COMBINING PREDICTIVE TECHNIQUES

Predictive Analytics with Alteryx & Tableau

TABLE OF CONTENTS

COMBINING PREDICTIVE TECHNIQUES

Business and Data Understanding

Your company currently has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores, similarly, shipping the same amount of product to each store, This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others

The 10 new stores opening at the beginning of the year, want to determine the format for each new store, there is currently no sales data for these new stores. Lastly the company want to have more have an accurate monthly forecast, on fresh produce due to the short life span.

To summarize the 3 tasks:

1. Determine Store Format

- Determine the optimal number of store formats based on sales data
- Sum sales data by StoreID and Year
- Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
- Use only 2015 sales data.
- Use a K-means clustering model.
- Segment the 85 current stores into the different store formats.
- Use the StoreSalesData.csv and StoreInformation.csv files.

2. Dertermine the Store Format for New Stores

- evelop a model that predicts which segment a store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
- Use a 20% validation sample with Random Seed = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
- Use the model to predict the best store format for each of the 10 new stores.
- Use the StoreDemographicData.csv file, which contains the information for the area around each store.

3. Forecasting Produce Sales

- You've been asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores

- 6 month holdout sample for the TS Compare tool (this is because we do not have that much data so using a 12 month holdout would remove too much of the d

Store Format for Existing Stores

To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection to better match local demand.

Filter the stores sales data 2015, as required then proceed to join both store information and store sales data, Next was to aggregate each produce by store and get a total sales, to find the percentage of sales for each produce type.

K-Means Test

The Optimal number of store formats is 3, This is derived from the K-means report the Adjusted Rand and Calinski Indices show that 3 clusters has the highest Median with smaller variations in spread and is more compact.

Report

K-Means Cluster Assessment Report

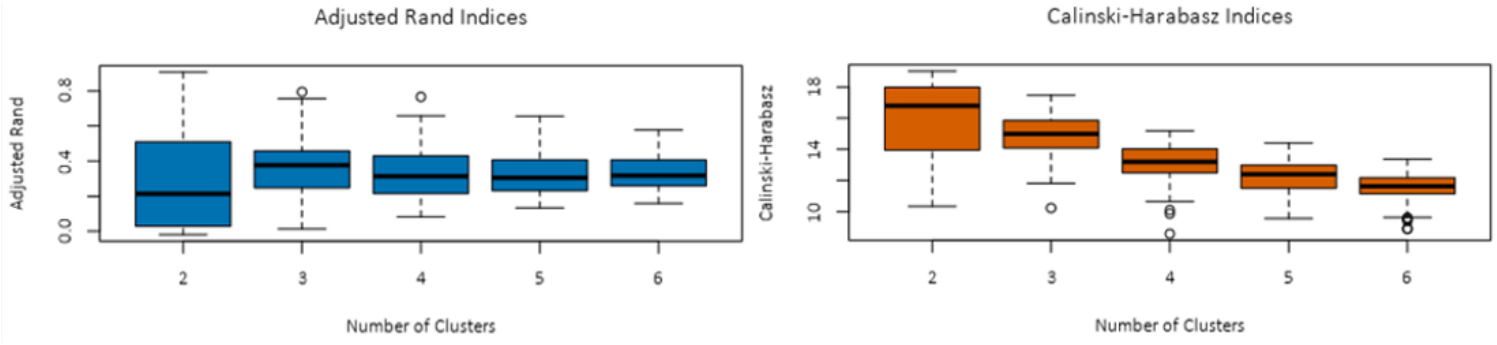
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.018714	0.013987	0.08275	0.133722	0.159965
1st Quartile	0.031182	0.250602	0.21908	0.233746	0.260979
Median	0.213435	0.375916	0.313216	0.304885	0.316444
Mean	0.286044	0.370071	0.33788	0.329179	0.32997
3rd Quartile	0.509439	0.455581	0.427949	0.405808	0.402038
Maximum	0.905582	0.793568	0.765626	0.65533	0.5774

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	10.33595	10.23213	8.584628	9.562864	8.890057
1st Quartile	13.95107	14.11439	12.497796	11.50724	11.154471
Median	16.78626	14.98704	13.195265	12.39996	11.629091
Mean	16.06439	14.89092	13.139321	12.258608	11.555379
3rd Quartile	17.95275	15.84016	14.015594	12.984948	12.154413
Maximum	18.99598	17.4659	15.176014	14.398966	13.364396



For each Segment / Cluster we have the follow store counts.

