Project 7: Predictive Analytics Capstone

By: Huong Ivy Nguyen

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. This value is chosen based on the results of the adjusted rand indices and the C-H indices obtained from running the Alteryx workflow 1 (Figures 1, 2).

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.02152	0.2941	0.2633	0.2877	0.3025	0.3497	0.3043
1st Quartile	0.3111	0.6158	0.4196	0.3965	0.3724	0.3812	0.3853
Median	0.4759	0.735	0.5094	0.4631	0.4476	0.4196	0.4365
Mean	0.5062	0.7214	0.5274	0.4753	0.4599	0.4527	0.4474
3rd Quartile	0.8195	0.823	0.6197	0.53	0.5357	0.5247	0.5063
Maximum	1	0.969	0.9184	0.7502	0.6878	0.6773	0.6124
	9	10					
Minimum	0.2804	0.2748					
1st Quartile	0.4035	0.3728					
Median	0.4503	0.4464					
Mean	0.4572	0.4364					
3rd Quartile	0.5157	0.4906					
Maximum	0.6371	0.602					

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	15.3	18.18	19.48	19.88	16.36	15.88	16.39
1st Quartile	28.26	30.33	25	23.01	21.38	20.09	18.84
Median	29.32	31.1	26.43	24.5	22.35	21.07	19.82
Mean	28.08	30.56	26.24	24.04	22.2	20.92	19.77
3rd Quartile	30.13	32.2	27.57	25.19	22.96	21.96	20.76
Maximum	31.58	33.57	30.14	26.85	24.69	24.23	23.16
	9	10					
Minimum	16.21	14.33					
1st Quartile	18.1	17.03					
Median	18.97	17.97					
Mean	19	17.87					
3rd Quartile	19.92	18.83					
Maximum	22.01	20.62					

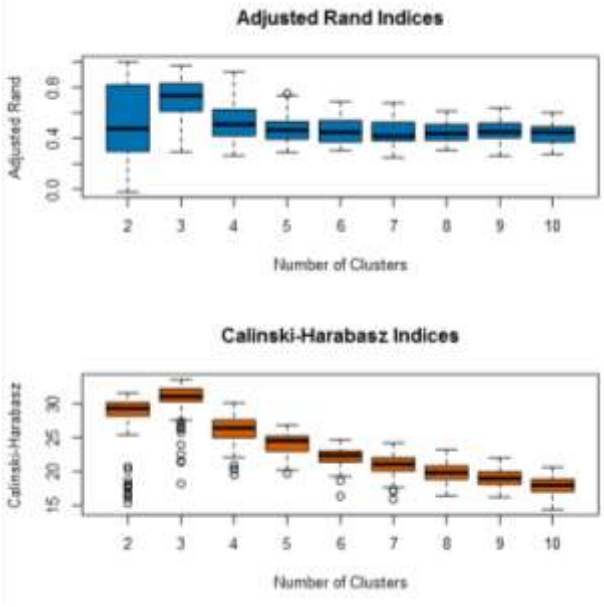


Figure 1. The adjusted rand indices and Calinski-Harabasz indices plots, which are used to determine the optimal number of clusters.

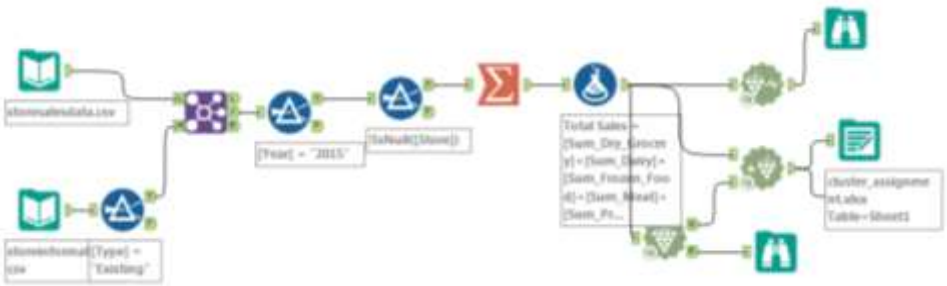


Figure 2. Evaluating the number of clusters that should be used in the K-mean clustering
Based on Figure 1, the following features can be pointed out:

By combining the features obtained from both the adjusted rand indices and the C-H indices, cluster 3 is chosen since it has the highest median value for both indices and its narrow IQR.

