Project: Predictive Analytics Capstone

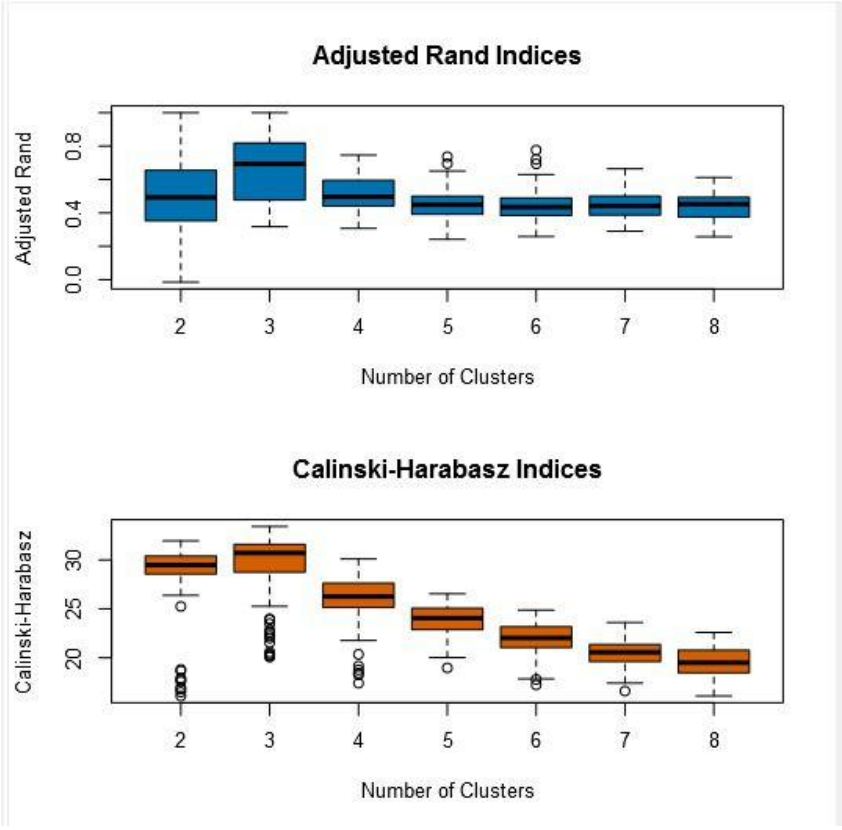
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

- Determine the optimal number of store formats based on sales data.
- Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
- Use only 2015 sales data.
- Use a K-means clustering model.
- Segment the 85 current stores into the different store formats.
- Use the StoreSalesData.csv and StoreInformation.csv files.

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

   Using K-mean clustering model , using year 2015 sales data, we can see that it is suggested that the optimal number of store format would **3 different formats**.
   We can see from K-mean centroid diagnostics, Adjusted Rand Indices and Calinski Harabasz indices shows highest mean and average in 3 cluster.
   Therefore , we will select 3 as optimal number of store format



2. How many stores fall into each store format?

Report

Summary Report of the K-Means Clustering Solution X

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + X.Sales.Dry.Grocery + X.Sales.FrozenFood + X.Sales.Dairy + X.Sales.Meat + X.Sales.Produce + X.Sales.Floral + X.Sales.Deli + X.Sales.Bakery + X.Sales.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.

	X.Sales.Dry.Grocery	X.Sales.FrozenFood	X.Sales.Dairy	X.Sales.Meat	X.Sales.Produce	X.Sales.Floral	X.Sales.Deli
1	0.327833	-0.389209	-0.761016	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.345898	0.702609	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.032704	-0.087039	0.48698	-0.53665	-0.538327	0.64952

	X.Sales.Bakery	X.Sales.General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Plots

