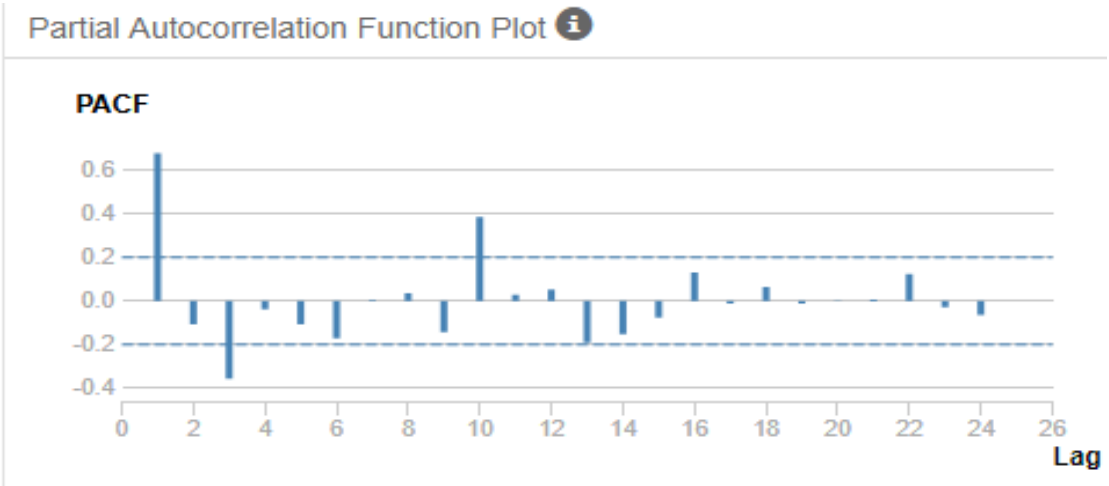
The ETS(M,N,M) model is shown below, providing predictions and confidence intervals. The AIC is 1279.4203.



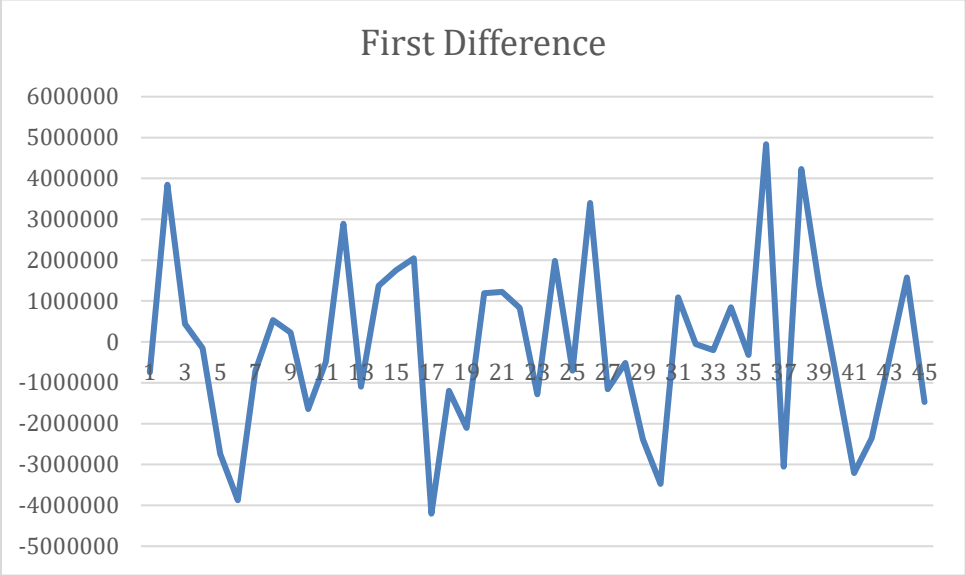
For the ARIMA model,
The Autocorrelation Plots and the Partial Autocorrelation Plots are shown below. The autocorrelation hints that the correlation is seasonal; also it is positive initially so the AR term should be used.



