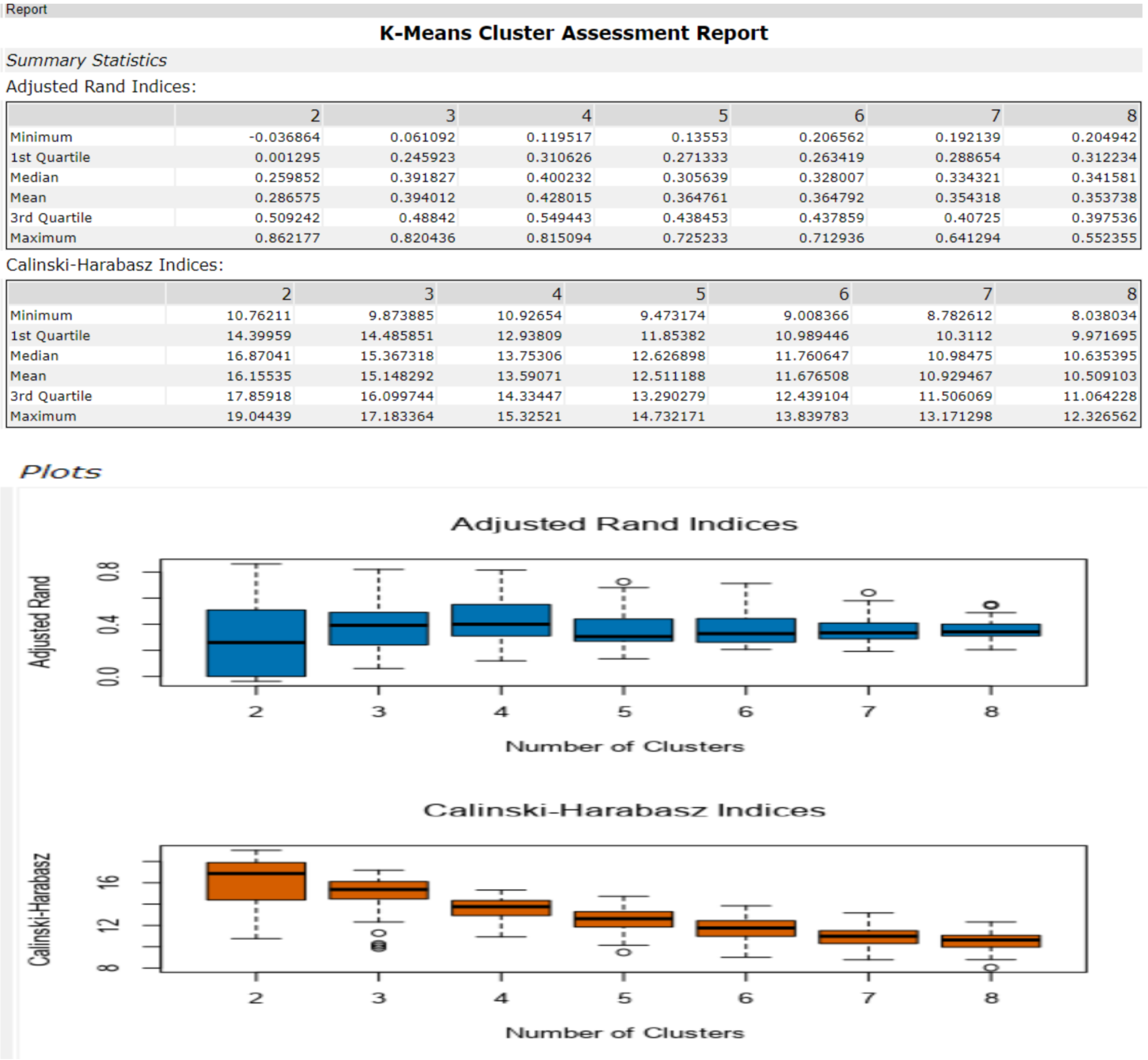
Project: Predictive Analytics Capstone

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 segments is 3, this I obtained from using K-Centroids diagnostic tool and K-means method.

As shown in the diagram below, K-Means Cluster assessment report which includes the Rand and Calinski-Harabasz indices. The median and spread was determined by each cluster. Although, the box-whisker plots show high median values between cluster 2 and 3, I selected cluster 3 because it has a tight or compact spread.



Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.036864	0.061092	0.119517	0.13553	0.206562	0.192139	0.204942
1st Quartile	0.001295	0.245923	0.310626	0.271333	0.263419	0.288654	0.312234
Median	0.259852	0.391827	0.400232	0.305639	0.328007	0.334321	0.341581
Mean	0.286575	0.394012	0.428015	0.364761	0.364792	0.354318	0.353738
3rd Quartile	0.509242	0.48842	0.549443	0.438453	0.437859	0.40725	0.397536
Maximum	0.862177	0.820436	0.815094	0.725233	0.712936	0.641294	0.552355

