Project: Predictive Analytics Capstone

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 3 and I arrived at this number based on the K-Means report, the AR Indices and CH Indices.
Here are the visuals proving that :

K-Means Cluster Assessment Report

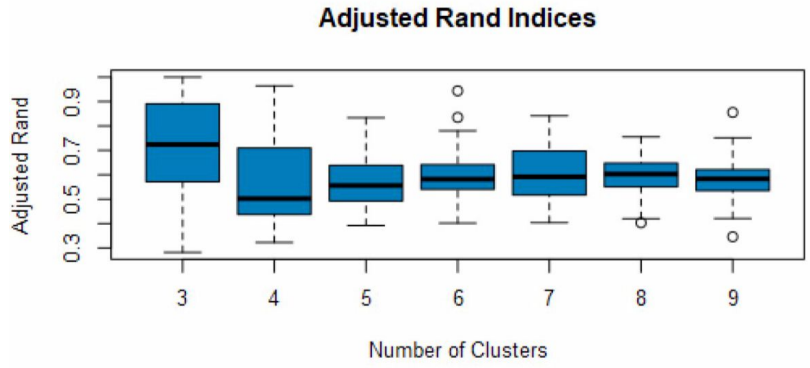
Summary Statistics

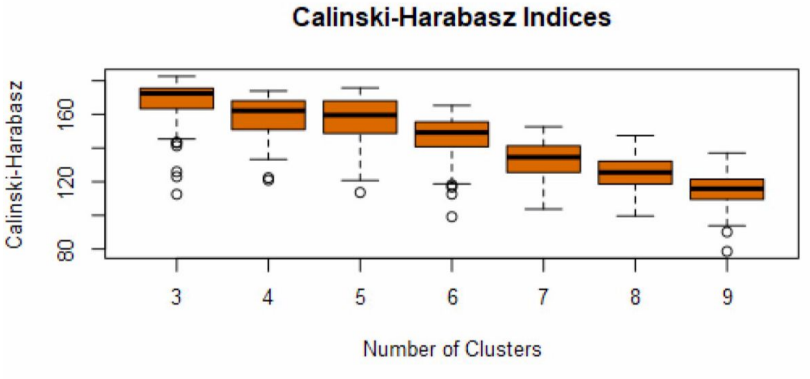
Adjusted Rand Indices:

	3	4	5	6	7	8	9
Minimum	0.282003	0.322485	0.391861	0.401568	0.403708	0.403088	0.345853
1st Quartile	0.580263	0.437322	0.494179	0.540373	0.517123	0.551005	0.535659
Median	0.723891	0.502234	0.555872	0.581788	0.591781	0.602762	0.583213
Mean	0.708809	0.5727	0.571462	0.600739	0.606374	0.588675	0.584364
3rd Quartile	0.88578	0.706955	0.63807	0.638672	0.697318	0.646922	0.621388
Maximum	1	0.964496	0.833759	0.94398	0.84182	0.755366	0.855467

Calinski-Harabasz Indices:

	3	4	5	6	7	8	9
Minimum	112.5134	120.9236	113.584	99.16435	103.7078	99.61324	78.60752
1st Quartile	163.3498	151.3745	148.7717	140.74356	125.5928	118.64541	109.58104
Median	172.4571	162.1756	159.5099	149.19875	134.5513	125.51749	115.82451
Mean	167.8003	159.3502	156.8519	146.54379	132.8127	125.00575	115.39174
3rd Quartile	175.4525	167.9654	168.0172	155.41139	141.2284	131.93227	121.43102
Maximum	182.5796	173.8978	175.6793	165.29257	152.6064	147.42403	137.05165



