Capstone Project: Combining Predictive Techniques

Business Problem #1: Store Format for Existing Stores
Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. You've been asked to provide analytical support to make decisions about store formats and inventory planning.

To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms "formats" and "segments" will be used interchangeably throughout this project. You've been asked to:

- Determine the optimal number of store formats based on sales data.
  - Sum sales data by StoreID and Year
  - Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
  - Use only 2015 sales data.
  - Use a K-means clustering model.
- Segment the 85 current stores into the different store formats.
- Use the StoreSalesData.csv and StoreInformation.csv files.

1. What is the optimal number of store formats? How did you arrive at that number?

A K-centroids analysis was conducted using K-means method to determine the number of clusters. According to our K-means assessment, Adjusted Rand Indices, and Calinski-Harabasz Indices, the optimal number of store formats is three as both the indices projected the highest median value at such and has smaller variation in its spread.

	2	3	4	5	6	7	8
Minimum	-0.016293	0.27351	0.335359	0.336327	0.318262	0.230196	0.27786
1st Quartile	0.352041	0.515917	0.445826	0.409773	0.366788	0.358895	0.377341
Median	0.526785	0.66768	0.538528	0.497192	0.423541	0.416509	0.428806
Mean	0.53781	0.664773	0.565975	0.50103	0.45115	0.432196	0.421514
3rd Quartile	0.734477	0.826692	0.644691	0.555087	0.499921	0.502931	0.458601
Maximum	1	1	0.975264	0.852076	0.8539	0.683894	0.647983

Figure 1: K-Means Cluster Assessment Report for Adjusted Rand Indices

	2	3	4	5	6	7	8
Minimum	16.61829	17.38103	20.28456	18.61989	17.8746	15.98702	16.16824
1st Quartile	28.17383	28.57484	25.20913	22.93454	21.30575	19.85155	18.71365
Median	29.46587	31.05384	26.53788	24.086	22.16245	20.97743	19.6662
Mean	28.45131	29.70664	26.41806	23.87003	22.02174	20.77195	19.65973
3rd Quartile	30.17907	32.08726	27.59305	25.10099	23.06602	21.72942	20.7099
Maximum	31.78345	33.63781	30.1583	26.63063	24.72038	24.63982	22.95166

Figure 2: K-Means Cluster Assessment Report for Calinski-Harabasz Indices

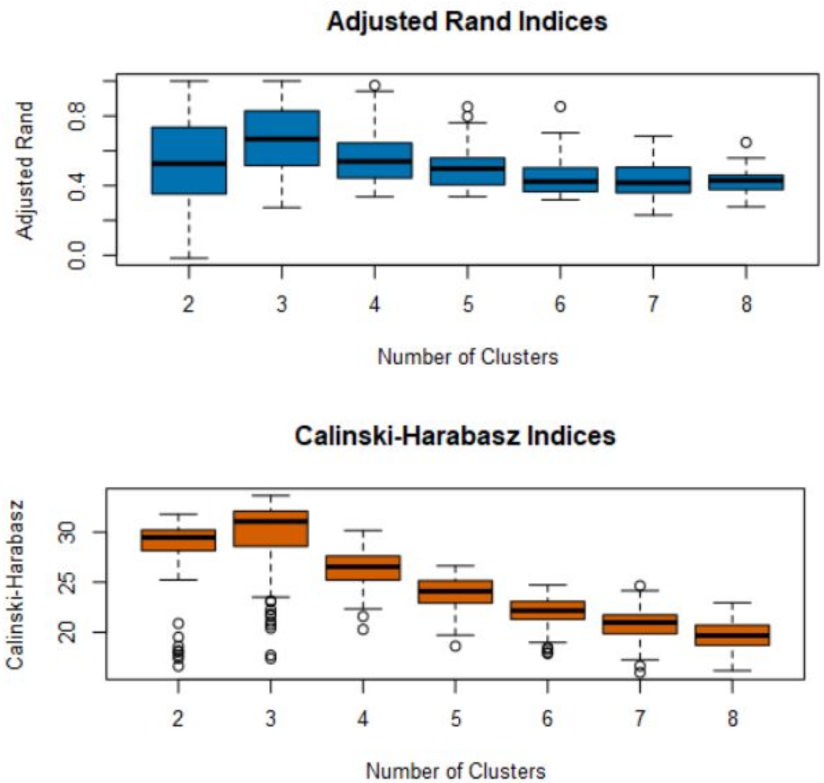


Figure 3: Plots for Adjusted Rand Indices and Calinski-Harabasz Indices

2. How many stores fall into each store format?

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

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 4: Cluster Information

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

Cluster category - Stores	**Differences from one another**
Cluster 1	Sold more General Merchandise & highest average total sales relative to cluster 1 and 2
Cluster 2	Sold more Produce, Floral
Cluster 3	Most similar in terms of sales due to more compact range

