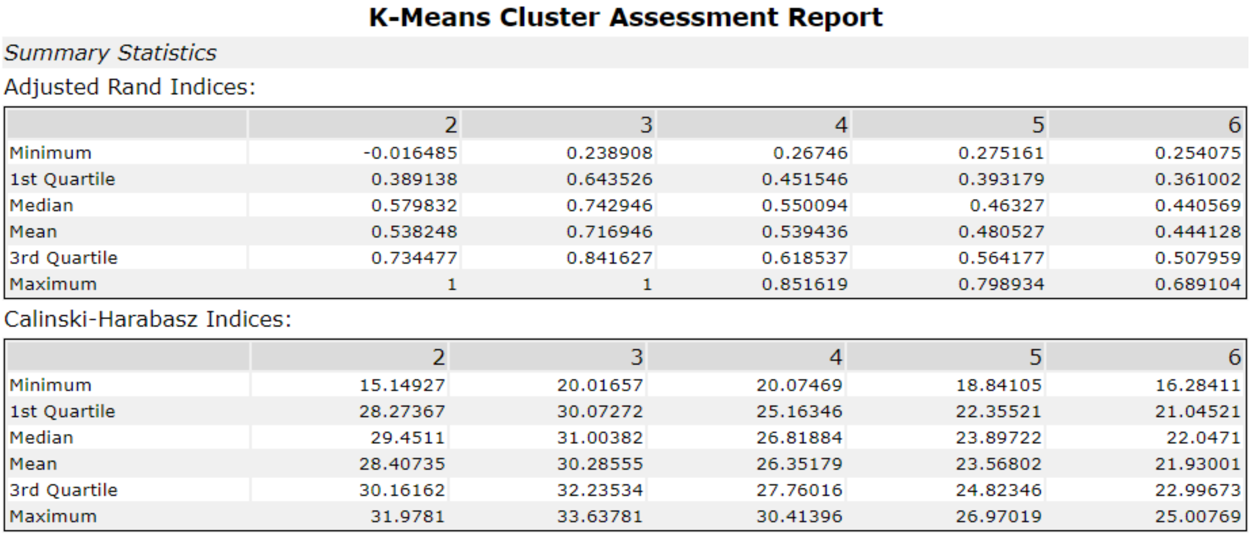
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

Thanks to my analysis, I have recognised that the optimal format number is 3. In the assessment report, we can observe that the median is highest for position number 3 for both Adjusted Rand Indices and Calinski-Harabasz Indices.



K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.016485	0.238908	0.26746	0.275161	0.254075
1st Quartile	0.389138	0.643526	0.451546	0.393179	0.361002
Median	0.579832	0.742946	0.550094	0.46327	0.440569
Mean	0.538248	0.716946	0.539436	0.480527	0.444128
3rd Quartile	0.734477	0.841627	0.618537	0.564177	0.507959
Maximum	1	1	0.851619	0.798934	0.689104

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.14927	20.01657	20.07469	18.84105	16.28411
1st Quartile	28.27367	30.07272	25.16346	22.35521	21.04521
Median	29.4511	31.00382	26.81884	23.89722	22.0471
Mean	28.40735	30.28555	26.35179	23.56802	21.93001
3rd Quartile	30.16162	32.23534	27.76016	24.82346	22.99673
Maximum	31.9781	33.63781	30.41396	26.97019	25.00769

Figure 1. K-Means Cluster Report

From the two plots below, we observe that the Compactness and distinctness have the best value for the number of clusters equal to 3.

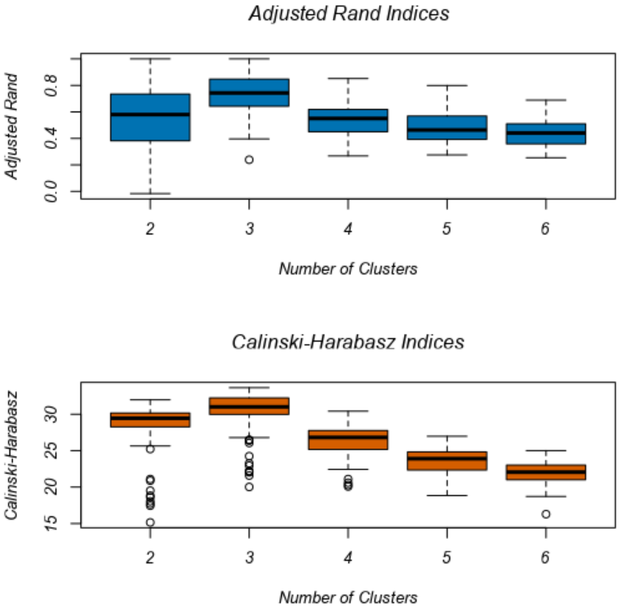
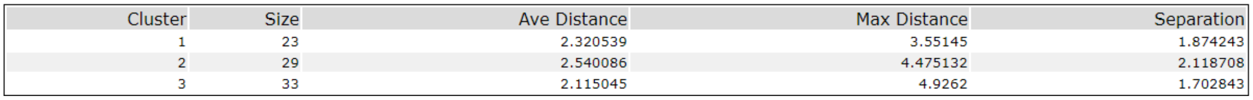


