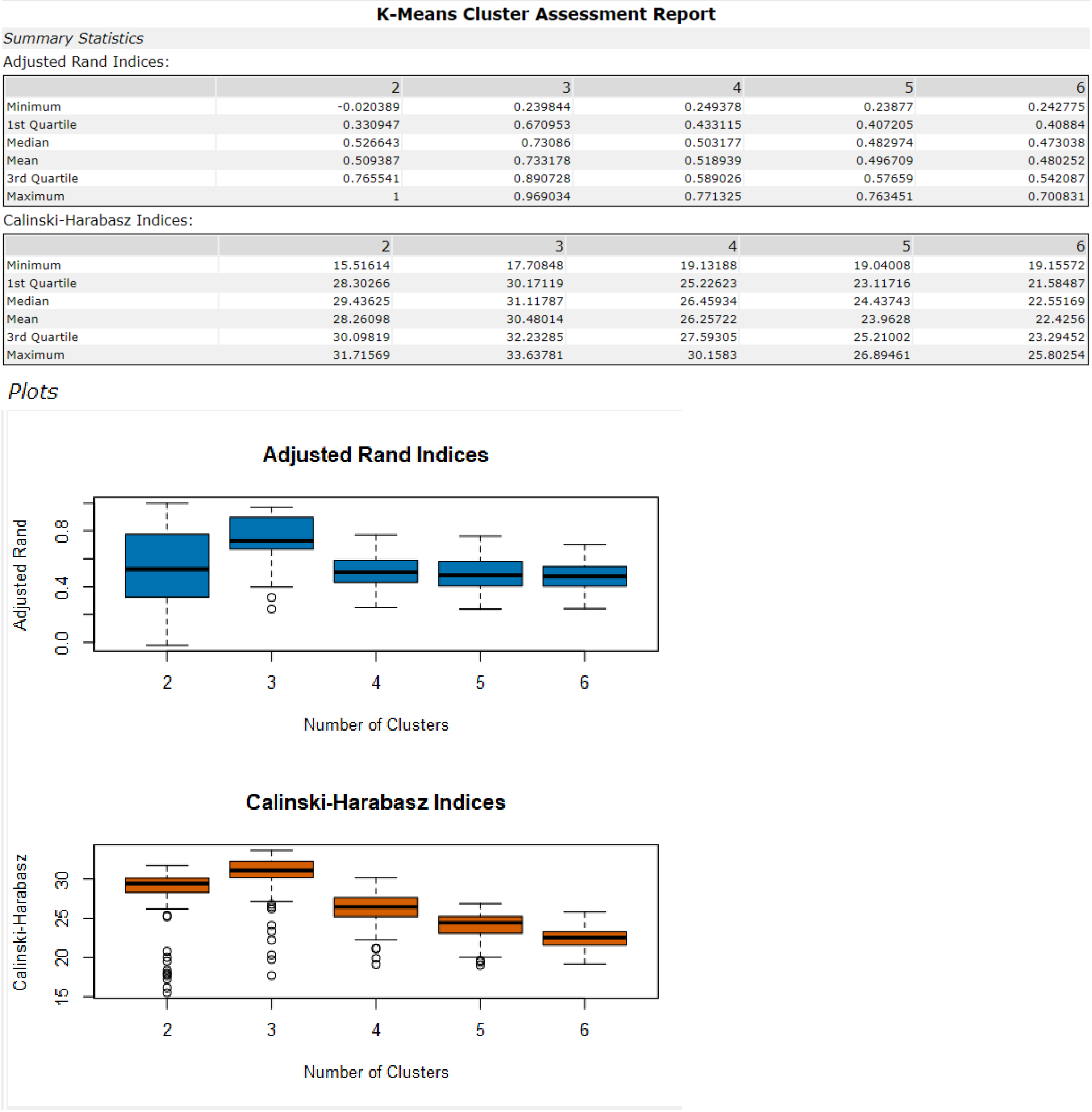
Project: Predictive Analytics Capstone BRIAN BERNS

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?