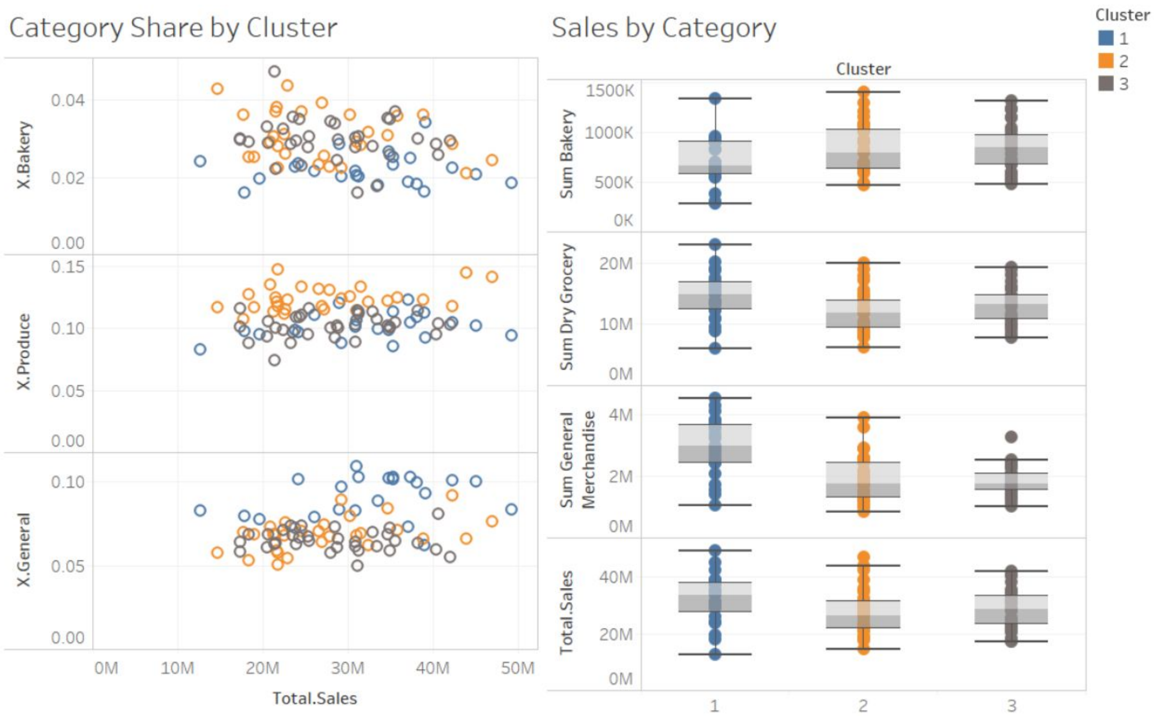


Figure 5: Plots for Cluster/Category

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_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

	Percent_Bakery	Percent_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Figure 6: Summary report of the K-means clustering

4. Please provide a map created in Tableau that shows the location of the existing stores, uses color to show cluster, and size to show total sales. Make sure to include a legend! Feel free to simply copy and paste the map into the submission template.

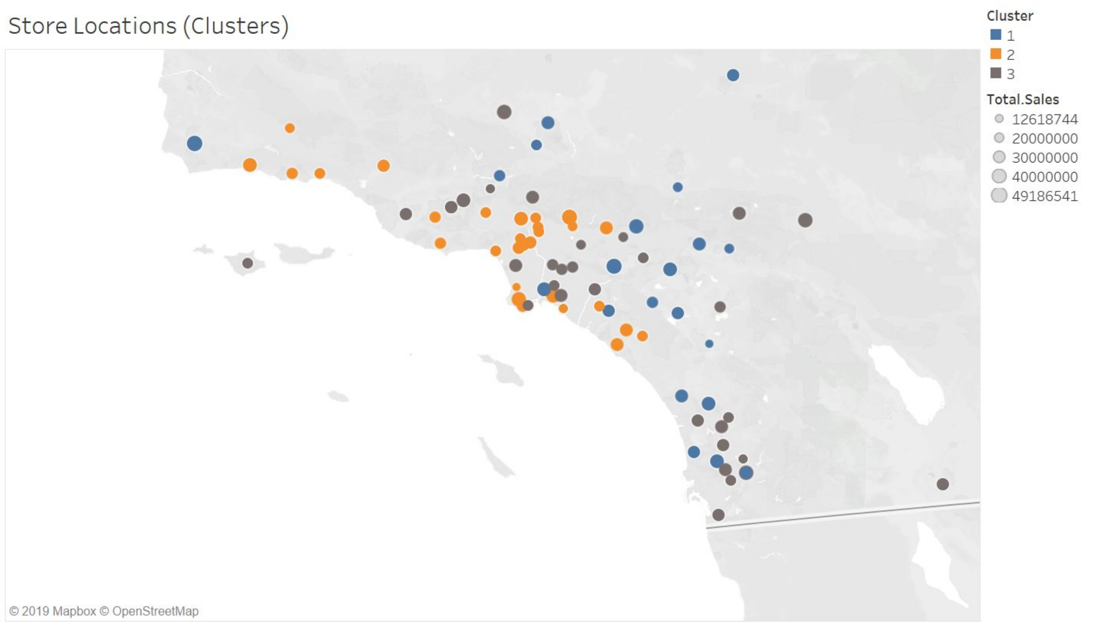


Figure 7: Store locations

Tableau Public:
https://public.tableau.com/profile/danny.lu6929#!/vizhome/StoreLocationbyClusterandSize/LocationofAllStores