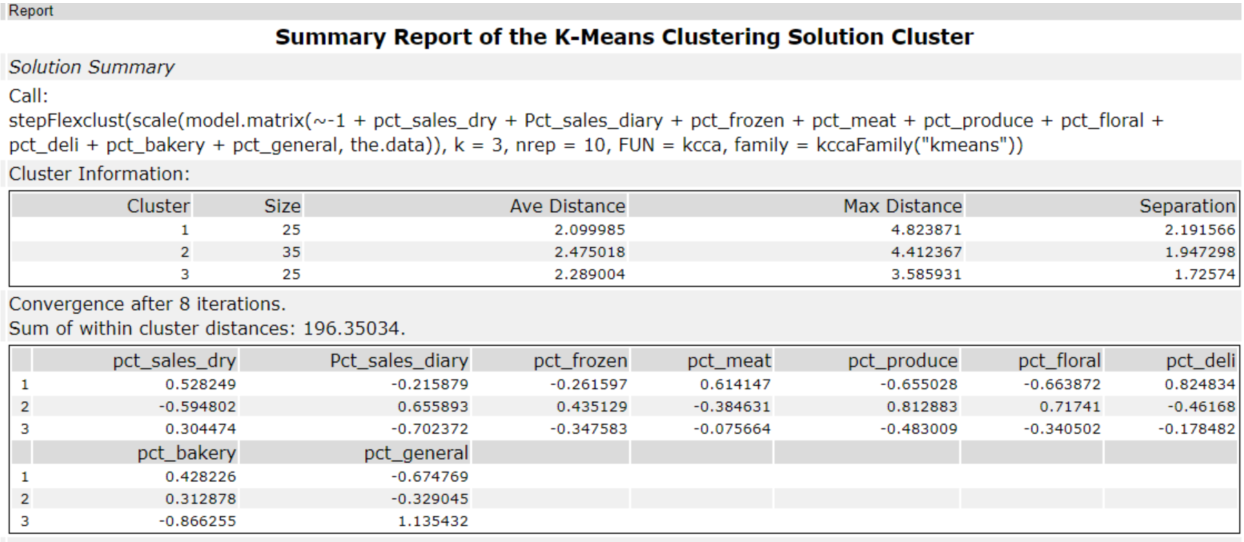
Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	10.76211	9.873885	10.92654	9.473174	9.008366	8.782612	8.038034
1st Quartile	14.39959	14.485851	12.93809	11.85382	10.989446	10.3112	9.971695
Median	16.87041	15.367318	13.75306	12.626898	11.760647	10.98475	10.635395
Mean	16.15535	15.148292	13.59071	12.511188	11.676508	10.929467	10.509103
3rd Quartile	17.85918	16.099744	14.33447	13.290279	12.439104	11.506069	11.064228
Maximum	19.04439	17.183364	15.32521	14.732171	13.839783	13.171298	12.326562

Plots

2. How many stores fall into each store format?

By running the K-Centroids Cluster Analysis tool as shown from the diagram below, Cluster 1 has 25 store, Cluster 2 has 35 stores and Clusters 3 has 25 stores



Report

Summary Report of the K-Means Clustering Solution Cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + pct_sales_dry + Pct_sales_diary + pct_frozen + pct_meat + pct_produce + pct_floral + pct_deli + pct_bakery + pct_general, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

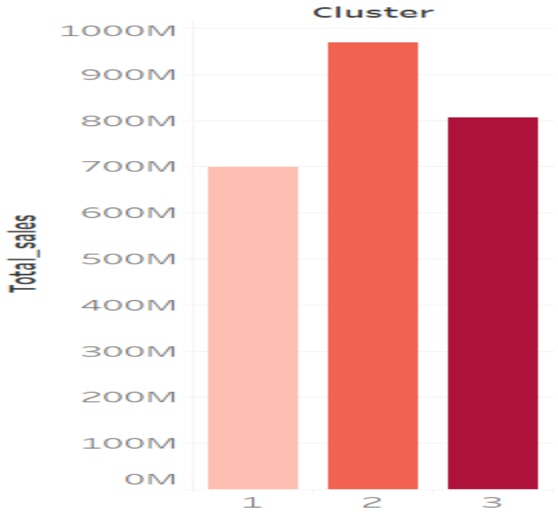
	pct_sales_dry	Pct_sales_diary	pct_frozen	pct_meat	pct_produce	pct_floral	pct_deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482
	pct_bakery	pct_general					
1	0.428226	-0.674769					
2	0.312878	-0.329045					
3	-0.866255	1.135432					

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

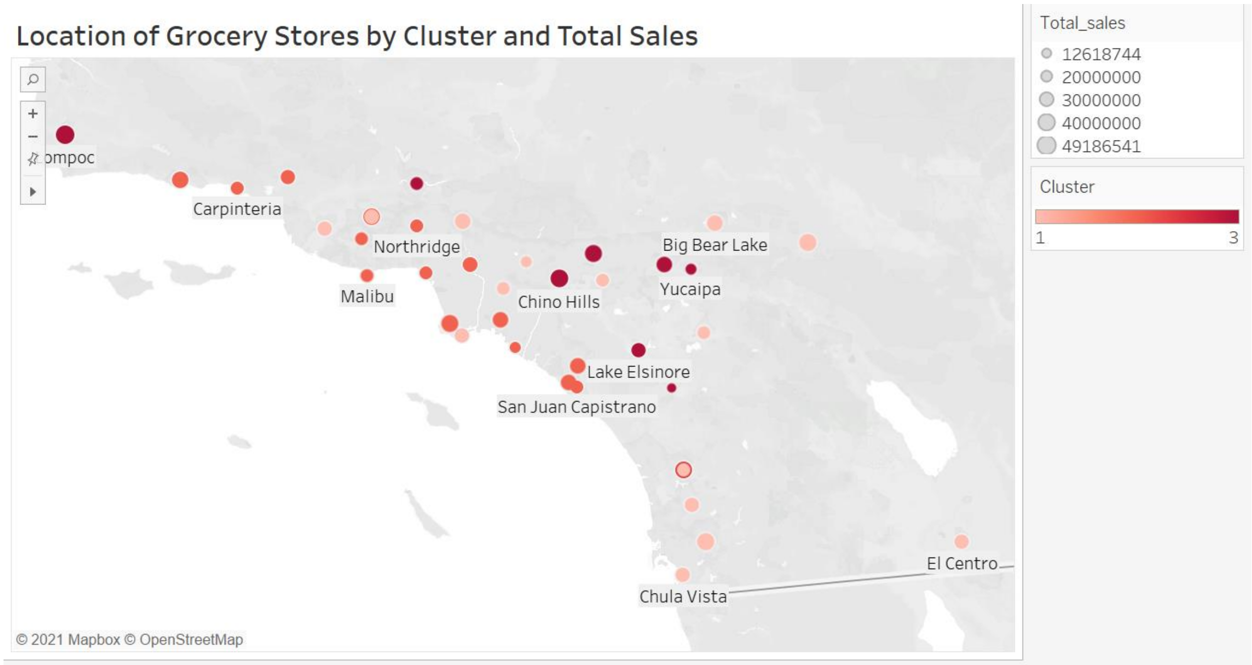
Cluster 1 – This cluster has 25 number of stores, least total sum of sales and the smallest average distance from the centroid which means it is the most compact and least variability of the 3 clusters.