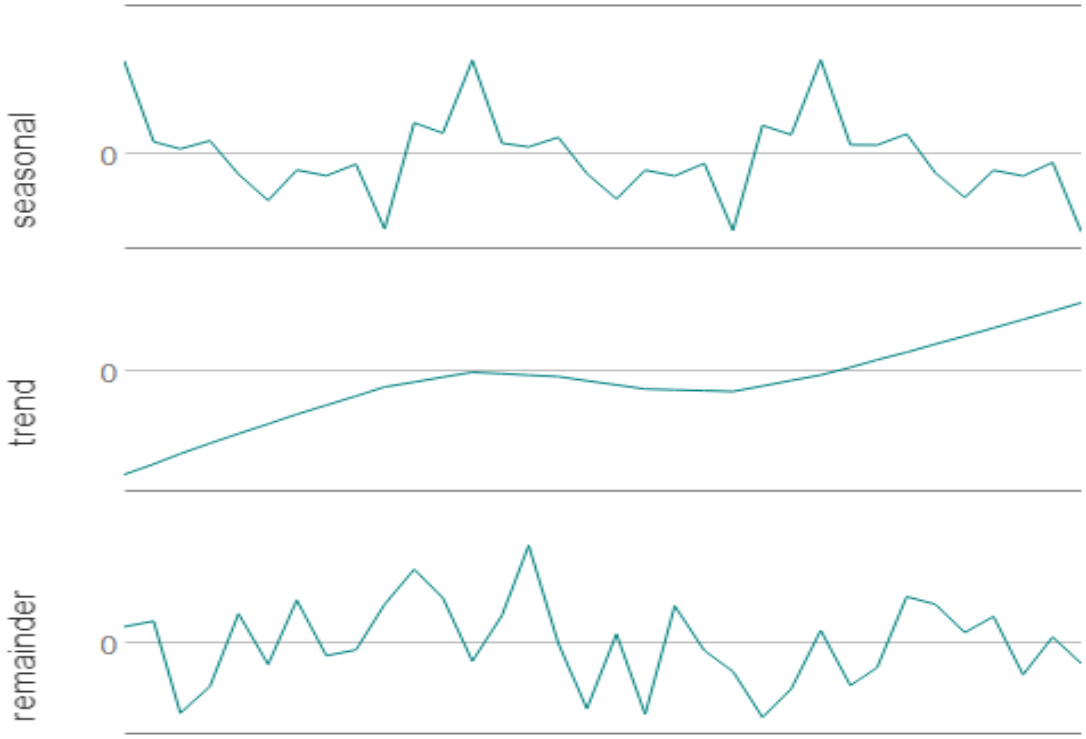
In the Partial Autocorrelation Plot, the correlation takes a huge drop after the first one, suggesting AR(1).



Using Excel, I produced a plot of the aggregate produce after taking the first difference. It is shown below. The first difference appears to have a mean of 0, so I(1).

