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

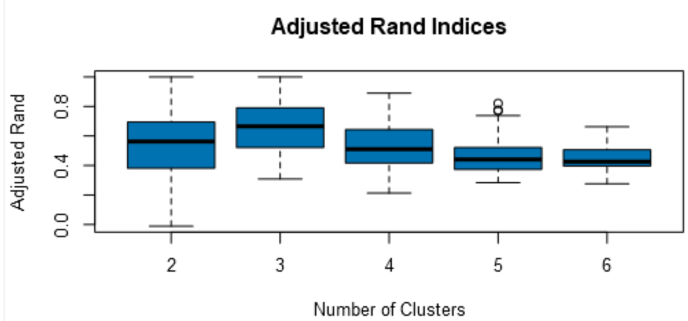
1. What is the optimal number of store formats? How did you arrive at that number?

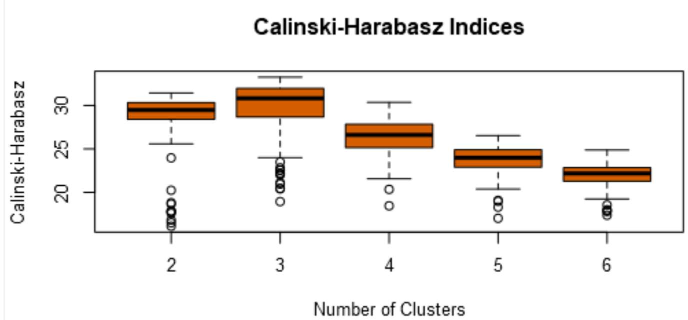
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.01155	0.3083	0.213	0.2837	0.2762
1st Quartile	0.3814	0.5258	0.4169	0.374	0.3965
Median	0.5619	0.6653	0.5107	0.4406	0.4256
Mean	0.5084	0.6594	0.5471	0.4704	0.4502
3rd Quartile	0.6942	0.7865	0.6427	0.5199	0.5067
Maximum	1	1	0.8902	0.8207	0.6626

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	16.1	18.94	18.45	17.02	17.37
1st Quartile	28.42	28.68	25.16	22.91	21.28
Median	29.47	30.83	26.61	23.98	22.17
Mean	28.24	29.58	26.34	23.7	21.95
3rd Quartile	30.31	31.97	27.85	24.9	22.84
Maximum	31.44	33.26	30.37	26.53	24.87





Based on the k-centroids diagnostics, with standardization of the fields, the clustering method: k-means we have to check the report and because of the adjusted rand indices, the Callinski-Harabasz Indices I recommend to use 3 clusters. At number of three clusters we're having the best median value.

2. How many stores fall into each store format?

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

Cluster 1 has 23 stores, Cluster 2 has 29 stores and 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?

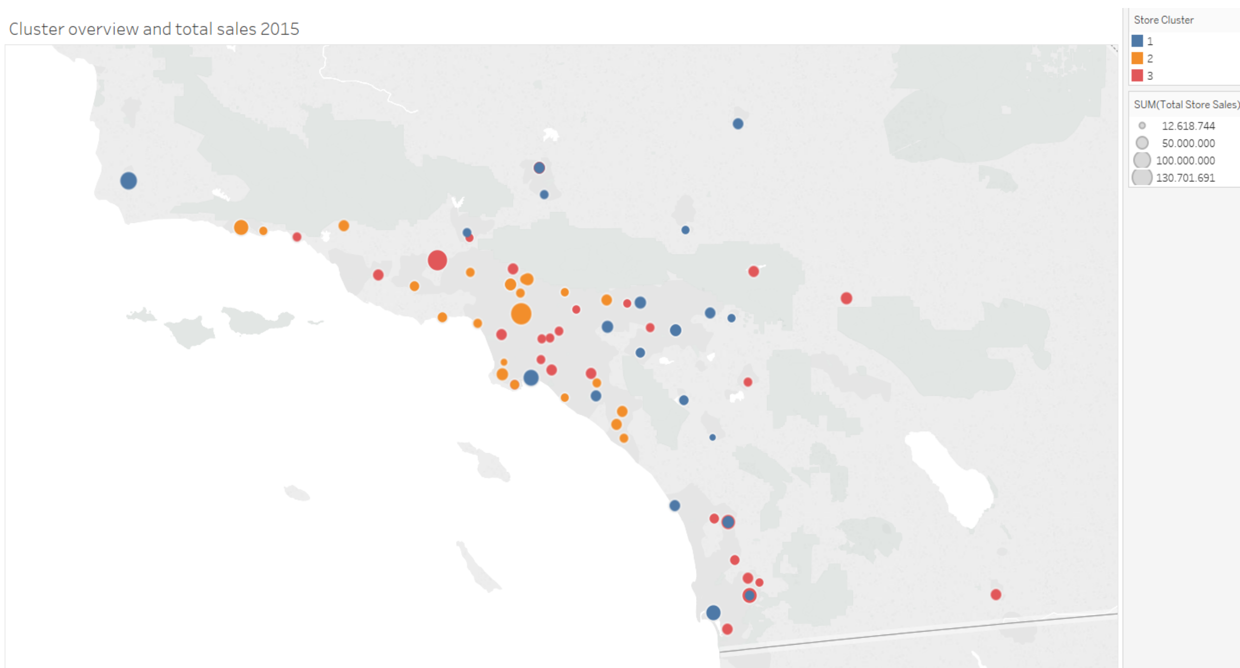
I compared the three clusters in Alteryx (K-Centroids Cluster Analysis) and I used the average sales for each product (Instead of the sum. Because a combination of 33 stores and the sum of the sold products are normally bigger as for 29 or 23 stores).

