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?

Figure 1: K-Means Cluster Assessment Report

K-Means Cluster Assessment Report

Summary Statistics

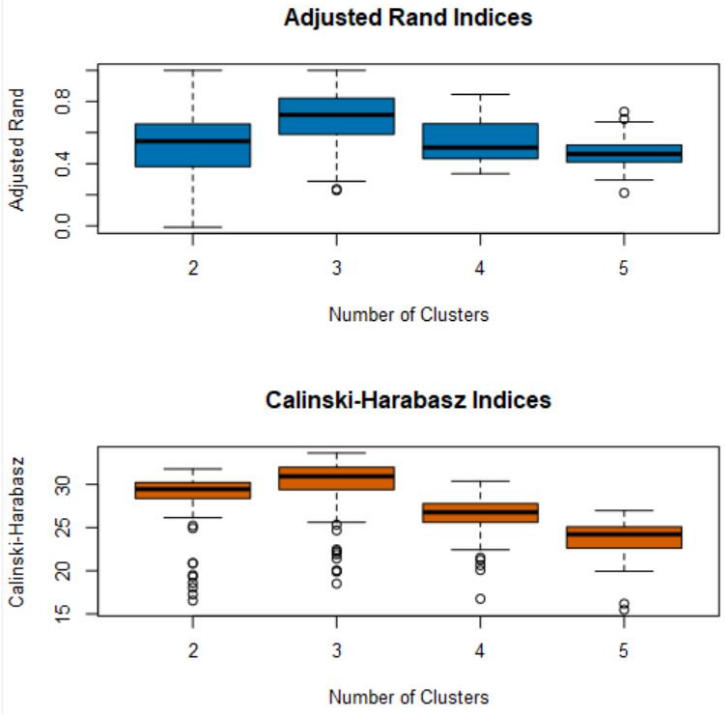
Adjusted Rand Indices:

	2	3	4	5
Minimum	-0.007972	0.228156	0.335359	0.212462
1st Quartile	0.381421	0.593906	0.434856	0.410809
Median	0.544002	0.713886	0.503544	0.462071
Mean	0.503774	0.688385	0.534777	0.471591
3rd Quartile	0.654956	0.820181	0.656084	0.516916
Maximum	1	1	0.845268	0.73396

Calinski-Harabasz Indices:

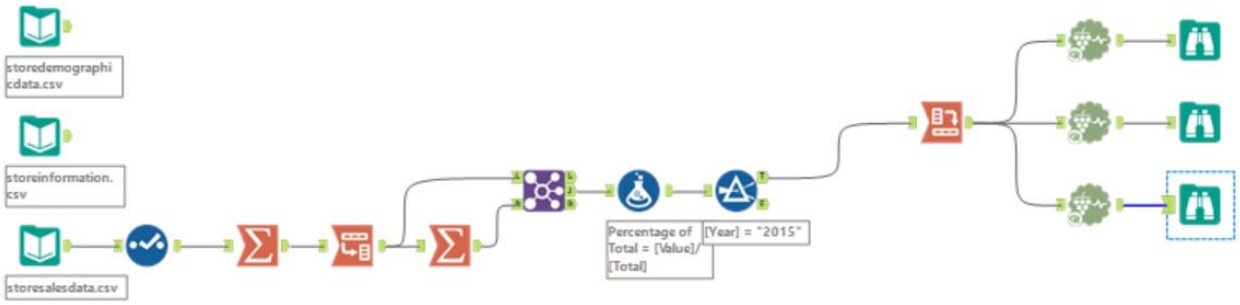
	2	3	4	5
Minimum	16.543	18.50776	16.75642	15.48421
1st Quartile	28.42185	29.41372	25.63743	22.63778
Median	29.45069	30.93662	26.76851	24.21592
Mean	28.56415	29.83325	26.41482	23.69305
3rd Quartile	30.21413	31.97449	27.76499	25.08265
Maximum	31.78345	33.63781	30.37935	26.97019

Figure 2: Adjusted Rand Indices and Calinski-Harabasz Indice



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.

