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: Based on the k-means plots shown in Figure 1, both Adjusted Rand indices and CH indices, indicate cluster number 3 is an excellent option for stability, since the median is high and the maximum, minimum and interquartile range are more compact.

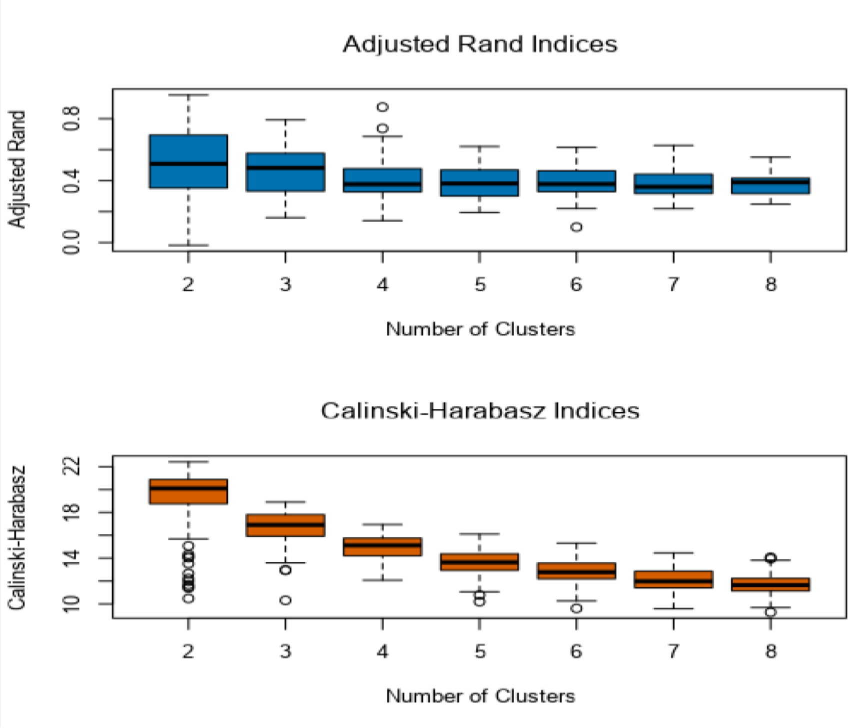
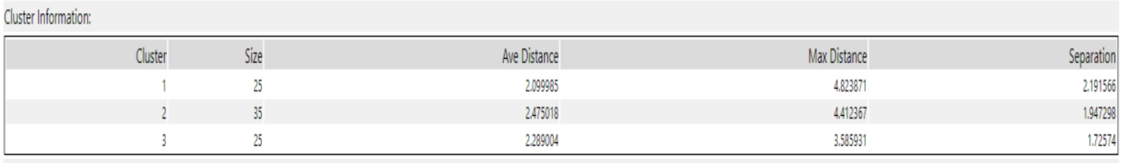


Figure 1 K-means plots

2. How many stores fall into each store format?



Cluster Information:

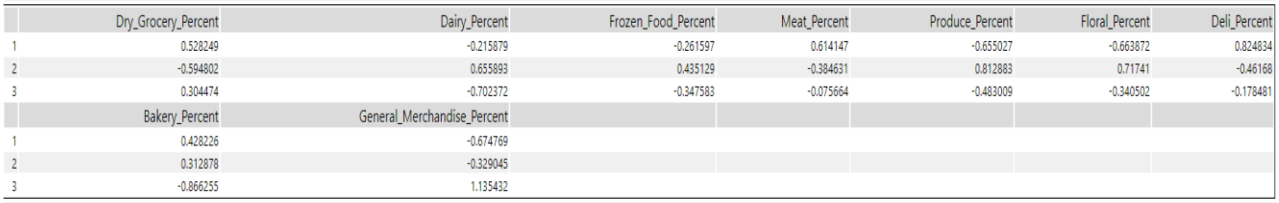
Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Figure 2 Clusters information

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Ans: As we can see on the Figure 3 below, we can conclude about the three different clusters based on the sales of each cluster product percentage.
   Cluster 1:
   - Dry_Grocery_Percent
   - Meat_percent
   - Deli_percent
   - Bakery_percent

   Cluster 2:
   - Dairy_Percent
   - Frozen_Food_Percent
   - Produce_percent
   - Floral_Percent
   - Bakery_percent

