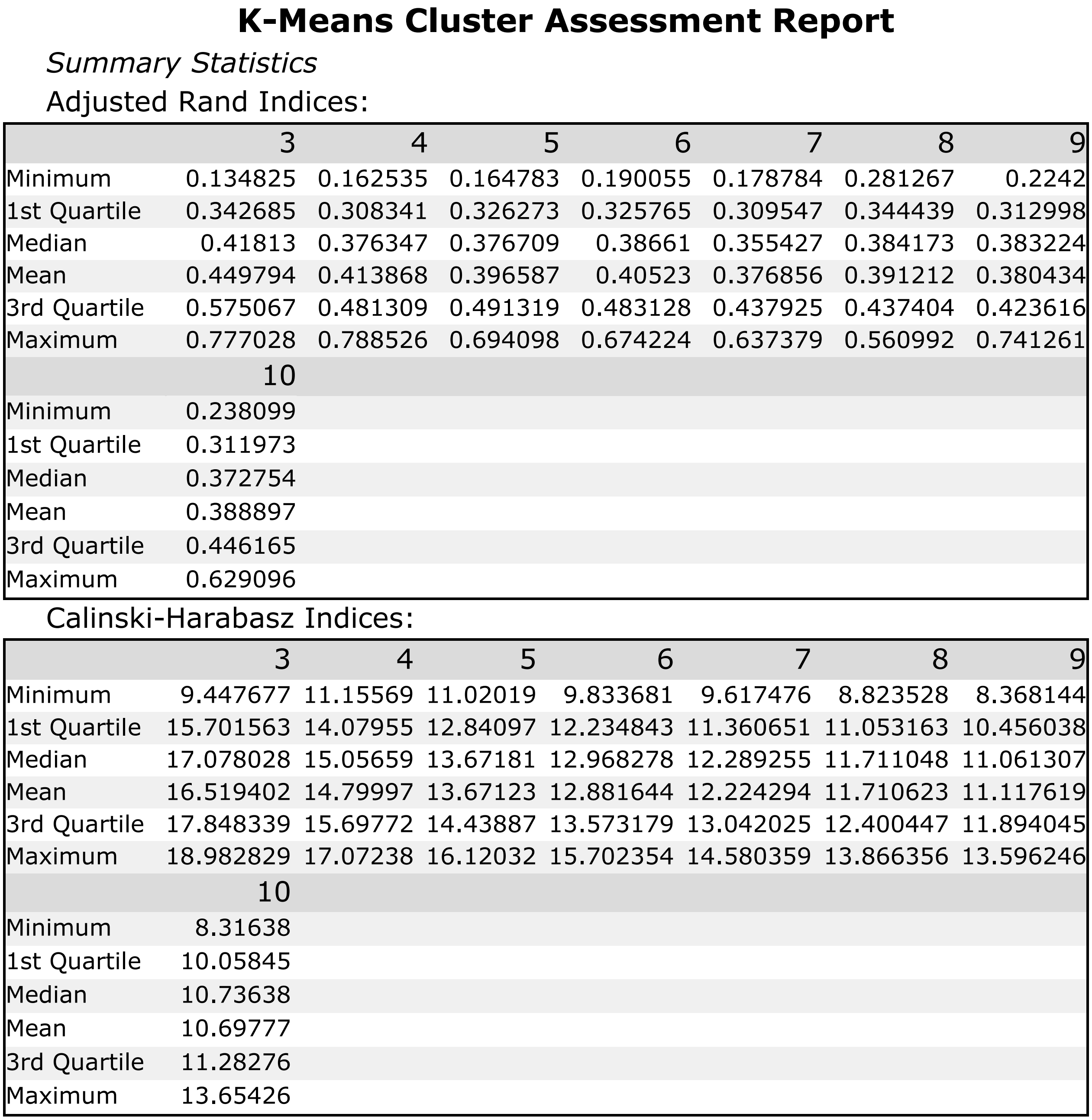
Project: Predictive Analytics Capstone

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?**



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	3	4	5	6	7	8	9	10
Minimum	0.134825	0.162535	0.164783	0.190055	0.178784	0.281267	0.2242	0.238099
1st Quartile	0.342685	0.308341	0.326273	0.325765	0.309547	0.344439	0.312998	0.311973
Median	0.41813	0.376347	0.376709	0.38661	0.355427	0.384173	0.383224	0.372754
Mean	0.449794	0.413868	0.396587	0.40523	0.376856	0.391212	0.380434	0.388897
3rd Quartile	0.575067	0.481309	0.491319	0.483128	0.437925	0.437404	0.423616	0.446165
Maximum	0.777028	0.788526	0.694098	0.674224	0.637379	0.560992	0.741261	0.629096

Calinski-Harabasz Indices:

