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 report, Adjusted Rand and Calinski-Harabasz, the optimal number of store formats is **3** when both the indices registered the highest median value.

2. How many stores fall into each store format?

   Ans. Cluster 1 has 23 stores, cluster 2 has 29 stores while cluster 3 has 33 stores.

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

   Ans. Cluster 1 stores sold more General Merchandise in terms of percentage while Cluster 2 stores sold more Produce. Cluster 1 stores have highest medial total sales when compared to the other 2. Its range of total sales and most of other categorical sales are also the largest. Cluster 3 stores are the most similar in terms of sales due to more compact

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.
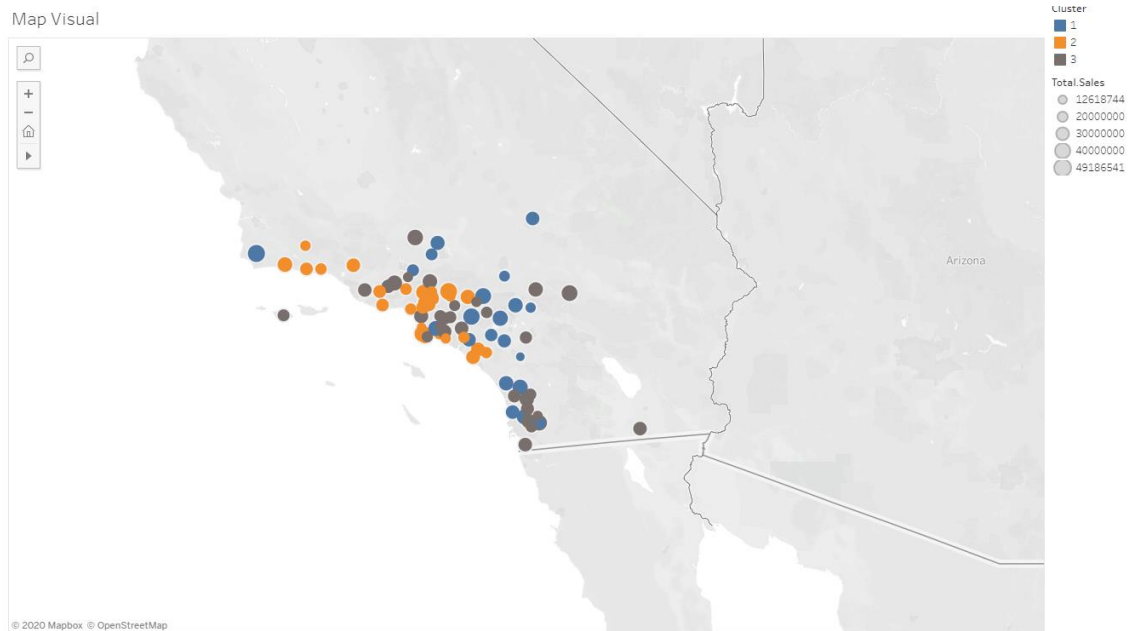
https://public.tableau.com/profile/shariq.hameed#!/vizhome/Task1_15871255474320/Task1?publish=yes

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.)

2. Ans. The model comparison report below shows comparison matrix of Decision Tree, Forest Model and Boosted Model.
   Boosted Model is chosen despite having same accuracy as Forest Model due to higher F1 value.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT	0.7059	0.7327	0.6000	0.6667	0.8333
FM	0.8235	0.8251	0.7500	0.8000	0.8750
BM	0.8235	0.8543	0.8000	0.6667	1.0000

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

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

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of BM

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

Confusion matrix of DT

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

Confusion matrix of FM

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

3. 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	3
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. The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and nothing should be applied. Its error is irregular and should be applied multiplicatively.

