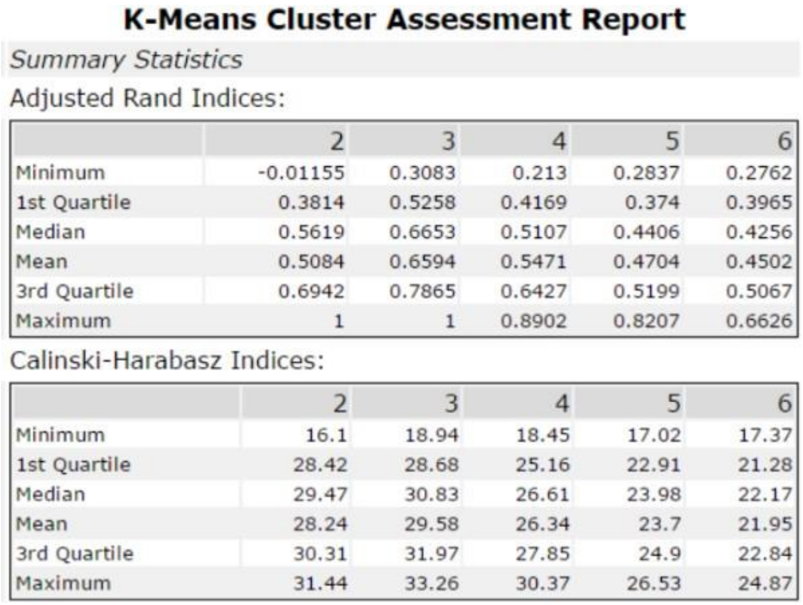
# Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

## Task 1: Determine Store Formats for Existing Stores

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

Ans – The appropriate number of clusters is 3 . This is based on the following information –



**K-Means Cluster Assessment Report**

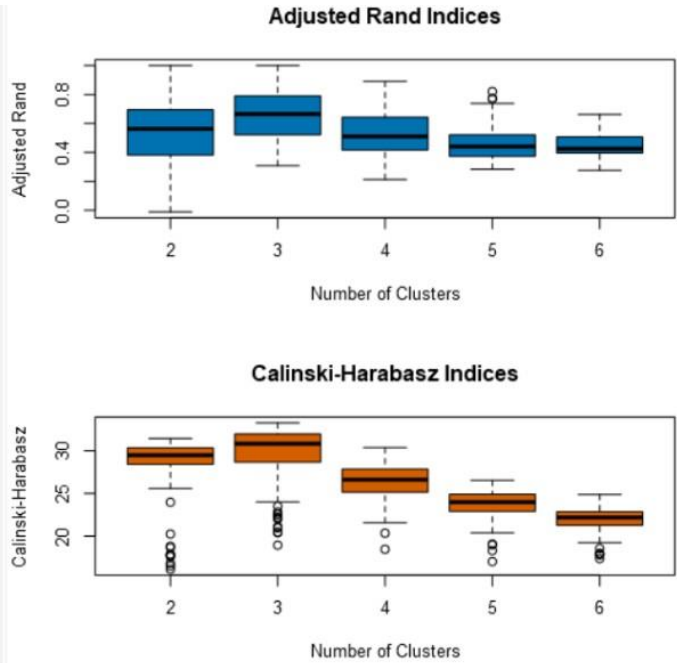
*Summary Statistics*

Adjusted Rand Indices:

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | -0.01155 | 0.3083 | 0.213 | 0.2837 | 0.2762 |
| 1st Quartile | 0.3814 | 0.5258 | 0.4169 | 0.374 | 0.3965 |
| Median | 0.5619 | 0.6653 | 0.5107 | 0.4406 | 0.4256 |
| Mean | 0.5084 | 0.6594 | 0.5471 | 0.4704 | 0.4502 |
| 3rd Quartile | 0.6942 | 0.7865 | 0.6427 | 0.5199 | 0.5067 |
| Maximum | 1 | 1 | 0.8902 | 0.8207 | 0.6626 |

Calinski-Harabasz Indices:

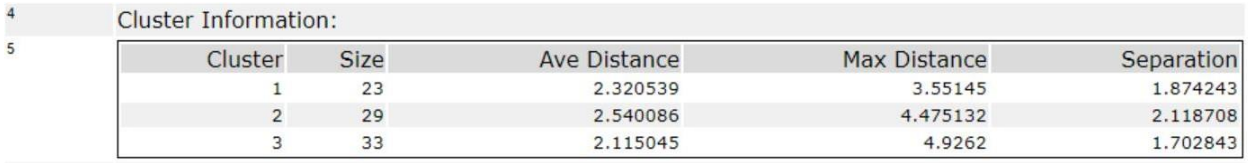
| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | 16.1 | 18.94 | 18.45 | 17.02 | 17.37 |
| 1st Quartile | 28.42 | 28.68 | 25.16 | 22.91 | 21.28 |
| Median | 29.47 | 30.83 | 26.61 | 23.98 | 22.17 |
| Mean | 28.24 | 29.58 | 26.34 | 23.7 | 21.95 |
| 3rd Quartile | 30.31 | 31.97 | 27.85 | 24.9 | 22.84 |
| Maximum | 31.44 | 33.26 | 30.37 | 26.53 | 24.87 |



