BENJAMIN OWUSU BEDIAKO

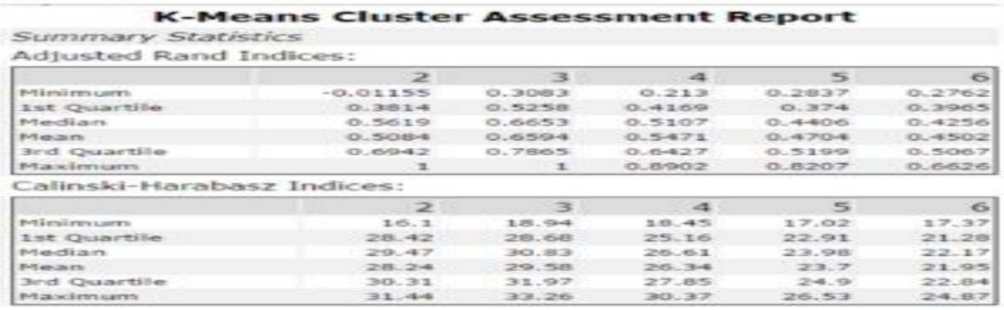
Project: Predictive Analytics Capstone

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?

Answer:



K-Means Cluster Assessment Report

Summary Statistics

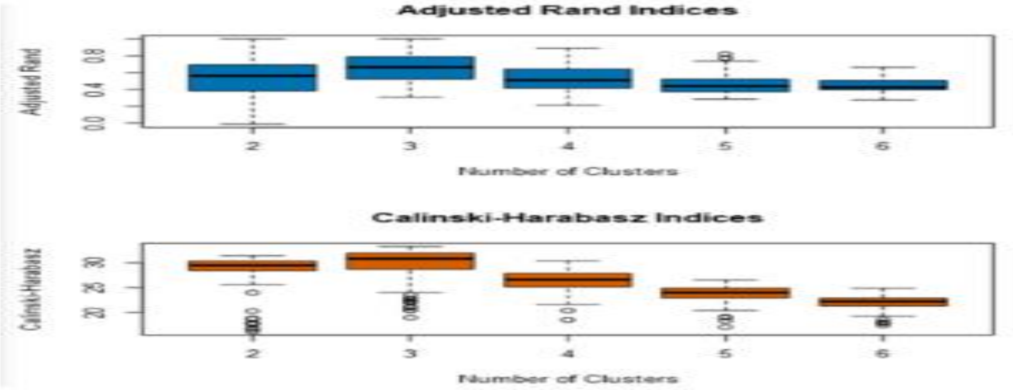
Adjusted Rand Indices:

	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

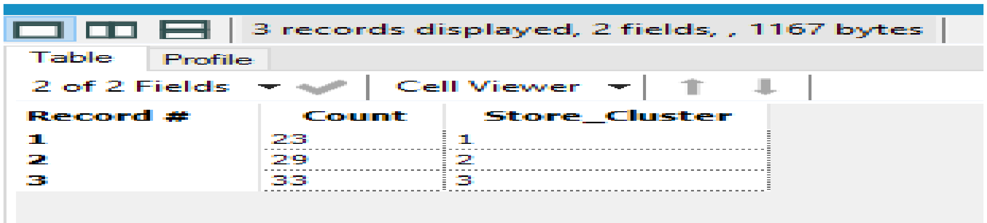
Figure 1: K-Means Cluster Assessment Report



Looking at the above visualization, the optimal number of store formats is 3 since the adjusted rand and Calinski-Harabasz indices had the highest median number.