2. How many stores fall into each store format?

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

Figure 3. Report for the number of existing stores fall into each cluster

There are 23 existing stores in cluster 1, 29 existing stores in cluster 2, and 33 existing stores in cluster 3.

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

Based on Figure 6 and Figure 7, it appears that cluster 2 sold more bakery while cluster 1 sold more dairy. Overall, cluster 3 sold more in total in comparison to the other two clusters.

Figure 6. Total sales distribution of all the stores in all clusters where blue is cluster 1, orange is cluster 2, and gray is cluster 3. The size of the circle represents the amount of total sale for each store.

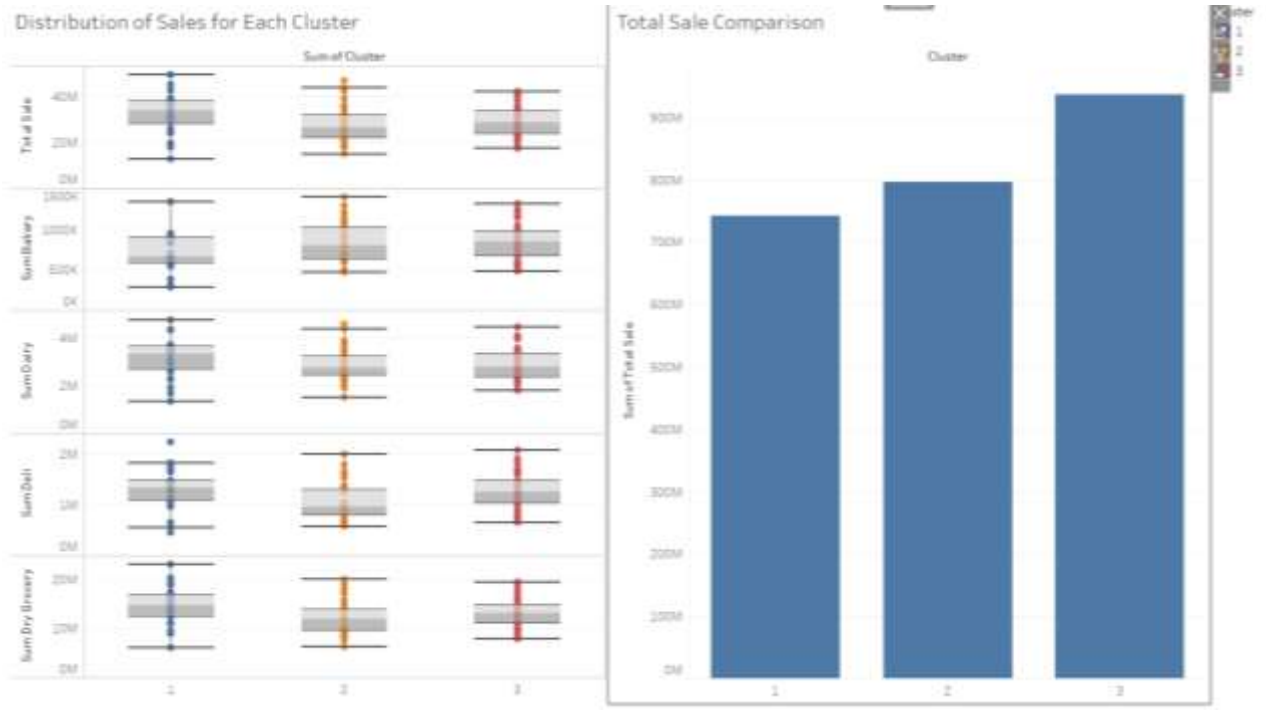


Figure 7. Left figure: whiskey-boxplot of the total sales, sum of bakery, sum of dairy, sum of deli, and sum of dry grocery for each cluster. (Blue: cluster 1, Orange: cluster 2, Gray: cluster 3). Right figure: bar plot shows the total sales of each cluster.

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 6:
https://public.tableau.com/profile/ivy.nguyen#!/vizhome/Project7_Business_Analytics_Udacity/Location_sales

Figure 7:
https://public.tableau.com/profile/ivy.nguyen#!/vizhome/Project7_Business_Analytics_Udacity/Whiskey_boxplots_sumsales?publish=yes

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 cluster assignment for the new stores are determined by using the Alteryx workflow 2 (Figure 5).

