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?

K-Means Cluster Assessment Report

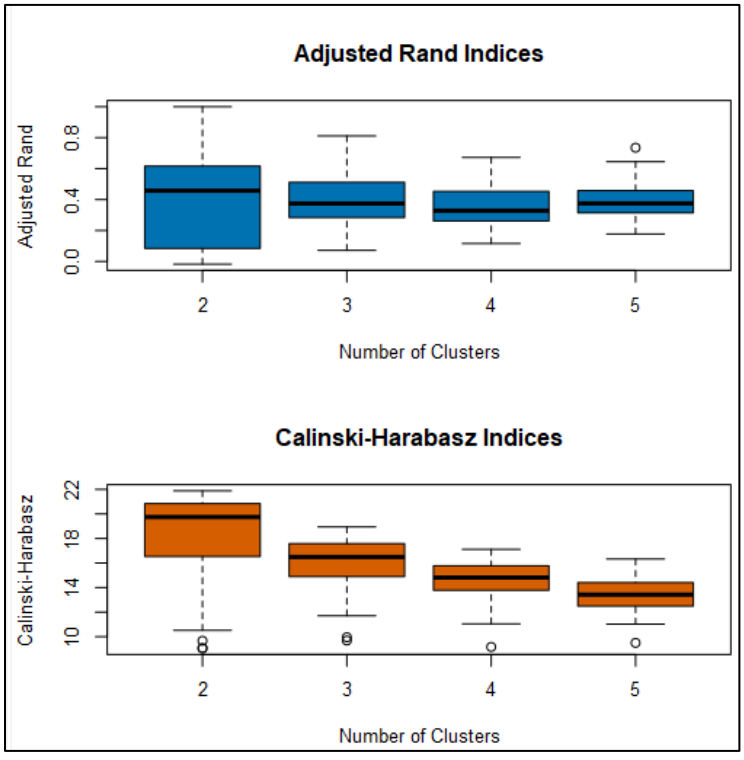
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5
Minimum	-0.017008	0.071616	0.116349	0.177641
1st Quartile	0.086202	0.287111	0.263166	0.31469
Median	0.457611	0.374366	0.328232	0.375368
Mean	0.394123	0.40335	0.358394	0.394509
3rd Quartile	0.607578	0.510747	0.447002	0.45718
Maximum	1	0.811418	0.672495	0.735269

Calinski-Harabasz Indices:

	2	3	4	5
Minimum	9.056197	9.683921	9.162814	9.495153
1st Quartile	16.705863	14.902161	13.821953	12.489711
Median	19.749349	16.487413	14.813432	13.421335
Mean	18.256151	16.035768	14.557953	13.356815
3rd Quartile	20.834246	17.560295	15.777219	14.375168
Maximum	21.878321	18.941224	17.112162	16.325684



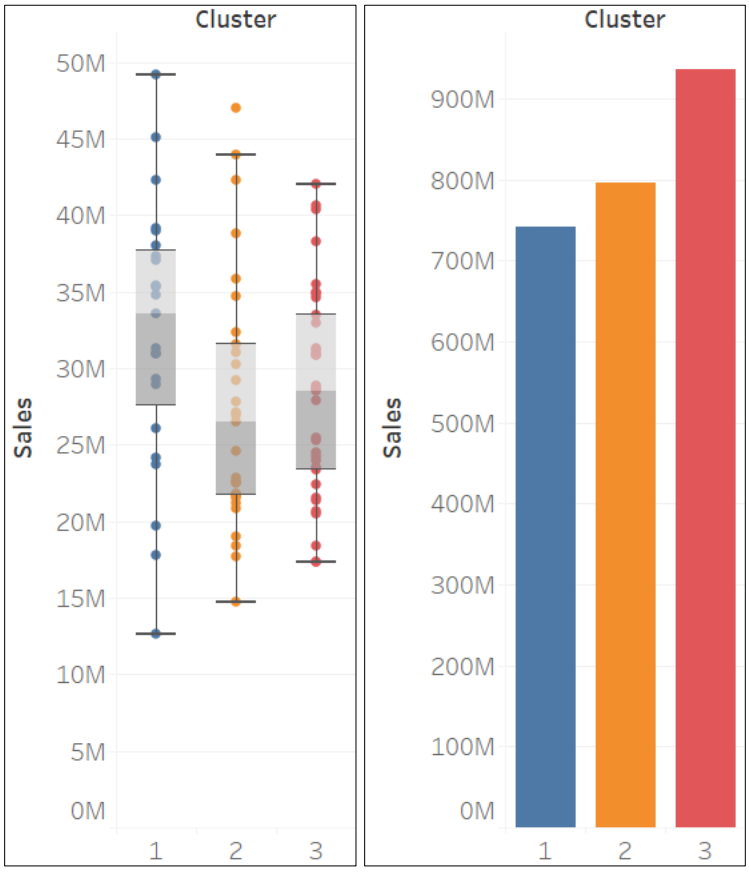
Analyzing with K-Centroid Diagnostics tool, seems n_clusters=3 is the best since it has enough mean adjusted rand and tight range.

