# 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**

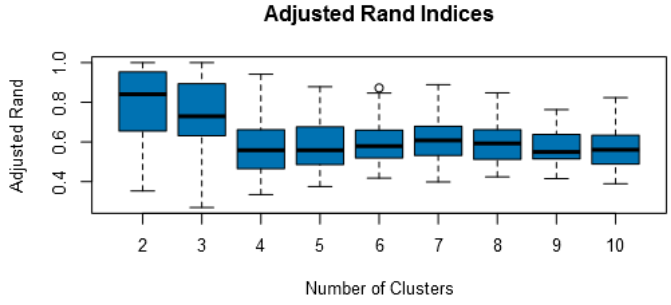
*Summary Statistics*

Adjusted Rand Indices:

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Minimum | 0.3527 | 0.2696 | 0.3342 | 0.3749 | 0.4177 | 0.3986 | 0.4239 |
| 1st Quartile | 0.6549 | 0.635 | 0.4689 | 0.4673 | 0.5197 | 0.5347 | 0.5136 |
| Median | 0.8403 | 0.7294 | 0.5583 | 0.5504 | 0.5709 | 0.6077 | 0.5922 |
| Mean | 0.7842 | 0.7308 | 0.576 | 0.5859 | 0.6004 | 0.6072 | 0.6007 |
| 3rd Quartile | 0.9529 | 0.8931 | 0.6574 | 0.672 | 0.6593 | 0.6782 | 0.6614 |
| Maximum | 1 | 1 | 0.9413 | 0.8784 | 0.8723 | 0.8888 | 0.8476 |
| | 9 | 10 | | | | | |
| Minimum | 0.415 | 0.3893 | | | | | |
| 1st Quartile | 0.5155 | 0.4901 | | | | | |
| Median | 0.5492 | 0.5607 | | | | | |
| Mean | 0.5744 | 0.5725 | | | | | |
| 3rd Quartile | 0.6371 | 0.6337 | | | | | |
| Maximum | 0.7633 | 0.8233 | | | | | |

Calinski-Harabasz Indices:

| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Minimum | 122.5 | 84.59 | 133.5 | 111 | 92.72 | 104.5 | 98.15 |
| 1st Quartile | 148.2 | 164.1 | 146.7 | 150.8 | 139.9 | 127.6 | 118.1 |
| Median | 154.4 | 173.3 | 161.6 | 160.7 | 148.6 | 135.4 | 123.6 |
| Mean | 151.9 | 167.8 | 158 | 157.1 | 145.2 | 133.7 | 123.5 |
| 3rd Quartile | 157.5 | 176.6 | 166.6 | 167.6 | 154.6 | 141.1 | 129.8 |
| Maximum | 158.7 | 182.7 | 173.6 | 174.9 | 164 | 151 | 147.5 |
| | 9 | 10 | | | | | |
| Minimum | 92.77 | 80.36 | | | | | |
| 1st Quartile | 109.3 | 101.7 | | | | | |
| Median | 116.1 | 107.3 | | | | | |
| Mean | 115.7 | 108.2 | | | | | |
| 3rd Quartile | 122.4 | 114.3 | | | | | |
| Maximum | 135.2 | 131.5 | | | | | |



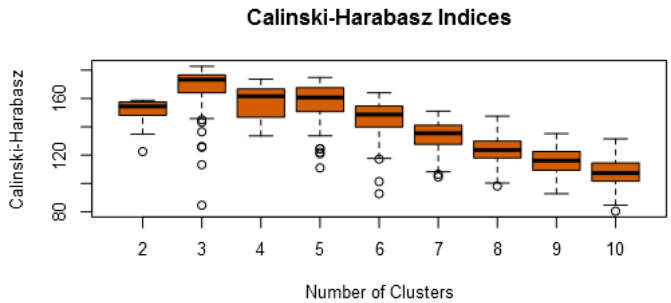




Figure 1. Adjusted rand indices and Calinski-Harabasz indices for the optimal number of clusters which should be used for the K-mean algorithm.