2. How many stores fall into each store format?

Answer:

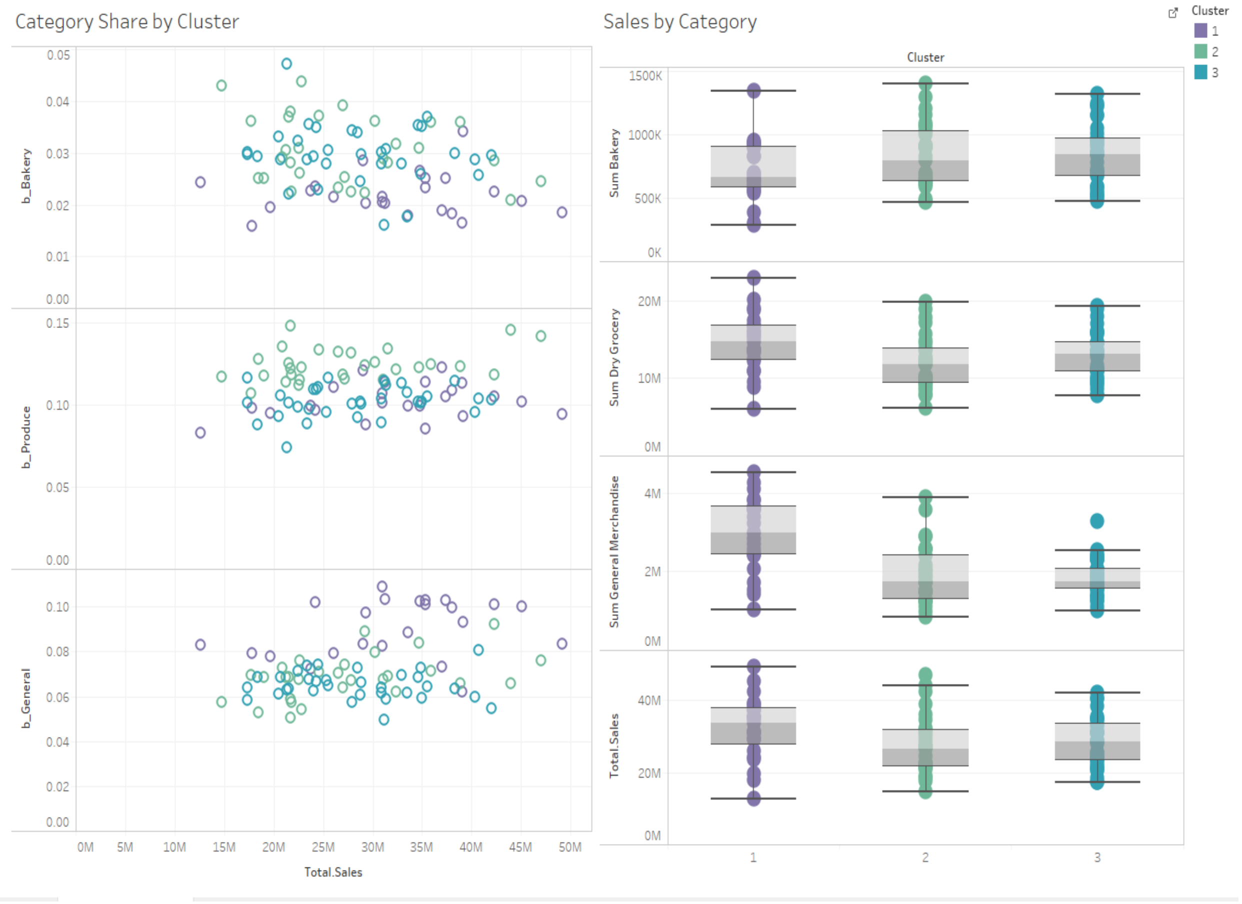


Record #	Count	Store_Cluster
1	23	1
2	29	2
3	33	3

The above visualization indicates that the Stores_Cluster 1, 2 and 3 have Store_count of 23, 29 and 33 respectively.