2. How many stores fall into each store format?

Cluster	CountDistinctNonNull_Store
1	23
2	29
3	33

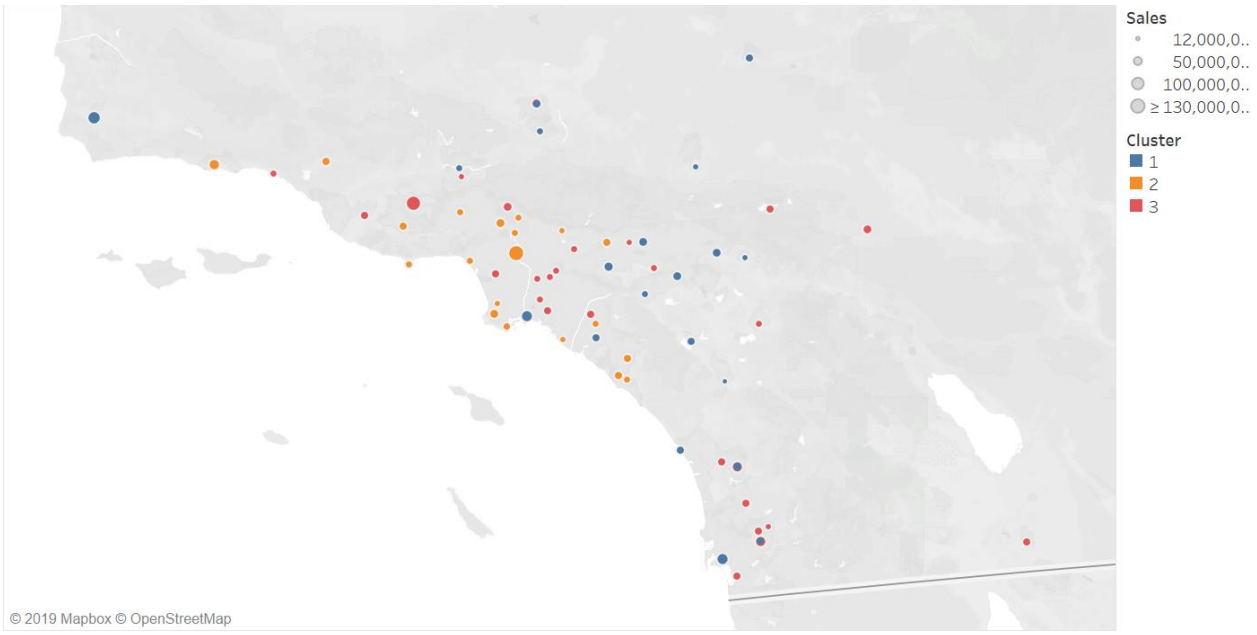
Using K-centroids cluster analysis, the final result is shown above.

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



Cluster 3 has much tighter sales range, but with higher total sales.

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.



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.

   We predict cluster using Decision Tree, Forest Model and Boosted Model. The model comparison report is shown below.

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	0.8235	0.8543	0.8000	0.6667	1.0000
Forest	0.8235	0.8251	0.7500	0.8000	0.8750

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 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 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

