<u>Project: Predictive Analytics Capstone</u>
<u>By : Alhanoof Alyabes</u>
<u>Date : 3/5/2018</u>

Task 1: Determine Store Formats for Existing Stores

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

By using percentage sales per category per store for clustering, we obtain the optimal number of store format based on 2015 sales data. the k-centroid diagnostic tool make assessment for appropriate clusters. We use tow measures: adjusted rand index and the Calinsk-Harabasz index, with k-means as clustering algorithms.

Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.0152	0.3171	0.3072	0.2412	0.2586	0.2903	0.2568
1st Quartile	0.352	0.4819	0.4431	0.3943	0.3896	0.3877	0.377
Median	0.4926	0.6936	0.4964	0.4487	0.4348	0.4417	0.4526
Mean	0.484	0.6575	0.5125	0.4623	0.4532	0.4498	0.4411
3rd Quartile	0.655	0.816	0.5913	0.4982	0.489	0.4997	0.491
Maximum	1	1	0.7458	0.7366	0.7762	0.6637	0.6118

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	20.09	17.41	18.98	17.24	16.61	16.11
1st Quartile	28.61	28.76	25.16	22.91	21.05	19.61	18.46
Median	29.47	30.7	26.25	24.05	22.02	20.56	19.5
Mean	28.41	29.47	25.99	23.88	21.96	20.48	19.62
3rd Quartile	30.39	31.58	27.62	25.06	23.14	21.35	20.77
Maximum	31.95	33.41	30.09	26.53	24.87	23.6	22.59

figure (1) K-means cluster assessment report

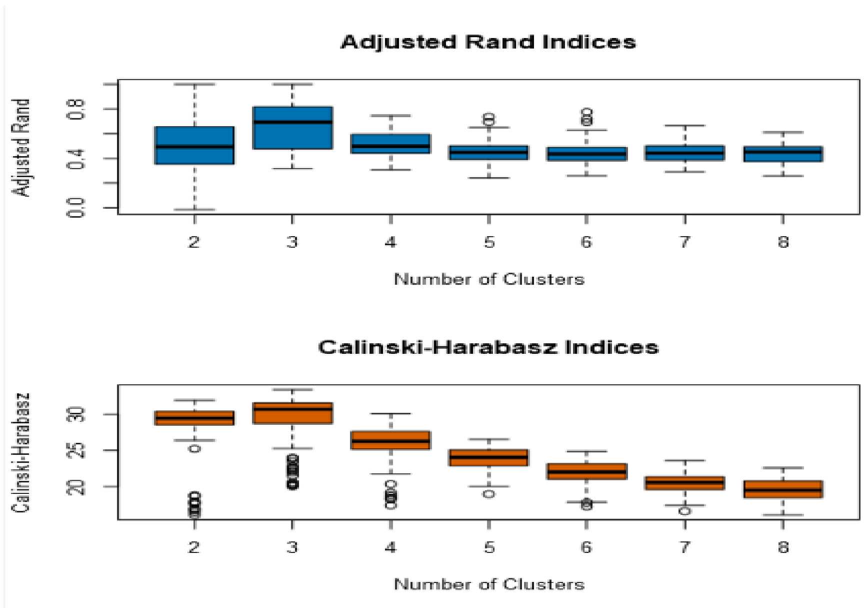


figure (2) Adjusted Rand and Calinski-Harabasz Indices

the k-means assessment report and adjusted rand, calinski-Harabasz indicates the optimal number format is **3**, based on the highest median value.

2. How many stores fall into each store format?

The k-centroid generate cluster information that indicates the size of cluster 1 is **23** , cluster 2 is **29** and 3 is **33**

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

figure(3) cluster information

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

By using tableau, we can observe that cluster 1 has highest percentage sales for general merchandise, while cluster 2 have high percentage sales for produce.