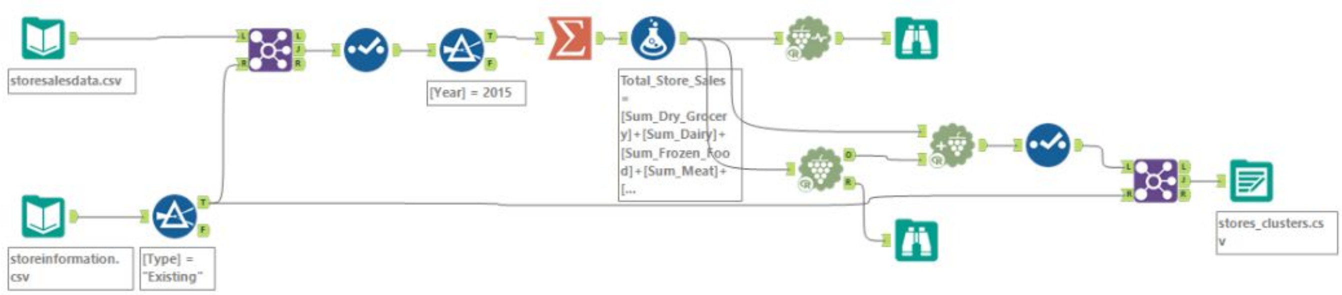


Figure 8: Alteryx workflow for calculating number of cluster on K-mean clustering model

Business Problem #2: Store Format for New Stores
The grocery store chain has 10 new stores opening up at the beginning of the year. The company wants to determine which store format each of the new stores should have. However, we don't have sales data for these new stores yet, so we'll have to determine the format using each of the new store's demographic data.

You've been asked to:

- Develop a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
- Use a 20% validation sample with Random Seed = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
- Use the model to predict the best store format for each of the 10 new stores.
- Use the StoreDemographicData.csv file, which contains the information for the area around each store.

1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology?

We do not run a logistic regression model because this a non-binary classification problem. A decision tree, forest, and boosted model were created to predict the store formats for the new stores. The boosted model and forest model exhibit the same accuracy at 82.35%, which is higher than that of the decision tree. **The boosted model is chosen** as the best classification model since its F1 value of 85.43% exceeds Forest model's F1 score of 82.51%.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8251	0.7500	0.8000	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

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 Boosted_Model

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

Confusion matrix of Decision_Tree

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of Forest_Model

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

Figure 1: Comparison Report and Confusion Matrix for three classification models

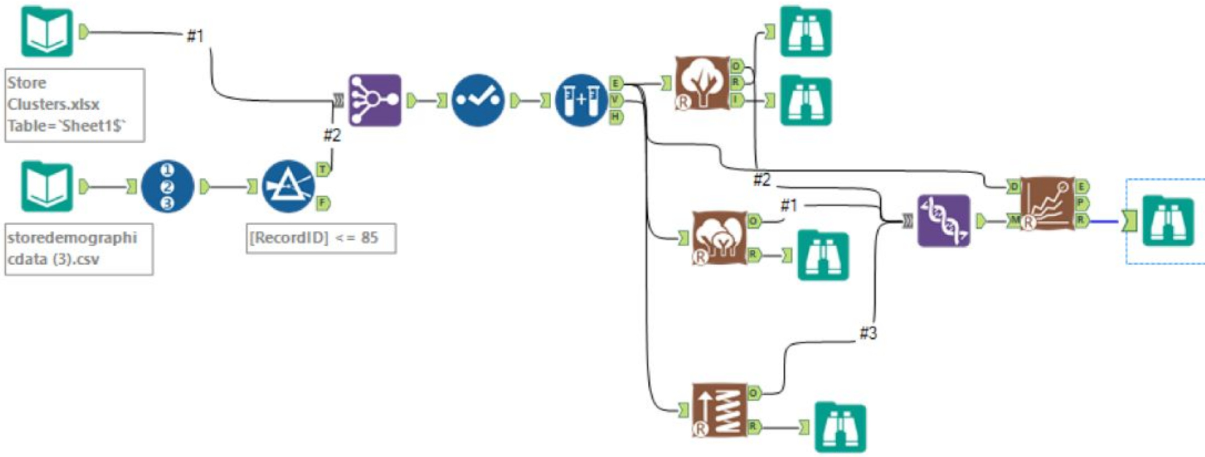


Figure 2: Alteryx Workflow for Model Comparison Report

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

Ave0to9, *HVal750KPlus* and *EdHSGrad* are the three most important variables.