Figure 2. Adjusted Rand Indices and Calinski-Harabasz Indices

2. How many stores fall into each store format?

The number of stores per cluster is displayed below.



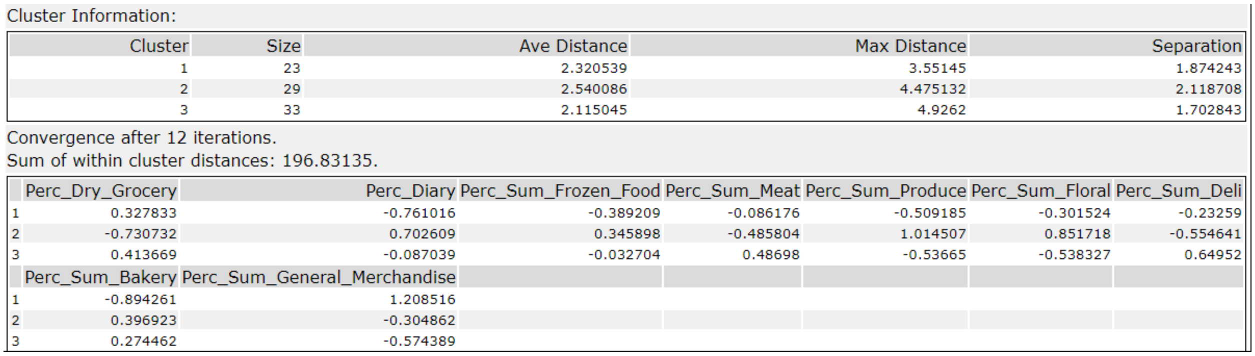
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Figure 3. Cluster Information

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

From the K-Centroids cluster analysis how each cluster is build. The more positive number the more sales for this particular product.

- For Cluster 1 the driver is: General Merchandise
- For Cluster 2 the driver is: Production
- For Cluster 3 the driver is: Meat and Deli



Cluster Information:

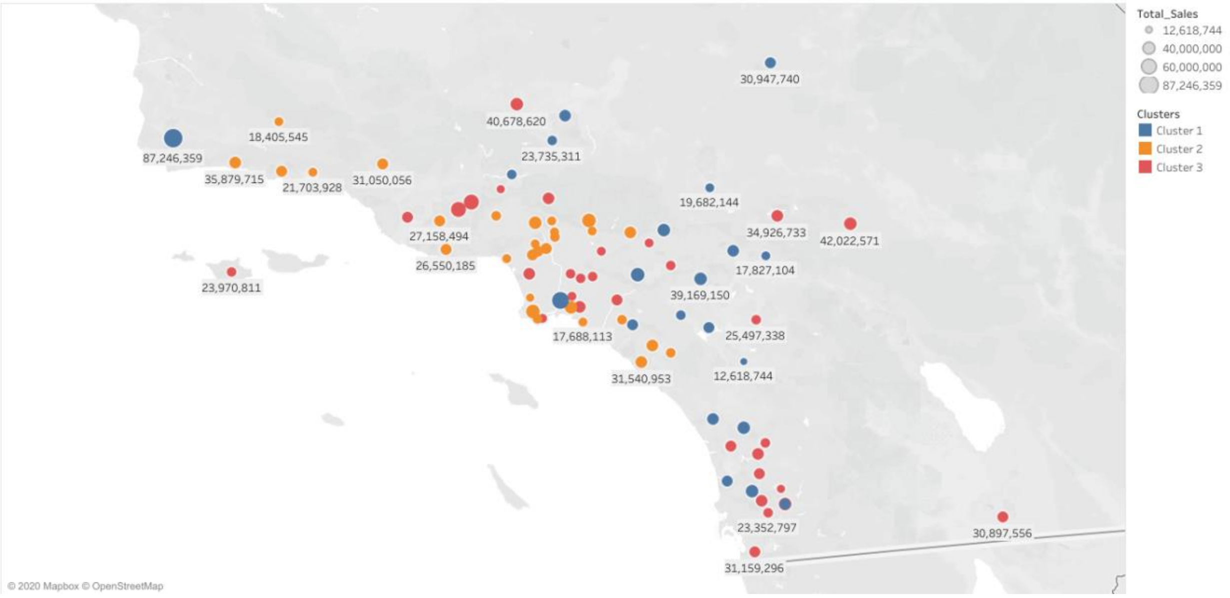
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