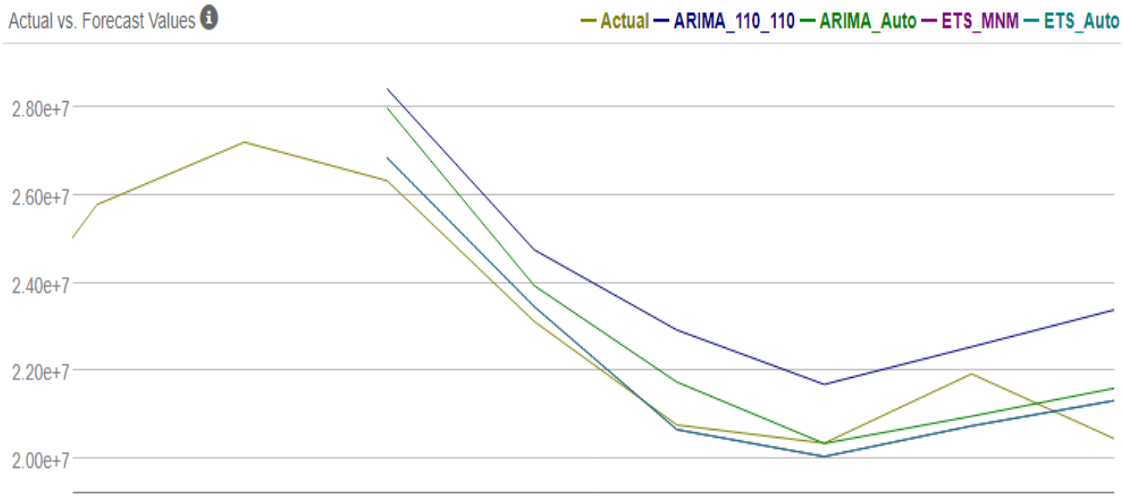


The MA term will be 0, so ARIMA(1,1,0). But, there does appear to be some seasonality, based on the Decomposition plots. Looking at the ACF plot above, there is a big correlation that slowly decreases. This means that an AR term should be used for the seasonal component. I took the seasonal difference and got these plots below in the Decomposition Plot.



Looking at the seasonal plot, there also appears to be an increasing trend. I will use a I(1) term in the seasonal model, with MA(0). Therefore, the final model will be ARIMA(1,1,0)(1,1,0)12.

I will compare 4 models: ETS(M,N,M), ETS_Auto, ARIMA(1,1,0)(1,1,0)12, and ARIMA_Auto. The results are shown below from the comparison tool.

