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 formats is 3 according to the K-Means, Adjusted Rand Indices and Calinski-Harbasz Indices.

K-Means Cluster Assessment Report

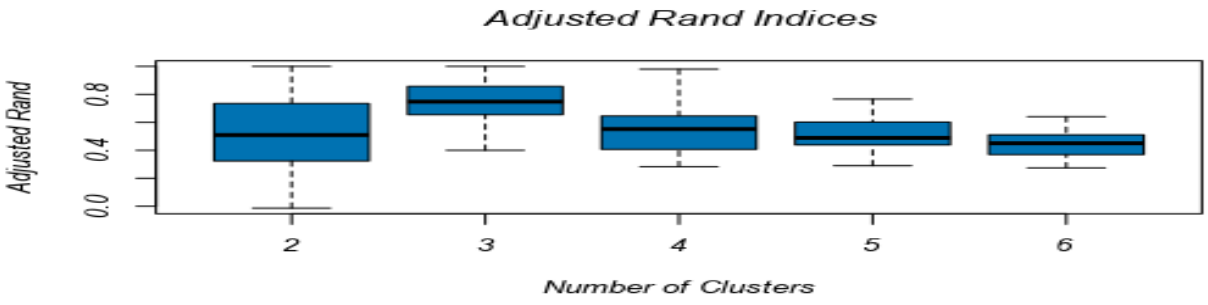
Summary Statistics

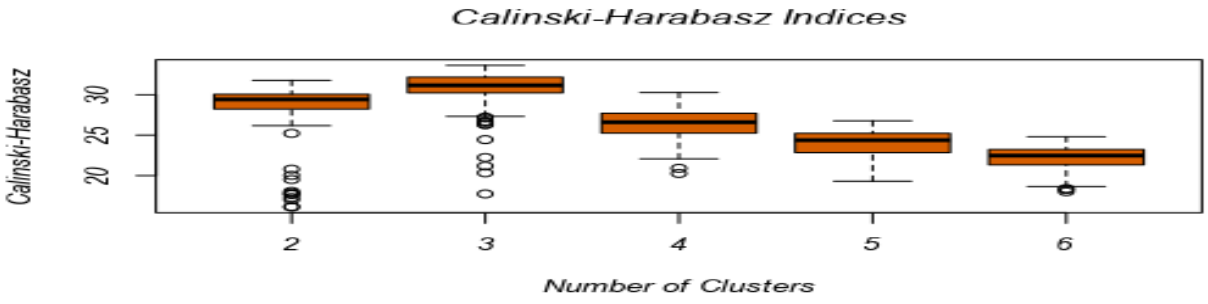
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.013285	0.40016	0.282983	0.290196	0.274408
1st Quartile	0.330717	0.658564	0.408503	0.438797	0.372238
Median	0.509315	0.748108	0.551657	0.488914	0.450276
Mean	0.497536	0.74604	0.552317	0.51303	0.450013
3rd Quartile	0.734461	0.853486	0.644338	0.600704	0.510104
Maximum	1	1	0.980253	0.765775	0.640469

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	16.09831	17.73061	20.26577	19.30878	18.00691
1st Quartile	28.23418	30.26749	25.28161	22.85841	21.33069
Median	29.42988	31.16865	26.60781	24.37966	22.47361
Mean	28.15119	30.46564	26.35459	24.03852	22.1983
3rd Quartile	30.05522	32.17675	27.68877	25.20925	23.2302
Maximum	31.78345	33.63781	30.28294	26.77603	24.80561