   Cluster 3:
   - Dry_Grocery_percent
   - General_Merchandise_percent



	Dry_Grocery_Percent	Dairy_Percent	Frozen_Food_Percent	Meat_Percent	Produce_Percent	Floral_Percent	Deli_Percent
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178481
	Bakery_Percent	General_Merchandise_Percent					
1	0.428226	-0.674769					
2	0.312878	-0.329045					
3	-0.866255	1.135432					

Figure 3 Clusters percentages

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/app/profile/khaled6919/viz/Book1-TASK1-Udacity/Sheet3?publish=yes

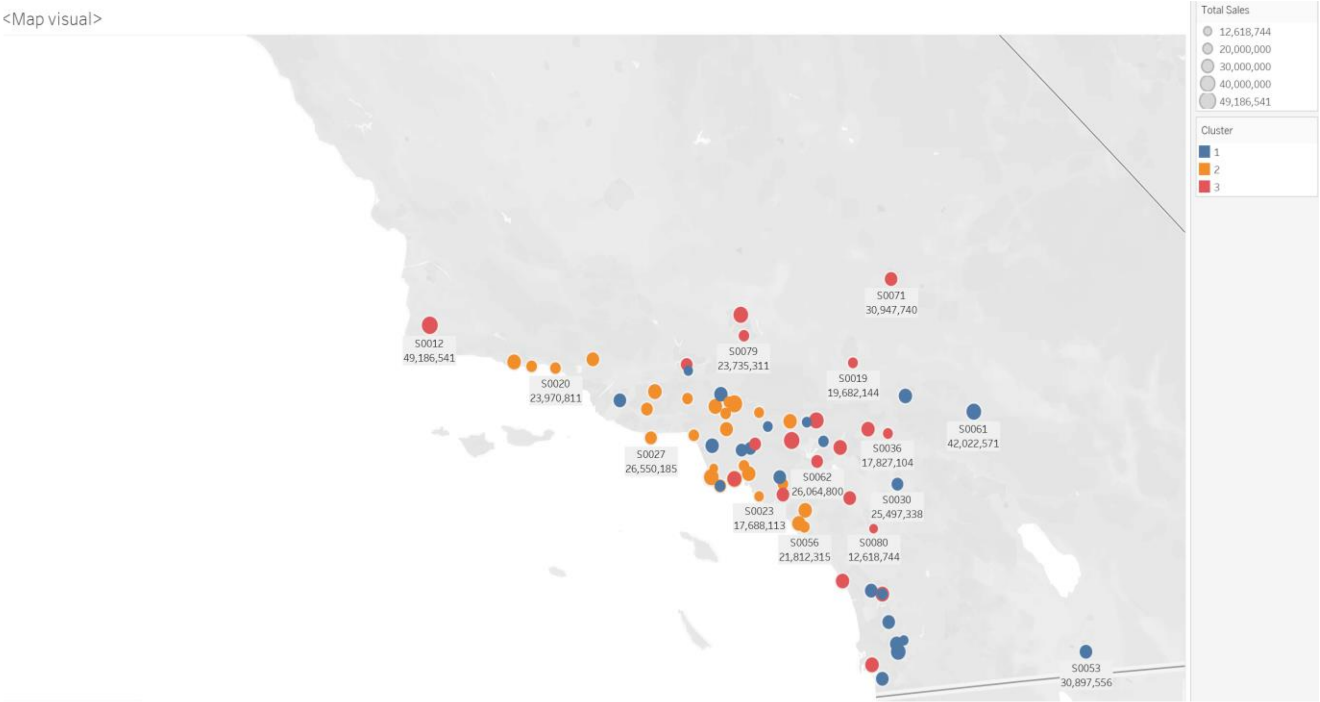
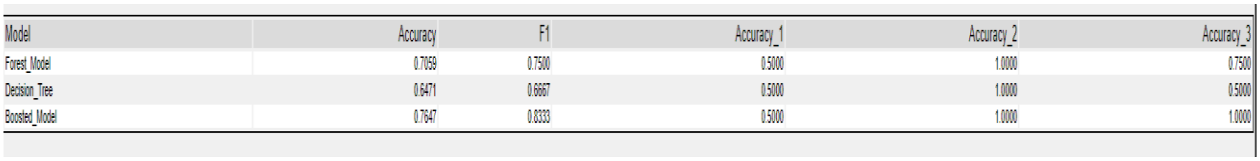


Figure 4 Clusters information using Tableau

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: Looking to model comparison report as show in Figure 4, we shall use Boosted model since its resulted in high accuracy in all 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	0.7647	0.8333	0.5000	1.0000	1.0000

Figure 5 Model comparison report

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

Table 1 New stores with corresponding segment

Store Number	Segment
S0086	1
S0087	2
S0088	3
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?

