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?

Referring to the K-means Cluster Assessment report, the AR and CH indices figures below it's clear that the optimal number of store formats is 3 since it has the highest median value and less variation as well.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5
Minimum	-0.022325	0.076475	0.091158	0.176746
1st Quartile	0.067912	0.282514	0.292629	0.286626
Median	0.427664	0.376728	0.380491	0.322002
Mean	0.384302	0.420872	0.386695	0.355639
3rd Quartile	0.607625	0.540652	0.443096	0.416883
Maximum	0.952941	0.910092	0.784611	0.666832

Calinski-Harabasz Indices:

	2	3	4	5
Minimum	9.056198	10.47407	10.47619	8.997843
1st Quartile	17.628844	15.07319	13.97253	12.792015
Median	19.922319	16.74288	14.88656	13.477705
Mean	18.565382	16.26929	14.59282	13.357049
3rd Quartile	20.868999	17.71547	15.71887	14.30476
Maximum	21.992649	19.16412	16.82131	15.598009

Figure 1

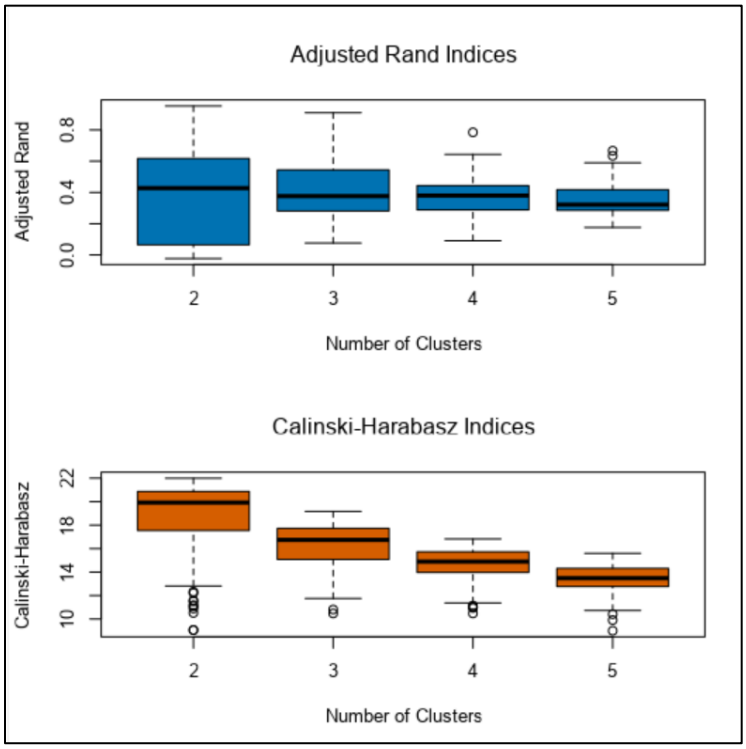


Figure 2

2. How many stores fall into each store format?

As shown below, cluster 1 has 23 stores, cluster 2 has 29 and cluster 3 has 33 stores.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540085	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Figure 3

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

Produce Sales in cluster 2 & 3 are opposite to each other while in cluster 1 & 3 General Merchandise Sales are opposite to each other.

	Per_Dry_Grocery	Per_Dairy	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.70261	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.08704	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Per_Bakery	Per_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Figure 4

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.

Clusters-Tableau visualization:
https://public.tableau.com/profile/sara6429#!/vizhome/Cluster_15945079471120/Clusters

The figure below shows a screenshot of the published visualization.