First I performed an analysis using K-Centroid Diagnostics and using the K-Means report, I concluded that the optimal number of store formats is 3. This is also because both indices on the Adjusted Rand and the Calinski-Harabasz have the highest median value. (As Shown Below)



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.020389	0.239844	0.249378	0.23877	0.242775
1st Quartile	0.330947	0.670953	0.433115	0.407205	0.40884
Median	0.526643	0.73086	0.503177	0.482974	0.473038
Mean	0.509387	0.733178	0.518939	0.496709	0.480252
3rd Quartile	0.765541	0.890728	0.589026	0.57659	0.542087
Maximum	1	0.969034	0.771325	0.763451	0.700831

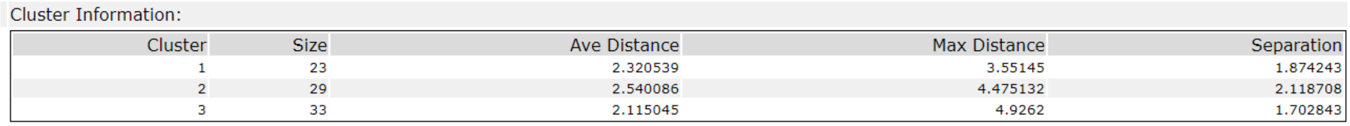
Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.51614	17.70848	19.13188	19.04008	19.15572
1st Quartile	28.30266	30.17119	25.22623	23.11716	21.58487
Median	29.43625	31.11787	26.45934	24.43743	22.55169
Mean	28.26098	30.48014	26.25722	23.9628	22.4256
3rd Quartile	30.09819	32.23285	27.59305	25.21002	23.29452
Maximum	31.71569	33.63781	30.1583	26.89461	25.80254

Plots

2. How many stores fall into each store format?

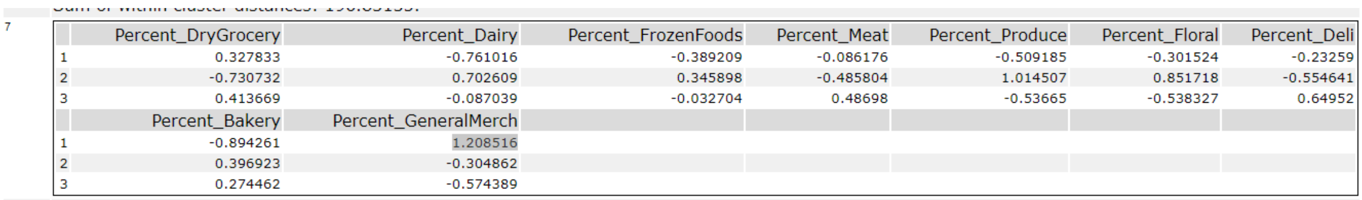
Cluster 1 has 23 Stores. Cluster 2 has 29 stores. Cluster 3 has 33 stores.



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

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



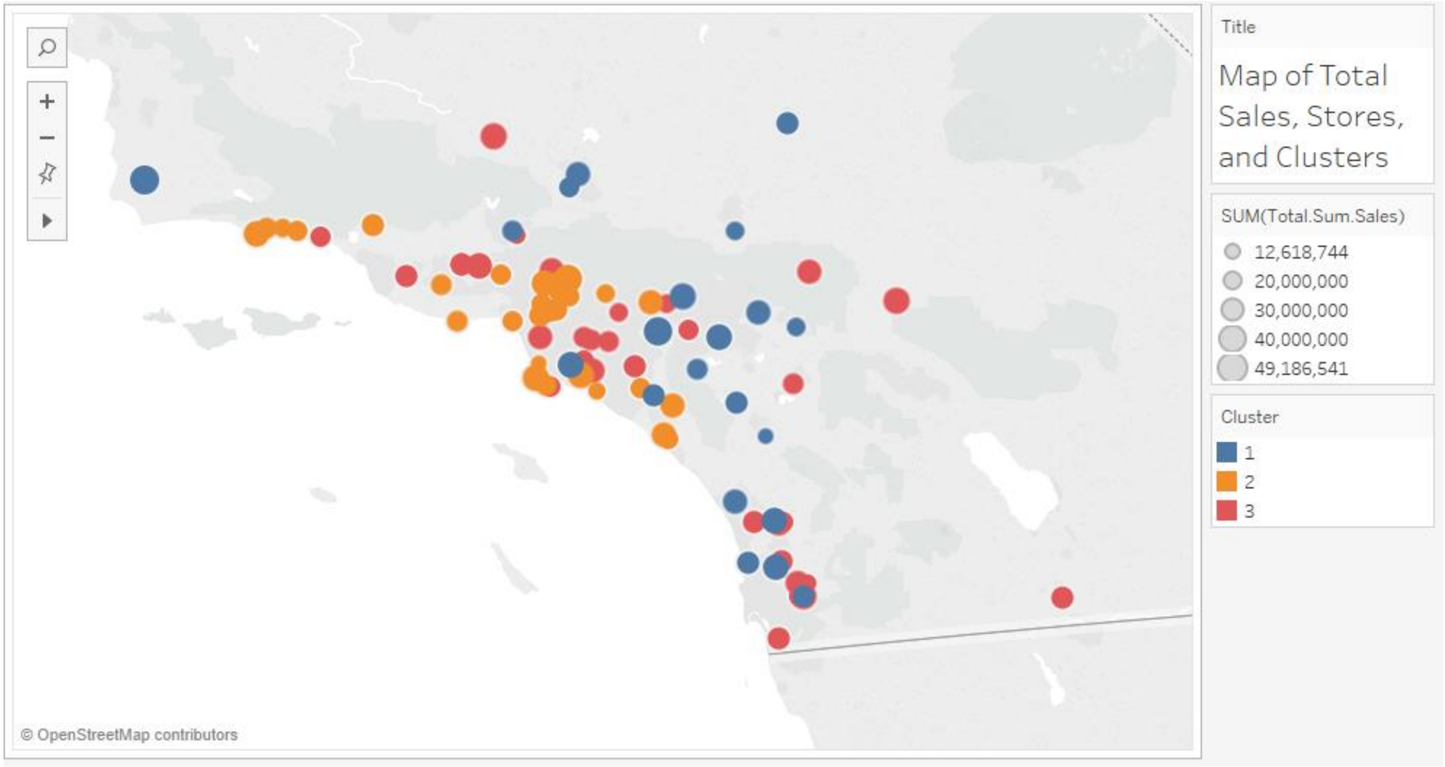
Sum of within cluster distances: 196.65155.