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

Answer:

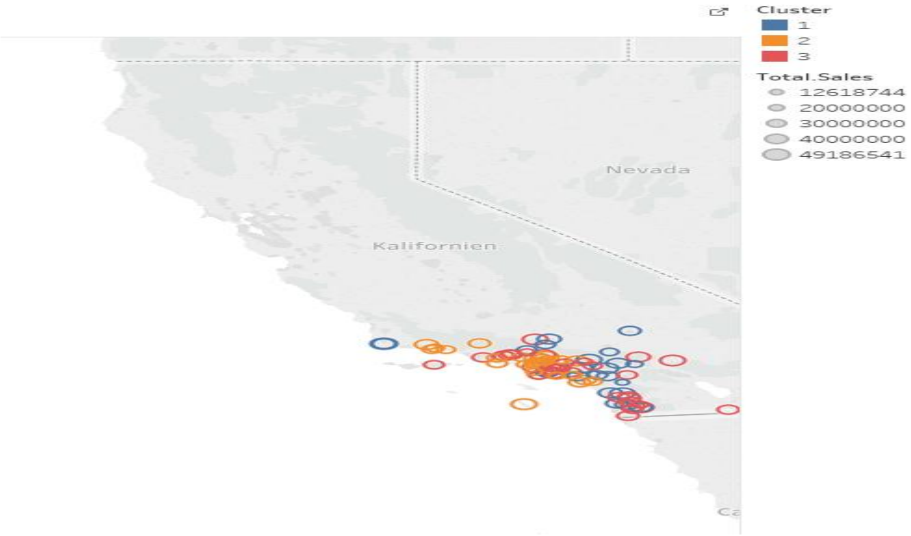


Based on the visualization, cluster_3 shows the highest sales on bakery with $1,243,333 whiles cluster_1 and cluster_2 with sales of $914,460 and $1,153,011 respectively. Again, cluster_1 gains the highest sales on Dry Grocery with $22,920,868 where as cluster_2 and cluster_3 had sales of $19,906,353 and $19,352,651 respectively. Furthermore, cluster_1 gains the highest sales on General Merchandise with $4,091,857 while cluster_2 and cluster_3 had sales of $3,571,912 and $2,297,763. Therefore it can be asserts that cluster_1 had the total highest sales of products than the cluster_2 and cluster_3.

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.

Answer:

https://public.tableau.com/profile/benjamin.bediako#!/vizhome/Task1_42/Clustervizualization?publish=yes

