# Project: Predictive Analytics Capstone

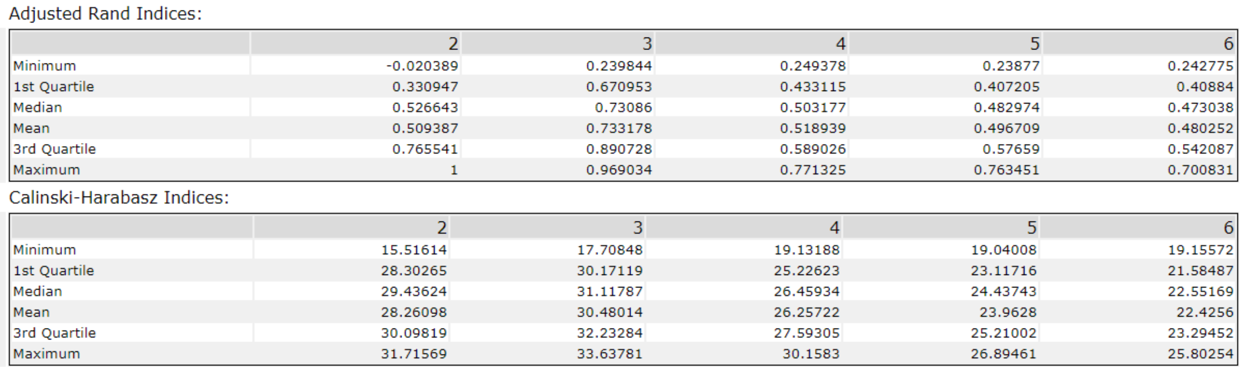
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

## Task 1: Determine Store Formats for Existing Stores

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

Considering the K-means report, Adjusted Rand and Calinski-Harabasz indices as depicted below, the optimal number of store formats is **3.**

Both AR and CH indices have highest median value at 3 clusters. So this corresponds to 3 store formats.

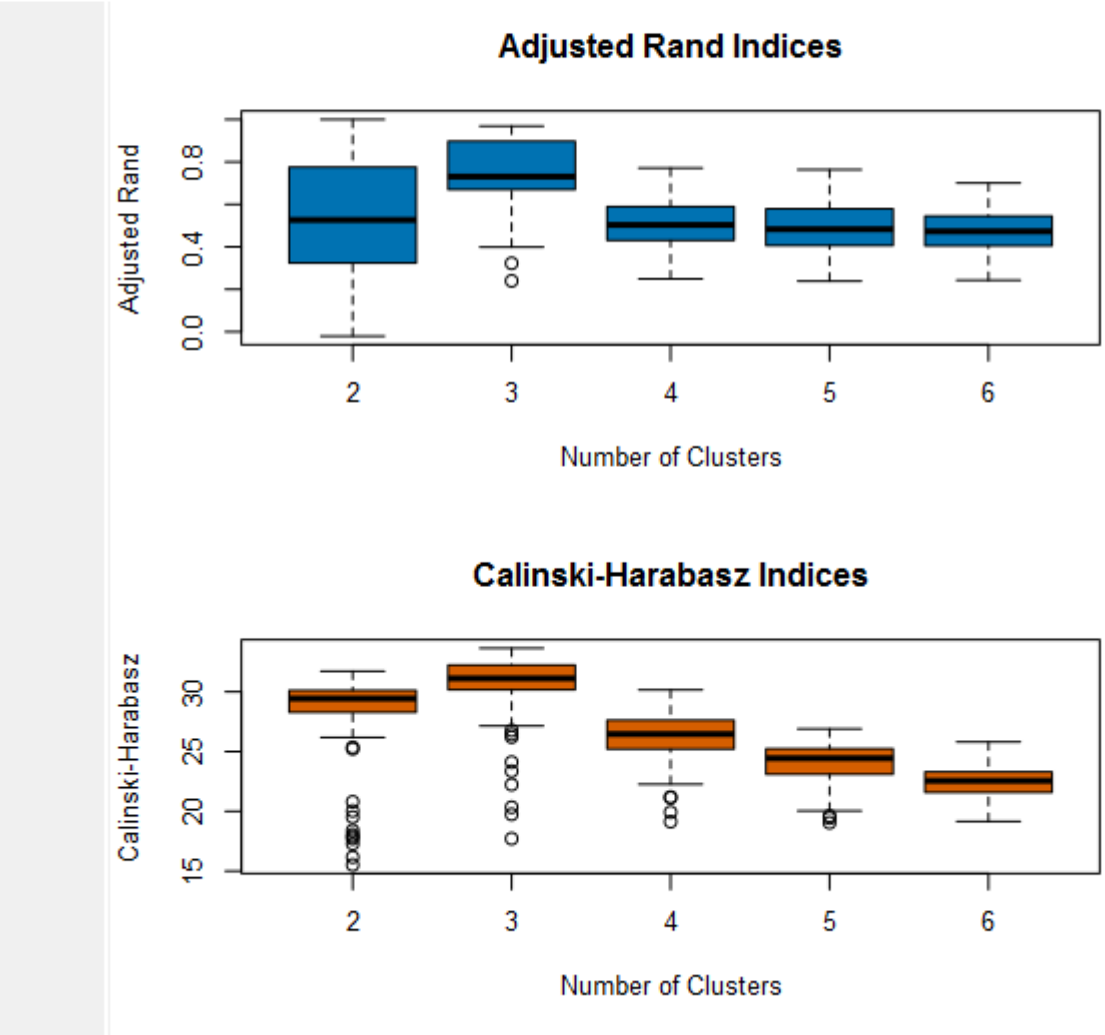


Adjusted Rand Indices:

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | -0.020389 | 0.239844 | 0.249378 | 0.23877 | 0.242775 |
| 1st Quartile | 0.330947 | 0.670953 | 0.433115 | 0.407205 | 0.40884 |
| Median | 0.526643 | 0.73086 | 0.503177 | 0.482974 | 0.473038 |
| Mean | 0.509387 | 0.733178 | 0.518939 | 0.496709 | 0.480252 |
| 3rd Quartile | 0.765541 | 0.890728 | 0.589026 | 0.57659 | 0.542087 |
| Maximum | 1 | 0.969034 | 0.771325 | 0.763451 | 0.700831 |

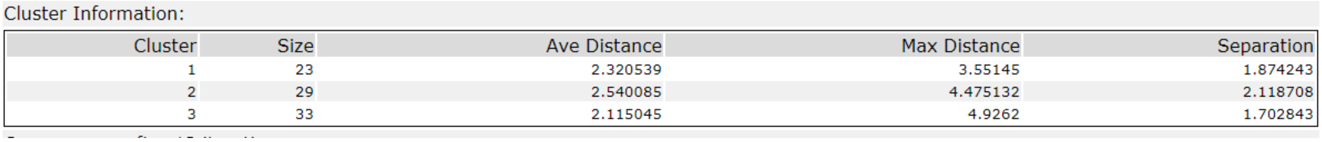
Calinski-Harabasz Indices:

| | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Minimum | 15.51614 | 17.70848 | 19.13188 | 19.04008 | 19.15572 |
| 1st Quartile | 28.30265 | 30.17119 | 25.22623 | 23.11716 | 21.58487 |
| Median | 29.43624 | 31.11787 | 26.45934 | 24.43743 | 22.55169 |
| Mean | 28.26098 | 30.48014 | 26.25722 | 23.9628 | 22.4256 |
| 3rd Quartile | 30.09819 | 32.23284 | 27.59305 | 25.21002 | 23.29452 |
| Maximum | 31.71569 | 33.63781 | 30.1583 | 26.89461 | 25.80254 |