Cluster 2 – This cluster has the largest number of stores 35, highest total sum of sales and the largest average distance from the centroid which means it is the least compact and high variability.

Cluster 3 – This cluster has 25 number of stores and also a moderately high average distance from the centroid.



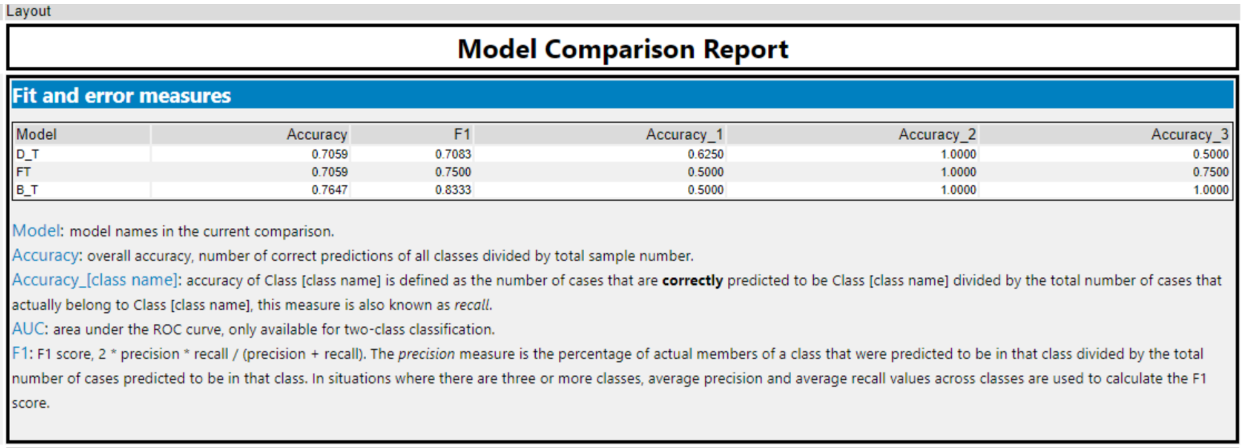
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.



https://public.tableau.com/profile/ndidi2972#!/vizhome/LocationofGroceryStoresbyClusterandTotalSalesinCalifornia/Sheet1?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.)



Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
D_T	0.7059	0.7083	0.6250	1.0000	0.5000
FT	0.7059	0.7500	0.5000	1.0000	0.7500
B_T	0.7647	0.8333	0.5000	1.0000	1.0000

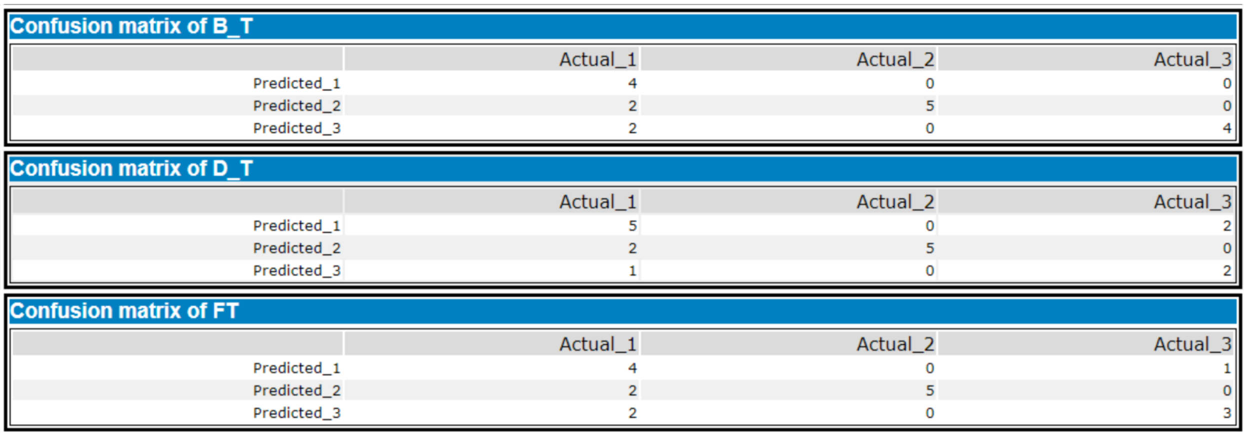
Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] 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.



