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?

Ans: The optimal number of store format is 3.

After cleaning the data with requirement, we used cluster diagnostics tool in order to identify the optimal number of store formats. We used K-Means clustering for analysis with minimum number of 3 clusters and maximum number of 8 clusters was used for analysis.

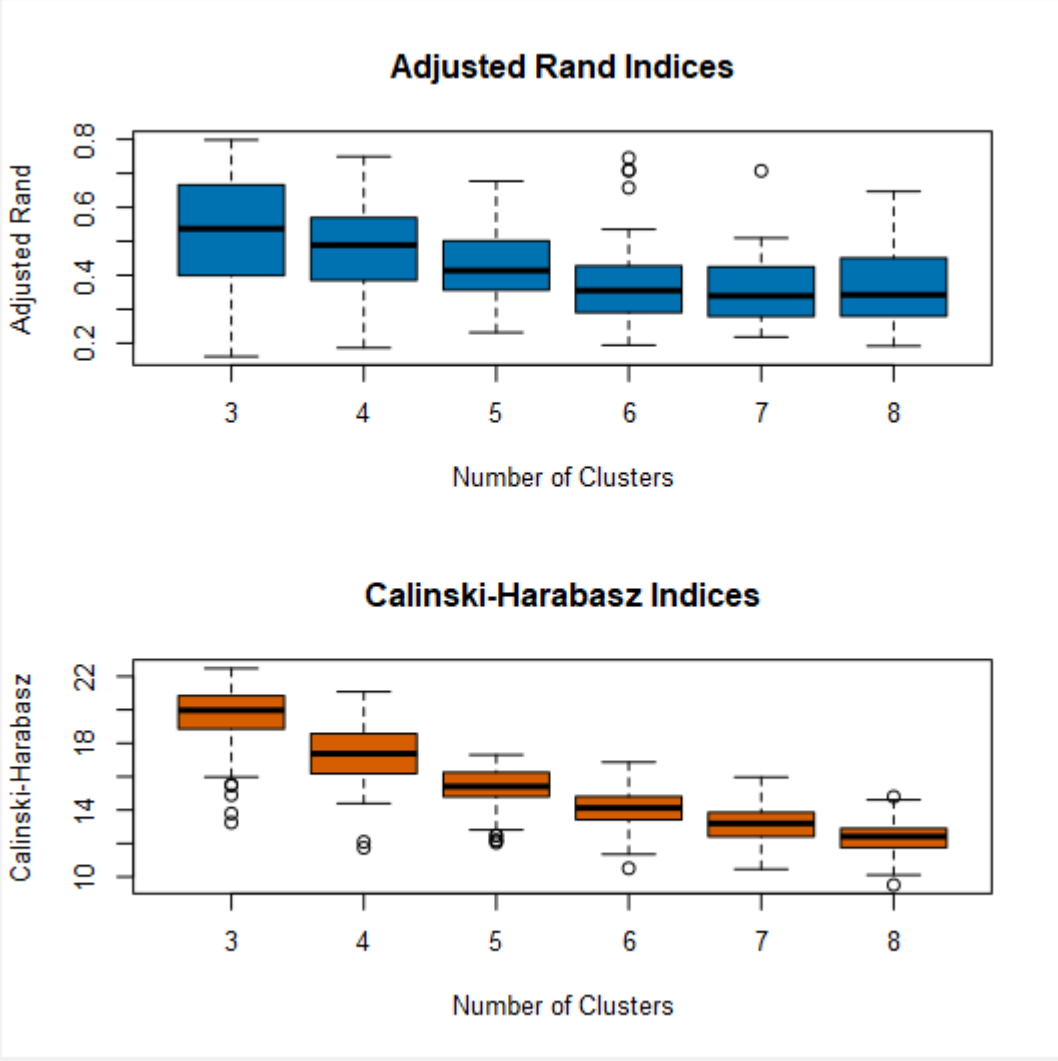


Figure 1: Diagnostic Cluster Number

From figure 1 we can see there are two indices: AR and CH. For AR indices the higher the indices, the better the stability of the cluster. For CH indices, the higher the indices the

better the distinctness of the clusters. From AR indices, we can select cluster number 3 as it is higher, highest median and good spread of the interquartile range. From CH indices, we can select cluster number 3 as it has the highest median and fairly compact spread. So, from both scale we can select cluster number of 3 which will be used in our analysis.

2. How many stores fall into each store format?

Ans:

Clusters	No. of Stores
1	**23**
2	**29**
3	**33**

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

Ans:

From figure 2 we can see, the avg. distance of each of the clusters are very close that means all the clusters are very much compact as cluster is the compact as it holds the lowest value of 2.12. Cluster 2 separated from other clusters with good distance 2.11. Some values have high positive value in a cluster and high negative value to other cluster. It indicates the clusters are opposite to each other. For Sum_Produce cluster 2 is high positive but cluster 1 and 3 are high negative. That means revenue is higher in high positive clusters and lower in high negative stores. For general merchandise, cluster 1 is opposite to cluster 2 because of their huge difference. For Sum-Dairy, cluster 2 (high positive) is opposite to cluster 1 (high negative). That means revenue is higher in high positive clusters and lower in high negative 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

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Pct_Sum_Dry_Grocery	Pct_Sum_Dairy	Pct_Sum_Frozen_Food	Pct_Sum_Meat	Pct_Sum_Produce	Pct_Sum_Floral	Pct_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

	Pct_Sum_Bakery	Pct_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Figure 2: Cluster Information

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.