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

	Perc_Dry_Grocery	Perc_Diary	Perc_Sum_Frozen_Food	Perc_Sum_Meat	Perc_Sum_Produce	Perc_Sum_Floral	Perc_Sum_Deli	Perc_Sum_Bakery	Perc_Sum_General_Merchandise
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259	-0.894261	1.208516
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641	0.396923	-0.304862
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952	0.274462	-0.574389

Figure 4. Alteryx K-Centroids Cluster Analysis Result

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.

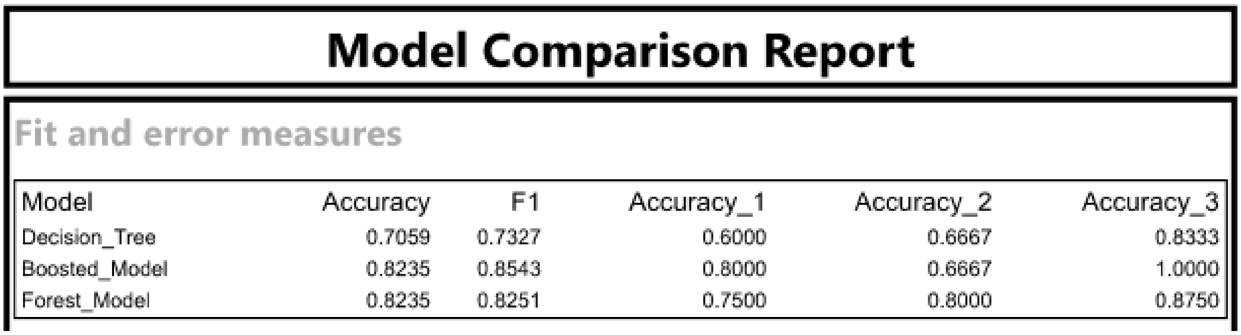


Map 1. Location of the Stores

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.)

The report comparison tool shows the same accuracy for both forest and boosted model. Looking at the F1 measure, we can see it have a slightly higher value than other models. That is why I have decided to use the boosted model.



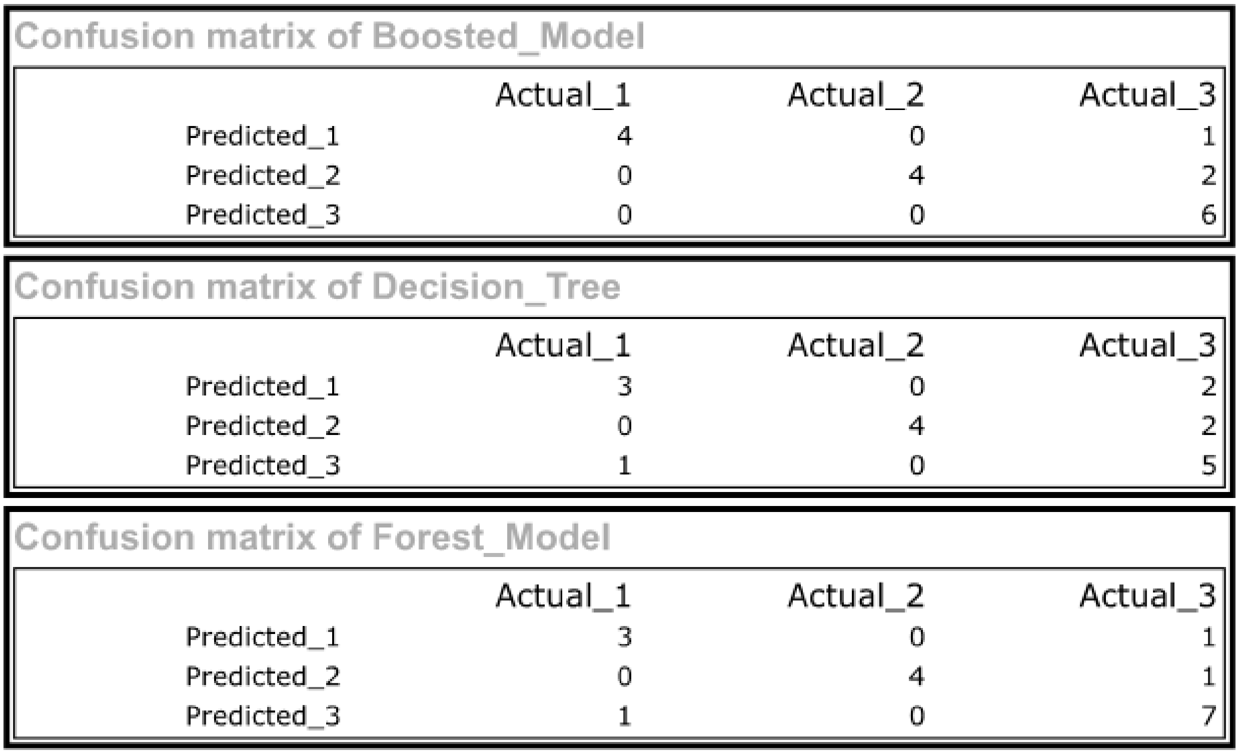
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7327	0.6000	0.6667	0.8333
Boosted_Model	0.8235	0.8543	0.8000	0.6667	1.0000
Forest_Model	0.8235	0.8251	0.7500	0.8000	0.8750

