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 store formats:
   Cluster 3 has the highest mean and median in both indices

K-Means Cluster Assessment Report

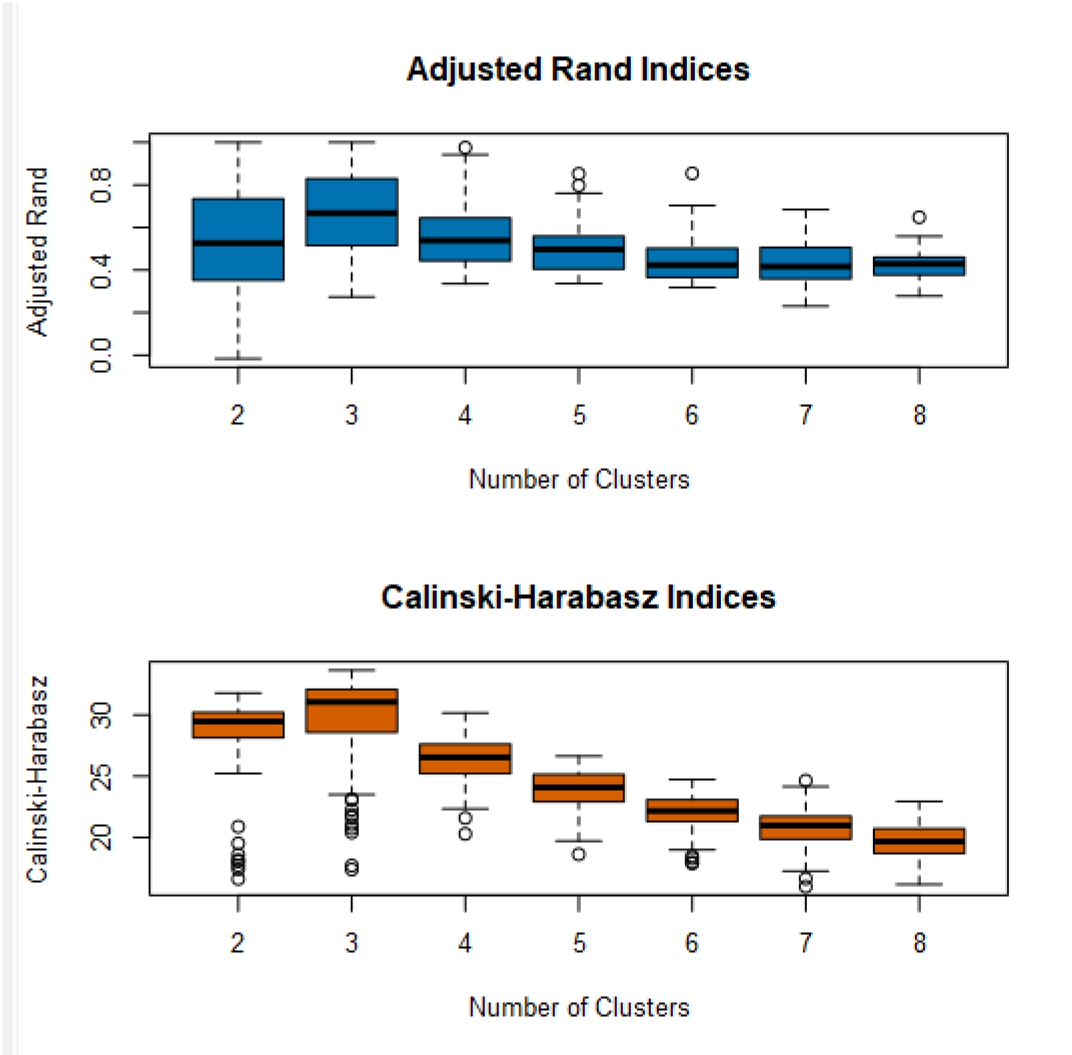
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.016293	0.27351	0.335359	0.336327	0.318262	0.230196	0.27786
1st Quartile	0.352041	0.515917	0.445826	0.409773	0.366788	0.358895	0.377341
Median	0.526785	0.66768	0.538528	0.497192	0.423541	0.416509	0.428806
Mean	0.53781	0.664773	0.565975	0.50103	0.45115	0.432196	0.421514
3rd Quartile	0.734477	0.826692	0.644691	0.555087	0.499921	0.502931	0.458601
Maximum	1	1	0.975264	0.852076	0.8539	0.683894	0.647983