	Percentage_Dry_Grocery	Percentage_Dairy	Percentage_Frozen_Food	Percentage_Meat	Percentage_Produce	Percentage_Floral	Percentage_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

	Percentage_Bakery	Percentage_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Here we can see for example, that the average sales for bakery products in cluster 1 is quite low. But cluster 1 is selling quite a lot of general merchandise. On the other hand cluster 2 performance quite good in selling floral but not in selling dry grocery. And for cluster 3 the sales of general merchandise aren't quite good but they are selling quit good for deli products.

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.



> [You can find the Tableau public file here: https://public.tableau.com/profile/bj.rn.gam#!/vizhome/UdacityBAND-FinalProject-/Clusteroverviewandtotalsales2015](https://public.tableau.com/profile/bj.rn.gam#!/vizhome/UdacityBAND-FinalProject-/Clusteroverviewandtotalsales2015)

> [You can find the Alteryx Worklow for Step 1 here: https://drive.google.com/open?id=1NG4X723teiE9N8vnqOjDqa8_b7Iv3zf4](https://drive.google.com/open?id=1NG4X723teiE9N8vnqOjDqa8_b7Iv3zf4)

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.)

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Model	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Forrest_Model	0.8235	0.8251	0.7500	0.8000	0.8750

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

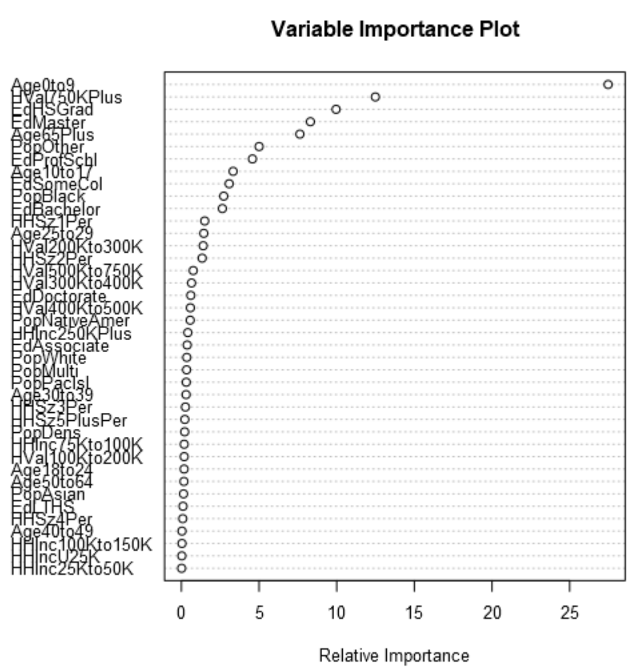
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

I used the boosted model for predict the segments/cluster for the new store, because of the accuracy at the accuracy and F1 both values are better as at the other models. 0.8235 = 0.8235 (if we compare the accuracy decision tree model and the boosted model) and 0.8543 > 0.8251 (if we compare the F1 value of the decision tree model and the F1 value of the boosted model).

Ave0to9, HVal750KPlus and EdHSGrad are the most important variables.



