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 format is 3. After running K-means clustering model and looking at the median and spread of Rand CH Indices that optimal number is 3. Below are the Table and plots for the same.

K-Means Cluster Assessment Report

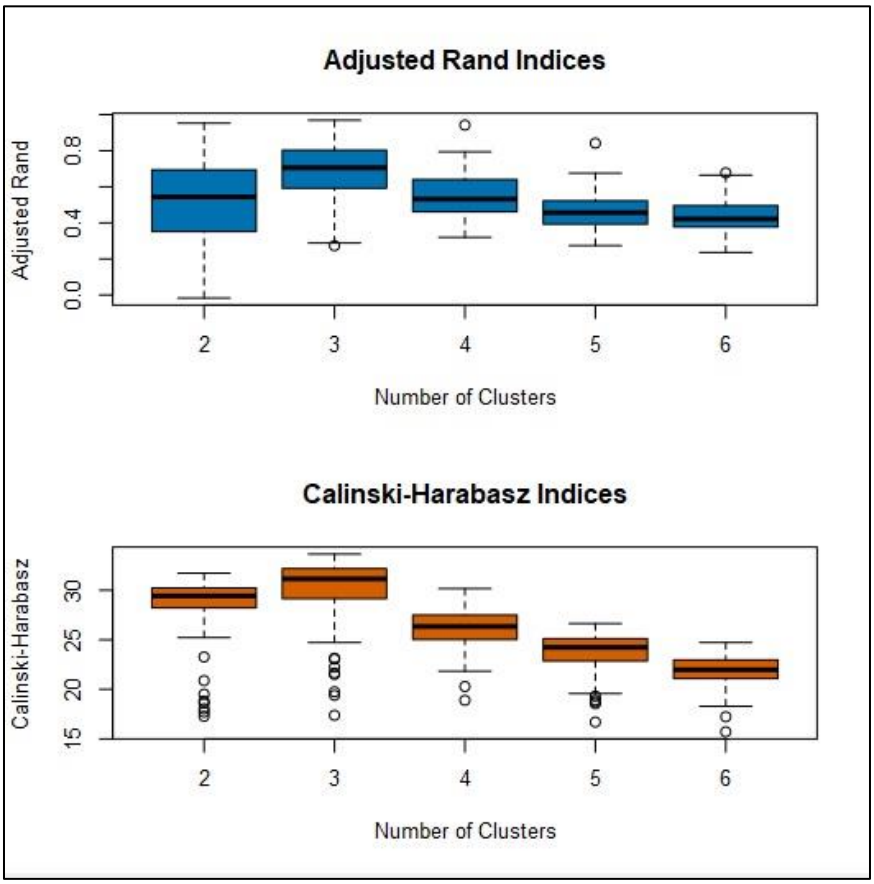
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.27351	0.31976	0.274316	0.235718
1st Quartile	0.35943	0.594017	0.46406	0.39294	0.377774
Median	0.544023	0.705326	0.53195	0.456588	0.421798
Mean	0.524263	0.69161	0.548167	0.470346	0.435429
3rd Quartile	0.694147	0.800179	0.635682	0.520656	0.493589
Maximum	0.952939	0.969034	0.942222	0.841981	0.677532

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.28099	17.38103	18.89399	16.69676	15.71092
1st Quartile	28.22121	29.21236	25.0347	22.86498	21.10249
Median	29.4157	31.14179	26.33467	24.22188	21.96958
Mean	28.56937	30.07118	26.18037	23.72205	21.92474
3rd Quartile	30.21867	32.17467	27.4999	25.09459	22.95561
Maximum	31.71569	33.63782	30.1583	26.63063	24.72038



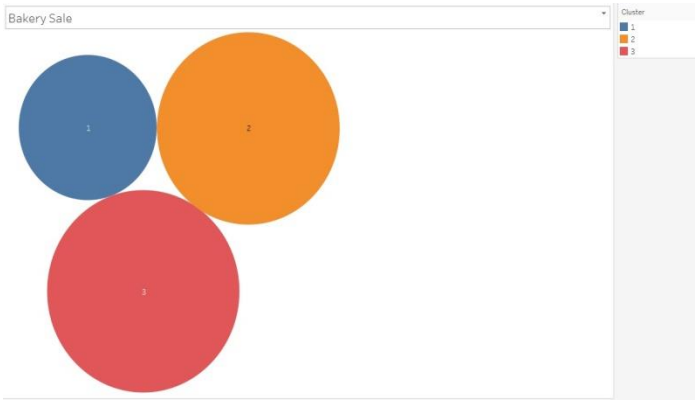
2. How many stores fall into each store format?