	3	4	5	6	7	8	9	10
Minimum	9.447677	11.15569	11.02019	9.833681	9.617476	8.823528	8.368144	8.31638
1st Quartile	15.701563	14.07955	12.84097	12.234843	11.360651	11.053163	10.456038	10.05845
Median	17.078028	15.05659	13.67181	12.968278	12.289255	11.711048	11.061307	10.73638
Mean	16.519402	14.79997	13.67123	12.881644	12.224294	11.710623	11.117619	10.69777
3rd Quartile	17.848339	15.69772	14.43887	13.573179	13.042025	12.400447	11.894045	11.28276
Maximum	18.982829	17.07238	16.12032	15.702354	14.580359	13.866356	13.596246	13.65426

Figure 1: K-Means Cluster Assessment Report

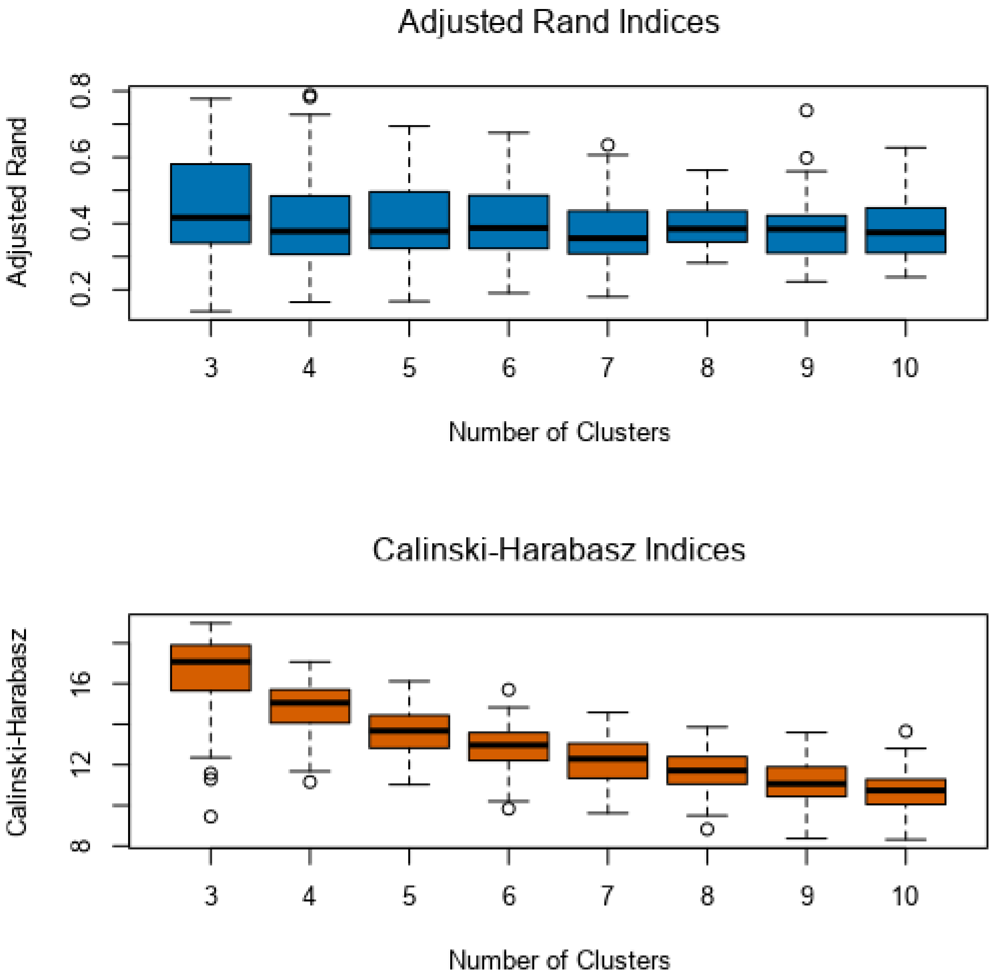
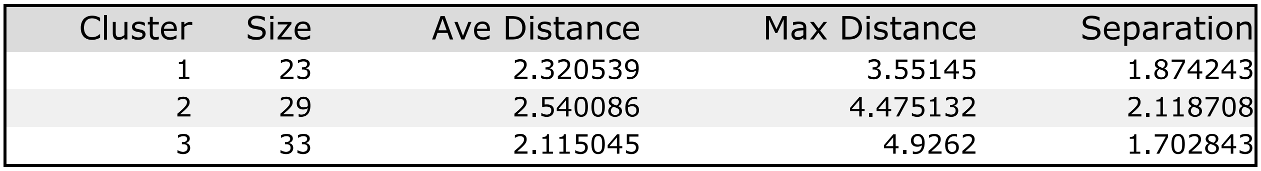


Figure 2: Adjusted Rand Indices and Calinski-Harabasz Indices

Based on the K-means report, Adjusted Rand and Calinski-Harabasz indices below, 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?