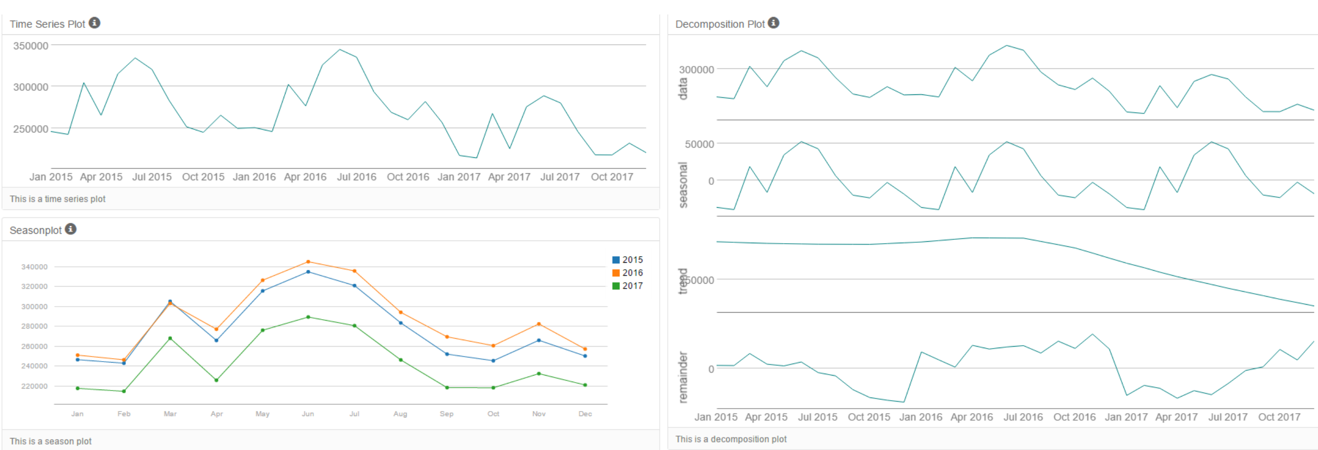
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

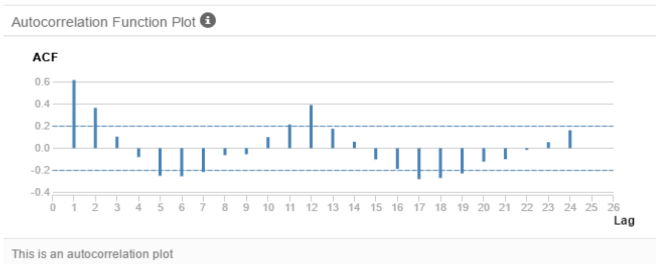
The link for the Alteryx workflow: https://drive.google.com/open?id=1NG4X723teiE9N8vnqOjDqa8_b7Iv3zf4

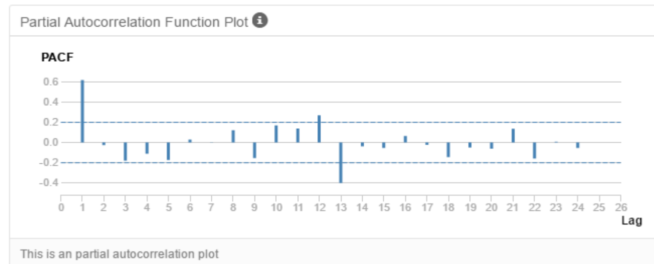
Task 3: Predicting Produce Sales

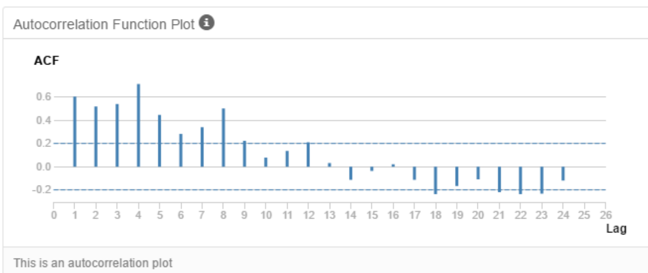
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

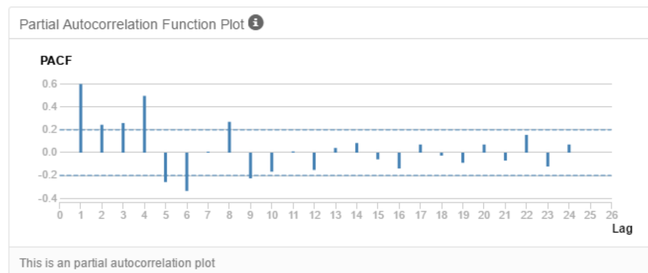


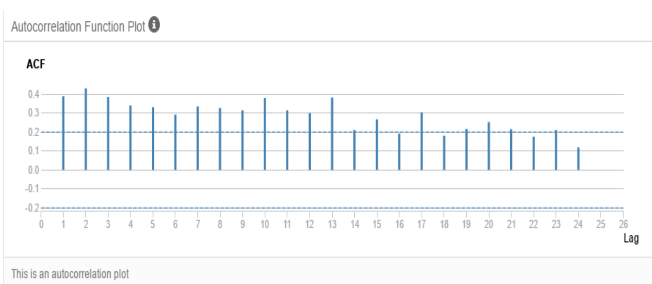
I used the no dampening ETS (M,N, M) model. The seasonality shows an increasing trend and should be set to multiplicative. The trend is not clear so we're using none for the trend. The error is irregular and so we use here also multiplicative.