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

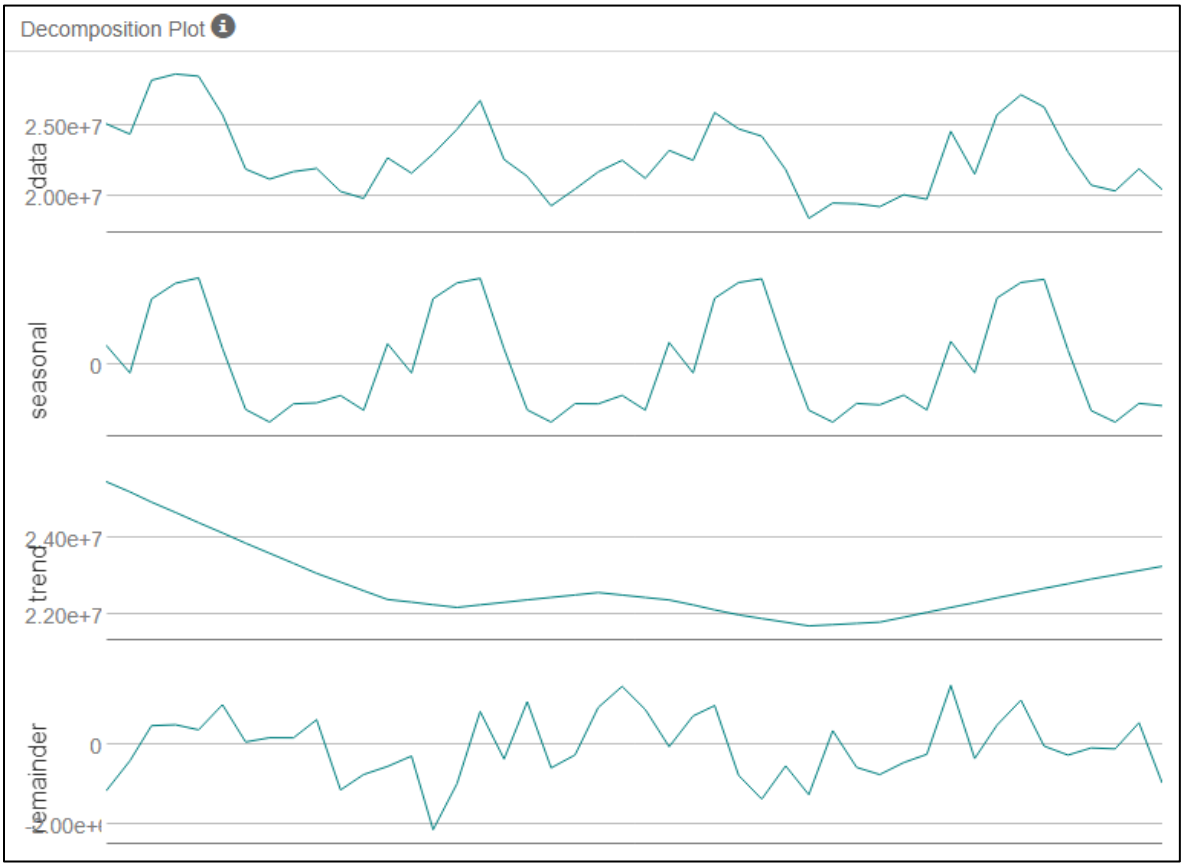
Judge from the report, the Boosted model has the high accuracy and high F1-score. Thus, we will use the Boosted model for store cluster prediction.

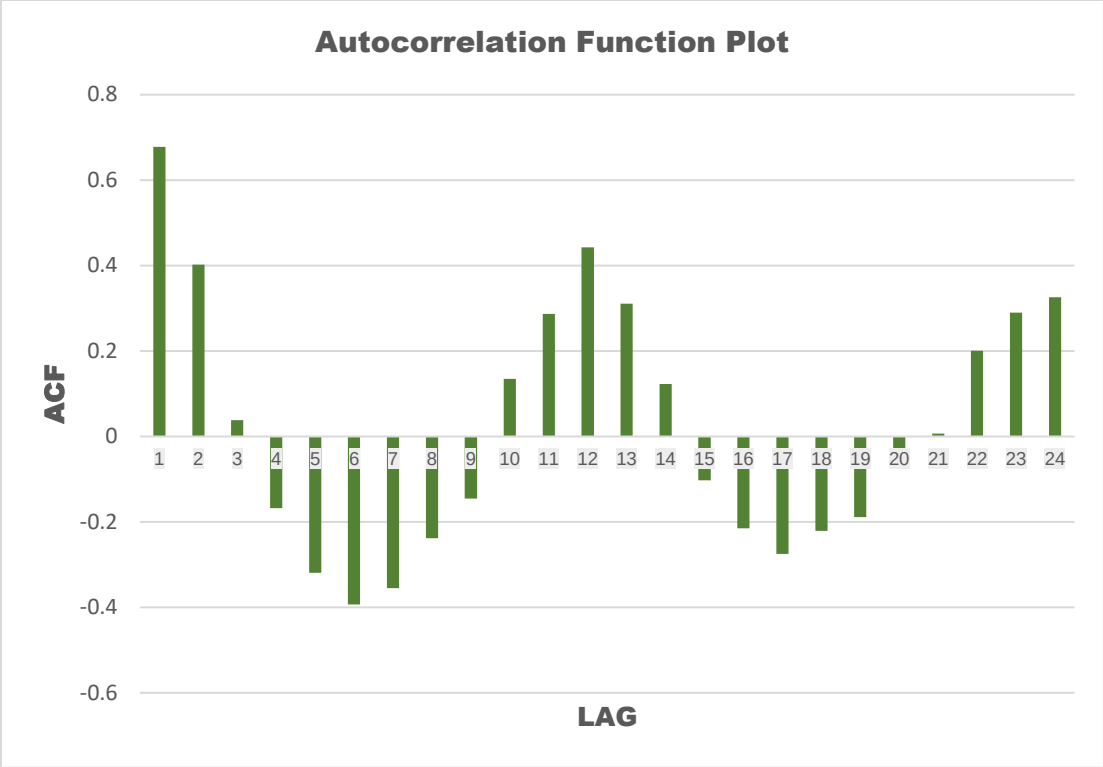
2. What format do each of the 10 new stores fall into? Please fill in the table below.

