

Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it

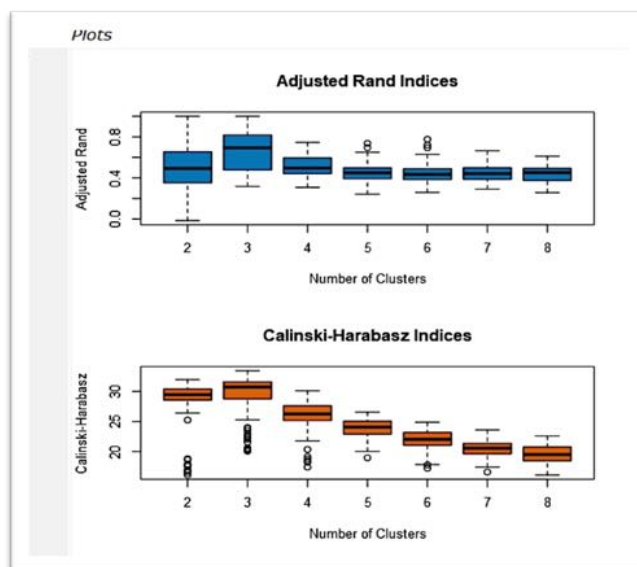
here: <https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project>

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 three. The conducted steps started by getting the AR and CH indices (via the K-Centroids Diagnostics as shown below) and thereby the K-Means are selected methodology.

K-Means Cluster Assessment Report								
Summary Statistics								
Adjusted Rand Indices:								
	2	3	4	5	6	7	8	
Minimum	-0.0152	0.3171	0.3072	0.2412	0.2586	0.2903	0.2568	
1st Quartile	0.352	0.4819	0.4431	0.3943	0.3896	0.3877	0.377	
Median	0.4926	0.6936	0.4964	0.4487	0.4348	0.4417	0.4526	
Mean	0.484	0.6575	0.5125	0.4623	0.4532	0.4498	0.4411	
3rd Quartile	0.655	0.816	0.5913	0.4982	0.489	0.4997	0.491	
Maximum	1	1	0.7458	0.7366	0.7762	0.6637	0.6118	
Calinski-Harabasz Indices:								
	2	3	4	5	6	7	8	
Minimum	16.1	20.09	17.41	18.98	17.24	16.61	16.11	
1st Quartile	28.61	28.76	25.16	22.91	21.05	19.61	18.46	
Median	29.47	30.7	26.25	24.05	22.02	20.56	19.5	
Mean	28.41	29.47	25.99	23.88	21.96	20.48	19.62	
3rd Quartile	30.39	31.58	27.62	25.06	23.14	21.35	20.77	
Maximum	31.95	33.41	30.09	26.53	24.87	23.6	22.59	



The highest AR of 0.6936 and CH of 29.47 indices are the results of a three-cluster solution.

2. How many stores fall into each store format?

By conducting a cluster analysis (using the K-Means method), the number of stores into each store format is as follows:

Cluster 1 = 23 stores

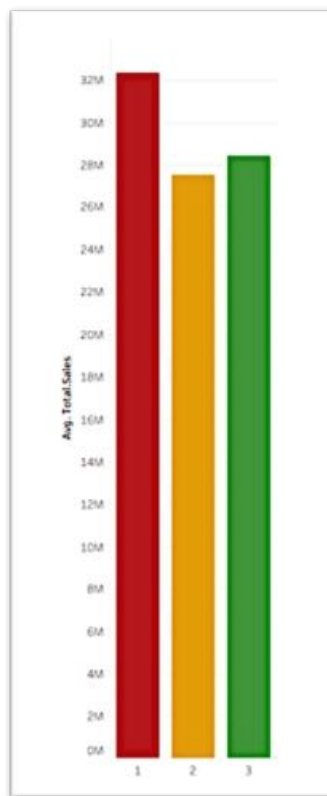
Cluster 2 = 29 stores

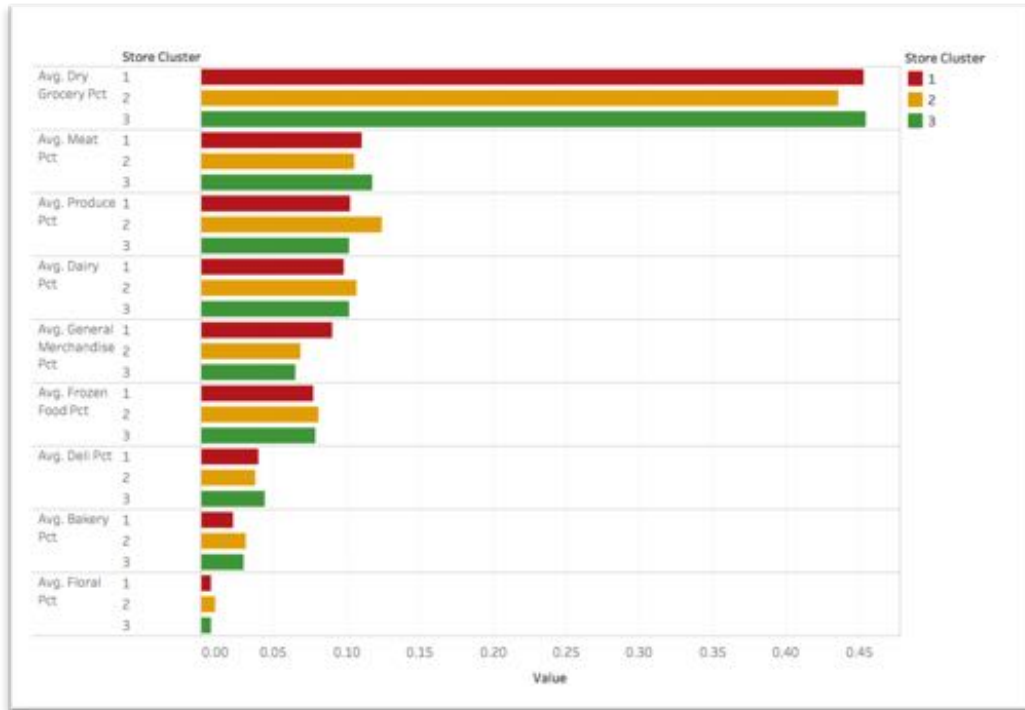
Cluster 3 = 33 stores

Hence each one of the existing stores was associated with the appropriate cluster.

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

One way that the clusters differ from one another is the average total sales. As shown in the visualization below. Other differences are shown in the second graph below:





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.

Here is the link:

https://public.tableau.com/profile/mohammed.assiri5000#!/vizhome/Task1_79/Task1?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.)

As a non-binary classification problem, three distinct methodologies can be used. However, to determine the most appropriate one, the following steps were followed. First, data preparation of existing and new stores to include demographic data. Then, model construction of three namely, decision tree model, forest model, and boosted model. A 20% validation sample was ensured.

The comparison of the three model is shown in the table below:

Fit and error measures					
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Forest_	0.8235	0.8251	0.7500	0.8000	0.8750

Hence, it can be concluded that, while all model performed equally regarding the validation sample, the boosted model was the most appropriate to use as it has the highest F1 score than the other two models.

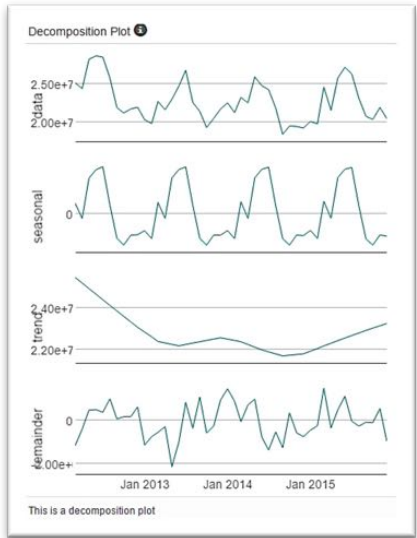
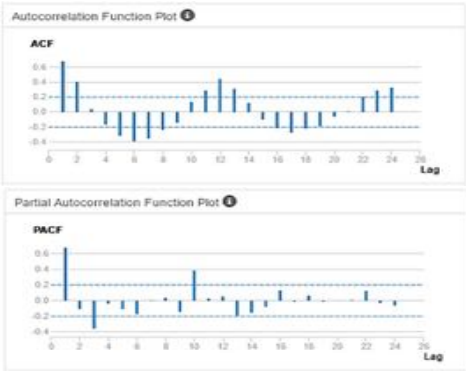
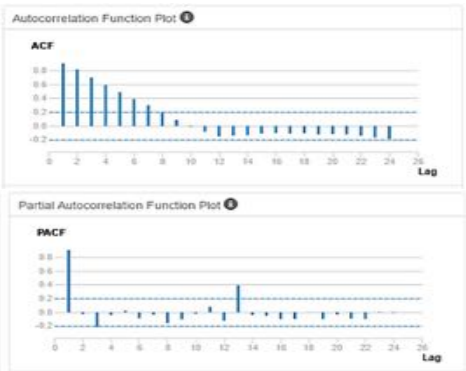
2. What format do each of the 10 new stores fall into? Please fill in the table below.