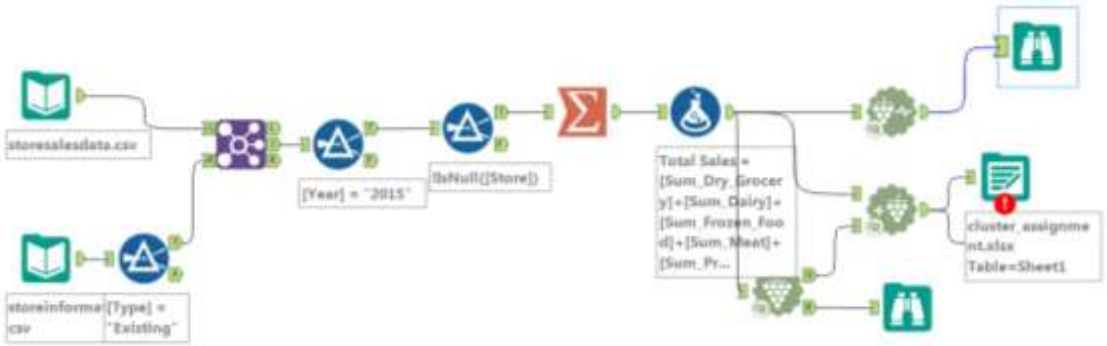



Figure 2. Evaluating the number of clusters that should be used in the K-mean clustering
Based on Figure 1, the following features can be pointed out:

- According to the adjusted rand indices, cluster 2 and 3 have a higher median value than the other clusters do, they also have a wider range of interquartile range (IQR).
- According to the C-H indices, cluster 3 has the highest median value with a narrow IQR range.

  By combining the features obtained from both the adjusted rand indices and the C-H indices, cluster 3 is chosen due to its median value from both indices and its narrow IQR from the C-H indices 'report.

2. How many stores fall into each store format?

Cluster Information:

| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 25 | 2.042736 | 4.648628 | 2.057697 |
| 2 | 32 | 1.235737 | 2.952974 | 1.785638 |
| 3 | 28 | 1.450403 | 2.511914 | 1.695148 |

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

There are 25 existing stores in cluster 1, 32 existing stores in cluster 2, and 28 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 the total sales and the sum of sales for each category are lowest for cluster 2 in comparison to the other clusters. The sales for cluster 3 is lower than that of cluster 1. Therefore, the clusters differ from each other based on the number of sales including total sales and the sale for each category. Specifically, cluster 1 has the highest number of sales followed by cluster 3 and then cluster 2.


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.