As we can see here the median values of adjusted rand index and CH index are maximum for number 3. Additionally the interquartile spread is fairly compact

2. How many stores fall into each store format?

ANS –



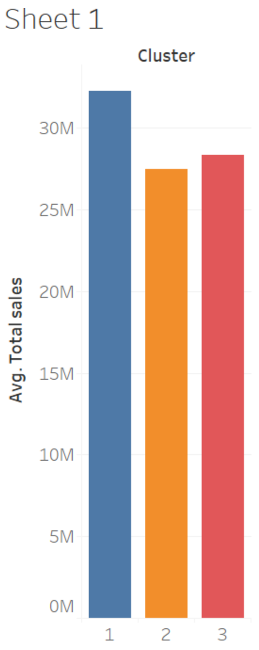
Cluster Information:

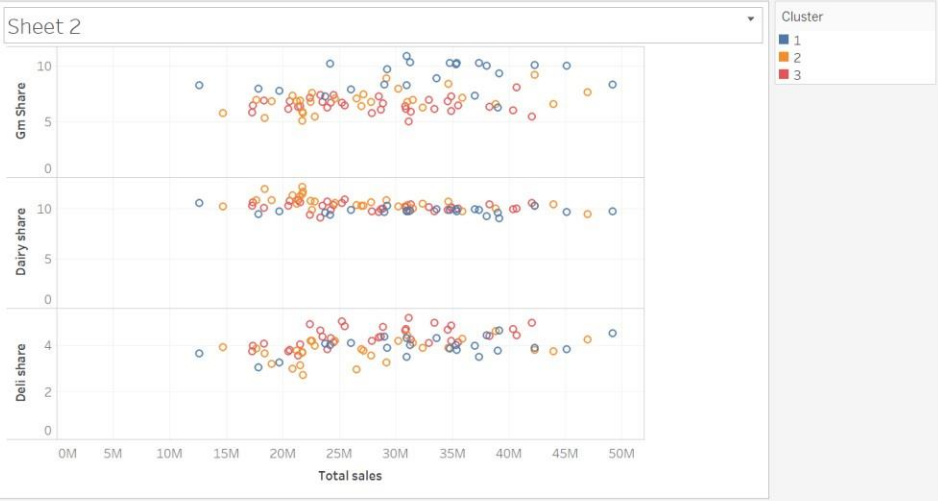
| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 23 | 2.320539 | 3.55145 | 1.874243 |
| 2 | 29 | 2.540086 | 4.475132 | 2.118708 |
| 3 | 33 | 2.115045 | 4.9262 | 1.702843 |

The first cluster has 23 stores, the second cluster has 29 stores and the third cluster has 33 stores.

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

ANS – Stores in Cluster 1 on an average have more total sales than stores in cluster 2 or cluster 3.

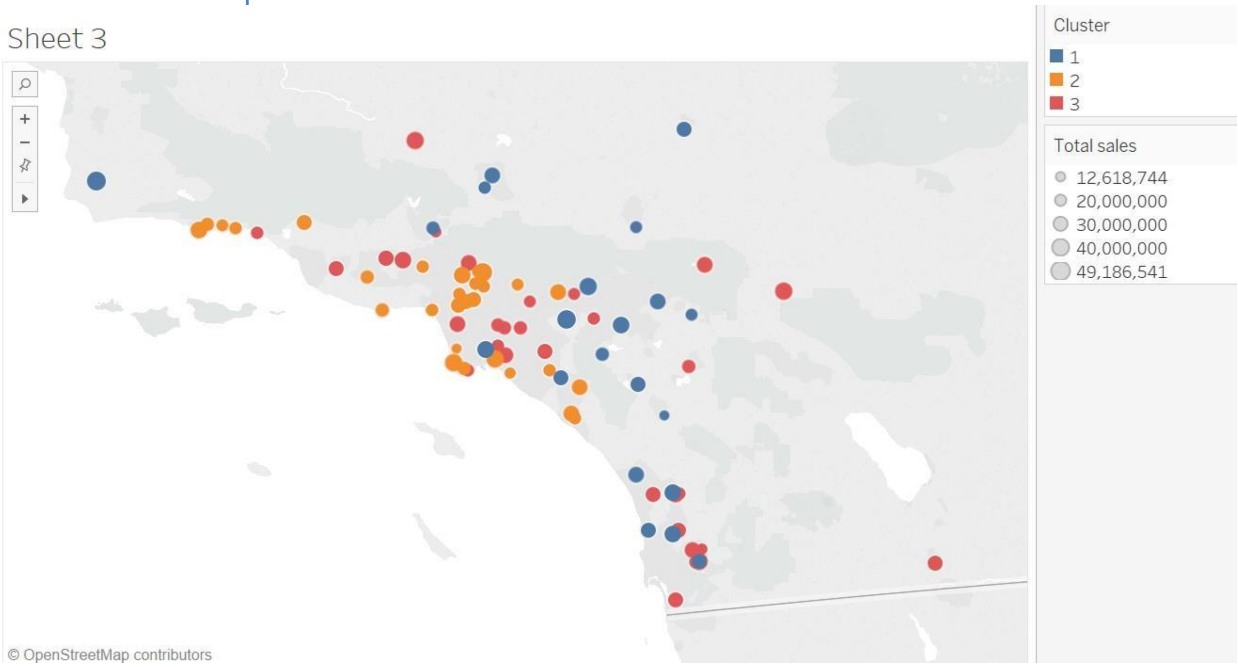




Stores of cluster 1 have a greater % share in selling general merchandise whereas cluster 2 stores have a greater percentage share in selling dairy products. Cluser 3 stores have a greater percentage share in selling deli products.

