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 segments is **three**, obtained by K-Centroids diagnostic tool selecting the minimum of clusters 2 and the maximum 8.
   From K-Means cluster assessment report, both Adjusted Rand and Calinski-Harabasz indices box plots registered the highest median value.

2. How many stores fall into each store format?
   According to previous three segments selected, cluster 1 has **23** stores, cluster 2 has **29** stores and 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?
   Cluster 1 has the lowest size of 23 and lowest max distance which is 3.55. Cluster 2 with size of 29 has the highest ave distance. Cluster 3 with the highest number of size has the lowest number of separation which is 1.7.

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.

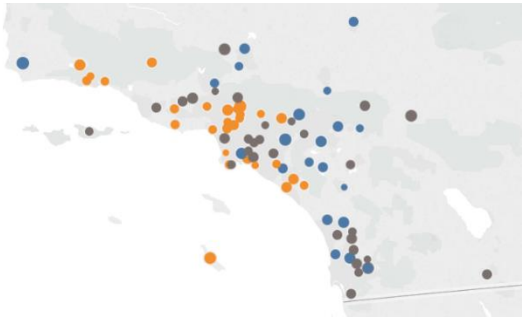


Figure 1: Stores location

Alteryx workflow:

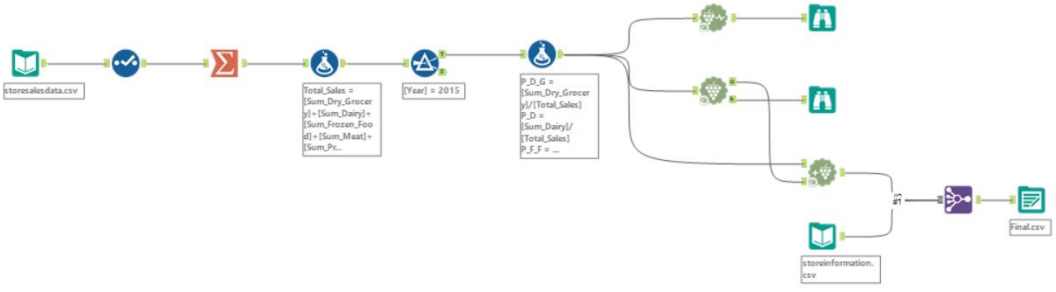
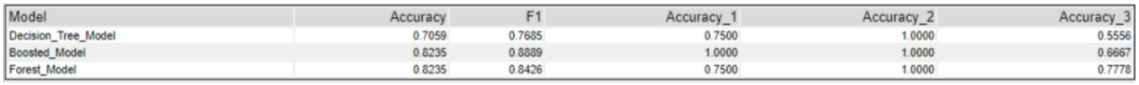


Figure 2

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.)
   After constructing and comparing the three models decision tree, **Boosted model** was chosen to predict the best store format for the new stores since it has a higher accuracy and F1 score among the other models.



Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Model	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778

Figure 3

2. 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

Table 1

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?

From the original dataset Store_Sales_Data, we compared the performance of ETS and ARIMA model to select one of them to be applied for the stores' forecast using TS tools.

Summary of Time Series Exponential Smoothing Model ETS_

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

Information criteria:

AIC	AICc	BIC
1279.4203	1299.4203	1304.7535

Smoothing parameters:

Parameter	Value
alpha	0.674884
gamma	0.000203

Initial states:

State	Value
l	23146230.586012
s0	0.90906
s1	0.938619
s2	0.926304
s3	0.901291
s4	0.870972
s5	0.897637
s6	1.019225
s7	1.166556
s8	1.167388
s9	1.137259
s10	0.997793

Figure 4

Summary of ARIMA Model ARIMA_

Method: ARIMA(1,0,0)(1,1,0)[12]

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616

Figure 5

Actual and Forecast Values:

Actual	ETS_	ARIMA_
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy 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 6