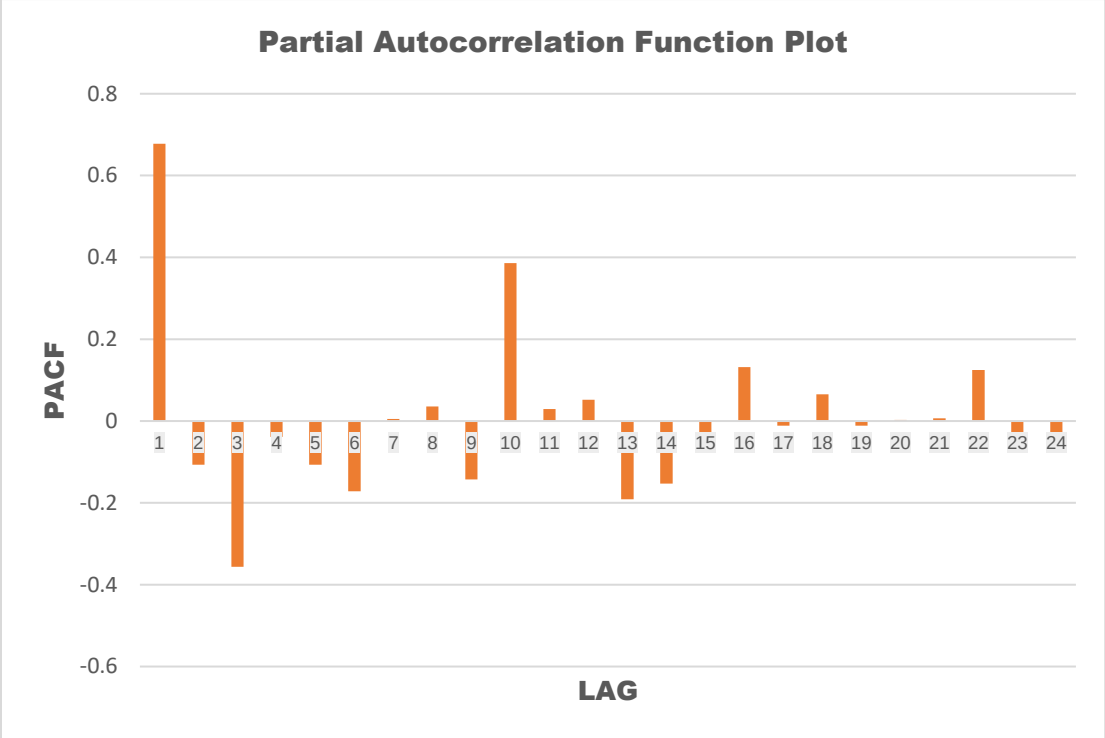
Store Number	***Segment***
S0086	*1*
S0087	*2*
S0088	*3*
S0089	*2*
S0090	*2*
S0091	*1*
S0092	*2*
S0093	*1*
S0094	*2*
S0095	*2*

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, we decomposed the time series plot. We can find that there is no obvious trend, a clear periodic seasonality and increased error element. Thus, the ETS model should be ETS(M,N,M). From the Autocorrelation plot, there is a high serial correlation and seems