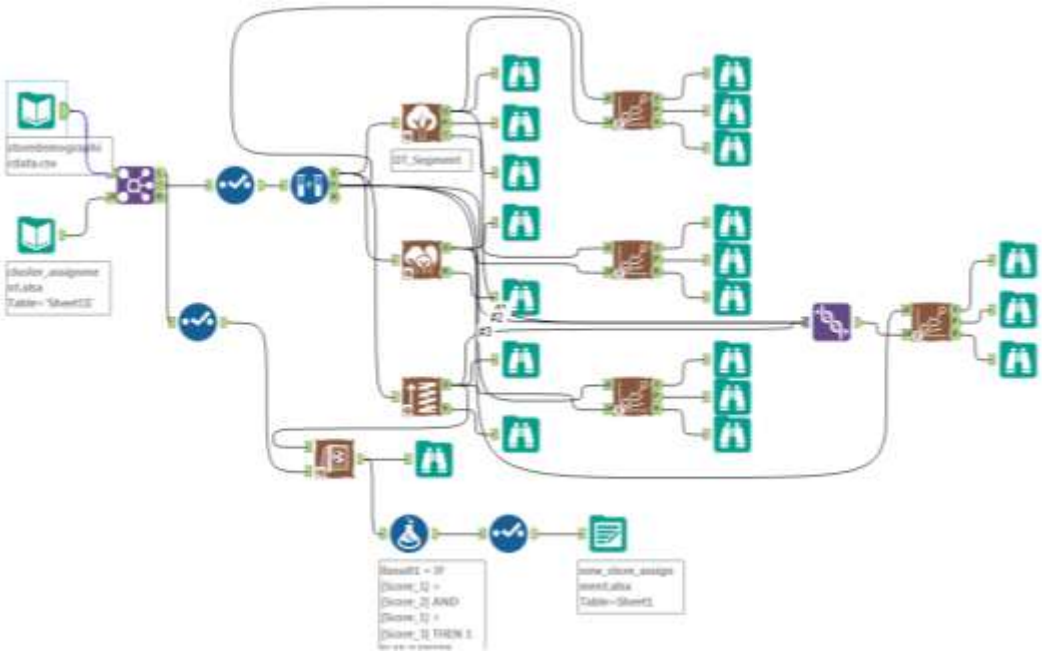
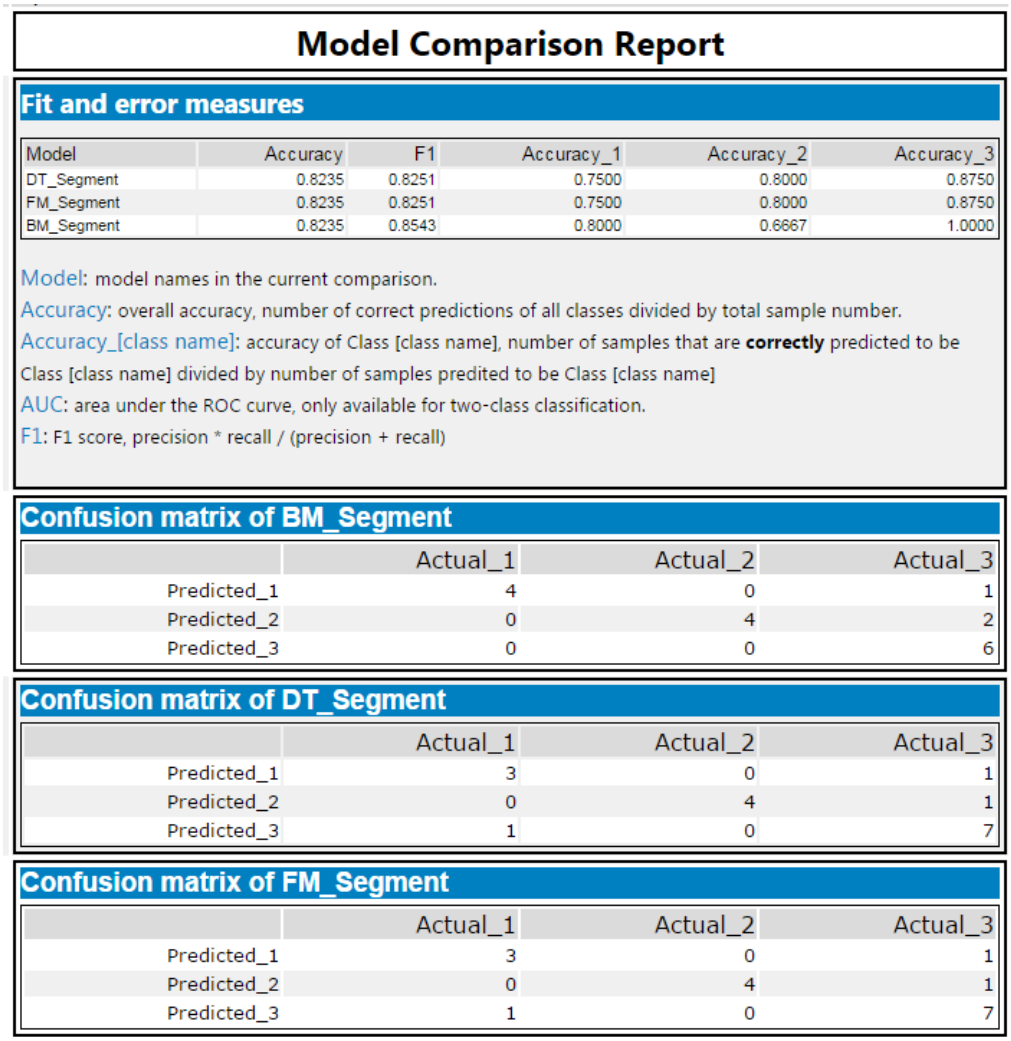


