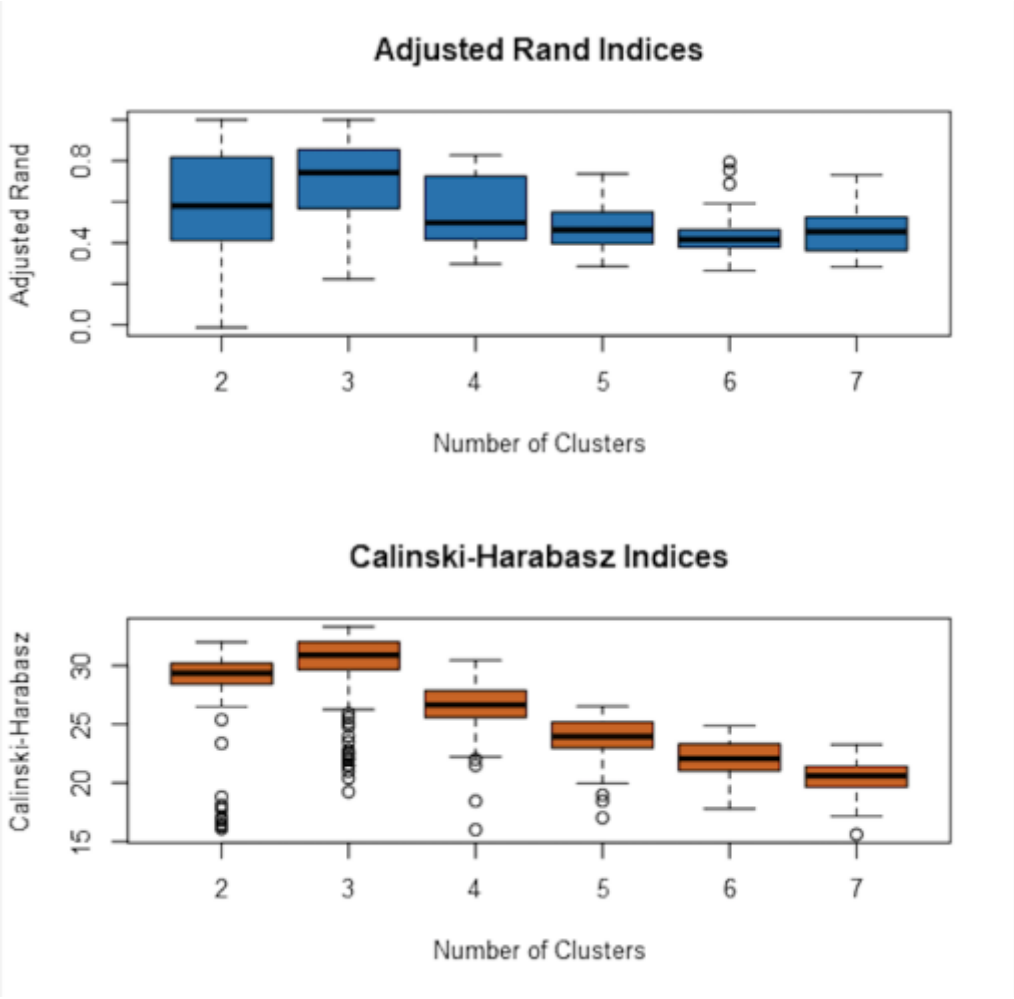
Predictive Analytics Capstone
Segmentation and Clustering and Time Series Analysis
Predicting Produce Sales

Determining Store Formats for Existing Stores

Using the sales data and information provided, stores can be segmented and clustered into 3 store clusters using the percentage of categorical product sales. This optimal cluster is achieved using a K-Means Clustering method. Through this analysis, visualisations and statistics of Calinksi-Harabasz (CH) and Adjusted Rand Index were calculated. From the summary statistics, the mean and medium of Adjusted Rand and CH index of Cluster #3 were the highest, indicating high stability and compactness of the cluster and high distinctness from others clusters. With 3 clusters, store formats are divided into 23 stores in Cluster 1, 29 in Cluster 2 and 33 in Cluster 3.



Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	-0.01304	0.2228	0.2966	0.2847	0.2642	0.282
1st Quartile	0.4117	0.5771	0.4162	0.397	0.3776	0.3651
Median	0.5798	0.7425	0.4977	0.4635	0.4164	0.4548
Mean	0.5397	0.705	0.5452	0.4709	0.4378	0.4513
3rd Quartile	0.7975	0.8492	0.717	0.5435	0.4609	0.5189
Maximum	1	1	0.8277	0.7366	0.7931	0.7308

Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	16.1	19.24	16.01	17.03	17.79	15.61
1st Quartile	28.42	29.68	25.58	22.98	21.05	19.64
Median	29.35	30.89	26.63	23.95	22.06	20.59
Mean	28.36	29.78	26.46	23.83	22.08	20.56
3rd Quartile	30.17	32	27.86	25.18	23.31	21.41
Maximum	32	33.31	30.45	26.53	24.87	23.27

K-Means Clustering Solution

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

The table shows that the size of each cluster is not that different, which means that no cluster is an outlier.

It also shows that average distance, which represents the average distance of the objects within the cluster to the centroid of cluster, of Cluster 3 is the smallest. The cluster with the smallest distance is the most compact.

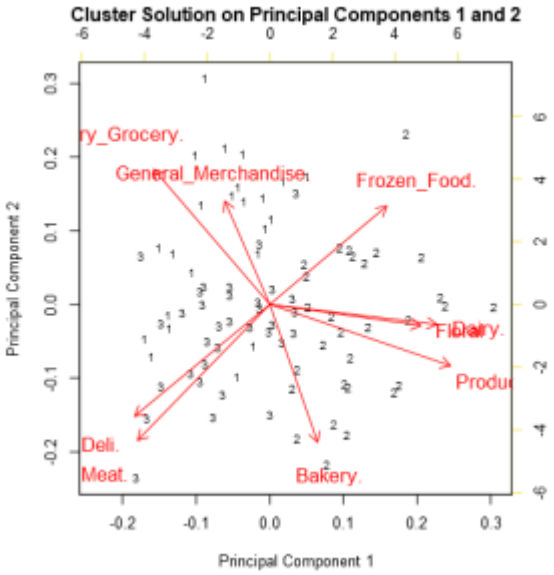
Max_distance shows that the object with the furthest distance from the centroid. This could just be a single point, showing the furthest outlier. And Cluster 3 has the object with further point.

Separation in the table shows the closet point not in the relevant cluster. Cluster 2 seems to have the furthest separation. The larger the number, the more separation it is from the other cluster and the better for the model.

	Dry_Grocery.	Dairy.	Frozen_Food.	Meat.	Produce.	Floral	Deli.
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Bakery.	General_Merchandise.
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

The variables shown generally follow a trend of high positive or high negative when they are in different cluster. This shows that they are opposites. Moreover, positivity and negativity do not have any direct correlation with sales.

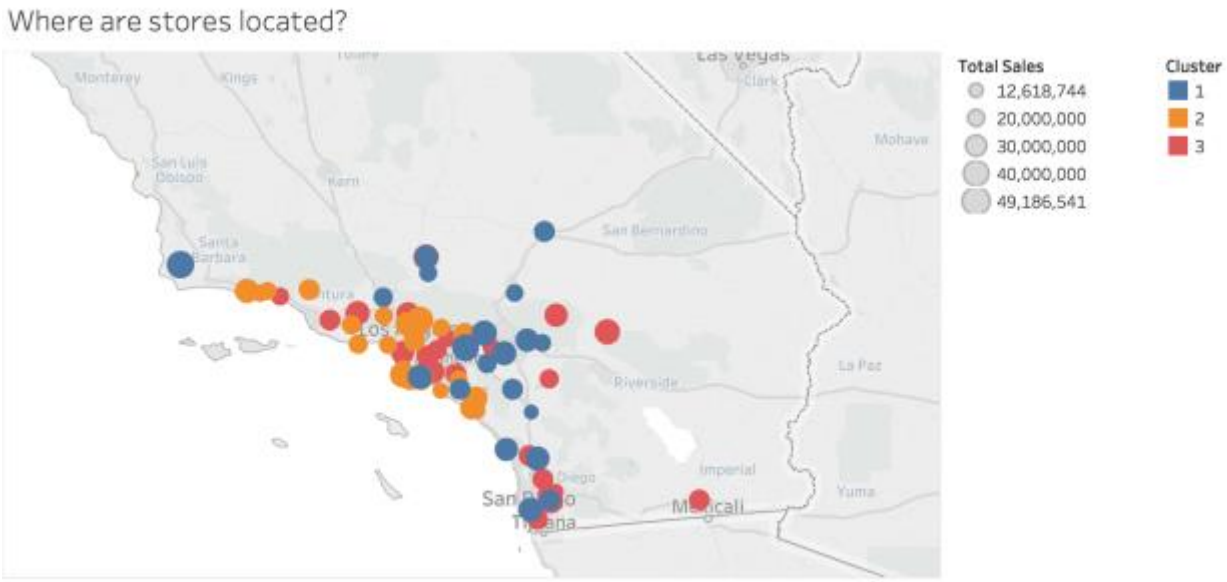


Evidence is further supported by the visualisation on the left.