2. How many stores fall into each store format?

This is as below:



Cluster Information:

| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 23 | 2.320539 | 3.55145 | 1.874243 |
| 2 | 29 | 2.540085 | 4.475132 | 2.118708 |
| 3 | 33 | 2.115045 | 4.9262 | 1.702843 |

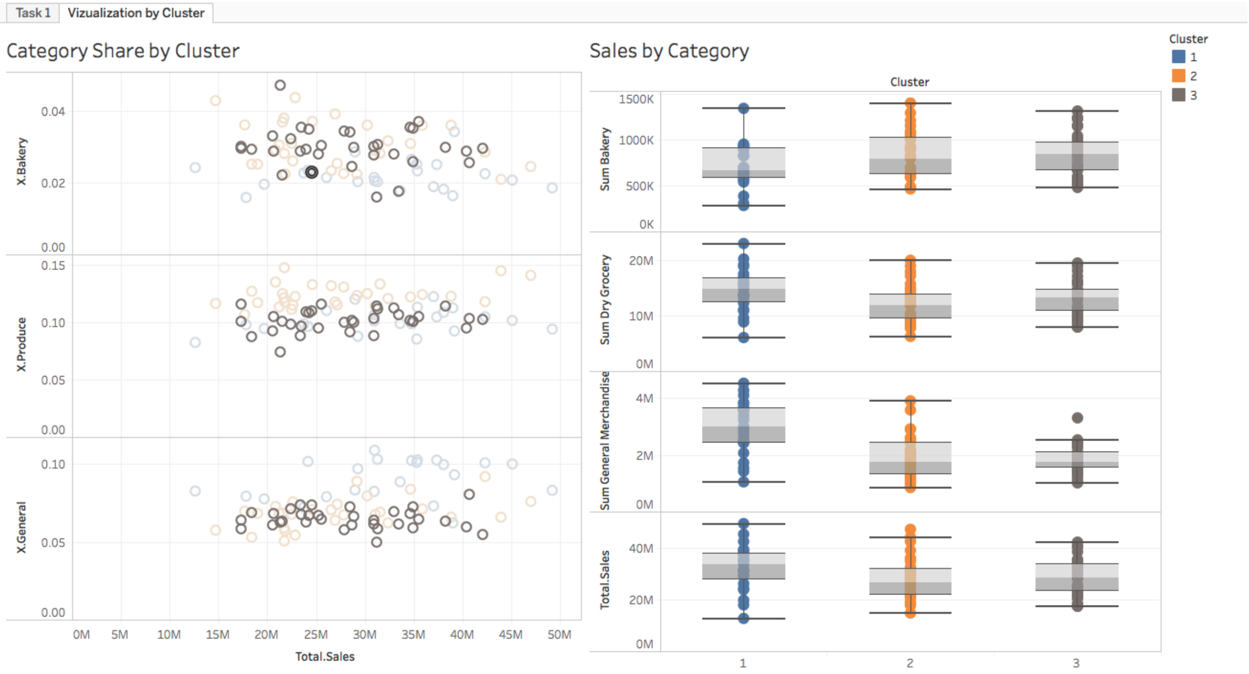
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?::

Basis the below graph, the below can be inferred:

- Cluster 1 stores:
  - Sold more General Merchandise in terms of percentage
  - Has highest medial total sales

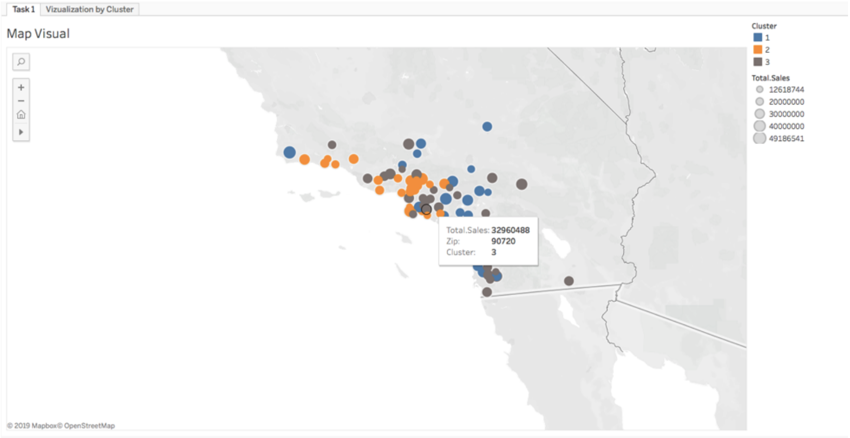
    - Range is also highest in Cluster 1
- Cluster 2
    - Sold more Produce in terms of percentage
- Cluster 3:
    - Most similar- most compact range

Cluster 1 stores have highest medial total sales when compared to the other 2. Its range of total sales and most of other categorical sales are also the largest. Cluster 3 stores are the most similar in terms of sales due to more compact range.



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.

This is as below:



The Tableau links are as below:

https://public.tableau.com/profile/siddharth3961#!/vizhome/Task1_15540608764730/VizualizationbyCluster

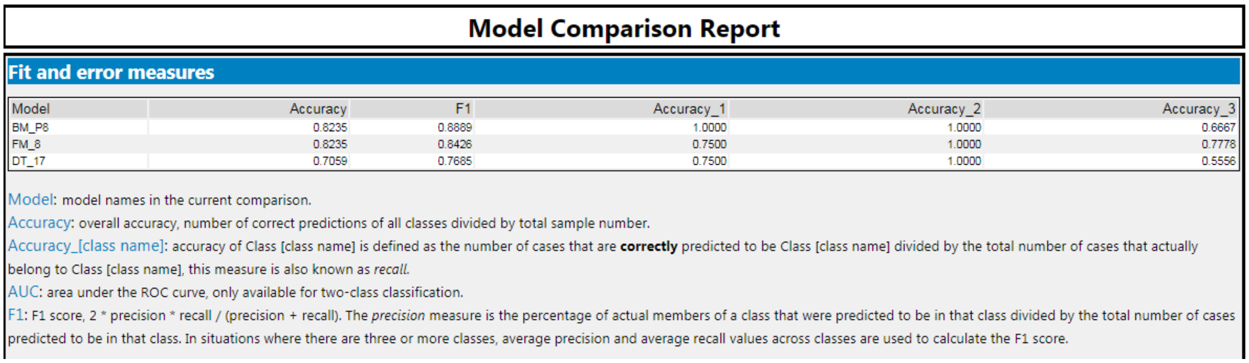
## Task 2: Formats for New Stores

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.)

Below is the model comparison report. I have compared three models:

a. Boosted Model
b. Forest Model
c. Decision Tree

I have chosen the **Boosted Model** since it is highest in both Accuracy and F1 score.