Workflow 1: Alteryx workflow for task 1



2. How many stores fall into each store format?

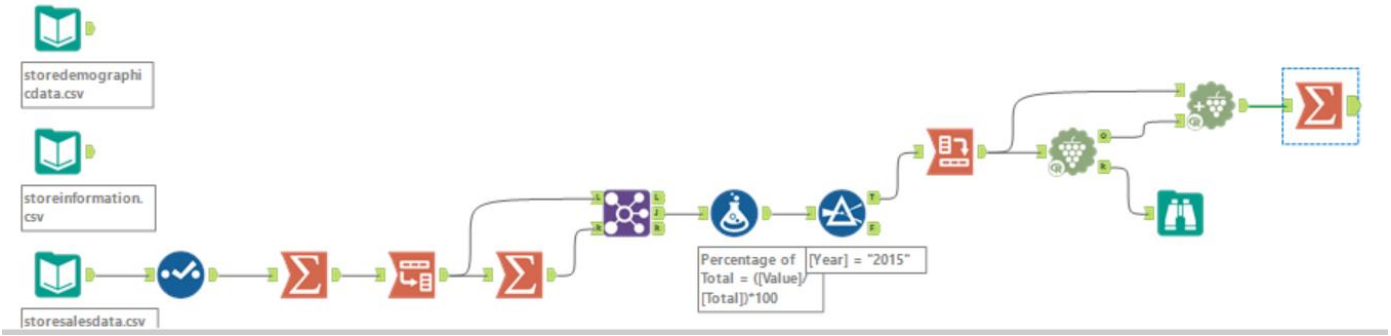
Cluster 1 has 23, Cluster 2 has 29 and Cluster 3 has 33

Figure 3: Cluster information

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

Workflow 2: Alteryx workflow for task 2



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

Cluster 3 has a higher average median than the other clusters. Stores in cluster 3 are most similar regarding sale volume as seen by their relatively tight compactness.
Cluster 1 sells a higher percentage of General Merchandise goods whereas cluster 2 sells a higher proportion of Produce.

Figure 3: Tableau visualization

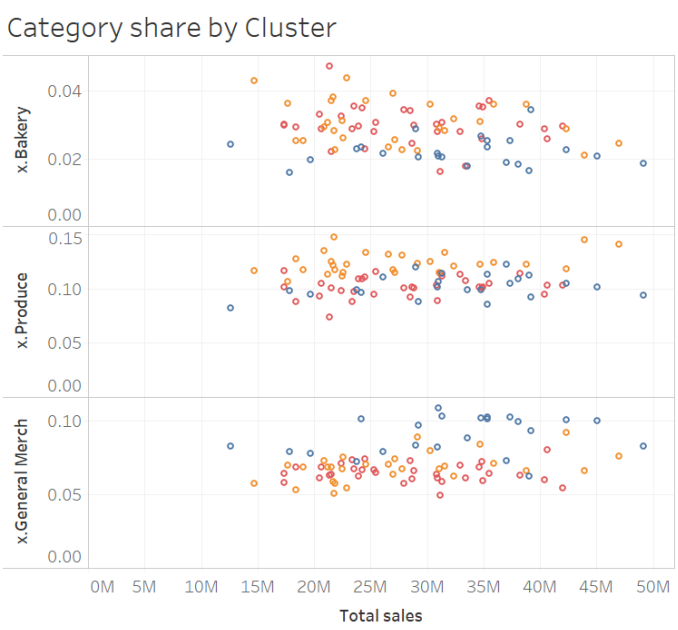


Tableau public link: https://public.tableau.com/shared/ZB29XXTXQ?:display_count=y&:origin=viz_share_link

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: Store location and size Tableau visualisation