Ans:

https://public.tableau.com/profile/fahad.munir#!/vizhome/StoreLocationsbyClusters/StoreLocationsbyClusters?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.)

Ans:

From figure 3 we can see, the overall accuracy measures of forest model and boosted model are the highest of the three models. However, we will select boosted model to predict the best store format for the new stores because it has a higher F1 value of 0.8889 compare to 0.8426 of forest model.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Cluster	0.8235	0.8426	0.7500	1.0000	0.7778
Decision_Tree_Cluster	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Cluster	0.8235	0.8889	1.0000	1.0000	0.6667

Figure 3: Comparative measures of different classification model

Again in confusion matrix of figure 4 we can see, boosted model has a higher rate of predicting actual value compare to other two.

Confusion matrix of Boosted_Cluster

	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_Cluster

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

Confusion matrix of Forest_Cluster

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

Figure 4: Confusion Matrix of different classification model

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?

Ans: ETS(m,n,m) model was used for the forecast. Let's explain the reasons below through data.

Actual and Forecast Values:

Actual	ARIMA	ETS
26338477.15	27997835.63764	26907095.61191
23130626.6	23946058.0173	22916903.07434
20774415.93	21751347.87069	20342618.32222
20359980.58	20352513.09377	19883092.31778
21936906.81	20971835.10573	20479210.4317
20462899.3	21609110.41054	21211420.14022

Figure 5: Comparison between actual and forecasted data

From figure 5 we can see, ETS model values are much closer to actual value than ARIMA model values.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822

Figure 6: Error measures comparison

In comparison between error measures, ARIMA has higher MASE and RMSE than ETS model that means ETS model can give us more significant and accurate values during forecasting.