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



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

Figure 3: Cluster Information

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

Cluster 2 stores sold more produce in terms of percentage while Cluster 3 stores sold more Dairy

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. But in produce sales cluster 2 was higher

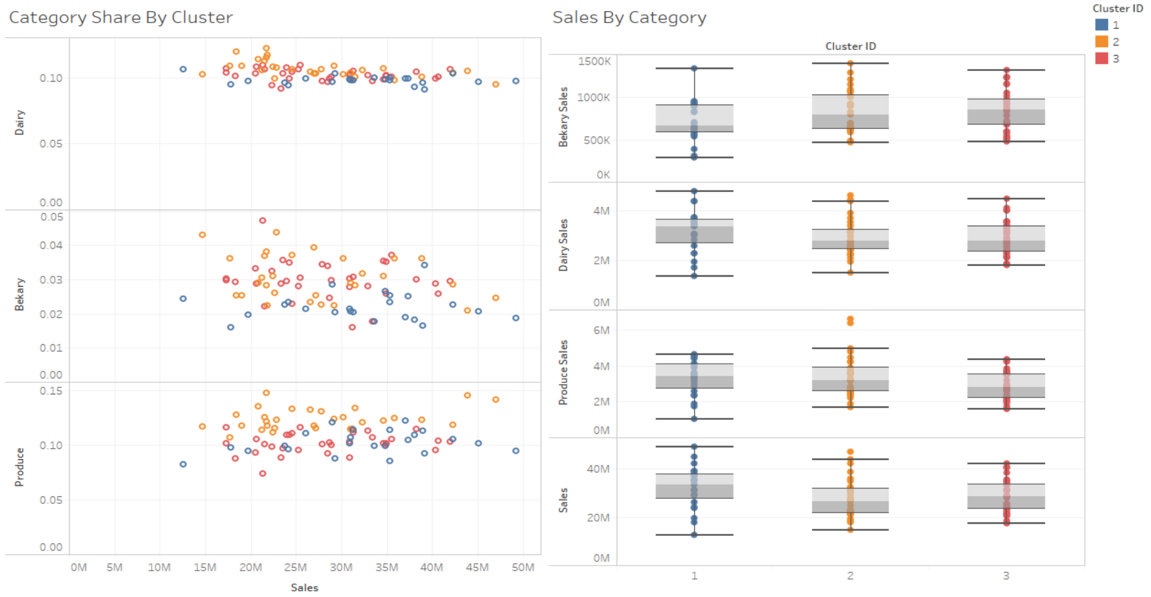


Figure 4: Tableau Visualization Tableau Dashboard

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.**