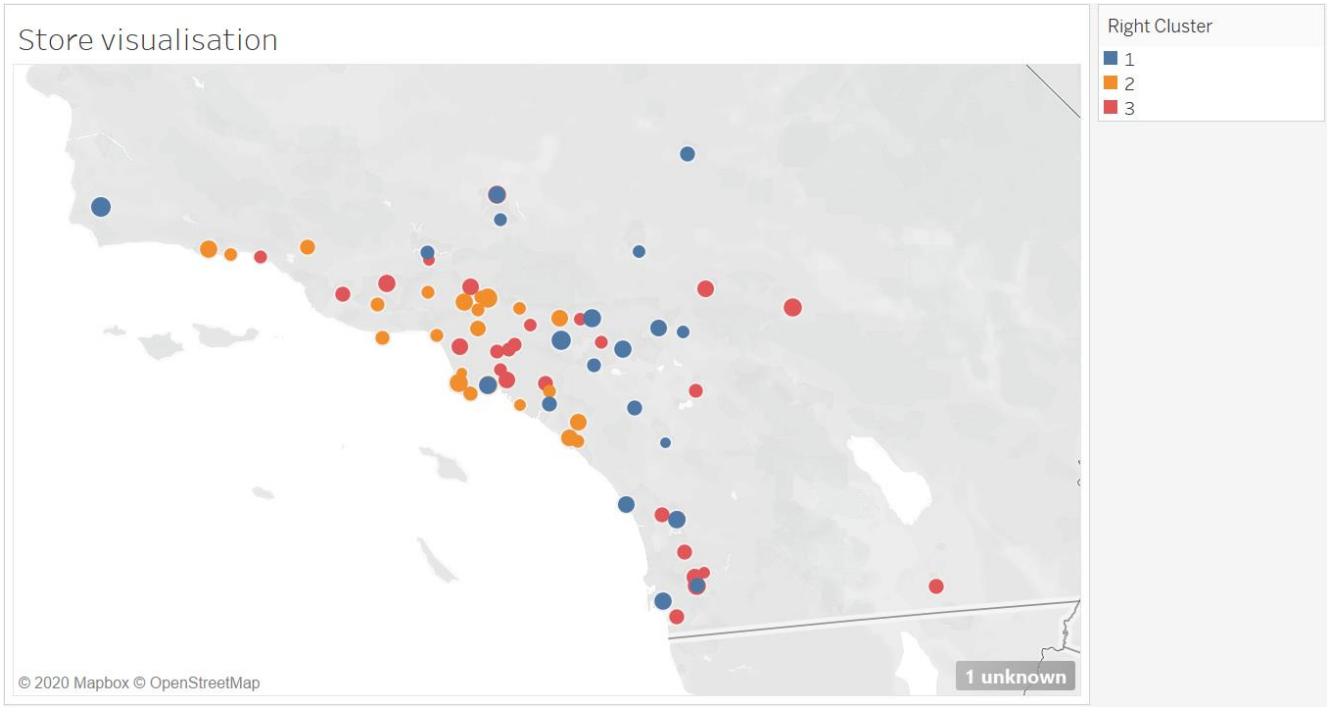


Tableau public link: https://public.tableau.com/views/Tableaumap_finalproject/Sheet1?:language=en-GB&:display_count=y&publish=yes&:origin=viz_share_link

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 was chosen as it had the highest F1 score, which shows the precision of the model and the cluster segment-specific accuracy was higher overall.

Figure 5: Model comparison report

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Cluster_forest	0.8235	0.8426	0.7500	1.0000	0.7778
Cluster_boosted	0.8235	0.8889	1.0000	1.0000	0.6667
Cluster_DT	0.8235	0.8426	0.7500	1.0000	0.7778

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 Cluster_DT

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

Confusion matrix of Cluster_boosted

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

Confusion matrix of Cluster_forest

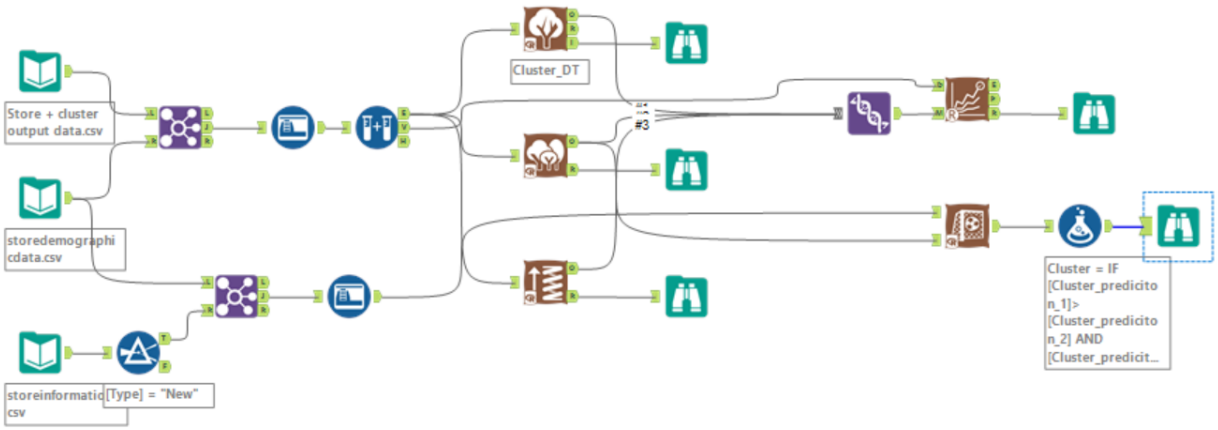
	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