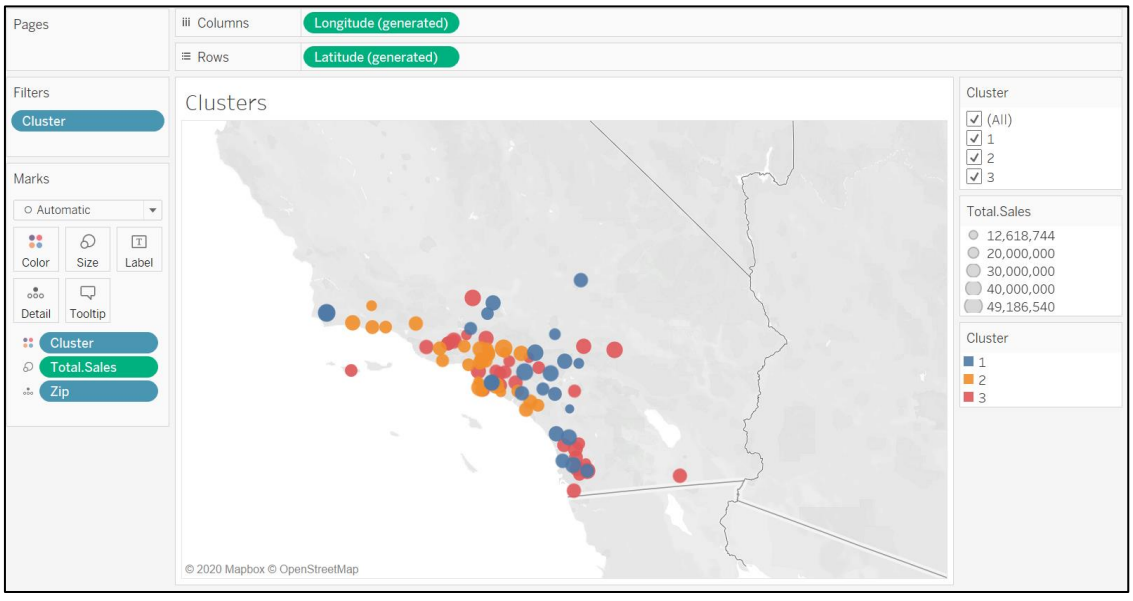


Figure 5

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?

Three predictive models have been tested out (Decision Tree, Forest Model & Boosted Model), the model comparison report below shows that the Boosted Model **BM_Store** has the highest overall accuracy and F1 value as well. Therefore, Boosted Model will be used to predict and score the new stores.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
FM_Store	0.8235	0.8426	0.7500	1.0000	0.7778
BM_Store	0.8235	0.8889	1.0000	1.0000	0.6667
DT_Store	0.7059	0.7685	0.7500	1.0000	0.5556

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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Figure 6

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

Ave0to9, HVal750KPlus and EdHSGrad are the most important variables as shown in the graph below.

