Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?
   Looking at both the K-Means Cluster Assessment Report and the Adjusted Rand Indices we see the 3 clusters gives us the highest Median value; therefore, we will model for 3 store formats.

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.01155	0.3083	0.213	0.2837	0.2762
1st Quartile	0.3814	0.5258	0.4169	0.374	0.3965
Median	0.5619	0.6653	0.5107	0.4406	0.4256
Mean	0.5084	0.6594	0.5471	0.4704	0.4502
3rd Quartile	0.6942	0.7865	0.6427	0.5199	0.5067
Maximum	1	1	0.8902	0.8207	0.6626

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	16.1	18.94	18.45	17.02	17.37
1st Quartile	28.42	28.68	25.16	22.91	21.28
Median	29.47	30.83	26.61	23.98	22.17
Mean	28.24	29.58	26.34	23.7	21.95
3rd Quartile	30.31	31.97	27.85	24.9	22.84
Maximum	31.44	33.26	30.37	26.53	24.87

Figure 1: K-Means Cluster Assessment

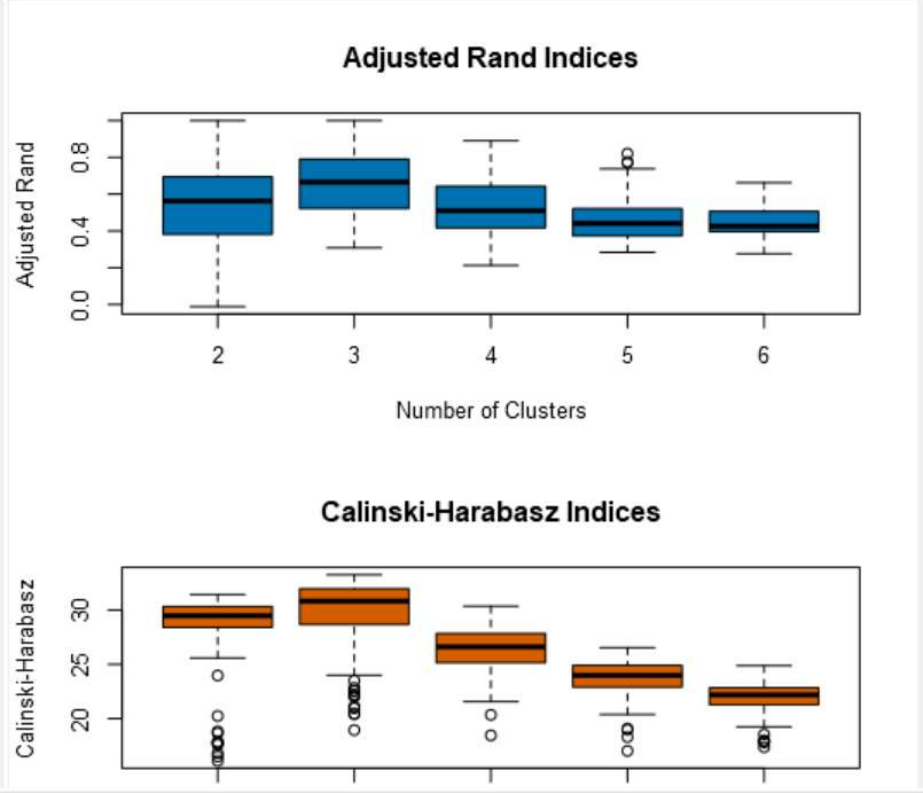

Figure 2: Adjusted Rand Boxplots

2. How many stores fall into each store format?
   The K-Means Cluster Solution gives us the following cluster sizes:
   - Cluster 1 = 23 stores
   - Cluster 2 = 29 stores
   - Cluster 3 = 33 stores

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.551451	1.874243
2	29	2.540085	4.475132	2.118707
3	33	2.115045	4.9262	1.702843

Figure 3: K-Means Clustering Solution

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Looking at Total Sales and the top three categories of goods sold we see that cluster 1 has wider variation in sales between stores and the compactness of the plots for cluster 3 tell us these stores are more closely related in terms of amount of sales dollars.