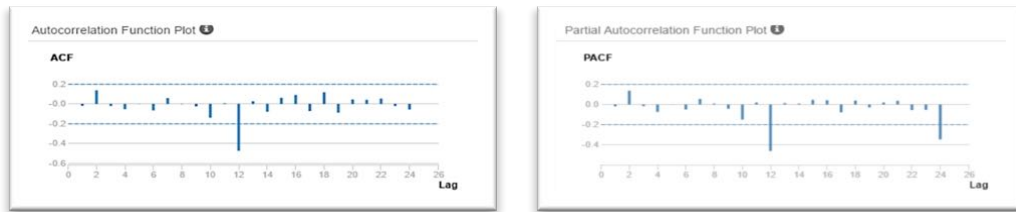
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?

In the next table, both ETS and ARIMA models are discussed.

<p>In this ETS, the error is irregular (i.e. multiplicatively(M)), the trend is not clear (i.e. none (N)), and the seasonal is increasing (i.e. multiplicatively(M))</p> <p>Therefore, ETS is ETS(M, N, M) model</p>	 <p>This is a decomposition plot</p>
<p>After non-seasonal and seasonal differencing, the configuration is as follows:</p> <ul style="list-style-type: none"> • Non-seasonal: <ul style="list-style-type: none"> ○ AR = 0 ○ I = 0 ○ MA = 2 <ul style="list-style-type: none"> ▪ since ACF is negative at lag-1 and hence MA term should be used • seasonal: <ul style="list-style-type: none"> ○ AR = 0 ○ I = 1 ○ MA = 0 <ul style="list-style-type: none"> ▪ since differencing is one seasonal • m = 12 (as the sample) <p>Hence, ARIMA(0,1,2)(0,1,0) 12 model</p>	<div> <p>ACF and PACF of ARIMA (non-seasonal)</p>  </div> <div> <p>ACF and PACF of ARIMA (seasonal)</p>  </div>



ACF and PACF of ARIMA after differencing

Now, the comparison between both models is as follows:

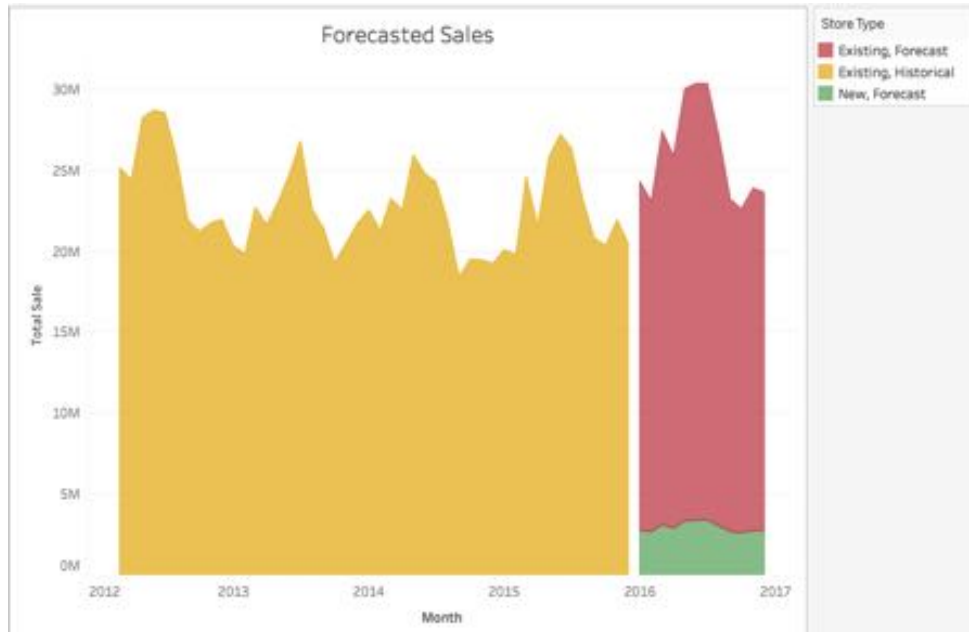
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	1068766	1590916	1337409	4.372	5.7523	0.833	NA
ARIMA	1303043	2181554	1874870	5.3495	8.2524	1.1678	NA

It can be noticed that ETS performs overly better than ARIMA with respect to the sample. Therefore, the forecast should be **ETS(M, N, M)**.

- 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".

Month	Grand Total	Existing	New
1	24,301,894	21,539,936	2,761,958
2	23,070,436	20,413,771	2,656,665
3	27,425,011	24,325,953	3,099,058
4	25,867,074	22,993,466	2,873,607
5	30,019,787	26,691,951	3,327,835
6	30,346,026	26,989,964	3,356,062
7	30,340,574	26,948,631	3,391,943
8	27,082,962	24,091,579	2,991,383
9	23,187,788	20,523,492	2,664,295
10	102,957,652	100,369,442	2,588,210
11	107,487,866	104,785,028	2,702,838
12	106,999,710	104,237,767	2,761,943



Here is the link:

https://public.tableau.com/profile/mohammed.assiri3979#!/vizhome/Task3v3_0/Task3?publish=yes