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.

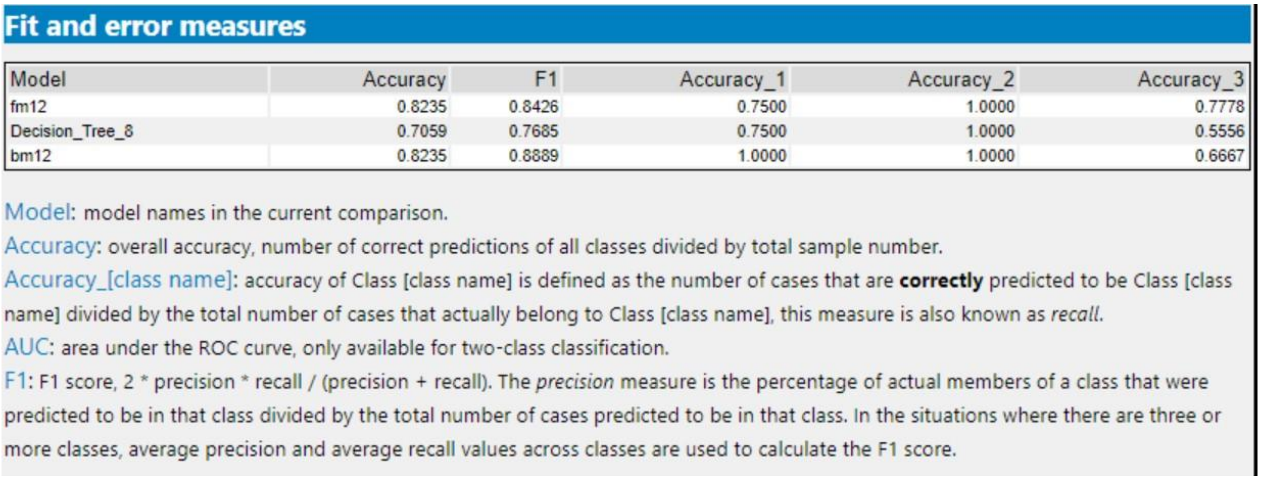
ANS – PFB the required visualization



## 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.)

ANS - I plan to use the boosted model because it has the highest accuracy . Although Forest model and boosted model have the same accuracy boosted model is chosen for higher f1 value



Fit and error measures

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| fm12 | 0.8235 | 0.8426 | 0.7500 | 1.0000 | 0.7778 |
| Decision_Tree_8 | 0.7059 | 0.7685 | 0.7500 | 1.0000 | 0.5556 |
| bm12 | 0.8235 | 0.8889 | 1.0000 | 1.0000 | 0.6667 |

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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

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

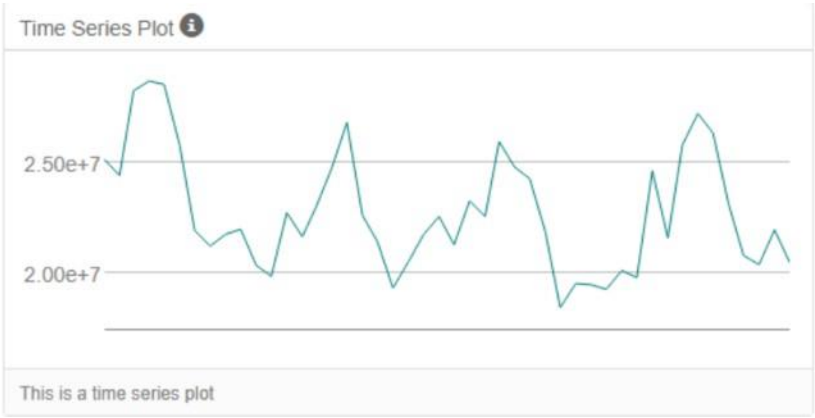
| Store Number | Segment |
|---|---|
| S0086 | 3 |
| S0087 | 2 |
| S0088 | 1 |
| S0089 | 2 |
| S0090 | 2 |
| S0091 | 1 |
| S0092 | 2 |
| S0093 | 1 |
| S0094 | 2 |
| S0095 | 2 |