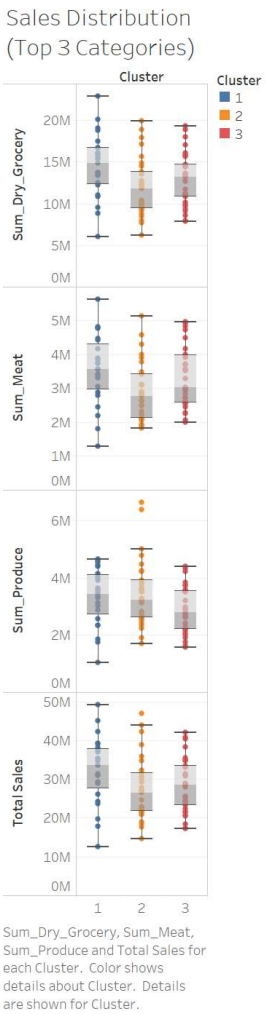

Figure 4: Sales by Category

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.

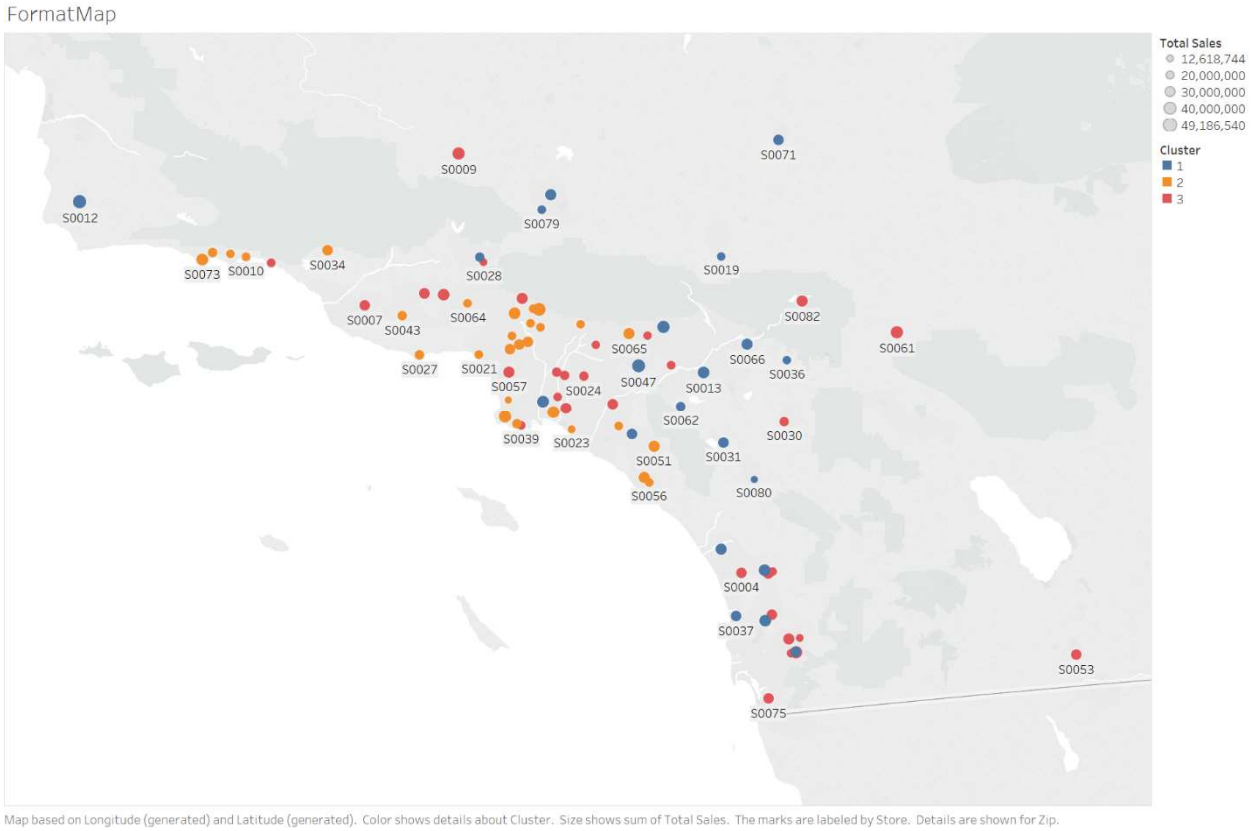

Figure 5: Store Format Map

[Tableau Format 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.)
   Looking at the Model Comparison Report we see an accuracy of 0.8235 across all three models; Decision Tree, Forrest Model, and the Boosted Model. Given this we look to the F1 value and we see the Boosted Model with the highest score therefore we will choose that model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DecisionTree	0.8235	0.8426	0.7500	1.0000	0.7778
ForrestModel	0.8235	0.8426	0.7500	1.0000	0.7778
BoostedModel	0.8235	0.8889	1.0000	1.0000	0.6667

Figure 6: Model Comparison Report

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?