Confusion matrix of B_T

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

Confusion matrix of D_T

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

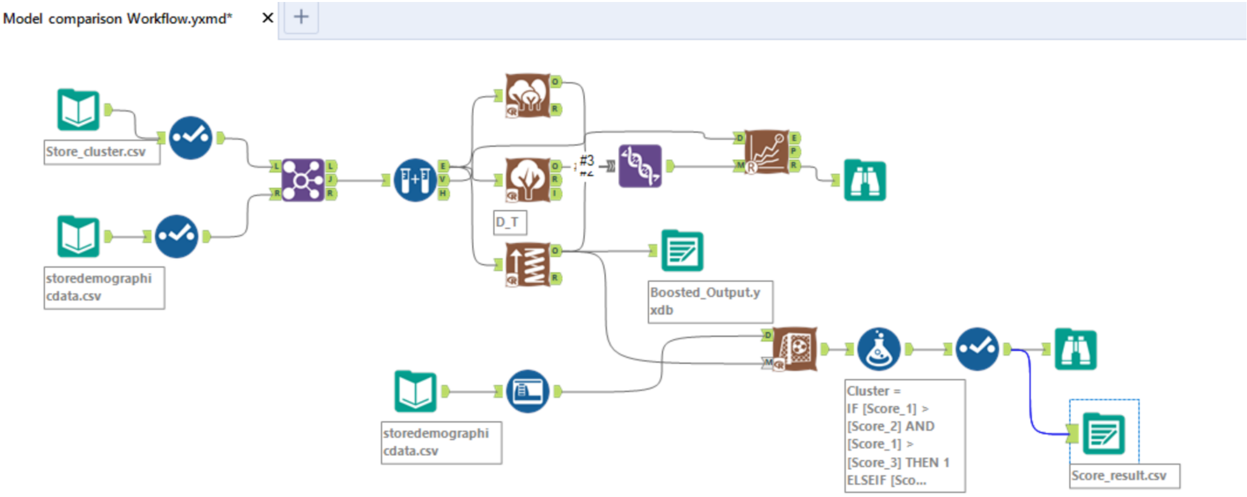
Confusion matrix of FT

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

Using the model comparison tool (as shown in the diagram above), I compared the results for Decision Tree, Forest Model and Boosted Model. Boosted Model is the best because it has a higher F1 score and the highest accuracy so I will use the Boosted Model to predict the best store format for the new stores.

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

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



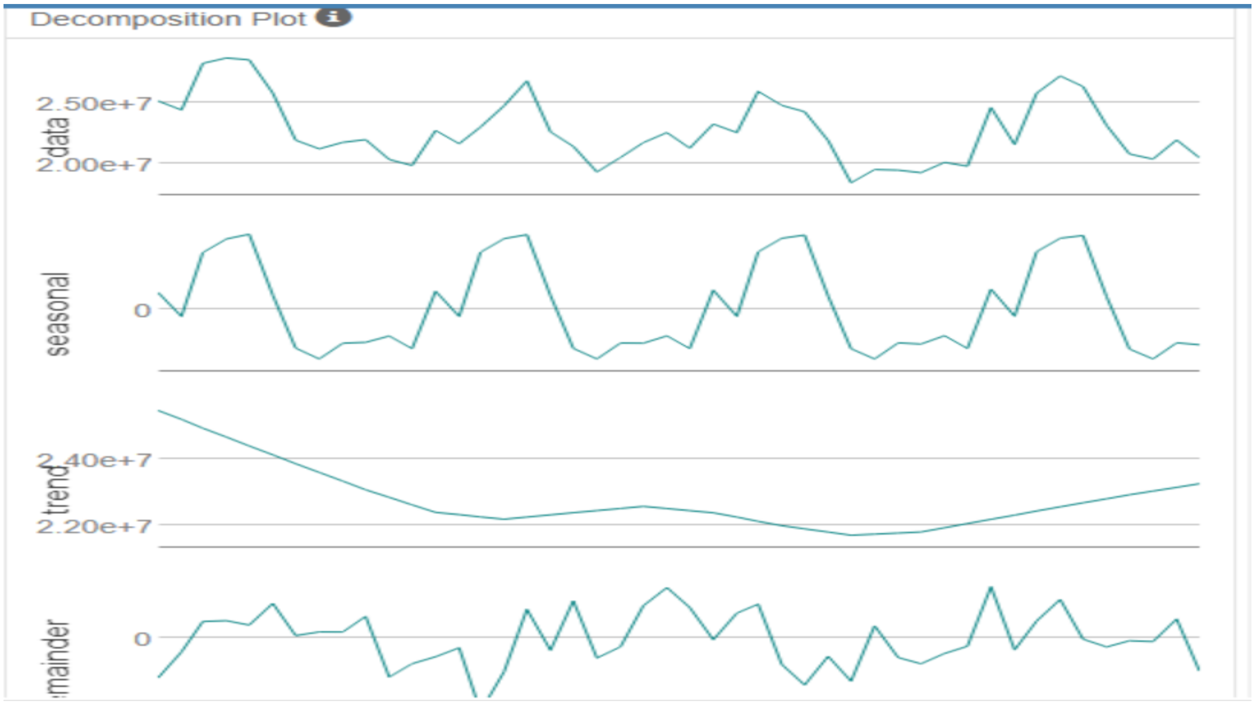
Task 3: Predicting Produce Sales

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 used (M,N,M) for the ETS model and (1,0,0)(1,1,0)12 for the ARIMA model.

ETS: from the Decomposition Plot, I observed the M, N, M pattern:

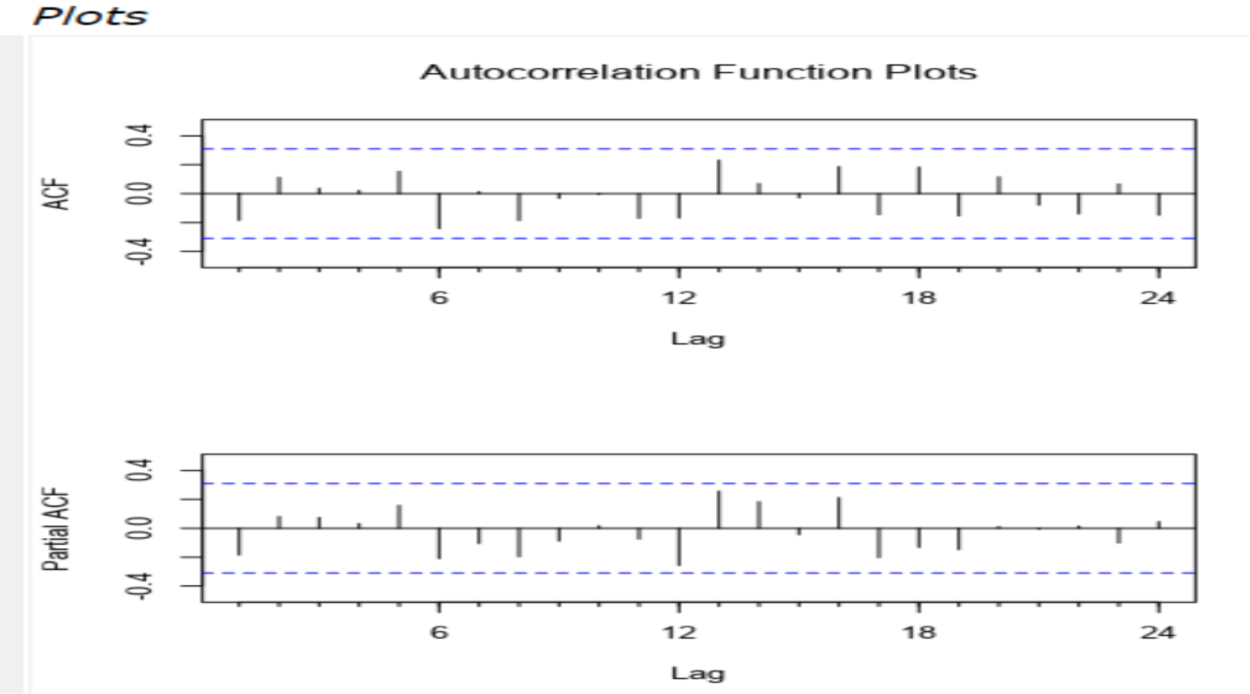
The Error plot shows variance along the years, this shows fluctuation with different sizes. Therefore, the error is multiplicatively (M.)
The Trend Plot shows the trend moves uptrend and downtrend. Therefore, the pattern is not clear and is neutral (N.)
The Seasonal plot shows peaks and valleys in similar periods of time, this suggests applying seasonality in a multiplicative method (M.)



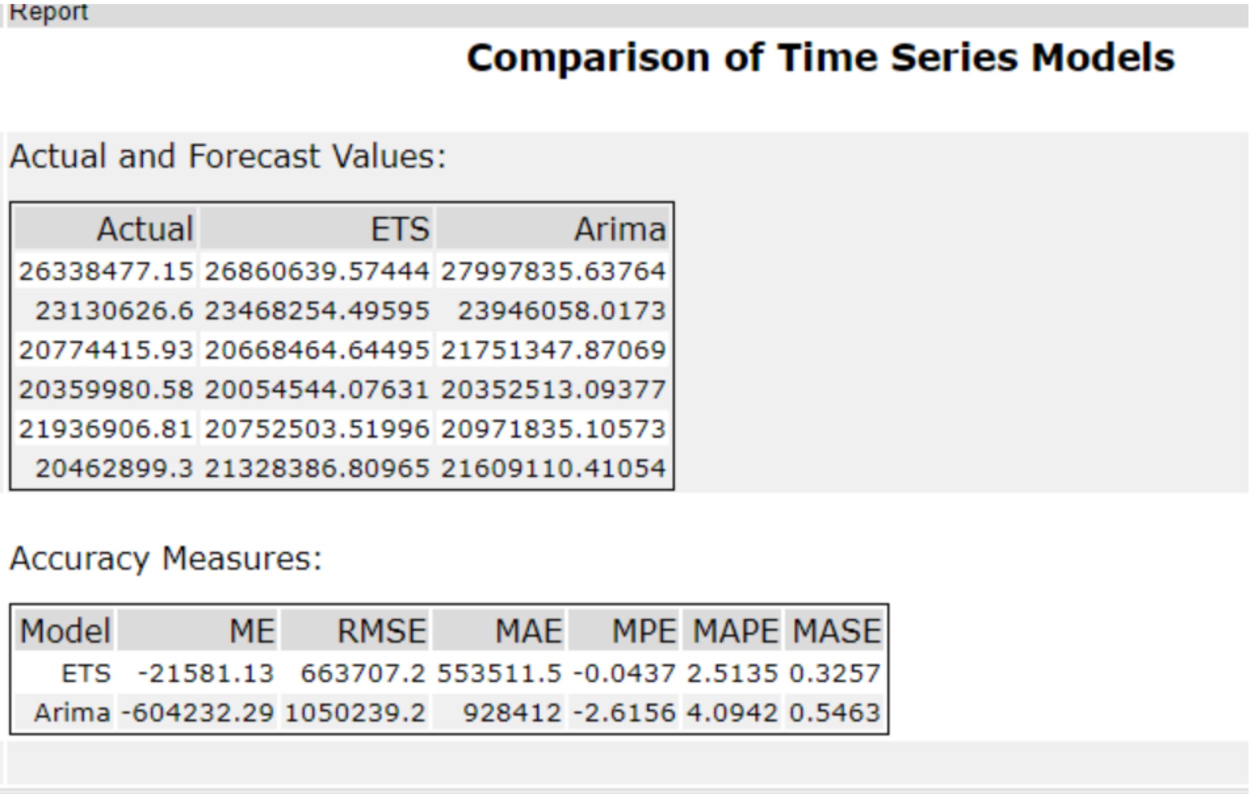
ARIMA:

The following charts show the ACF and PACF plots after applying the (1,0,0)(1,1,0)12 format to the ARIMA model.

Plots



From the diagram below: Using TS Compare tool against the holdout sample of ETS and ARIMA, ETS(M,N,M) turns out to be the better one.