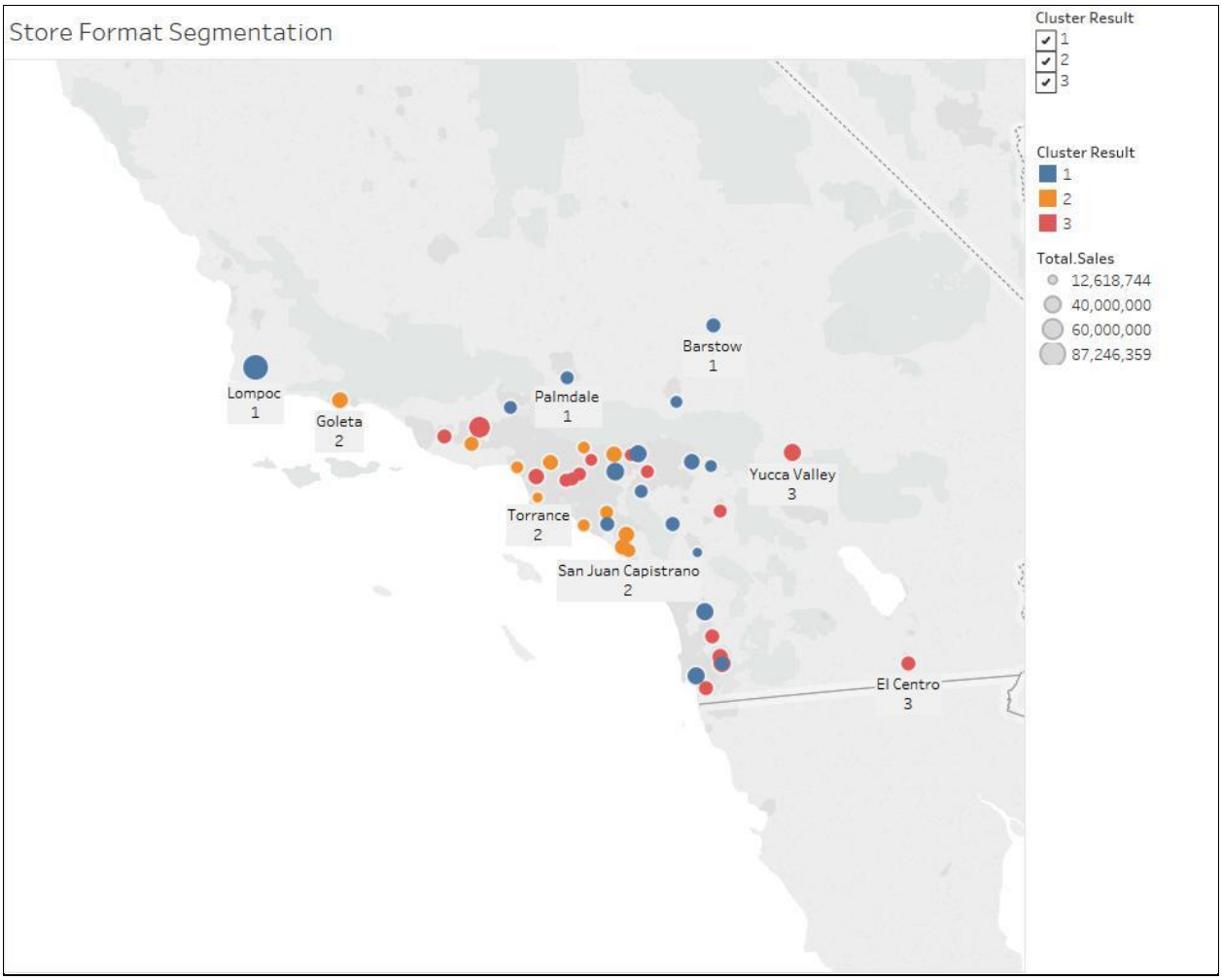
23 Stores Falls into format 1
29 Stores Falls into format 2
33 Stores Falls Into format 3

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

Format	Dry Grocery	Frozen Food	Dairy	Meat	Produce	Floral	Deli	Bakery	G.Merchandise
1	High	High	Low	Med	Low	Low	Low	Low	High
2	Low	Low	High	Low	High	High	Low	High	Low
3	High	Low	Med	High	Low	Low	High	High	Low

I can generalize 3 types format into, Dry, Frozen &General Merchandise focus, Fresh n Flowery Focus, the 3 one is Meat & Eat.