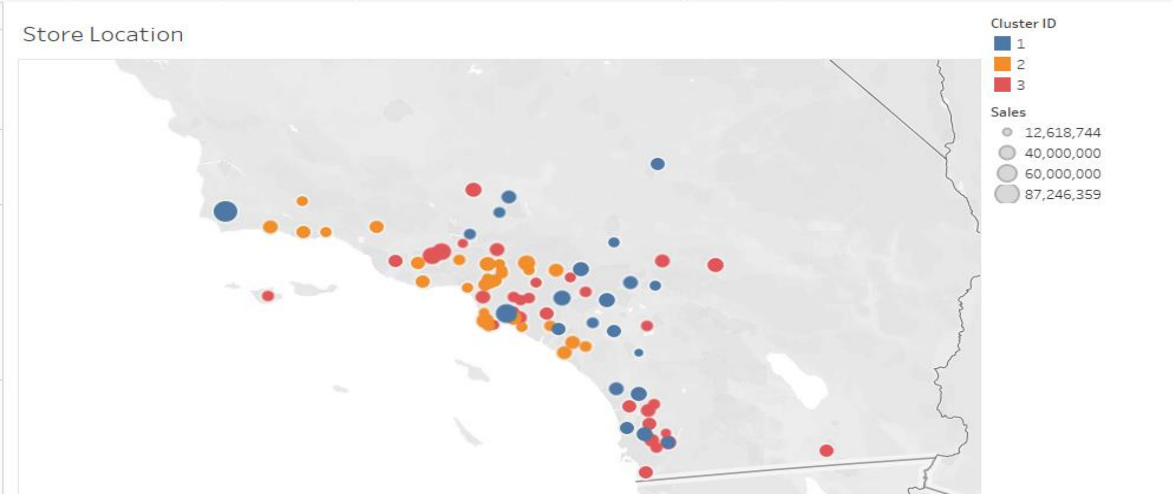


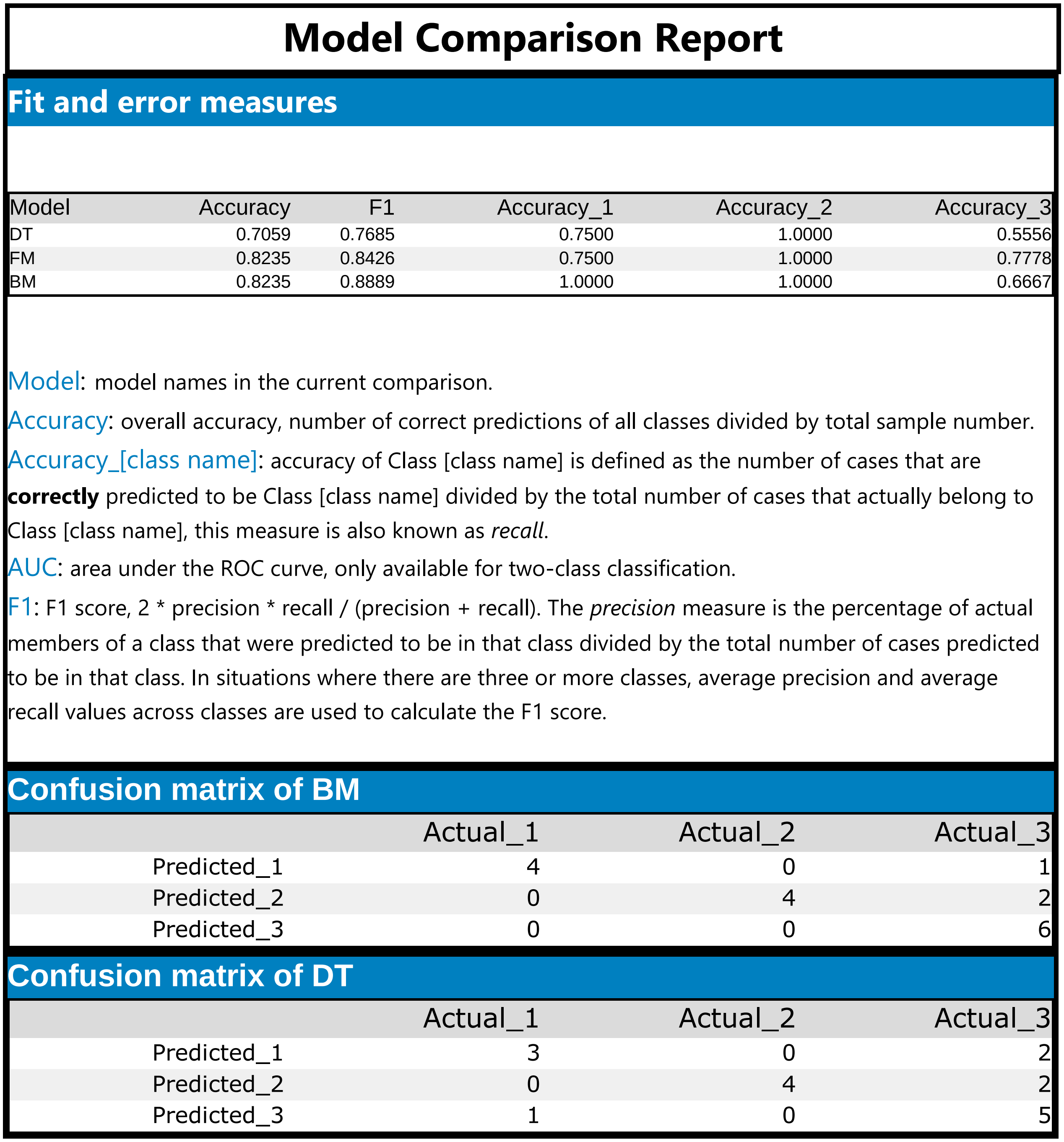
Figure 4: Location of the stores Tableau Dashboard

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.)**

   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	0.7685	0.7500	1.0000	0.5556
FM	0.8235	0.8426	0.7500	1.0000	0.7778
BM	0.8235	0.8889	1.0000	1.0000	0.6667

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

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

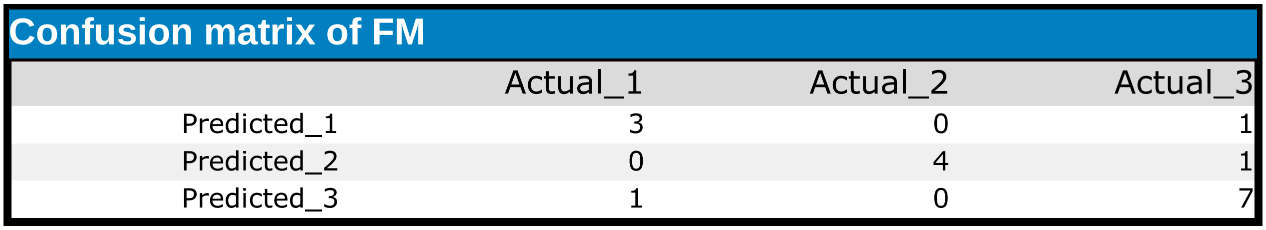
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