First, looking at the TS Plot we see that the seasonal component of the Decomposition Plot we see a slight increase and thus will be applied multiplicatively. No clear trend is present therefore nothing will be used here. The remainder (error) is showing variation over the x axis and will be applied multiplicatively. Given these attributes the notation we will use for predicting the time series is ETS(N,M,N) Model.

Figure 7: Time Series Plot

The time series given to us shows seasonality and the seasonality should be removed before running the ARIMA model. A seasonal difference is taken but still shows correlations in the lags which are significant. We need to perform a first seasonal

difference. The first seasonal difference smooths the significant lags therefore no further differencing is needed as the series is stationary.

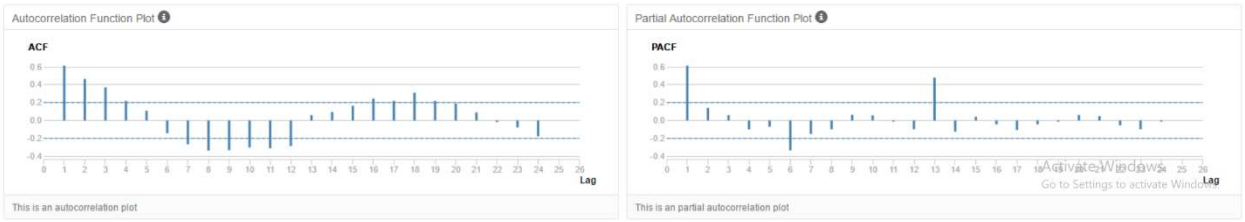

Figure 8: Seasonal Difference ACF/PACF Plots

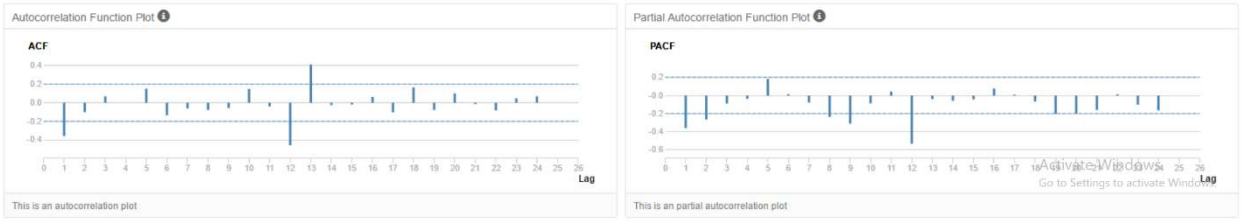

Figure 9: First Seasonal Difference ACF/PACF Plots

Removing the seasonality in the first differencing gives us D(1) and d(1) for our model. The strong negative correlation at lag 1 in the ACF and PCAF plots suggest MA(2) for the non-seasonal component. And the lags at 12 and 24 are not significant and therefore will no require a seasonal component. m should be 12. We will use the following notation for the ARIMA model: ARIMA(0,1,2)(0,1,0)12.

Comparing the two models against the 20% holdout sample we see a lower RMSE and a lower MSAE value in the ETS(M,N,M) model. Given this information on the accuracy measures of the errors we will choose the ETS(M,N,M) model over the ARIMA(0,1,2)(0,1,0)12 model.

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS_NoDampen	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA
ARIMA	584382.3	846863.9	664382.6	2.5998	2.9927	0.3909	NA

Figure 10: Model Comparison

3. 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.