## 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?

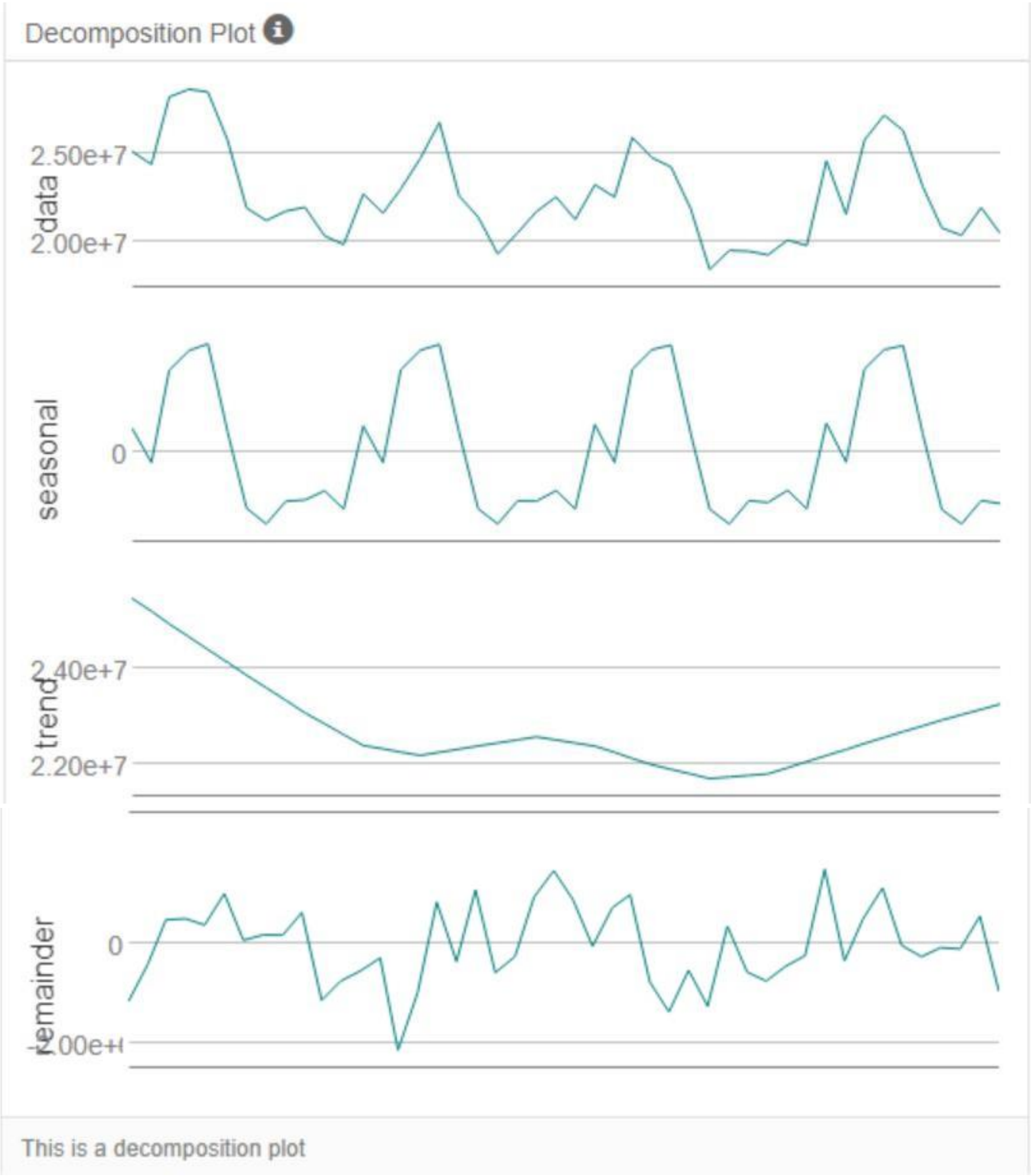
Ans:

For the ETS model – Error is showing an irregular pattern so it should be applied multiplicatively. Trend is showing unclear pattern so no trend should be applied. Seasonality is showing an increasing pattern so it should be applied multiplicatively. So we take ETS (M,N,M) model with no dampening



Time Series Plot

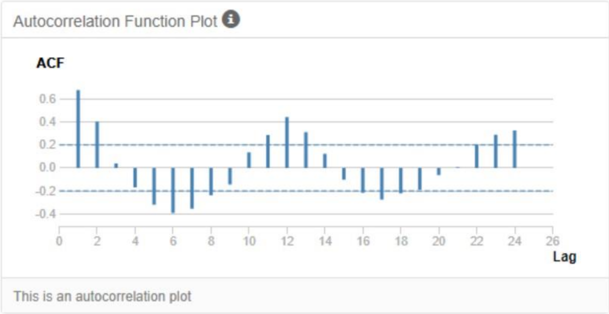
This is a time series plot



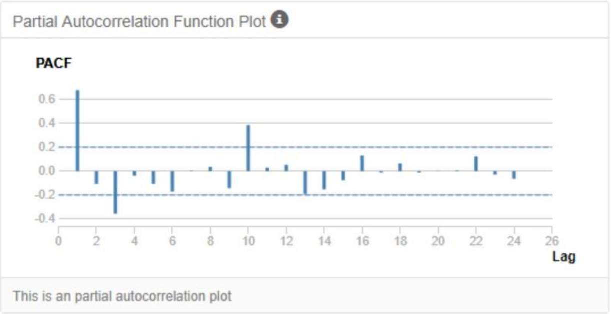
Decomposition Plot

This is a decomposition plot

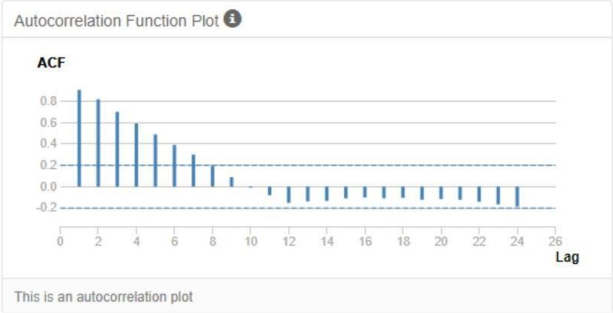
For The ARIMA model we have taken ARIMA(0,1,2)(0,1,0) because seasonal difference and seasonal first difference have been taken and there is a lag-2