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

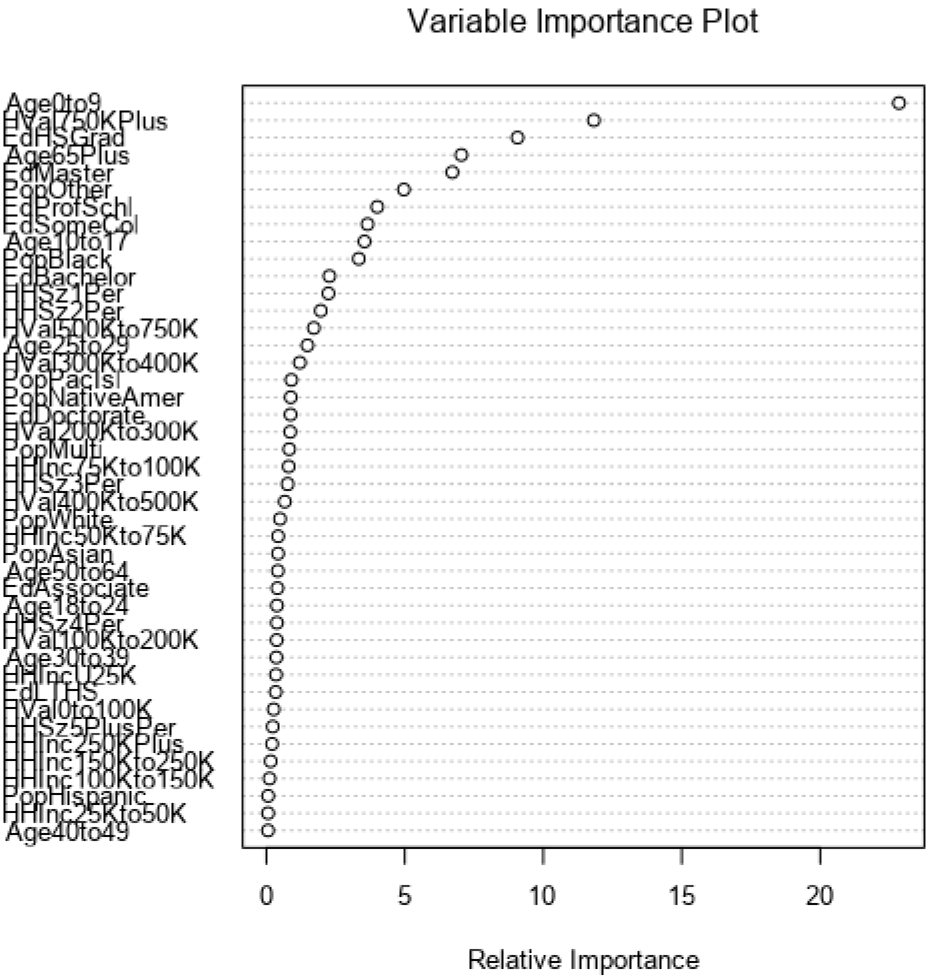
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5