Moreover, cluster 1 have high total sales than the other two clusters. While cluster 3 are more compact range, which means the stores in cluster 3 are more similar on sales

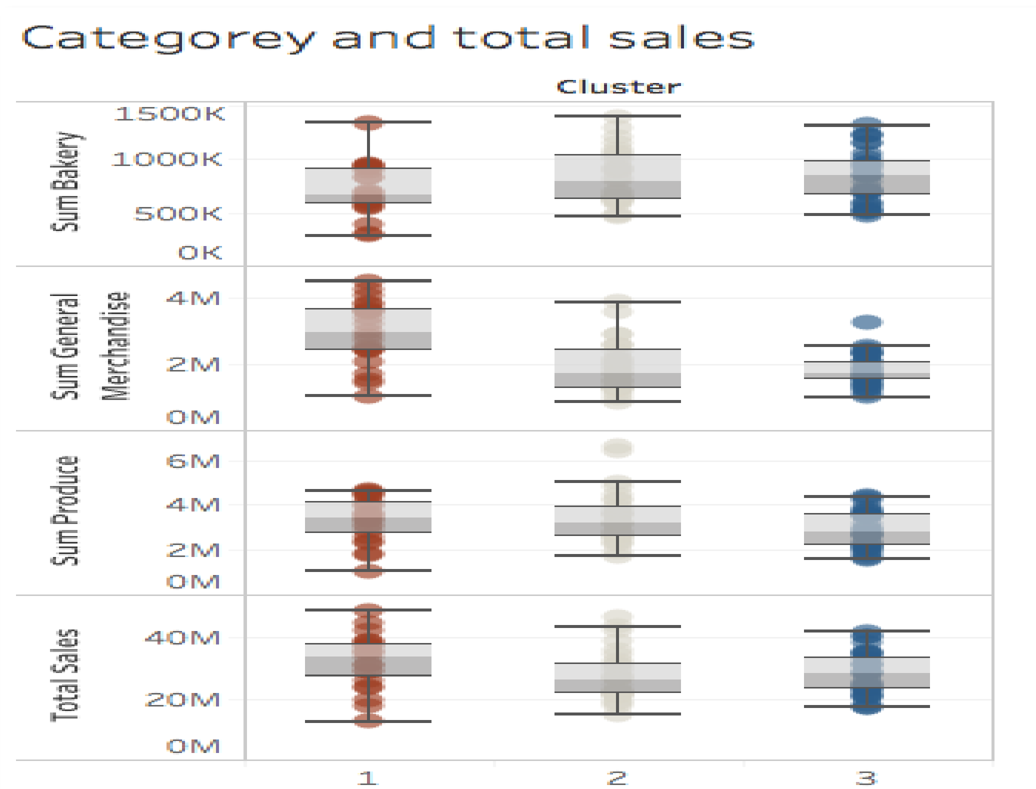


figure (4) total sales for category

figure (5) category for clusters

https://public.tableau.com/views/Task1_126/Q3?:embed=y&:display_count=yes&publish=yes

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.