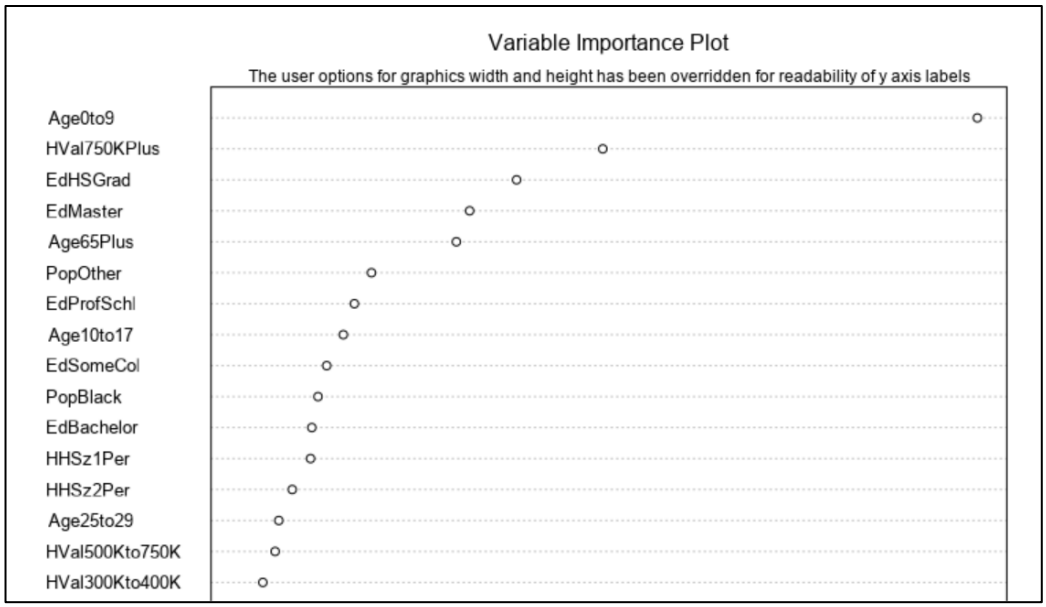


Figure 7

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?

The figure below will be used to examine and observe the three main components -seasonality, trend & error-.

- The seasonality is slightly changing in magnitude over time and hence it'll be applied multiplicatively (M).
- No trend is observed below, so no trend component will be included (N).
- The error is inconstantly fluctuating over time and it'll be applied multiplicatively (M).

Therefore, **ETS (M,N,M)** will be used for the forecasting.

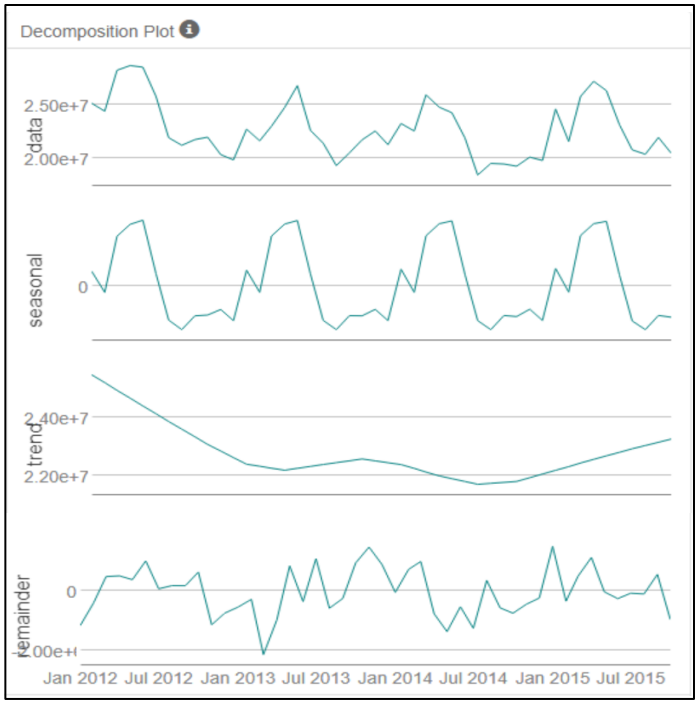


Figure 8

Based on the data behavior "seasonality & error" I suggest using ETS model instead of ARIMA model. However, the figure below confirms this as the EST model has a better ability to predict the holdout sample.

Accuracy Measures:

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

Figure 9

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.

The table below shows the forecasts for both new and existing stores and figure 10 shows the total sales visualization.

Month	New Store	Existing Store
Jan-16	2,696,429.48	21,829,060.03
Feb-16	2,537,249.91	21,146,329.63
Mar-16	2,459,553.62	23,735,686.94
Apr-16	2,499,426.40	22,409,515.28
May-16	2,710,067.90	25,621,828.73
Jun-16	2,520,389.87	26,307,858.04
Jul-16	2,501,498.90	26,705,092.56
Aug-16	2,537,125.52	23,440,761.33
Sep-16	2,698,680.98	20,640,047.32
Oct-16	2,557,965.67	20,086,270.46
Nov-16	2,472,589.21	20,858,119.96
Dec-16	2,553,384.27	21,255,190.24