Cluster 1: 25, Cluster 2: 35, Cluster 3: 25

Report

Summary Report of the K-Means Clustering Solution cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Percent_Dry_Grocery + Percent_Dairy + Percent_Frozen_Food + Percent_Meat + Percent_Produce + Percent_Floral + Percent_Deli + Percent_Bakery + Percent_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.099985	4.823871	2.191566
2	35	2.475018	4.412367	1.947298
3	25	2.289004	3.585931	1.72574

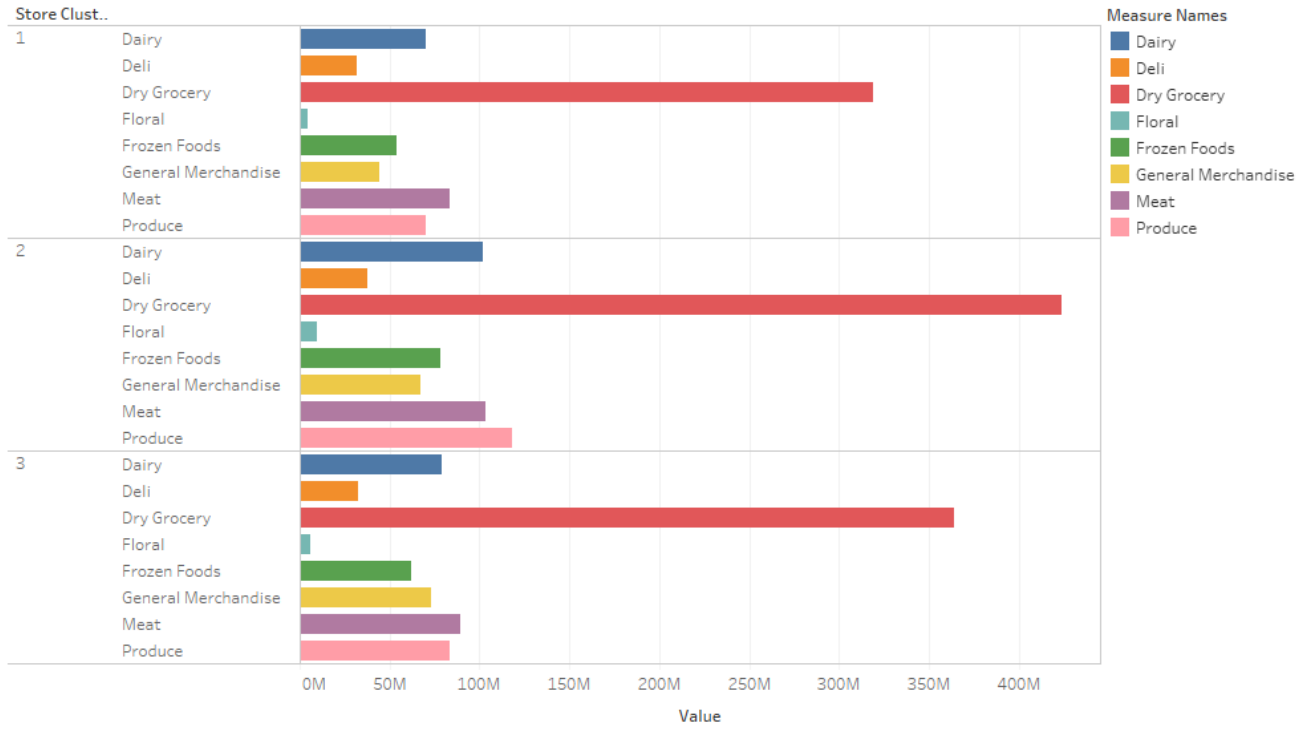
Convergence after 8 iterations.
Sum of within cluster distances: 196.35034.

