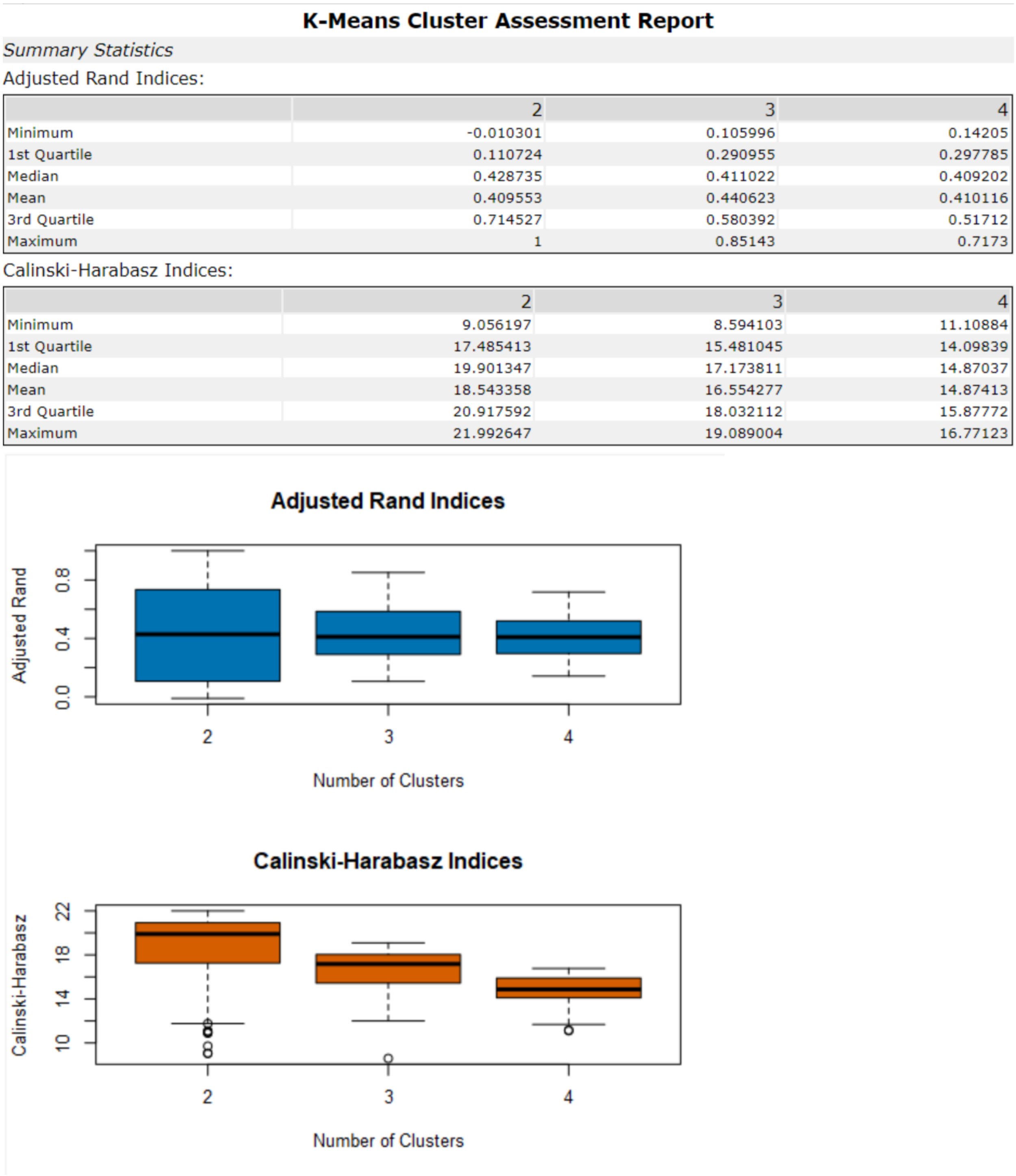
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Store

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



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

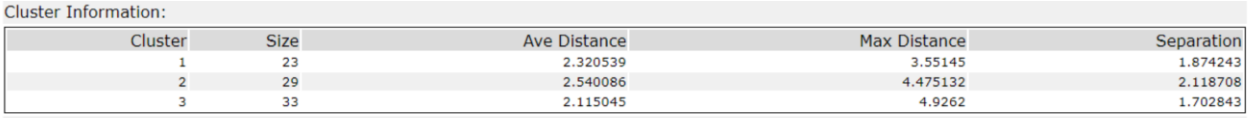
	2	3	4
Minimum	-0.010301	0.105996	0.14205
1st Quartile	0.110724	0.290955	0.297785
Median	0.428735	0.411022	0.409202
Mean	0.409553	0.440623	0.410116
3rd Quartile	0.714527	0.580392	0.51712
Maximum	1	0.85143	0.7173

Calinski-Harabasz Indices:

	2	3	4
Minimum	9.056197	8.594103	11.10884
1st Quartile	17.485413	15.481045	14.09839
Median	19.901347	17.173811	14.87037
Mean	18.543358	16.554277	14.87413
3rd Quartile	20.917592	18.032112	15.87772
Maximum	21.992647	19.089004	16.77123

Based on the, Adjusted Rand and Calinski-Harabasz indices of the K-means Cluster Assessment report, the optimal number of store formats is 3 since both indices have fairly high medians with small spreads (the interquartile range is compact).