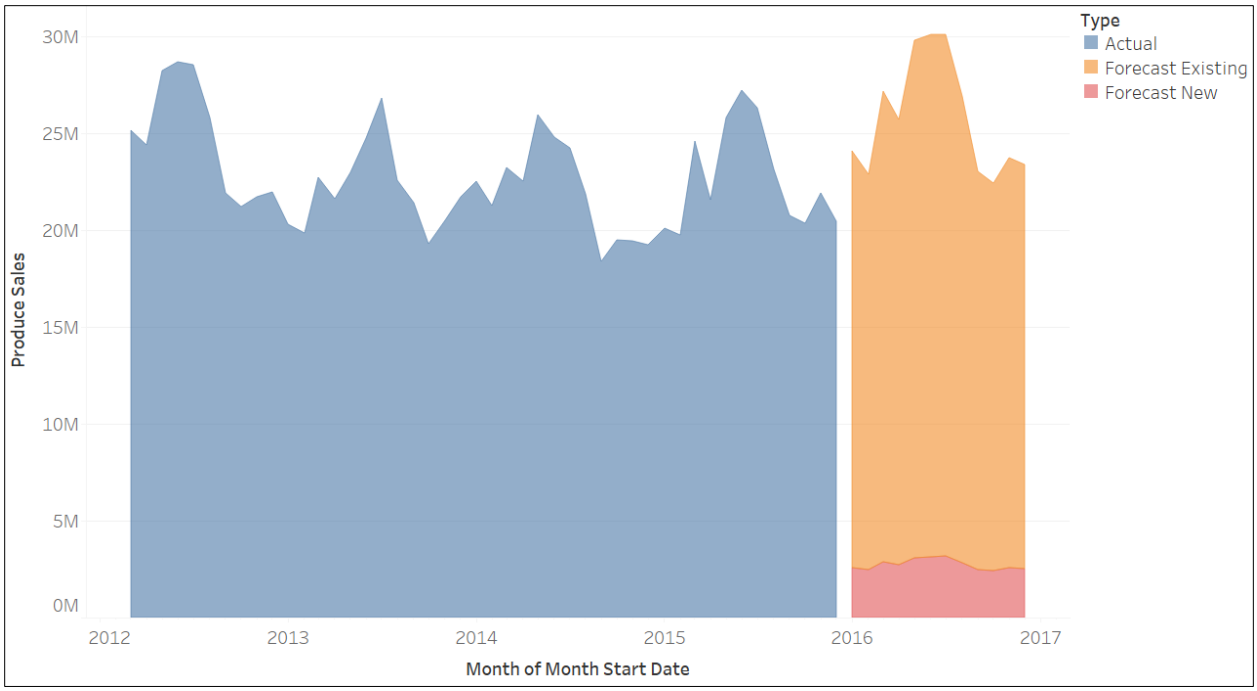
to be seasonal. The PACF plot also indicates seasonal trend. The ACF plot indicates an exponential decay and sine oscillation. This indicates seasonal difference. Thus we choose ARIMA (0,1,1)(0,1,1) [12] model.

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	1983592.6926	2226512.5538	1983592.6926	8.4729	8.4729	1.2691
ARIMA	2878344.1382	3061362.1418	2878344.1382	12.5815	12.5815	1.8416

Judge from the comparison table above, it can be observed that ETS model has lower ME, RMSE, MASE, MPE, MAPE and MASE values. Consequently, ETS model is chosen for forecasting.

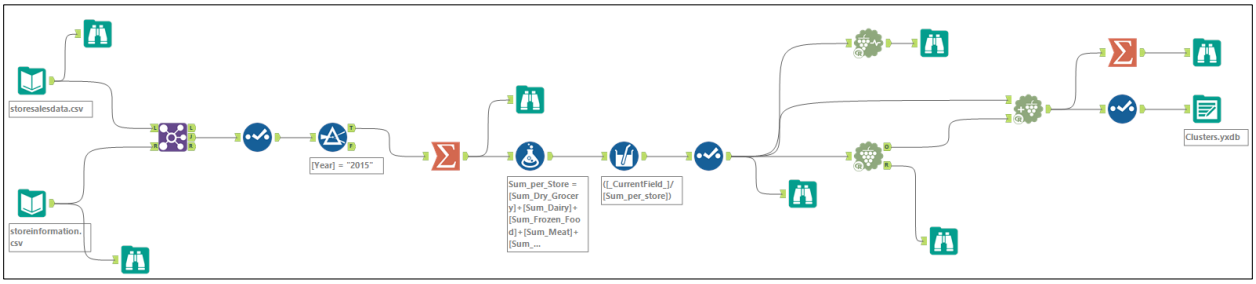
2. Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.