Cluster 2 objects are more spread out and further apart from Cluster 1 and 3.

Cluster 3 objects are less spread out. The stores in Cluster 3 are more compacted compared to the stores in the other two clusters.

Tableau Visualisation of stores location



https://public.tableau.com/profile/julie.ly#!/vizhome/ProduceSales_3/Dashboard1

Formating for New Stores

In order to predict which of the 3 clusters the new stores will fall into, the new stores' demographic data was passed into a variety of classification model. After comparing Decision Tree, Forest Model and Boosted Model, Forest Model and Boosted Model has equal overall

accuracy rate at 0.8235. However, the Boosted Model was chosen to classify new stores category because it has higher accuracy for predicting Cluster 1 and 3 at 0.800 and 1.00.

Model	Accuracy	F1	Accuracy 1	Accuracy 2	Accuracy 3
Forest Model	0.8235	0.8251	0.7500	0.800	0.8750
Decision Tree	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted Model	0.8235	0.8543	0.8000	0.6667	1.0000

Store Format

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

Predicting Produce Sales of Existing and New stores

In order to pick the better model to predict the sum of produce sales, both ETS and ARIMA models were built and compared against each other using in-sample error measures and validated against the holdout sample of 12 months (2015 data).

ETS Model

The characteristics of Error, Trend and Seasonality were determined using decomposition plot.

Image of decomposition plot

The seasonal graph looks constant at first but upon a closer inspection, there is a small decrease over time. This suggests that any ARIMA models used for analysis will need seasonal differencing and any ETS models will use a multiplicative method.

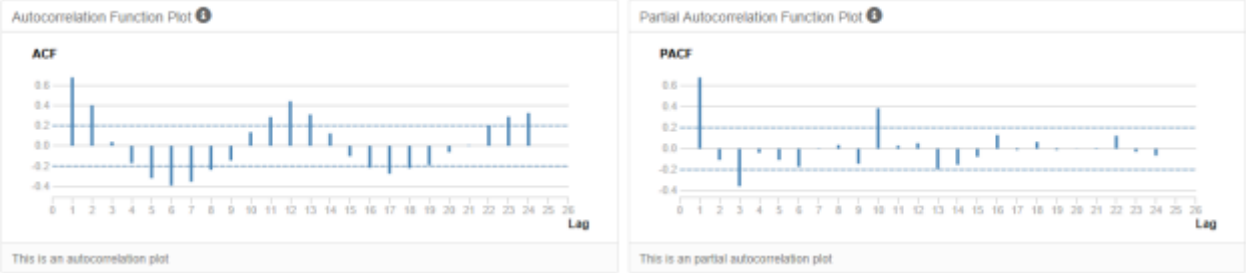
The trend graph shows an upward and downward trend. Given more dataset, one might be able to spot any linear or increasing trend. But with the amount of data given, there is no significant linear or increasing trend. Thus, ETS trend would be None.

The remainder of the sales data or the error component shows a general trend of increasing in error over time. This shows that a multiplicative method needs to be applied in an ETS model.