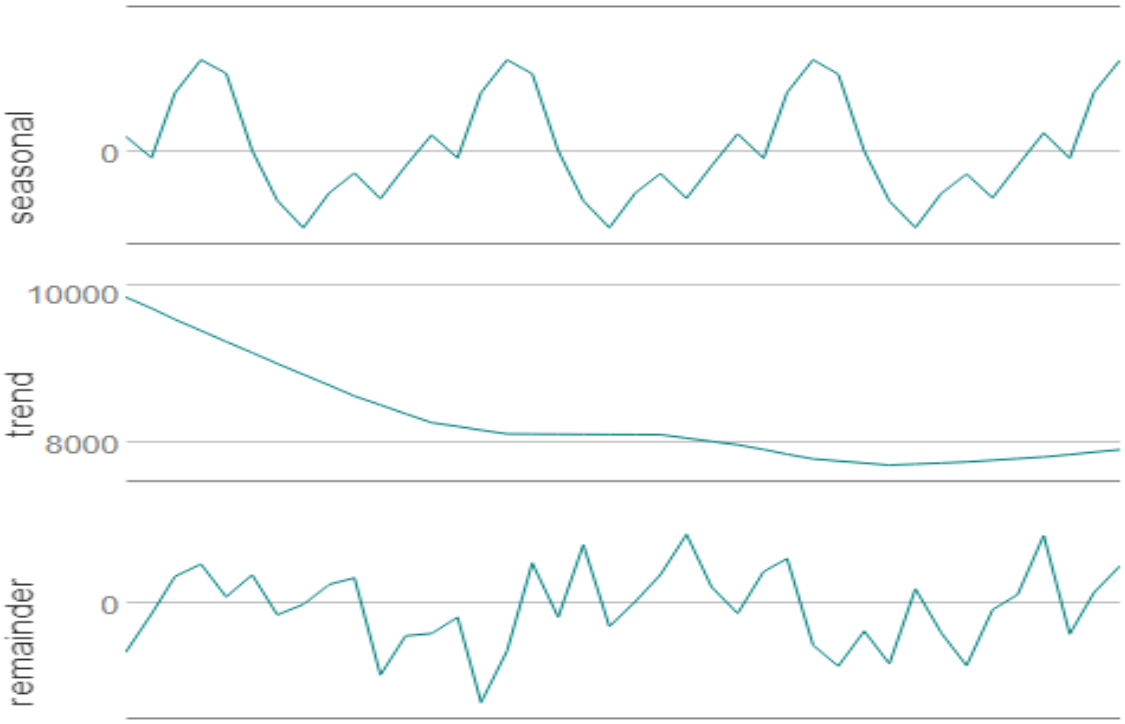


The metrics are provided below for analysis. The ARIMA model that was automatically generated is ARIMA(1,0,0)(1,1,0)12. The ETS model that was automatically generated is ETS(M,N,M). Across the board, ETS(M,N,M) performs better than all the other models, having the lowest scores. So, for existing stores, the ETS(M,N,M) will be used.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_110_110	-1795372.98	1935635.6	1795373	-8.1855	8.1855	1.0564
ARIMA_Auto	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463
ETS_MNM	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ETS_Auto	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

The AIC of the ETS(M,N,M)(1279.4203) is the highest to the AICs of ARIMA_Auto(880.4445), and ARIMA_110_110(848.8506). However, we do see better fitting with ETS(M,N,M), so that will be used.

I provided decomposition plots below for clusters 1, 2, and 3, respectively. It is apparent their decomposition plots are similar to the plots of the aggregate sum total stores. Therefore, ETS(M,N,M) will be used for the clusters also.