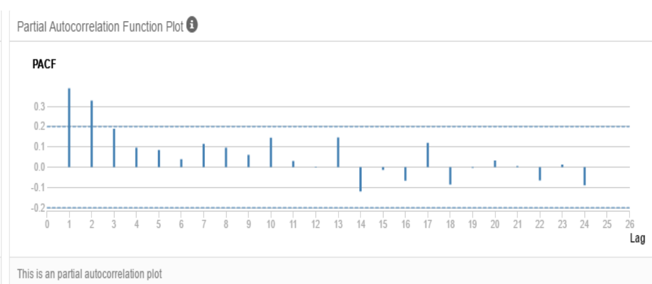












So thanks to our observations we can built our ARIMA model: ARIMA (0,1,2)(0,1,0)12.At the ARIMA model we're having the following structure: ARIMA (p, d, q) or for the seasonal ARIMA model (p, d, q)(P, D, Q) m.

So p is the AR order, d is the integration order, q the MA order and m is the period. After we have stationary the values ones we have to set d = 1(none stationary time series plot = 0), p has to be set to = 0, because we're having a negative lag_1 and q has set to 2 because of the negative correlation at lag_1. The period (m) will set to 12 (it's monthly data).

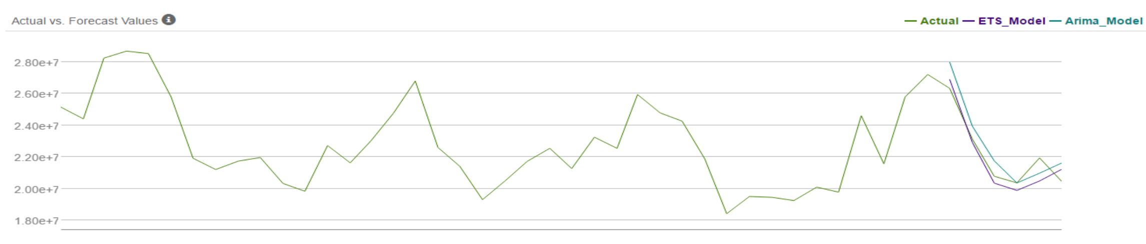
The seasonal first difference of the series has removed most of the significant lags from the ACF and PACF model so there is no further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms and the differencing terms will be D(1).

If we compared the ARIMA and the ETS model with each other:

Accuracy Measures:

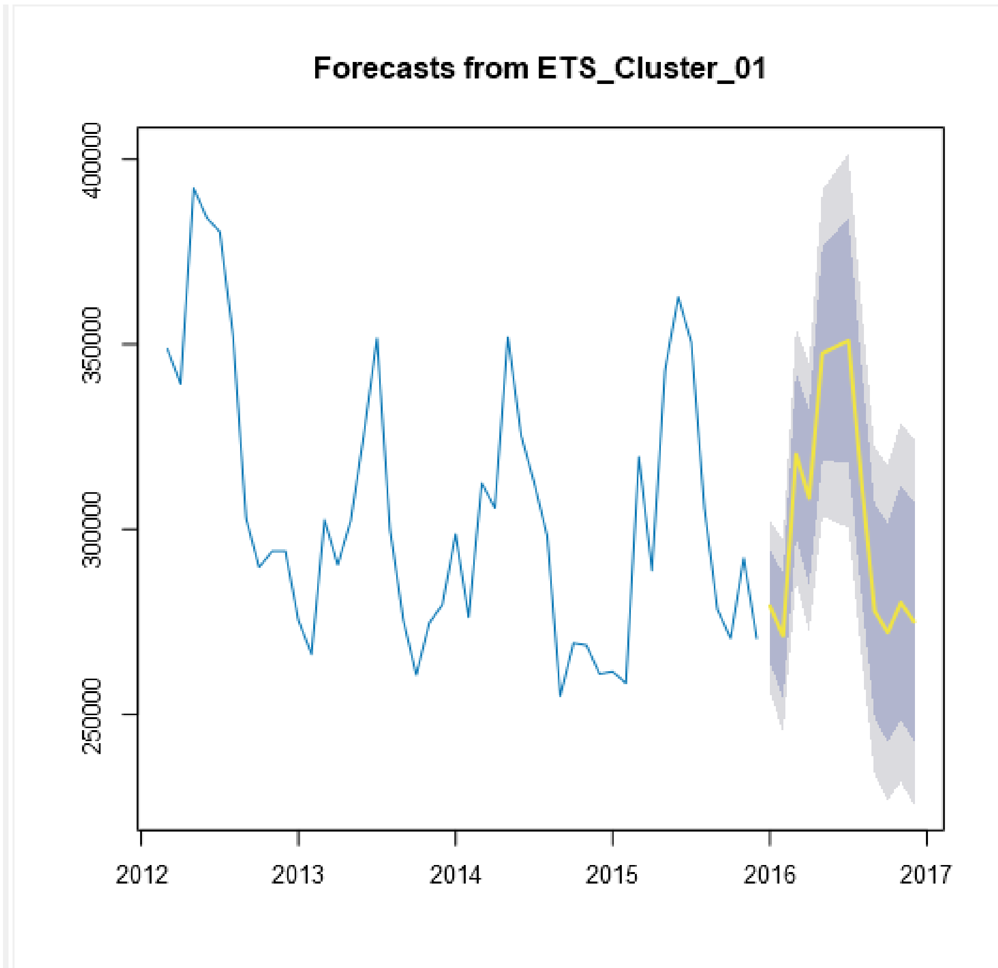
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_Model	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA
Arima_Model	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463	NA

ETS model's accuracy is higher when compared to ARIMA model. Its RMSE of 76027.3 is lower than ARIMA's 1050239,2 while its MASE is 0.3822 compared to ARIMA's 0.5463. I used a holdout sample of six month for comparing both models.