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

Figure 6: Model Comparison Report

2. **What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.**



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

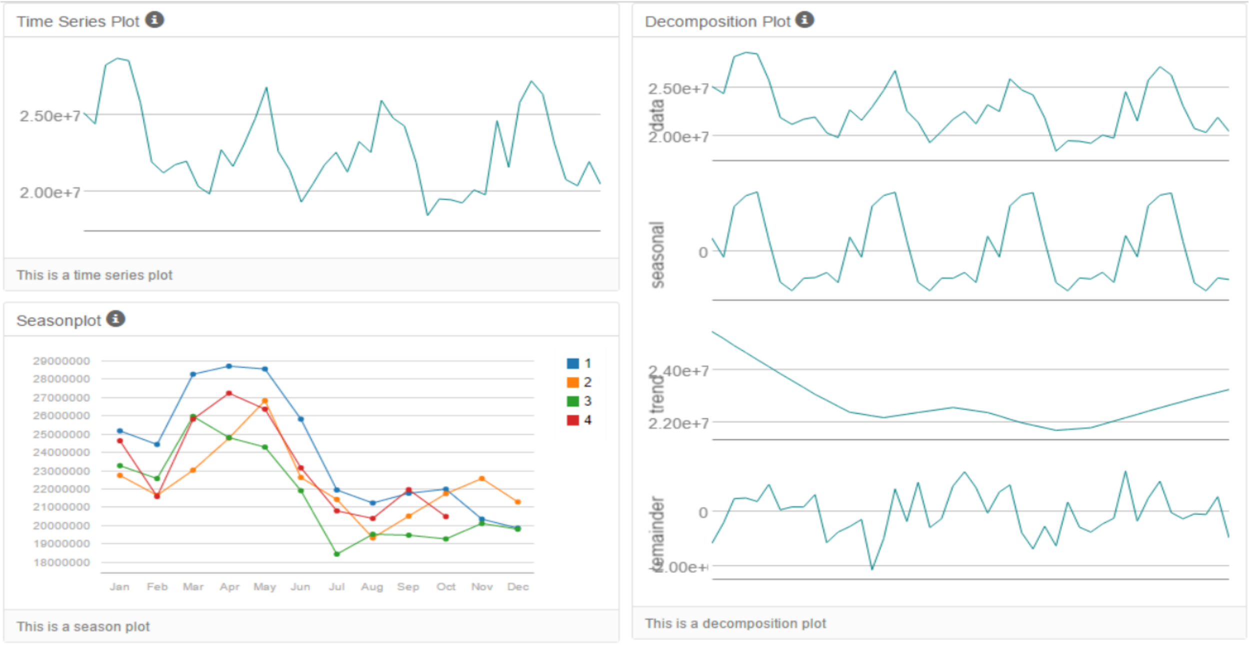
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

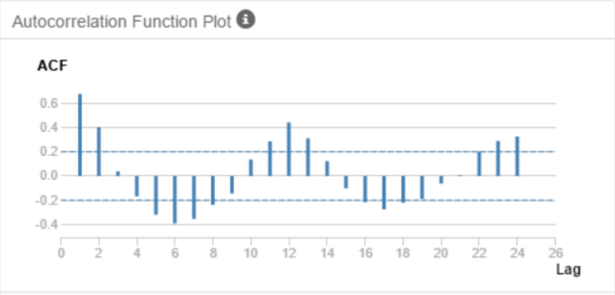
Task 3: Predicting Produce Sales

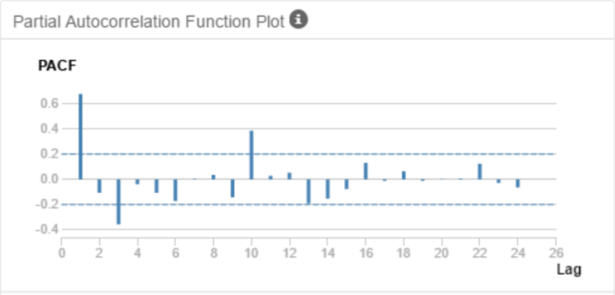
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

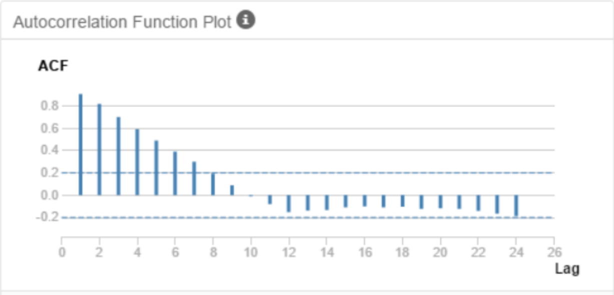
ETS(M,N,M) with no dampening is used for ETS model.

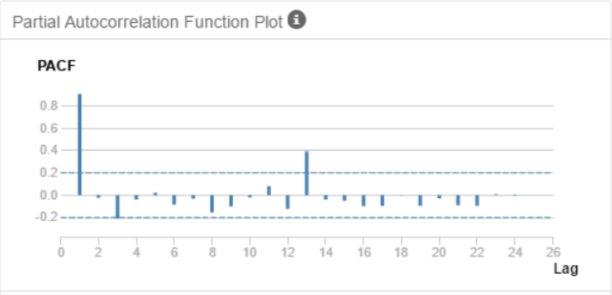
The seasonality shows increasing trend and should be applied multiplicatively. The trend is not clear and nothing should be applied. Its error is irregular and should be applied multiplicatively.