Autocorrelation Function Plot

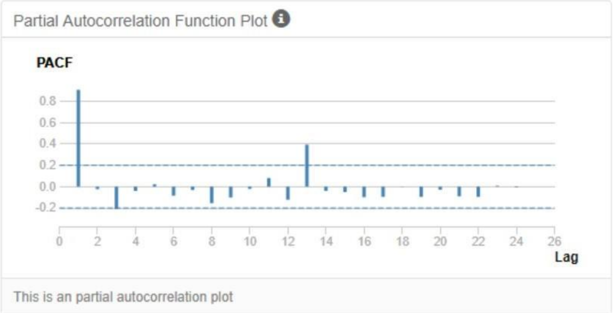
This is an autocorrelation plot


Partial Autocorrelation Function Plot

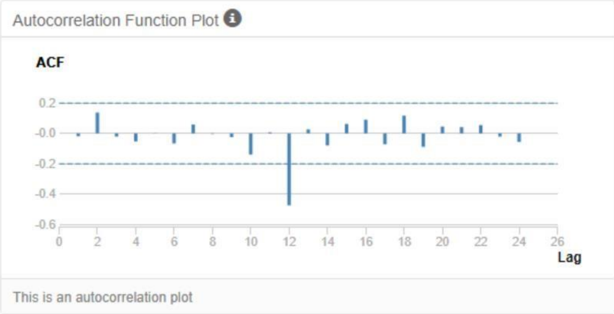
This is an partial autocorrelation plot


Autocorrelation Function Plot

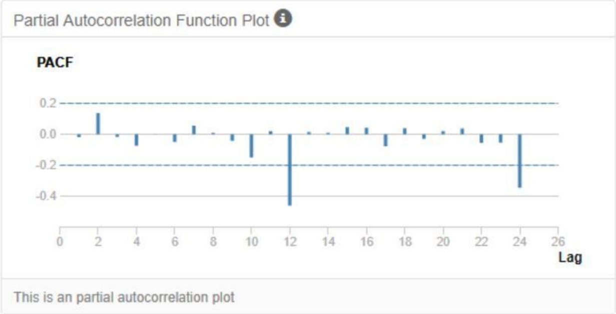
This is an autocorrelation plot


Partial Autocorrelation Function Plot

This is an partial autocorrelation plot


Autocorrelation Function Plot

This is an autocorrelation plot


Partial Autocorrelation Function Plot

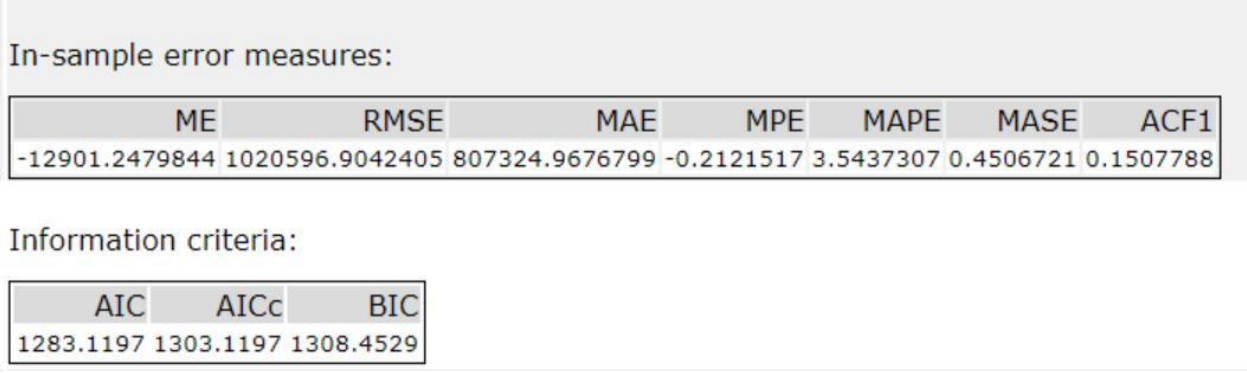
This is an partial autocorrelation plot

Below are the error metrics of the two models. A holdout sample of 6 months was used. RMSE of ETS model is 1020596.9042405 whereas the RMSE of ARIMA model is 1429296.2983494. MASE of ETS is 0.45 compared to ARIMA's 0.53 . Hence Accuracy of ETS is higher