7

	Percent_DryGrocery	Percent_Dairy	Percent_FrozenFoods	Percent_Meat	Percent_Produce	Percent_Floral	Percent_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
	Percent_Bakery	Percent_GeneralMerch					
1	-0.894261	1.208516					
2	0.396923	-0.304862					
3	0.274462	-0.574389					

Cluster 1 sold more merchandise in terms on percentage over the other two clusters, but Cluster 2 sold the most Produce in terms of percentage over the rest.

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.

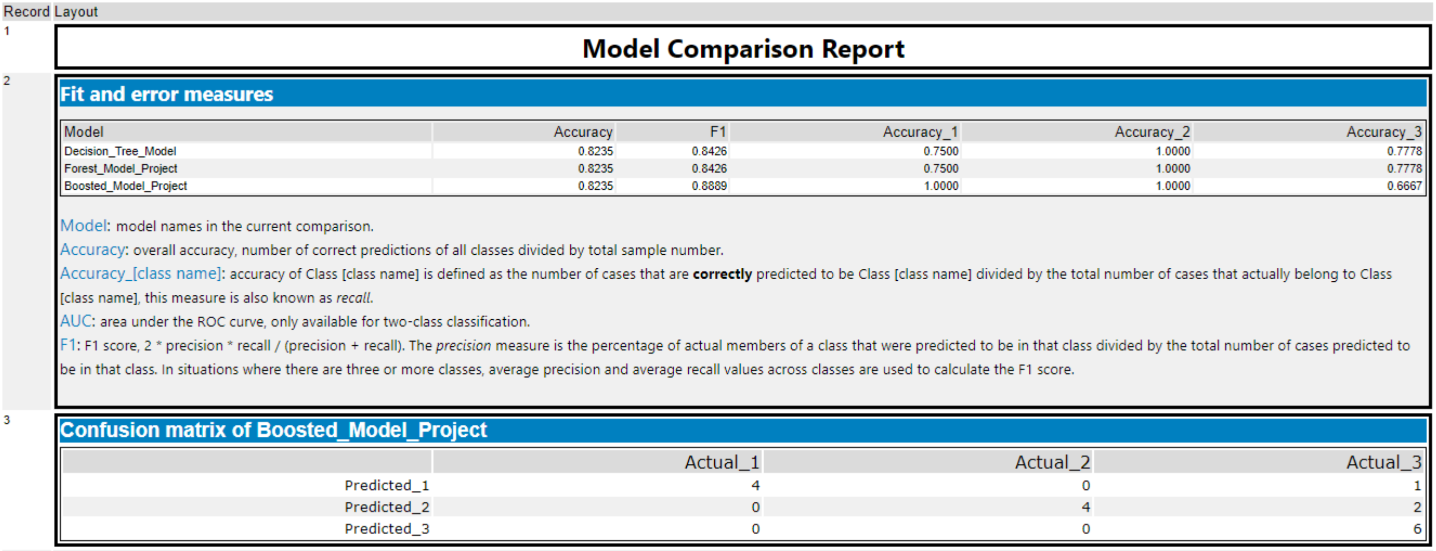


https://public.tableau.com/profile/brian.berns#!/vizhome/ClusterPart1Project/Sheet1?publish=yes

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.)