ARIMA(0,1,2)(0,1,0) is used as seasonal difference and seasonal first difference were performed. There is a lag-2.
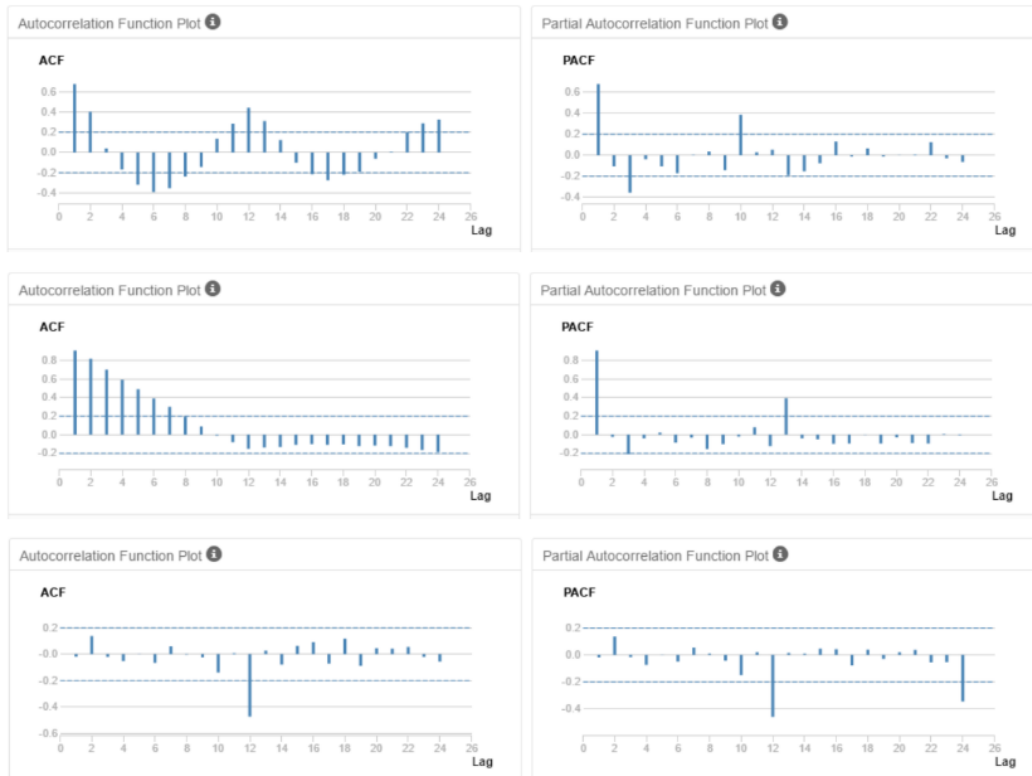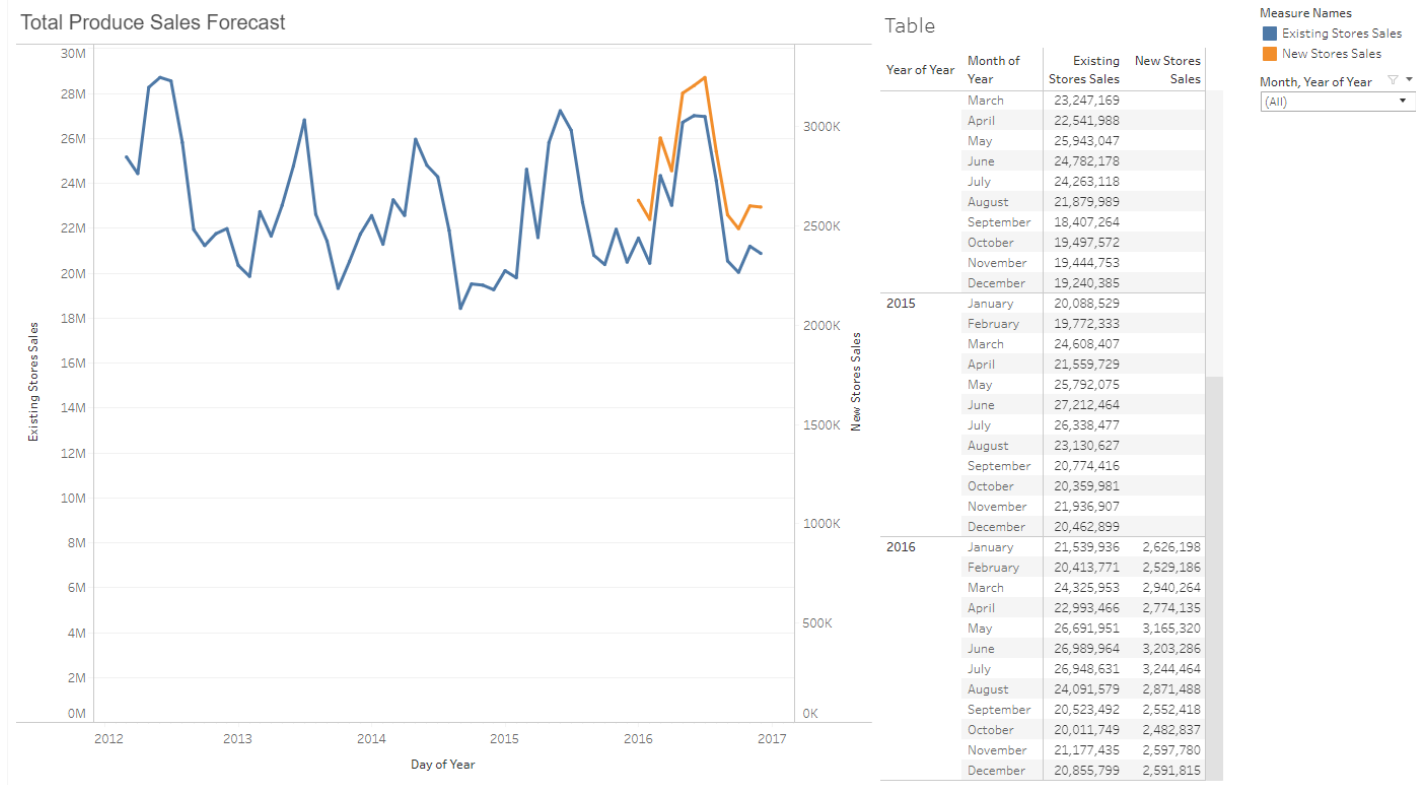