Ans:

As shown in figure 6, the seasonality changes in magnitude each year so multiplicative will be used, the trend line is not clear therefore we don't apply any method, the reminder is irregular therefore we will used multiplicative method for it.
ETS(M,N,M)

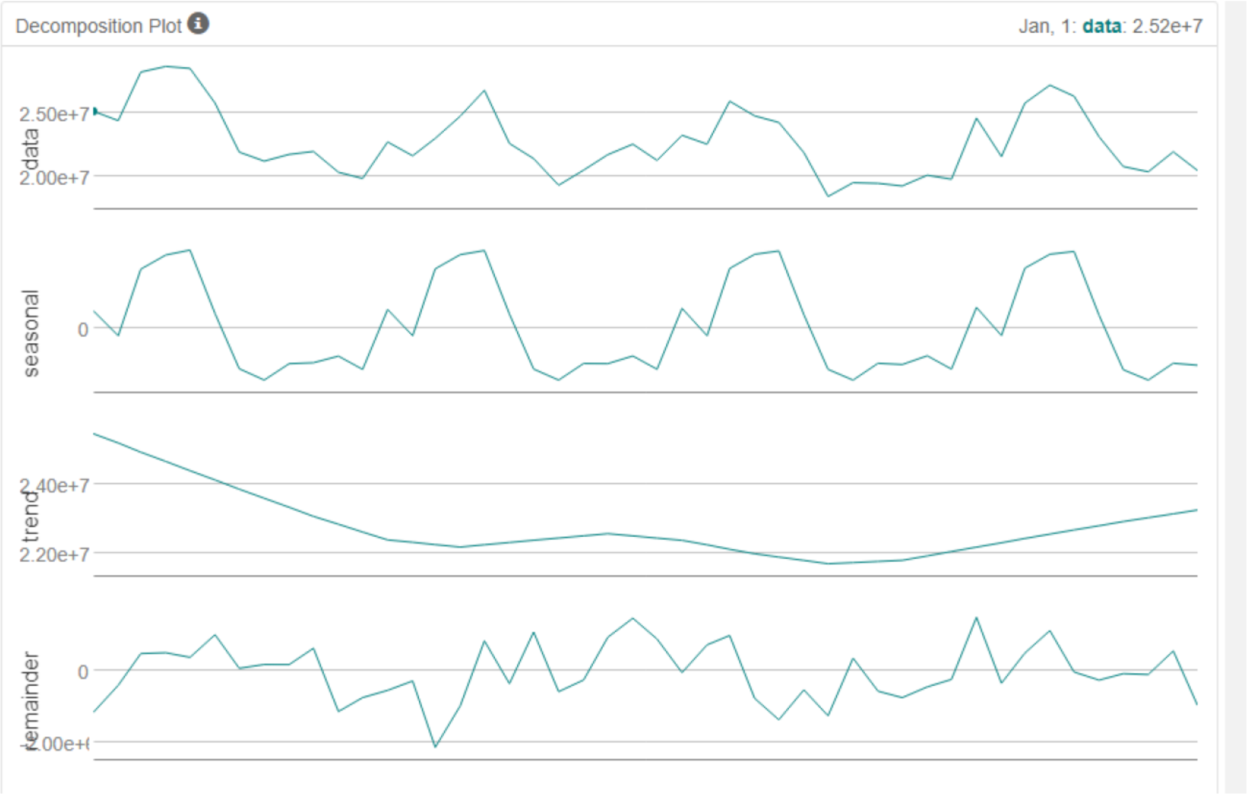


Figure 6 Seasonality, trend. error plots

ARIMA (0,1,2)(0,1,0) was chosen, seasonal difference and seasonal first difference were performed. There is a lag -2 see Figure 7.