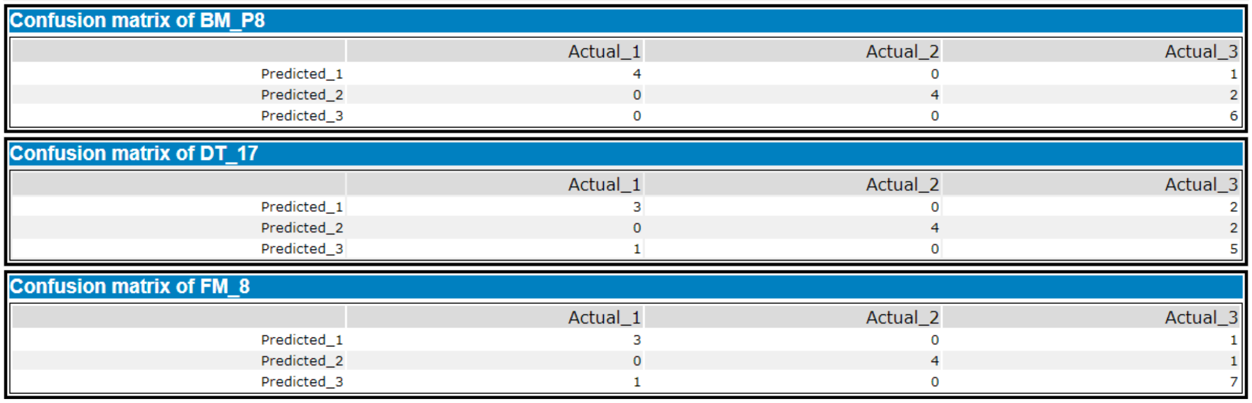
**Model Comparison Report**

**Fit and error measures**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| BM_P8 | 0.8235 | 0.8889 | 1.0000 | 1.0000 | 0.6667 |
| FM_8 | 0.8235 | 0.8426 | 0.7500 | 1.0000 | 0.7778 |
| DT_17 | 0.7059 | 0.7685 | 0.7500 | 1.0000 | 0.5556 |

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

The confusion matrix is as below:



**Confusion matrix of BM_P8**

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

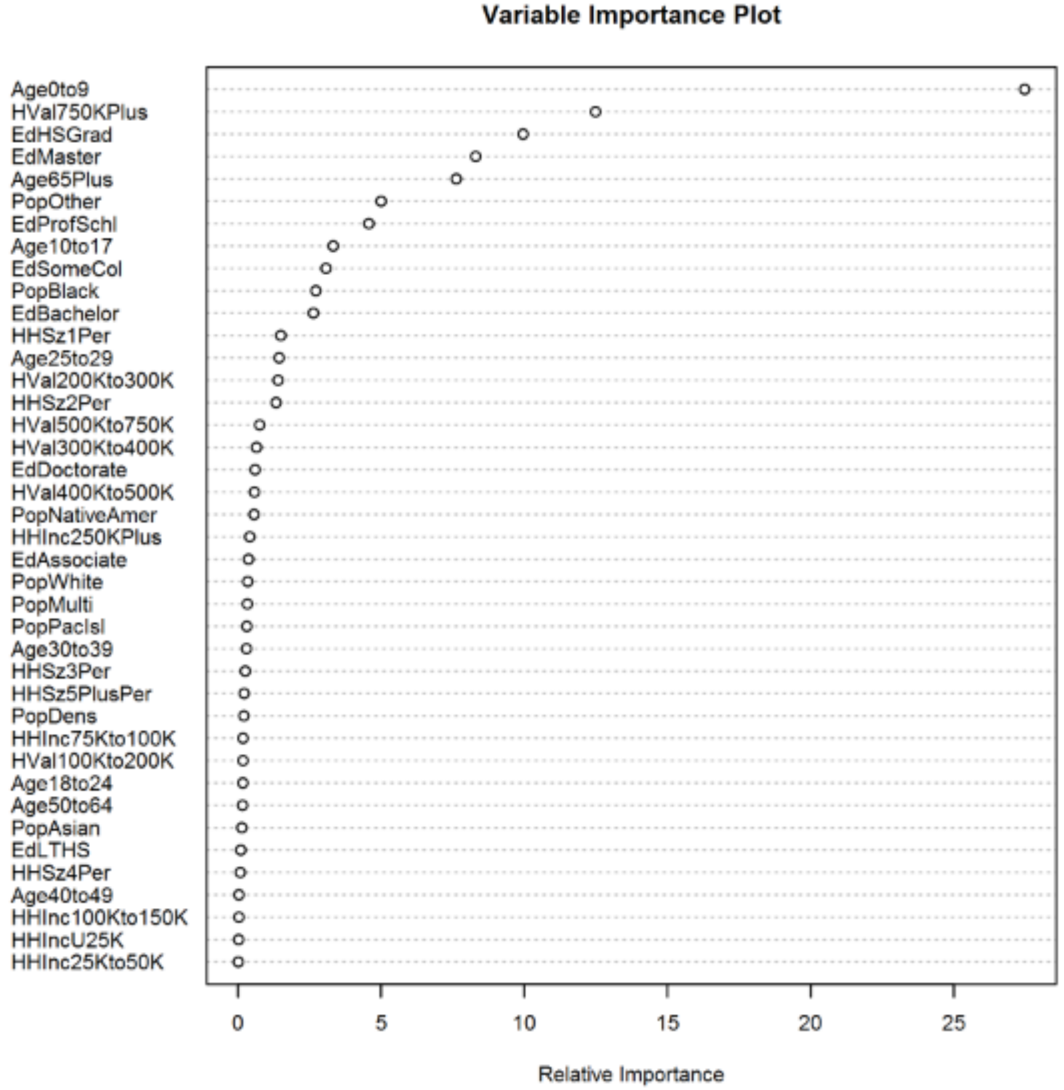
**Confusion matrix of DT_17**

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

**Confusion matrix of FM_8**

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

**The most important factors are as below:**



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

Basis the score tool, the segments for the new stores have been predicted as below:

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

| | |
|---|---|
| S0093 | 1 |
| S0094 | 2 |
| S0095 | 2 |

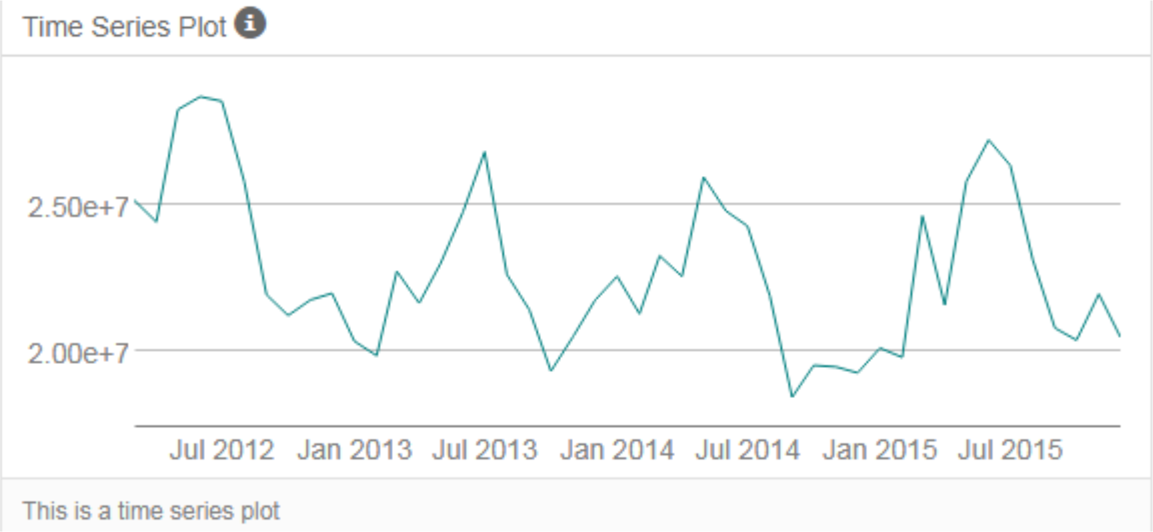
## Task 3: Predicting Produce Sales

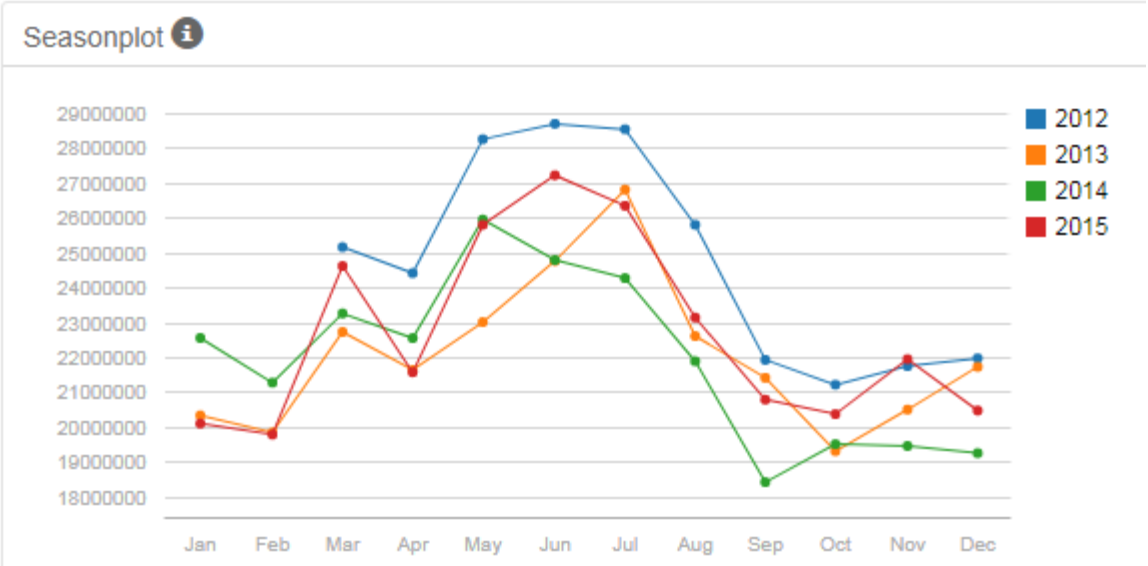
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

I have used the TS Plot tool to generate the below graphs.

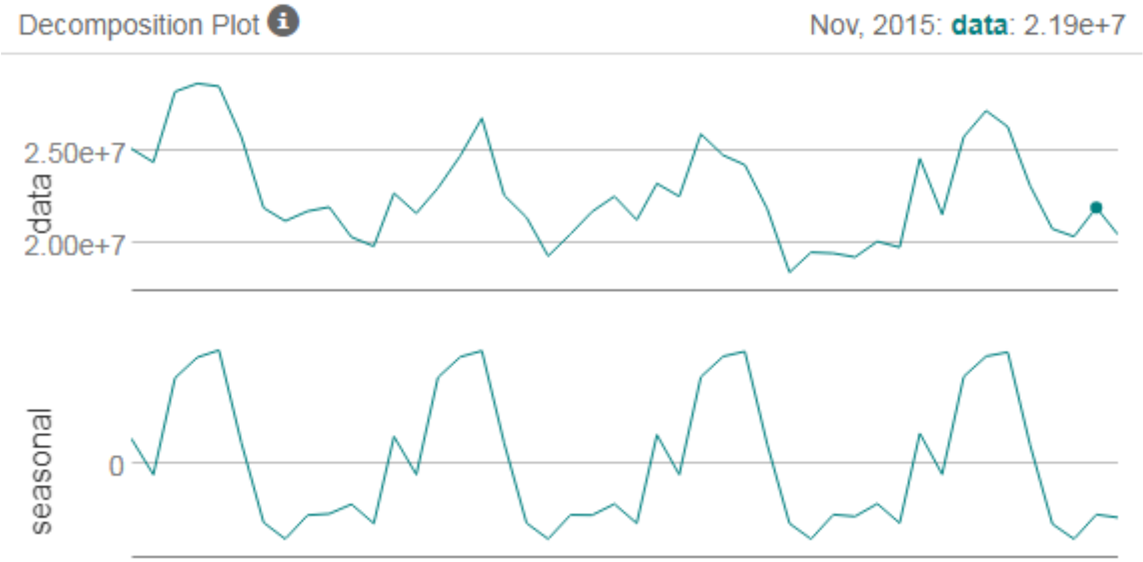
Considering the ETS plot:

1. Seasonality: Is increasing and hence applied multiplicatively
2. Trend: No clear trend and hence not taken
3. Error: Error is applied multiplicatively



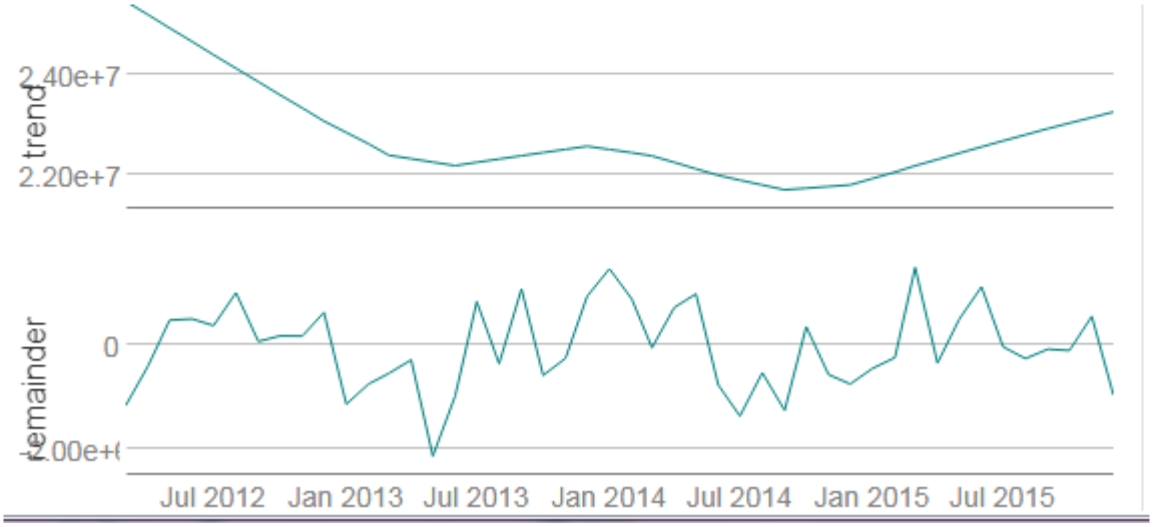


Seasonplot



Decomposition Plot

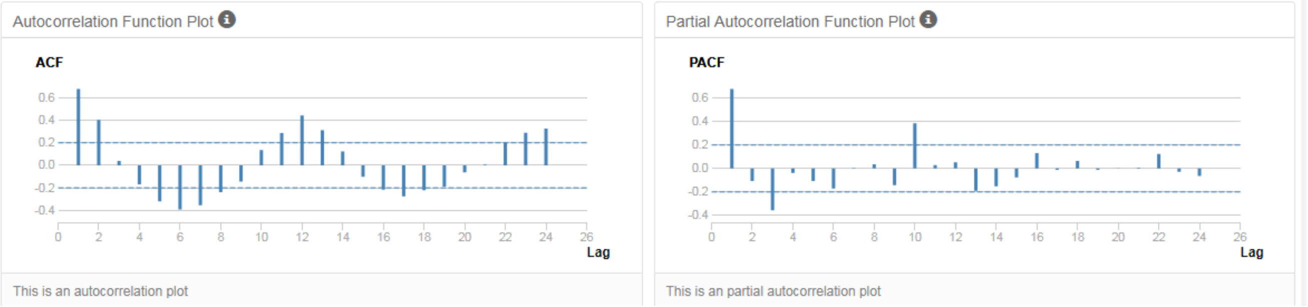
Nov, 2015: **data**: 2.19e+7



**Considering the ARIMA Plot:**

ARIMA models are displayed in the terms (p,d,q). These are explained as below:

- p - periods to lag for
- d - number of transformations used to make the data stationary
- q - lags of the error component

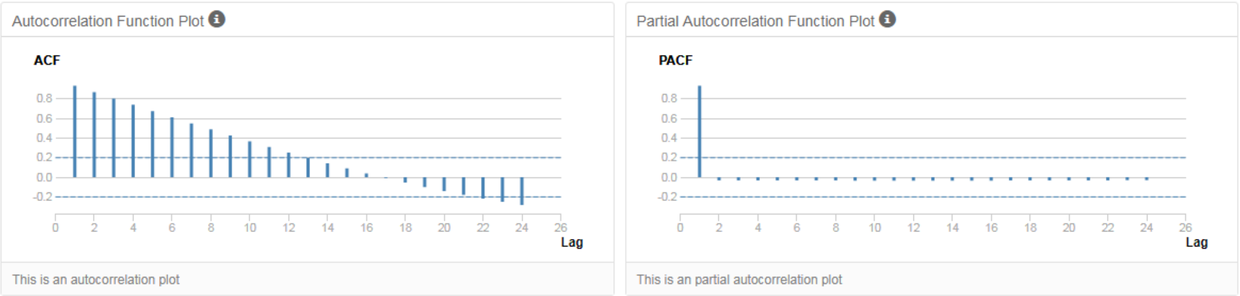

Autocorrelation Function Plot

This is an autocorrelation plot

Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

I have re-plotted the ACF and PACF graphs after taking 1 seasonal difference. Even after this, the ACF still shows high co-relation.

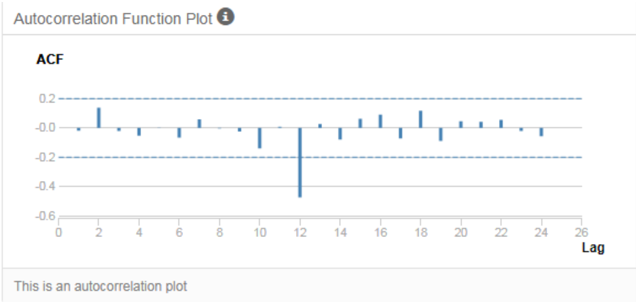

Autocorrelation Function Plot

This is an autocorrelation plot

Partial Autocorrelation Function Plot

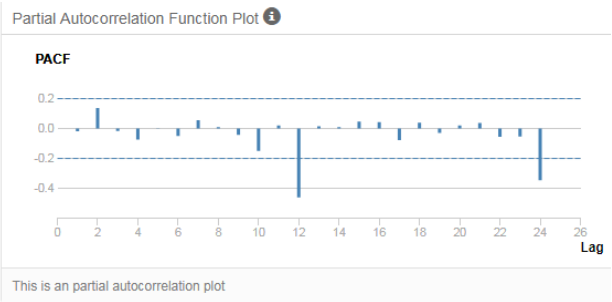
This is an partial autocorrelation plot

I have re-plotted the ACF and PACF graphs after taking the 1st difference. There is no significant co-relation now.



Autocorrelation Function Plot

This is an autocorrelation plot



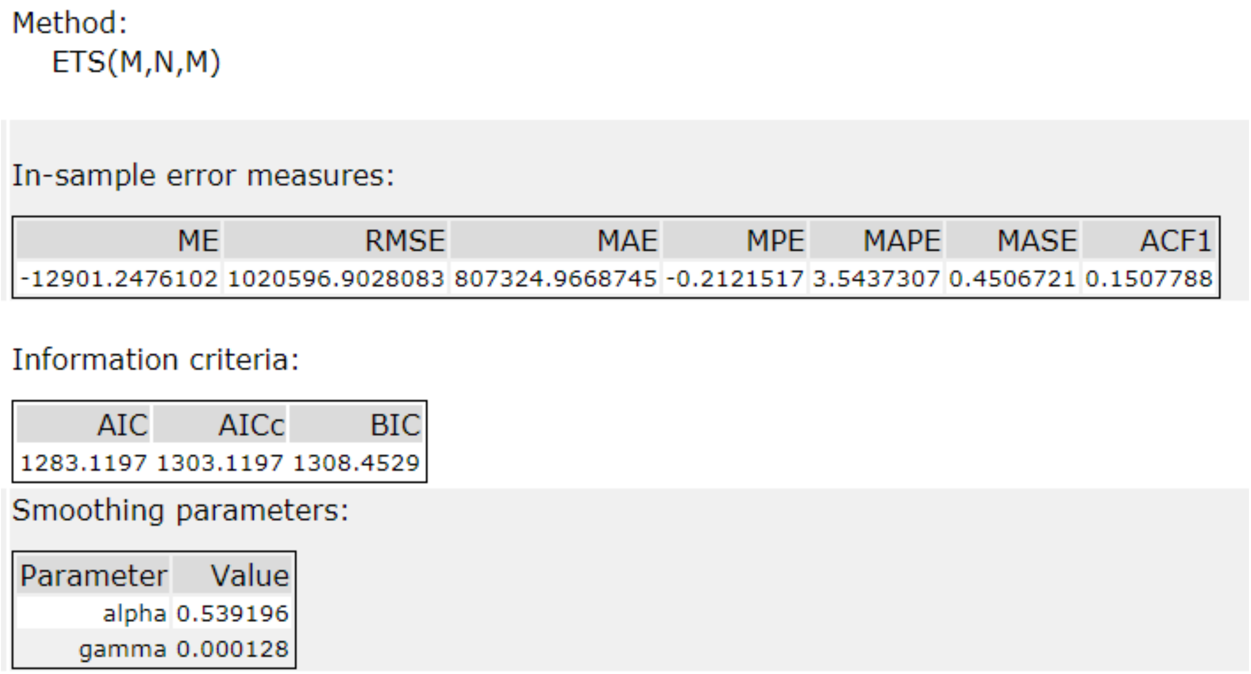
Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

The accuracy of ETS model is higher compared to ARIMA model. I have used a holdout sample of 6 months data.