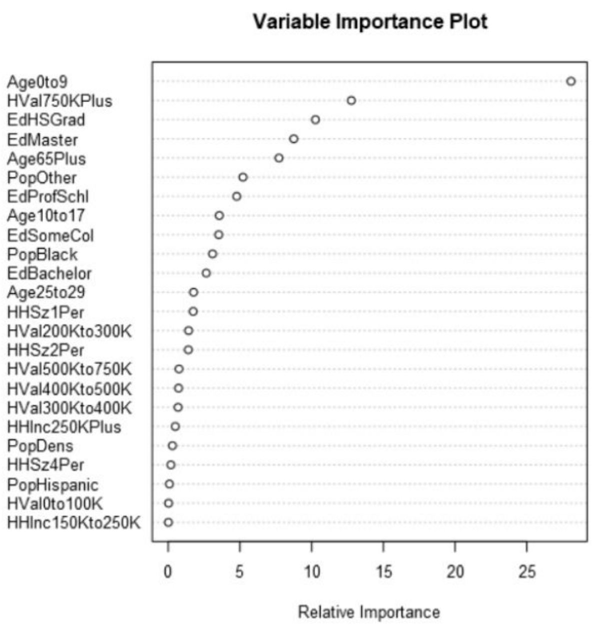


Figure 3: Variable Importance Plot for Boosted Model

3. What format do each of the 10 new stores fall into? Please provide a data table.

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

Figure 4: Store Number and Segment

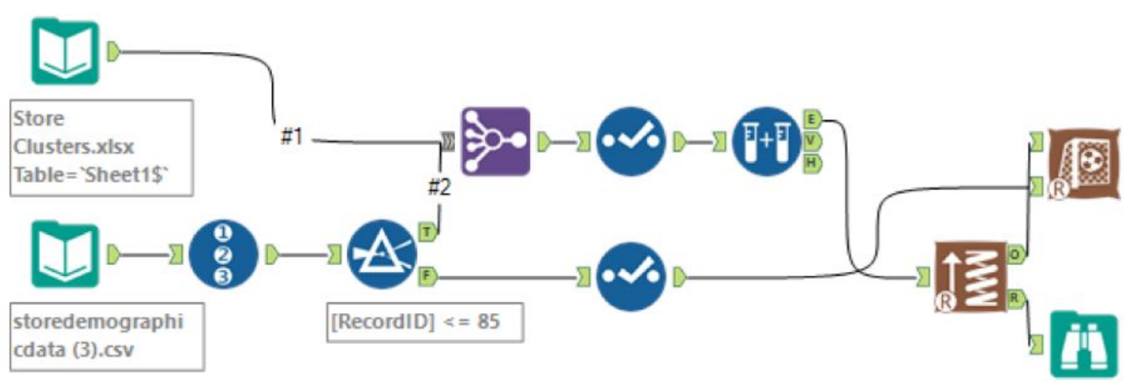


Figure 5: Alteryx Workflow with Scoring Tool on assigning cluster to new stores

Business Problem #3: Forecasting Produce Sales
Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast.

You've been asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To do so, follow the steps below.

Note: Use a 6 month holdout sample for the TS Compare tool (this is because we do not have that much data so using a 12 month holdout would remove too much of the data)

Step 1: To forecast produce sales for existing stores you should aggregate produce sales across all stores by month and create a forecast.

Step 2: To forecast produce sales for new stores:

- Forecast produce sales (not total sales) for the average store (rather than the aggregate) for each segment.
- Multiply the average store produce sales forecast by the number of new stores in that segment.
- For example, if the forecasted average store produce sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
- Sum the new stores produce sales forecasts for each of the segments to get the forecast for all new stores.

Step 3: Sum the forecasts of the existing and new stores together for the total produce sales forecast.

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?

The decision to use ETS or ARIMA model can be clarified with Time Series model with a holdout sample of 12 months.

Based on the decomposition plot below, our ETS(M,N,M) models shows the following:

(1) The seasonality has an increasing trend and multiplicative as the peaks change over time.
(2) The trend is zero as the trend seems inconsistent.
(3) The error is irregular and multiplicative since the errors are abruptly growing and shrinking over time.

ETS(M,N,M) with no dampening is used for ETS model.