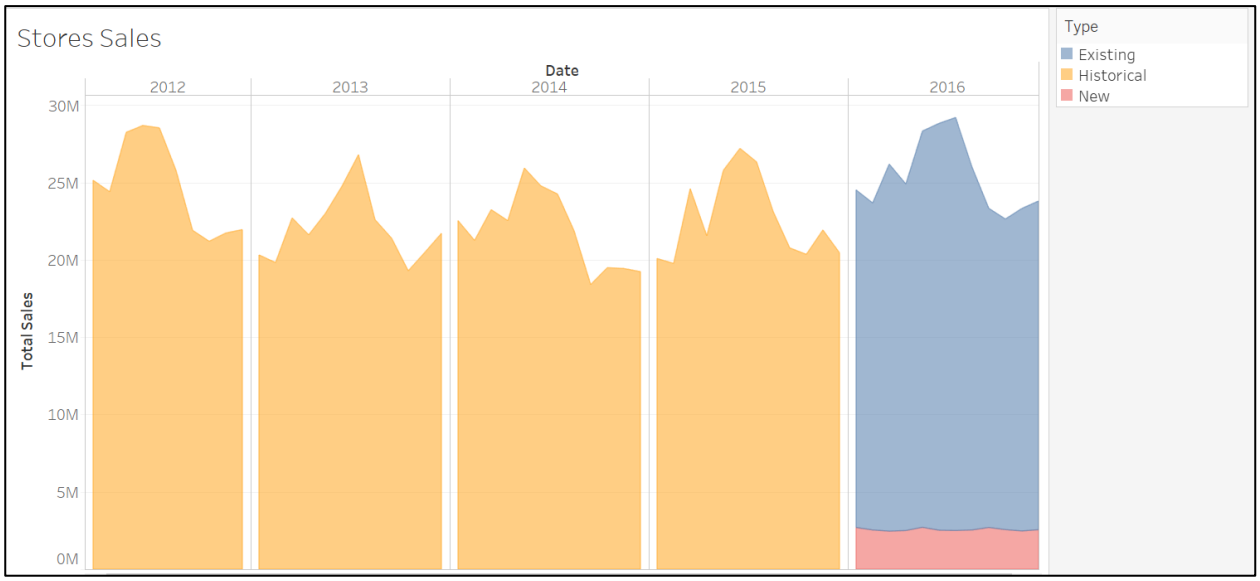
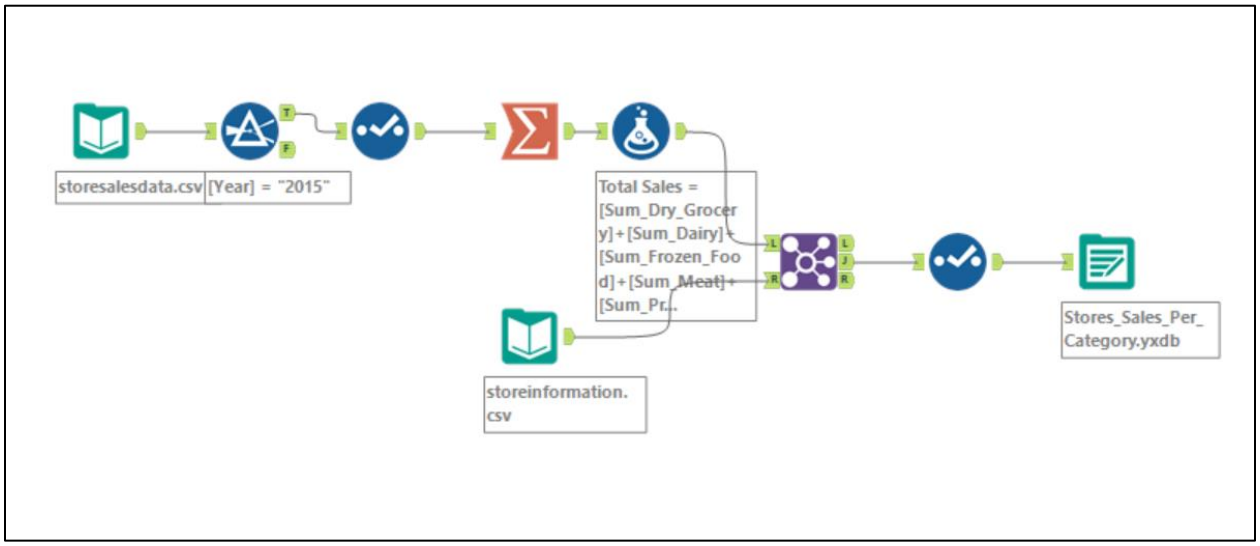