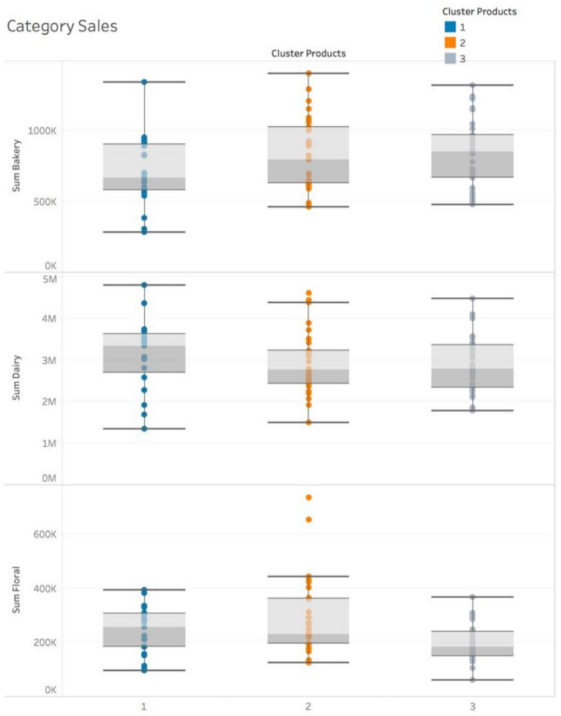


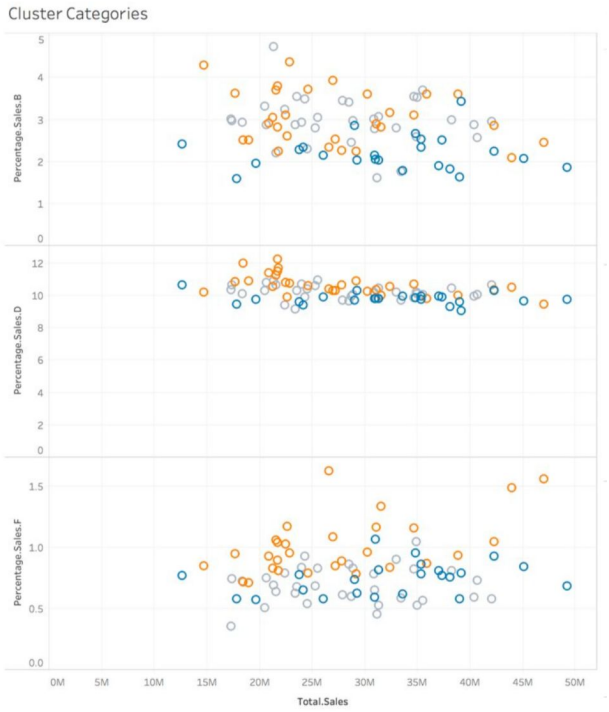
2. How many stores fall into each store format?
   Cluster 1 has 23 stores, cluster 2 has 29 stores while cluster 3 has 33 stores.

Cluster Information:

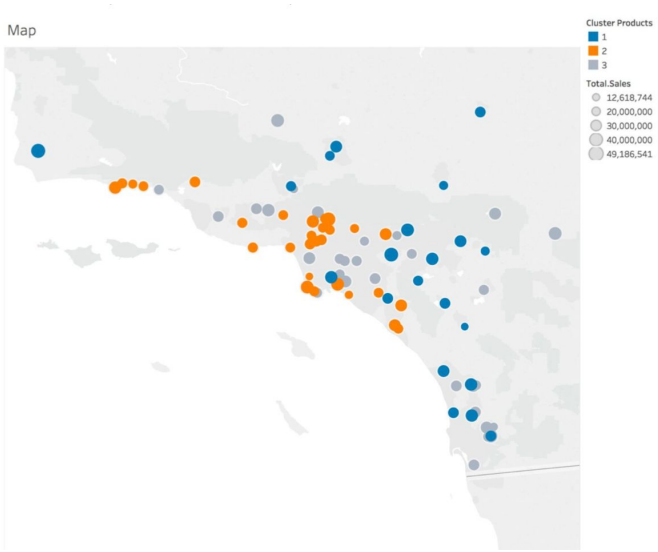
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Cluster 2 had slightly more percentage sales in Bakery compared to cluster 1 and 3.
   In terms of sales, cluster 2 has large sales in Bakery and Floral compared to cluster 1 and 3.





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.



Link to 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.)

   My choice went to Boosted model because it had a greater F1 percentage compared to Decision Tree and Forest Model.

