

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?

The optimal number of store formats is 3.

Using the Alteryx K-Diagnostic tool I analyzed the effects of applying various numbers of clusters to the data set and used the generated Adjusted Rand Indices and Calinski-Harabasz Indices to arrive at the optimal number of clusters.

As depicted in the diagrams, although 2 clusters had higher mean, it was best to compromise and select 3 clusters since it had a more compact interquartile range and had higher mean than all the other numbers of clusters that followed.

K-Means Cluster Assessment Report

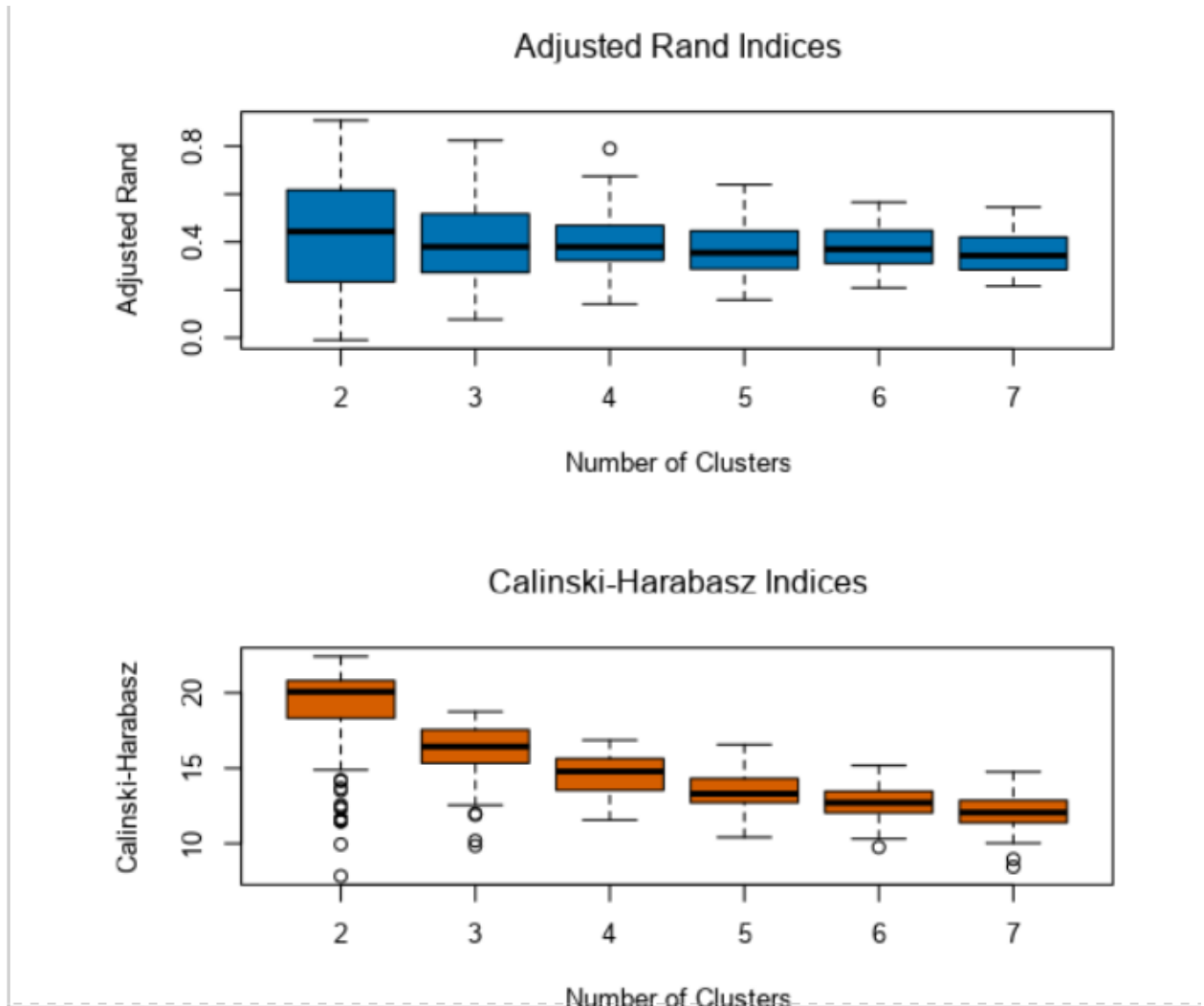
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	-0.009675	0.076235	0.140656	0.157999	0.208442	0.215517
1st Quartile	0.237245	0.273359	0.324062	0.28911	0.310322	0.283793
Median	0.443127	0.379958	0.379205	0.354445	0.369622	0.343121
Mean	0.42889	0.410693	0.396973	0.372638	0.379017	0.355602
3rd Quartile	0.607523	0.513414	0.465973	0.444893	0.445965	0.419453
Maximum	0.907005	0.823811	0.789549	0.639632	0.565878	0.54505

Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	7.838511	9.845155	11.56778	10.41516	9.754192	8.452392
1st Quartile	18.329049	15.370633	13.54646	12.70601	12.042498	11.37346
Median	20.072097	16.43124	14.78233	13.31046	12.703326	12.062457
Mean	18.866108	16.214792	14.61573	13.44934	12.742937	12.071297
3rd Quartile	20.790946	17.532122	15.63393	14.31965	13.470937	12.865362
Maximum	22.415549	18.750421	16.86351	16.57168	15.173243	14.756313



2. How many stores fall into each store format?

Cluster Information:

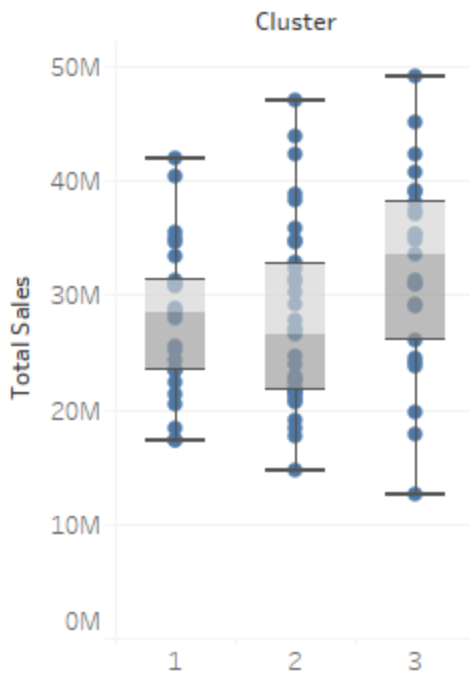
Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Cluster 1 had 25 stores, cluster 2 had 35 stores and cluster 3 had 25 stores.

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

The total sales in cluster 1 were the least varying as depicted by lowest range whereas cluster 2 total sales varied moderately and the highest variation in total sales was in cluster 3 as depicted by the highest range below.

Sheet 2

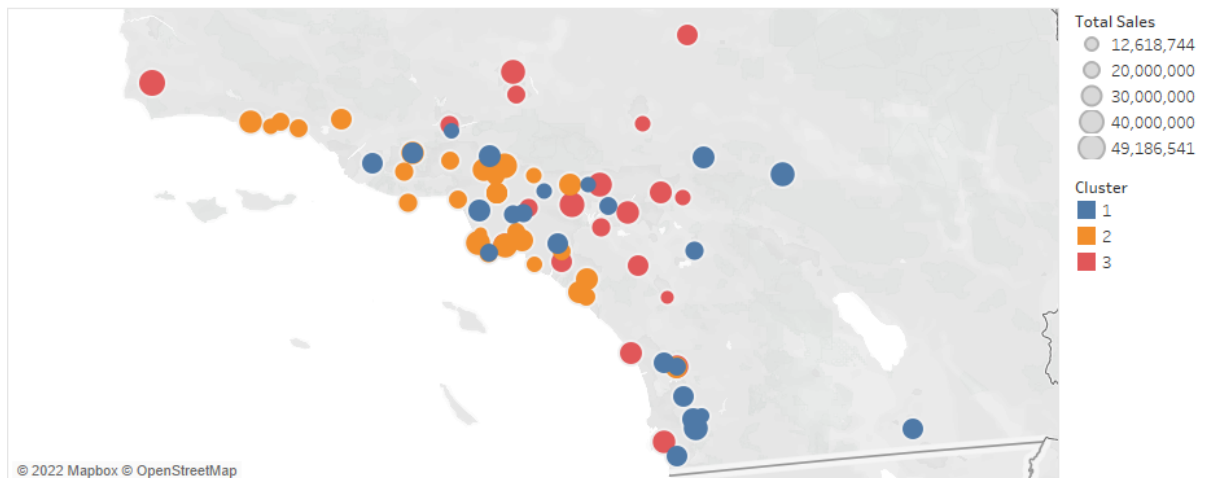


Total Sales for each Cluster.