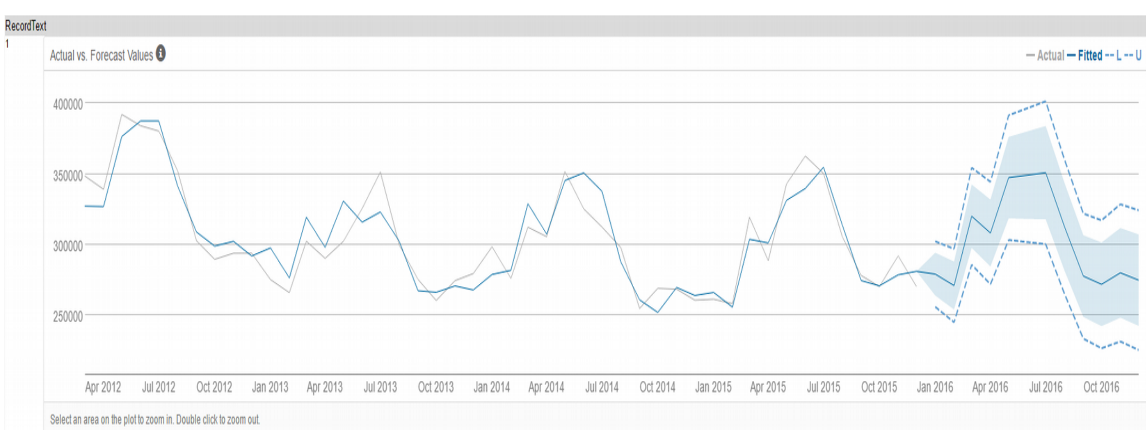


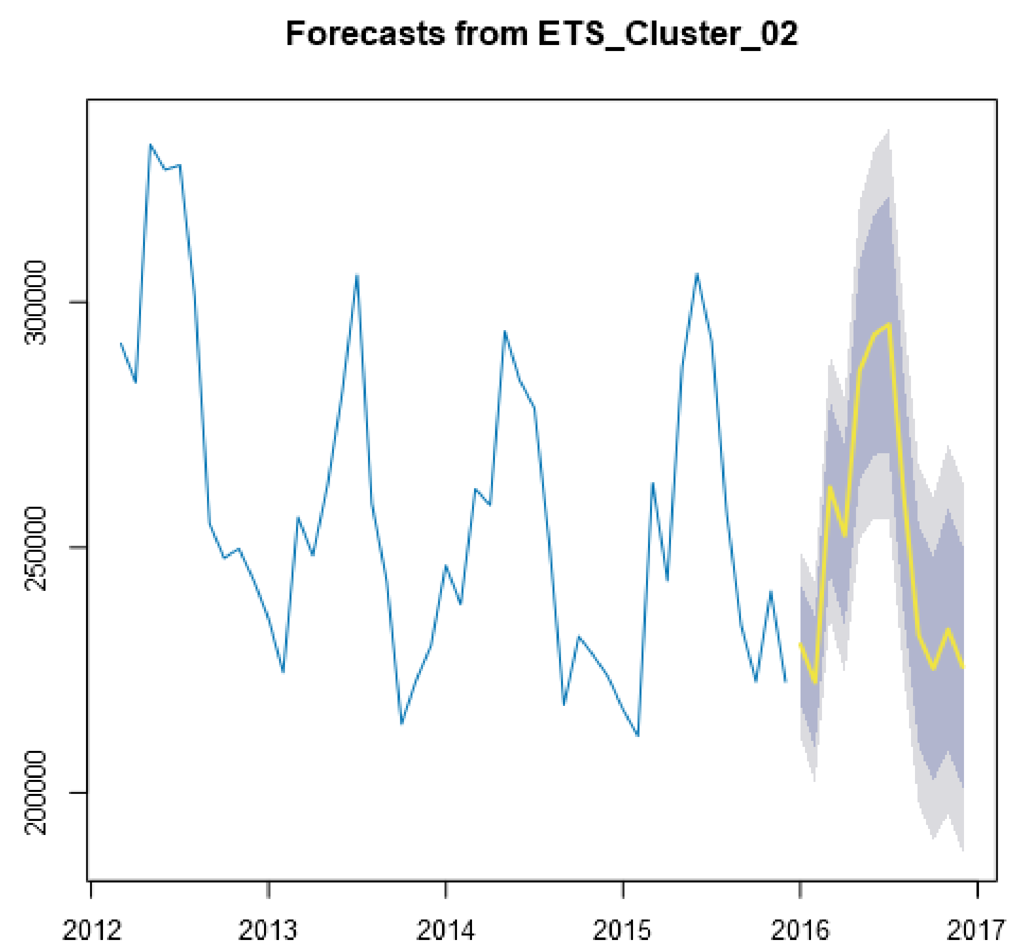
Also in the graph above we can see, that the ETS line is closer to the actual line.

The predicted forecast for the three clusters:



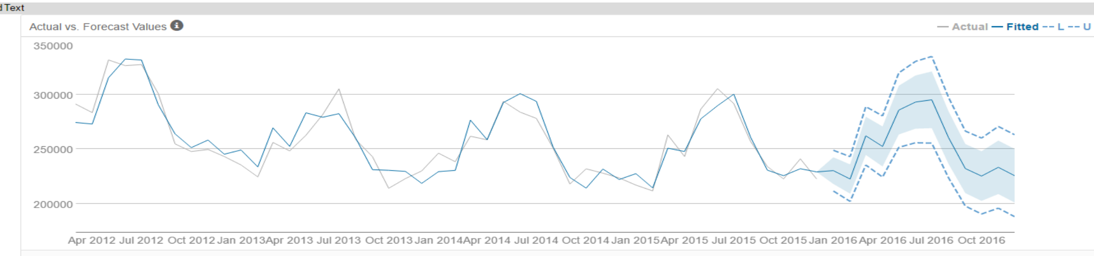
Period	Sub_Period	forecast_cluster01	forecast_cluster01_high_95	forecast_cluster01_high_80	forecast_cluster01_low_80	forecast_cluster01_low_95
2016	1	279359.951231	302491.485714	294484.84866	264235.053802	256228.416748
2016	2	271276.711509	297233.530044	288248.962883	254304.460135	245319.892974
2016	3	320292.322453	354574.141234	342707.998902	297876.646004	286010.503672
2016	4	308385.507855	344557.497073	332037.100322	284733.915388	272213.518636
2016	5	347561.575041	391609.339374	376362.860057	318760.290024	303513.810707
2016	6	349273.129635	396607.291396	380223.274427	318322.984843	301938.967874
2016	7	351036.383804	401507.948659	384037.966214	318034.801395	300564.818949
2016	8	312807.059744	360225.798673	343812.506573	281801.612914	265388.320814
2016	9	277964.883996	322168.053488	306867.782984	249061.985009	233761.714505
2016	10	272144.056763	317355.201313	301706.035215	242582.07831	226932.912212
2016	11	280249.651075	328717.973509	311941.38367	248557.918481	231781.328642
2016	12	275095.472535	324477.737812	307384.800234	242806.144836	225713.207258