2. How many stores fall into each store format?

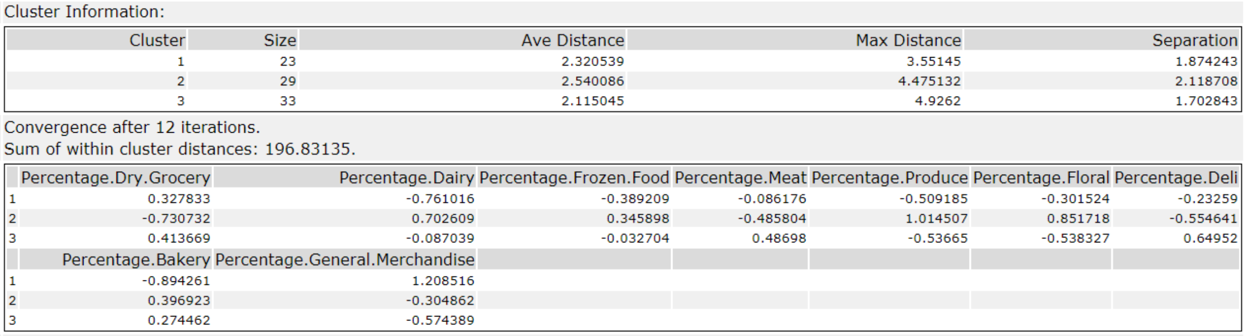


Cluster Information:

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

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

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



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

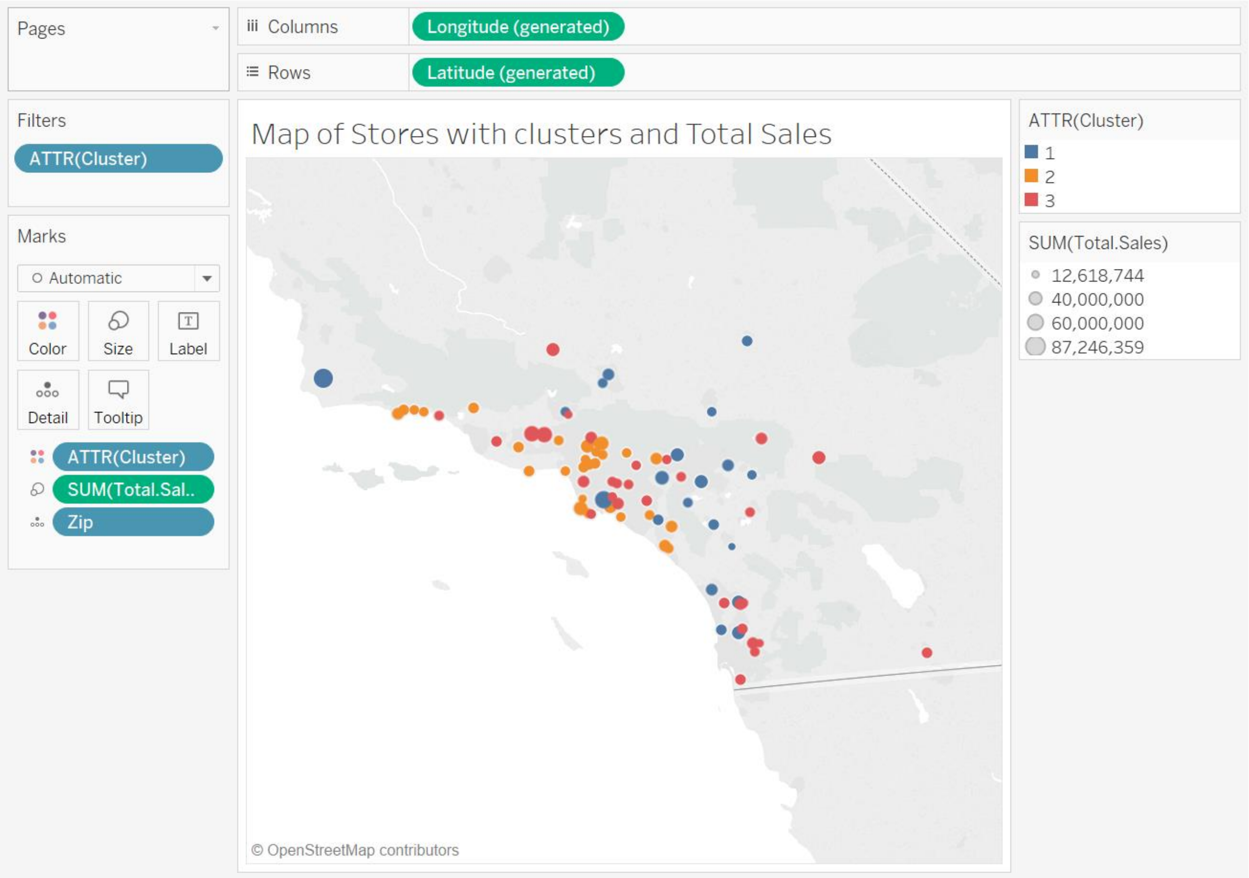
Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Percentage.Dry.Grocery	Percentage.Dairy	Percentage.Frozen.Food	Percentage.Meat	Percentage.Produce	Percentage.Floral	Percentage.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

	Percentage.Bakery	Percentage.General.Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

While Cluster 2 has the highest average distance, which are less compact and might show more variability, Cluster 3 has the smallest average distance which are the most compact of the clusters. Cluster 1 has the highest total sales for General Merchandise in terms of percentage while Cluster 2 has the highest total sales for Produce.