		Type	
Year of..	Month o..	Forecast Existi..	Forecast New
2016	January	21,539,940	2,587,451
	February	20,413,770	2,477,353
	March	24,325,950	2,913,185
	April	22,993,470	2,775,746
	May	26,691,950	3,150,867
	June	26,989,960	3,188,922
	July	26,948,630	3,214,746
	August	24,091,580	2,866,349
	Septem..	20,523,490	2,538,727
	October	20,011,750	2,488,148
	Novemb..	21,177,440	2,595,270
	Decemb..	20,855,800	2,573,397

Figure 11: Forecast Table

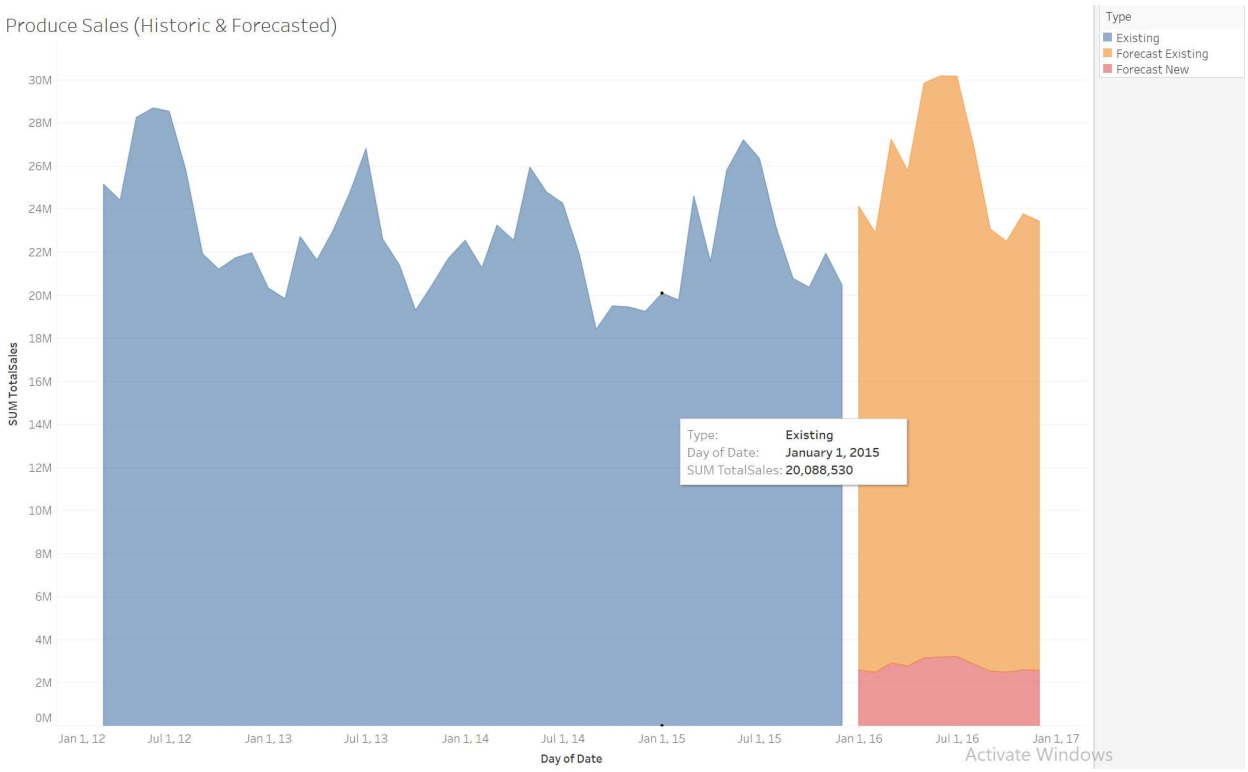

Figure 12: Historical and Forecasted Sales

[Task 3 Tableau Dashboard]