Report

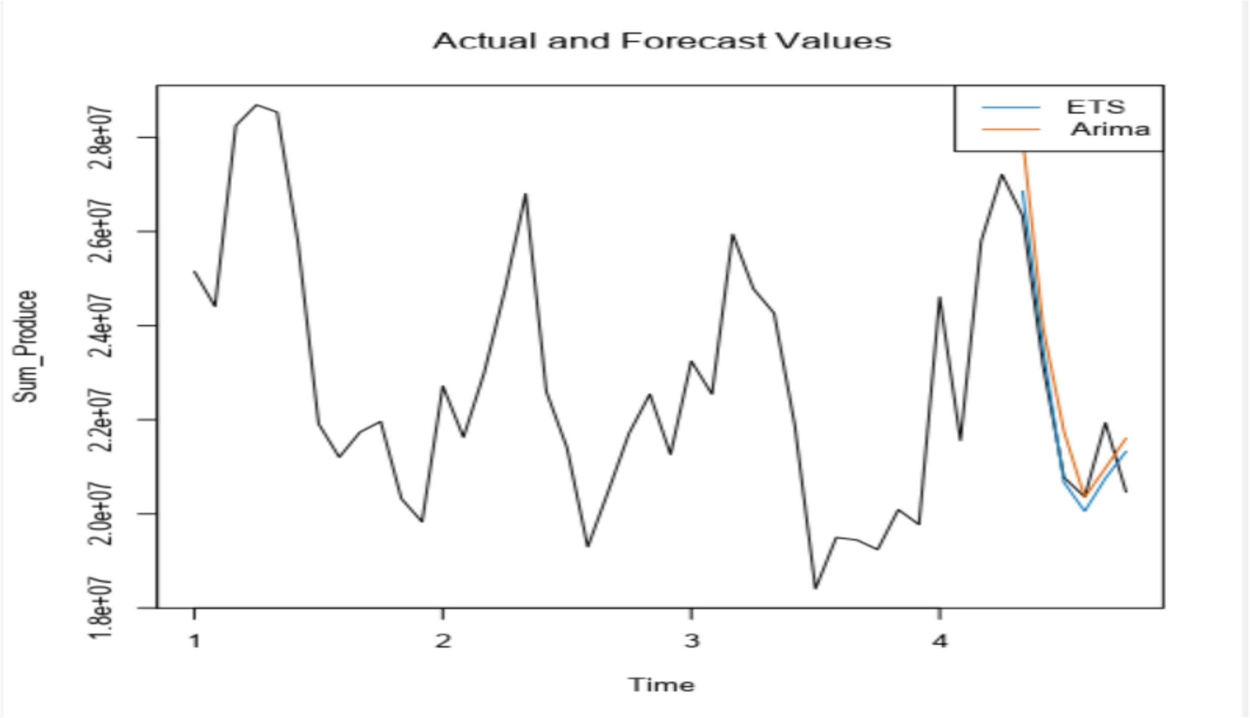
Comparison of Time Series Models

Actual and Forecast Values:

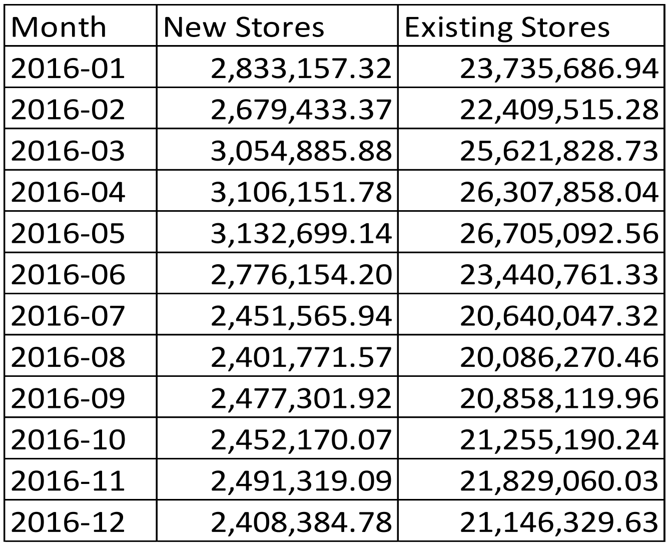
Actual	ETS	Arima
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
Arima	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

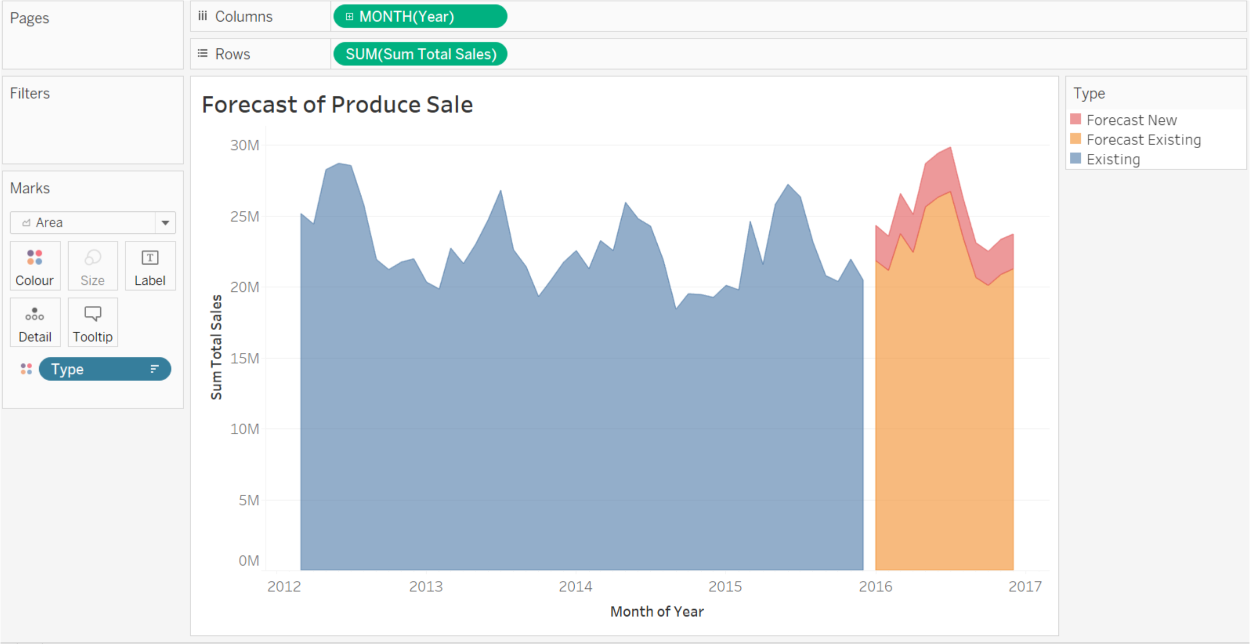


2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.