2. How many stores fall into each store format?

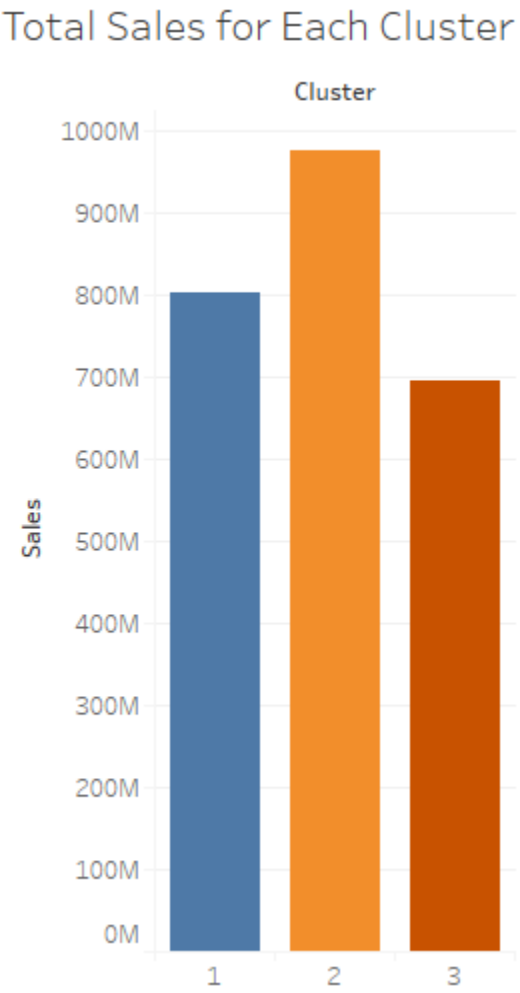
   Cluster 1 has 23 stores, cluster 2 has 29 and cluster 3 has 33 as following:

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

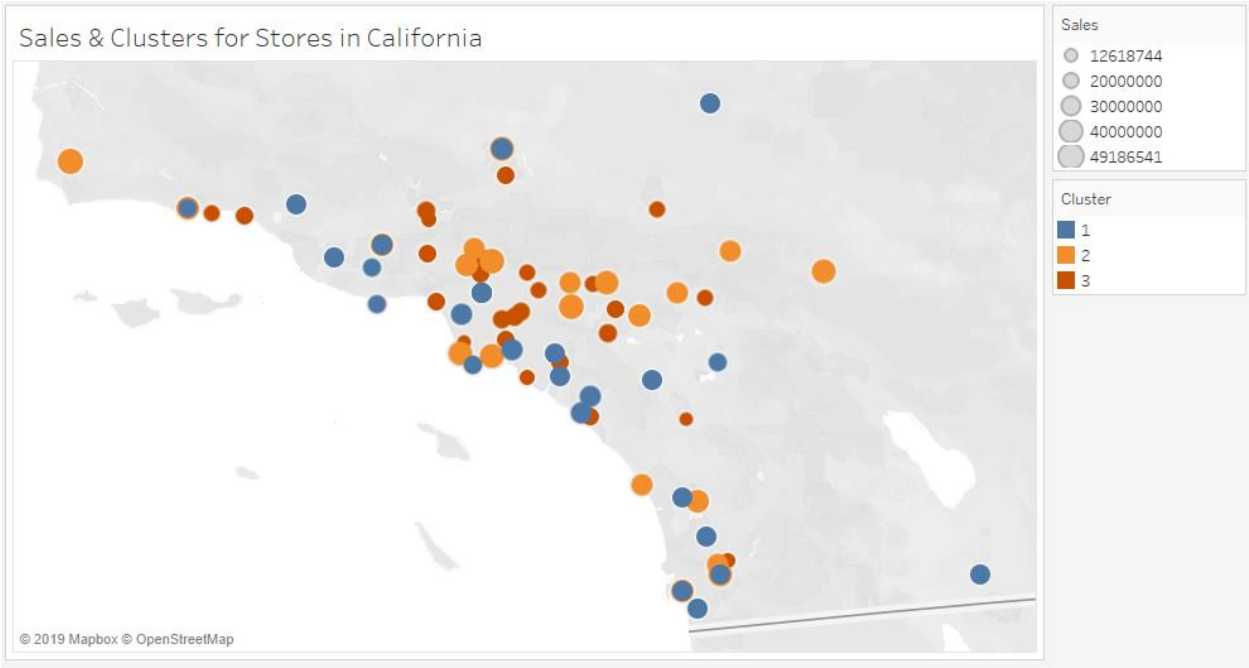
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

   One way to differ clusters from one to another is the total of sales of each one. For instance, we can figure which one has the highest sales! As following:



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



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?