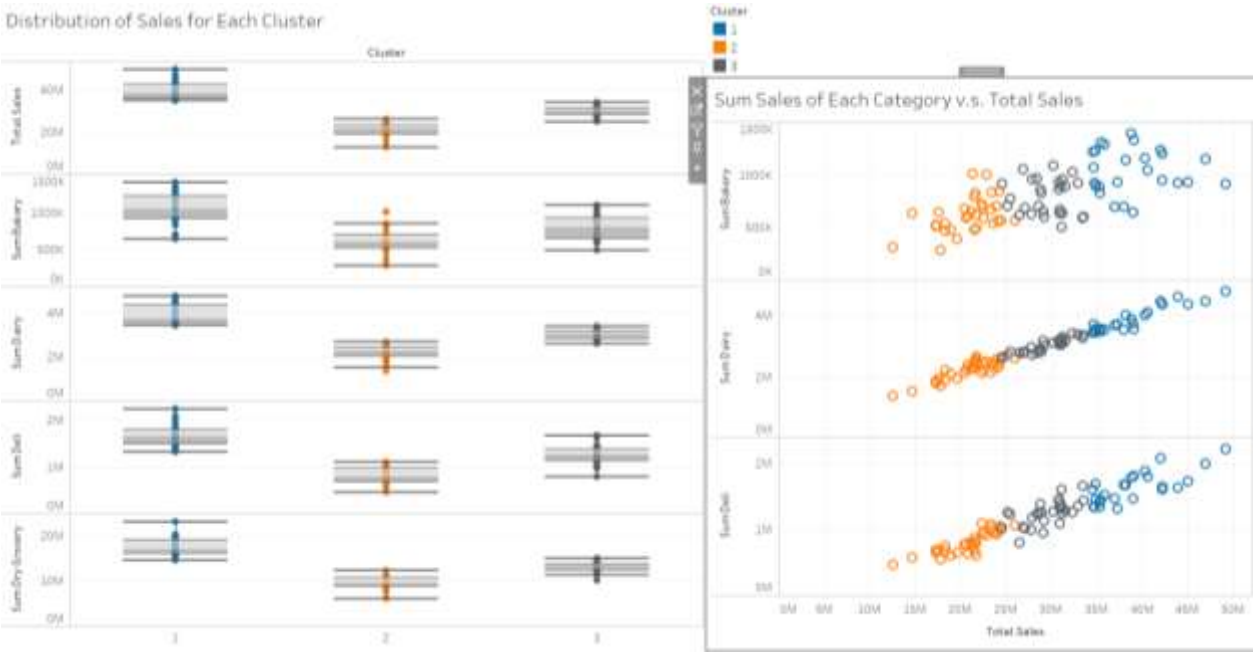



Figure 7. Left figure: whiskey-boxplot of the total sales, sum of bakery, sum of dairy, sum of deli, and sum of dry grocery for each cluster. Right figure: the relationship between total sales and sum sales for each category for each cluster. Blue: cluster 1, Orange: cluster 2, Gray: cluster 3).

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).

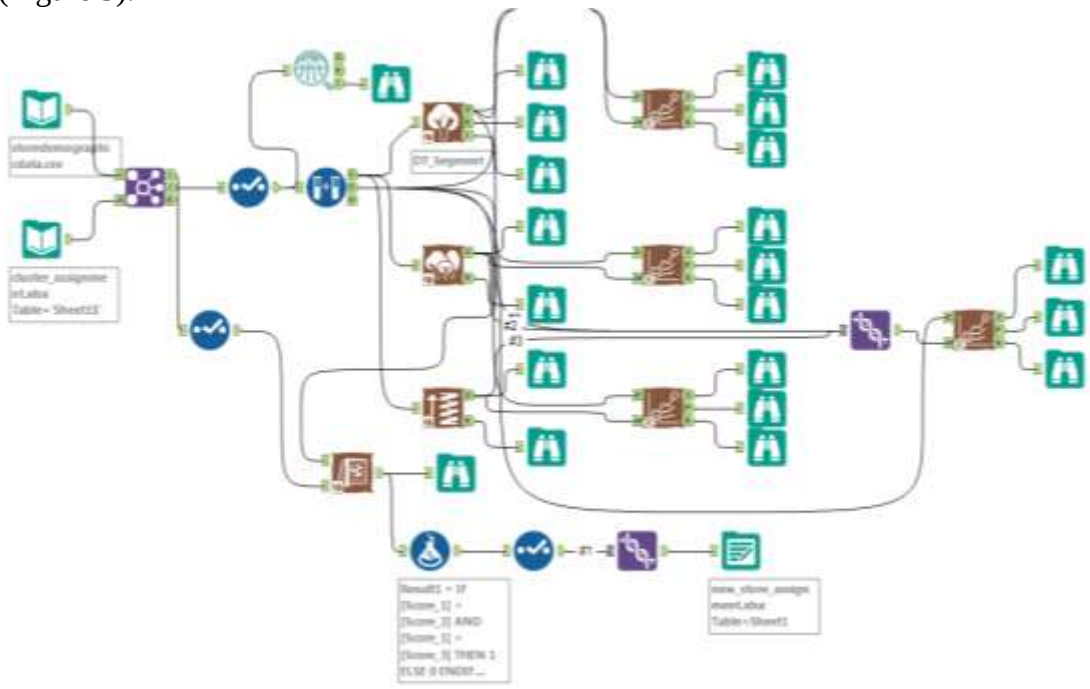


Figure 8. The Alteryx workflow that is used in assigning cluster for the new stores.