Figure 5. Model Comparison Tool

Using the confusion matrix, we can also observe where the models have been correct and where they didn't predict cluster accurately. From the tables below, we can see that the boosted model predicted cluster number 1 and cluster number 2 100% correctly. It predicted incorrectly 3 positions in cluster 3.



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 6. Confusion matrixes of all 3 models

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?

Using original data dataset, I have compared the performance of ETS and ARIMA model to select which one will be better for forecasting the store's performance.

I have used ETS(M, N, M) without dampening for the ETS model.

The error plot shows variance over the years. It is fluctuating with different seizes; this means we need to use the error multiplicatively(M).

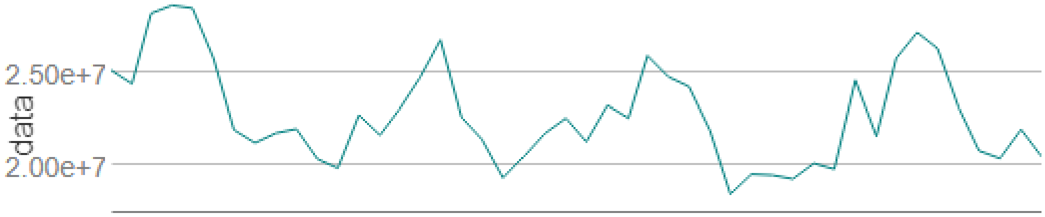


Figure 7. Decomposition Plot – Data Graph

We aren't able to clearly say if there is a pattern in the below data, that is why we have applied neutral trend(N).

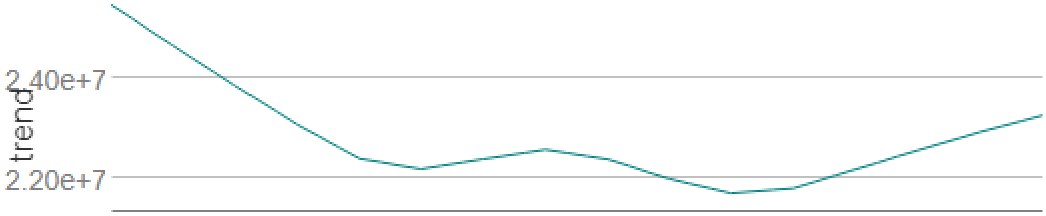


Figure 8. Decomposition Plot – Trend Graph

The seasonal plot shows seasonality in similar periods. That is why I have applied seasonality in the multiplicative method(M).

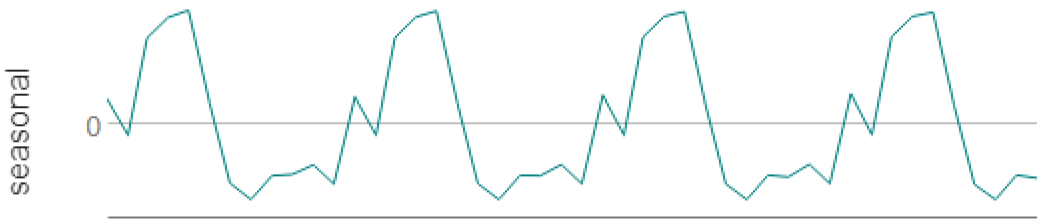


Figure 9. Decomposition Plot – Seasonal Graph

Using a time series plot, we can identify that the plot isn’t stationary, and we will need to apply some changes to it to use the ARIMA model effectively.