Figure 8. The Alteryx workflow that is used in assigning cluster for the new stores.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_Segment	0.8235	0.8251	0.7500	0.8000	0.8750
FM_Segment	0.8235	0.8251	0.7500	0.8000	0.8750
BM_Segment	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_Segment

	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_Segment

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

Confusion matrix of FM_Segment

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

Figure 9. Model comparison between decision tree, forest model, and boosted model

Based on the results obtained from the model comparison report, accuracy for the decision tree model, forest model, and boosted model is 0.8235. The F1_score for decision tree model, forest model, and boosted model are 0.8251, 0.8251, and 0.8543 respectively. Based on the accuracy and F1 score, the boosted model is chosen due to its higher value in F1 metric. The assignment for the new stores is as described in Table 1.

2. The most three important predictor variables are determined by the variance importance plot of the forest model. The three most important predictor variables are: Age0to9, HVal750K, and EdHSGrad.

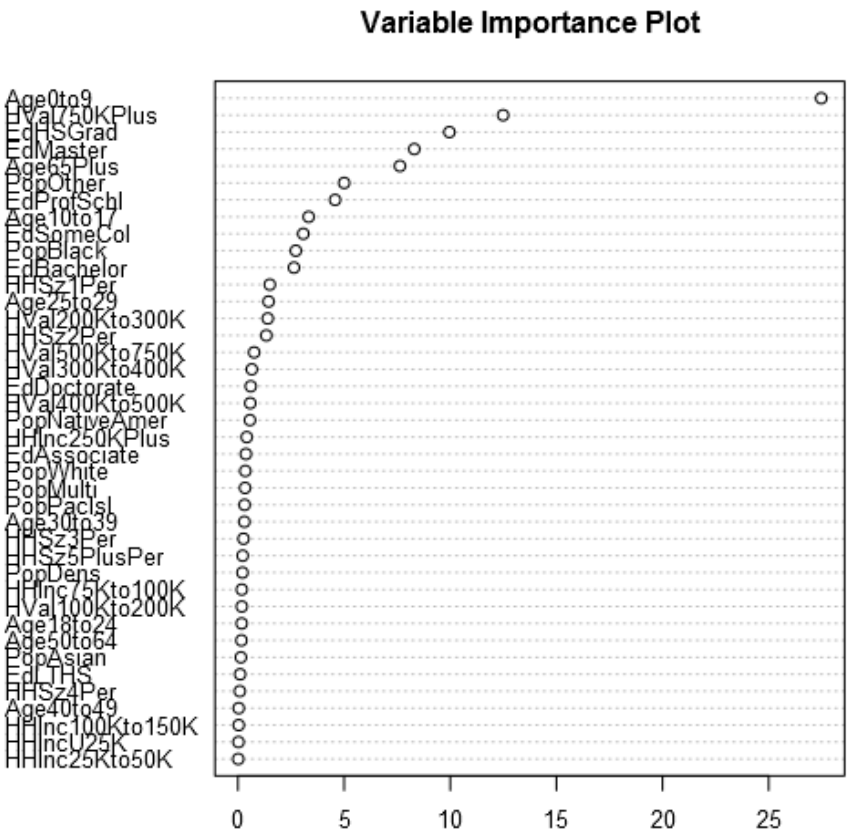


Figure 10. The variable importance plot obtained from running the forest model.

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

Table 1. Cluster assignment for the new stores

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?

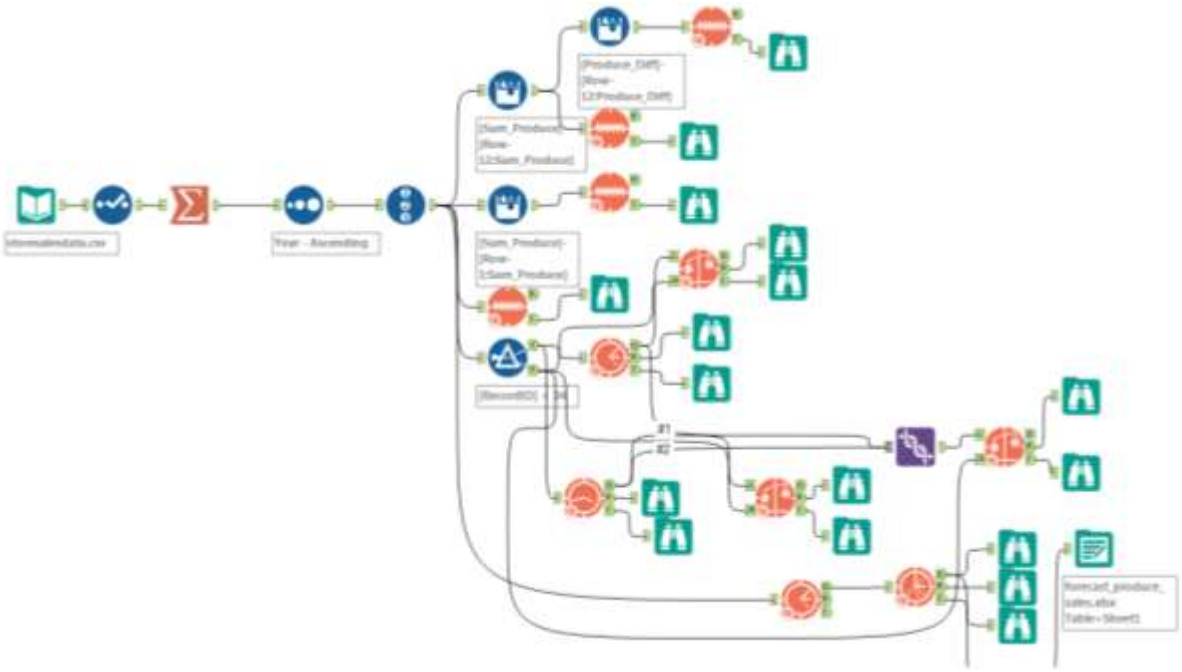


Figure 11. The Alteryx workflow that is used in forecasting the sale value for the average store in the year 2016.