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 1 Guide



Based on the location we can see Cluster 2 Fresh and Flowery locates in more coastal area where Cluster 1 Dry, Frozen & General Merchandising more locates in the inner area, while the rest is in cluster 3 Meat & Eat

Task 2: Formats for New Stores

- Develop a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
- Use a 20% validation sample with Random Seed = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
- Use the model to predict the best store format for the each of 10 new stores.
- Use the StoreDemographicData.csv file, which contains the information for the area around each store.
- Note: In a real world scenario, you could use PCA to reduce the number of predictor variables. However, there is no need to do so in this project. You can leave all predictor variables in the model.

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.)

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Random_Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750
DecisionTree	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted_Model	0.8235	0.8543	0.8000	0.6667	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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Based on Model Comparison report, Random Forest Model and Boosted Model Has same accuracy in predicting the clusters, but in the end I will choose Boosted Model Since it has higher F1 score compared by random forest model

.

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

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

Task 3: Predicting Produce Sales

Step 1: To forecast sales for existing stores you should aggregate sales across all stores by month and produce a forecast.

Step 2: To forecast sales for new stores:

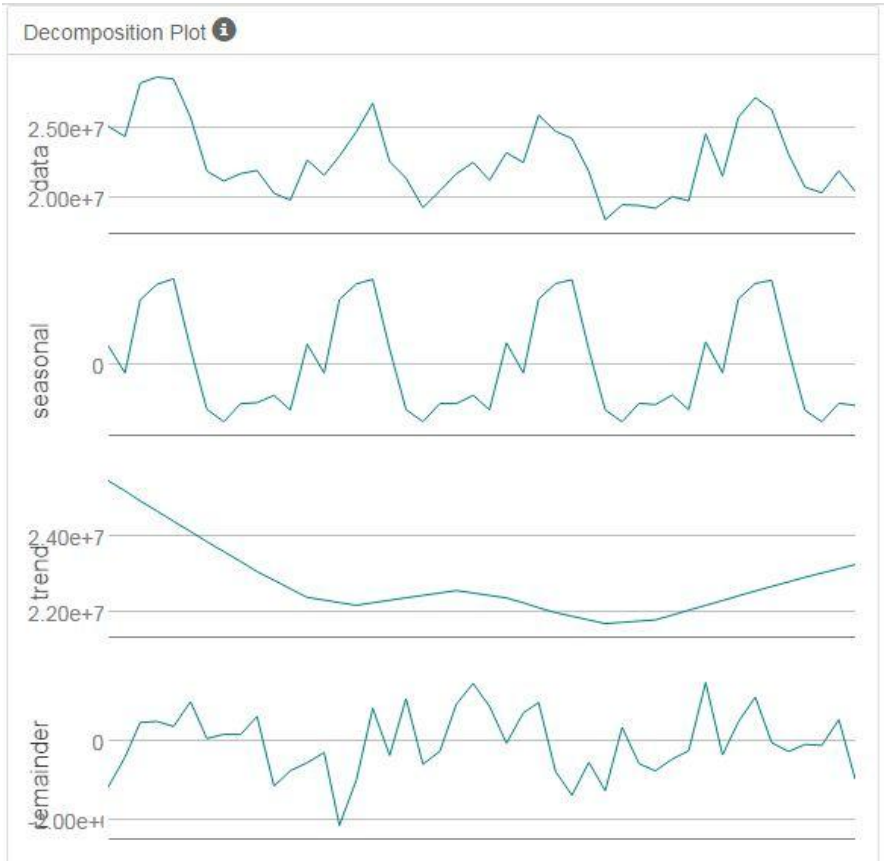
- Forecast produce sales for the average store (rather than the aggregate) for each segment.
- Multiply the average store sales forecast by the number of new stores in that segment.
- For example, if the forecasted average store sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
- Sum the new stores sales forecasts for each of the segments to get the forecast for all new stores.