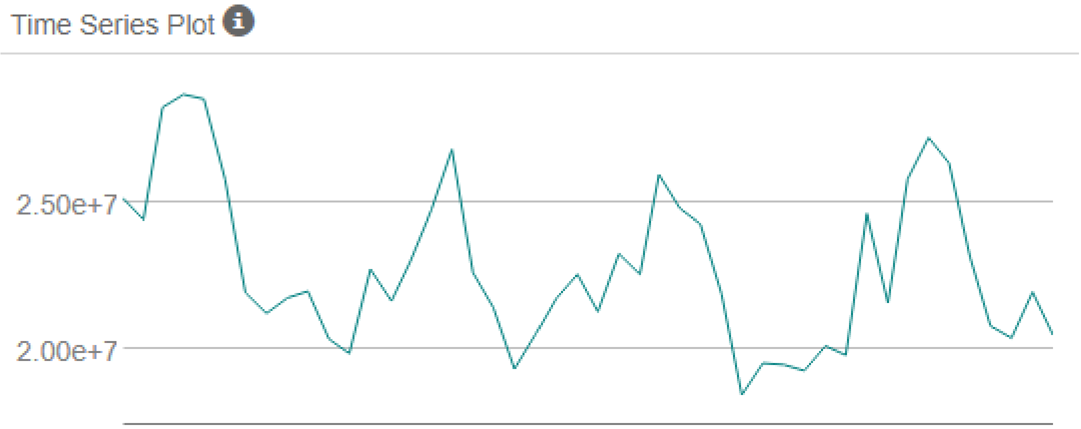


Figure 10. Time Series Plot.

The same is observed on the ACF and PACF function plots.

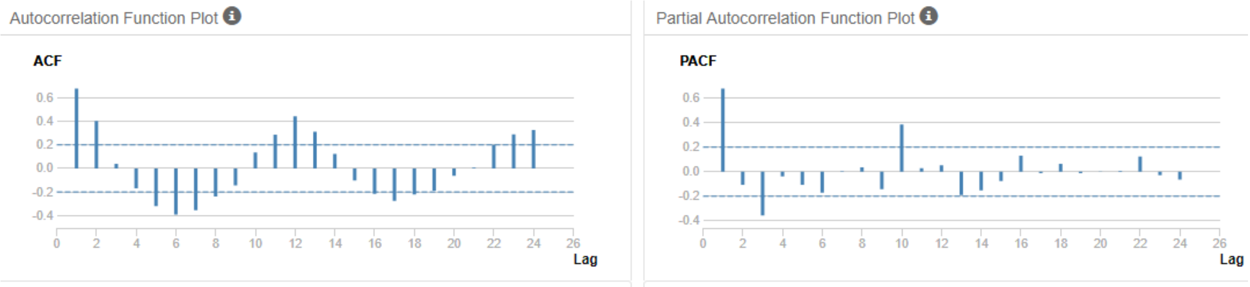
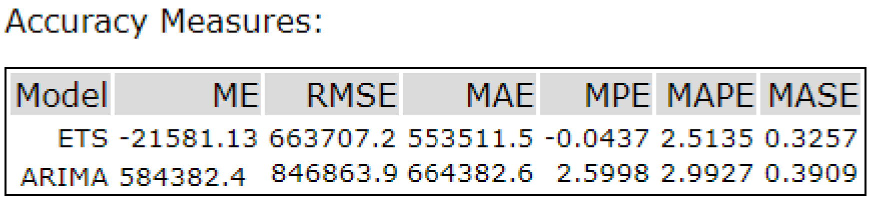


Figure 11. Time Series Plot.

Using the TS plot, I have discovered that I should use the models with these parameters: (0,1,2)(0,1,0).

After the two models have been complete, we can compare how good are their predictions.



Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

Figure 12. Accuracy Measures – TS Compare Tool Results.

Using the TS compare tool, we have obtained compassion for the two models. ETS model has the best accuracy values. That is why I believe the ETS model to forecast product sales for the new and existing stores.

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.