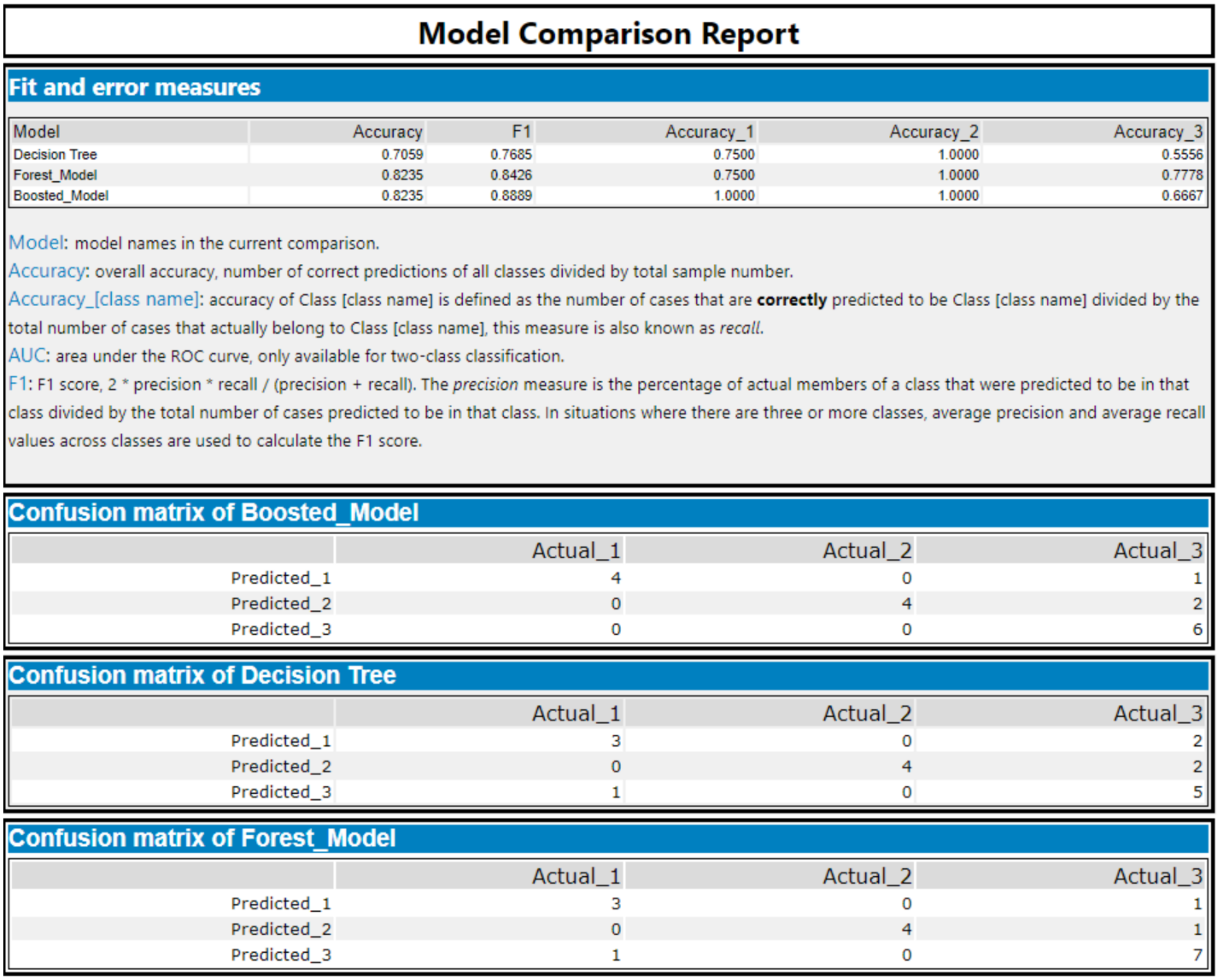
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.



https://public.tableau.com/profile/dolly.yu#!/vizhome/FinalProject_710/Map

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.)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Boosted_Model	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 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 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