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

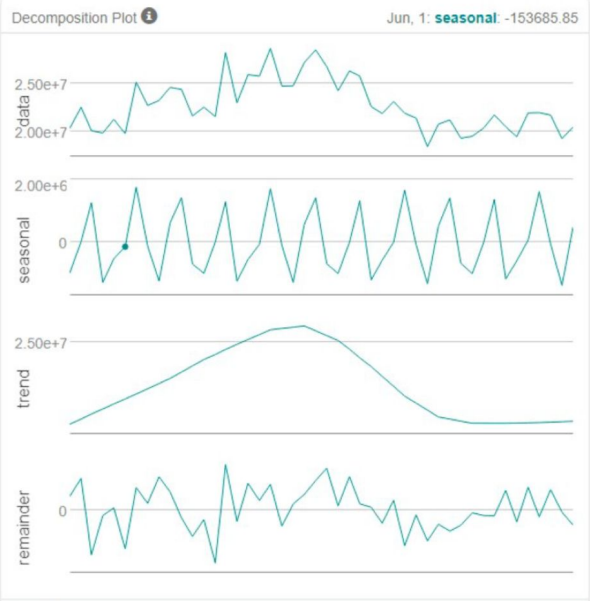
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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?

ETS(M,N,M)is used because Its error is irregular and should be applied multiplicatively,the trend is not clear therefore nothing should be applied and the seasonality is increasing then it 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.

The comparaison :

The ETS model is more accurate compared to the ARIMA model, it also has a high AIC value compared to ARIMA's AIC value. The RMSE value of ETS is much lower compared to the value of ARIMA's. The same applies to MASE values.
Therefore I believe I should choose the ETS model.

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-120308.6319739	1591855.1342896	1177589.1377921	-0.8708746	5.1105069	0.286632	-0.1136451

Information criteria:

AIC	AICc	BIC
1317.3523	1337.3523	1342.6855

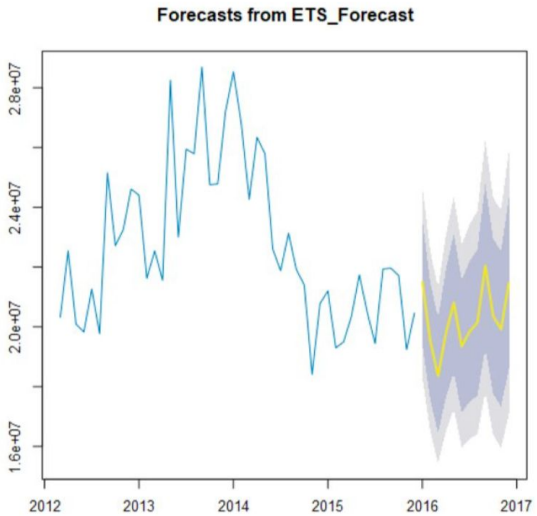
Information Criteria:

AIC	AICc	BIC
876.0533	877.0968	879.9408

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-104067.978831	1916864.2704849	1151718.4279995	-0.5104726	4.7640996	0.2803349	-0.044534

For the Forecast, the percentage value of the larger confidence interval is 95%, the percentage value of the smaller confidence interval is 80%, and the number of periods is 12.