   I chose to predict the best format for each store based on predictable cluster with use of module that has the highest accuracy from Boosted Module, Forest Module and Decision Tree. I chose Boosted Module since its F1 is the highest. Following report shows more details of all modules accuracies.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
B_Clustring	0.8235	0.8889	1.0000	1.0000	0.6667
DT_Clustring	0.7059	0.7685	0.7500	1.0000	0.5556
FM_Clustring	0.8235	0.8426	0.7500	1.0000	0.7778

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_Clustring

	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_Clustring

	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_Clustring

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

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?

I chose non dampening ETS (M,N,M) rather than ARIMA (0,1,2)(0,1,0)[12] with 6 month holdout sample. ETS has been chosen since its accuracy is higher than ARIMA. Its **RMSE** is **1044018.8** which is lower than **1429296.29** that **ARIMA** has and Its **MASE** is **0.4555** which is lower than **ARIMA's 0.5311**. Also, its **AIC** is **1479.4** which is higher than **858.7** that **ARIMA** has. Following tables show the mentioned information more clearly:

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-14783.6612202	1044018.8940828	809742.8924252	-0.2664397	3.5527937	0.4555978	0.3283229

Information criteria:

AIC	AICc	BIC
1479.4048	1495.4048	1506.8344

Method: ARIMA(0,1,2)(0,1,0)[12]

Call:
Arima(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	0.234438

sigma^2 estimated as 3268620653560.66: log likelihood = -426.38872

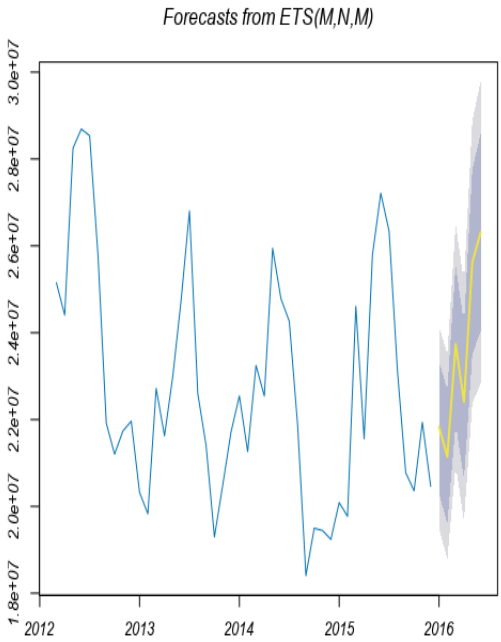
Information Criteria:

AIC	AICc	BIC
858.7774	859.8209	862.665

In-sample error measures:

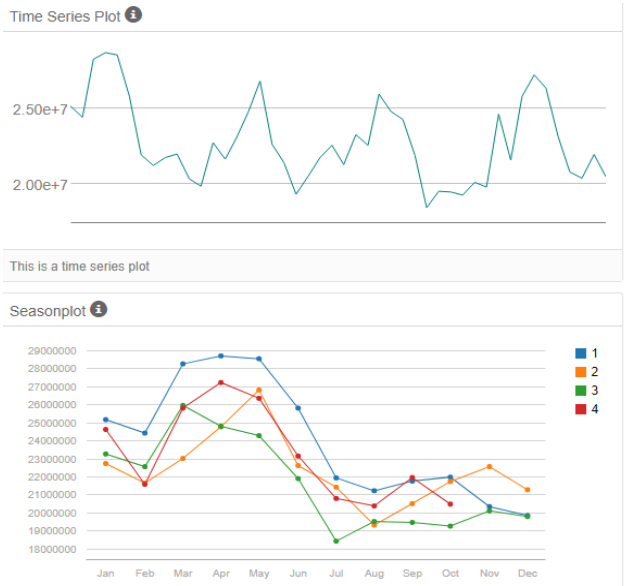
ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
170664.054315	1429296.2983494	951432.2560696	0.6151859	4.2022854	0.531117	-0.0260961