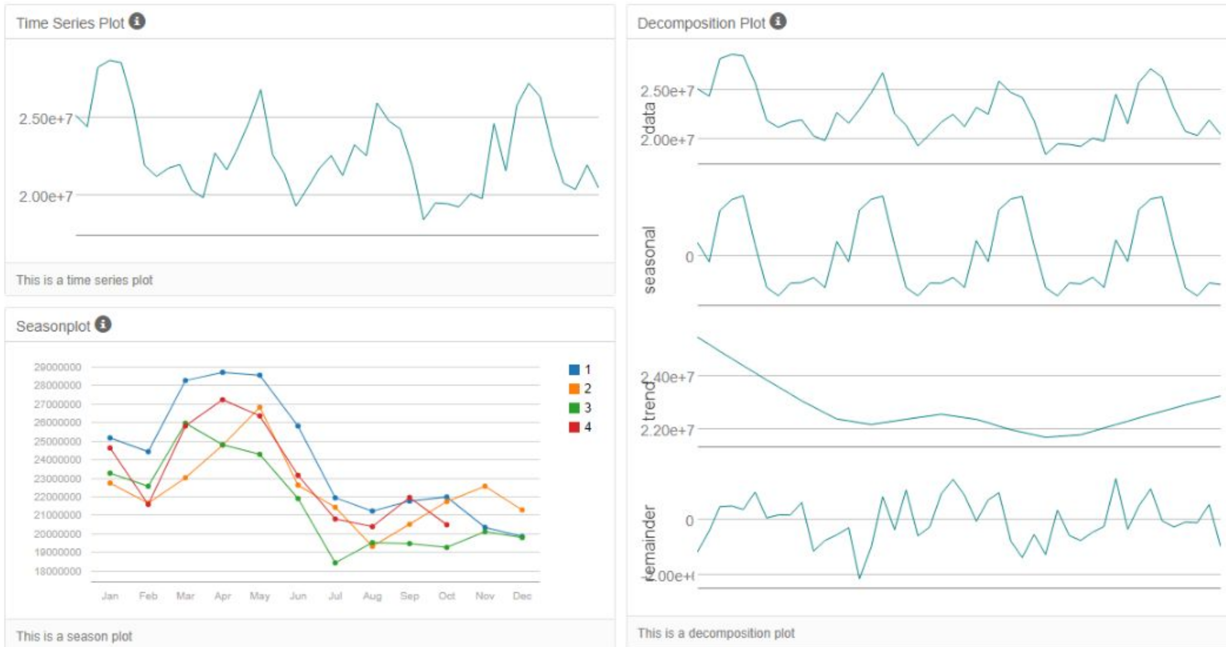
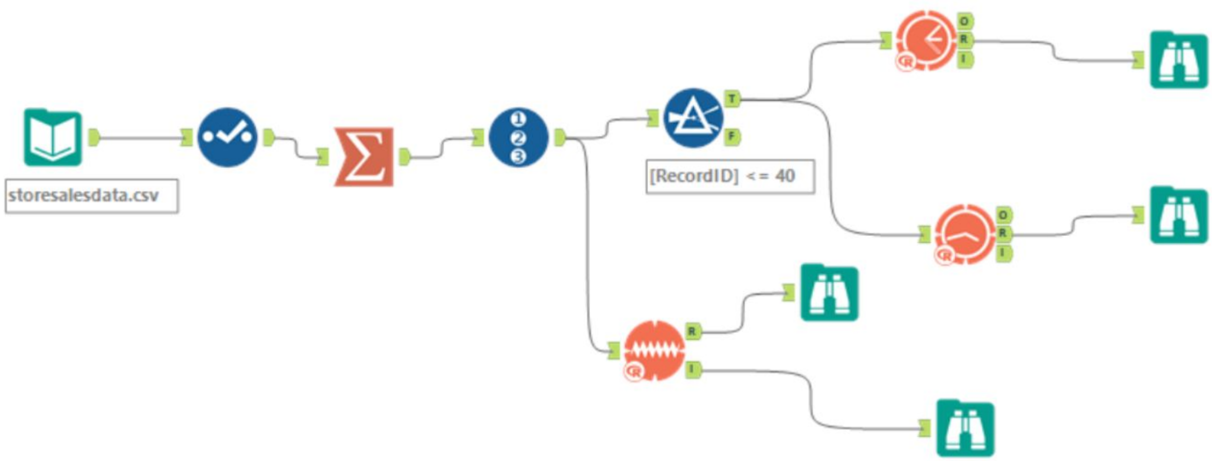
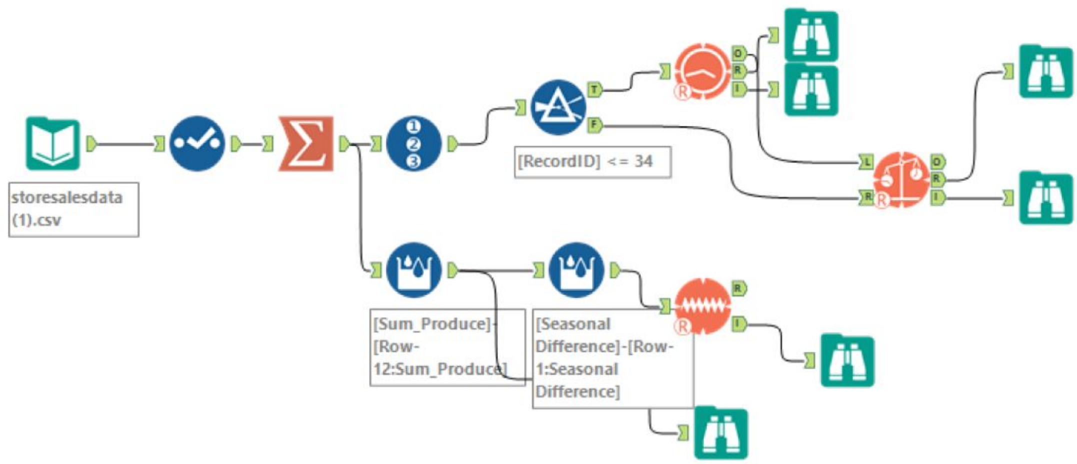