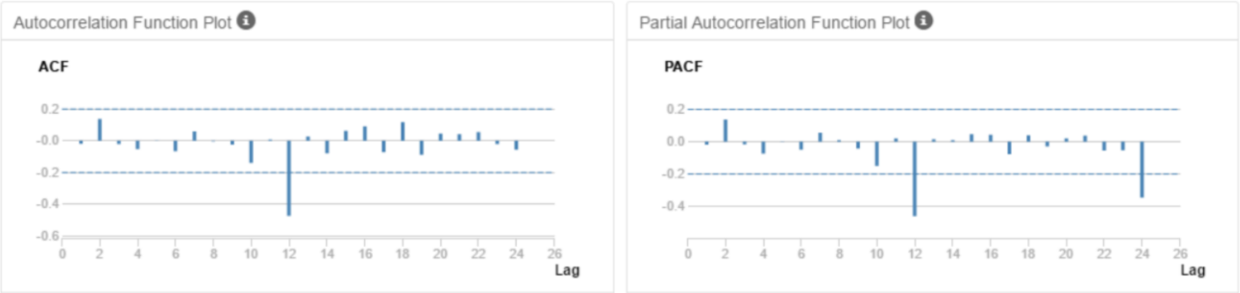
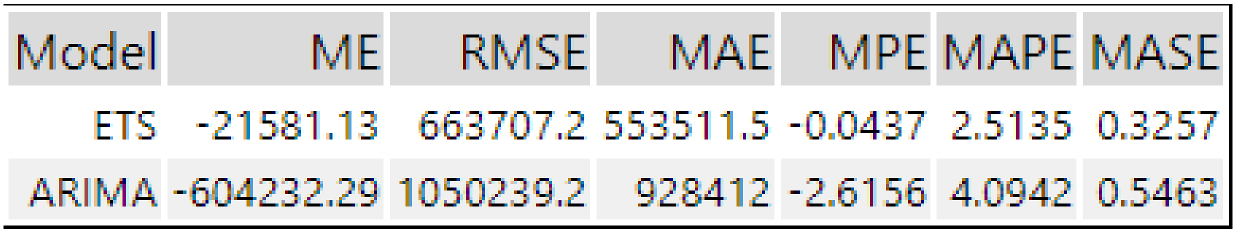


Figure 7 ACF, PACF PLOTS

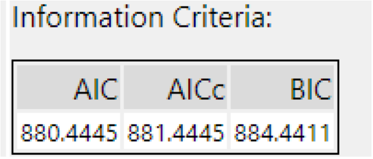
When comparing the two models, ETS performed better since it has the lowest values in all measures.



Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

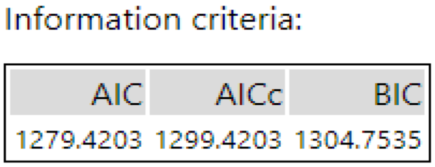
Figure 8 Accuracy measures

While ETS (1279.42) has high value in AIC compared to ARIMA (858.78).



Information Criteria:

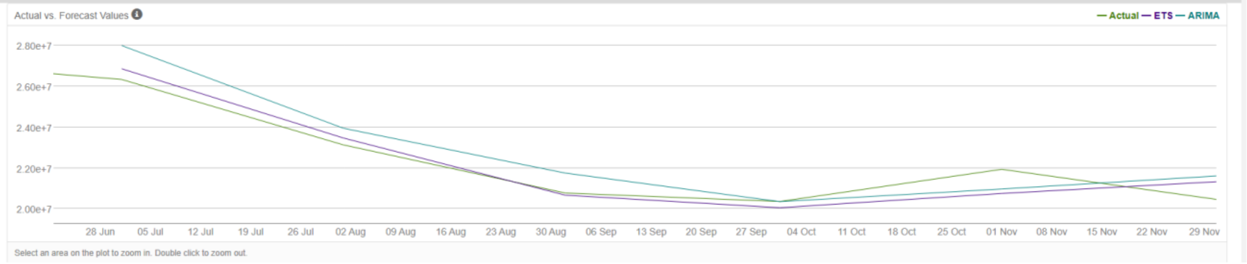
AIC	AICc	BIC
880.4445	881.4445	884.4411



Information criteria:

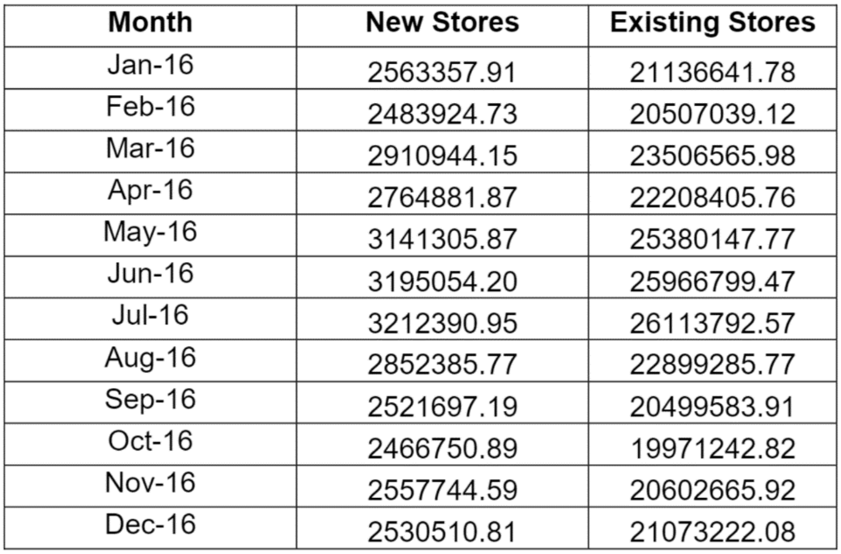
AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

And the actual plot vs forecast values, we can see that ETS performed better than ARIMA



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.

Table 2 Forecasts for existing and new stores



Month	New Stores	Existing Stores
Jan-16	2563357.91	21136641.78
Feb-16	2483924.73	20507039.12
Mar-16	2910944.15	23506565.98
Apr-16	2764881.87	22208405.76
May-16	3141305.87	25380147.77
Jun-16	3195054.20	25966799.47
Jul-16	3212390.95	26113792.57
Aug-16	2852385.77	22899285.77
Sep-16	2521697.19	20499583.91
Oct-16	2466750.89	19971242.82
Nov-16	2557744.59	20602665.92
Dec-16	2530510.81	21073222.08

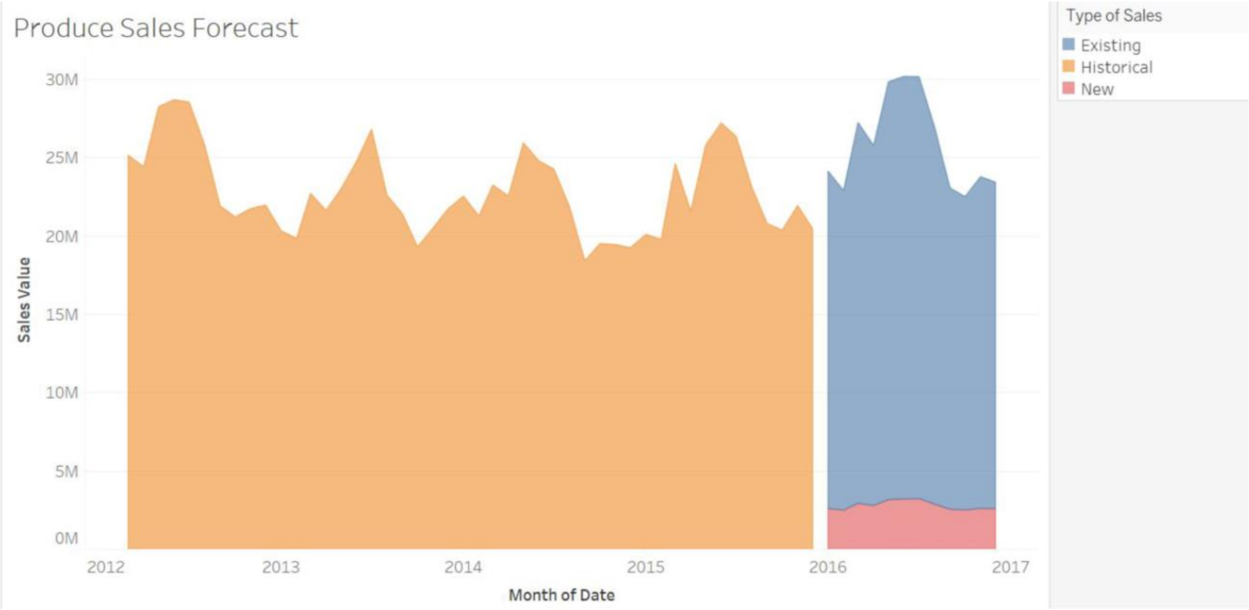


Figure 9 Produce sales forecast using Tableau

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.