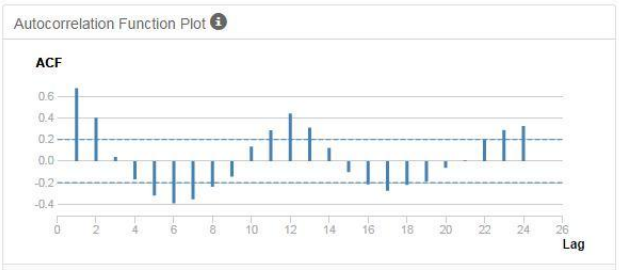
Step 3: Sum the forecasts of the existing and new stores together for the total produce sales forecast.

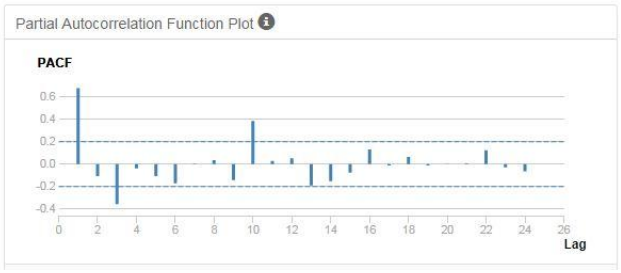
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

ETS Decomposition Plot



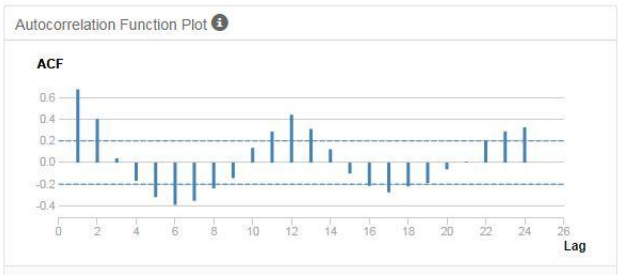


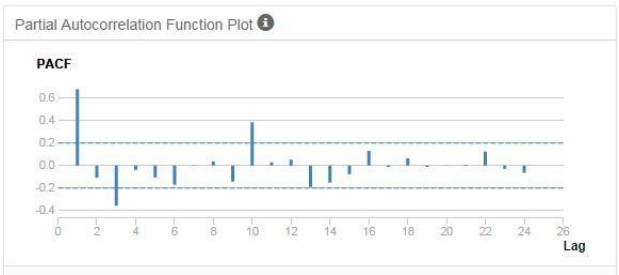


Based on the decomposition plot, the seasonal are very slightly decreasing over time, there is no trend, and increasing in error. This would suggest a ETS (M,N,M) model

ARIMA Method Result

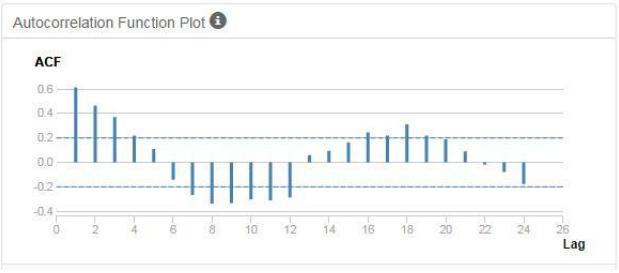
Time Series ACF & PACF

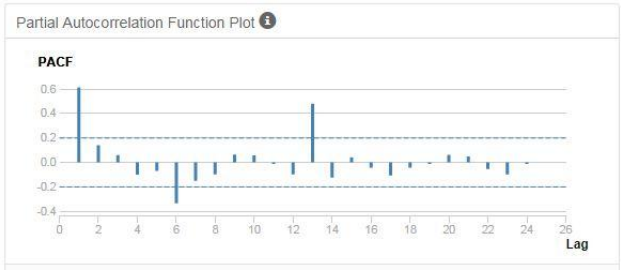




We should be able to see that the ACF presents slowly decaying serial correlations towards 0 with increases at the seasonal lags. Since serial correlation is high we will need to seasonally difference the series