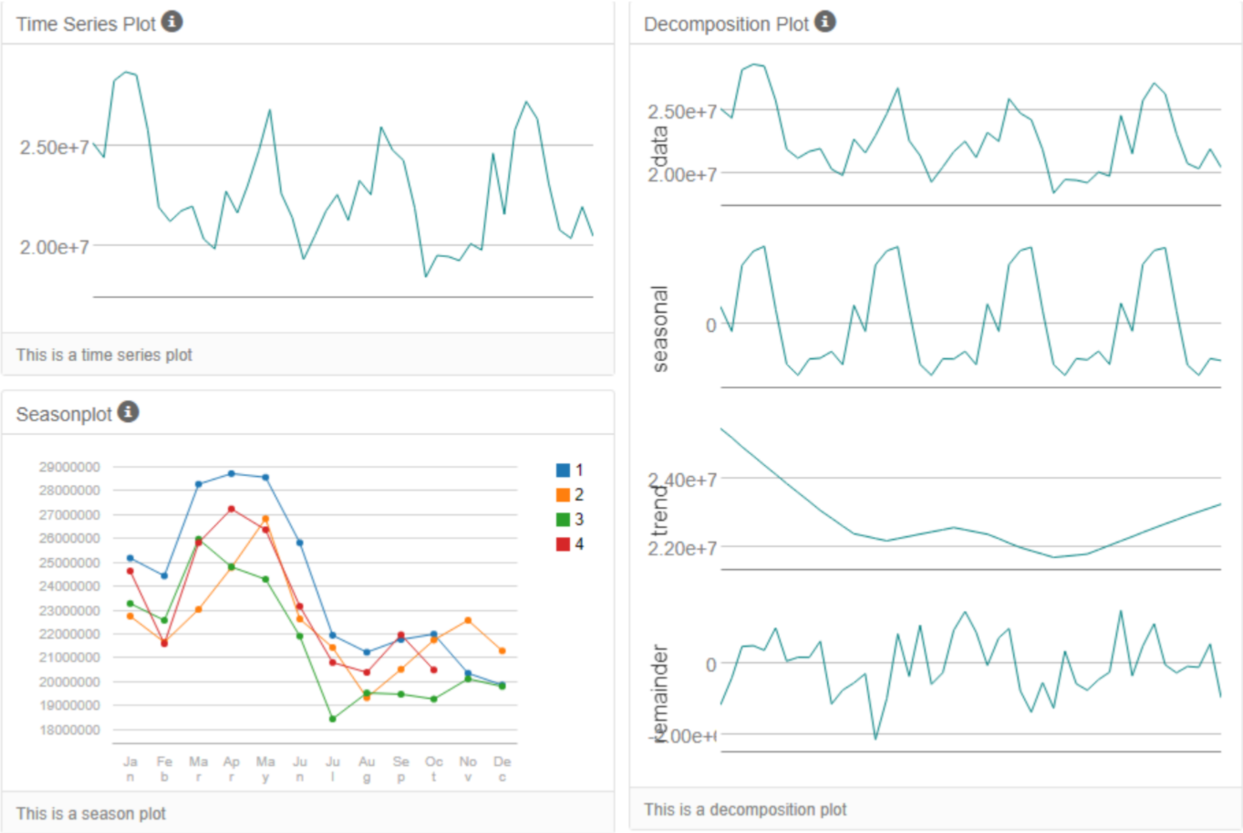
The model comparison report shows comparison between Decision Tree, Forest Model and Boosted Model. **Boosted Mode**l is chosen since it has the highest accuracy with a higher F1 value than the Forest Model.

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?



Time Series Plot

This is a time series plot

Seasonplot

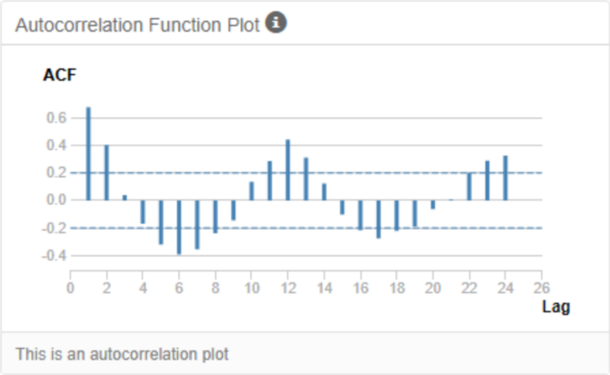
This is a season plot

Decomposition Plot

This is a decomposition plot

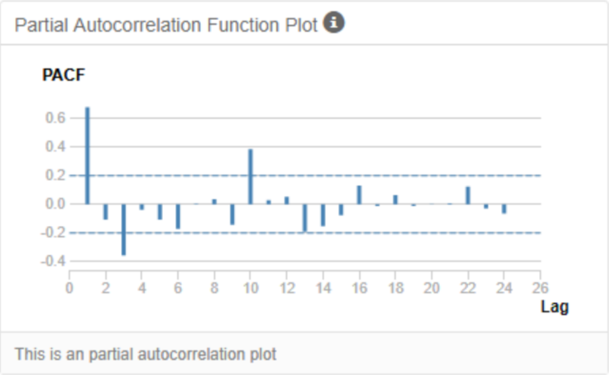
Based on the time series decomposition plot above, ETS(M,N,M) with no dampening is chosen. For Error, we see the remainder plot fluctuating between large and small errors over time, so we apply multiplicatively (M). For Trend, there is no clear trend, so no trend component is included (N). For Seasonal, size of the seasonal fluctuations tends to increase with the level of time series, so we apply it multiplicatively (M).