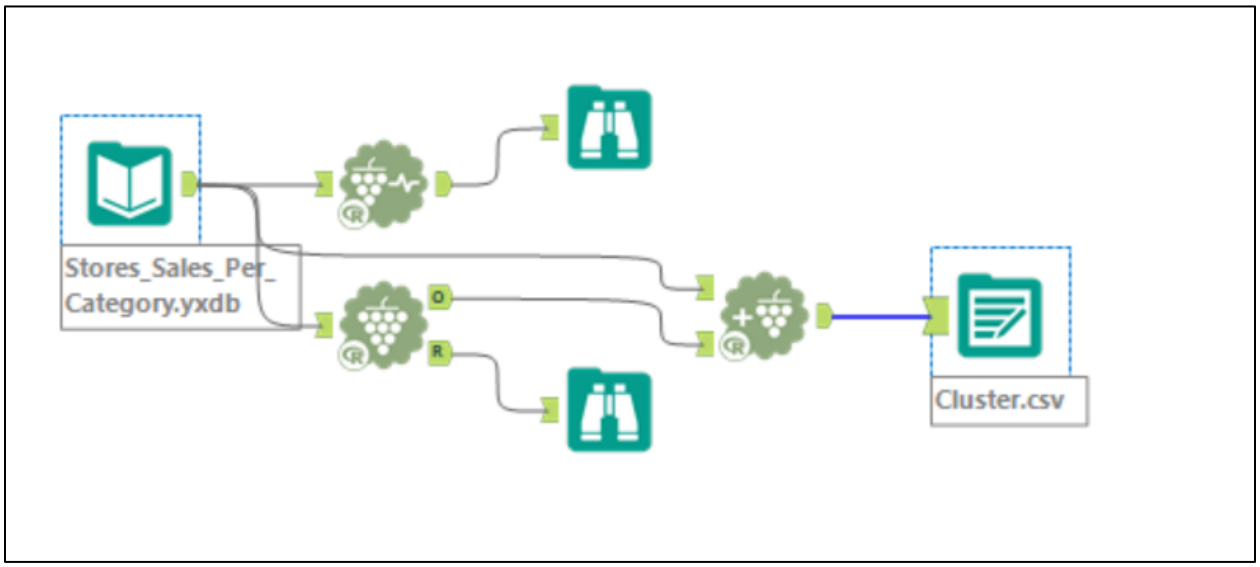
Figure 10

Total Sales-Tableau visualization:
https://public.tableau.com/profile/sara6429#!/vizhome/StoresSales_15950298651610/StoresSales