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

Table 1: Store number and segment

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

Workflow 2: Alteryx workflow for task 2

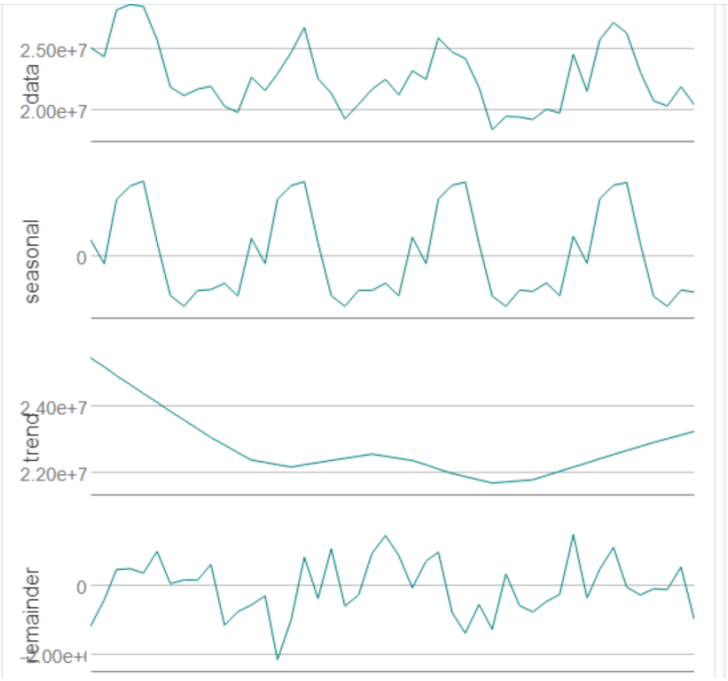


Task 3: Predicting Produce Sales

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?

An ETS(M,N,M) with no damping was used as we can see from the decomposition graph seasonality shows increasing trend and should be applied multiplicatively, trend is unclear and error is irregular and should be applied multiplicatively.
The ARIMA model was set to auto and was optimised to ARIMA (2,0,2)(1,0,1)

Figure 6: Decomposition graphs



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_forecast	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463
ETS_forecast	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257

As we can see from the accuracy measures, RSMA, MAPE and MASE are lower for ETS than for ARIMA, indicating is it the stronger model. A 6-month holdout sample was used.

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 2: 2016 monthly sales forecast for new and existing stores

Year	Month	New	Existing
2016	1	2558242	21829060
2016	2	2468197	21146330
2016	3	2883620	23735687
2016	4	2762836	22409515
2016	5	3129542	25621829
2016	6	3170534	26307858
2016	7	3199198	26705093
2016	8	2842411	23440761
2016	9	2512051	20640047
2016	10	2460445	20086270
2016	11	2555392	20858120
2016	12	2534378	21255190

The chart below shows the historical and forecast sales for existing stores and new stores from Mar-12 to Dec-16.

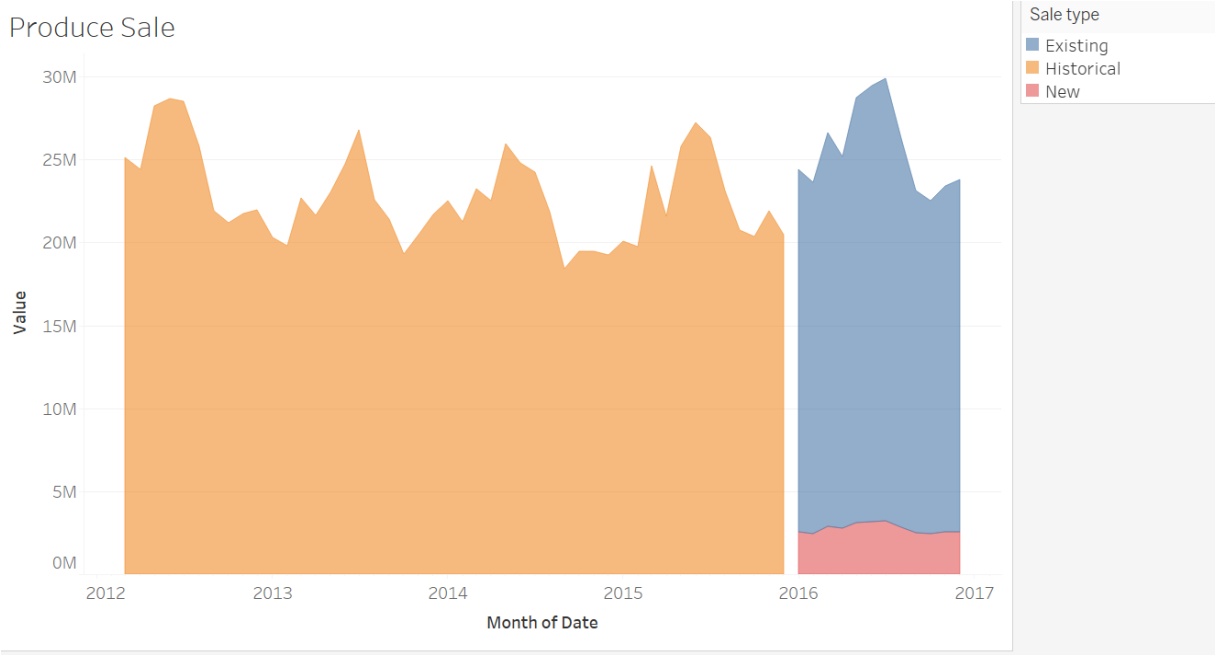


Tableau public: https://public.tableau.com/views/Producesalegraphs/Sheet2?:language=en-GB&:display_count=y&publish=yes&:origin=viz_share_link

Workflow 3: Alteryx workflow for task 3