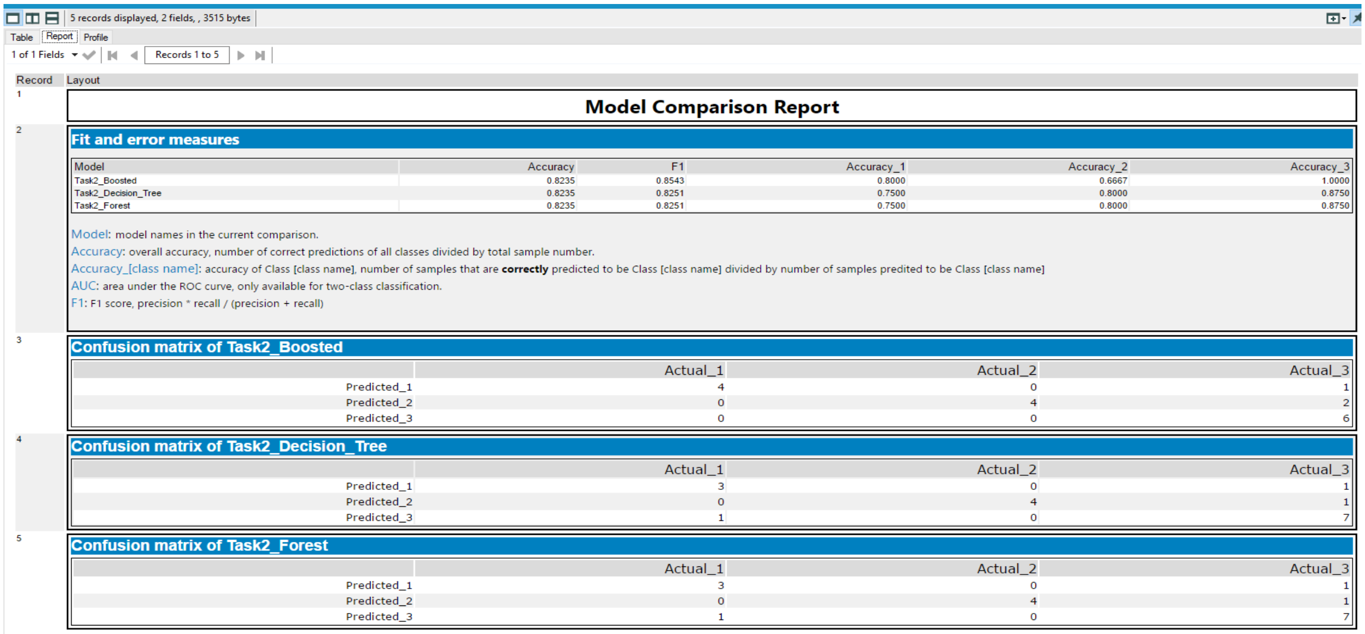


Task 2: Formats for New Stores

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.)

Answer:

In predicting the best store format for the new stores, the model comparison report exhibits both the fit and error measures; and the confusion matrices of Boosted, decision tree and Forest models. Under the fit and error measures, all the Boosted, decision tree and forest models had the same accuracy number but looking at the F1 value, the Boosted model had the highest. Therefore the Boosted model is chosen.



1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Task2_Boosted	0.8235	0.8543	0.8000	0.6667	1.0000
Task2_Decision_Tree	0.8235	0.8251	0.7500	0.8000	0.8750
Task2_Forest	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)

3

Confusion matrix of Task2_Boosted

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

4

Confusion matrix of Task2_Decision_Tree

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

5

Confusion matrix of Task2_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.

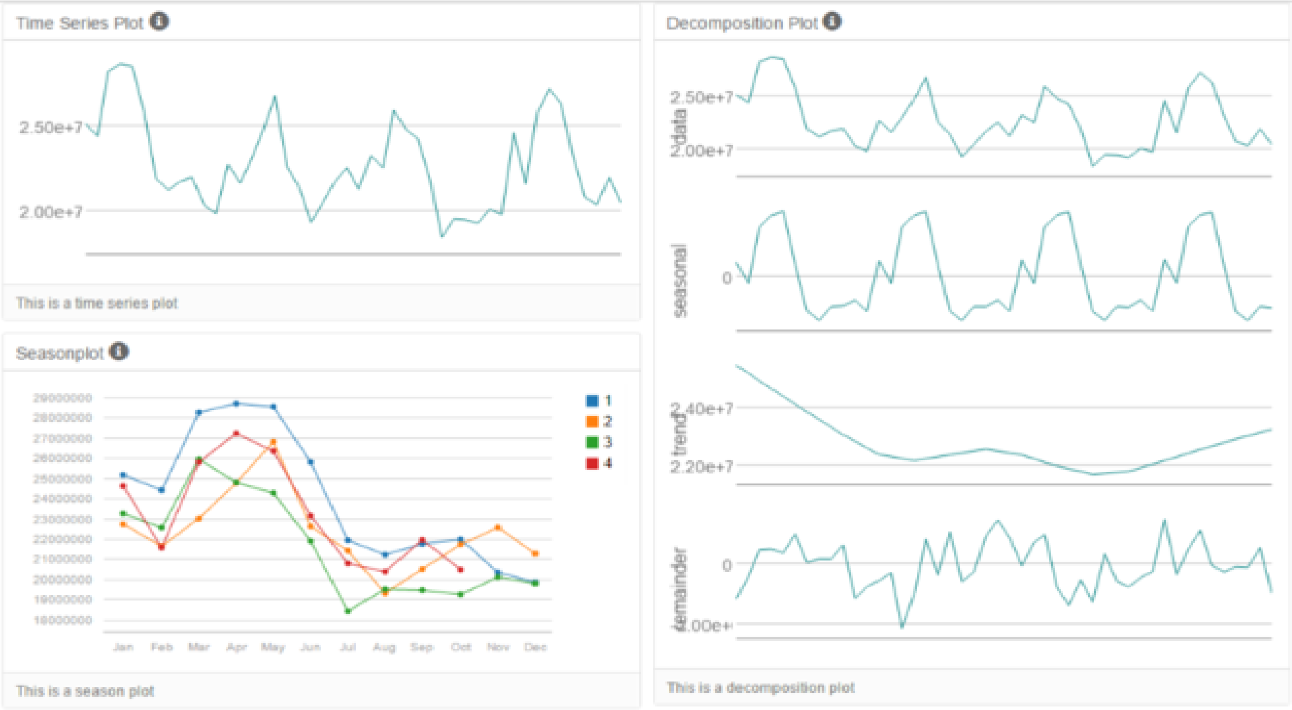
Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