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.61829	17.38103	20.28456	18.61989	17.8746	15.98702	16.16824
1st Quartile	28.17383	28.57484	25.20913	22.93454	21.30575	19.85155	18.71365
Median	29.46587	31.05384	26.53788	24.086	22.16245	20.97743	19.6662
Mean	28.45131	29.70664	26.41806	23.87003	22.02174	20.77195	19.65973
3rd Quartile	30.17907	32.08726	27.59305	25.10099	23.06602	21.72942	20.7099
Maximum	31.78345	33.63781	30.1583	26.63063	24.72038	24.63982	22.95166



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

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

Record #	Cluster	Count
1	1	23
2	2	29
3	3	33

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Stores in cluster 1 sold more General Merchandise
   Stores in cluster 2 sold more Produce and floral
   Stores in cluster 3 sold more meat and deli

Report

Summary Report of the K-Means Clustering Solution Cluster_Analysis

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.

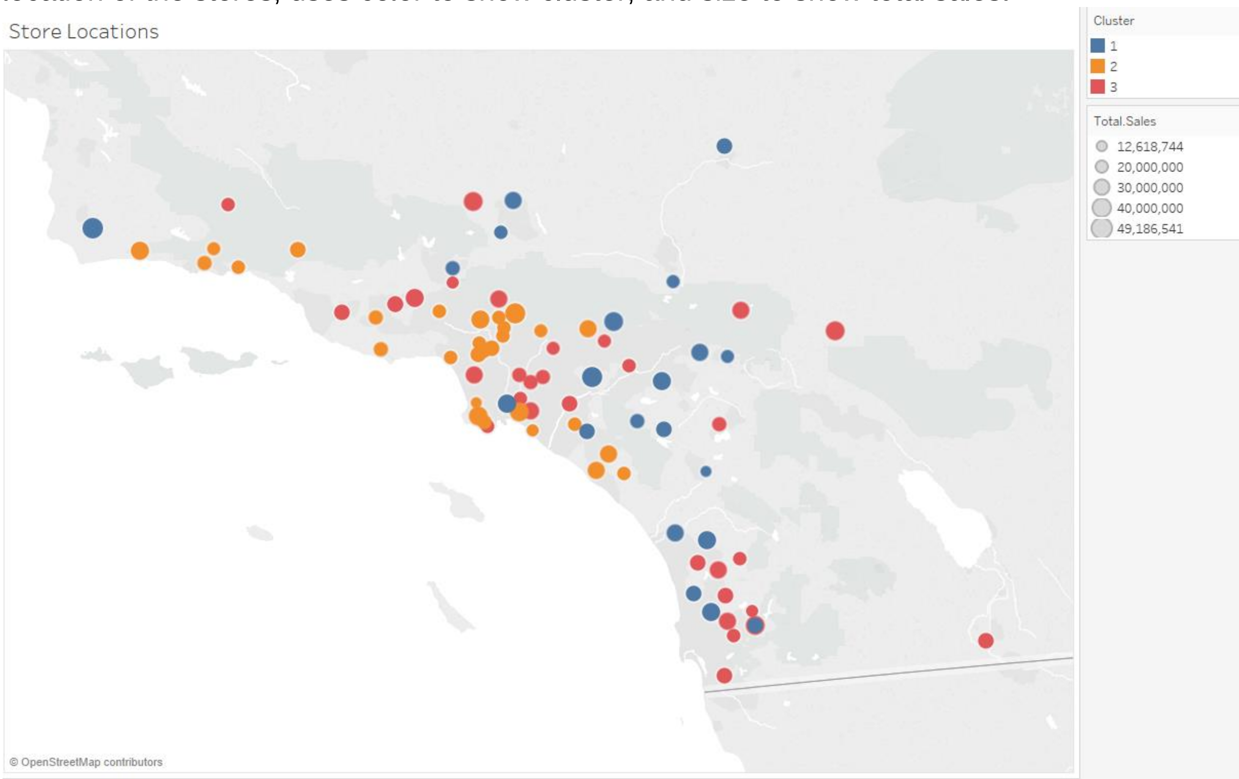


Tableau Public Link: https://public.tableau.com/profile/johan.tibeus#!/vizhome/Store_Locations_map/StoreLocations?publish=yes

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 model comparison report below shows comparison matrix of Decision Tree, Forest Model and Boosted Model.
I choose boosted model. Even though it has the same accuracy as the forest model (.8235) Boosted model has a higher F1 value (.8889 vs .8426)