ETS model's accuracy is higher when compared to ARIMA model. A holdout sample of 6 months data is used. Its RMSE of 1,020,597 is lower than ARIMA's 1,429,296 while its MASE is 0.45 compared to ARIMA's 0.53. ETS also has a higher AIC at 1,283 while ARIMA's AIC is 859. The graph below shows actual and forecast value with 80% & 95% confidence level interval.

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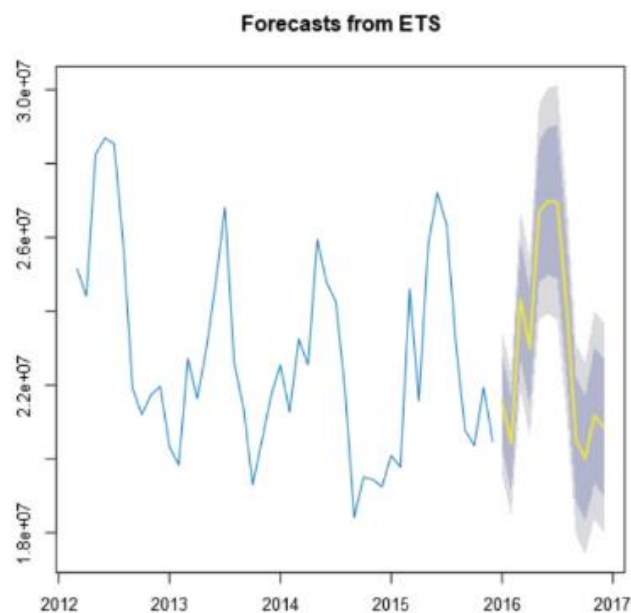
2. Please **provide** a table of your forecasts **for** existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Ans. Table below shows the forecast sales for existing stores and new stores. New store sales is obtained by using ETS(M,N,M) analysis with all the 3 individual cluster to obtain the average sales per store. The average sales value (x3 cluster 1, x6 cluster 2, x1 cluster 3) are added up produce New Store Sales.

Total Produce Sales Forecast

Table

Year of Year	Month of Year	Existing Stores Sales	New Stores Sales
	March	23,247,169	
	April	22,541,988	
	May	25,943,047	
	June	24,782,178	
	July	24,263,118	
	August	21,879,989	
	September	18,407,264	
	October	19,497,572	
	November	19,444,753	
	December	19,240,385	
2015	January	20,088,529	
	February	19,772,333	
	March	24,608,407	
	April	21,559,729	
	May	25,792,075	
	June	27,212,464	
	July	26,338,477	
	August	23,130,627	
	September	20,774,416	
	October	20,359,981	
	November	21,936,907	
	December	20,462,899	
2016	January	21,539,936	2,626,198
	February	20,413,771	2,529,186
	March	24,325,953	2,940,264
	April	22,993,466	2,774,135
	May	26,691,951	3,165,320
	June	26,989,964	3,203,286
	July	26,948,631	3,244,464
	August	24,091,579	2,871,488
	September	20,523,492	2,552,418
	October	20,011,749	2,482,837
	November	21,177,435	2,597,780
	December	20,855,799	2,591,815

Measure Names
- Existing Stores Sales
- New Stores Sales

Month, Year of Year
(All)