Figure 1: Time Series Plot w/ Decomposition Plot of historical monthly sales (no differencing)



The ARIMA model has a dataset that is seasonal on its series, we apply a seasonal difference in our time series in order to stationize the dataset.



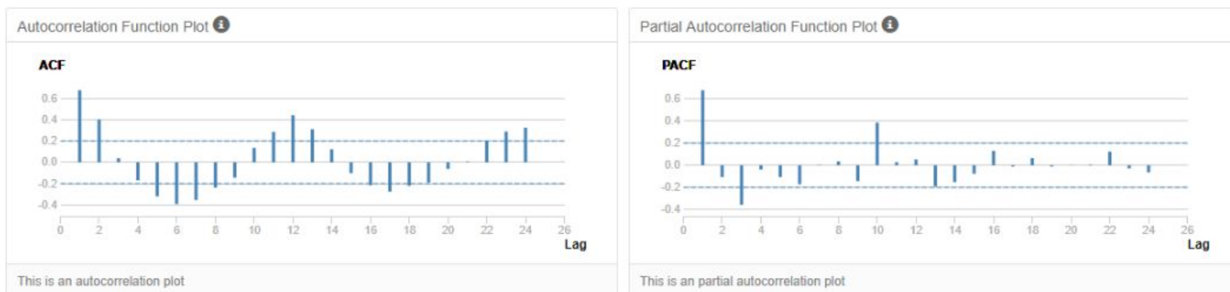


Figure 2: ACF and PACF plot (non-seasonal part of ARIMA model with one differencing)

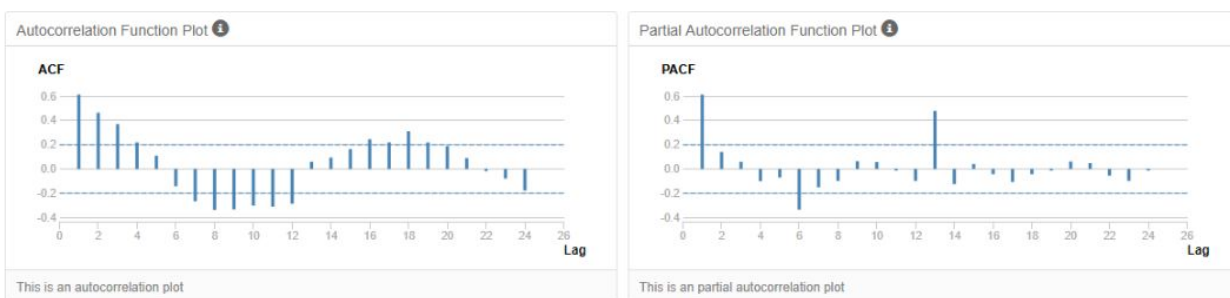


Figure 3: ACF and PACF plot (seasonal part of ARIMA model)

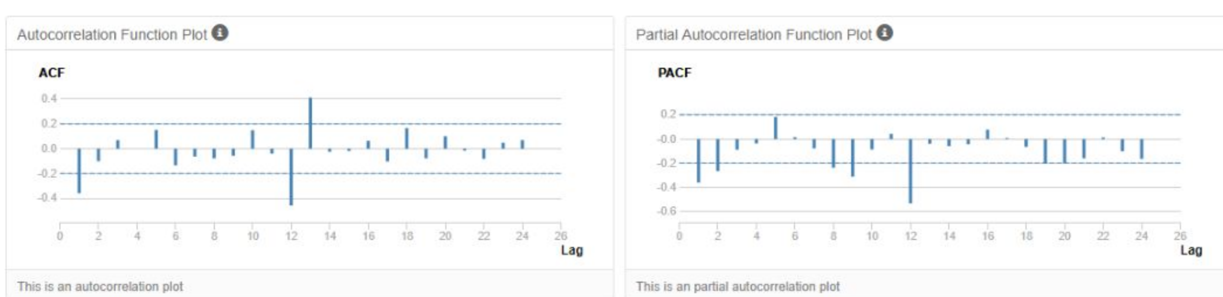


Figure 4: ACF and PACF plot (seasonal first difference)