Cluster 1:

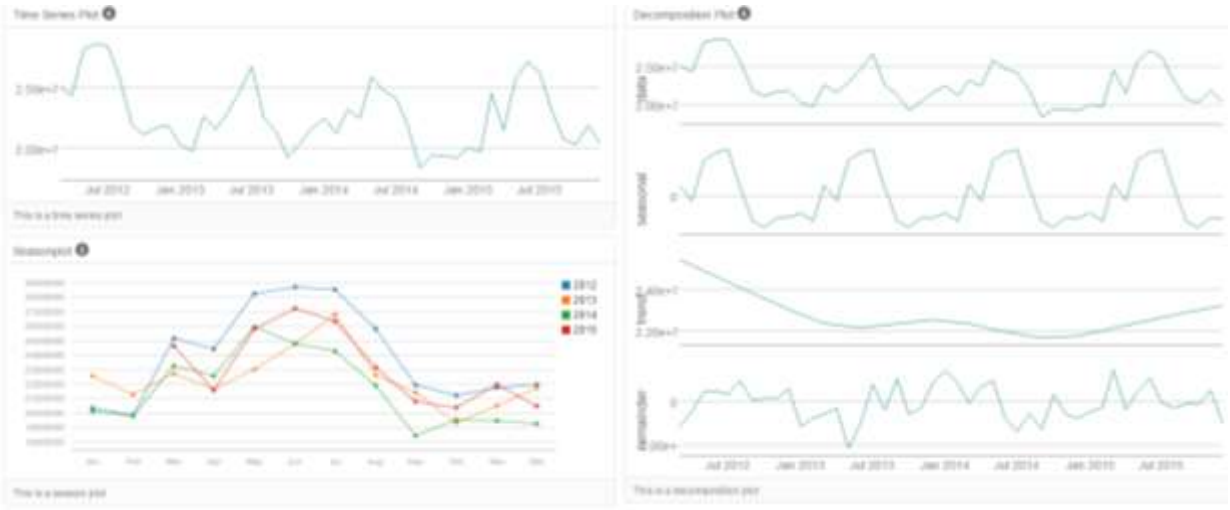


Figure 12. Time series plot, season plot, and the decomposition plot for dataset without differencing.

Based on Figure 12, the dataset of is not very stationary and therefore differencing will need to be performed.

ETS (M,N,M)

The features of the ETS model were figured out using the decomposition plot. According to the decomposition plot, the seasonal component does show to increase and therefore should be used multiplicatively. The trend component does not show to have any particular behavior and therefore neither multiplicative nor additive should be used. The error show variation along the 0 axis and thus multiplicatively should be applied. For the ETS model, the dampening effect was chosen by using the auto option. The final result shows that no-dampening seems to have a better accuracy in forecasting.

ARIMA(0,1,2)

For the ARIMA model, the set of (0,1,2)(0,1,0) was chosen. The parameters determined for the ARIMA are based on the ACF and PACF plots (Figure 13a,b,c).

For the non-seasonal part, it takes one time differencing in order to obtain a stationary series, therefore I=1. The ACF plot has a negative correlation at lag-1 which signals the use of MA term. In this case, MA=2 is used since there is a lag-2 (Figure 13a).
For the seasonal part, I=1 since there is one seasonal differencing. No particular signature for both AR and MA terms so AR=0 and MA=0.

Figure 13 a. The ACF plot and PACF plot of the non-seasonal component of the ARIMA with one differencing.

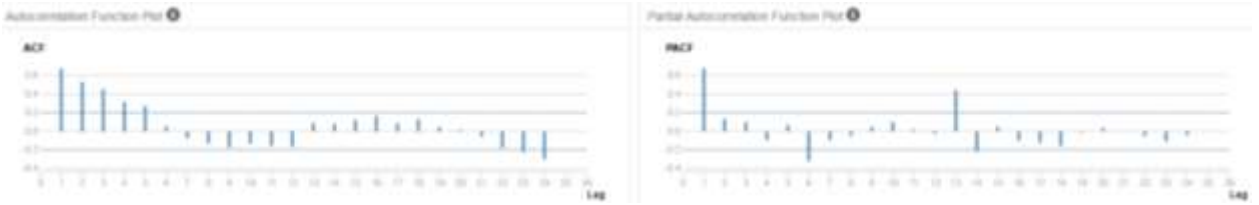


Figure 13 b. The ACF plot and PACF plot of the seasonal component of the ARIMA.