   I used a model comparison report shows the comparison matrix between the Decision Tree, Forest Model, and the Boosted Model. Even though all models have the same accuracy, I chose the Boosted Model due to the highest F1 value.



Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Forest_Model_Project	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model_Project	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*.

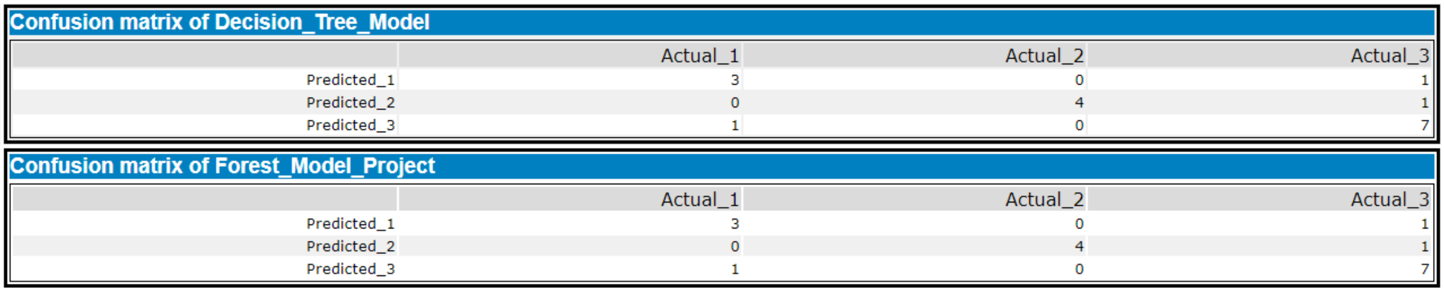
AUC: area under the ROC curve, only available for two-class classification.

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.

3

Confusion matrix of Boosted_Model_Project

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



Confusion matrix of Decision_Tree_Model

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

Confusion matrix of Forest_Model_Project

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

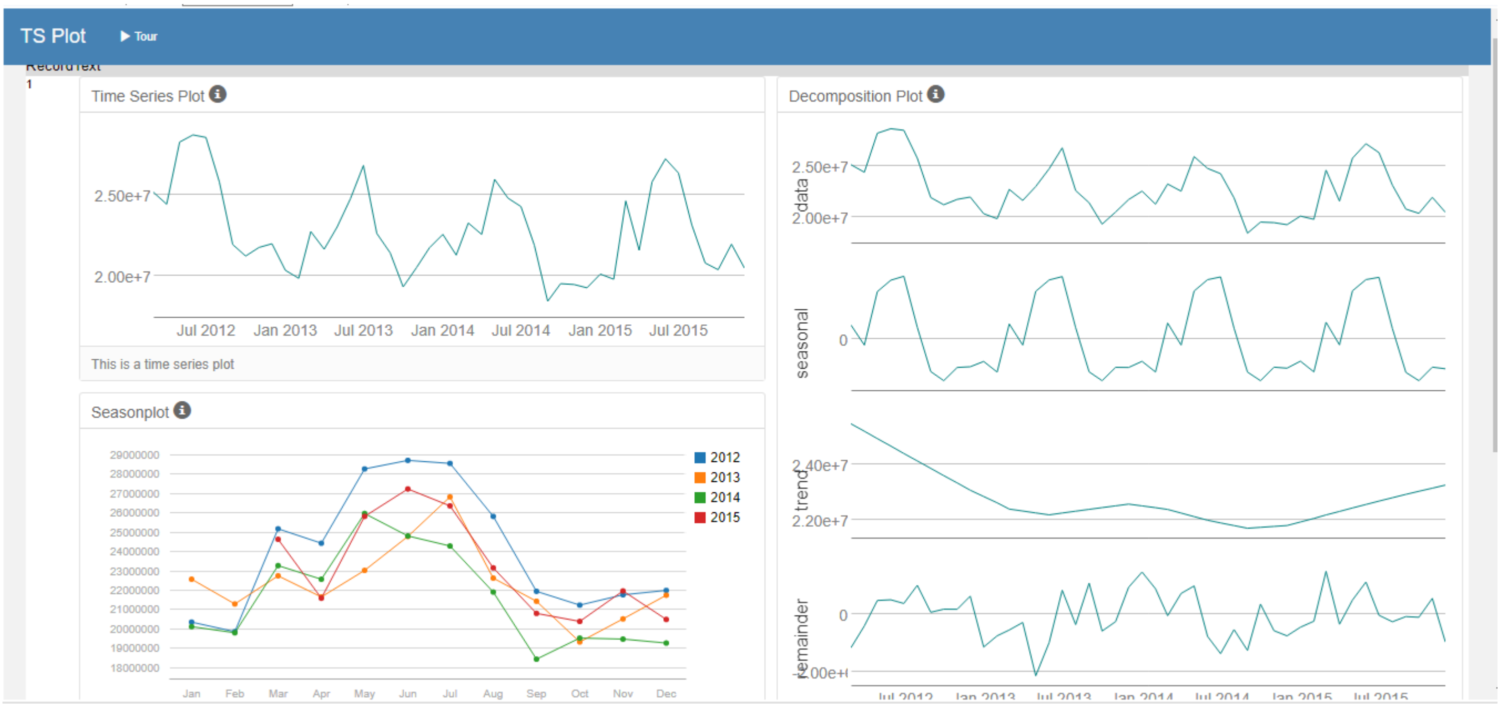
2. What format do each of the 10 new stores fall into? Please fill in the table below.