Original time series plot without differencing



Autocorrelation Function Plot

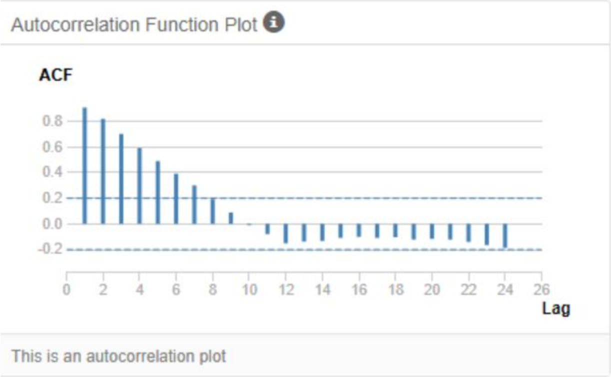
This is an autocorrelation plot



Partial Autocorrelation Function Plot

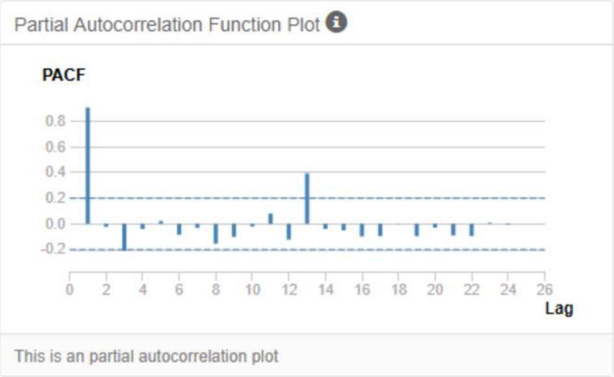
This is an partial autocorrelation plot

Seasonal difference



Autocorrelation Function Plot

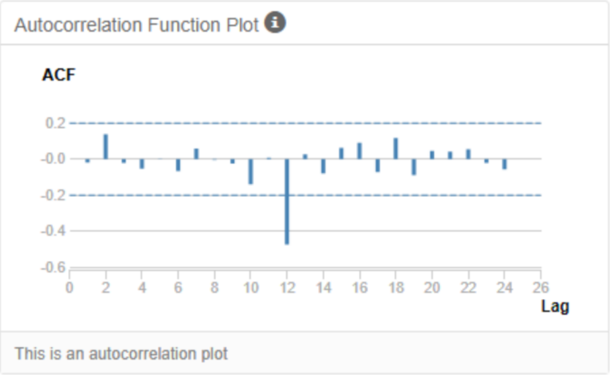
This is an autocorrelation plot



Partial Autocorrelation Function Plot

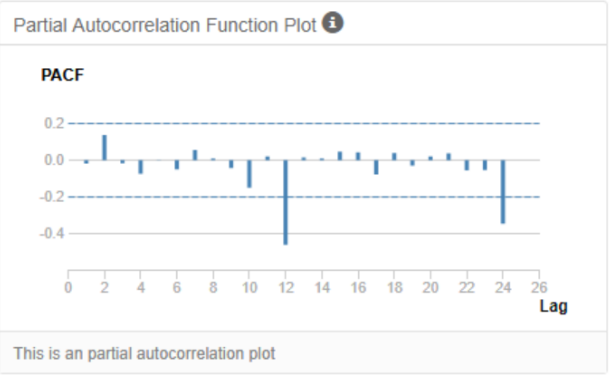
This is an partial autocorrelation plot

Seasonal first difference



Autocorrelation Function Plot

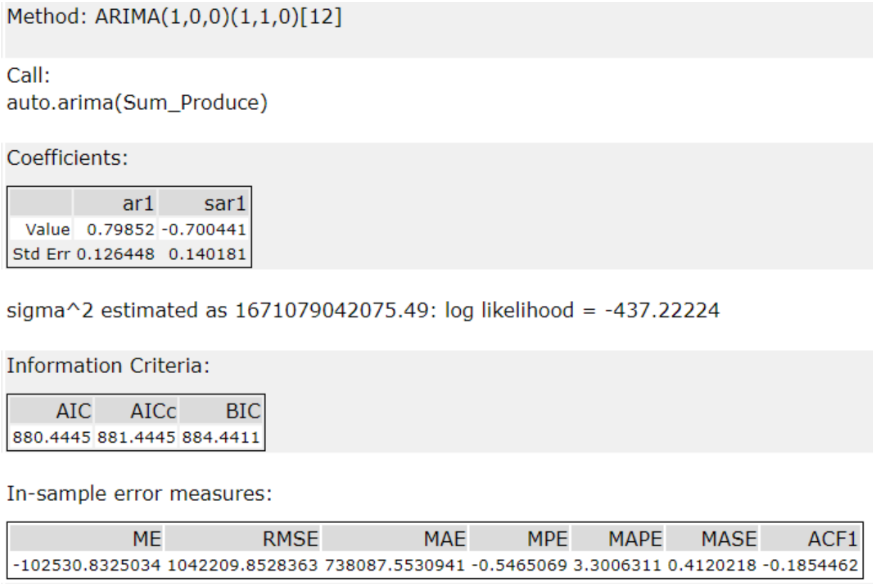
This is an autocorrelation plot



Partial Autocorrelation Function Plot

This is an partial autocorrelation plot

For the ARIMA model, the model is set to calculate automatically and **ARIMA(1,0,0)(1,1,0)12** is used