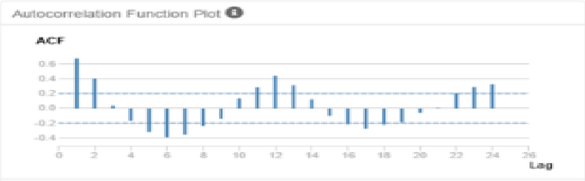
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS (a, m, n) or ARIMA (ar, i, ma) notation. How did you come to that decision?

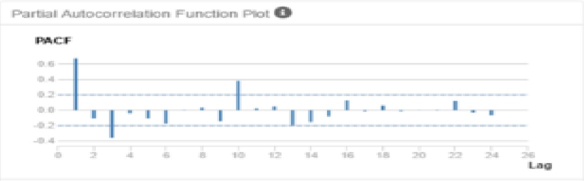
Answer:

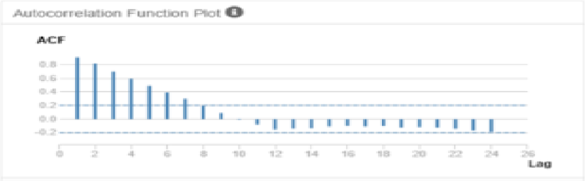
The ETS (M, N, M) without dampening, the seasonal plot is multiplicative that shows an increased periodic fluctuations. There should not be an effect on trend plot and reminder plot is multiplicative since it has irregularities.

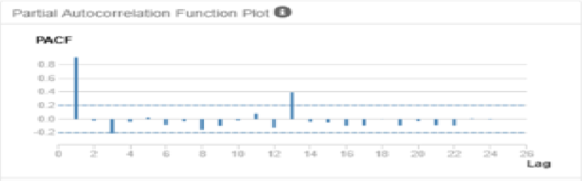


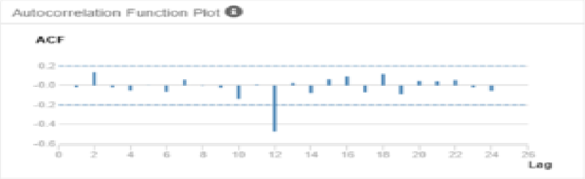
The seasonal difference of ARIMA (0, 1, 2) (0, 1, 0) lagging at 2.