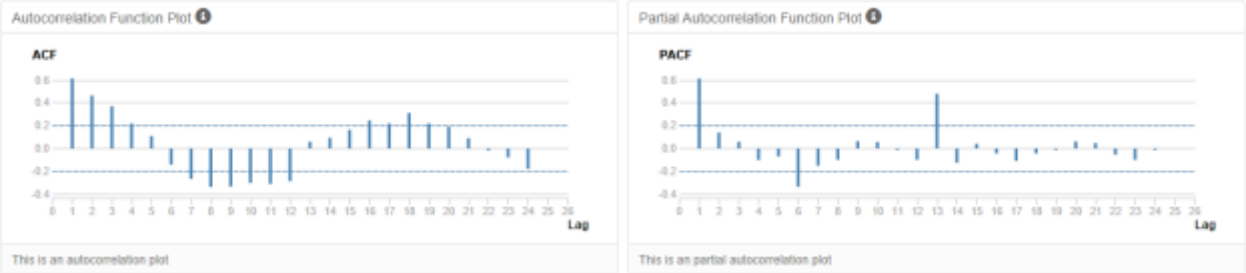
ARIMA Model

The ARIMA model chosen was ARIMA(0,1,1)(0,1,1)[12].
Such conclusion is arrived after inspecting the ACF and PACF plots after first seasonal differencing. Without seasonal difference, ACF and PACF do not exhibit stationary behavior and could be confused for AR(1) as ACF slowly decay towards 0 with increasing seasonal lags and PACF plot quickly cuts off at lag-1. After seasonal differencing (of 12 months), ACF did not present much changes and PACF decays towards 0. After the first seasonal difference, ACF and PACF show a strong negative correlation at lag 1 and 12(and decaying towards 0 at PACF). This suggested an MA(1) term on non-seasonal and seasonal ARIMA. After adding MA(1), there is no significant correlation lags in the ACF and PACF plots. Thus, resulting in ARIMA(0,1,1)(0,1,1)[12].

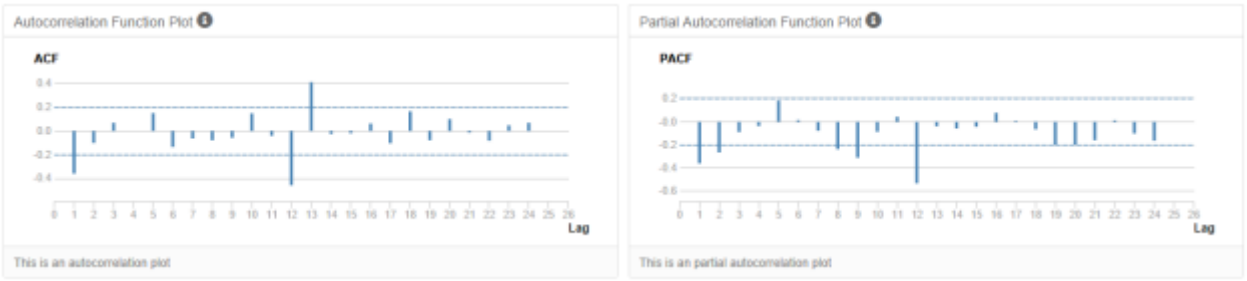
ACF/PACF before seasonal difference



ACF/PACF after seasonal difference



ACF/PACF after first seasonal difference



After ARIMA(0,1,1)(0,1,1)[12]

ETS and ARIMA comparison

ETS(M,N,M) in-sample error measures

ME	RMSE	MAE	MPE	MAPE	MASE	AIC
-241658.32	886787.76	699047.47	-1.158	3.13	0.37	1078.95

ARIMA(0,1,1)(0,1,2)[12] in-sample error measures

ME	RMSE	MAE	MPE	MAPE	MASE	AIC
44805.79	862271.08	533233.61	0.203	2.38	0.28	661.85

Comparing ETS and ARIMA models with their in-sample error measures, ARIMA exhibits criteria that makes it a better model than ETS. ARIMA has lower RMSE, MAPE, MASE, AIC and MPE. This show that on average it has lower percentage difference and standard deviation (residual errors) between its actual and forecasted values. Both models have an ideal value of MASE (<1) but ARIMA's MASE is relatively better. Its low MASE indicates that it yields a smaller in-sample absolute error than ETS(M,N,M) Model compared to a naïve random walk model.

However, in-sample errors between models are not necessarily an efficient way to decide which model is better than the other. Thus, both models were validated against its holdout sample to test for accuracy.

Accuracy Measures - ETS(M,N,M) and ARIMA(0,1,1)(0,1,1)[12]

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	1978789	2200153	1978789	8.4769	8.4769	1.266
ARIMA	2878344	3061362	2878344	12.58	12.58	1.84