1. The RMSE of ETS is 1,020,597, which is lower than ARIMA's 1,429,296
2. MASE of ETS  is 0.45 compared to ARIMA's 0.53

**The in-sample error measure for ETS Time Series are as below:**



Method:
ETS(M,N,M)

In-sample error measures:

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| -12901.2476102 | 1020596.9028083 | 807324.9668745 | -0.2121517 | 3.5437307 | 0.4506721 | 0.1507788 |

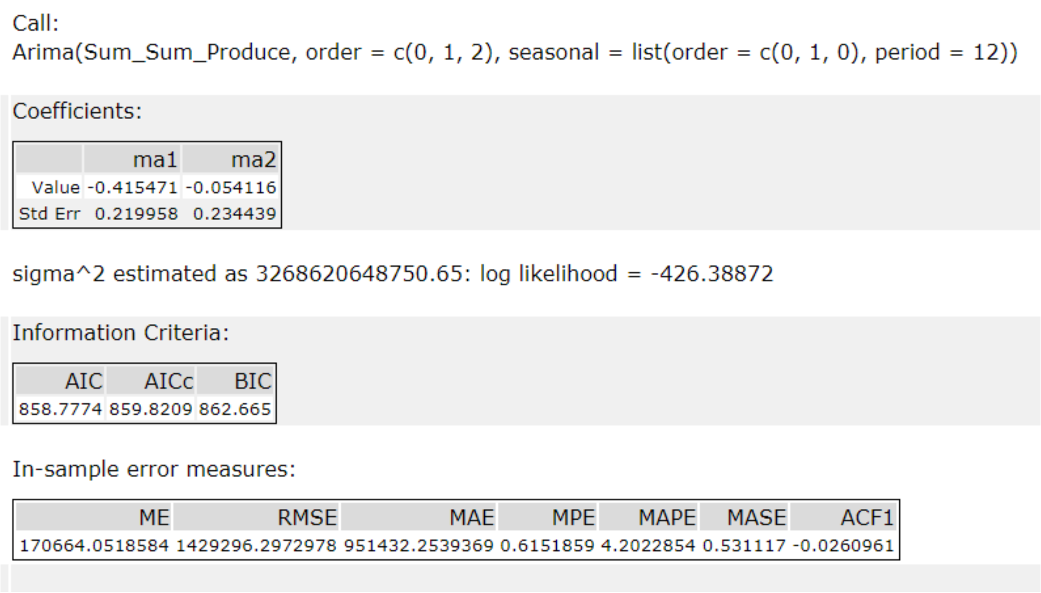
Information criteria:

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

Smoothing parameters:

| Parameter | Value |
|---|---|
| alpha | 0.539196 |
| gamma | 0.000128 |

**For ARIMA the details are as below:**



Call:
Arima(Sum_Sum_Produce, order = c(0, 1, 2), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:

|  | ma1 | ma2 |
|---|---|---|
| Value | -0.415471 | -0.054116 |
| Std Err | 0.219958 | 0.234439 |

sigma^2 estimated as 3268620648750.65: log likelihood = -426.38872

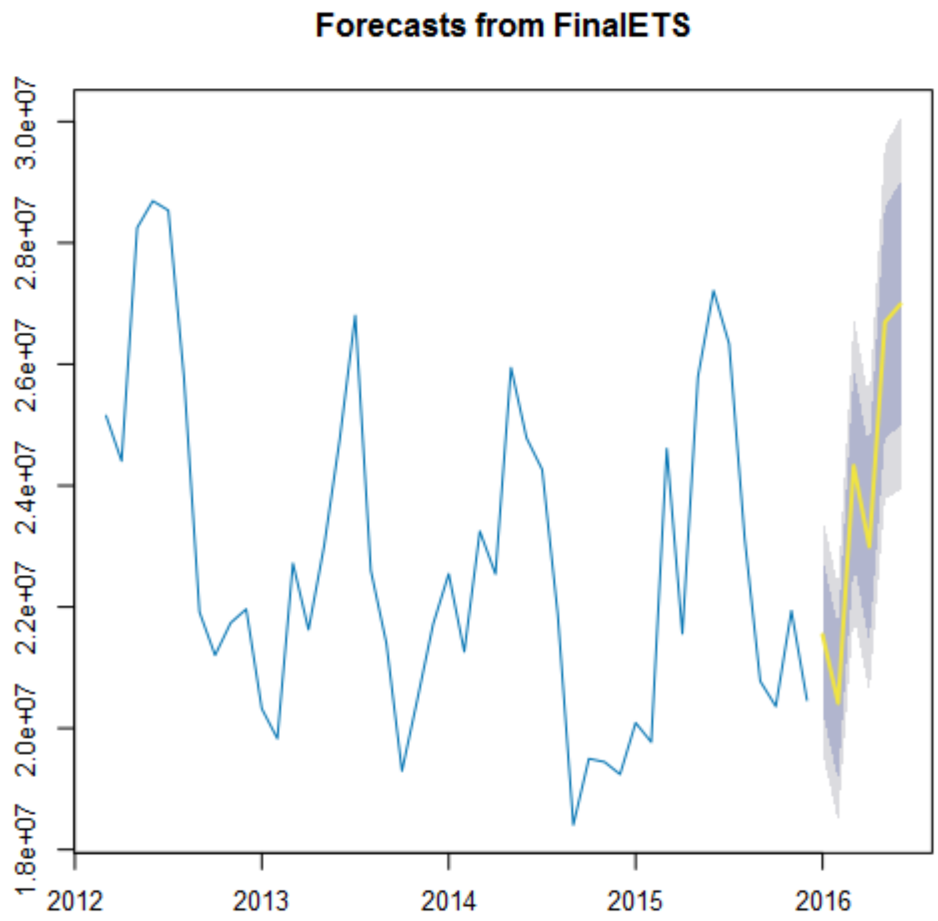
Information Criteria:

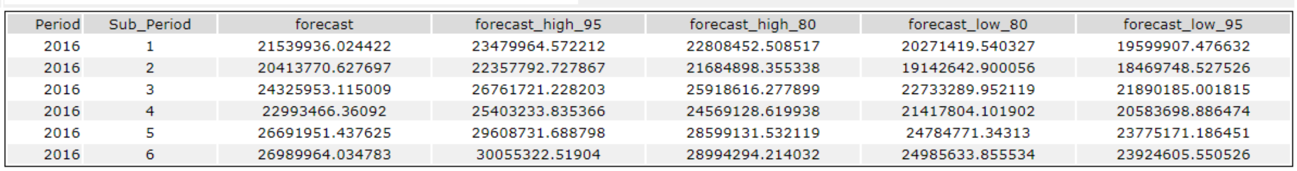
| AIC | AICc | BIC |
|---|---|---|
| 858.7774 | 859.8209 | 862.665 |

In-sample error measures:

| ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|
| 170664.0518584 | 1429296.2972978 | 951432.2539369 | 0.6151859 | 4.2022854 | 0.531117 | -0.0260961 |

**The forecast using TS forecast using the ETS series is as below. The actual and forecast values are with 80% & 95% confidence level intervals.**





| Period | Sub_Period | forecast | forecast_high_95 | forecast_high_80 | forecast_low_80 | forecast_low_95 |
|---|---|---|---|---|---|---|
| 2016 | 1 | 21539936.024422 | 23479964.572212 | 22808452.508517 | 20271419.540327 | 19599907.476632 |
| 2016 | 2 | 20413770.627697 | 22357792.727867 | 21684898.355338 | 19142642.900056 | 18469748.527526 |
| 2016 | 3 | 24325953.115009 | 26761721.228203 | 25918616.277899 | 22733289.952119 | 21890185.001815 |
| 2016 | 4 | 22993466.36092 | 25403233.835366 | 24569128.619938 | 21417804.101902 | 20583698.886474 |
| 2016 | 5 | 26691951.437625 | 29608731.688798 | 28599131.532119 | 24784771.34313 | 23775171.186451 |
| 2016 | 6 | 26989964.034783 | 30055322.51904 | 28994294.214032 | 24985633.855534 | 23924605.550526 |