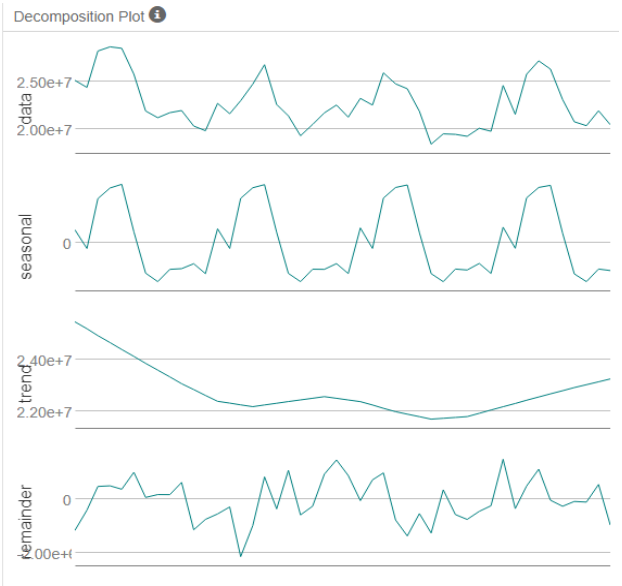
This chart shows the historic data, expected value in both 90% interval and 95% interval for ETS(M,N,M) Model:

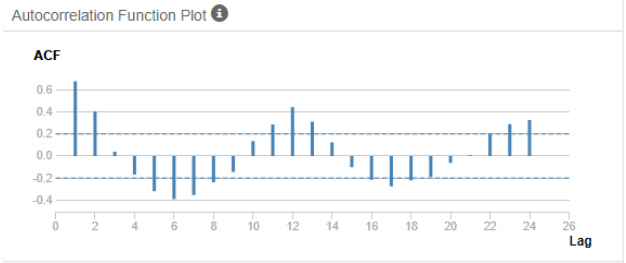


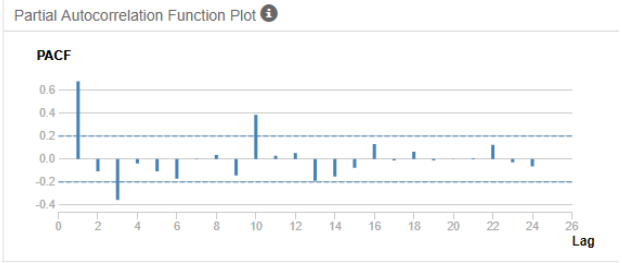
The Forecast Plot shows the historic data in black and the expected value in blue. The orange in the plot shows the 90% confidence interval, and the yellow shows the 95% confidence interval.

Since the increasing trend here is not clear and seasonal with not regular error, I chose the multiplicative application.