Period	Sub_Period	New Store Sales Forecast	Existing Store Sales Forecast
2016	1	2 588 356.56	21 829 060.03
2016	2	2 498 567.17	21 146 329.63
2016	3	2 919 067.02	23 735 686.94
2016	4	2 797 280.08	22 409 515.28
2016	5	3 163 764.86	25 621 828.73
2016	6	3 202 813.29	26 307 858.04
2016	7	3 228 212.24	26 705 092.56
2016	8	2 868 914.81	23 440 761.33
2016	9	2 538 372.27	20 640 047.32
2016	10	2 485 732.28	20 086 270.46
2016	11	2 583 447.59	20 858 119.96
2016	12	2 562 181.70	21 255 190.24

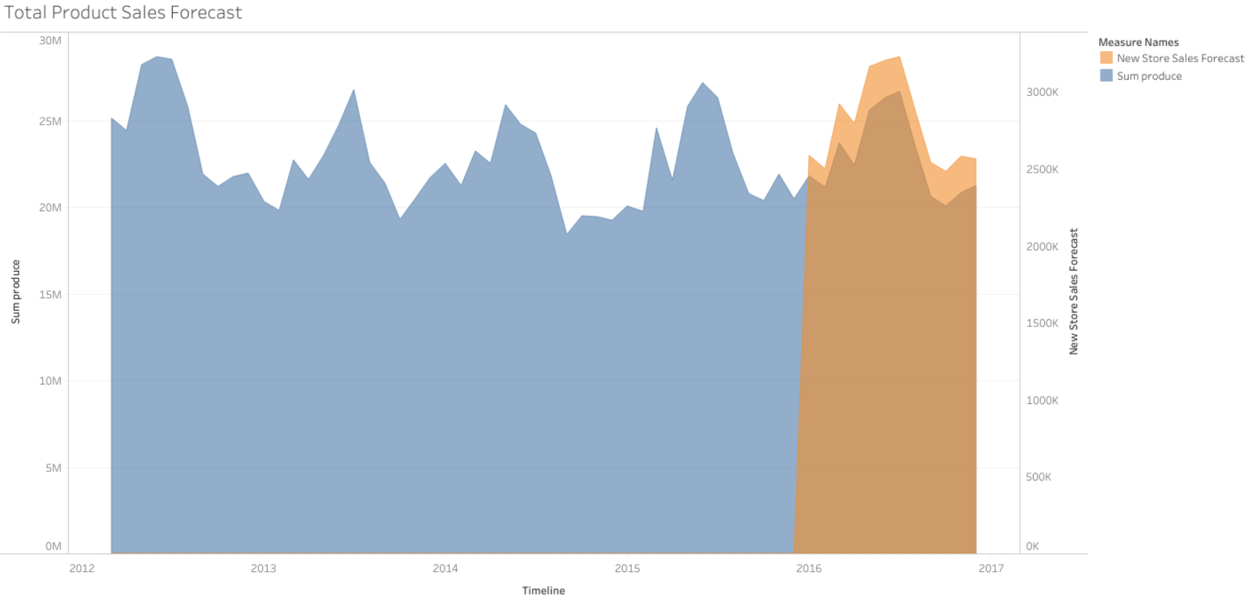
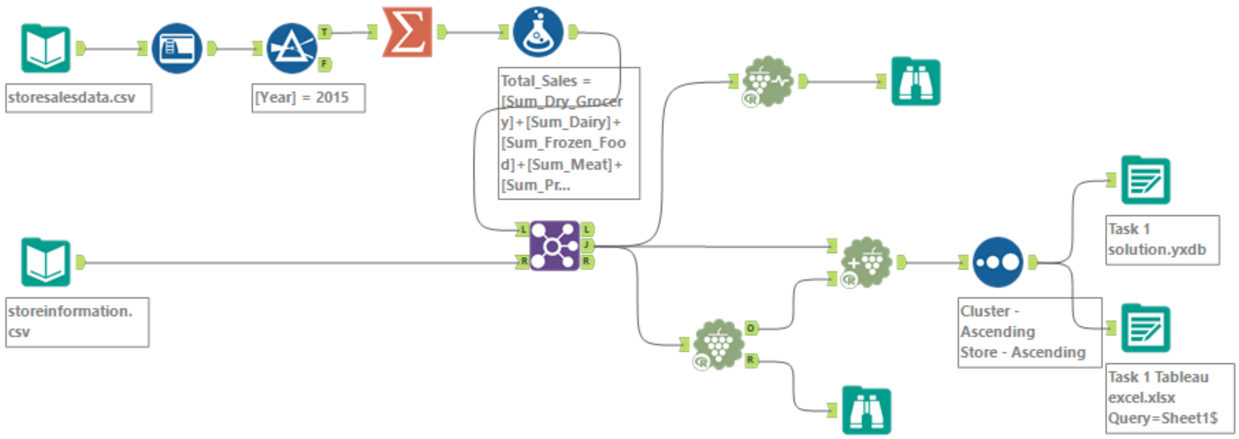
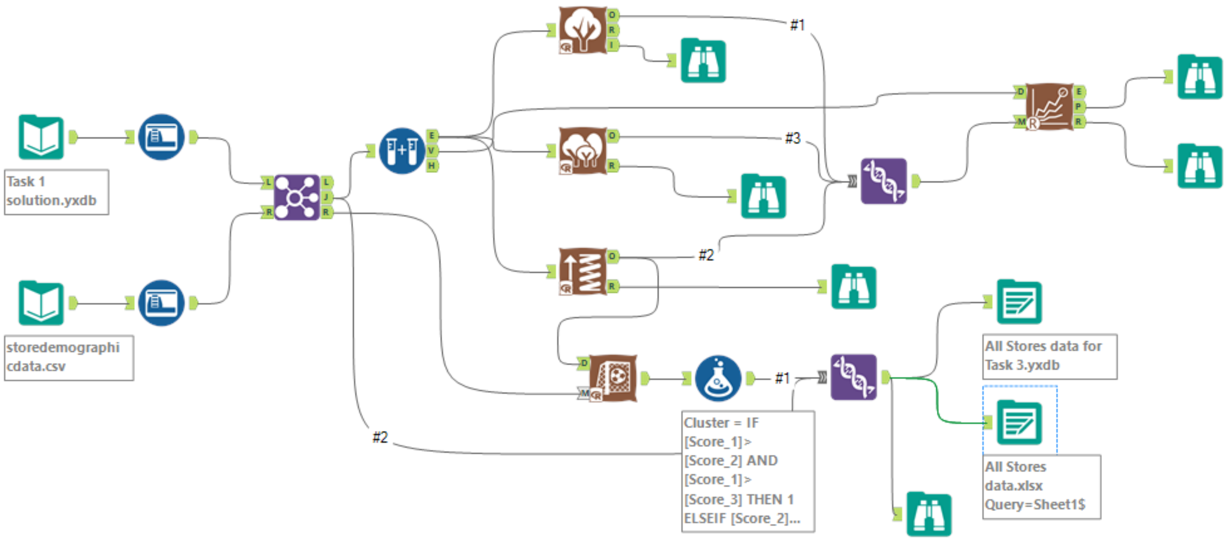