Method: ARIMA(1,0,0)(1,1,0)[12]

Call:
auto.arima(Sum_Produce)

Coefficients:

	ar1	sar1
Value	0.79852	-0.700441
Std Err	0.126448	0.140181

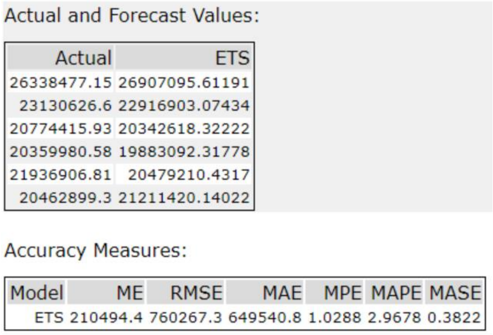
sigma^2 estimated as 1671079042075.49: log likelihood = -437.22224

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

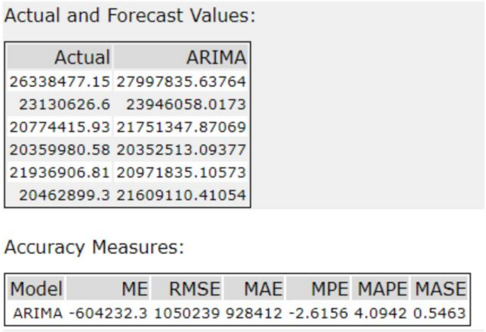


Actual and Forecast Values:

Actual	ETS
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822



Actual and Forecast Values:

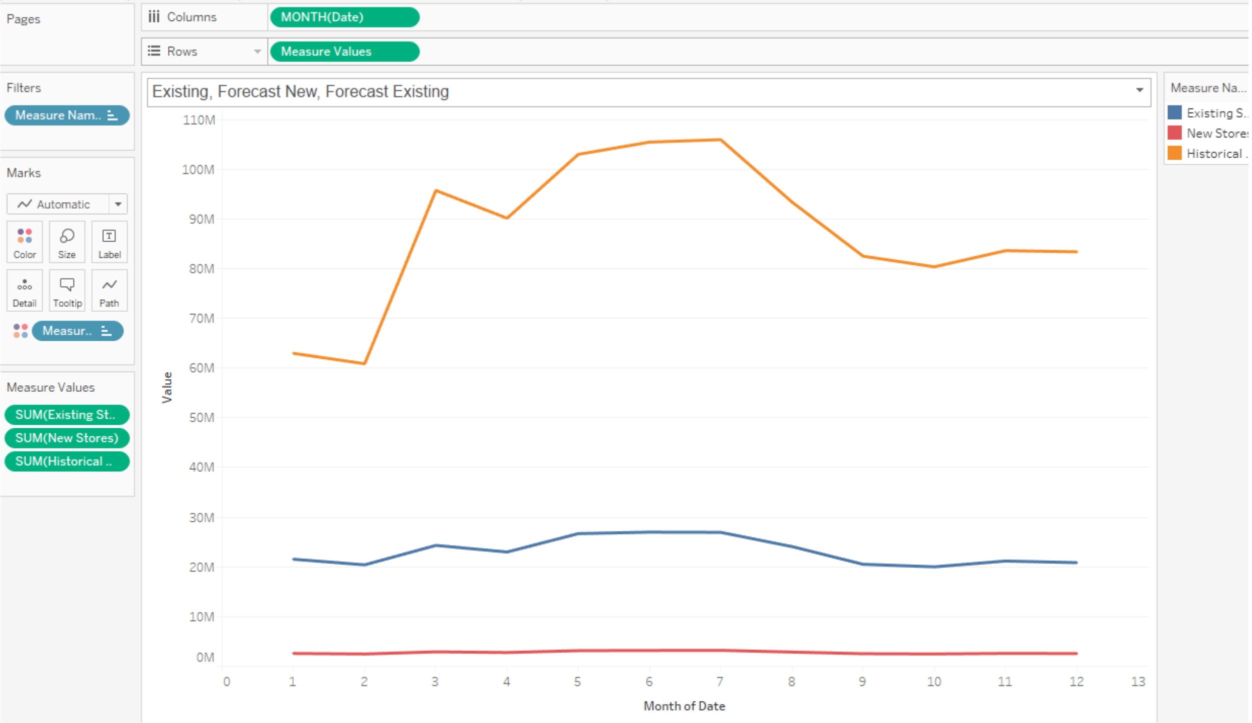
Actual	ARIMA
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239	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
1	2587450.851495	21539936.007499
2	2477352.892393	20413770.60136
3	2913185.23625	24325953.097628
4	2775745.609767	22993466.348585
5	3150866.835326	26691951.419156
6	3188922.00336	26989964.010552
7	3214745.646251	26948630.764764
8	2866348.663392	24091579.349106
9	2538726.84886	20523492.408643
10	2488148.287462	20011748.6686
11	2595270.386448	21177435.485838
12	2573396.62905	20855799.10961



Pages

Columns: MONTH(Date)

Rows: Measure Values

Filters: Measure Nam..