- 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/app/profile/kevin2888/viz/proj6_16453872356710/Sheet1?publish=yes

Location of the existing stores in cities in California



Map based on Longitude (generated) and Latitude (generated). Color shows details about Cluster. Size shows Total Sales. Details are shown for various dimensions.

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 Forest Model and Boosted Model were the best performing model over the decision tree model. The Forest Model and Boosted Model had the same accuracy levels and similar confusion matrices as depicted below. I choose the Forest model to predict the best store formats for the new stores.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Decision_Tree_	0.6471	0.6667	0.5000	1.0000	0.5000
Boosted_Model	0.7059	0.7500	0.5000	1.0000	0.7500

Confusion matrix of Boosted_Model			
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	2	5	0
Predicted_3	2	0	3

Confusion matrix of Decision_Tree_			
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	2
Predicted_2	3	5	0
Predicted_3	1	0	2

Confusion matrix of Forest_Model			
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	2	5	0
Predicted_3	2	0	3

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

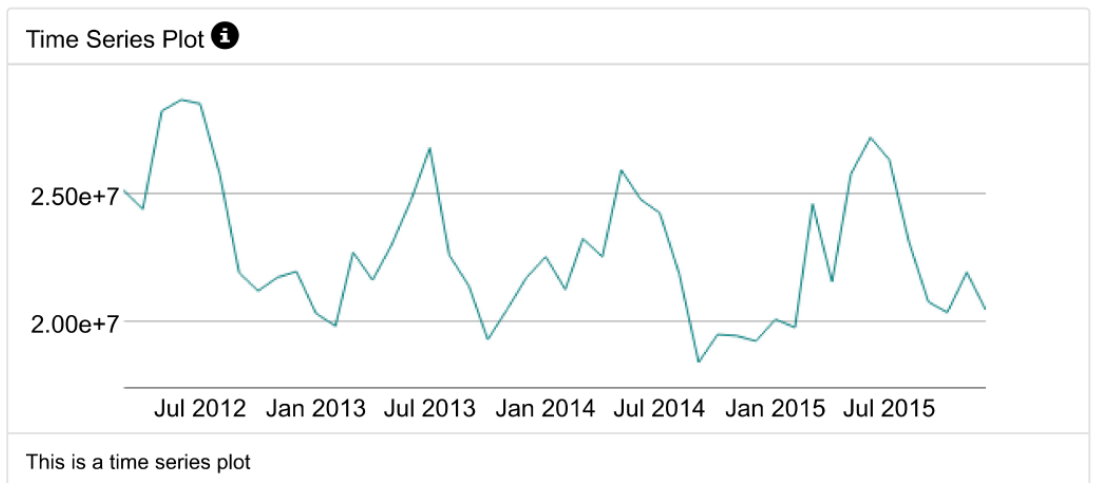
Store Number	Segment
S0086	1
S0087	2
S0088	1
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

Task 3: Predicting Produce Sales

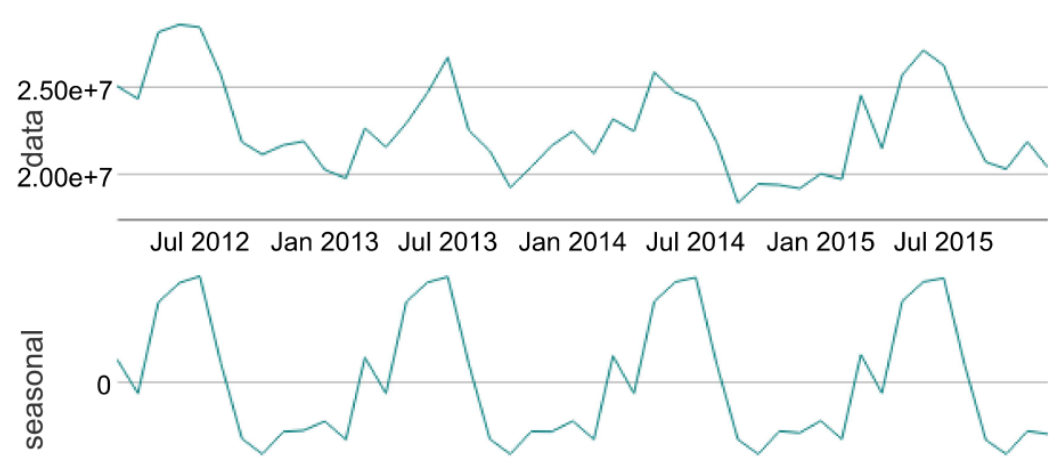
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