Month	New Stores	Existing Stores
2016-01	2,833,157.32	23,735,686.94
2016-02	2,679,433.37	22,409,515.28
2016-03	3,054,885.88	25,621,828.73
2016-04	3,106,151.78	26,307,858.04
2016-05	3,132,699.14	26,705,092.56
2016-06	2,776,154.20	23,440,761.33
2016-07	2,451,565.94	20,640,047.32
2016-08	2,401,771.57	20,086,270.46
2016-09	2,477,301.92	20,858,119.96
2016-10	2,452,170.07	21,255,190.24
2016-11	2,491,319.09	21,829,060.03
2016-12	2,408,384.78	21,146,329.63

https://public.tableau.com/profile/ndidi2972#!/vizhome/ForecastofProduceSale/Sheet1?publish=yes



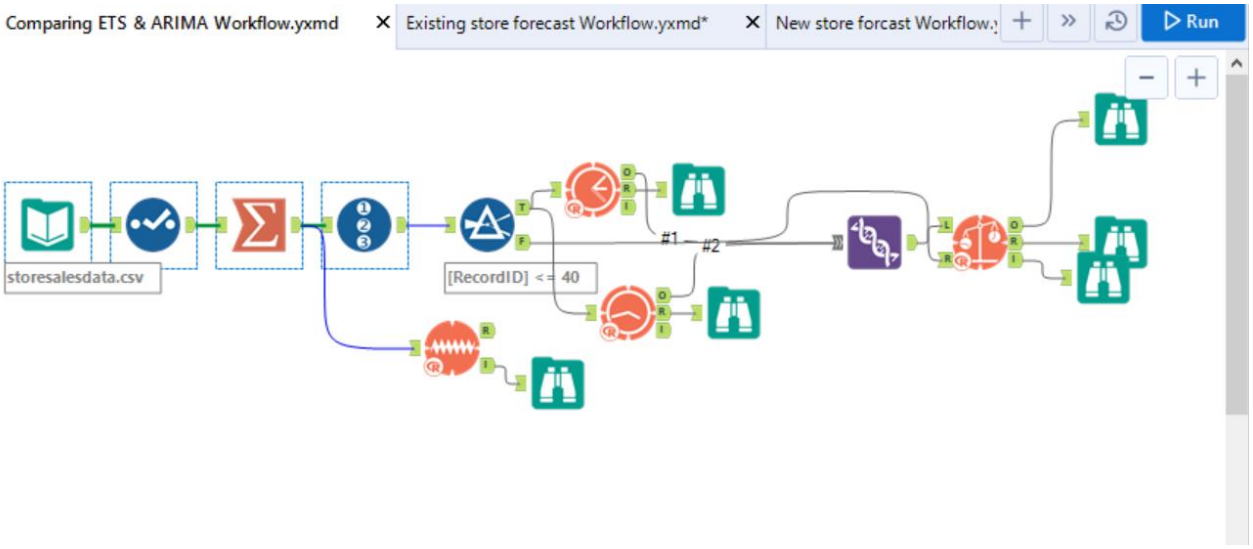
Columns: MONTH(Year)

Rows: SUM(Sum Total Sales)