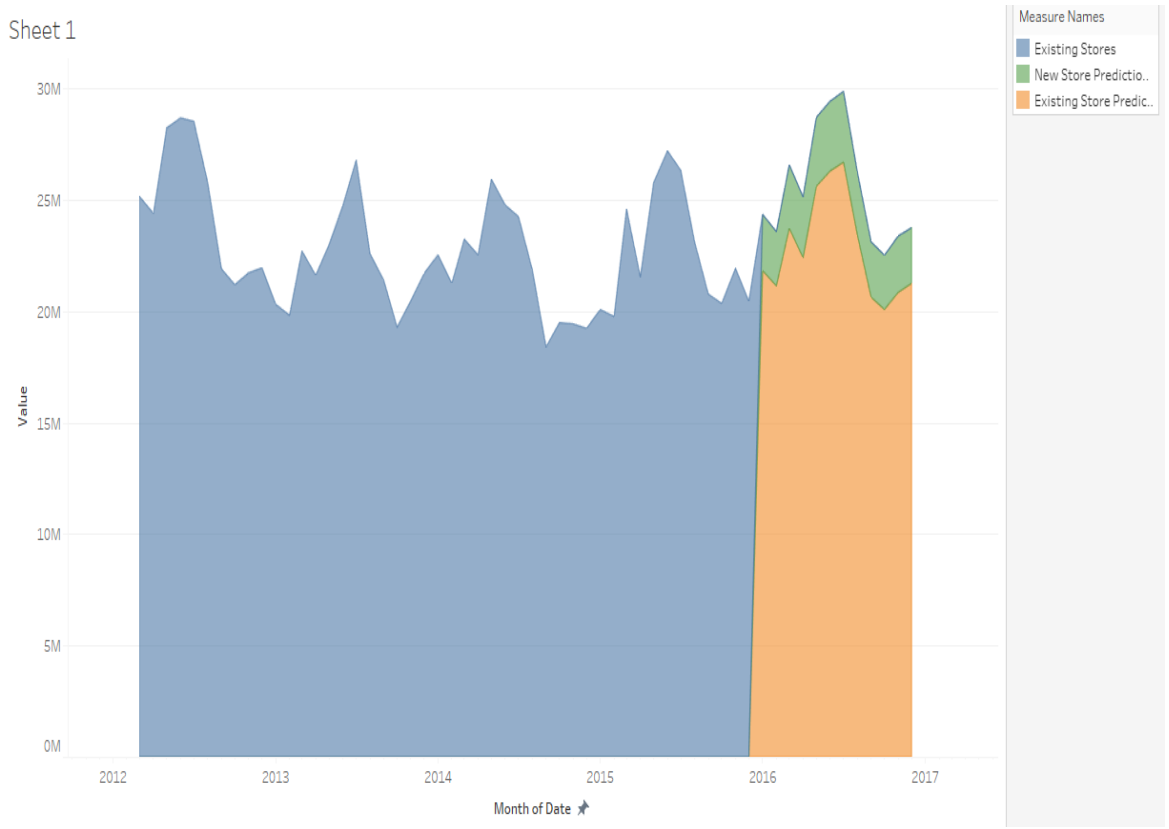


3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

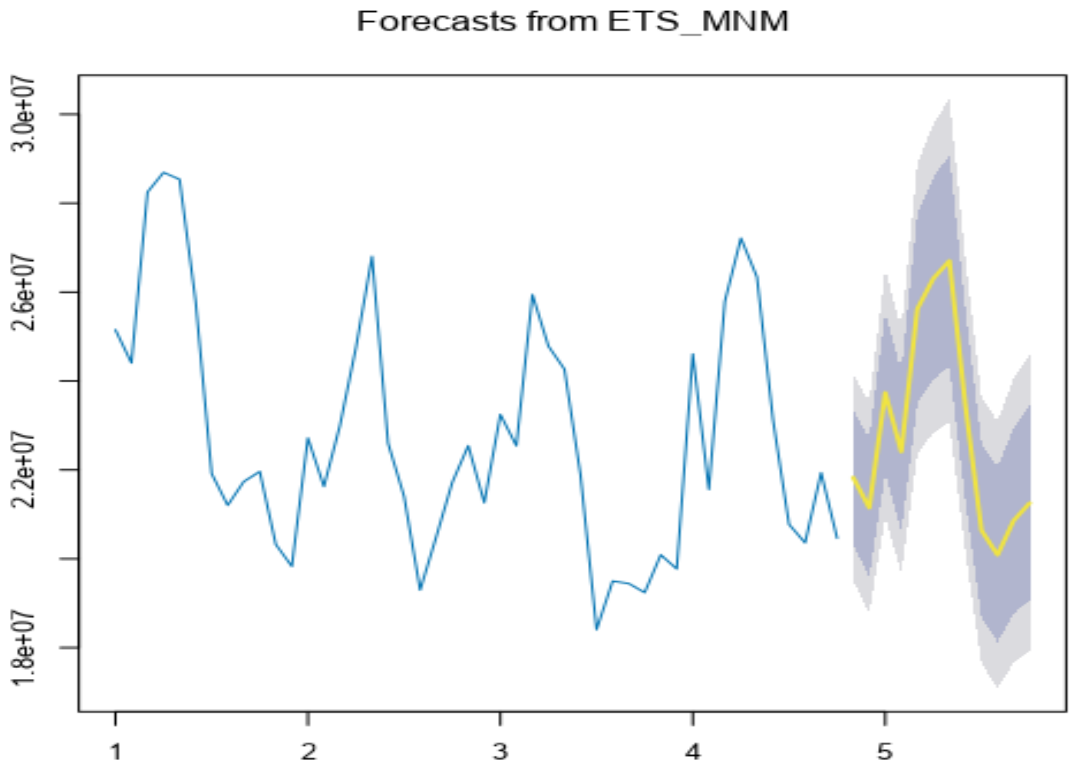
   Below is a table of the forecasts for the existing stores and new stores.

Month	New Stores	Existing Stores
January	2528128.347	21829060.03
February	2437826.838	21146329.63
March	2848173.808	23735686.94
April	2728391.659	22409515.28
May	3095318.243	25621828.73
June	3138254.046	26307858.04
July	3170183.677	26705092.56
August	2815908.177	23440761.33
September	2485729.145	20640047.32
October	2435157.478	20086270.46
November	2527336.262	20858119.96
December	2506574.213	21255190.24