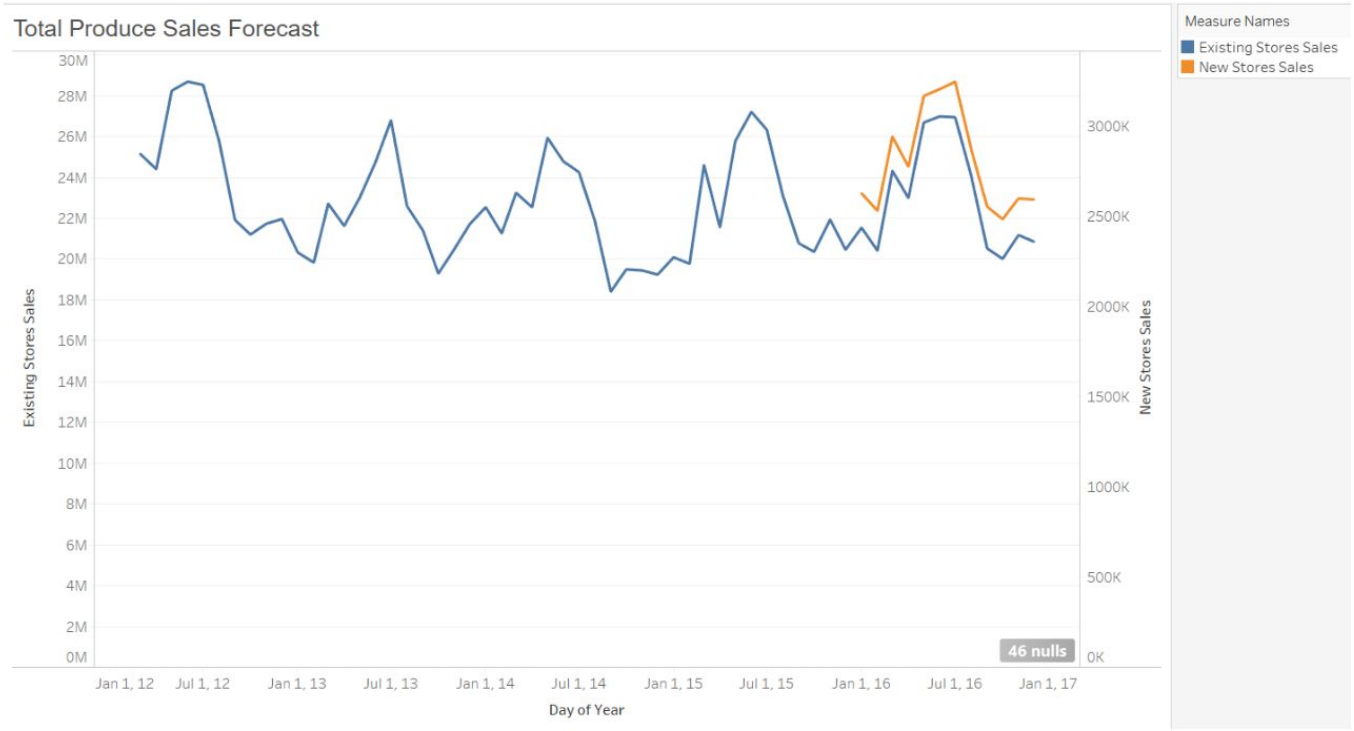
Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21504677.708111	24692012.137579	23588763.671026	19420591.745196	18317343.278642
2016	2	19544107.449771	22555645.336566	21513246.231893	17574968.66765	16532569.562976
2016	3	18363161.442448	21296667.111059	20281277.699582	16445045.185315	15429655.773837
2016	4	19788317.456897	23057710.55224	21926058.697433	17650576.216361	16518924.361555
2016	5	20801880.931059	24348898.794169	23121151.249972	18482610.612145	17254863.067948
2016	6	19349167.145946	22747905.396084	21571482.6373	17126851.654591	15950428.895808
2016	7	19841355.83156	23425652.266027	22185001.298109	17497710.365011	16257059.397093
2016	8	20145195.102866	23882336.98811	22588780.837846	17701609.367885	16408053.217621
2016	9	22044999.484622	26239078.716145	24787360.491035	19302638.47821	17850920.253099
2016	10	20359244.652728	24326816.07179	22953500.161307	17764989.144149	16391673.233666
2016	11	19922751.007689	23895350.130265	22520293.954894	17325208.060484	15950151.885113
2016	12	21465767.637647	25841002.884322	24326580.195546	18604955.079748	17090532.390973

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.

The table below shows the forecast sales for existing stores and new stores.

Year	Month	New Store Sales	Existing Store Sales
2016	1	2,626,198	21,539,936
2016	2	2,529,186	20,413,771
2016	3	2,940,264	24,325,953
2016	4	2,774,135	22,993,466
2016	5	3,165,320	26,691,951
2016	6	3,203,286	26,989,964
2016	7	3,244,464	26,948,631
2016	8	2,871,488	24,091,579
2016	9	2,552,418	20,523,492
2016	10	2,482,837	20,011,749
2016	11	2,597,780	21,177,435
2016	12	2,591,815	20,855,799

Visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.



Link to Tableau.