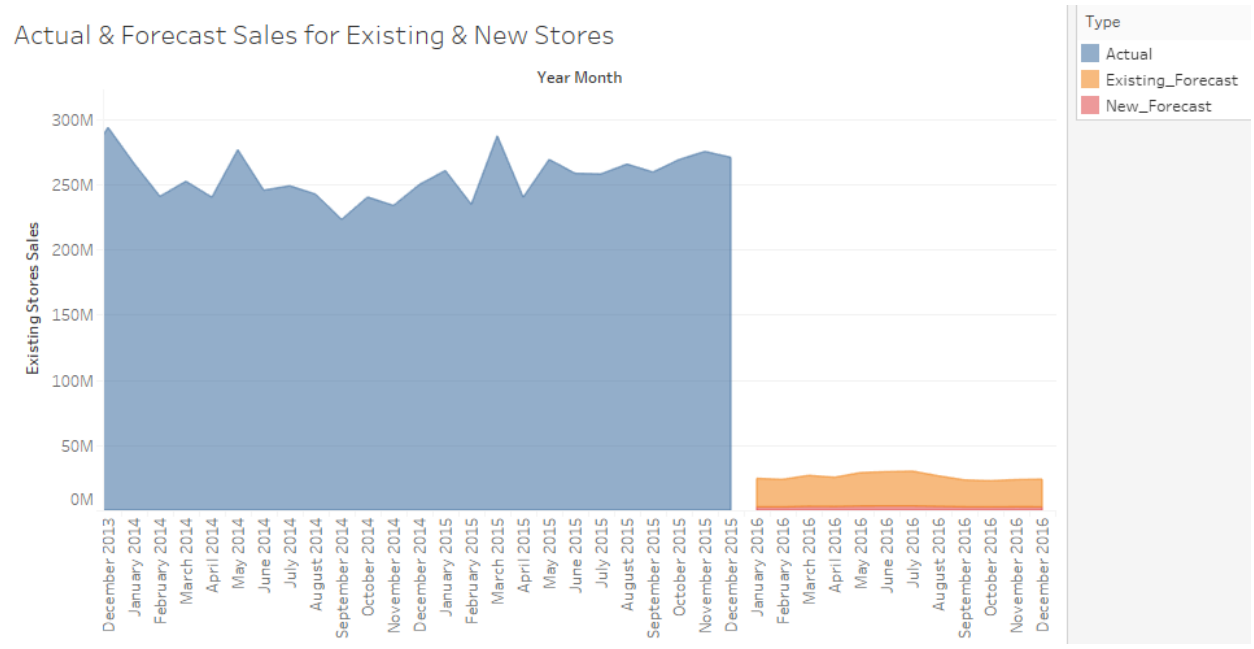




2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Year - Month	Existing Stores Forecast	New Stores Forecast
2016 - 01	21829060	2603262
2016 - 02	21146330	2508878
2016 - 03	23735687	2989458
2016 - 04	22409515	2849287
2016 - 05	25621829	3224711
2016 - 06	26307858	3269623
2016 - 07	26705093	3288334
2016 - 08	23440761	2937302
2016 - 09	20640047	2606592
2016 - 10	20086270	2536270
2016 - 11	20858120	2631293
2016 - 12	21255190	2586562

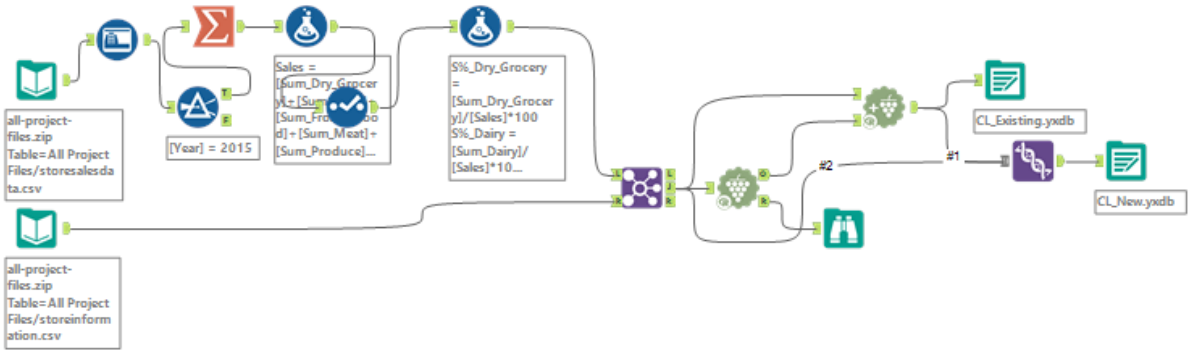
Visualization



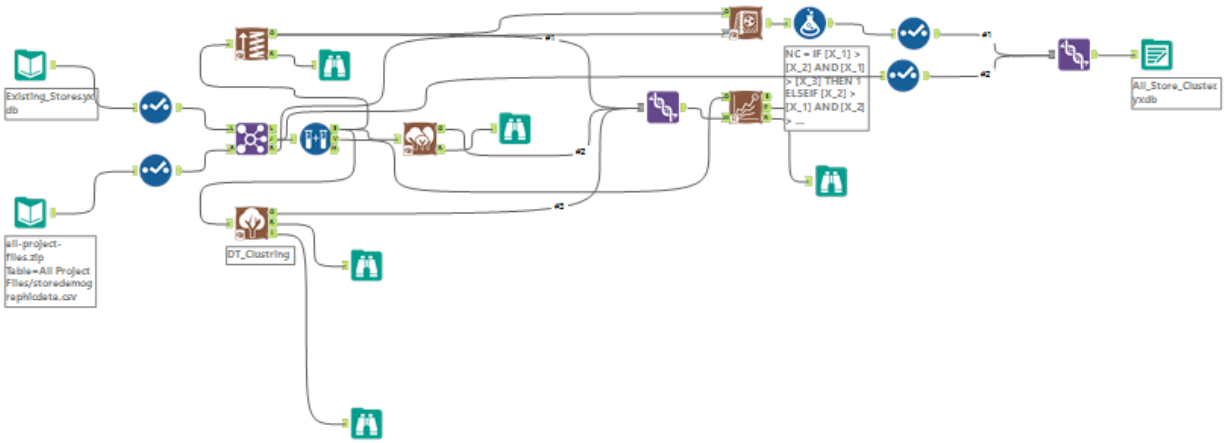
This chart shows Actual existing stores sales data from Jan - 2012 to Dec – 2015 and existing stores sales forecast with new stores sales forecast from Jan – 2016 to Dec – 2016. Check it in Tableau Public, It more clear there!