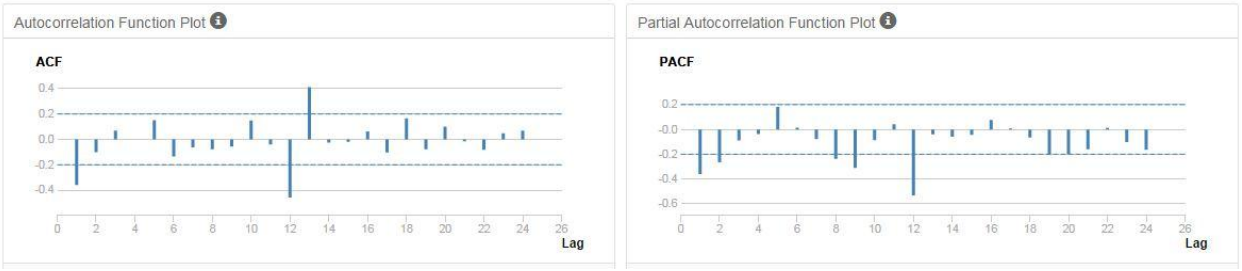
Seasonal Difference ACF & PACF





We can see that the seasonal difference presents similar ACF and PACF results as the initial plots without differencing, only slightly less correlated. In order to remove correlation we will need to difference further.

Seasonal First Difference ACF & PACF



The seasonal lags (lag 12, 24, etc.) in the ACF and PACF do not have any significant correlation so there will be no need for seasonal autoregressive or moving average terms. This means the P & Q would be zero. And since the know that the forecast is monthly, we found that m would be 12.

Suggested ARIMA model : ARIMA(1,0,0)(0,1,0)[12]

ARIMA(1,0,0)(0,1,0)[12] Result

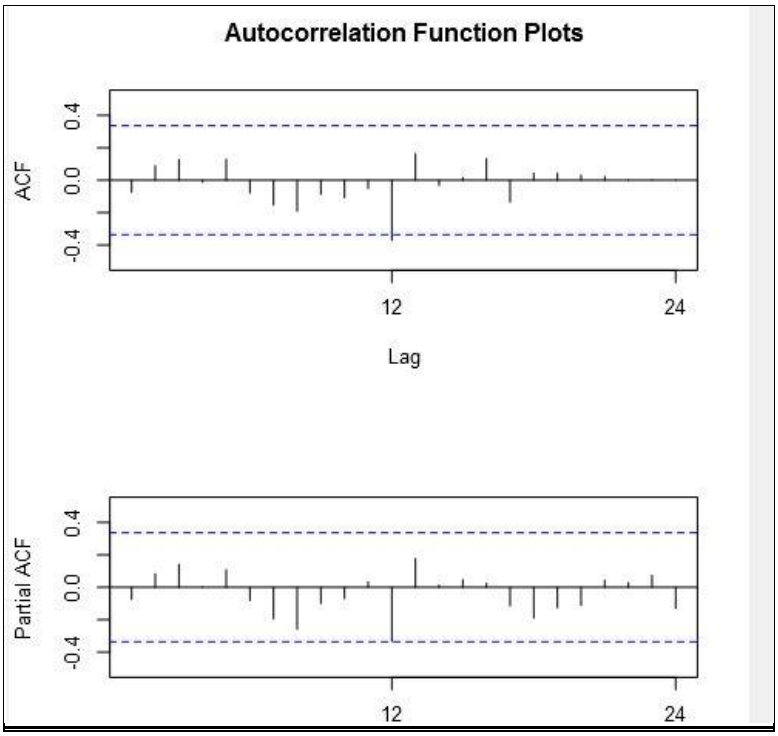
Information Criteria:

AIC	AICc	BIC
698.5919	699.9252	701.865

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
33580.6473571	1323644.3516029	927254.9963853	0.072035	4.2044525	0.4940823	-0.0747419

Ljung-Box test of the model residuals:
Chi-squared = 6.045, df = 10, p-value = 0.81147



The ACF and PACF results for the ARIMA (0,1,1) (0,1,0) 12 model show no significantly correlated lags suggesting no need for adding additional AR() or MA() terms