## Model Comparison Report

### Fit and error measures

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| DT_Segment | 0.2353 | 0.2388 | 0.1667 | 0.2500 | 0.3333 |
| FM_Segment | 0.3529 | 0.3712 | 0.2857 | 0.2000 | 0.6000 |
| BM_Segment | 0.2353 | 0.2488 | 0.2500 | 0.1667 | 0.3333 |

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 | 2 | 4 | 2 |
| Predicted_2 | 3 | 1 | 2 |
| Predicted_3 | 0 | 2 | 1 |

### Confusion matrix of DT_Segment

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 1 | 3 | 2 |
| Predicted_2 | 4 | 2 | 2 |
| Predicted_3 | 0 | 2 | 1 |

### Confusion matrix of FM_Segment

| | Actual_1 | Actual_2 | Actual_3 |
|---|---|---|---|
| Predicted_1 | 2 | 5 | 0 |
| Predicted_2 | 2 | 1 | 2 |
| Predicted_3 | 1 | 1 | 3 |

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

Based on the results obtained from the model comparison report, accuracies for the decision tree model, forest model, and boosted model are 0.2353, 0.3529, and 0.2353 respectively. The F1_score for decision tree model, forest model, and boosted model are 0.2388, 0.3712, and 0.2488 respectively. These values are not very high in predicting the cluster however they can help to determine which the best model to be used is. Based on the accuracy and F1 score, the forest model is chosen. 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: HVal200Kto300K, Age10to17, and HVal500Kto750K.

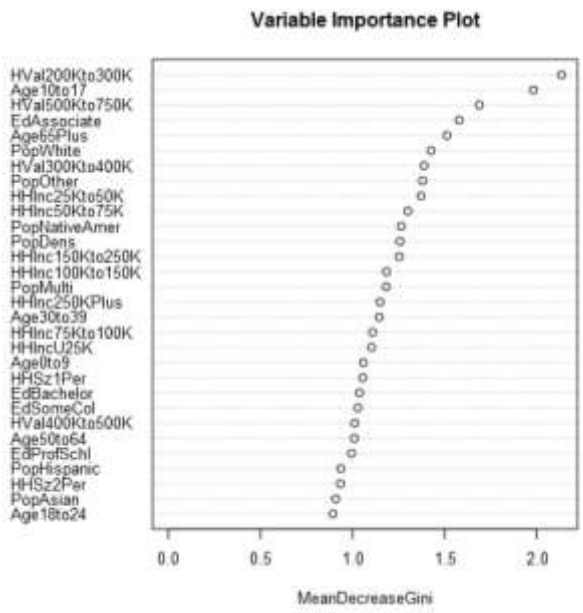



Figure 10. The variable importance plot obtained from running the forest model.

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 | 3 |
| S0089 | 2 |
| S0090 | 3 |
| 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?
   Since the forecasted values will need to be done for each cluster segment. The best time-series model is chosen by running the Alteryx workflow in Figure 11 with different dataset separately for each cluster. The data for each cluster is filtered out using the filter tool available in Alteryx.