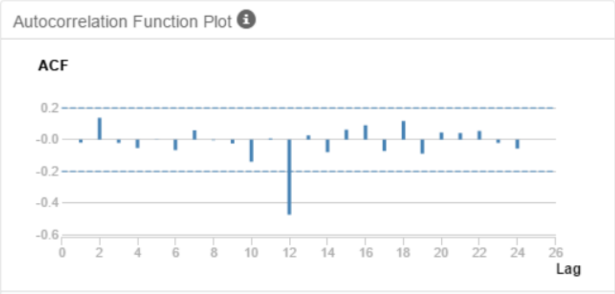


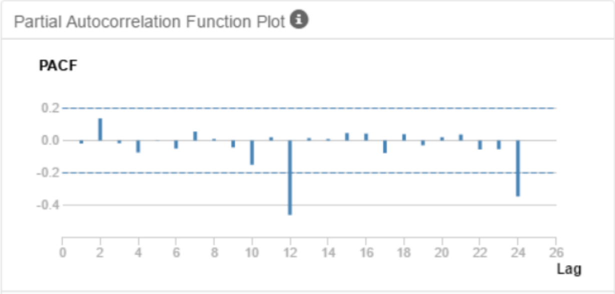




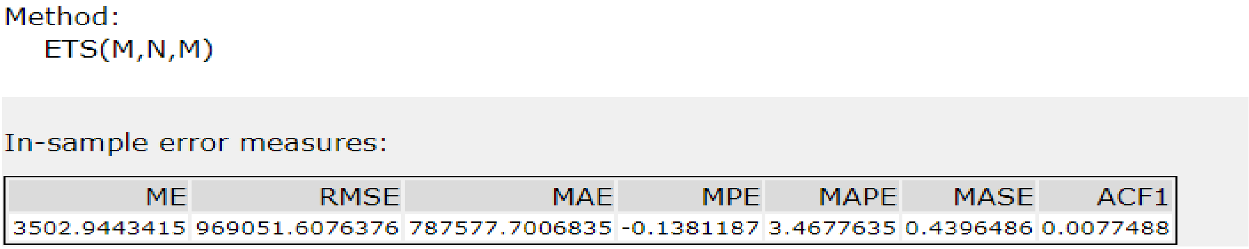








ETS model's accuracy is higher when compared to ARIMA model. A holdout sample of 6 months data is used. Its RMSE of **969,051.60** is lower than ARIMA's **1,429,296** while its MASE is **0.44** compared to ARIMA's **0.53**. ETS is also lower in other error ratings hence

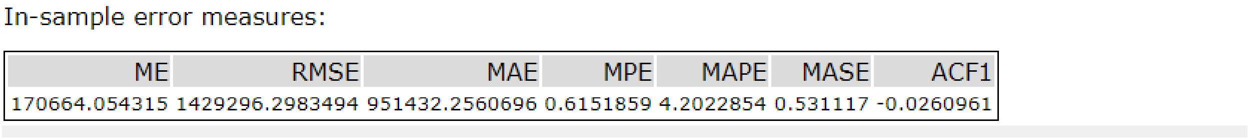


Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Arima

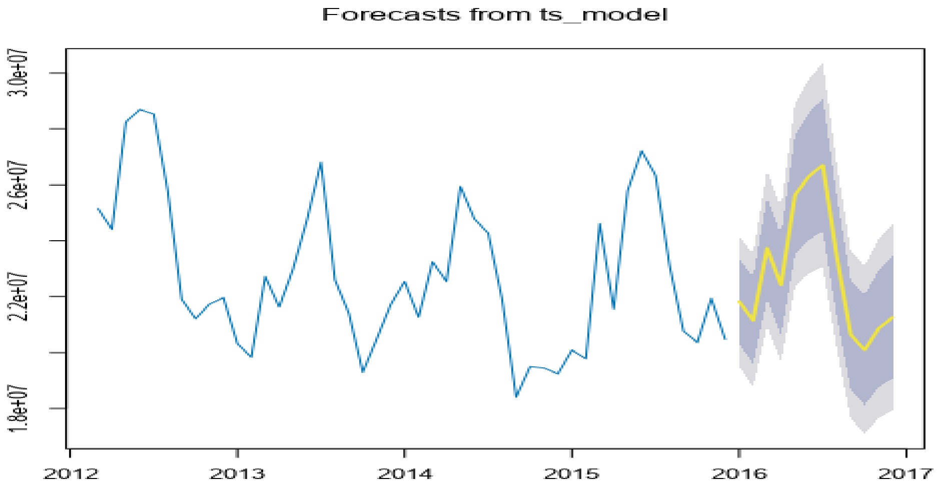


In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961

The graph and table below shows actual and forecast value with 80% & 95% confidence level interval.