ETS (M,N,M) & ARIMA (0,1,1) (0,1,0) 12 TS comparison result

ETS M,N,M TS comparison

Report

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS__M_N_M_
20088529	19592857.92774
19772333	18699731.89634
24608406	21034791.45839
21559729	20376418.05772
25792074	23165380.05115
27212464	23673094.79871
26338477	24203475.5952
23130626	21294826.26085
20774415	19013559.92342
20359980	18512113.57484
21936906	19125133.00384
20462899	19599984.59824

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS__M_N_M_	1978789	2200153	1978789	8.4769	8.4769	1.266	NA

ARIMA TS (0,1,1) (0,1,0) 12 comparison

Report

Comparison of Time Series Models

Actual and Forecast Values:

Actual	ARIMA
20088529.29	20612945.68065
19772333.34	19631548.64226
24608406.71	21782519.33267
21559729.45	21169231.70749
25792074.59	24621094.18052
27212464.15	23488312.9623
26338477.15	22984780.24665
23130626.6	20610236.33425
20774415.93	17142257.22266
20359980.58	18235189.54062
21936906.81	18183821.42614
20462899.3	17980255.01536

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	2132888	2569526	2220290	9.201	9.6361	1.4205	NA

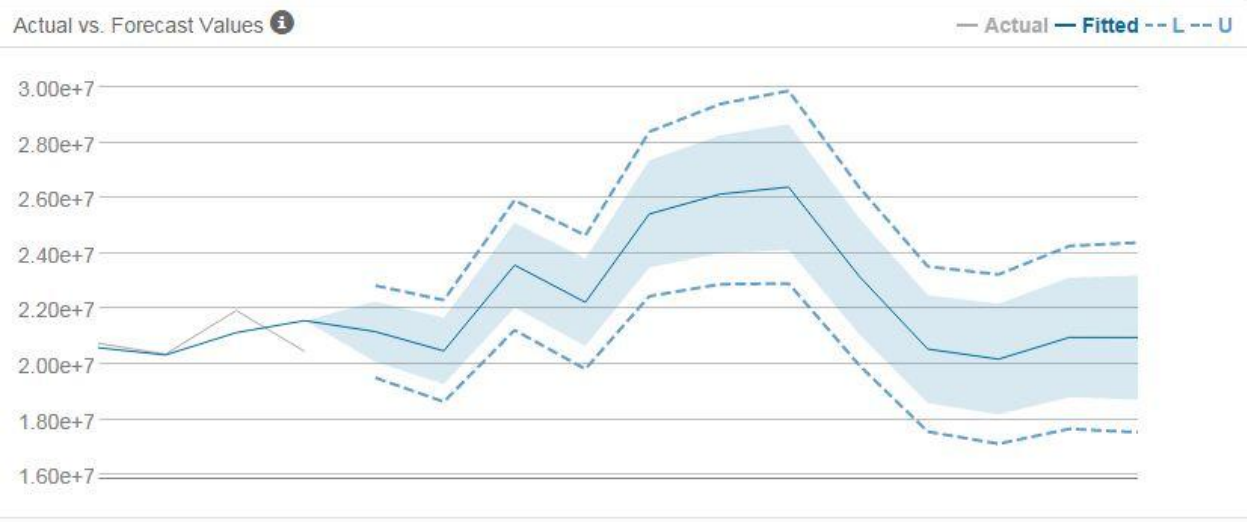
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS__M_N_M_	1978789	2200153	1978789	8.4769	8.4769	1.266	NA

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS__M_N_M_	1978789	2200153	1978789	8.4769	8.4769	1.266
ARIMA	2132888	2569526	2220290	9.201	9.6361	1.4205

Based on TS comparison on both model, we can see the result for both model are suite similar, but ETS (M,N,M) has the overall lower errors across all the variables. Mean Error Average Percentage Error (MPE) for ETS is less at 8.47% while the ARIMA model gives 9.20% errors. And Mean Absolute Scale Error (MASE) for ETS is also lower at 1.266 while the ARIMA models gives 1.42 errors. Ideally we should look for model that gives error less than MASE 1.0, that can be done by digging for more data, but since we only have 3-4 years monthly sales data, so this model is good enough.