	Percent_Dry_Grocery	Percent_Dairy	Percent_Frozen_Food	Percent_Meat	Percent_Produce	Percent_Floral	Percent_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	Percent_Bakery	Percent_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

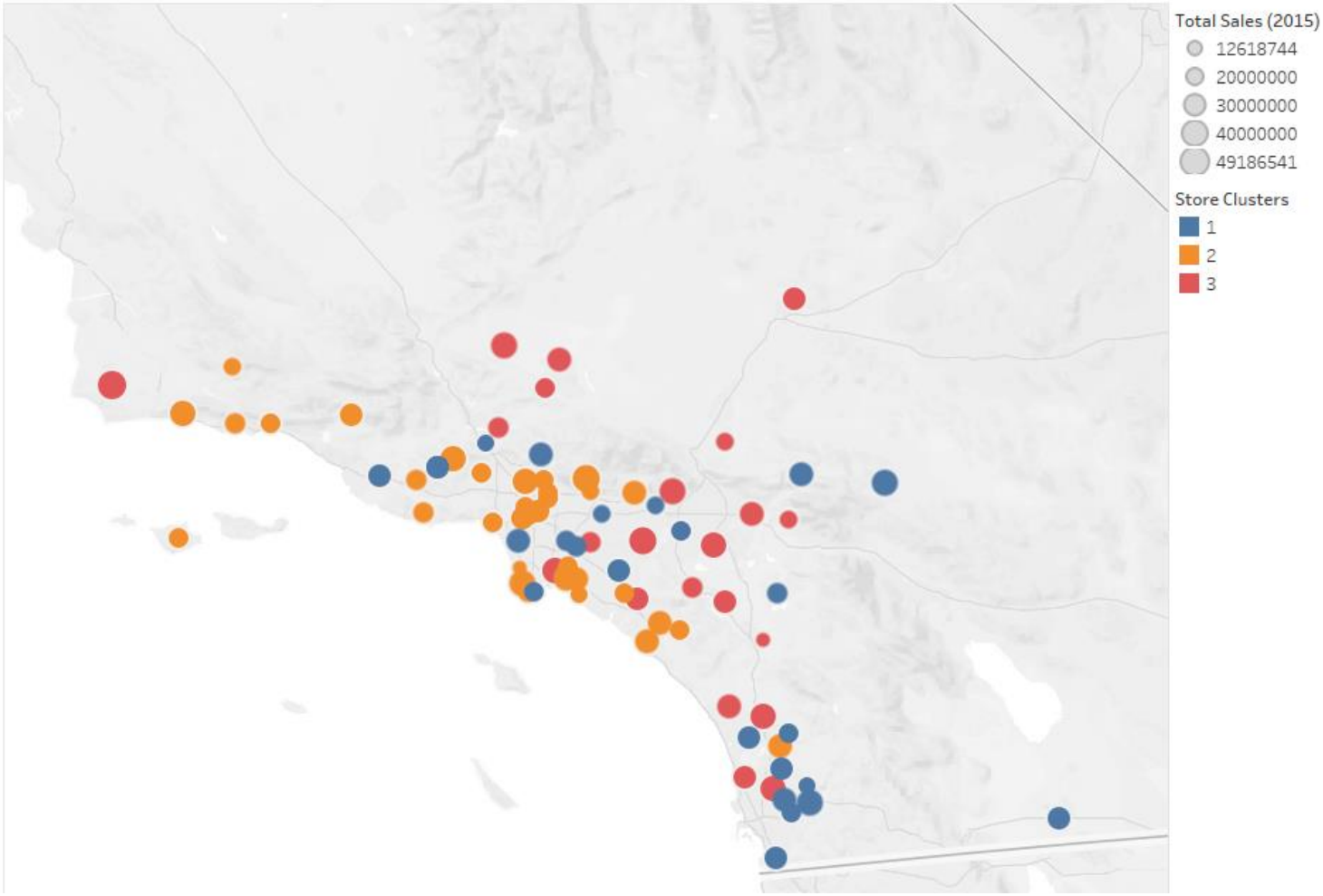
Looking at the summary of the K-Means Clustering solution, we can see how each cluster differs and percentage of each category sold.