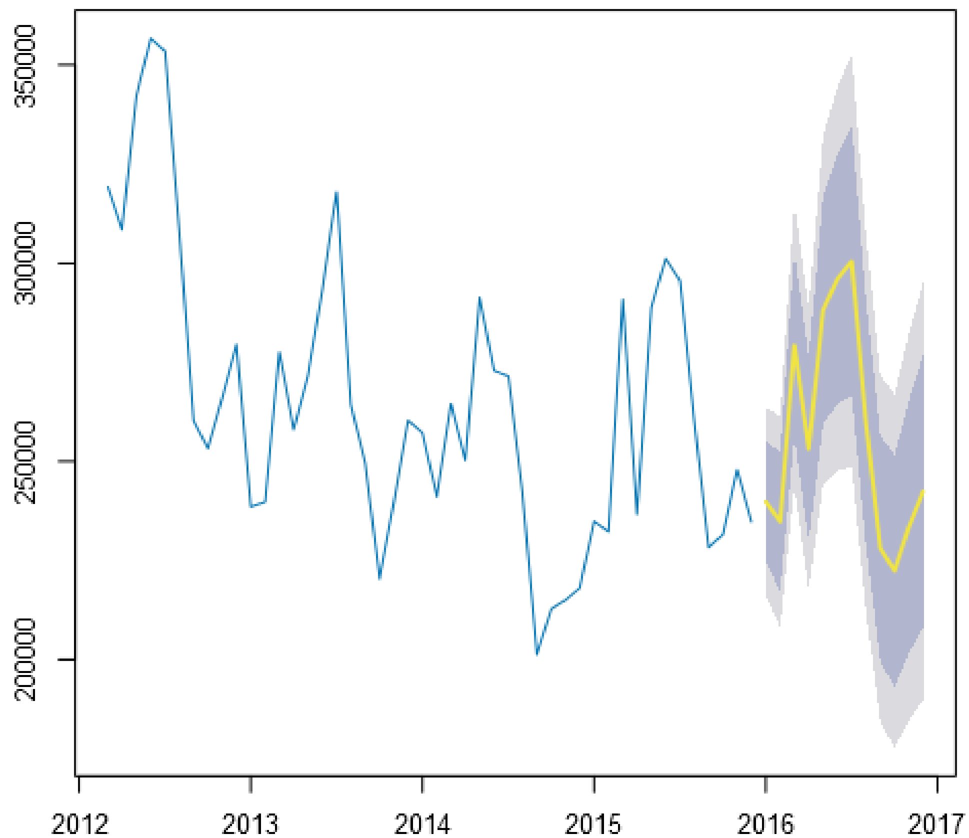


Period	Sub_Period	forecast_cluster02	forecast_cluster02_high_95	forecast_cluster02_high_80	forecast_cluster02_low_80	forecast_cluster02_low_95
2016	1	230279.308724	249020.729897	242533.665352	218024.952096	211537.887552
2016	2	222541.592231	243123.861867	235999.614972	209083.569491	201959.322595
2016	3	262308.721179	289168.353438	279871.290718	244746.151641	235449.088921
2016	4	252507.231641	280644.934047	270905.48633	234108.976951	224369.529234
2016	5	285958.946717	320215.060827	308357.815785	263560.077649	251702.832606
2016	6	293377.59829	330815.979632	317857.240058	268897.956522	255939.216948
2016	7	295468.100509	335350.65389	321545.900472	269390.300546	255585.547128
2016	8	260629.866341	297631.915254	284824.205669	236435.527013	223627.817428
2016	9	232181.54451	266692.358673	254746.952991	209616.136029	197670.730347
2016	10	225290.324677	260214.728039	248126.164599	202454.484755	190365.921315
2016	11	233333.701974	270935.971761	257920.504636	208746.899311	195731.432187
2016	12	225637.8003	263333.538105	250285.718414	200989.882185	187942.062495