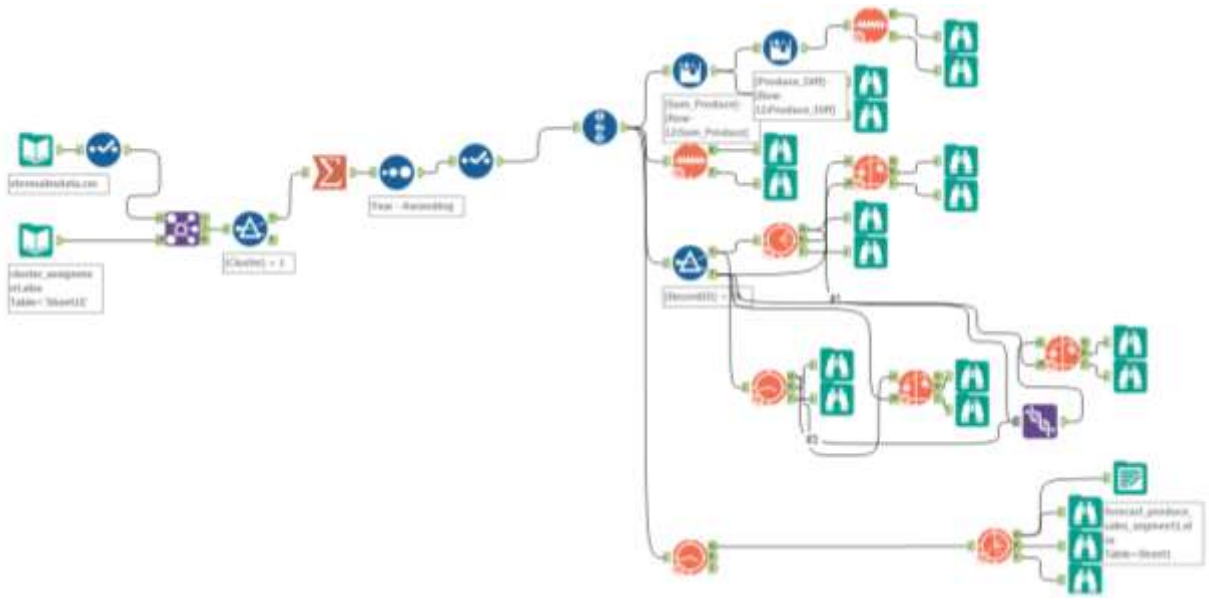


Figure 11. The Alteryx workflow that is used in forecasting the sale value for the average store in the year 2016.

***Cluster 1:***

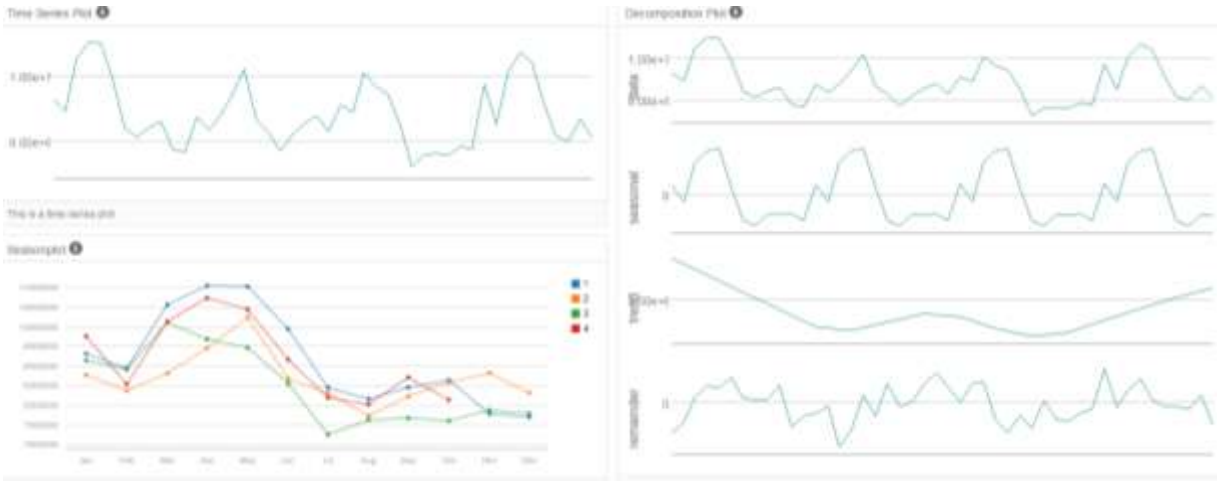


Figure 12. Time series plot, season plot, and the decomposition plot for cluster 1 dataset.

Based on Figure 12, the dataset of cluster 1 is not very stationary and therefore differencing will need to be performed. According to the decomposition plot, the seasonal component does not show to increase and therefore should be used addictively. 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. For the ARIMA model, the set of (1,0,0)(0,1,0) was

chosen. The parameters determined for the ARIMA are based on the ACF and PACF plots (Figure 13a,b,c).

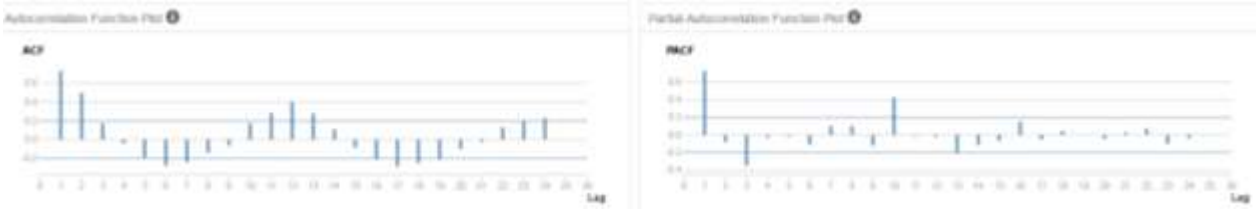


Figure 13 a. The ACF plot and PACF plot of the non-seasonal component of the ARIMA.