Alteryx Workflows

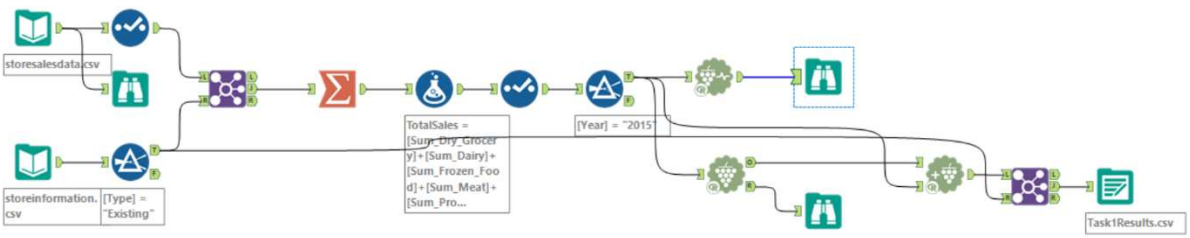

Figure 13: Task 1 Workflow

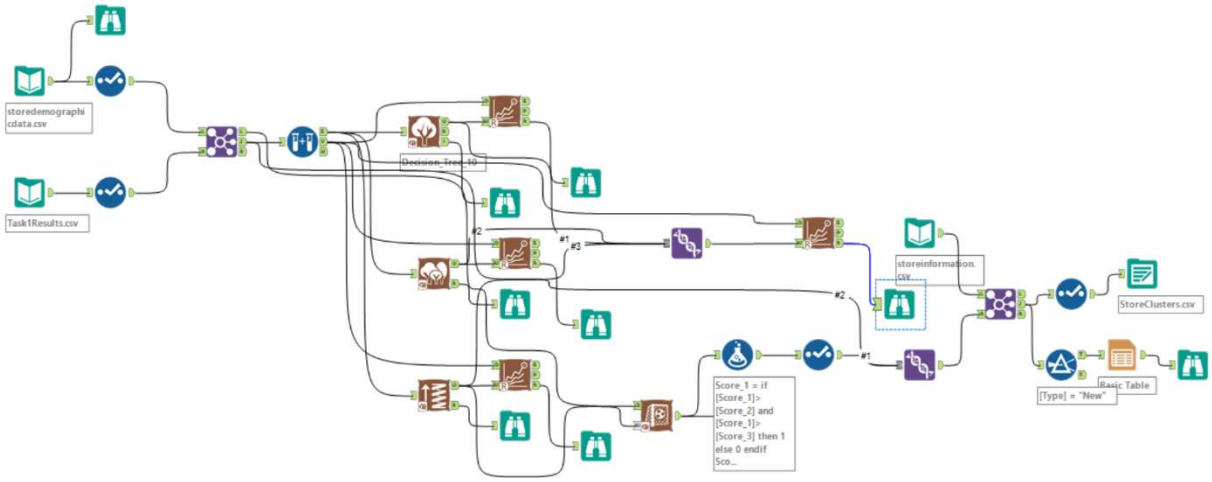

Figure 14: Task 2 Workflow

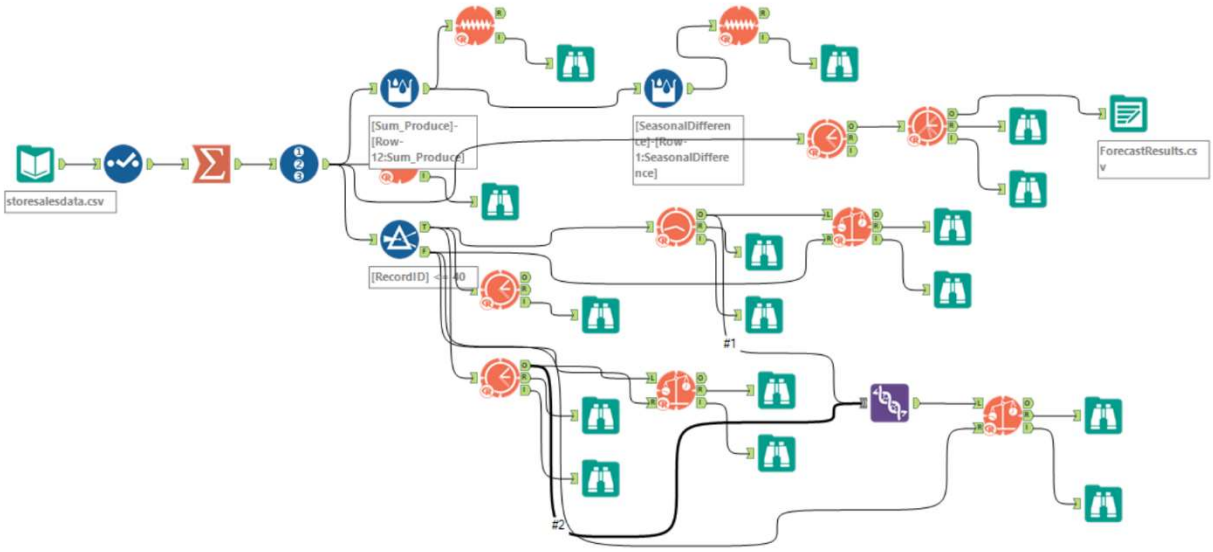