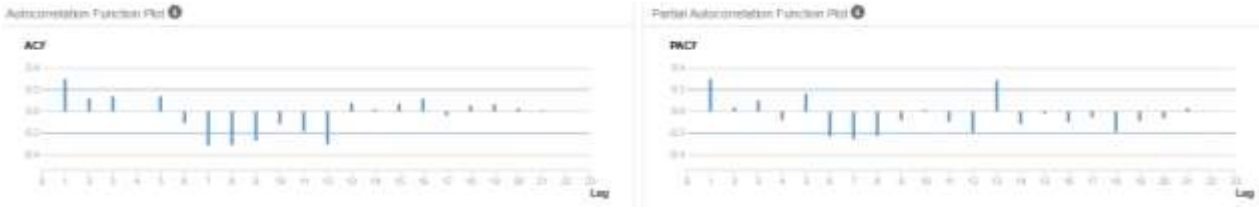


Figure 13 c. The ACF plot and PACF plot after taking the first differencing of the seasonal component of the ARIMA.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	1068766	1590916	1337409	4.372	5.7523	0.833	NA
ARIMA	1303043	2181554	1874870	5.3495	8.2524	1.1678	NA

Figure 14. Accuracy measures between the two time-series models

Based on the results obtained from running the two time-series models against the hold-out sample, the ETS (M,N,N) model has a lower RMSE value and a lower MASE value. Specifically, whereas the RMSE and MASE values for ETS model are 1068766 and 0.833 respectively, the ARIMA model has a RMSE value of 2181554 and the MASE value of 1.1678. Therefore**, the ETS model is used to forecast** the total sale values of the next 12 months for all the existing stores and the average sales of all segments as listed in Table 2-5. The hold-out sample has a 12 months of data since the forecast will be done for the next 12 months.

Table 2. Forecasted sale values in the next 12 months for all the existing stores

Period	Sub_Period	forecast
2016	1	21539936
2016	2	20413771
2016	3	24325953
2016	4	22993466
2016	5	26691951
2016	6	26989964
2016	7	26948631
2016	8	24091579
2016	9	20523492
2016	10	20011749
2016	11	21177435
2016	12	20855799

The monthly total produce sales of all the new stores are calculated by using the workflow in Figure 15 and Figure 16. Specifically, the average sale of each segment of each month is calculated by diving the forecasted sum produce sales of each segment by the number of existing stores in that segment. The monthly average sale of each segment is then multiplied by the number of new stores in that segment. The monthly total produce sale for all the new stores are then calculated by summing the total sales of all segments in each month.

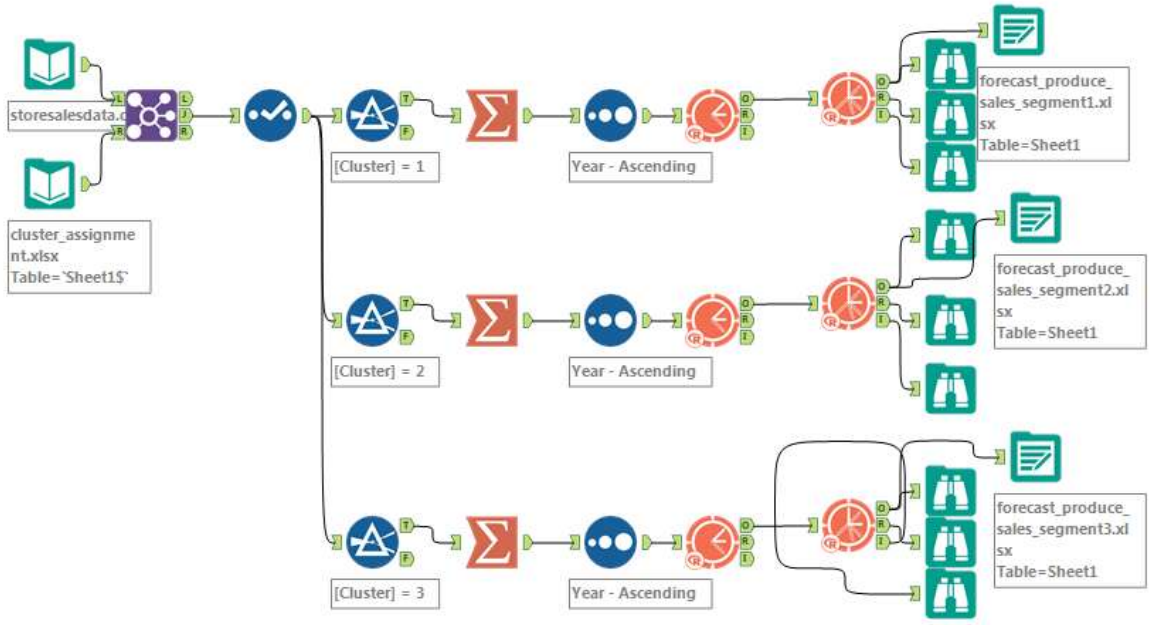


Figure 15. The Alteryx workflow that was used in forecasting the sum produce sale for each segment, which will be used later to find the total produce sale for each month of the new stores.