And in this case, for the forecasting we should use the **ETS model** as our forecasting model.

ETS Forecast result



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
4	11	21174989.403661	22840074.392097	22263729.957327	20086248.849994	19509904.415225
4	12	20479354.577584	22316289.407716	21680461.702984	19278247.452184	18642419.747453
5	1	23580340.680462	25927572.132269	25115112.832138	22045568.528787	21233109.228656
5	2	22236546.234723	24649240.998411	23814122.545296	20658969.92415	19823851.471035
5	3	25427255.45708	28396631.129606	27368825.849103	23485685.065056	22457879.784554
5	4	26143967.404052	29399195.651373	28272446.748298	24015488.059806	22888739.156731
5	5	26399993.267031	29879368.950321	28675034.74188	24124951.792183	22920617.583742
5	6	23172393.880014	26386378.249505	25273905.302084	21070882.457944	19958409.510523
5	7	20544268.638819	23528908.819356	22495819.956152	18592717.321486	17559628.458283
5	8	20182471.08571	23241644.32162	22182756.948449	18182185.222971	17123297.849799
5	9	20966876.352467	24271769.849047	23127830.057693	18805922.647242	17661982.855887
5	10	20965097.001691	24391891.031043	23205757.181039	18724436.822343	17538302.972339

Final Sales Forecast – Existing Stores

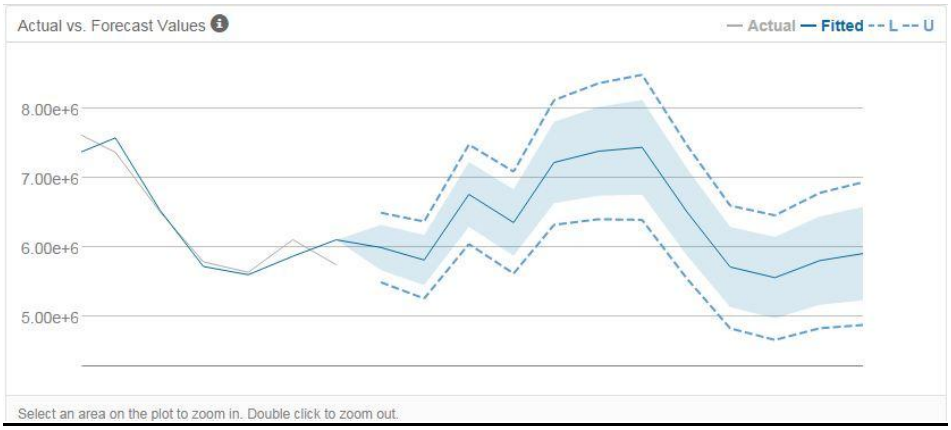
Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
Jan-16	$21,174,989	$22,840,074	$22,263,730	$20,086,249	$19,509,904
Feb-16	$20,479,355	$22,316,289	$21,680,462	$19,278,247	$18,642,420
Mar-16	$23,580,341	$25,927,572	$25,115,113	$22,045,569	$21,233,109
Apr-16	$22,236,546	$24,649,241	$23,814,123	$20,658,970	$19,823,851
May-16	$25,427,255	$28,396,631	$27,368,826	$23,485,685	$22,457,880
Jun-16	$26,143,967	$29,399,196	$28,272,447	$24,015,488	$22,888,739
Jul-16	$26,399,993	$29,879,369	$28,675,035	$24,124,952	$22,920,618
Aug-16	$23,172,394	$26,386,378	$25,273,905	$21,070,882	$19,958,410
Sep-16	$20,544,269	$23,528,909	$22,495,820	$18,592,717	$17,559,628
Oct-16	$20,182,471	$23,241,644	$22,182,757	$18,182,185	$17,123,298
Nov-16	$20,966,876	$24,271,770	$23,127,830	$18,805,923	$17,661,983
Dec-16	$20,965,097	$24,391,891	$23,205,757	$18,724,437	$17,538,303

Sales forecast for new stores

since all new stores locates in the cluster 1, we will filter monthly sales from all cluster 1 stores and aggregates it into sum of sales.