AIC of arima is 859 whereas AIC of ETS is 1283

Hence we are choosing ETS (M,N,M) model

ETS -



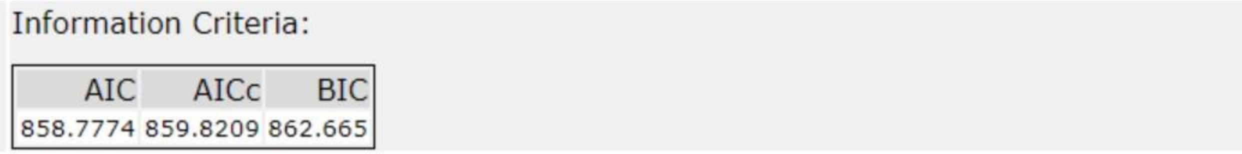
In-sample error measures:

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| -12901.2479844 | 1020596.9042405 | 807324.9676799 | -0.2121517 | 3.5437307 | 0.4506721 | 0.1507788 |

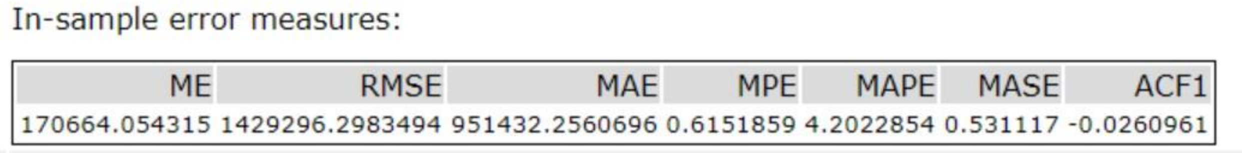
Information criteria:

| AIC | AICc | BIC |
|---|---|---|
| 1283.1197 | 1303.1197 | 1308.4529 |

Arima –

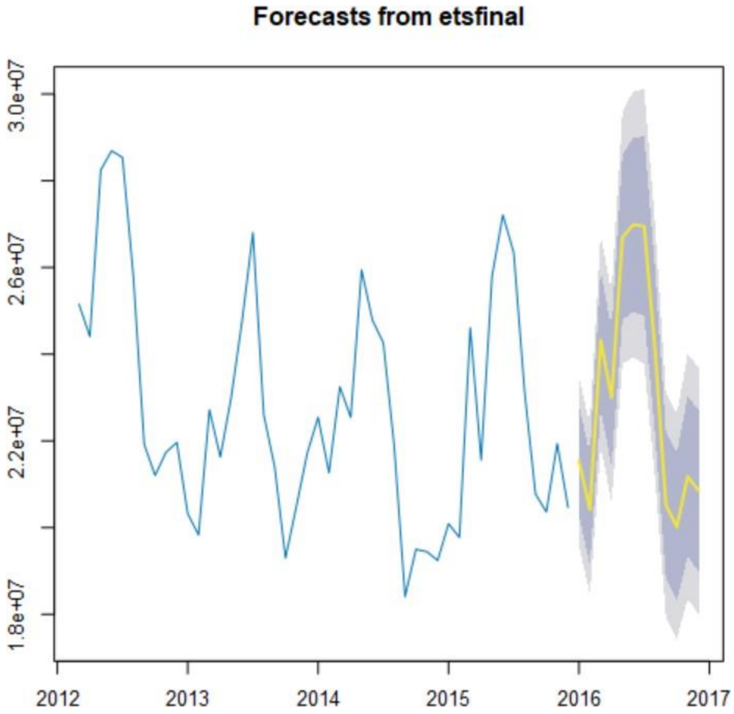

Information Criteria:

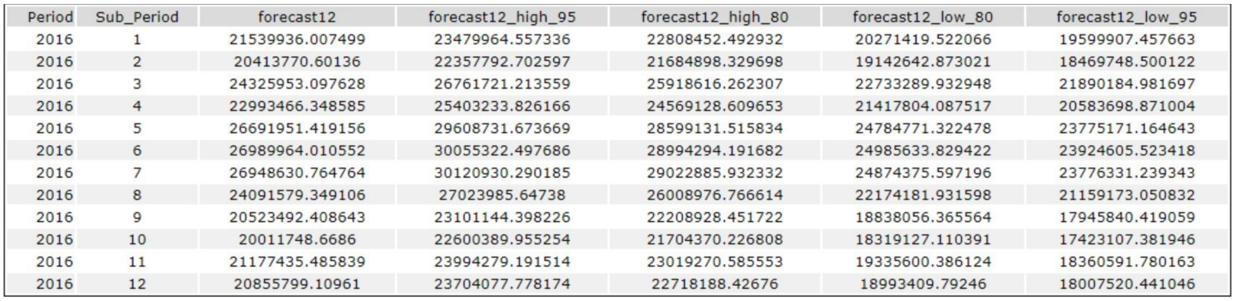
| AIC | AICc | BIC |
|---|---|---|
| 858.7774 | 859.8209 | 862.665 |


In-sample error measures:

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| 170664.054315 | 1429296.2983494 | 951432.2560696 | 0.6151859 | 4.2022854 | 0.531117 | -0.0260961 |

The below graph shows actual values as well as the forecasted values with a 80% and 95% confidence interval. The table shows the same values numerically





| Period | Sub_Period | forecast12 | forecast12_high_95 | forecast12_high_80 | forecast12_low_80 | forecast12_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 |