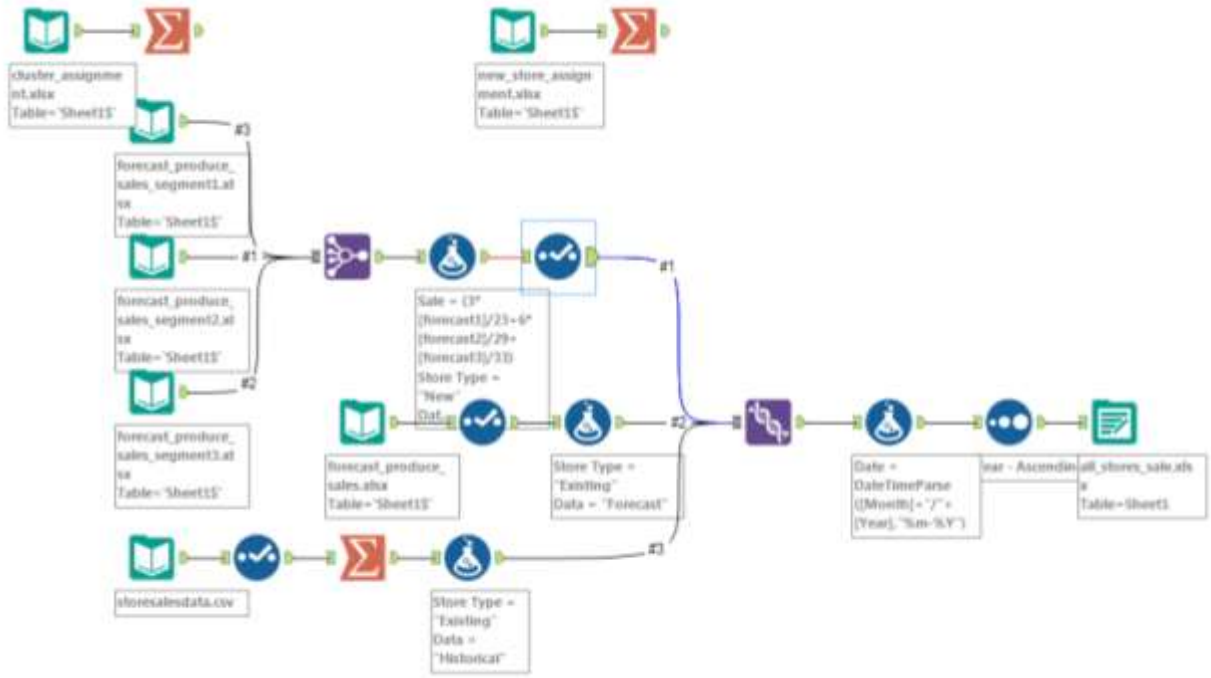


Figure 16. The Alteryx workflow that was used in calculating the monthly total produce sales of all the new stores. This workflow is also used to generate the dataset for Tableau.

Table 5. The forecasted sale in the next 12 months for both existing and new stores

Year	Month	Existing Stores Sale	New Stores Sale
2016	1	21,539,936	2,761,958
2016	2	20,413,771	2,656,665
2016	3	24,325,953	3,099,058
2016	4	22,993,466	2,873,607
2016	5	26,691,951	3,327,835
2016	6	26,989,964	3,356,062
2016	7	26,948,631	3,391,943
2016	8	24,091,579	2,991,383
2016	9	20,523,492	2,664,295
2016	10	20,011,749	2,588,210
2016	11	21,177,435	2,702,838
2016	12	20,855,799	2,761,943

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

https://public.tableau.com/profile/ivy.nguyen#!/vizhome/Project7_Task3_1/Sheet1?publish=yes