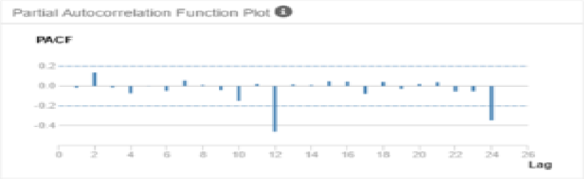

Autocorrelation Function Plot


Partial Autocorrelation Function Plot


Autocorrelation Function Plot

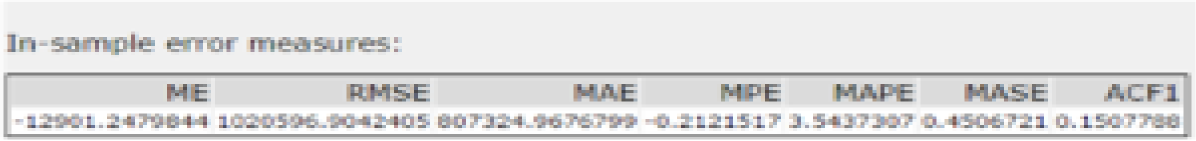

Partial Autocorrelation Function Plot


Autocorrelation Function Plot


Partial Autocorrelation Function Plot

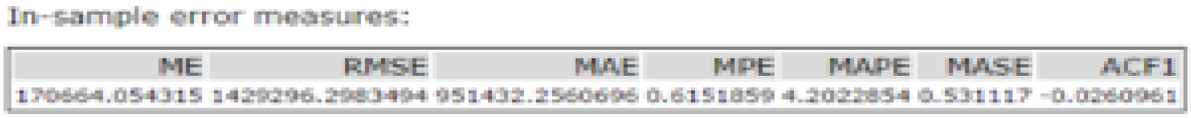
When comparing the ARIMA model and ETS model, the EST model had the best or higher accuracy. With a holdout sample of 6 months.

ETS (m, n, m)


In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

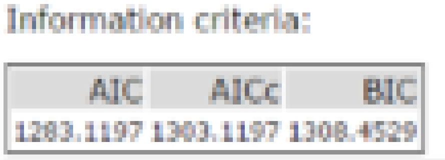
ARIMA (m, n, m)


In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961

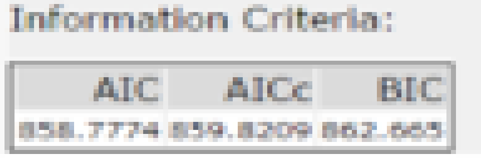
According to the above visualizations, the RMSE and the MASE of ETS had the lowest values (1020596.9042405 and 0.4506721 respectively) when compared to ARIMA (RMSE and MASE with 1429296.2983494 and 0.531117 respectively). But looking at the visualizations below:

EST


Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

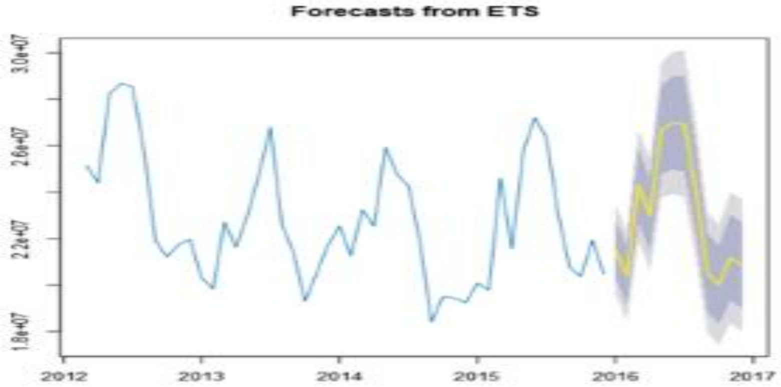
ARIMA

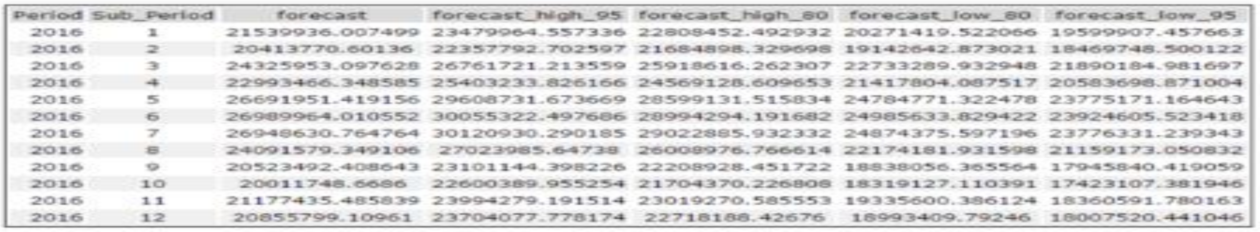

Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

It can be asserts that the ETS has the highest value (128.3.1197) for AIC than the ARIMA of AIC with 858.7774.