Marks: Area

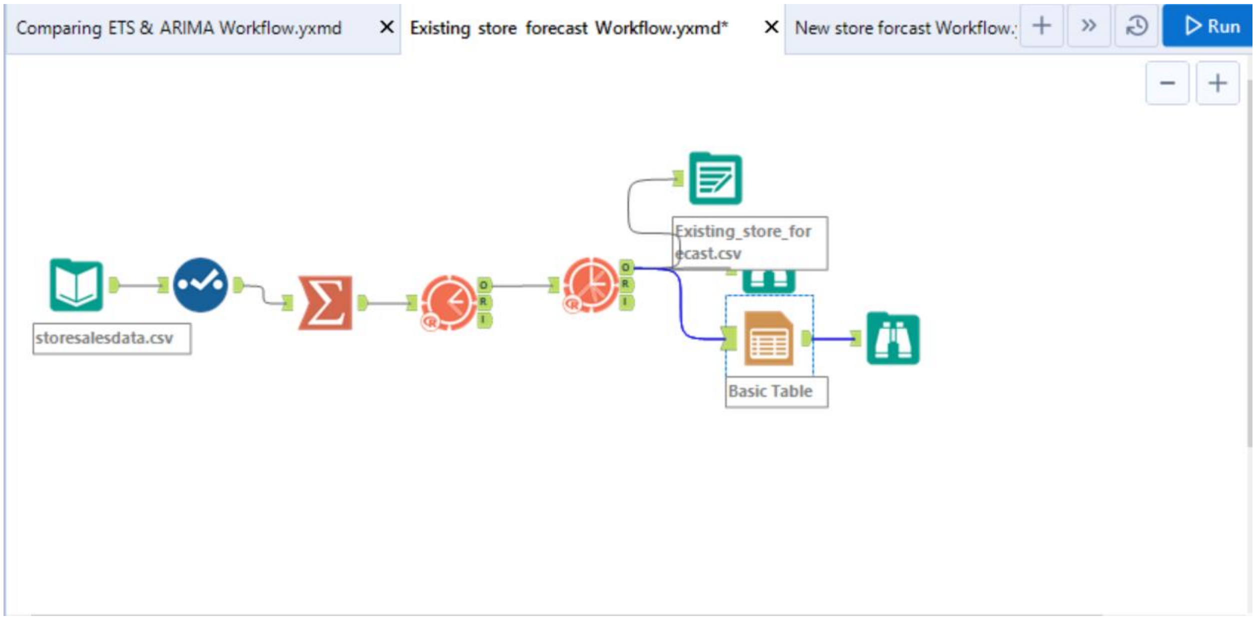
Colour: Type

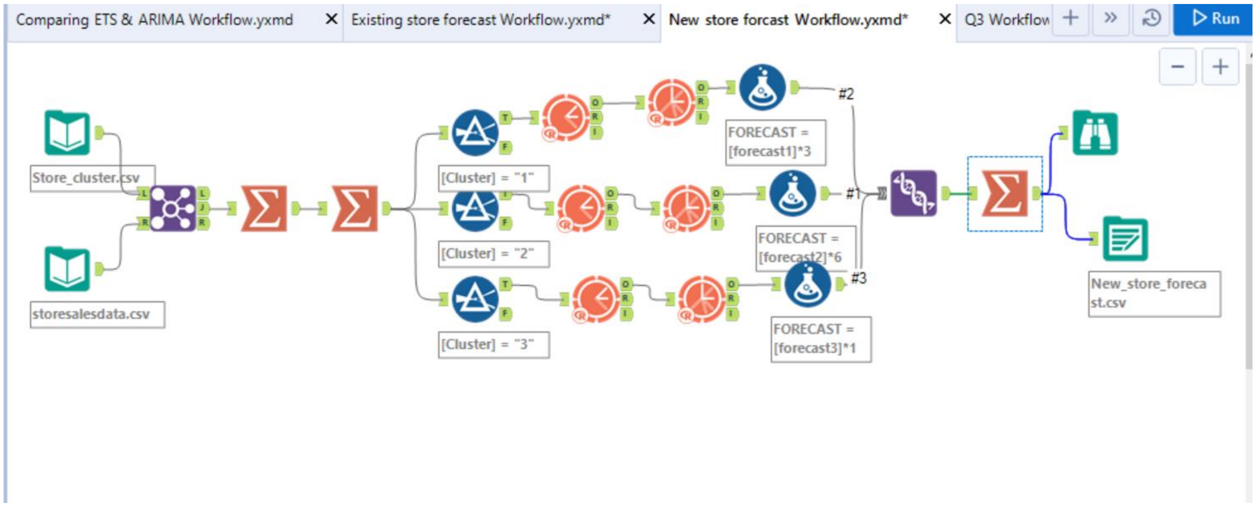
Below are the Alteryx workflow for Task 3

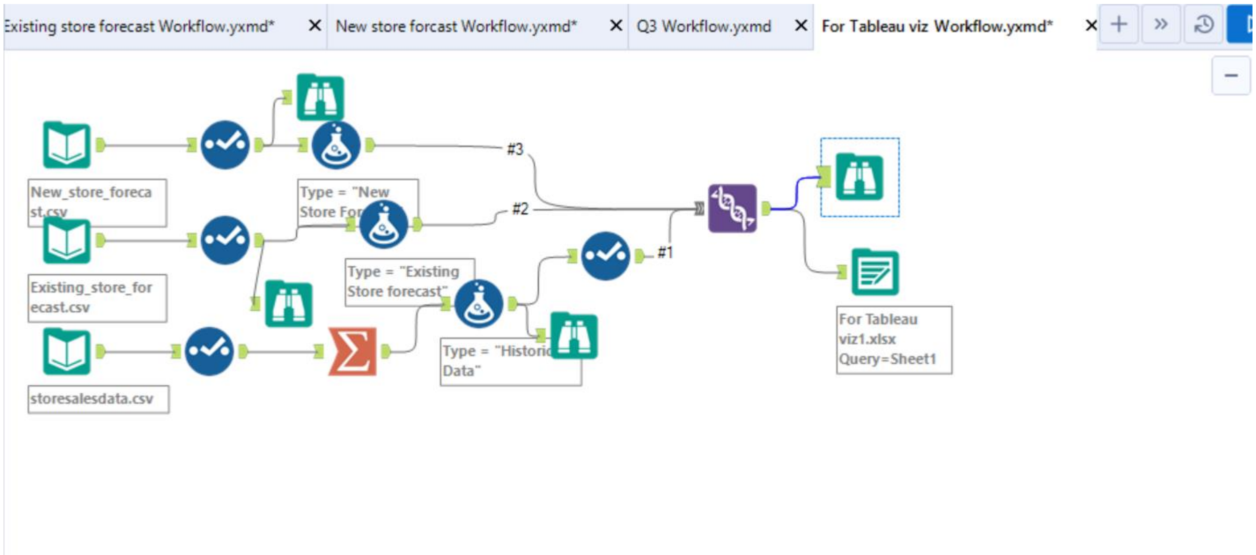


[RecordID] <= 40

Comparing ETS & ARIMA Workflow.yxmd
Existing store forecast Workflow.yxmd*
New store forcast Workflow.y
Run
storesalesdata.csv
T
F
O
R
I
#1
#2
L
R
O
R
I


Comparing ETS & ARIMA Workflow.yxmd
Existing store forecast Workflow.yxmd*
New store forcast Workflow.
Run
storesalesdata.csv
Existing_store_forecast.csv
Basic Table


Comparing ETS & ARIMA Workflow.yxmd
Existing store forecast Workflow.yxmd*
New store forcast Workflow.yxmd*
Q3 Workflow
Run
Store_cluster.csv
storesalesdata.csv
[Cluster] = "1"
[Cluster] = "2"
[Cluster] = "3"
FORECAST = [forecast1]*3
FORECAST = [forecast2]*6
FORECAST = [forecast3]*1
#1
#2
#3
New_store_forecast.csv


Existing store forecast Workflow.yxmd*
New store forcast Workflow.yxmd*
Q3 Workflow.yxmd
For Tableau viz Workflow.yxmd*
New_store_forecast.csv
Existing_store_forecast.csv
storesalesdata.csv
Type = "New Store For
Type = "Existing Store forecast"
Type = "Historic Data"
#1
#2
#3
For Tableau viz1.xlsx Query=Sheet1