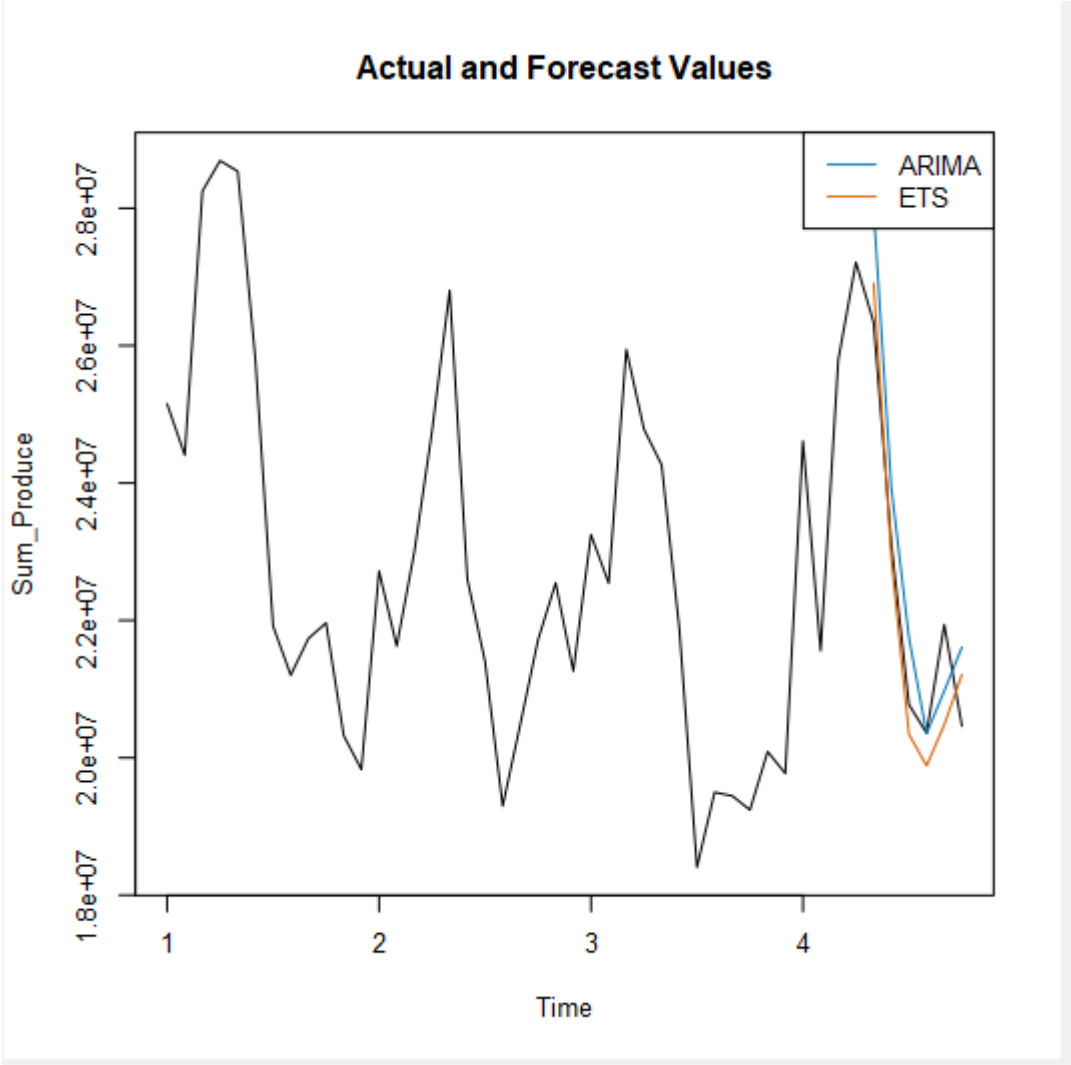


Figure 7: Comparison Graph

Finally, the graph showing the comparative measurement.

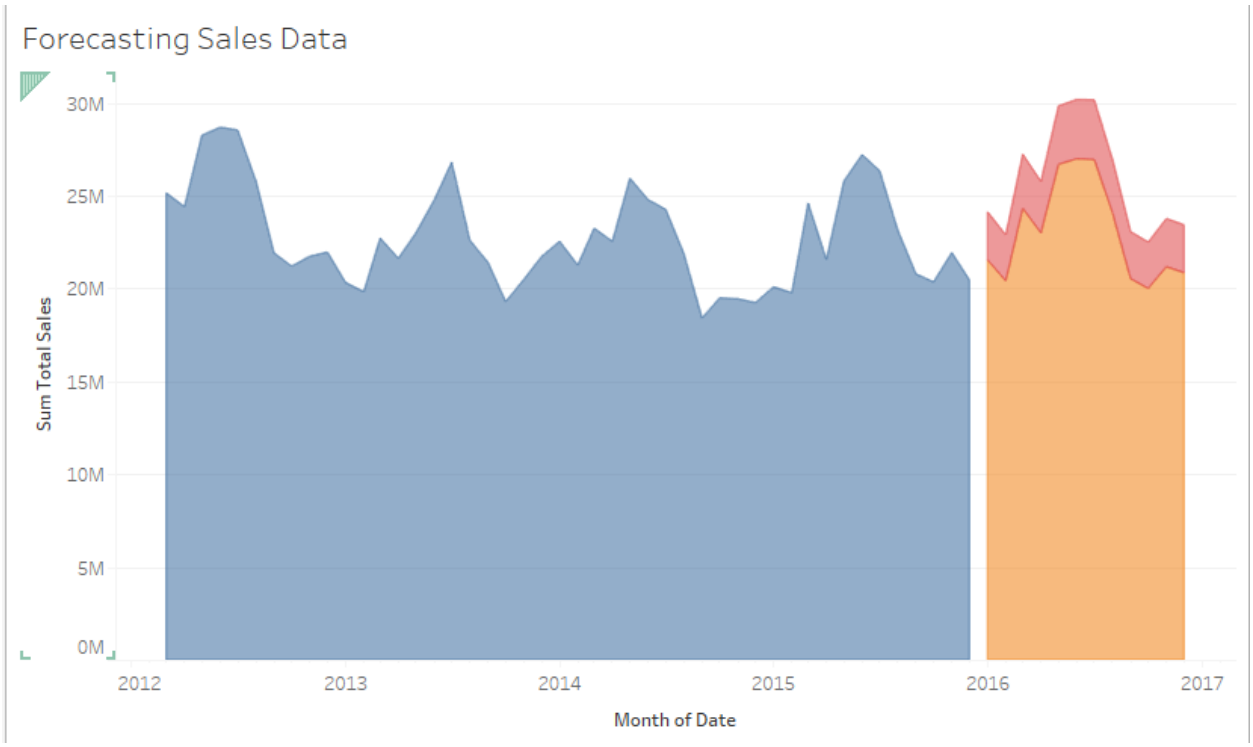
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Ans:

Month	New	Existing
Jan-16	$2,587,451	$21,539,936
Feb-16	$2,477,353	$20,413,771
Mar-16	$2,913,185	$24,325,953
Apr-16	$2,775,746	$22,993,466
May-16	$3,150,867	$26,691,951
Jun-16	$3,188,922	$26,989,964
Jul-16	$3,214,746	$26,948,631
Aug-16	$2,866,349	$24,091,579
Sep-16	$2,538,727	$20,523,492
Oct-16	$2,488,148	$20,011,749
Nov-16	$2,595,270	$21,177,435
Dec-16	$2,573,397	$20,855,799

Visualization of forecast

https://public.tableau.com/profile/fahad.munir#!/vizhome/ForecastingSalesData/ForecastingSalesData?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.