Figure 13. Historical data + Forecast for existing and new stores for the year 2016

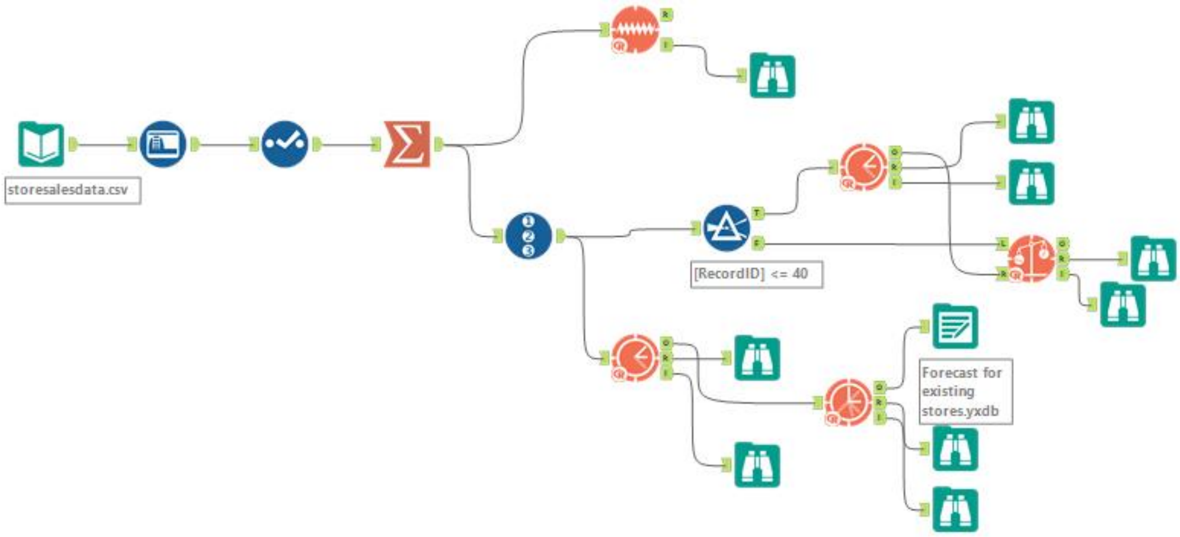
Alteryx Workflows


storesalesdata.csv
[Year] = 2015
T
F
Total_Sales = [Sum_Dry_Grocery]+[Sum_Dairy]+[Sum_Frozen_Food]+[Sum_Meat]+[Sum_Pr...
storeinformation.csv
L
J
R
O
R
Cluster - Ascending Store - Ascending
Task 1 solution.yxdb
Task 1 Tableau excel.xlsx Query=Sheet1$