Year	Month	New Stores	Existing Stores
2016	1	2580411	21539936
2016	2	2494753	20413771
2016	3	2876480	24325953
2016	4	2742890	22993466
2016	5	3103562	26691951
2016	6	3124176	26989964
2016	7	3168777	26948631
2016	8	2820029	24091579
2016	9	2491912	20523492
2016	10	2442136	20011749
2016	11	2551509	21177435
2016	12	2520758	20855799

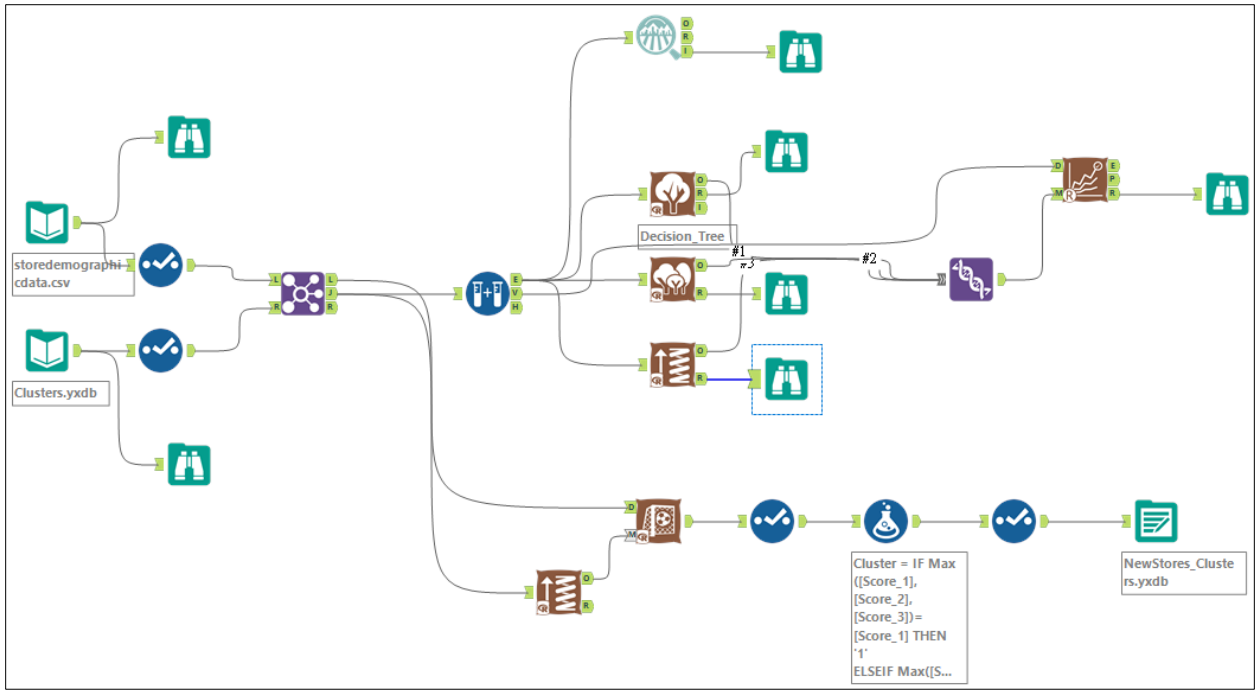


Appendix: Alteryx Workflows

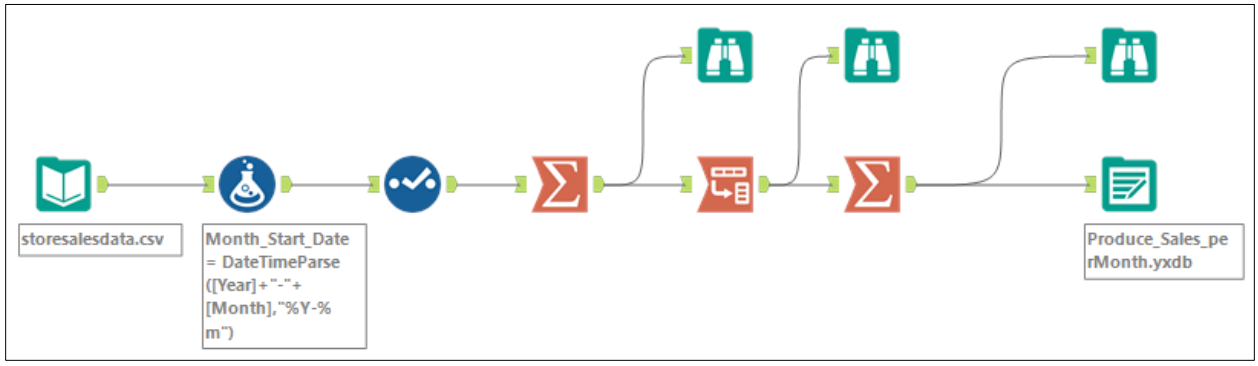