Record #	Model	Accuracy	Accuracy_1	Accuracy_2	Accuracy_3	F1
1	NewStores_DT	0.705882	0.75	1	0.555556	0.768519
2	NewStores_Forest	0.823529	0.75	1	0.777778	0.842593
3	NewStores_Boosted	0.823529	1	1	0.666667	0.888889

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
NewStores_DT	0.7059	0.7685	0.7500	1.0000	0.5556
NewStores_Forest	0.8235	0.8426	0.7500	1.0000	0.7778
NewStores_Boosted	0.8235	0.8889	1.0000	1.0000	0.6667

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of NewStores_Boosted

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of NewStores_DT

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of NewStores_Forest

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

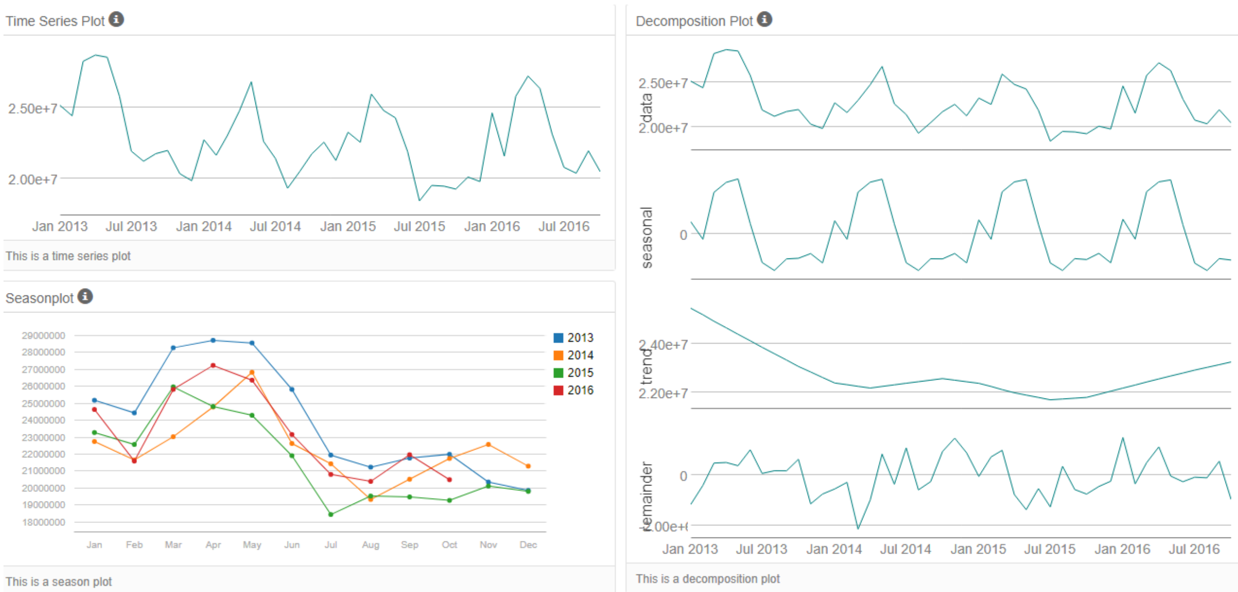
49 of 49 Fields | Cell Viewer | 10 records displayed | Data | Metadata

Record #	PopPacIsl	PopWhite	HVal0to100K	HVal100Kto200K	HVal200Kto300K	HVal300Kto400K	HVal400Kto500K	HVal500Kto750K	HVal750KPlus	PopDens	Score_1	Score_2	Score_3	Cluster
1	0.000756	0.179619	0.130383	0.13756	0.088517	0.113038	0.121411	0.325359	0.083732	2094.407018	0.348417	0.013522	0.638061	3
2	0.003119	0.506368	0.017774	0.018442	0.092075	0.113324	0.09702	0.34545	0.315916	6256.72792	0.078987	0.804431	0.116582	2
3	0.004828	0.043139	0.047129	0.028211	0.094922	0.155659	0.13309	0.375705	0.165284	8043.562891	0.486943	0.064498	0.448559	1
4	0.005308	0.453006	0.035694	0.060048	0.060978	0.080312	0.068786	0.415691	0.27849	7547.025711	0.026597	0.935435	0.037968	2
5	0.00659	0.527099	0.022281	0.019634	0.017648	0.070152	0.054269	0.291198	0.524818	7621.043926	0.019654	0.939601	0.040745	2
6	0.006286	0.405789	0.101091	0.210909	0.368727	0.167273	0.044364	0.074909	0.032727	1054.522398	0.887418	0.003833	0.108749	1
7	0.001625	0.471739	0.027035	0.048877	0.137926	0.136857	0.151062	0.35711	0.141133	8639.436528	0.028199	0.94173	0.030071	2
8	0.004384	0.469771	0.013704	0.192849	0.346456	0.197459	0.108135	0.112744	0.028653	3207.438094	0.857561	0.005592	0.136847	1
9	0.001893	0.713645	0.008479	0.019272	0.121025	0.162074	0.099827	0.189632	0.399692	4435.823519	0.00871	0.955864	0.035426	2
10	0.002157	0.567129	0.196016	0.053418	0.183794	0.184699	0.189679	0.181077	0.011317	2663.834099	0.080423	0.641377	0.2782	2