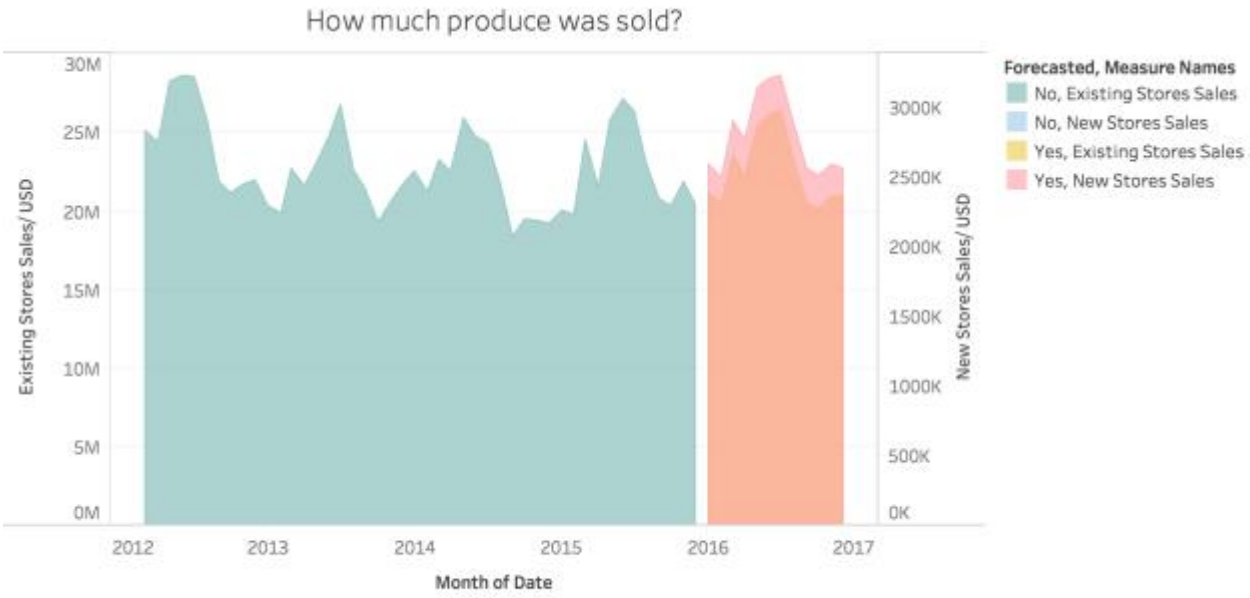
When comparing both models against its holdout sample, ETS has lower ME, RMSE, MAE, MPE, MAPE and MASE. In terms of predicting actual sales values, ETS also has less percentage difference between the actual value and its forecasted value compared to ARIMA model. This made ETS(M,N,M) a better model for predicting 2016 produce sales data.

Percentage difference table

Actual	ETS	% Difference (ETS)	ARIMA	% Difference (ARIMA)
20088529.29	19592858.5	0.024674319	19195237.94	0.044467732
19772333.34	18699732.55	0.054247558	18307740.59	0.074072833
24608406.71	21034791.5	0.145219284	20661282.29	0.160397399
21559729.45	20376418.34	0.054885249	19813841.49	0.080979122
25792074.59	23165379.97	0.101841153	22686956.9	0.120390381
27212464.15	23673094.98	0.130064266	23032374.34	0.153609382
26338477.15	24203475.91	0.081060162	23490052.6	0.108146896
23130626.6	21294826.72	0.07936663	20380673.34	0.118887971
20774415.93	19013560.45	0.08476077	17530847.03	0.156132857
20359980.58	18512113.98	0.090759743	16955389.33	0.167219769
21936906.81	19125133.79	0.128175455	17513078.26	0.201661455
20462899.3	19599985.21	0.042169688	17929240.16	0.123817212
	Average Difference %	0.08476869		0.125815251

After running ETS(M,N,M) Model on existing stores, the sum of historical produce sales data was grouped by store, year, month and cluster. With no previous data for new stores, the new stores were clustered into 3 different groups using demographic data around its area. With that information, the average produce sales was calculated after grouping by year, month and cluster again. After running the ETS on each cluster, the forecasted produce sales were multiply by the number of new stores in that cluster, where the total sales of each cluster are added together and taken as New Stores Forecast for 2016. The table and visualisation showing the existing and new stores forecast for the 2016 produce sales is shown below.

Year	Month	Existing Store Forecast	New Stores Forecast
2016	1	21174989.4	2590566.586
2016	2	20479354.58	2503135.097
2016	3	23580340.68	2910154.08
2016	4	22236546.23	2772193.192
2016	5	25427255.46	3142262.476
2016	6	26143967.4	3203694.415
2016	7	26399993.27	3233436.116
2016	8	23172393.88	2884618.003
2016	9	20544268.64	2562088.683
2016	10	20182471.09	2506670.54
2016	11	20966876.35	2598150.832
2016	12	20965097	2566314.036



The plots of Existing Stores Sales and New Stores Sales for Date Month. Color shows details about Forecasted, Existing Stores Sales and New Stores Sales.

https://public.tableau.com/profile/julie.ly#!/vizhome/ProduceSales_3/Dashboard1

After using K-Means to segment existing stores into 3 different clusters based on their percentage categorical sales, the new stores were classified into the existing clusters based on the similar demographic data that they shared with existing stores. This is done through Boosted Model classification. With stores segmented into their distinctive group, ETS(M,N,M) model was then used to predict the 2016 produce sales of both existing and new stores.