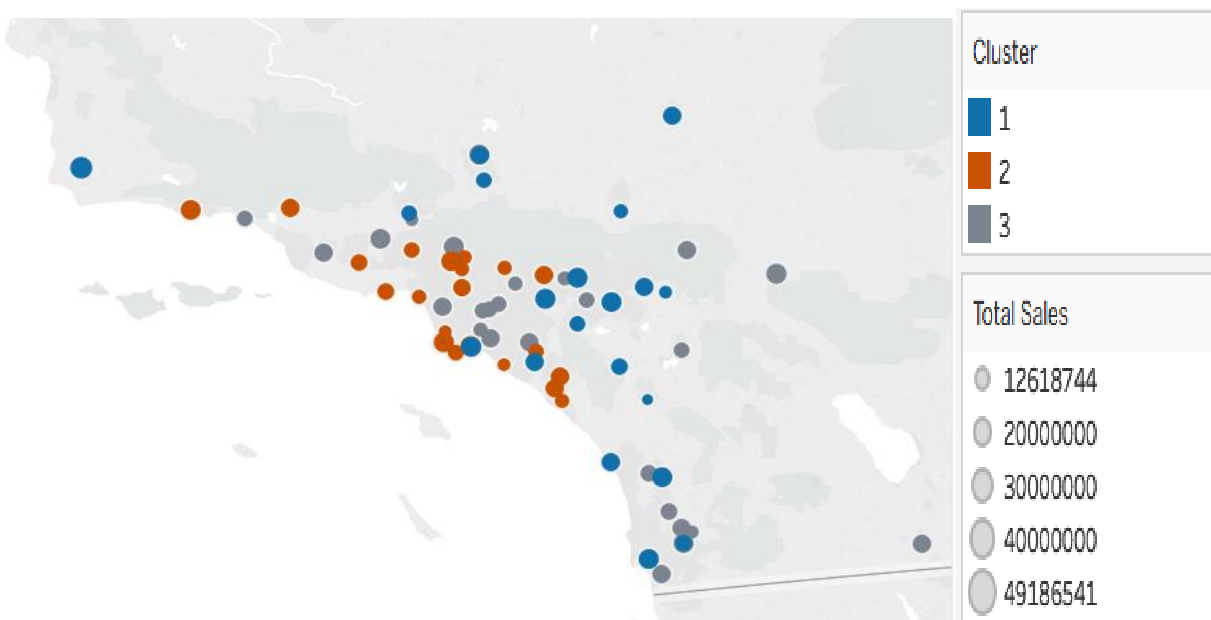


figure (6) locations of the stores

https://public.tableau.com/shared/KNB83W6JP?:display_count=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.

By using model comparison tool, report generated with comparison between decision tree, forest and boosted model. As the report shows, the boosted model chosen due the higher accuracy an F1 than forest model

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	.7059	.7327	.6000	.6667	.8333
forest	.8235	.8251	.7500	.8000	.8750
boosted	.8235	.8543	.8000	.6667	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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

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 boosted

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

Confusion matrix of forest

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

figure (7) model comparison report

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.

Variable importance plot shows that **Age0to9 , HVal750Plus and EdHSGrad** are the three most importance variables.

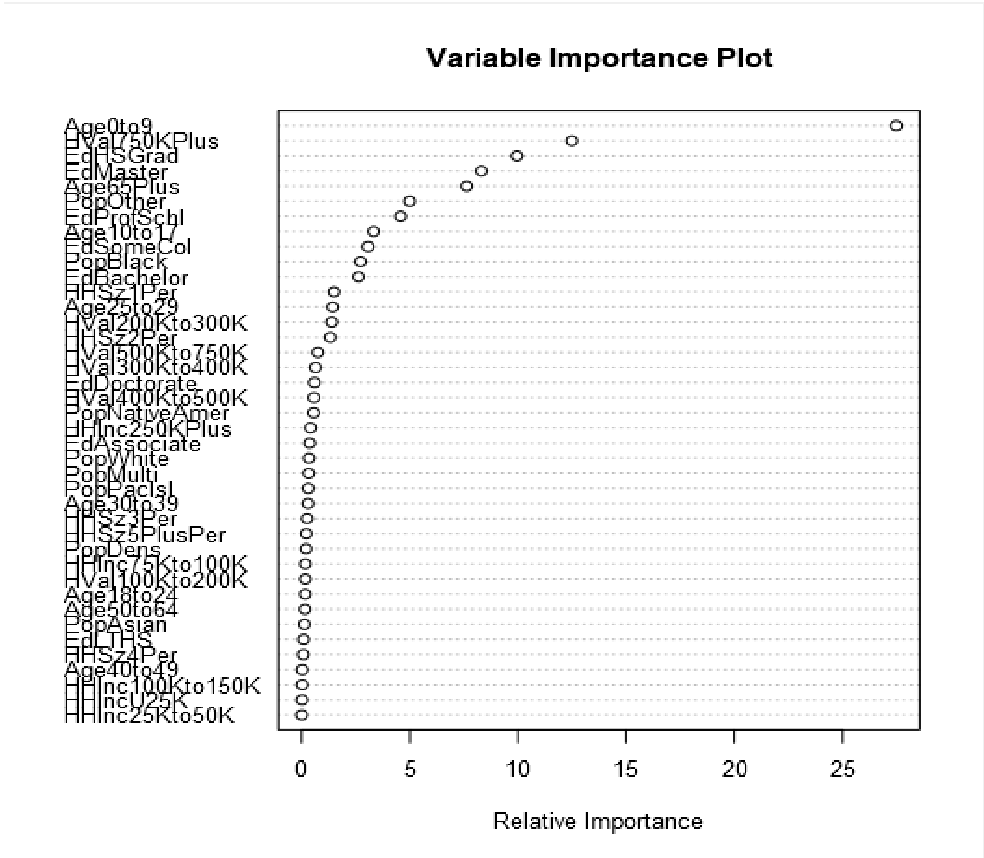


figure (8) variable importance plot

3. 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?