Task 3: Predicting Produce Sales

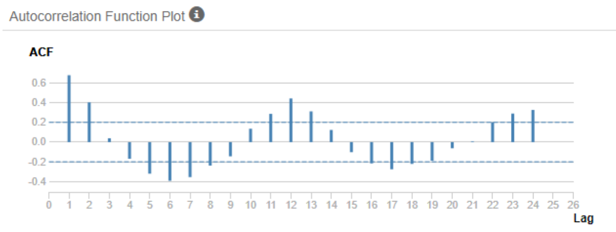
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

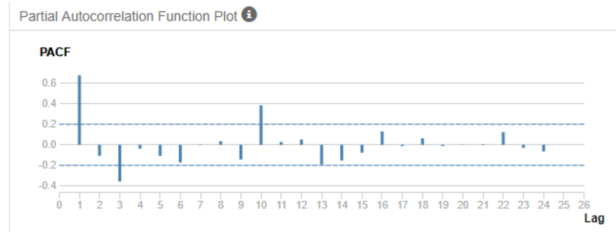
ETS(M,N,M)



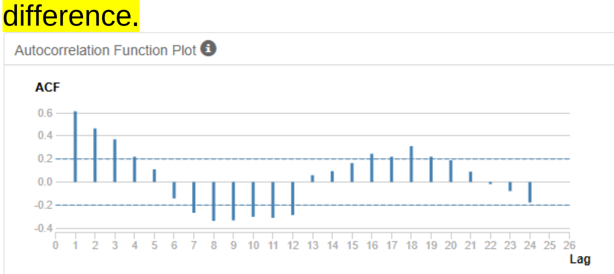
Error: Graph changes variance as the time series moves along. Multiplicative.
Trend: No clear trend, None.
Seasonality: Trending shows in each seasonal period. Multiplicative.
Model used: ETS(M,N,M) See below how I came to choose ETS model.

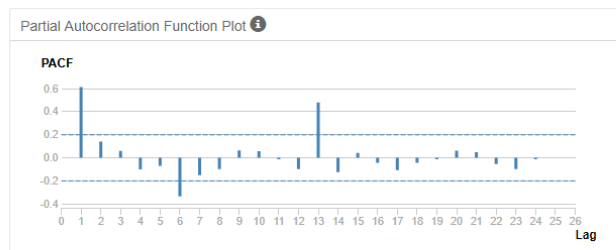
1): This is the ACF/PACF plots before any differencing, series is not stationary. Model is seasonal. Next step is to apply seasonal differencing.



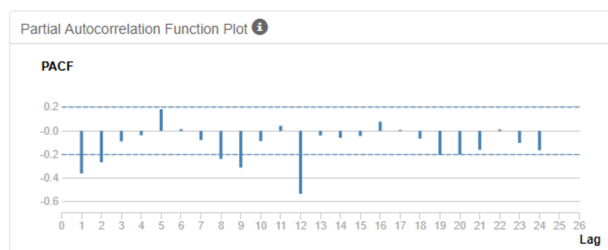


2): Taking the seasonal difference using [Sum_Produce]-[Row-12: Sum_Produce], we see that series is not yet stationary but the seasonal component is gone. Next is first difference.