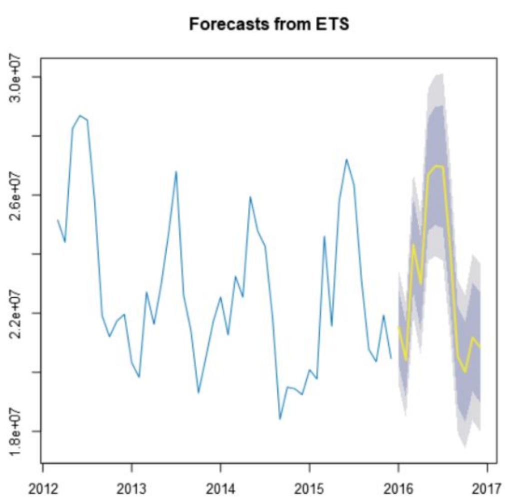
After plotting the first seasonal difference, the series is stationized. The plot above demonstrate that the serial correlation has disappeared.

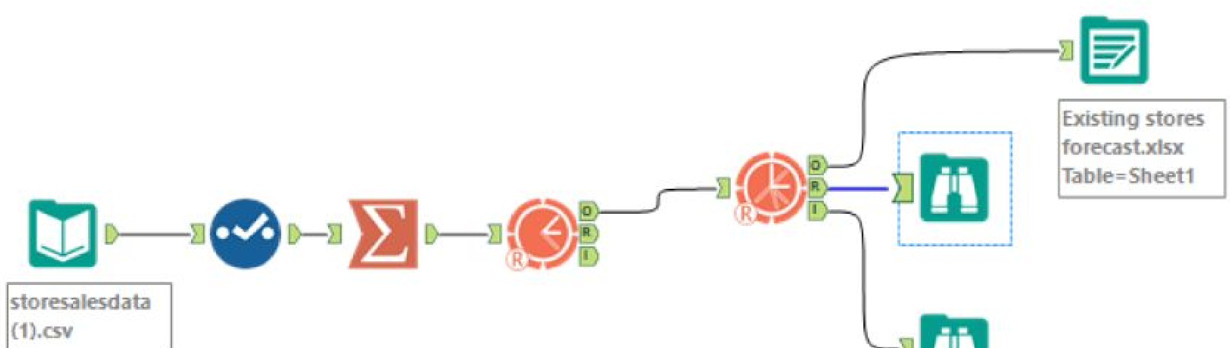
The ARIMA model(0,1,2)(0,1,0) is selected, seasonal differences and seasonal first difference were conducted. There is a lag-2. The parameters determined for the ARIMA are based on the ACF and PACF plots.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

Figure 5: accuracy differences in ETS and ARIMA models

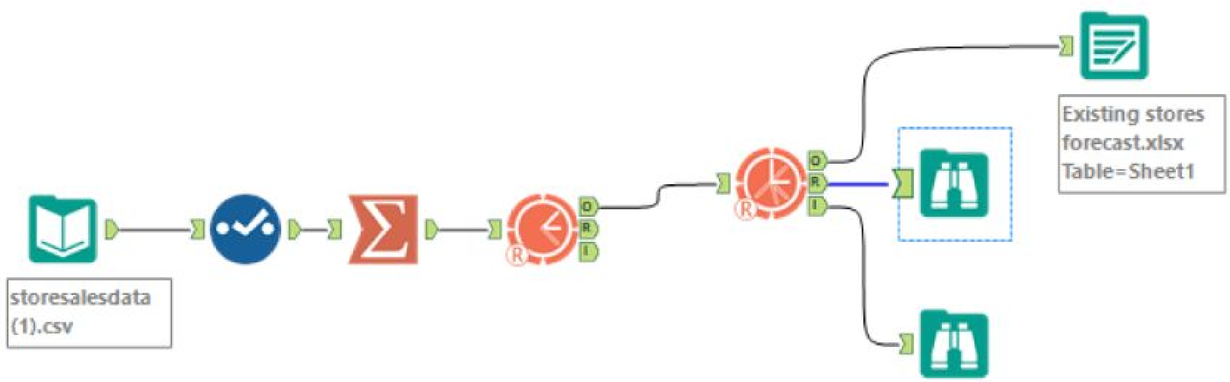
ETS's model accuracy is greater compared to ARIMA model based on running the two time-series models against the holdout sample of 6 months. The RMSE and MASE is also lower than that of the ARIMA.





Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21539936.007499	23479964.557336	22808452.492932	20271419.522066	19599907.457663
2016	2	20413770.60136	22357792.702597	21684898.329698	19142642.873021	18469748.500122
2016	3	24325953.097628	26761721.213559	25918616.262307	22733289.932948	21890184.981697
2016	4	22993466.348585	25403233.826166	24569128.609653	21417804.087517	20583698.871004
2016	5	26691951.419156	29608731.673669	28599131.515834	24784771.322478	23775171.164643
2016	6	26989964.010552	30055322.497686	28994294.191682	24985633.829422	23924605.523418
2016	7	26948630.764764	30120930.290185	29022885.932332	24874375.597196	23776331.239343
2016	8	24091579.349106	27023985.64738	26008976.766614	22174181.931598	21159173.050832
2016	9	20523492.408643	23101144.398226	22208928.451722	18838056.365564	17945840.419059
2016	10	20011748.6686	22600389.955254	21704370.226808	18319127.110391	17423107.381946
2016	11	21177435.485839	23994279.191514	23019270.585553	19335600.386124	18360591.780163
2016	12	20855799.10961	23704077.778174	22718188.42676	18993409.79246	18007520.441046

Figure 6 : Forecasts from ETS Model - Graph and table with actual and forecast value with 80% and 95% confidence level interval



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.

The chart and Tableau visualization shows the forecast sales for new stores and existing stores. New Store Sales is calculated by using the ETS(M,N,M) analysis with all three individual cluster to obtain the average sales per store. The average sales value (x3 cluster 1, x6 cluster 2, x1 cluster 3) are added up to produce New Store Sales.

Month	**New Stores**	**Existing Stores**
Jan 2016	2,587,451	21,539,936
Feb 2016	2,477,353	20,413,771
Mar 2016	2,913,185	24,325,953
April 2016	2,775,746	22,993,466
May 2016	3,150,867	26,691,951
June 2016	3,188,922	26,989,964
July 2016	3,214,746	26,948,631
August 2016	2,866,349	24,091,579
September 2016	2,538,727	20,523,492
October 2016	2,488,148	20,011,749
November 2016	2,595,270	21,177,435
December 2016	2,573,397	20,855,799
Total annual sales	$33,370,160	$276,563,727

Figure 7: Sales for Existing and New Stores for the next 12 months

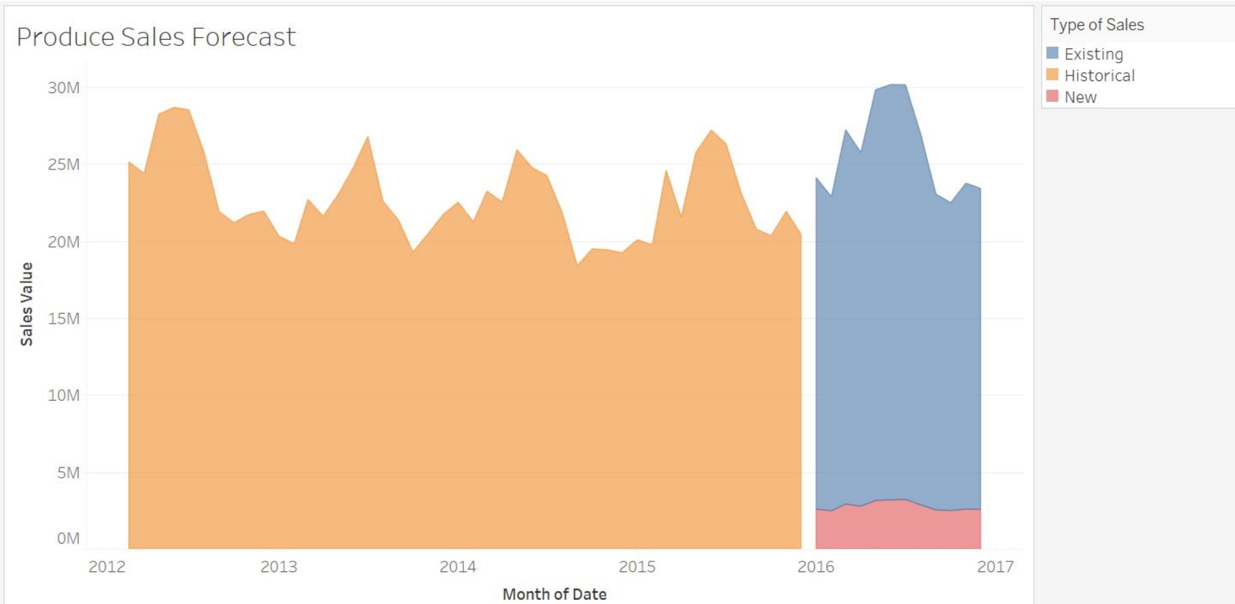


Figure 8: Historical and forecast sales for existing stores and new stores from Mar-12 to Dec-16

Tableau Public:
https://public.tableau.com/profile/danny.lu6929#!/vizhome/producesalesforecast/Sheet1?publish=yes