Record
1

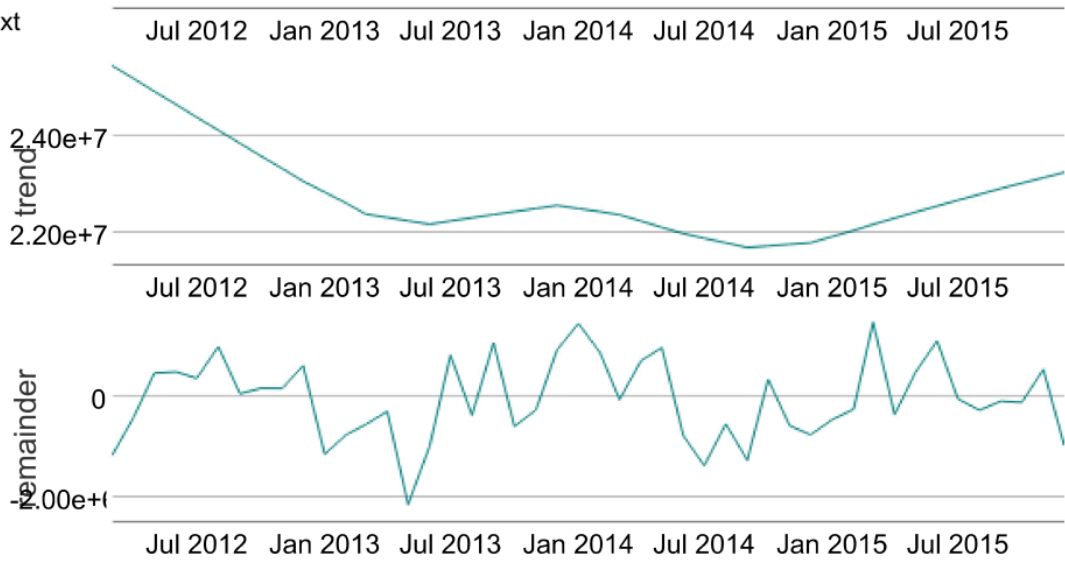
Text



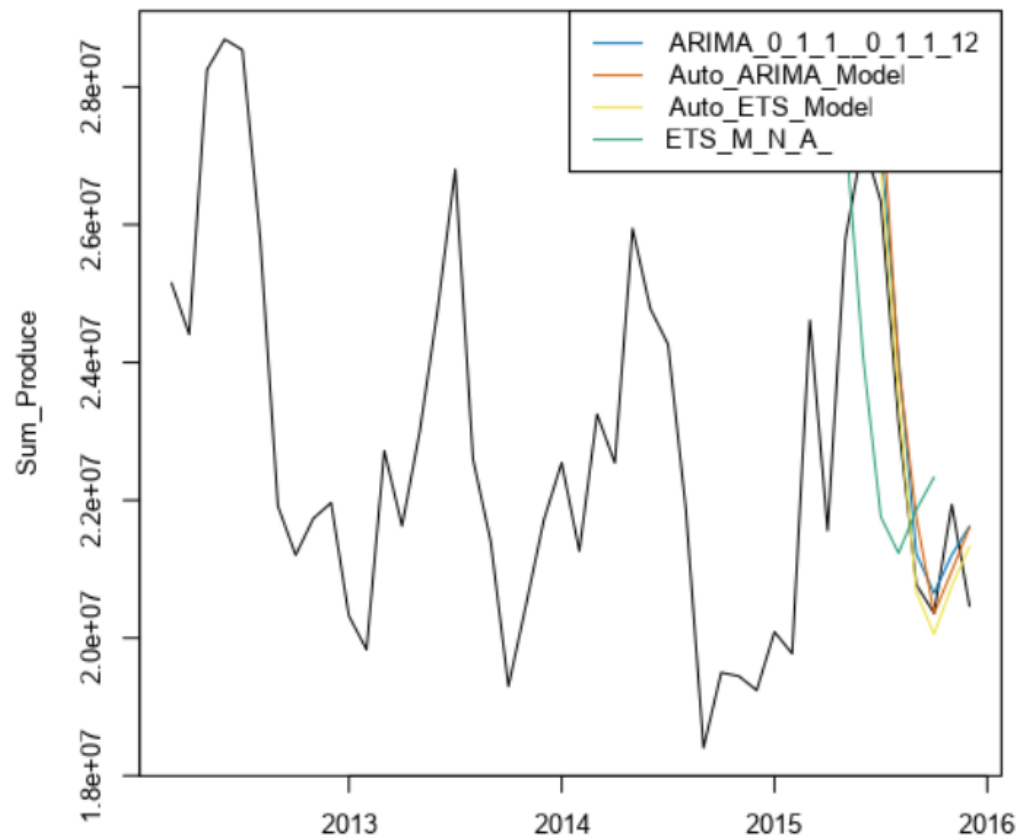
Decomposition Plot 



Text



Actual and Forecast Values



Actual and Forecast Values:

Actual	ARIMA_0_1_1_0_1_1_12	Auto_ARIMA_Model	Auto_ETS_Model	ARIMA_0_1_1_0_1_2_12
26338477.15	27182961.17184	27997835.66704	26860639.57443	28106648.91535
23130626.6	24073582.2774	23946058.0479	23468254.49595	24546897.01689
20774415.93	21223756.44966	21751347.9177	20668464.64495	22098517.83001
20359980.58	20648299.23877	20352513.12727	20054544.07631	20510896.2131
21936906.81	21205988.81563	20971835.14524	20752503.51996	21441137.38521
20462899.3	21622151.4136	21609110.45558	21328386.80965	22239453.70879

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_0_1_1_0_1_1_12	-492238.83	792197.3	735878.2	-2.1992	3.3098	0.433
Auto_ARIMA_Model	-604232.33	1050239.2	928412	-2.6156	4.0942	0.5463
Auto_ETS_Model	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA_0_1_1_0_1_2_12	-990040.78	1310865.3	1155297.3	-4.3955	5.1488	0.6798

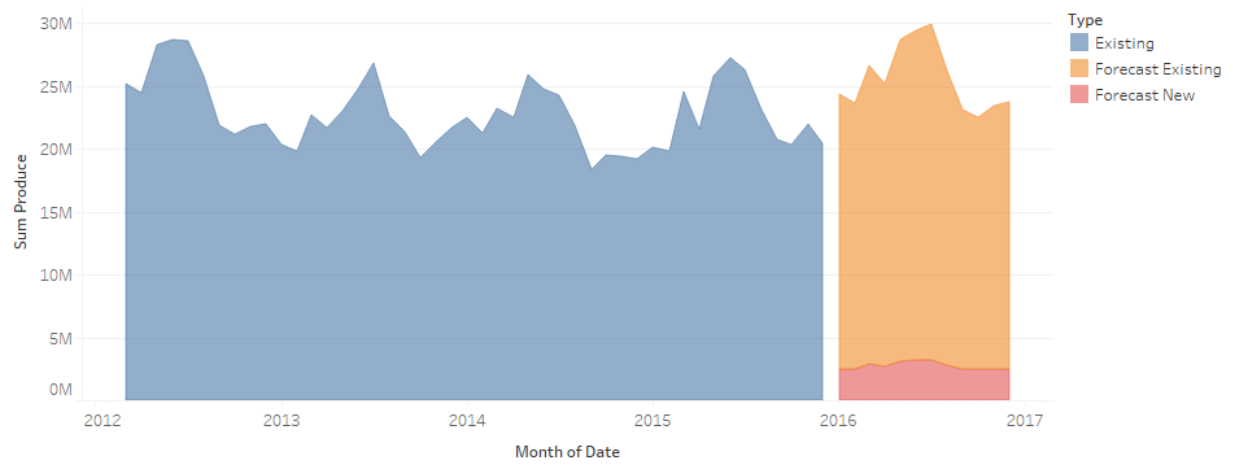
I selected ETS(M,N,M) that resulted from running an ETS tool with Auto configurations. ETS(M,N,M)/Auto_ET_S_Model as shown above performed best when compared to other models using the TS Compare tool as it had the least error measures.

2. 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 Stores Forecast	New Stores Forecast
Jan-16	21829060.03	2527338.50
Feb-16	21146329.63	2446154.76
Mar-16	23735686.94	2872050.73
Apr-16	22409515.28	2722157.62
May-16	25621828.73	3098095.87
Jun-16	26307858.04	3150602.99
Jul-16	26705092.56	3172545.05
Aug-16	23440761.33	2814269.98
Sep-16	20640047.32	2486631.56
Oct-16	20086270.46	2434261.23
Nov-16	20858119.96	2517523.25
Dec-16	21255190.24	2491340.44

https://public.tableau.com/app/profile/kevin2888/viz/Sales_Forecast_Viz/Sheet1?publish=yes

Sheet 1



The plot of Sum Produce for Month of Date. Color shows details about Type.