Figure 15: Task 3 Workflow

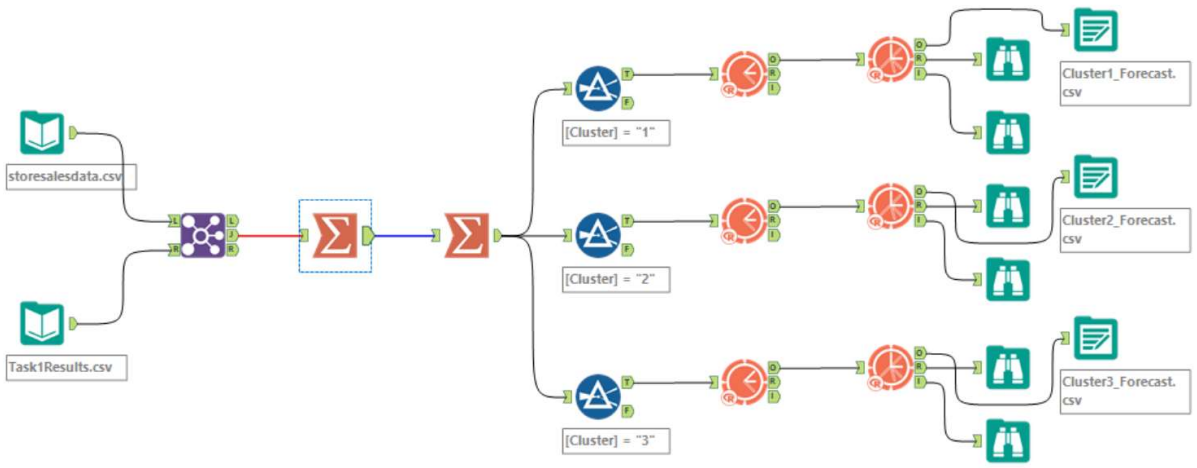

Figure 16: Task 3.1 Workflow

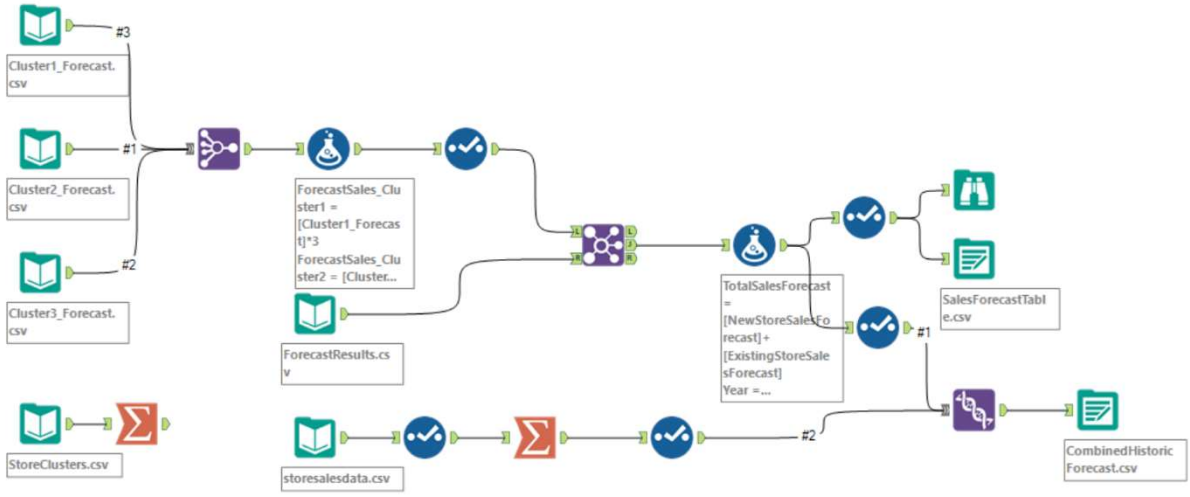

Figure 17: Task 3.2 Workflow