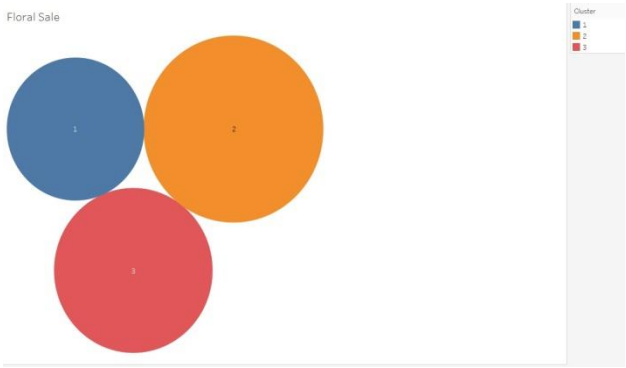
Cluster Information:

Cluster	Size
1	23
2	29
3	33

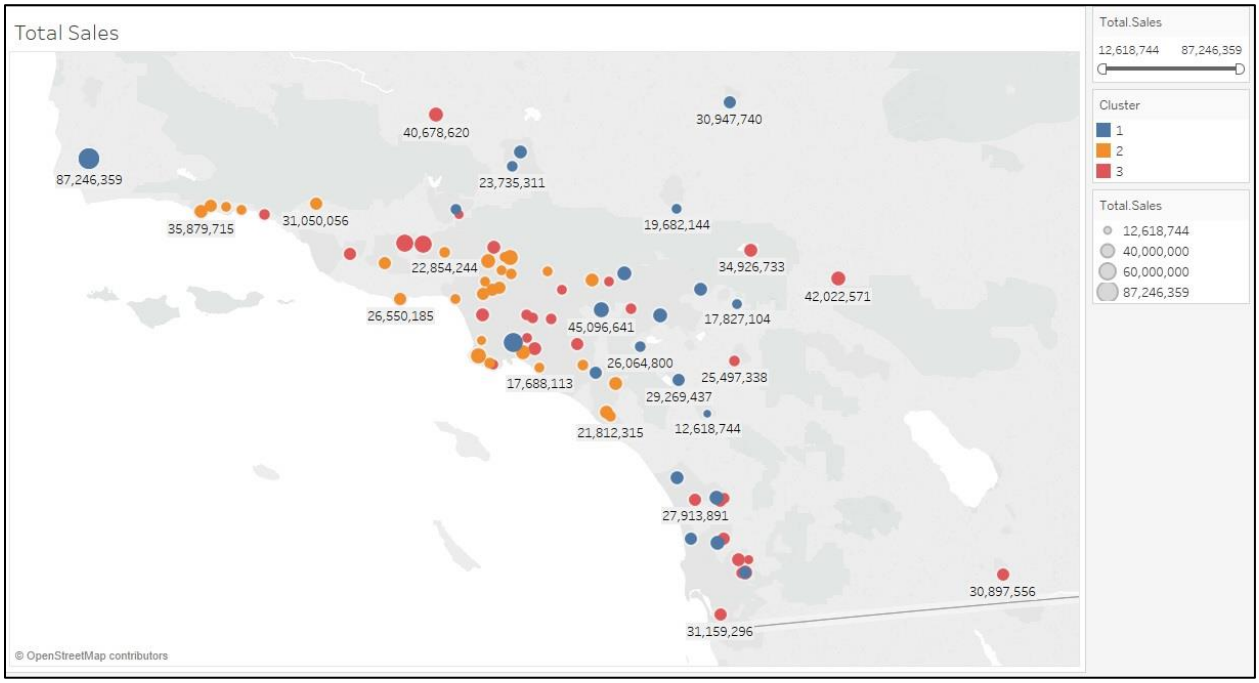
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Different clusters have different category sales for example, in the Bakery category, we can see that cluster 3 is more dominant whereas for Floral category sale we can see that cluster 2 is more dominant.