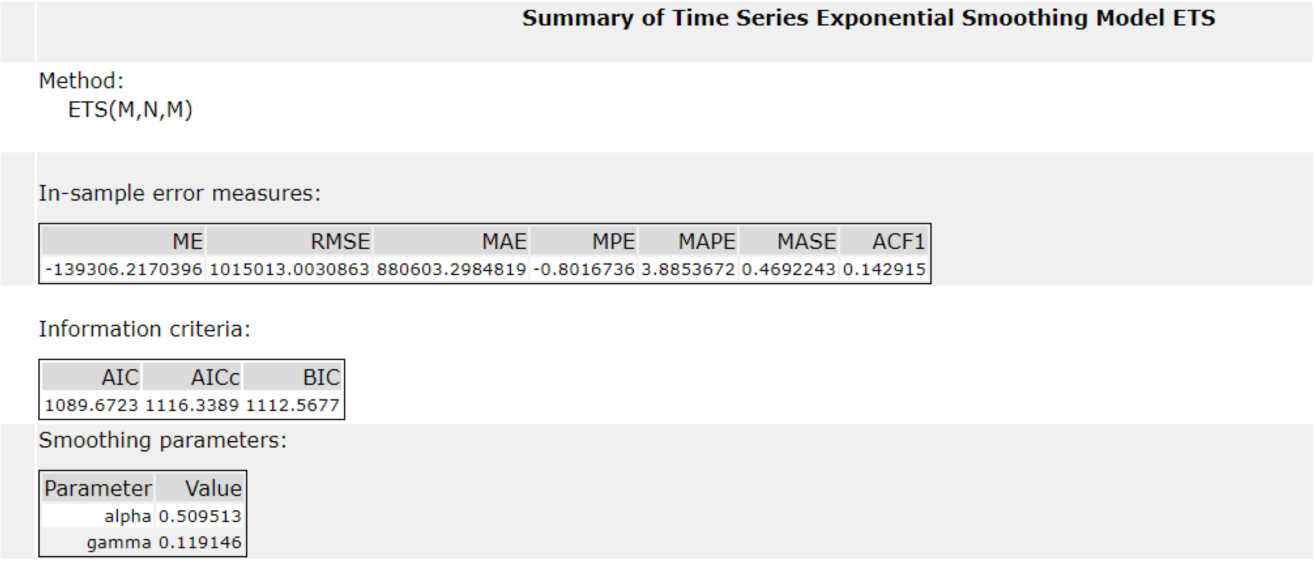
Store Number	Segment
S0086	**3**
S0087	**2**
S0088	**1**
S0089	**2**
S0090	**2**
S0091	**1**
S0092	**2**
S0093	**1**
S0094	**2**
S0095	**2**

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 the ETS model, I used ETS(M,N,M) with no dampening. I noticed a couple things when looking at the TS Plot in Alteryx. The error has an irregular pattern, so it should be applied multiplicatively. There is no trend, or it is not clear, so it should not be applied. Finally, the seasonality shows increasing peaks and valleys so it should also be applied multiplicatively.





Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,N,M)

In-sample error measures:

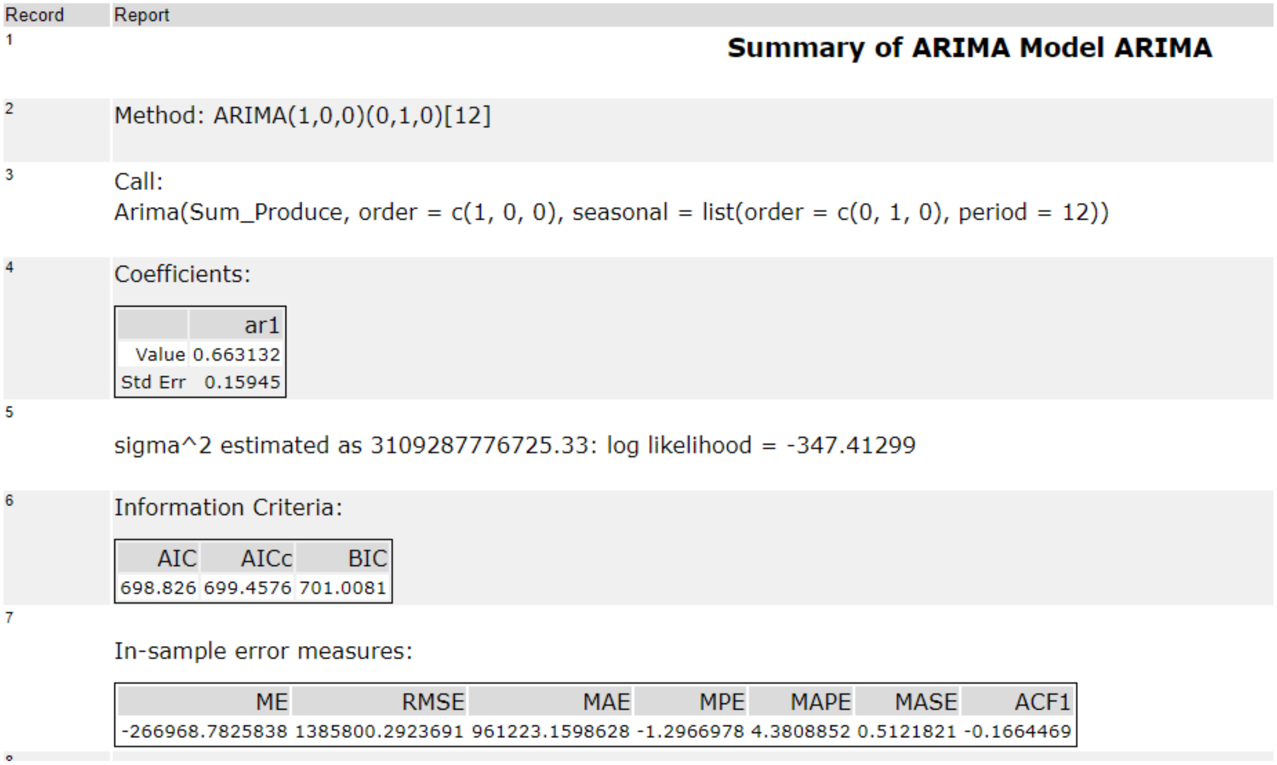
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-139306.2170396	1015013.0030863	880603.2984819	-0.8016736	3.8853672	0.4692243	0.142915

Information criteria:

AIC	AICc	BIC
1089.6723	1116.3389	1112.5677

Smoothing parameters:

Parameter	Value
alpha	0.509513
gamma	0.119146



Record	Report
1	**Summary of ARIMA Model ARIMA**
2	Method: ARIMA(1,0,0)(0,1,0)[12]
3	Call: Arima(Sum_Produce, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 0), period = 12))
4	Coefficients:
5	sigma^2 estimated as 3109287776725.33: log likelihood = -347.41299
6	Information Criteria:
7	In-sample error measures:

Coefficients:

	ar1
Value	0.663132
Std Err	0.15945

Information Criteria:

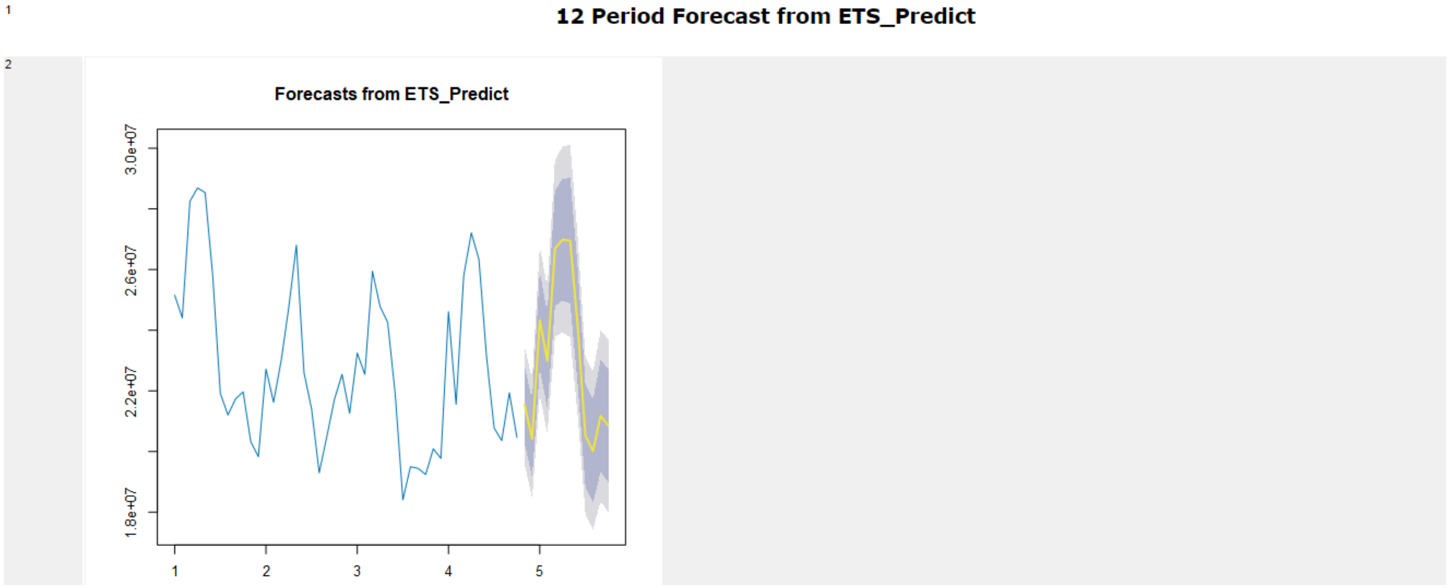
AIC	AICc	BIC
698.826	699.4576	701.0081

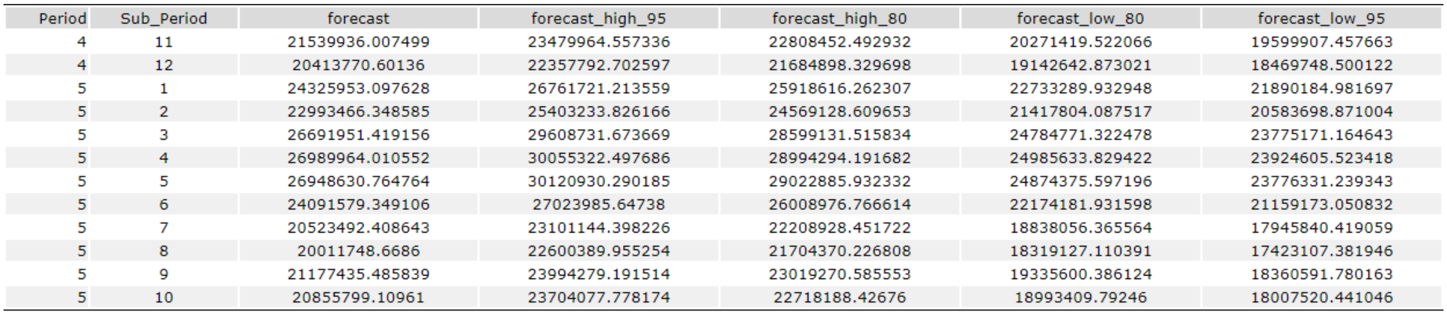
In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-266968.7825838	1385800.2923691	961223.1598628	-1.2966978	4.3808852	0.5121821	-0.1664469

First of all, I used a holdout sample of 12 months. ETS model's accuracy is higher than the ARIMA model. When I compare MASE values, ETS = .46 and ARIMA = .51. When I compare AIC values, ETS = 1089 and ARIMA = 698. Finally, when I compare RMSE values, ETS = 1015013 and ARIMA = 1385800.

The graph shows the forecast values, and actual values between 80-95% confidence levels.