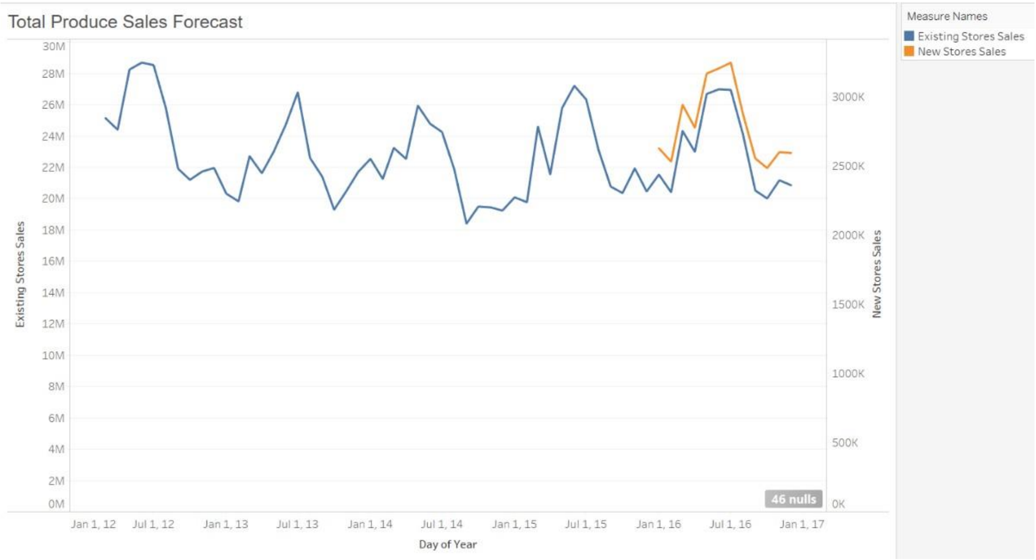
3. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and

new stores forecasts.

ANS

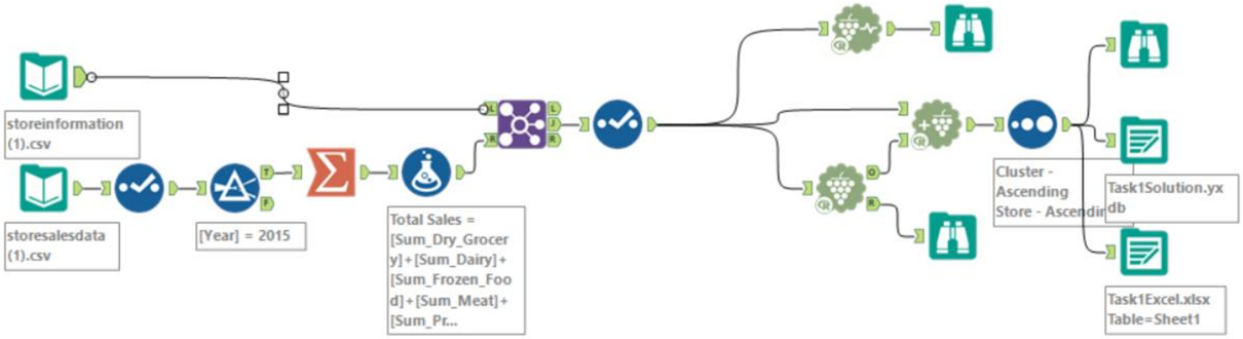
| Year | Month | New Store Sales | Existing Store Sales |
|---|---|---|---|
| 2016 | 1 | 26,26,198 | 2,15,39,936 |
| 2016 | 2 | 25,29,186 | 2,04,13,771 |
| 2016 | 3 | 29,40,264 | 2,43,25,953 |
| 2016 | 4 | 27,74,135 | 2,29,93,466 |
| 2016 | 5 | 31,65,320 | 2,66,91,951 |
| 2016 | 6 | 32,03,286 | 2,69,89,964 |
| 2016 | 7 | 32,44,464 | 2,69,48,631 |
| 2016 | 8 | 28,71,488 | 2,40,91,579 |
| 2016 | 9 | 25,52,418 | 2,05,23,492 |
| 2016 | 10 | 24,82,837 | 2,00,11,749 |
| 2016 | 11 | 25,97,780 | 2,11,77,435 |
| 2016 | 12 | 25,91,815 | 2,08,55,799 |

The above table shows the forecasted sales for existing and new stores



The above chart shows historical sales for existing stores as well as forecasted sales for existing and new stores