Record Text
1

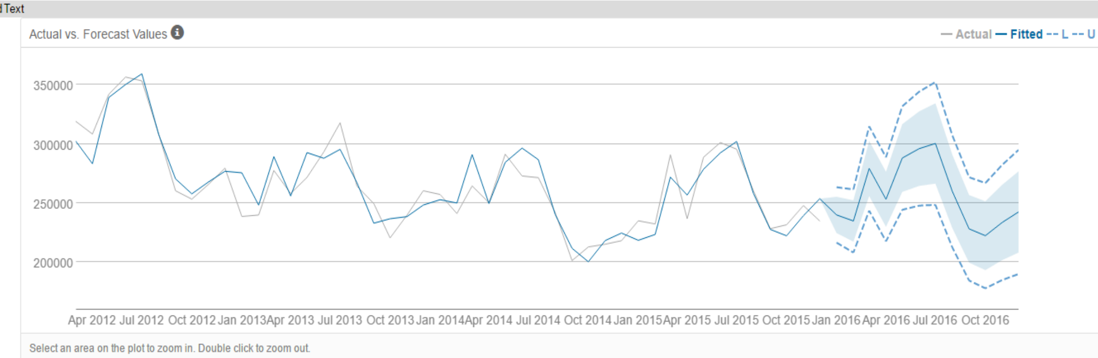


Forecasts from ETS_Cluster_03



Period	Sub_Period	forecast_cluster03	forecast_cluster03_high_95	forecast_cluster03_high_80	forecast_cluster03_low_80	forecast_cluster03_low_95
2016	1	239909.276978	263389.237841	255261.998171	224556.555786	216429.316116
2016	2	234783.792647	261522.407102	252267.232897	217300.352397	208045.178192
2016	3	279260.228617	314984.273449	302618.925992	255901.531241	243536.183785
2016	4	253158.109329	288746.383406	276428.03098	229888.187679	217569.835253
2016	5	288099.930594	331946.443074	316769.624096	259430.237091	244253.418113
2016	6	296029.610059	344279.579756	327578.569474	264480.650645	247779.640363
2016	7	300458.757652	352475.703309	334470.810254	266446.70505	248441.811996
2016	8	259986.416018	307487.670718	291045.81701	228927.015026	212485.161318
2016	9	228090.3674	271842.221116	256698.166864	199482.567937	184338.513684
2016	10	222319.486909	266899.73714	251468.945739	193170.028078	177739.236677
2016	11	233553.513603	282336.615565	265451.069469	201655.957737	184770.411641
2016	12	242410.822809	294991.923927	276791.75684	208029.888778	189829.721692

Record Text



The forecast for the existing stores:

t_cluster01	forecast01	Input_#2_Period	Input_#2_Sub_Period	forecast_cluster02	forecast_cluster02_high_80	forecast02	Input_#3_Period	Input_#3_Sub_Period	forecast_cluster03	forecast03
1231	838079.853693	2016	1	230279.308724	242533.665352	1381675.852345	2016	1	239909.276978	239909.276978
1509	813830.134526	2016	2	222541.592231	235999.614972	1335249.553388	2016	2	234783.792647	234783.792647
2453	960876.967359	2016	3	262308.721179	279871.290718	1573852.327077	2016	3	279260.228617	279260.228617
7855	925156.523565	2016	4	252507.231641	270905.48633	1515043.389845	2016	4	253158.109329	253158.109329
5041	1042684.725122	2016	5	285958.946717	308357.815785	1715753.680301	2016	5	288099.930594	288099.930594
9635	1047819.388905	2016	6	293377.59829	317857.240058	1760265.589741	2016	6	296029.610059	296029.610059
3804	1053109.151413	2016	7	295468.100509	321545.900472	1772808.603052	2016	7	300458.757652	300458.757652
9744	938421.179231	2016	8	260629.866341	284824.205669	1563779.198045	2016	8	259986.416018	259986.416018
3996	833894.651988	2016	9	232181.54451	254746.952991	1393089.267059	2016	9	228090.3674	228090.3674
6763	816432.170288	2016	10	225290.324677	248126.164599	1351741.948061	2016	10	222319.486909	222319.486909

The forecast for the existing stores and the new stores. For the new stores I used the average product value of the average product sales of the existing stores. And also the ETS model (m,n,m).

Period	Sub_Period	Forecast Existing Stores	Forecast New Stores
2016	1	$21.539.936,01	$2.459.664,98
2016	2	$20.413.770,60	$2.383.863,48
2016	3	$24.325.953,10	$2.813.989,52
2016	4	$22.993.466,35	$2.693.358,02
2016	5	$26.691.951,42	$3.046.538,34
2016	6	$26.989.964,01	$3.104.114,59
2016	7	$26.948.630,76	$3.126.376,51
2016	8	$24.091.579,35	$2.762.186,79
2016	9	$20.523.492,41	$2.455.074,29
2016	10	$20.011.748,67	$2.390.493,61
2016	11	$21.177.435,49	$2.474.304,68
2016	12	$20.855.799,11	$2.421.524,04

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".

You can find the file here:

https://public.tableau.com/profile/bj.rn.gam#!/vizhome/Task3_150/Dashboard1?publish=yes

And all project related files can be found here: https://drive.google.com/open?id=1NG4X723teiE9N8vnqOjDqa8_b7Iv3zf4

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.