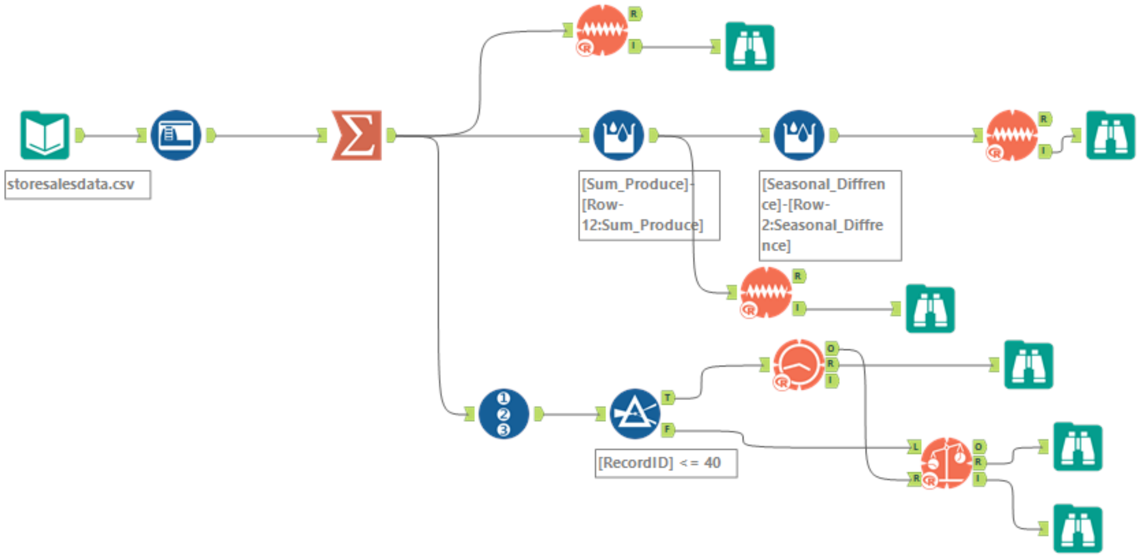
Workflow 1. Task 1: Determine Store Formats for Existing Stores



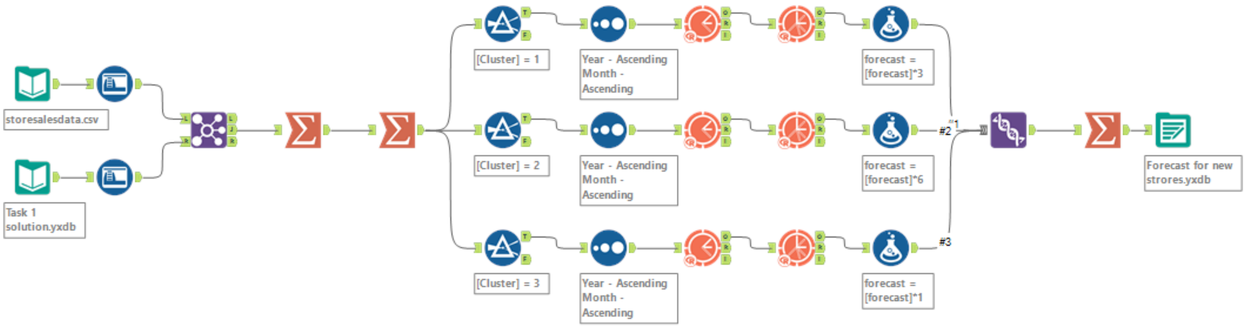
Workflow 2. Task 2: Formats for New Stores



Workflow 3. Task 3: ETS model – Validation + Forecast for existing stores


storesalesdata.csv
[Sum_Produce]-[Row-12:Sum_Produce]
[Seasonal_Diffrence]-[Row-2:Seasonal_Diffrence]
[RecordID] <= 40

Workflow 4. Task 3: Arima model - Validation



Workflow 5. Task 3: Forecast For New Stores