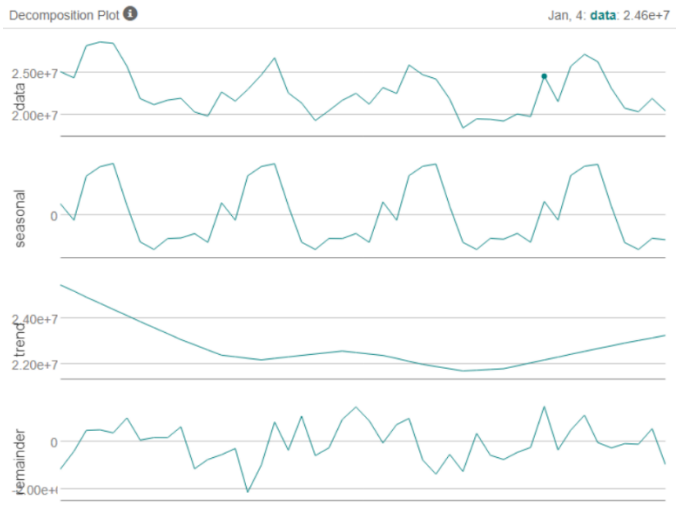


Figure 7

Based on reports and the decomposition plot above, EST model is the best model to forecast the produce sales for the new and existing store.

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.

Date	Forecast	New Stores Sales
2016/1	21829060	2603262
2016/2	21146330	2508878
2016/3	23735687	2989458
2016/4	22409515	2849287
2016/5	25621829	3224711
2016/6	26707858	3269623
2016/7	26705093	3288334
2016/8	23440761	2937302
2016/9	20640047	2606592
2016/10	20086270	2536270
2016/11	20858120	2631293
2016/12	21255190	2586562

Table 2: Forecasts

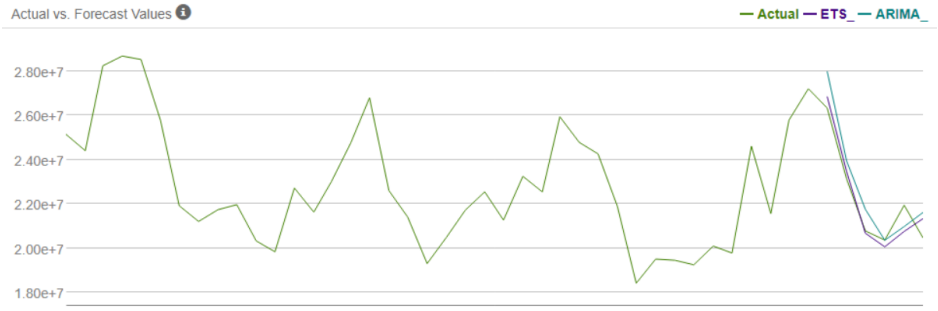


Figure 8: Forecast visualization

Alteryx Workflow:

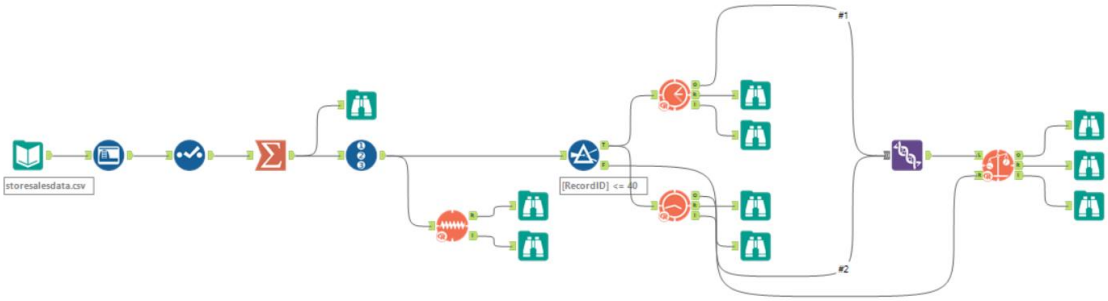


Figure 9