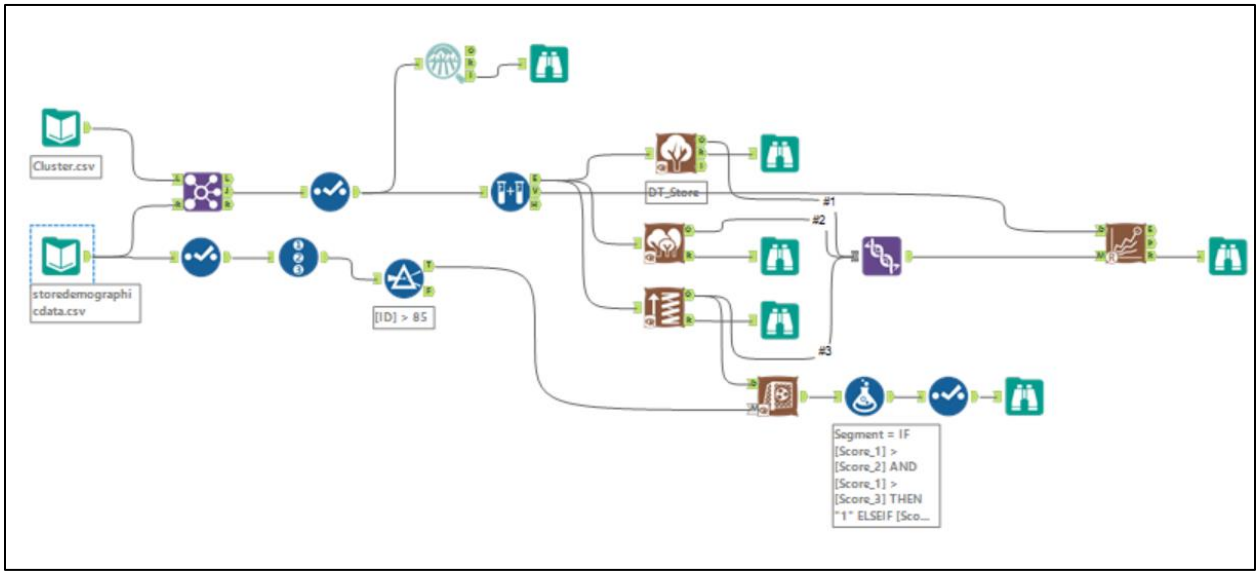
- Data Preparation Workflow- Task 1:



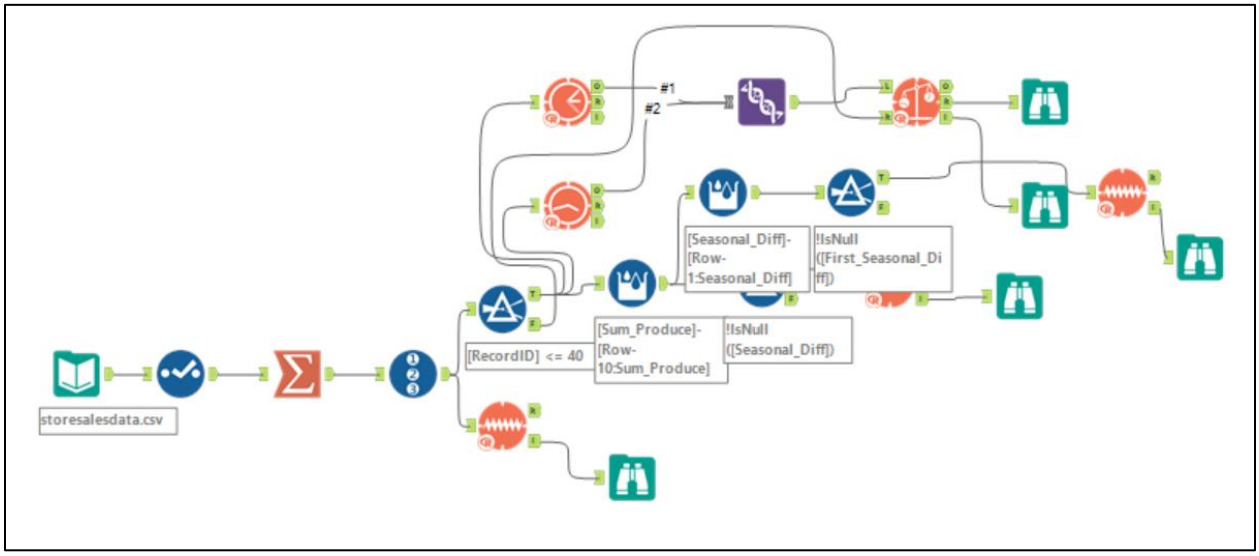
- Clustering Workflow- Task 1:

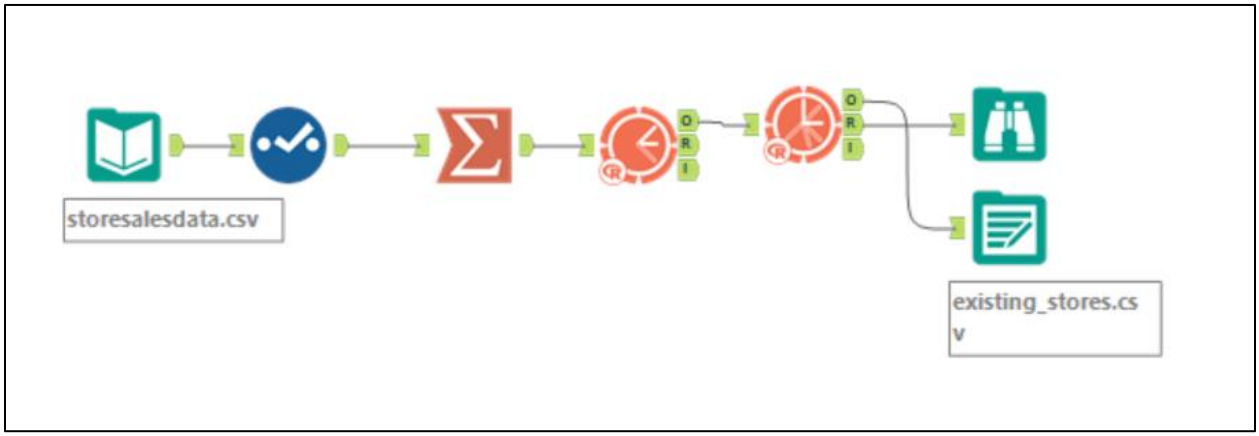


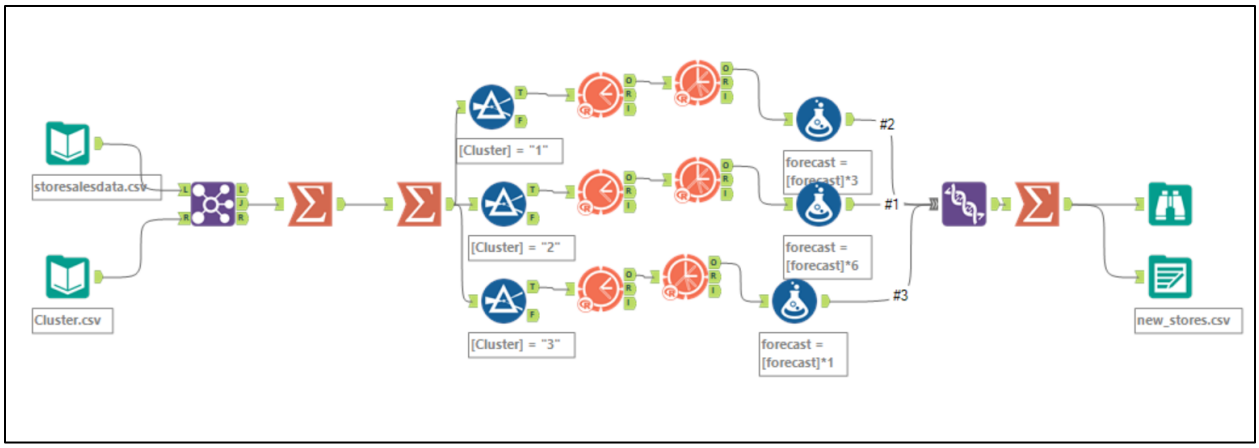
- Data Modeling & Scoring Workflow- Task 2:



- Predicting Produce Sales - Task 3 (three workflows):






storesalesdata.csv
Cluster.csv
[Cluster] = "1"
[Cluster] = "2"
[Cluster] = "3"
forecast = [forecast]*3
forecast = [forecast]*6
forecast = [forecast]*1
#2
#1
#3
new_stores.csv