Below is the forecast from ETS with a confidence level interval of 80% and 90%.

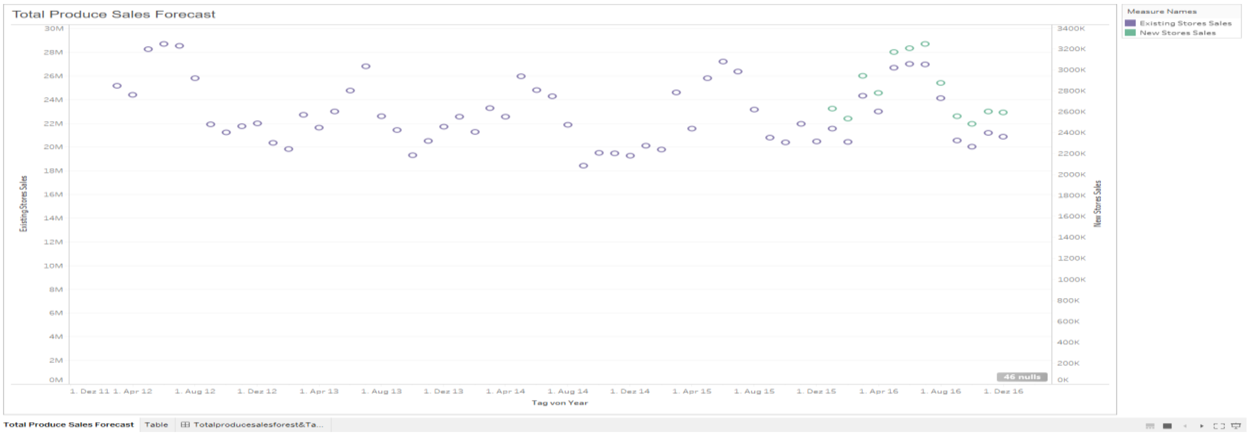


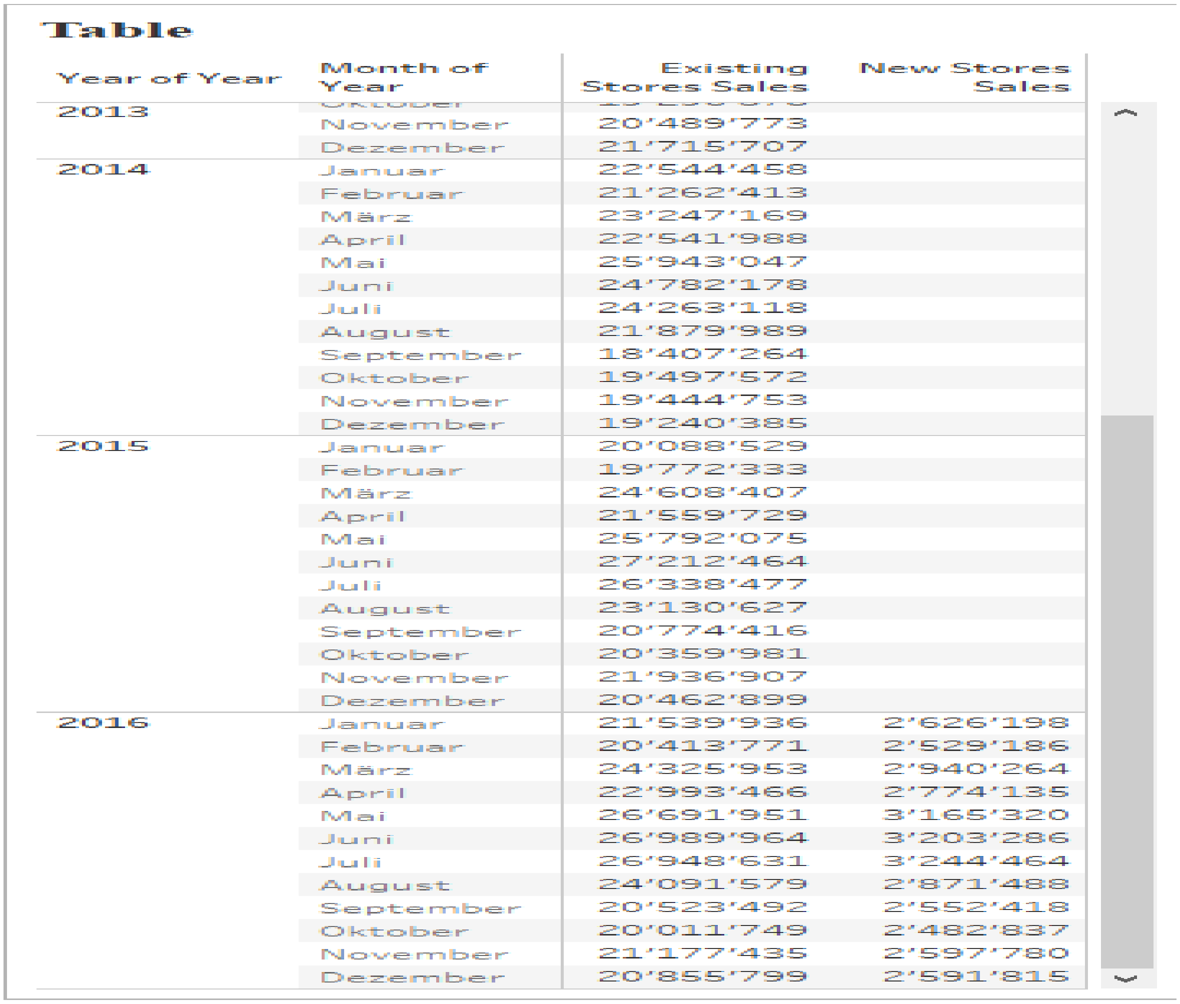


Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	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

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".

Answer:
https://public.tableau.com/profile/benjamin.bediako#!/vizhome/Task3_73/Dashboard1





Table

Year of Year	Month of Year	Existing Stores Sales	New Stores Sales
2013	Oktober	[illegible]	
	November	20'489'773	
	Dezember	21'715'707	
2014	Januar	22'544'458	
	Februar	21'262'413	
	März	23'247'169	
	April	22'541'988	
	Mai	25'943'047	
	Juni	24'782'178	
	Juli	24'263'118	
	August	21'879'989	
	September	18'407'264	
	Oktober	19'497'572	
	November	19'444'753	
	Dezember	19'240'385	
2015	Januar	20'088'529	
	Februar	19'772'333	
	März	24'608'407	
	April	21'559'729	
	Mai	25'792'075	
	Juni	27'212'464	
	Juli	26'338'477	
	August	23'130'627	
	September	20'774'416	
	Oktober	20'359'981	
	November	21'936'907	
	Dezember	20'462'899	
2016	Januar	21'539'936	2'626'198
	Februar	20'413'771	2'529'186
	März	24'325'953	2'940'264
	April	22'993'466	2'774'135
	Mai	26'691'951	3'165'320
	Juni	26'989'964	3'203'286
	Juli	26'948'631	3'244'464
	August	24'091'579	2'871'488
	September	20'523'492	2'552'418
	Oktober	20'011'749	2'482'837
	November	21'177'435	2'597'780
	Dezember	20'855'799	2'591'815

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.