Step 1. Aggregate monthly sales of stores in cluster 1
Step 2. Conduct ETS model forecast for next 12 months
Step 3. Averaged it by number of cluster 1 existing stores
Step 4. Multiply it by number of new stores (10)

ETS Cluster 1 Forecast Result



Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
Jan-16	$5,993,193	$6,496,568	$6,322,332	$5,664,054	$5,489,818
Feb-16	$5,812,153	$6,369,475	$6,176,566	$5,447,740	$5,254,831
Mar-16	$6,763,382	$7,483,488	$7,234,234	$6,292,530	$6,043,277
Apr-16	$6,354,144	$7,091,847	$6,836,502	$5,871,786	$5,616,442
May-16	$7,224,490	$8,127,480	$7,814,923	$6,634,057	$6,321,501
Jun-16	$7,388,125	$8,372,924	$8,032,050	$6,744,200	$6,403,327
Jul-16	$7,443,989	$8,494,453	$8,130,851	$6,757,127	$6,393,524
Aug-16	$6,510,524	$7,477,529	$7,142,815	$5,878,233	$5,543,519
Sep-16	$5,711,427	$6,600,092	$6,292,494	$5,130,359	$4,822,761
Oct-16	$5,556,692	$6,458,852	$6,146,583	$4,966,801	$4,654,532
Nov-16	$5,802,982	$6,782,815	$6,443,661	$5,162,304	$4,823,149
Dec-16	$5,905,241	$6,939,290	$6,581,369	$5,229,113	$4,871,192

Since we have known that Cluster 1 consist 23 stores , we will divide forecast to 23, and then multiply the result by 10

Sub_Period	forecast	Divide by 23	Multiply by 10	New Stores Forecasts
Jan-16	$5,993,193	$260,573.61	$2,605,736.06	$2,605,736.06
Feb-16	$5,812,153	$252,702.31	$2,527,023.12	$2,527,023.12
Mar-16	$6,763,382	$294,060.10	$2,940,600.95	$2,940,600.95
Apr-16	$6,354,144	$276,267.15	$2,762,671.47	$2,762,671.47

May-16	$7,224,490	$314,108.27	$3,141,082.71	$3,141,082.71
Jun-16	$7,388,125	$321,222.83	$3,212,228.33	$3,212,228.33
Jul-16	$7,443,989	$323,651.69	$3,236,516.88	$3,236,516.88
Aug-16	$6,510,524	$283,066.26	$2,830,662.64	$2,830,662.64
Sep-16	$5,711,427	$248,322.90	$2,483,228.98	$2,483,228.98
Oct-16	$5,556,692	$241,595.31	$2,415,953.08	$2,415,953.08
Nov-16	$5,802,982	$252,303.57	$2,523,035.69	$2,523,035.69
Dec-16	$5,905,241	$256,749.60	$2,567,496.04	$2,567,496.04

Sales Forecast Summary

Period	Existing Stores	New Stores	All Stores
16-Jan	$21,174,989	$2,605,736.06	$23,780,725
16-Feb	$20,479,355	$2,527,023.12	$23,006,378
16-Mar	$23,580,341	$2,940,600.95	$26,520,942
16-Apr	$22,236,546	$2,762,671.47	$24,999,217
16-May	$25,427,255	$3,141,082.71	$28,568,338
16-Jun	$26,143,967	$3,212,228.33	$29,356,195
16-Jul	$26,399,993	$3,236,516.88	$29,636,510
16-Aug	$23,172,394	$2,830,662.64	$26,003,057
16-Sep	$20,544,269	$2,483,228.98	$23,027,498
16-Oct	$20,182,471	$2,415,953.08	$22,598,424
16-Nov	$20,966,876	$2,523,035.69	$23,489,912
16-Dec	$20,965,097	$2,567,496.04	$23,532,593

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.