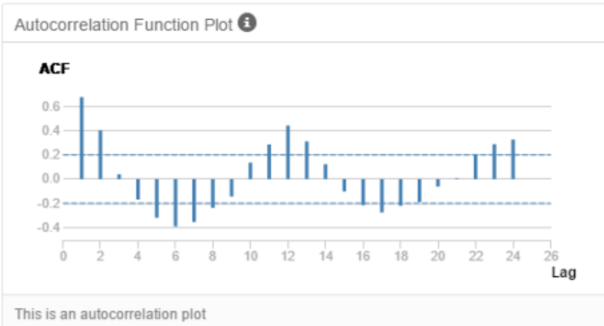
We used decomposition plot to forecast produce sales for 2016 for exist and new stores. To forecast sales for new stores we aggregate sales and make the forecast.

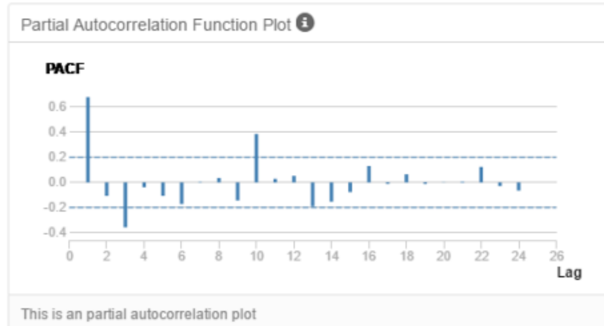
As at shown in plot below, the seasonal is slightly growing which should be applied as multiplicatively. there is no trend, so it will not be applied. The error shows shrinking or growing over time, so it applied as multiplicatively.

Then, the ETS model is (M, N, M) .

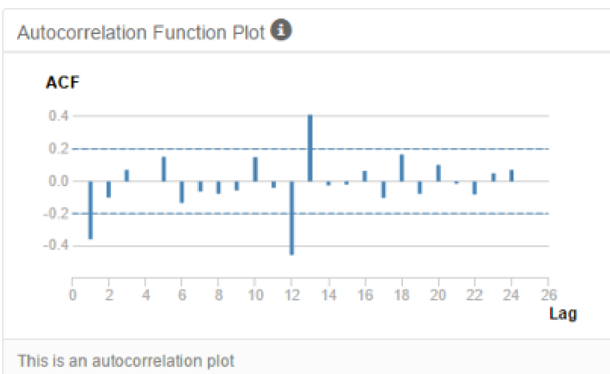
figure(9) decomposition plot

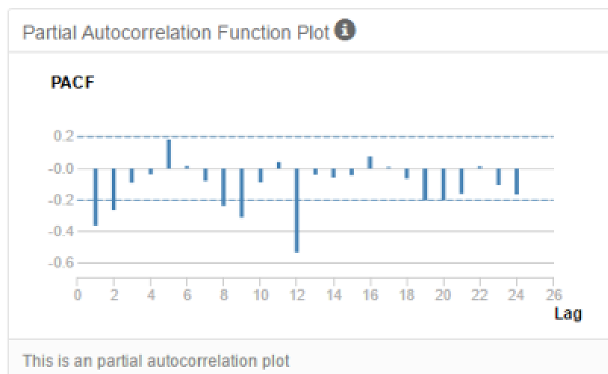
Autocorrelation function (ACF) and partial autocorrelation (PACF) for ARIMA model are shown below, it's shows that (ACF) is more fluctuation, indicate that more multiple seasonal period. there is strong correlation as lag 1 in ACF plot indicates. Peaks occur at lag 12,24.





the non-stationary series corrected by differences. the seasonal first difference as shown below, has been stationeries. for non-seasonal terms, we can observe early lags with two negative spikes in ACF, which indicates MA terms. for seasonal series, negative peaks at lag 12 indicates MA terms. ARIMA model is (0,1,1) (0,1,1) 12.





we use holdout data as test and the rest of the data to choose the model. The hold out sample is the same number of period that we want to forecast, which is 1/2015 to 12/2015 used as holdout.

By comparing the two models, ETS model has been chosen due to the highest accuracy measures. the RMSE (**1983593**) and MASE (**1.2691**) for ETS are lower than ARIMA model

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	1983593	2226513	1983593	8.4729	8.4729	1.2691	NA
arima	2878344	3061362	2878344	12.5815	12.5815	1.8416	NA

figure (10) actual vs forecast values plot

as it shown in the plot, the ETS model is the best model to be used

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.