Marks: Automatic

Measur..

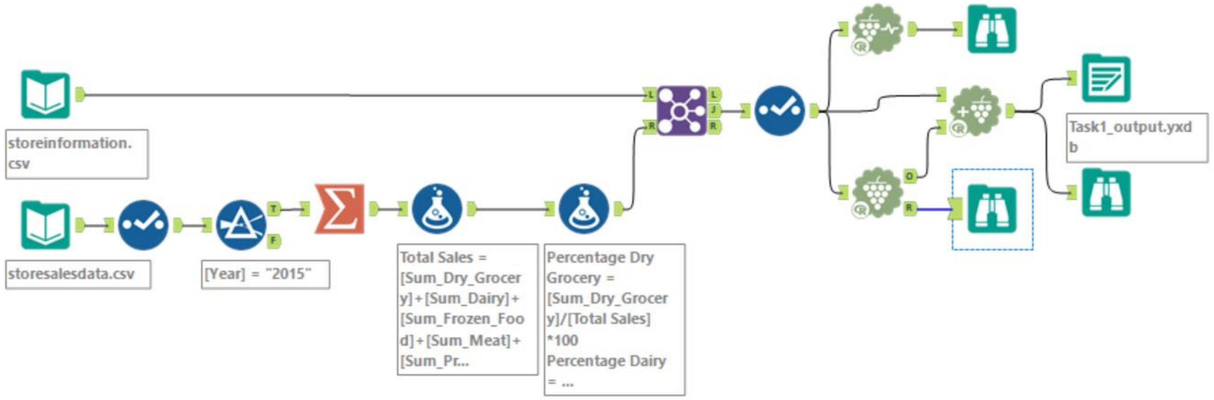
Measure Values: SUM(Existing St.. | SUM(New Stores) | SUM(Historical ..

Existing, Forecast New, Forecast Existing

https://public.tableau.com/profile/dolly.yu#!/vizhome/FinalProjectTask3Forecast/Sheet1?publish=yes