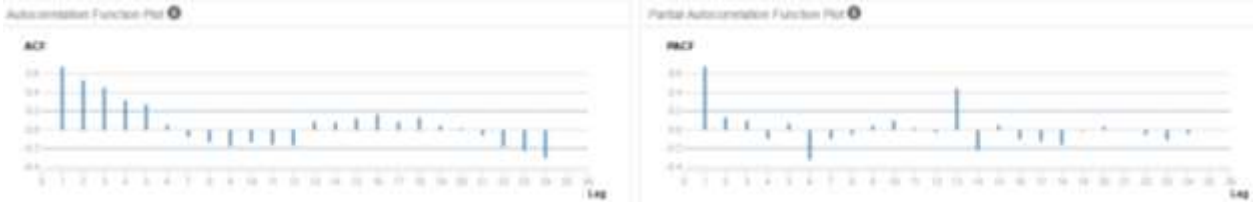


Figure 13 b. The ACF plot and PACF plot of the seasonal component of the ARIMA.

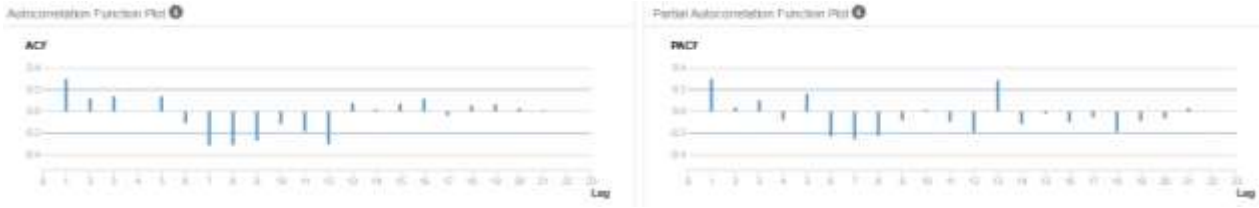


Figure 13 c. The ACF plot and PACF plot after taking the first differencing of the seasonal component of the ARIMA.

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE | NA |
|---|---|---|---|---|---|---|---|
| ETS | 597976.5 | 796209.3 | 646520.3 | 6.6718 | 7.3111 | 1.0458 | NA |
| ARIMA | 279421.3 | 698011.2 | 614550.9 | 2.6673 | 6.9263 | 0.9941 | 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 ARIMA model has a lower RMSE value and a lower MASE value. Specifically, whereas the RMSE and MASE values for ETS model are 796209.3 and 1.0458 respectively, the ARIMA model has a RMSE value of 698011.2 and the MASE value of 0.9941. Therefore, the ARIMA model is used to forecast the values of the next 12 months for cluster 1 as listed in Table 2.

Table 2. Forecasted sale values in the next 12 months for cluster 1

| Period | Sub_Period | forecast |
|---|---|---|
| 2016 | 1 | 8607906 |
| 2016 | 2 | 8221744 |
| 2016 | 3 | 9582768 |
| 2016 | 4 | 8902208 |
| 2016 | 5 | 10281367 |
| 2016 | 6 | 10315055 |
| 2016 | 7 | 10064658 |
| 2016 | 8 | 8985247 |
| 2016 | 9 | 7829629 |
| 2016 | 10 | 7957772 |
| 2016 | 11 | 8290248 |
| 2016 | 12 | 7998026 |

***Cluster 2 and 3:***

The same procedure was used to forecast the sale values in the next 12 months for both cluster 2 and cluster 3. The accuracy measures of both the ETS and the ARIMA models used in fitting the dataset of both clusters are listed in Figure 15. By using the dataset of cluster 2, the ETS model shows to have better accuracy measures against the hold-out sample (Figure 15a). In contrast, the ARIMA model appears to be better at forecasting the values against the hold-out sample using dataset of cluster 3 (Figure 15b).