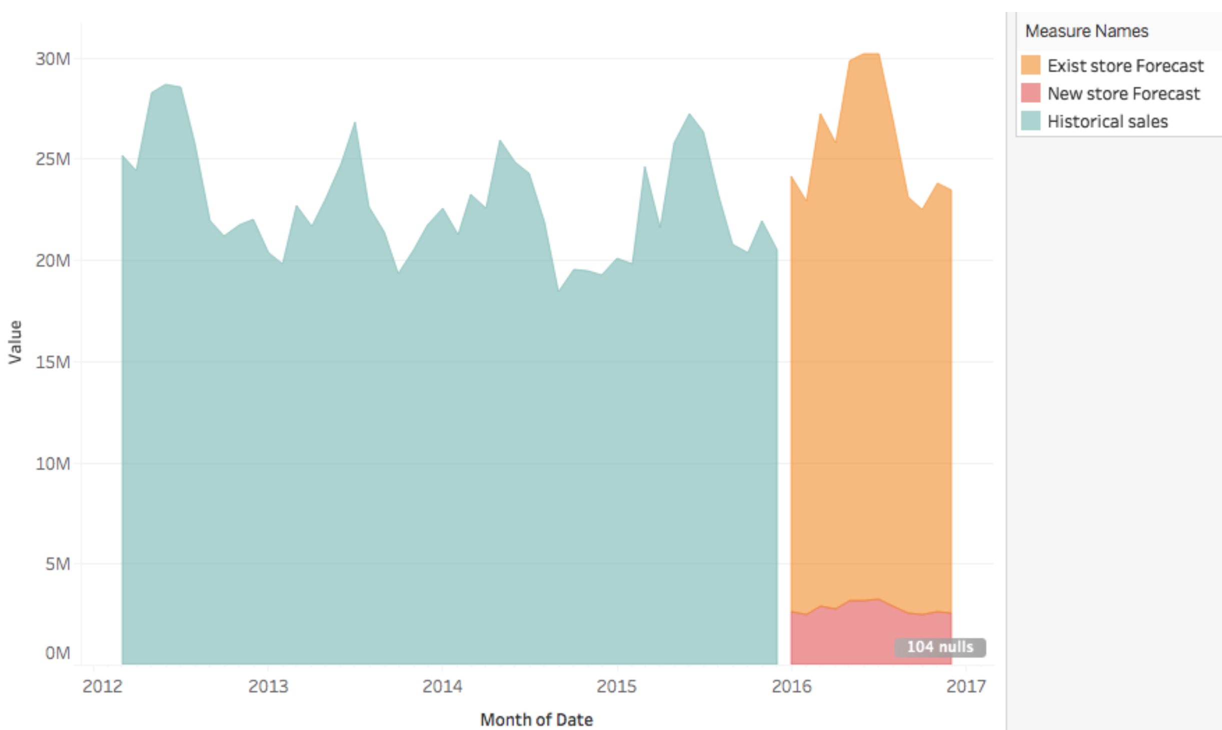


figure (11) tableau visualization

https://public.tableau.com/views/Task3_203/storessales?:embed=y&:display_count=yes

Year of Date	Month of D..	Historical s..	Exist store ..	New store ..
2012	March	25,151,526		
	April	24,406,048		
	May	28,249,539		
	June	28,691,364		
	July	28,535,707		
	August	25,793,521		
	September	21,915,642		
	October	21,203,563		
	November	21,736,159		
	December	21,962,977		
2013	January	20,322,684		
	February	19,829,621		
	March	22,717,070		
	April	21,625,385		
	May	23,000,152		
	June	24,755,406		
	July	26,803,106		
	August	22,600,217		
	September	21,401,266		
	October	19,296,578		
	November	20,489,773		
	December	21,715,707		
2014	January	22,544,458		
	February	21,262,413		
	March	23,247,169		
	April	22,541,988		
	May	25,943,047		
	June	24,782,178		
	July	24,263,118		
	August	21,879,989		
	September	18,407,264		
	October	19,497,572		
	November	19,444,753		
	December	19,240,385		
2015	January	20,088,529		
	February	19,772,333		
	March	24,608,407		
	April	21,559,729		
	May	25,792,075		
	June	27,212,464		
	July	26,338,477		
	August	23,130,627		
	September	20,774,416		
	October	20,359,981		
	November	21,936,907		
	December	20,462,899		
2016	January		21,539,936	2,587,451
	February		20,413,771	2,477,353
	March		24,325,953	2,913,185
	April		22,993,466	2,775,746
	May		26,691,951	3,150,867
	June		26,989,964	3,188,922
	July		26,948,631	3,214,746
	August		24,091,579	2,866,349
	September		20,523,492	2,538,727
	October		20,011,749	2,488,148
	November		21,177,435	2,595,270
	December		20,855,799	2,573,397