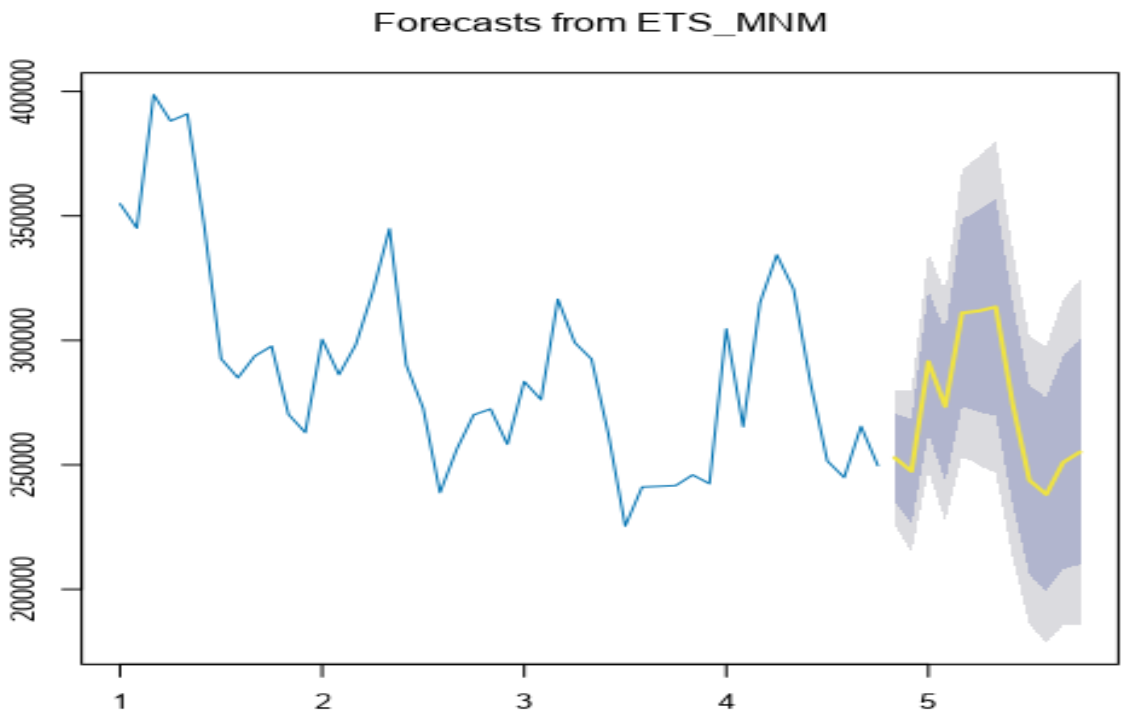
   Here is an area visualization of the forecasts.



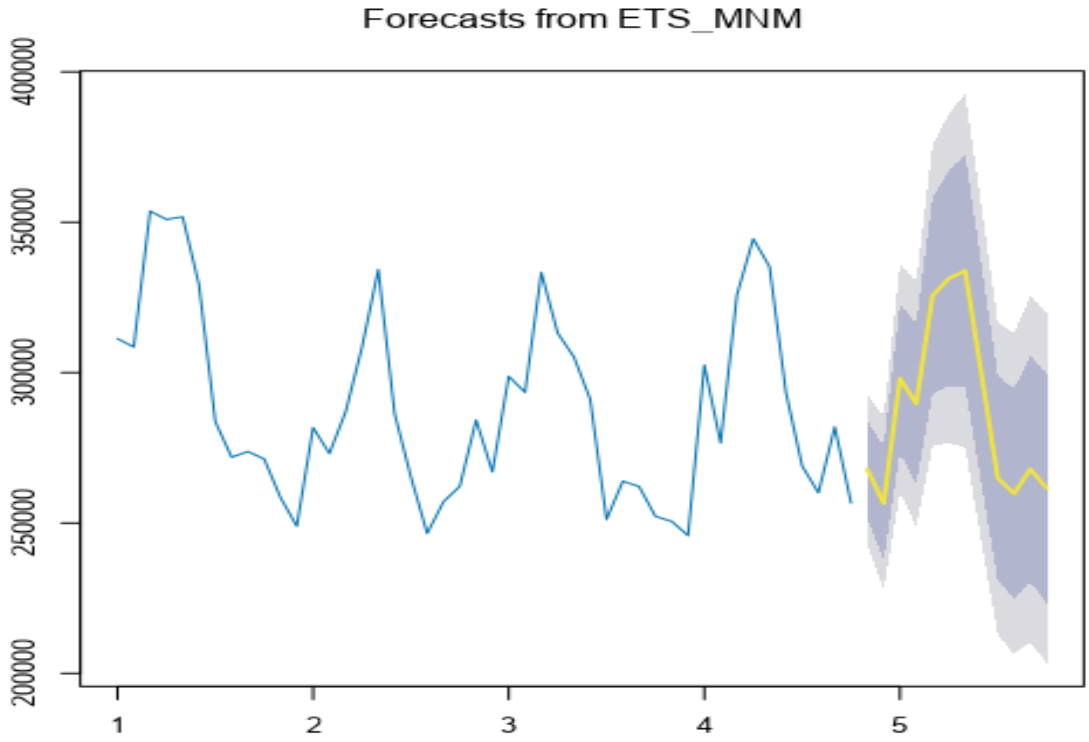
The forecast for the existing stores is shown below.