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE | NA |
|---|---|---|---|---|---|---|---|
| ETS | 371741 | 523189.8 | 418263.9 | 6.0745 | 6.9597 | 0.9508 | NA |
| ARIMA | 297038.9 | 541452.2 | 469734.8 | 4.4675 | 7.7953 | 1.0678 | NA |

Figure 15a. Accuracy measures obtained from running both ETS and ARIMA model for cluster 2's dataset.

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE | NA |
|---|---|---|---|---|---|---|---|
| ETS | 210866.14 | 431577.9 | 348038.3 | 2.8135 | 4.661 | 0.6249 | NA |
| ARIMA | 74592.88 | 581704.9 | 538354.6 | 0.5444 | 7.2096 | 0.9666 | NA |

Figure 15b. Accuracy measures obtained from running both ETS and ARIMA model for cluster 2's dataset.

The forecasted values for cluster 2 and cluster 3 are listed in Table 3 and Table 4.

Table 3. Forecasted sale values for cluster 3 in the next 12 months.

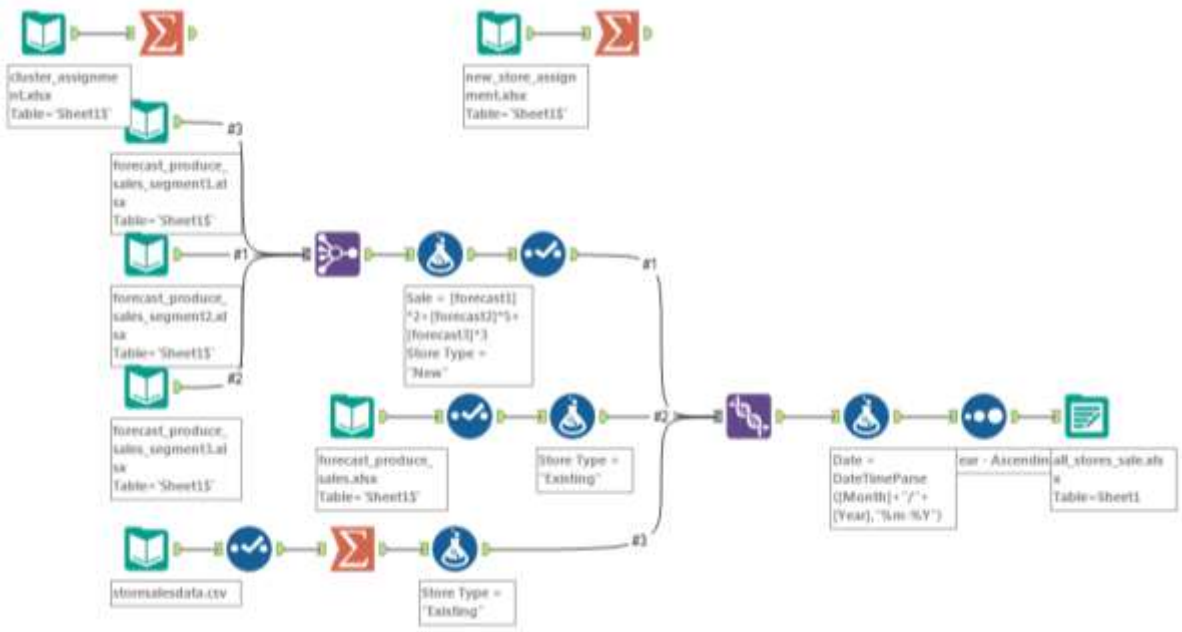
| Period | Sub_Period | forecast |
|---|---|---|
| 2016 | 1 | 5543356 |
| 2016 | 2 | 5338630 |

| | | |
|---|---|---|
| 2016 | 3 | 6174804 |
| 2016 | 4 | 5832207 |
| 2016 | 5 | 6750774 |
| 2016 | 6 | 6877591 |
| 2016 | 7 | 6973347 |
| 2016 | 8 | 6055028 |
| 2016 | 9 | 5382732 |
| 2016 | 10 | 5255635 |
| 2016 | 11 | 5341924 |
| 2016 | 12 | 5473541 |