## ALTERYX WORKFLOWS



storeinformation (1).csv

storesalesdata (1).csv

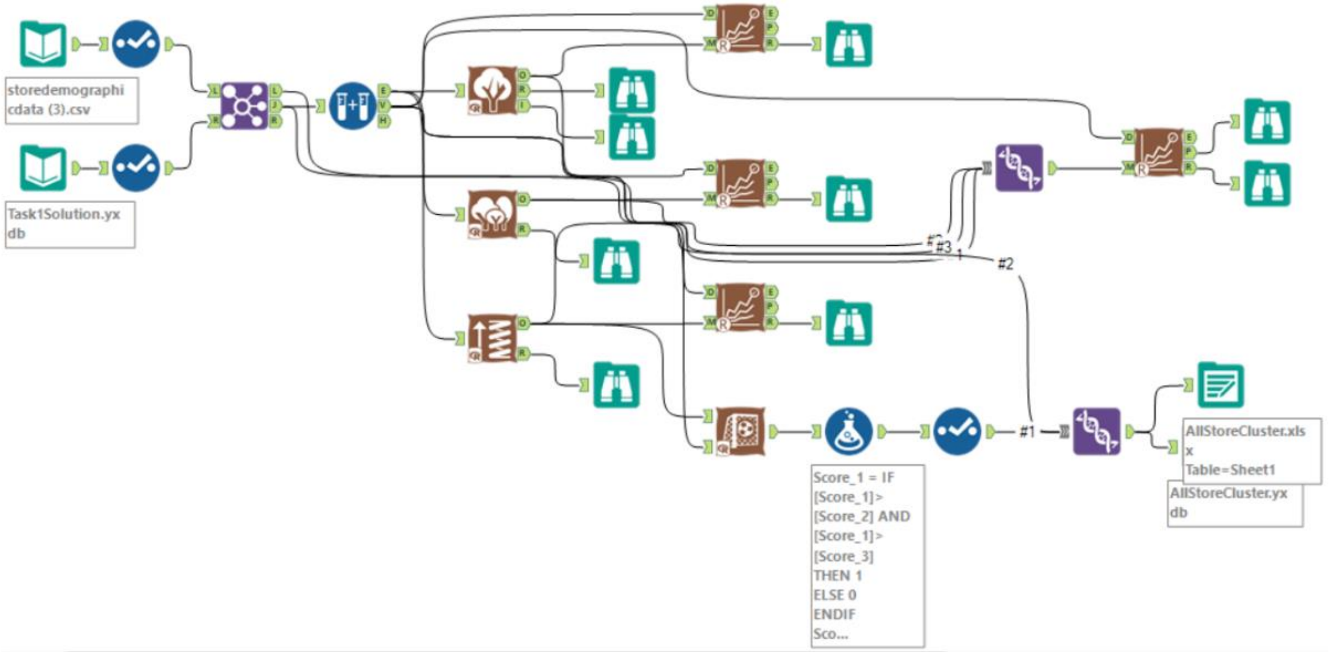
[Year] = 2015

Total Sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...

Cluster - Ascending Store - Ascendir

Task1Solution.yxdb

Task1Excel.xlsx Table=Sheet1



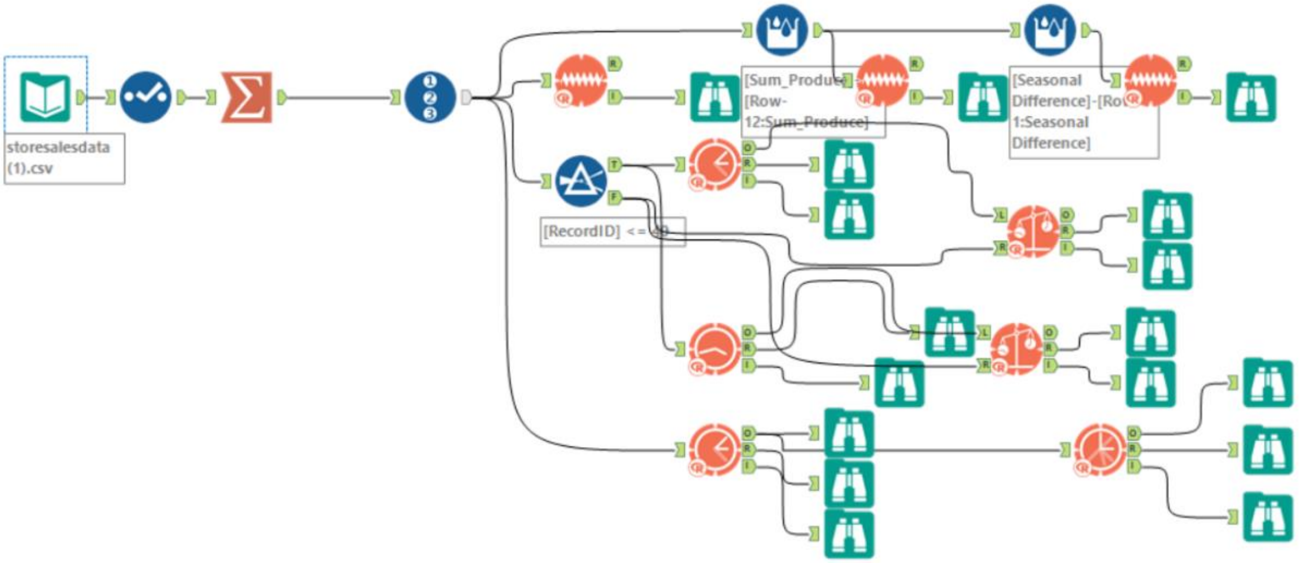
storedemographicdata (3).csv

Task1Solution.yxdb

Score_1 = IF [Score_1]> [Score_2] AND [Score_1]> [Score_3] THEN 1 ELSE 0 ENDIF Sco...

AllStoreCluster.xlsx Table=Sheet1

AllStoreCluster.yxdb



storesalesdata (1).csv

[Sum_Produce] - [Row-12:Sum_Produce]

[Seasonal Difference]-[Row-1:Seasonal Difference]

[RecordID] <= 40

## Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.