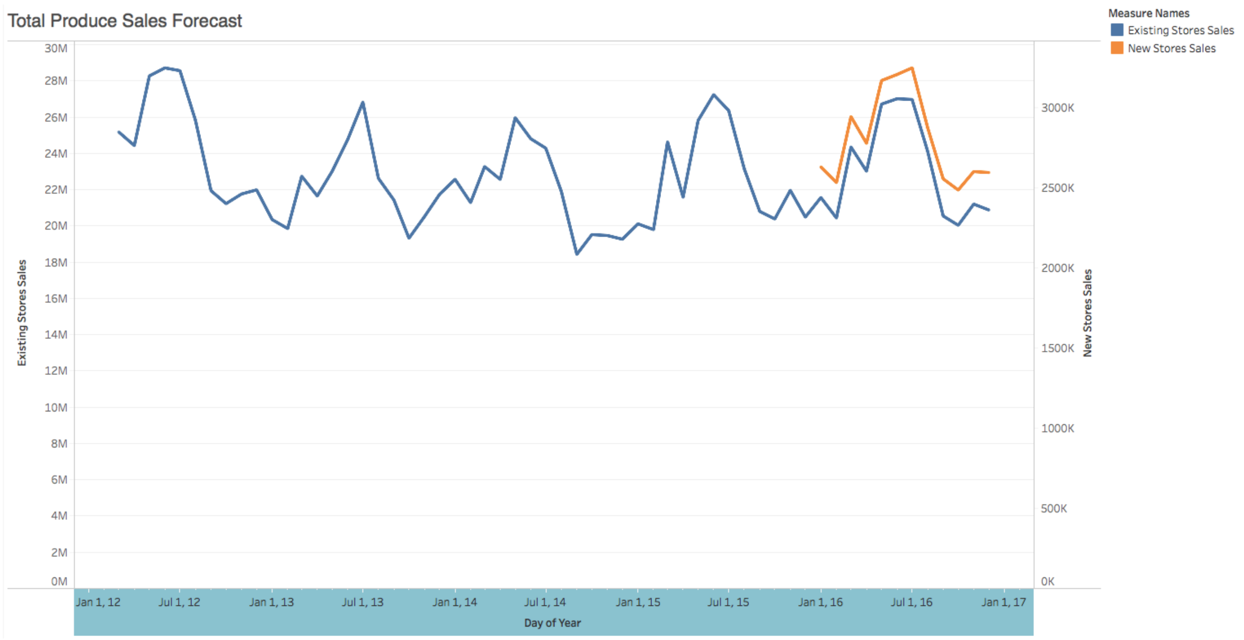
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

**The forecast table is as below having forecasts from both existing and new stores: This is for the year 2016.**

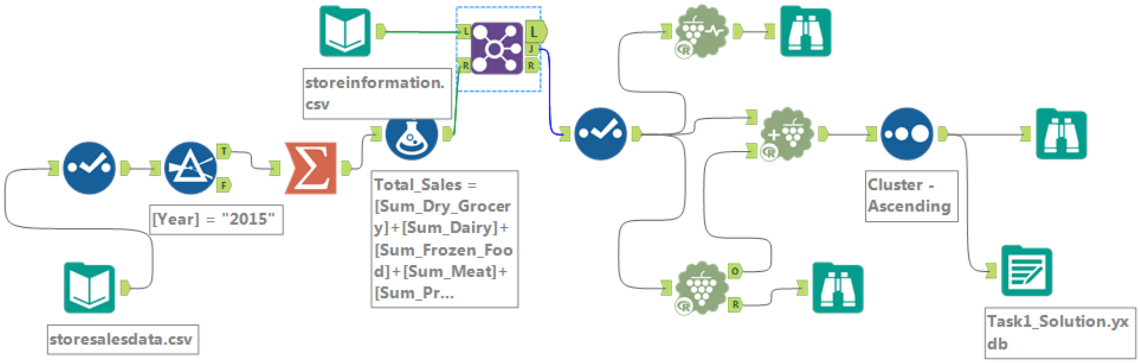
| Month | Existing Stores | New Stores | Total |
|---|---|---|---|
| Jan-16 | 21539936 | 2587451 | 24127387 |
| Feb-16 | 20413771 | 2477353 | 22891124 |
| Mar-16 | 24325953 | 2913185 | 27239138 |
| Apr-16 | 22993466 | 2775746 | 25769212 |
| May-16 | 26691951 | 3150867 | 29842818 |
| Jun-16 | 26989964 | 3188922 | 30178886 |
| Jul-16 | 26948631 | 3214746 | 30163376 |
| Aug-16 | 24091579 | 2866349 | 26957928 |
| Sep-16 | 20523492 | 2538727 | 23062219 |
| Oct-16 | 20011749 | 2488148 | 22499897 |
| Nov-16 | 21177436 | 2595270 | 23772706 |
| Dec-16 | 20855799 | 2573397 | 23429196 |

The Tableau graph is as below:

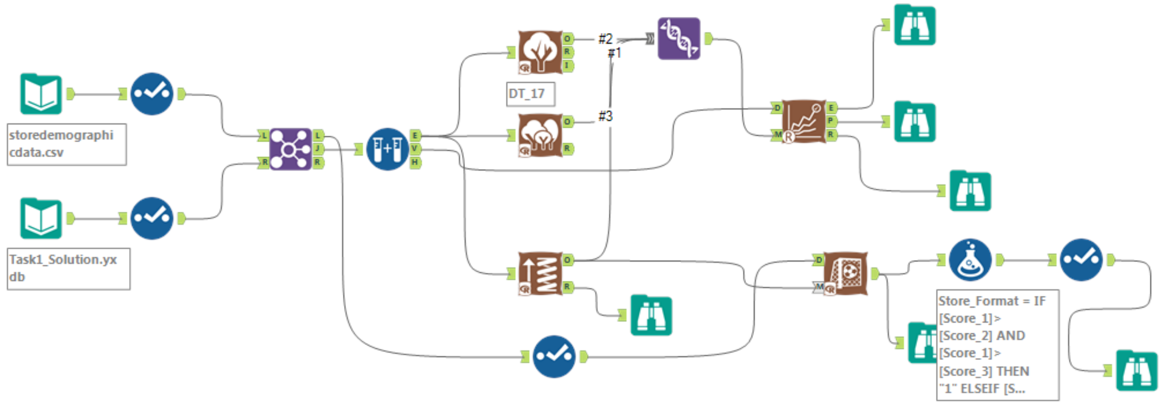
https://public.tableau.com/profile/siddharth3961#!/vizhome/Task3_15540600989490/TotalProduceSalesForecast?publish=yes