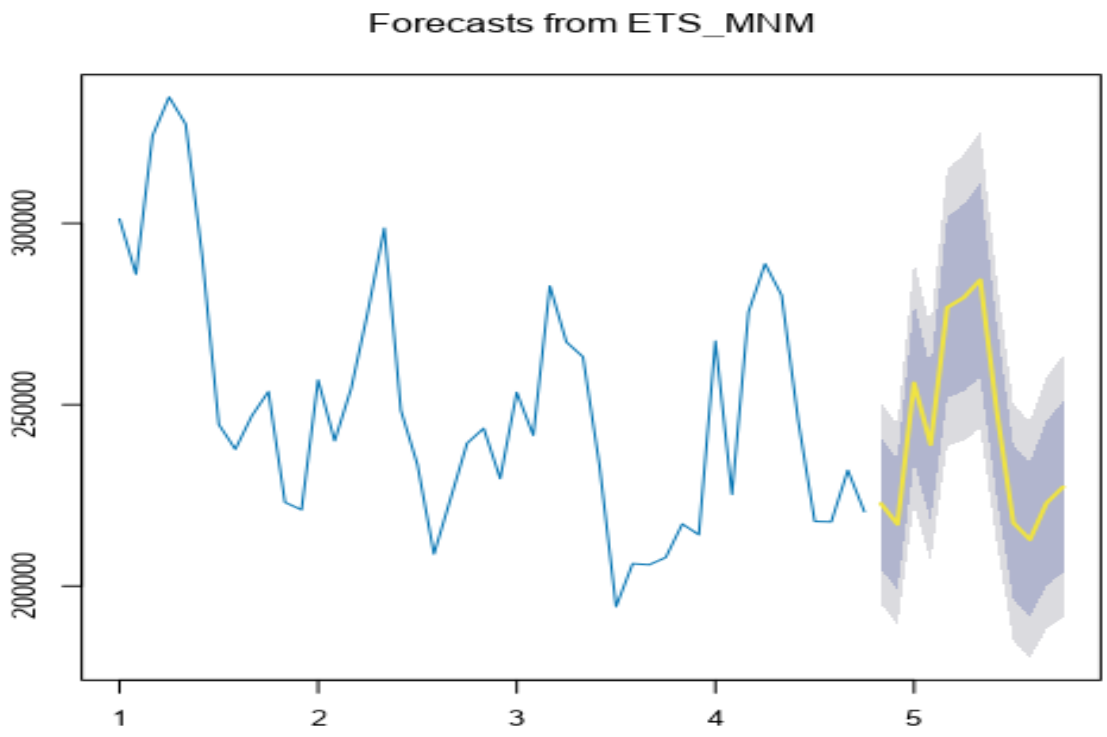
The forecast for the average store in cluster 1 is show below.



The forecast for the average store in cluster 2 is produced from ARIMA(1,1,0)(1,1,0)12.



The forecast for the average store in cluster 3 is shown below.



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.