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/kanchanprasad#!/vizhome/Task1Question4/Sheet1?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.)

Forest Model, Boosted Model, and Decision tree model were developed and compared to find the best model to predict the best format for the new stores. The Boosted model was chosen over the other two since the F1 was higher in case of Boosted model even though Boosted Model and Forest model show the same accuracy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_18	0.7059	0.7327	0.6000	0.6667	0.8333
FM_model	0.8235	0.8251	0.7500	0.8000	0.8750
BM_model	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)

2. What format does 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?

ETS (MNM) with no dampening effect was considered for the forecast.

Both ARIMA model and ETS model were developed but ETS was chosen over the ARIMA (0,1,1)(0,1,0)12 model since the model accuracy of ETS was better with lower RMSE value of 1020596.90 whereas the RMSE value of ARIMA model is 1430610.61. We can also see that the MASE value of ETS model is 0.45 which is lower than the MASE value of ARIMA model which is 0.52. We can also see that the model forecast for ETS is better and closer to the actual value.

Summary of Time Series Exponential Smoothing Model ETS_nodampening

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-12901.2479844	1020596.9042405	807324.9676799	-0.2121517	3.5437307	0.4506721	0.1507788

Information criteria:

AIC	AICc	BIC
1283.1197	1303.1197	1308.4529

ARIMA (0,1,1)(0,1,0)12

Information Criteria:

AIC	AICc	BIC
856.8308	857.3308	859.4225

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
166650.6641392	1430610.6110795	934256.056469	0.6051296	4.1163464	0.5215287	-0.0019723

Ljung-Box test of the model residuals:
Chi-squared = 23.9532, df = 12, p-value = 0.025655

Record Report

1

Comparison of Time Series Models

2

Actual and Forecast Values:

Actual	ETS_nodampening
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

3

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_nodampening	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

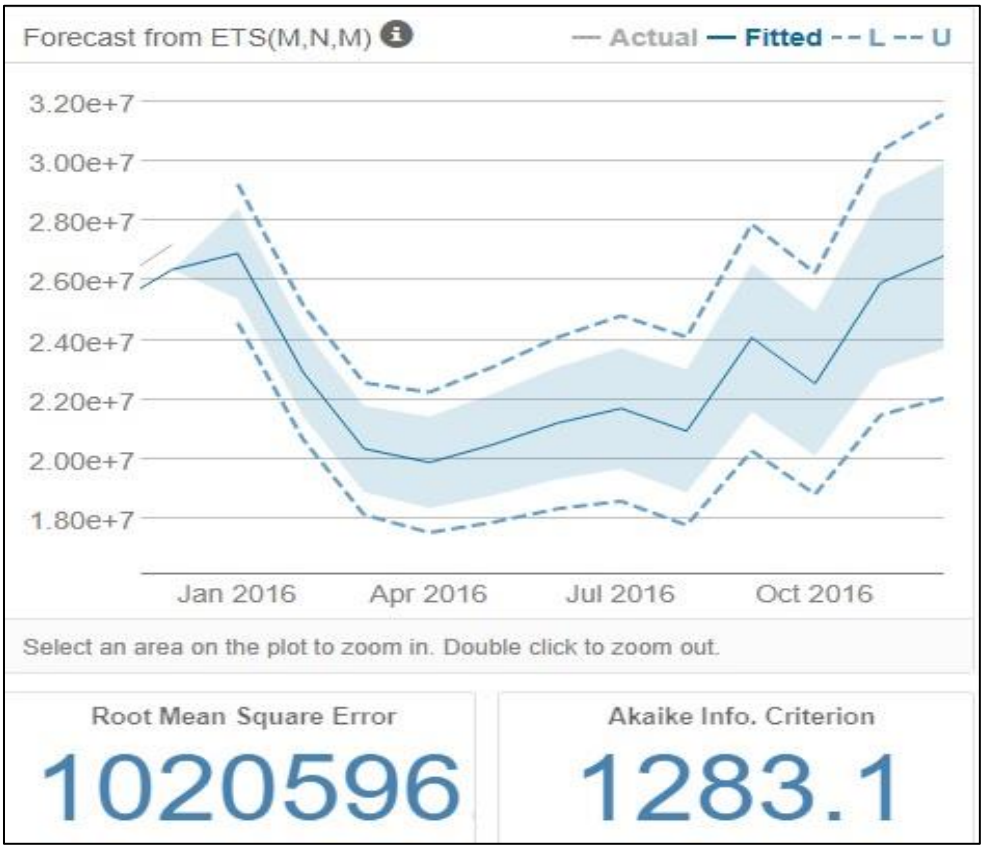
Comparison of Time Series Models

Actual and Forecast Values:

Actual	ARIMA
26338477.15	25454225.03787
23130626.6	23071096.30787
20774415.93	19598371.02787
20359980.58	20688679.39787
21936906.81	20635860.61787
20462899.3	20431492.19787

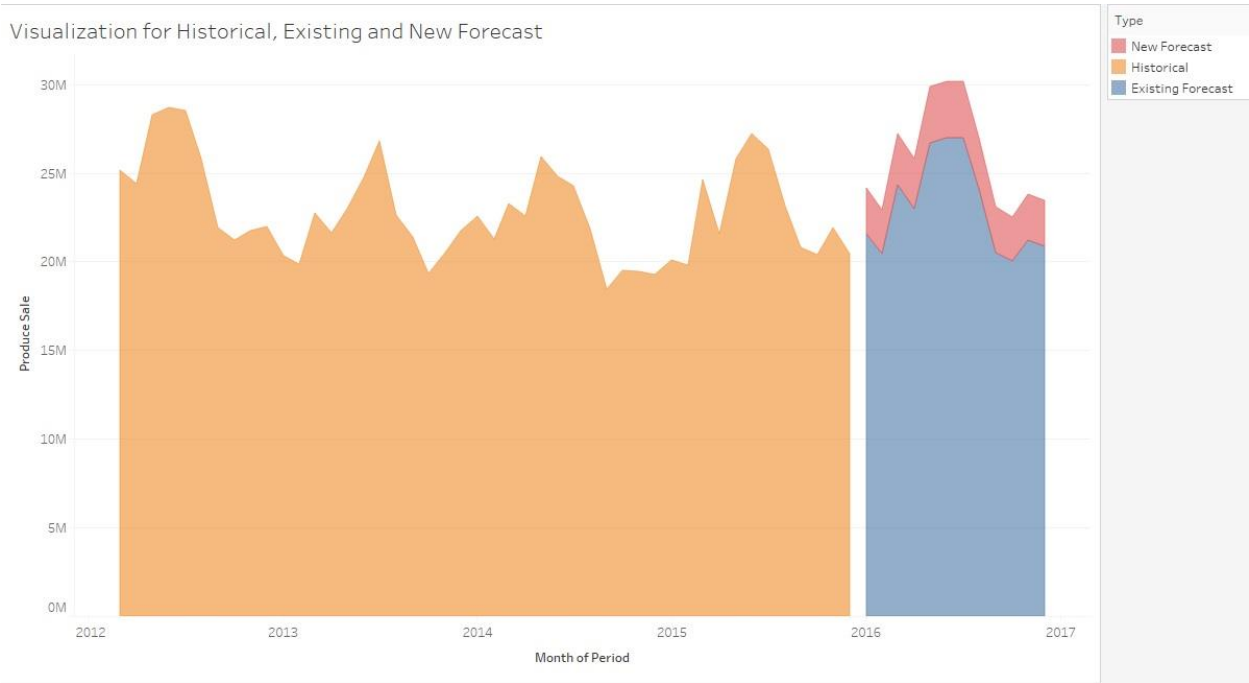
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	520597	813457.4	630163.2	2.2909	2.8291	0.3708	NA



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.

Period	Sub_Period	Existing forecast	New forecast
2016	1	21539936	2587450.851
2016	2	20413771	2477352.892
2016	3	24325953	2913185.236
2016	4	22993466	2775745.61
2016	5	26691951	3150866.835
2016	6	26989964	3188922.003
2016	7	26948631	3214745.646
2016	8	24091579	2866348.663
2016	9	20523492	2538726.849
2016	10	20011749	2488148.287
2016	11	21177435	2595270.386
2016	12	20855799	2573396.629



https://public.tableau.com/profile/kanchanprasad#!/vizhome/Task3_234/Sheet1?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.

References

Udacity forum for understanding some issues.