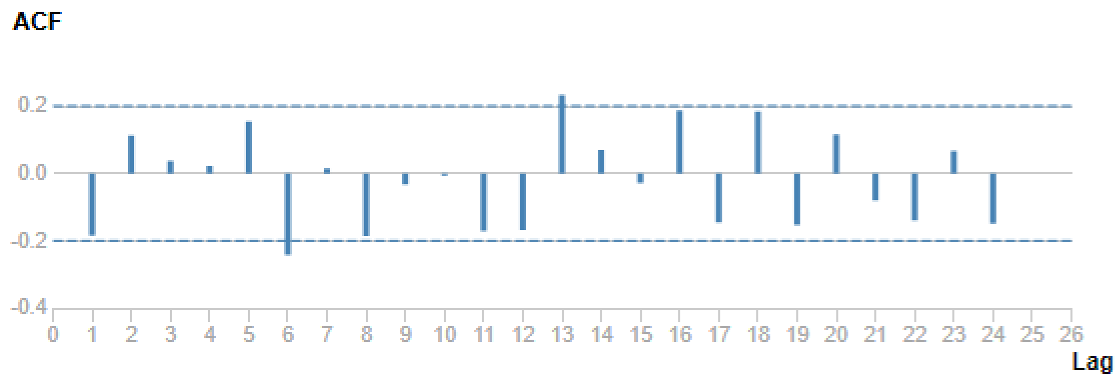


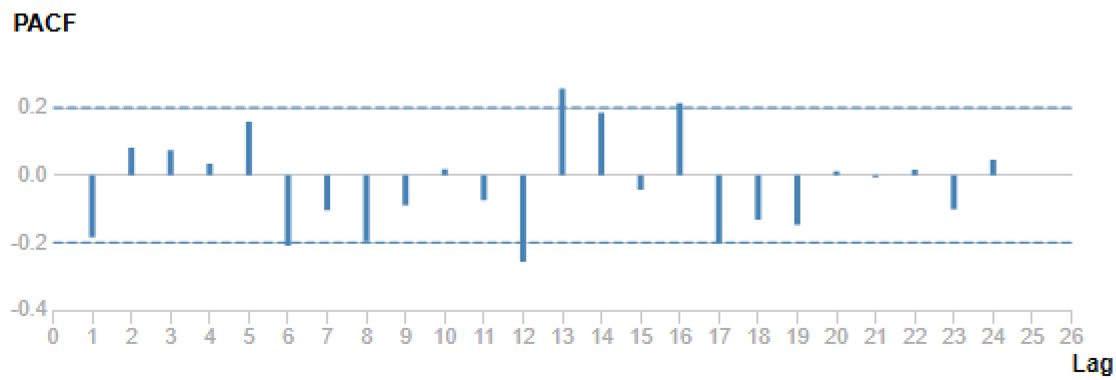
3): Taking the first difference using [Seasonal Difference]-[Row-1:Seasonal Difference], we see that series is now stationary excepts for lag 1. Next is to add RA term in ARIMA tool, since PACF cuts off.



Seasonal ARIMA models are denoted (p,d,q)(P,D,Q)m.
From the graphs we need to use a ARIMA(1,0,0)(1,1,0)12
Graph below after Running ARIMA(1,0,0)(1,1,0)12

Autocorrelation Function Plot





ETS(M,N,M) comparison

Actual and Forecast Values:

Actual	ETS
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822

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

Actual and Forecast Values:

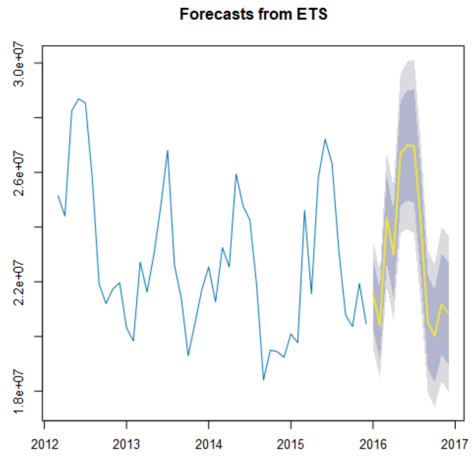
Actual	ARIMA
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

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

Comparing all 3 models I choose the ETS(M,N,M) for the forecast due to RMSE and MASE is the lowest in the ETS method.