Workflows for All tasks.

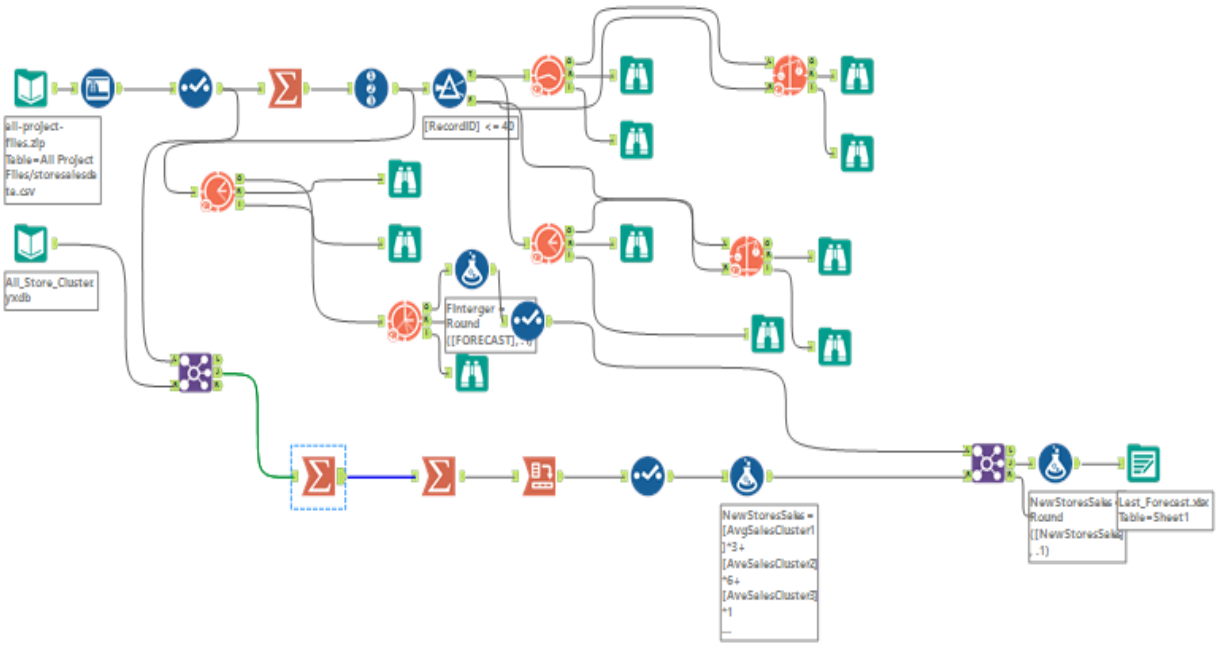
Task 1



Task 3



Task 3


all-project-files.zip|Table=All Project Files/storesalesdata.csv
All_Store_Clusteryxdb
[RecordID] <= 40
FInterger = Round([FORECAST], 0
NewStoresSales = [AvgSalesCluster1]*3+[AvgSalesCluster2]*6+[AvgSalesCluster3]*1
NewStoresSales = Round([NewStoresSales], -1)
Last_Forecast.xlsx|Table=Sheet1