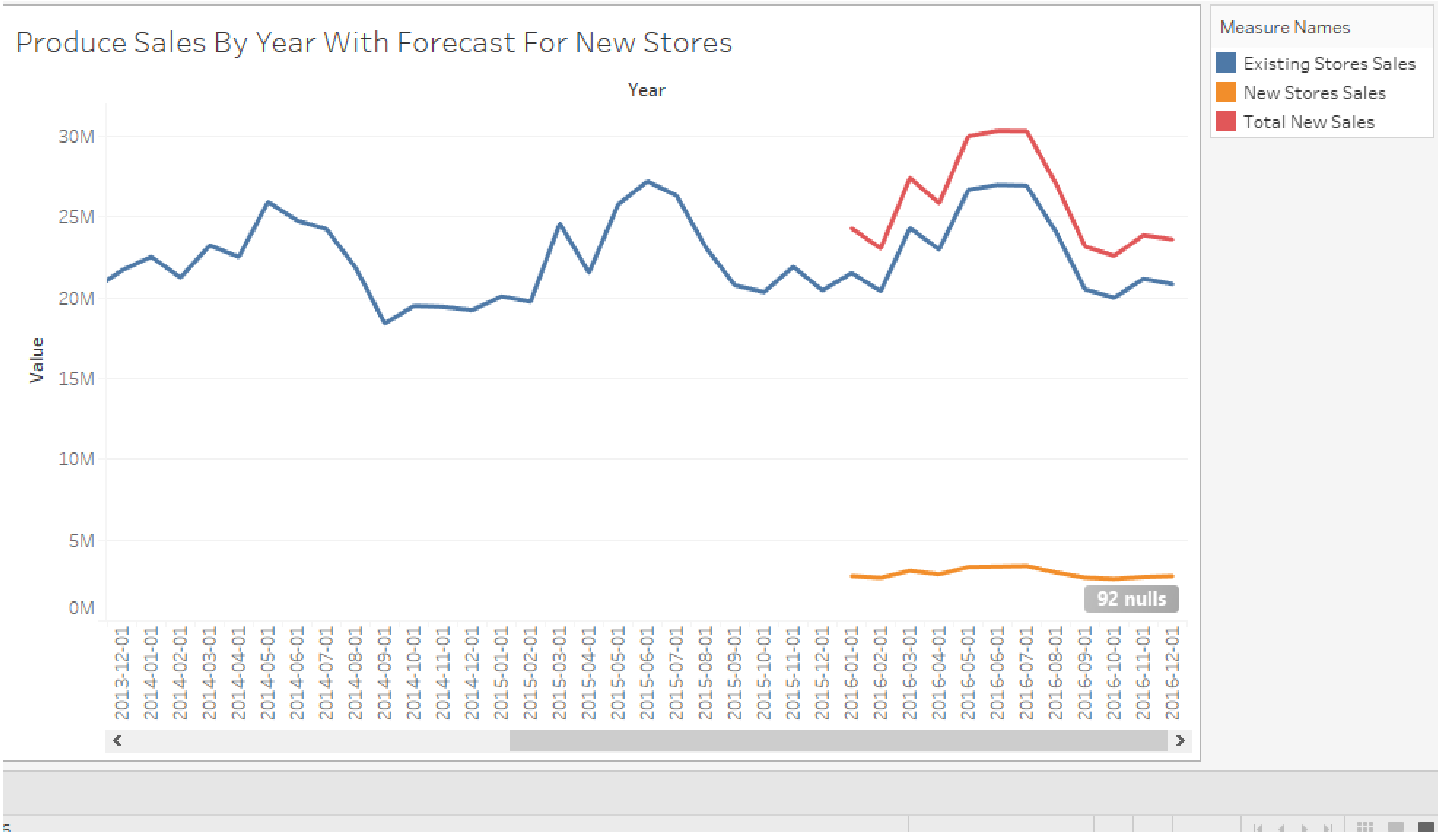




Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
4	11	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
4	12	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
5	1	24325953.097628	26761721.213559	25918616.262307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

Month	Existing Store Sales	New Store Sales
1	**$21,539,936**	**$2,761,958**
2	**$20,413,770**	**$2,656,665**
3	**$24,325,953**	**$3,099,057**
4	**$22,993,466**	**$2,873,607**
5	**$26,691,951**	**$3,327,835**
6	**$26,989,964**	**$3,356,062**
7	**$26,948,630**	**$3,391,942**
8	**$24,091,579**	**$2,991,382**
9	**$20,523,492**	**$2,664,295**
10	**$20,011,748**	**$2,588,209**
11	**$21,177,435**	**$2,702,838**
12	**$20,855,799**	**$2,761,943**



https://public.tableau.com/profile/brian.berns#!/vizhome/Task3_266/Sheet5?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.