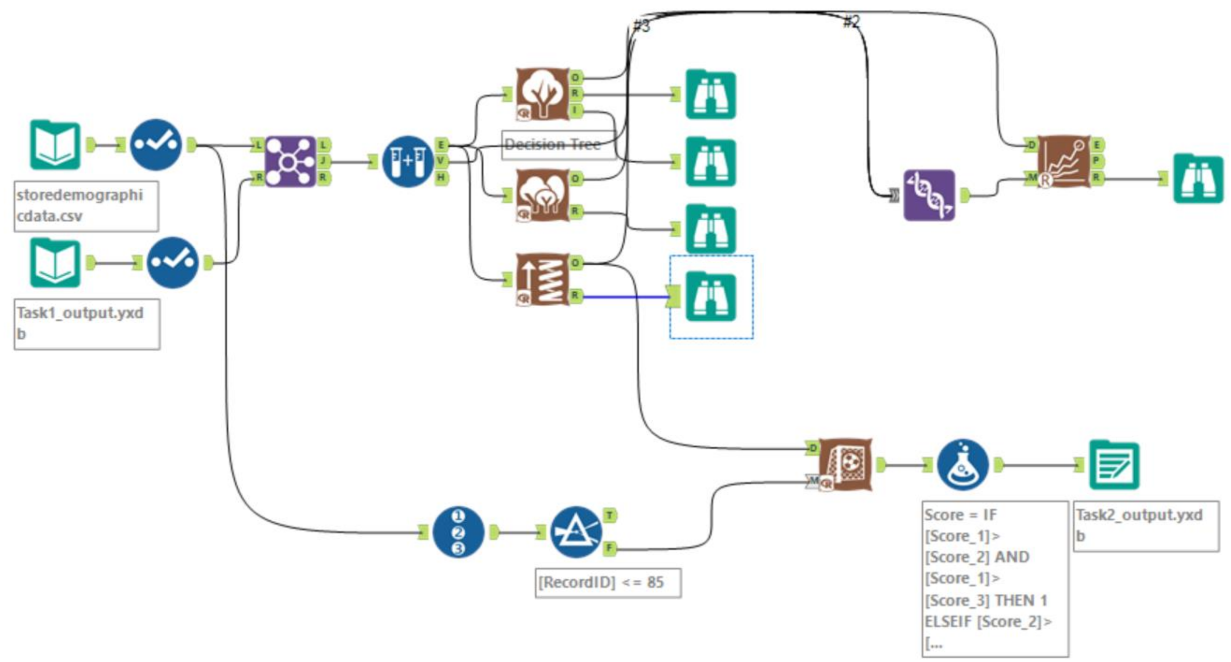
Alteryx Workflow

Task 1: Store Format

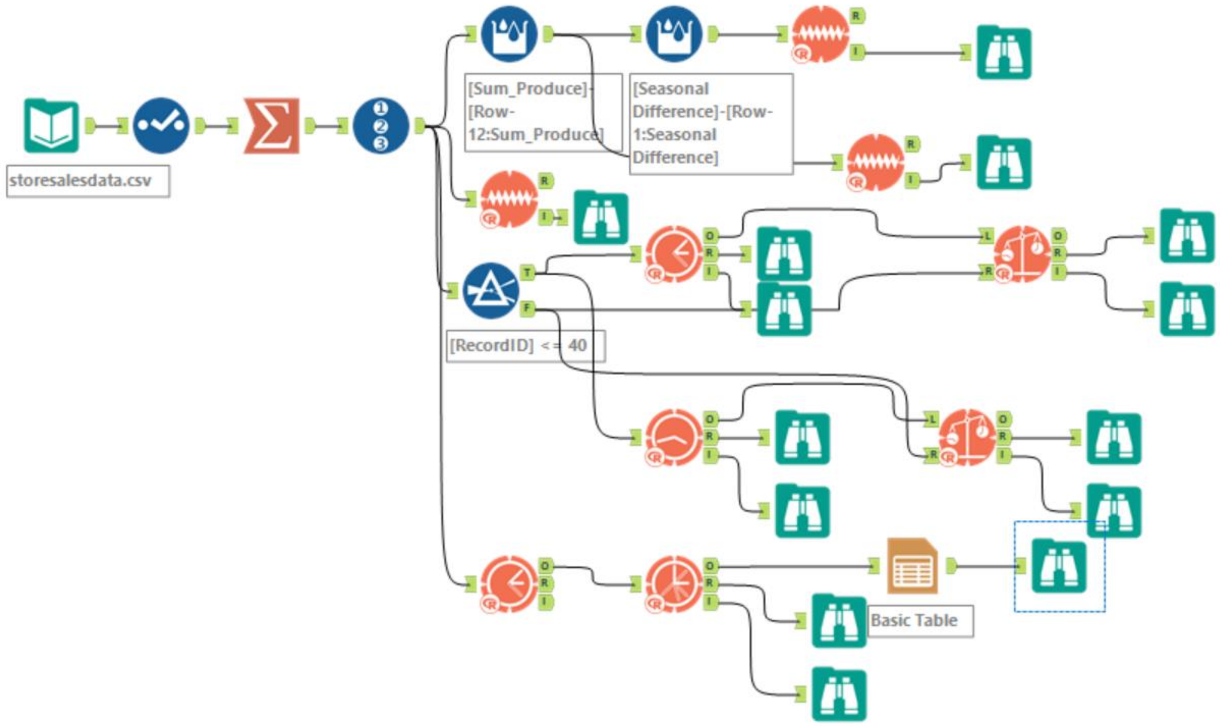


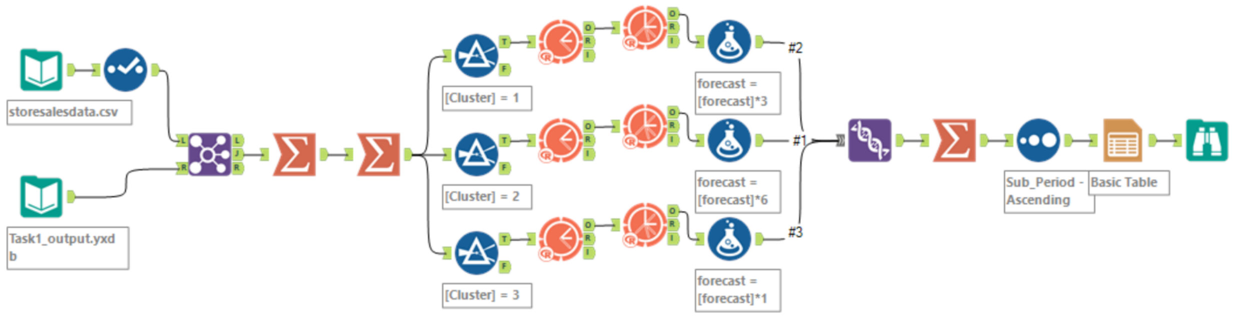
storeinformation.csv
storesalesdata.csv
[Year] = "2015"
T
F
Total Sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...
Percentage Dry Grocery = [Sum_Dry_Grocery]/[Total Sales]*100 Percentage Dairy = ...
L
J
R
O
Task1_output.yxdb

Task 2: New Stores



Task 3: Forecasting


storesalesdata.csv
[Sum_Produce]-[Row-12:Sum_Produce]
[Seasonal Difference]-[Row-1:Seasonal Difference]
[RecordID] <= 40
Basic Table


storesalesdata.csv
Task1_output.yxdb
[Cluster] = 1
[Cluster] = 2
[Cluster] = 3
forecast = [forecast]*3
forecast = [forecast]*6
forecast = [forecast]*1
#2
#1
#3
Sub_Period - Ascending
Basic Table