## Task 1: Workflow

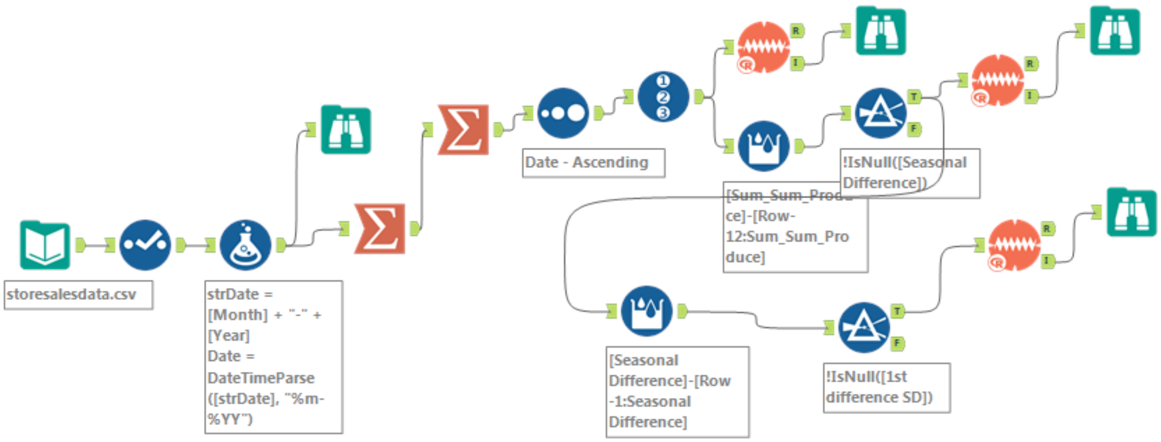


## Task 2: Workflow

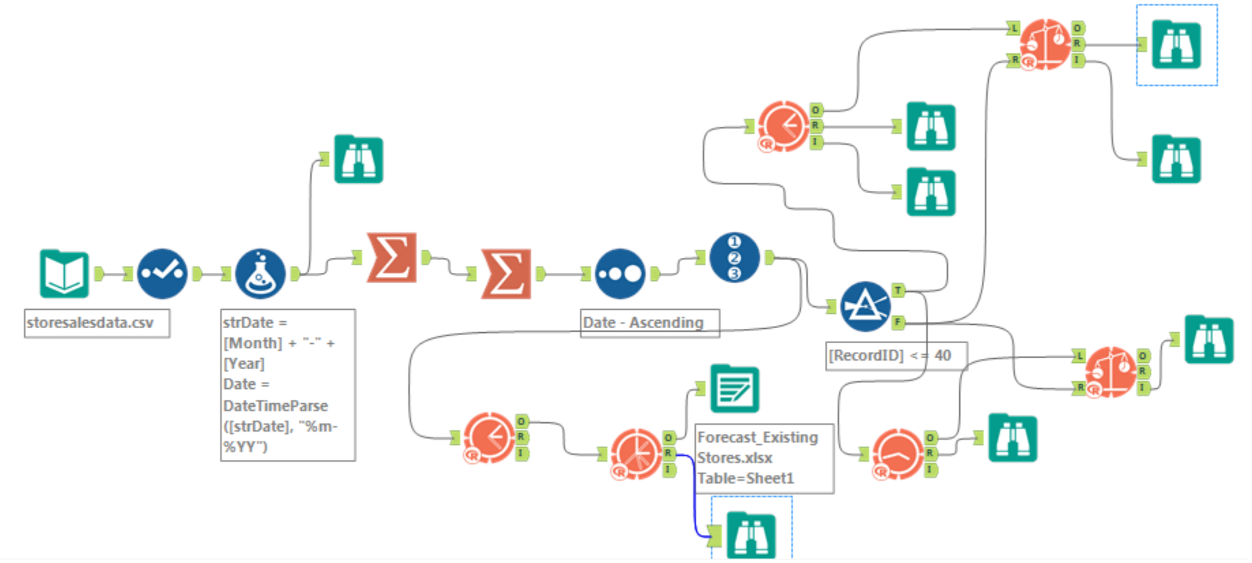

storedemographicdata.csv
Task1_Solution.yxdb
DT_17
#2
#1
#3
Store_Format = IF [Score_1]> [Score_2] AND [Score_1]> [Score_3] THEN "1" ELSEIF [S...

## Task 3 Workflows:

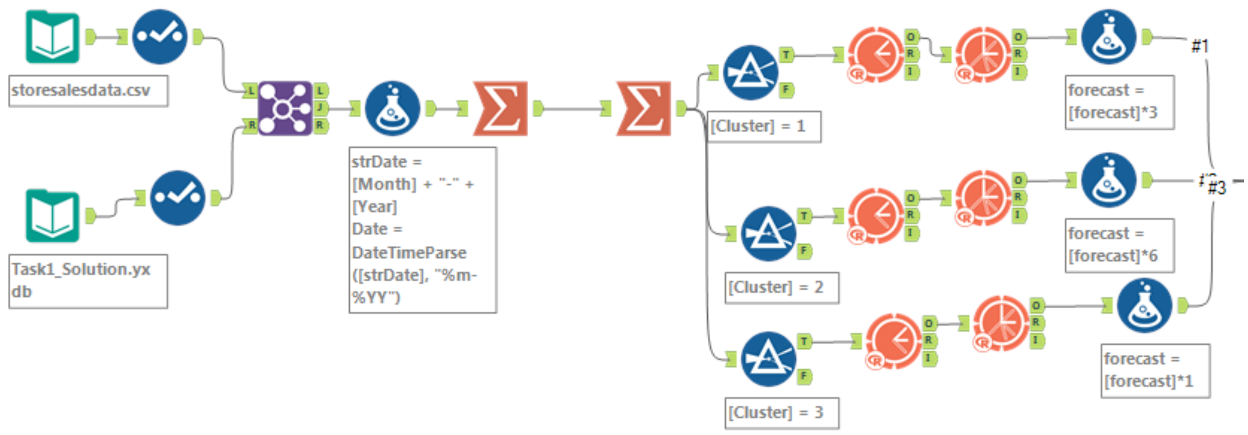
### a) Time Series Analysis using TS Plot


storesalesdata.csv
strDate = [Month] + "-" + [Year]
Date = DateTimeParse ([strDate], "%m-%YY")
Date - Ascending
[Sum_Sum_Produce]-[Row-12:Sum_Sum_Produce]
!IsNull([Seasonal Difference])
[Seasonal Difference]-[Row -1:Seasonal Difference]
!IsNull([1st difference SD])

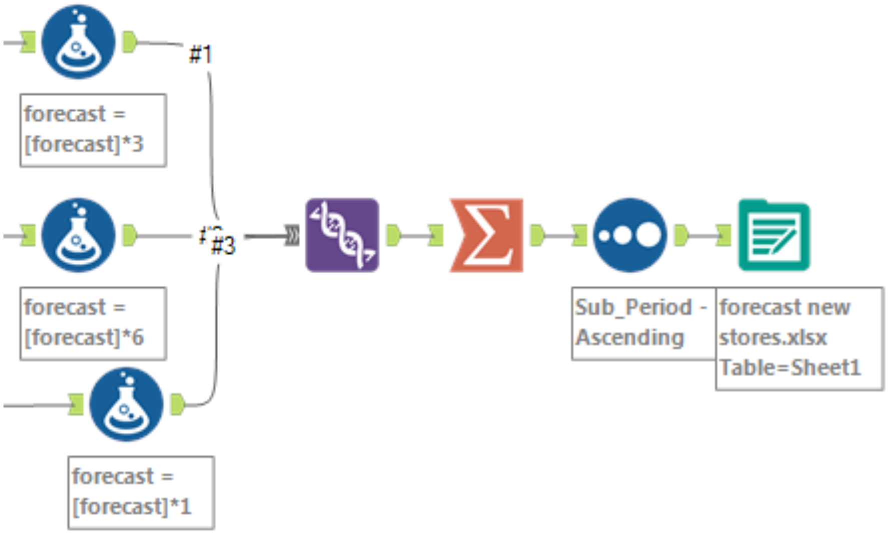
### b) Forecast from existing stores


storesalesdata.csv
strDate = [Month] + "-" + [Year] Date = DateTimeParse ([strDate], "%m-%YY")
Date - Ascending
[RecordID] <= 40
Forecast_Existing Stores.xlsx Table=Sheet1

### c) Forecast from New stores



Cont..


#1
forecast =
[forecast]*3
#3
forecast =
[forecast]*6
forecast =
[forecast]*1
Sub_Period -
Ascending
forecast new
stores.xlsx
Table=Sheet1