Task 1

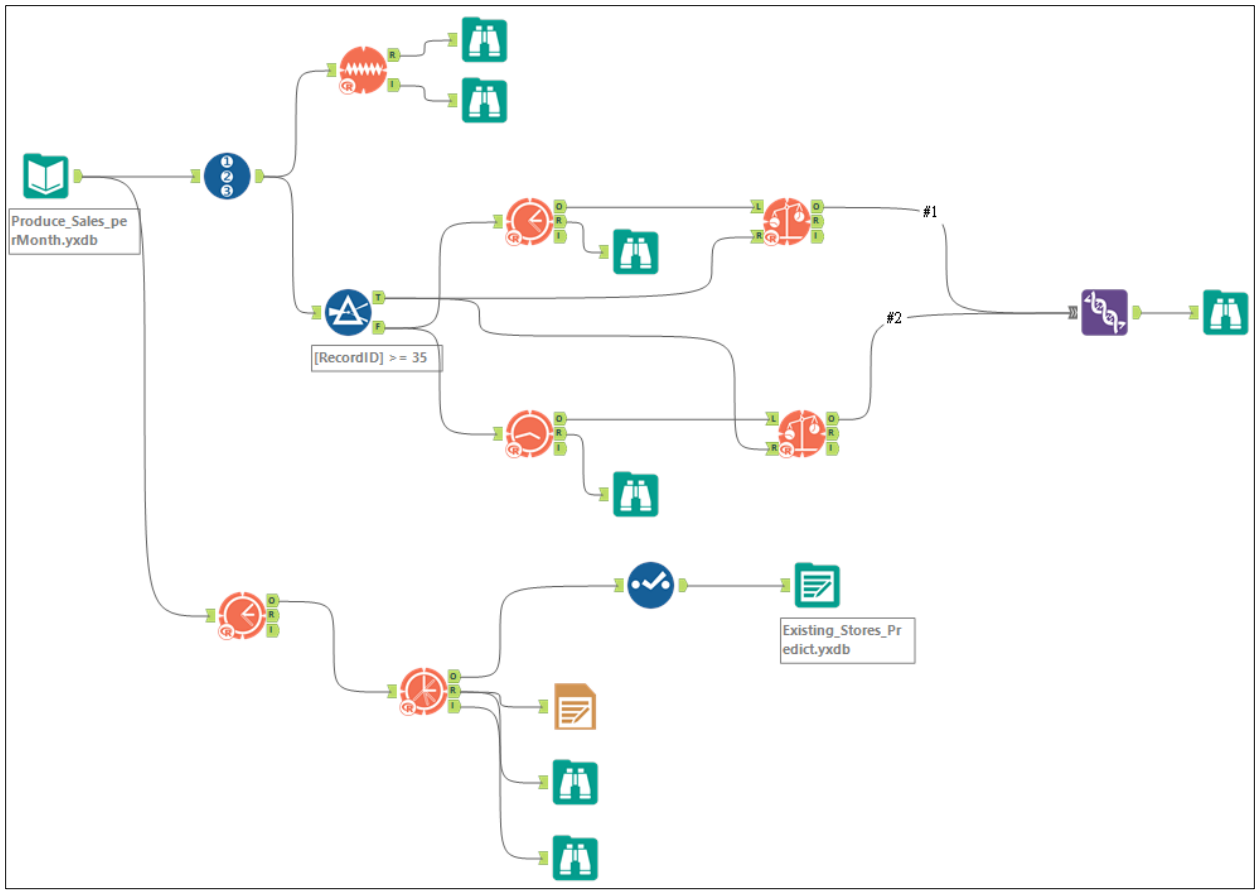


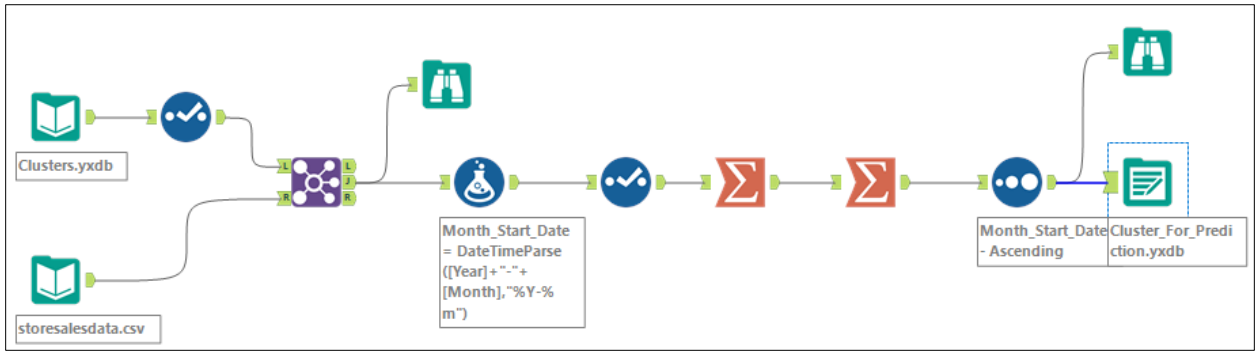
Task 2

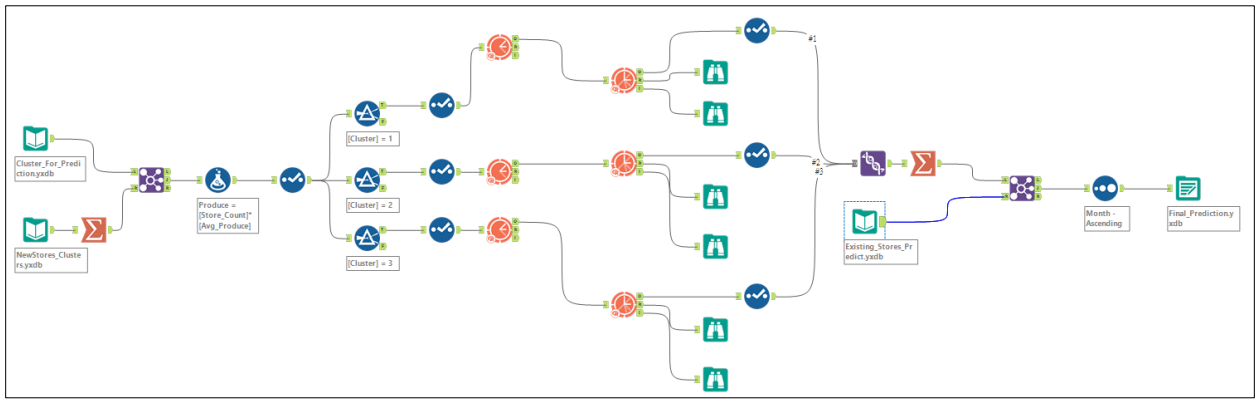

storedemographicdata.csv
Clusters.yxdb
Decision_Tree
#1
#2
#3
Cluster = IF Max([Score_1],[Score_2],[Score_3])=[Score_1] THEN '1' ELSEIF Max([S...
NewStores_Clusters.yxdb

Task 3


storesalesdata.csv
Month_Start_Date = DateTimeParse([Year]+"-"+[Month],"%Y-%m")
Produce_Sales_perMonth.yxdb


Produce_Sales_perMonth.yxdb
[RecordID] >= 35
#1
#2
Existing_Stores_Predict.yxdb


Clusters.yxdb
storesalesdata.csv
Month_Start_Date = DateTimeParse([Year]+"-"+[Month],"%Y-%m")
Month_Start_Date - Ascending
Cluster_For_Prediction.yxdb


Cluster_For_Prediction.yxdb
NewStores_Clusters.yxdb
Produce = [Store_Count]* [Avg_Produce]
[Cluster] = 1
[Cluster] = 2
[Cluster] = 3
#1
#2
#3
Existing_Stores_Predict.yxdb
Month - Ascending
Final_Prediction.yxdb