12 Period Forecast from ts_model



Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21829060.031666	24149899.115321	23346575.14138	20311544.921952	19508220.948011
2016	2	21146329.631982	23512577.365832	22693535.862148	19599123.401815	18780081.898131
2016	3	23735686.93879	26517865.796798	25554855.912929	21916517.964651	20953508.080782
2016	4	22409515.284474	25150243.401256	24201581.075733	20617449.493214	19668787.167691
2016	5	25621828.725097	28880596.484529	27752622.431914	23491035.018279	22363060.965665
2016	6	26307858.040046	29777680.067343	28576652.715009	24039063.365084	22838036.01275
2016	7	26705092.556349	30348682.320364	29087507.847195	24322677.265503	23061502.792334
2016	8	23440761.329527	26742106.733295	25599395.061562	21282127.597491	20139415.925758
2016	9	20640047.319971	23635033.372194	22598363.439189	18681731.200753	17645061.267747
2016	10	20086270.462075	23084199.797487	22046511.090727	18126029.833423	17088341.126662
2016	11	20858119.95754	24055437.105831	22948733.269445	18767506.645635	17660802.809249
201	12	21255190.244	24596988.126	23440274.430	19070106.059	17913392.363

6	976	893	75	202	058

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.

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.

Year	Month	New Store Sales	Existing Store
2016	1	2,588,249.61	21,829,060.0
2016	2	2,499,158.58	21,146,329.6
2016	3	2,916,908.19	23,735,686.9
2016	4	2,791,560.12	22,409,515.3
2016	5	3,156,890.12	25,621,828.7
2016	6	3,200,940.33	26,307,858.0
2016	7	3,224,857.58	26,705,092.6
2016	8	2,861,958.21	23,440,761.3
2016	9	2,534,352.63	20,640,047.3
2016	10	2,481,117.23	20,086,270.5
2016	11	2,578,335.98	20,858,120.0
2016	12	2,561,916.53	21,255,190.2

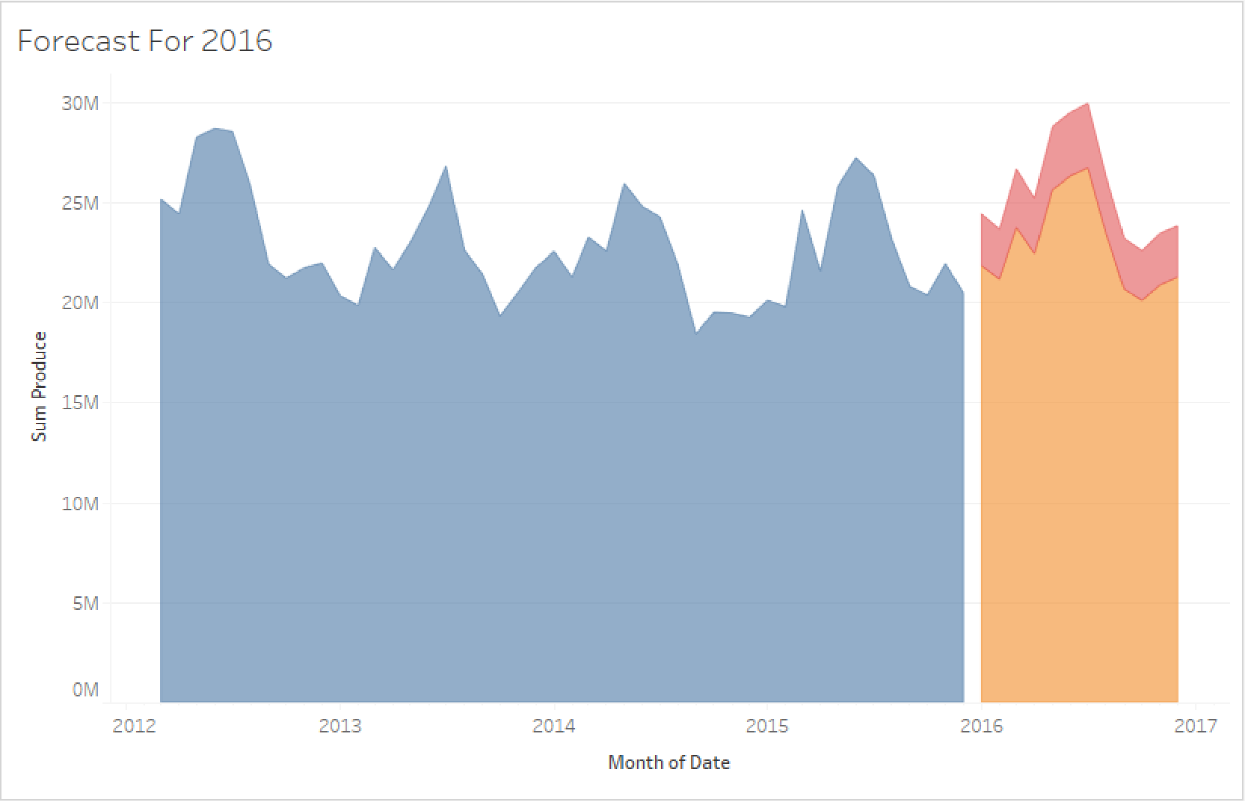


Tableau Dashboard