Forecasting from ETS with 95% & 80% confidence intervals:



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

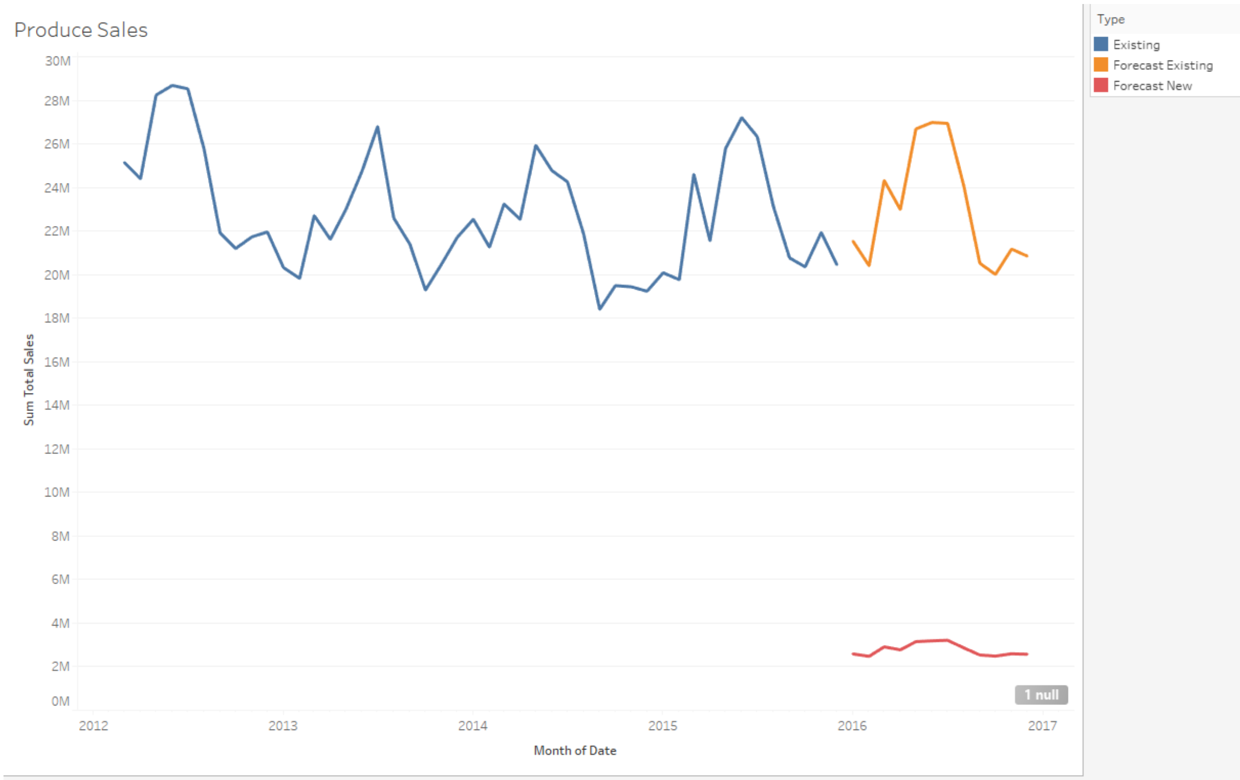
Record #	Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
1	2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2	2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
3	2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981696
4	2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
5	2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
6	2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
7	2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
8	2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
9	2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
10	2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
11	2016	11	21177435.485838	23994279.191514	23019270.585553	19335600.386124	18360591.780163
12	2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

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.

New Store Sales:

Record #	Sum_forecast	Sub_Period
1	2587450.851495	1
2	2477352.892393	2
3	2913185.23625	3
4	2775745.609767	4
5	3150866.835326	5
6	3188922.00336	6
7	3214745.646251	7
8	2866348.663392	8
9	2538726.84886	9
10	2488148.287462	10
11	2595270.386448	11
12	2573396.62905	12

Year	Month	New Store Sales	Existing Store Sales
2016	1	$2,587,451	$21,539,936
2016	2	$2,477,353	$20,413,771
2016	3	$2,913,185	$24,325,953
2016	4	$2,775,746	$22,993,466
2016	5	$3,150,867	$26,691,951
2016	6	$3,188,922	$26,989,964
2016	7	$3,214,746	$26,948,631
2016	8	$2,866,349	$24,091,579
2016	9	$2,538,727	$20,523,492
2016	10	$2,488,148	$20,011,749
2016	11	$2,595,270	$21,177,435
2016	12	$2,573,397	$20,855,799



https://public.tableau.com/profile/johan.tibeus#!/vizhome/ProduceSales_15598879775890/ProduceSales?publish=yes

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.