Year	Month	New store sales	Exist store sales
2016	1	2587450.85149522	21539936.0074994
2016	2	2477352.8923928	20413770.6013595
2016	3	2913185.23624958	24325953.0976278
2016	4	2775745.60976656	22993466.3485849
2016	5	3150866.83532587	26691951.4191559
2016	6	3188922.00335955	26989964.0105518
2016	7	3214745.64625064	26948630.7647638
2016	8	2866348.66339173	24091579.3491059
2016	9	2538726.84885954	20523492.4086428
2016	10	2488148.28746187	20011748.6685998
2016	11	2595270.38644805	21177435.4858385
2016	12	2573396.6290496	20855799.1096099

table (1) existing and new store sales

Alteryx workflow

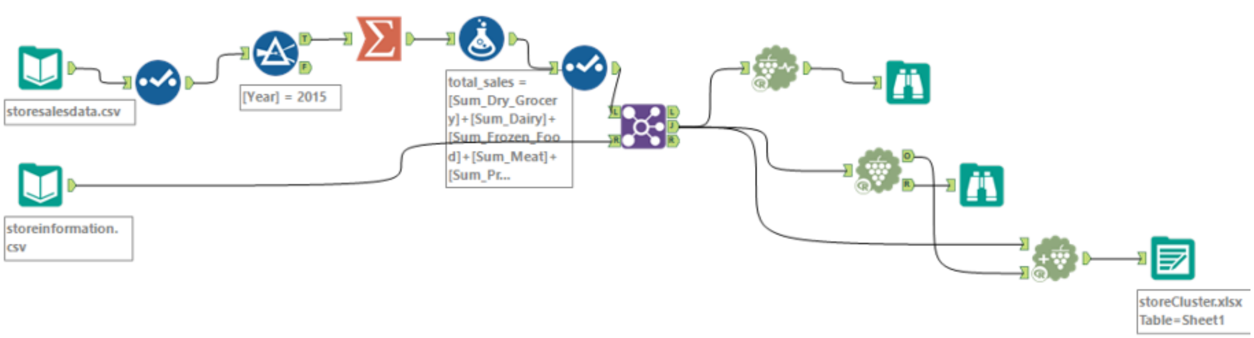


figure (12) task 1 workflow

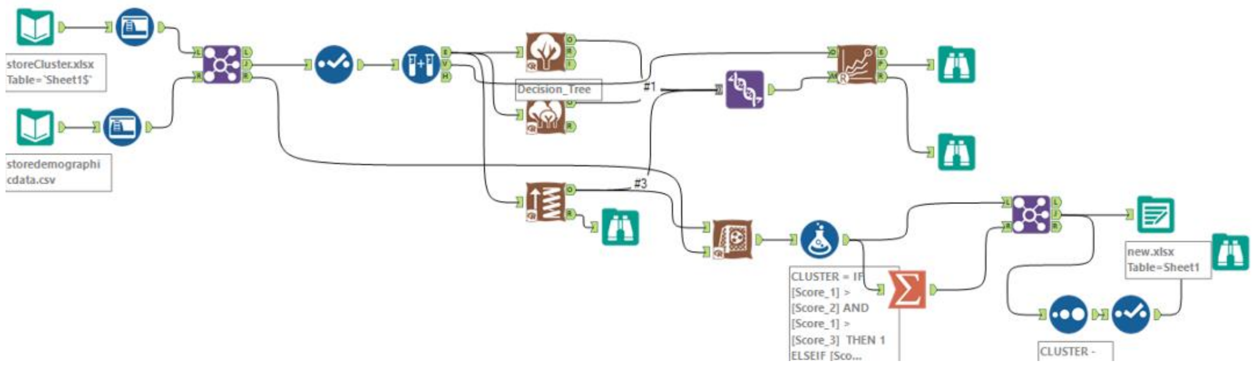


figure (13) Task 2 workflow

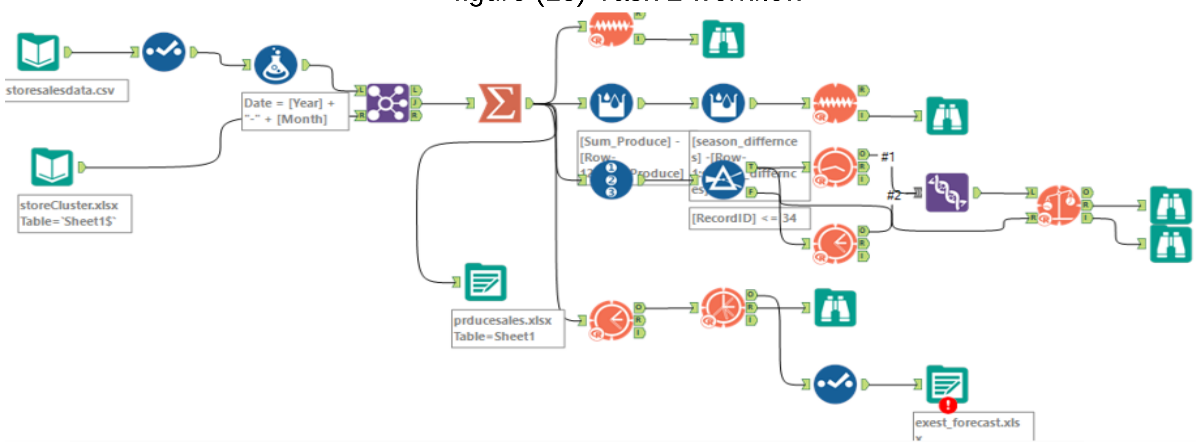
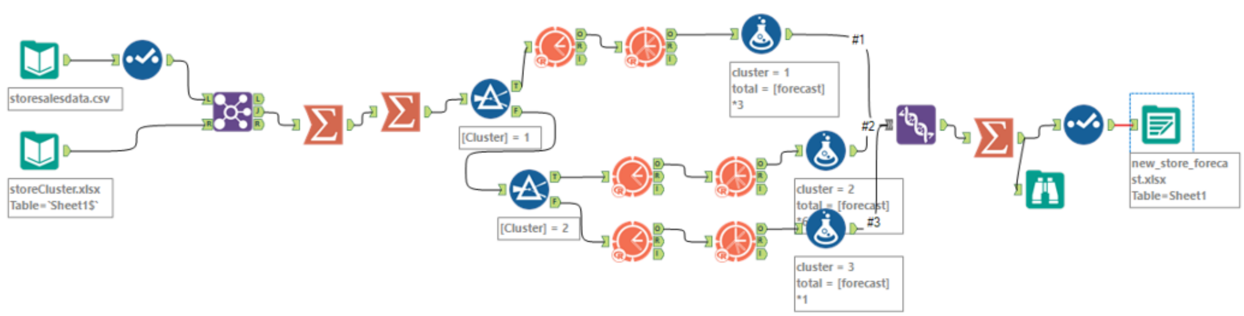
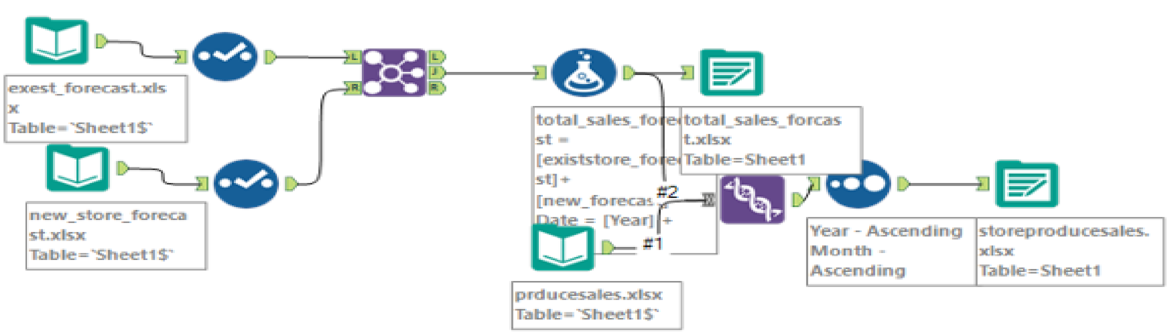


figure (14) task 3 , step 1 workflow



task 3-step 2



task 3-step 3