Table 4. Forecasted sale values for cluster 3 in the next 12 months.

| Period | Sub_Period | forecast |
|---|---|---|
| 2016 | 1 | 7371488 |
| 2016 | 2 | 7093694 |
| 2016 | 3 | 8076891 |
| 2016 | 4 | 7538891 |
| 2016 | 5 | 8835881 |
| 2016 | 6 | 8702911 |
| 2016 | 7 | 8478784 |
| 2016 | 8 | 7606253 |
| 2016 | 9 | 6545782 |
| 2016 | 10 | 6748301 |
| 2016 | 11 | 6963044 |
| 2016 | 12 | 6752773 |

The total sales for each month for the existing stores are calculated by using the workflow in Figure 16. The forecasted values for the existing stores are predicted in a similar fashion but without any filtration in term of cluster. The produce sales for both existing and new stores are listed in Table 5.



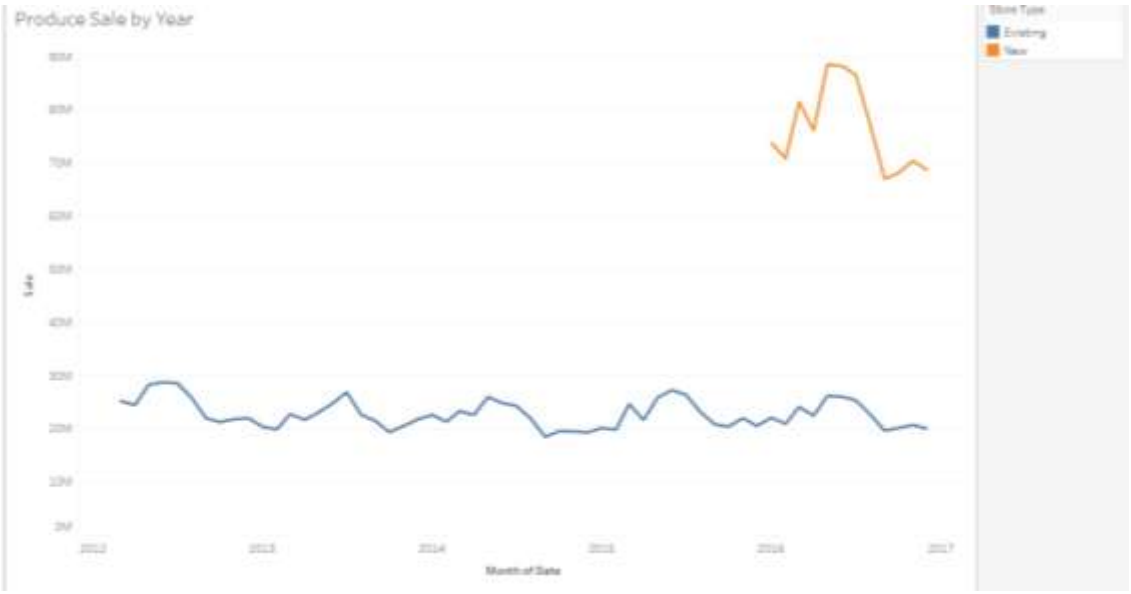
Figure

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 | 22016607 | 67047055.36 |
| 2016 | 2 | 20870322 | 64417720.55 |
| 2016 | 3 | 24045144 | 74270228.49 |
| 2016 | 4 | 22401007 | 69582125.99 |
| 2016 | 5 | 26134595 | 80824245.91 |
| 2016 | 6 | 25975297 | 81126797.22 |
| 2016 | 7 | 25319114 | 80432406.52 |
| 2016 | 8 | 22614678 | 71064395.16 |
| 2016 | 9 | 19576596 | 62210263.87 |
| 2016 | 10 | 20081244 | 62438623.86 |
| 2016 | 11 | 20662709 | 64179250.01 |
| 2016 | 12 | 19964213 | 63622078.39 |

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


Produce Sale by Year
Sale
Month of Date
2012
2013
2014
2015
2016
2017