Category Sales for each Cluster

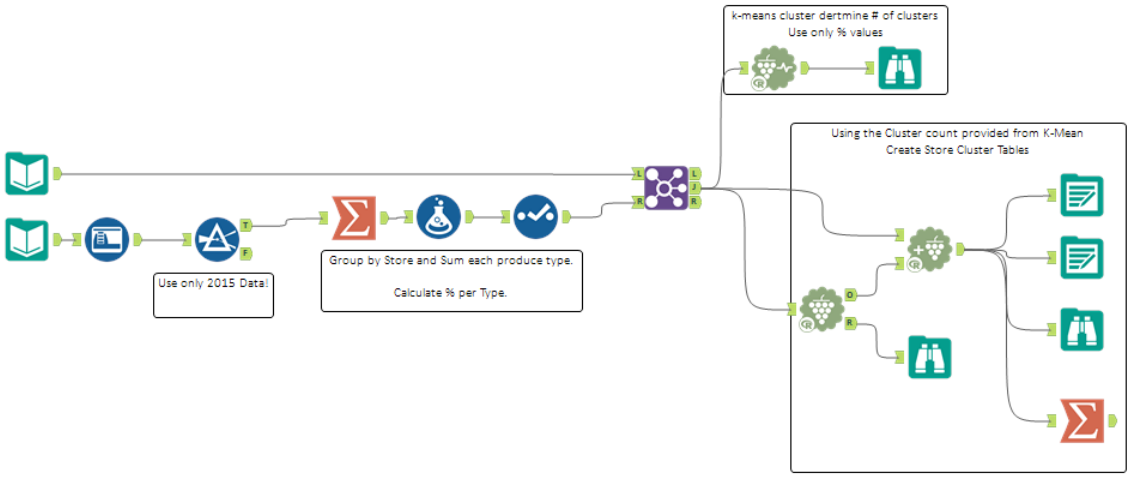


Cluster 2 stores sells more in general that Cluster 1 and 3 stores, with exceptions to Meat and General Merchandise.

Below is Cluster Map showing Store locations by Totals Sales and by Clusters.



Atleryx Workflow



Store Format for New Stores.

For the models all demographic variables were taken for each model type, with Cluster as the target variable

- Decision Tree Model
- Random Forest Model
- Boosted Model

80% from the set was used for the estimation sample, and 20% was used for the validation.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.6471	0.6667	0.5000	1.0000	0.5000
Random_Forest	0.7059	0.7917	0.3750	1.0000	1.0000
Boosted_Model	0.7647	0.8333	0.5000	1.0000	1.0000

Confusion matrix of Boosted_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	0
Predicted_2	2	5	0
Predicted_3	2	0	4

Confusion matrix of Decision_Tree

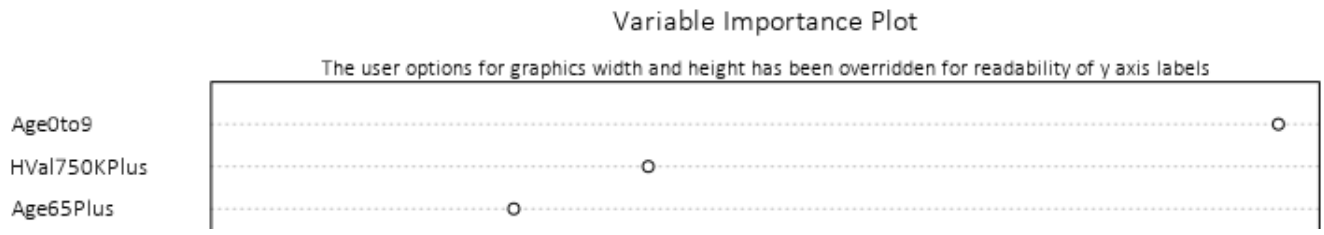
	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	2
Predicted_2	3	5	0
Predicted_3	1	0	2

Confusion matrix of Random_Forest

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	0
Predicted_2	3	5	0
Predicted_3	2	0	4

The Boosted model performs the best with an *Accuracy* of 76.4% and *F1* 83.3%

The variables with most significance are; *Age0to9*, *HVaI750Plus*, and *Age65Plus*.



Boosted model will be chosen to score against and choose the format for the New Stores.

New Store Clusters

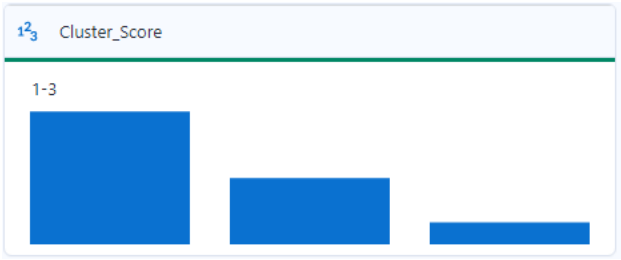
To predict which Clusters the New Store fall into the following formula was used after Scoring the models.

IF [Score_1] > [Score_2] AND [Score_1] > [Score_3] THEN "1 ELSEIF [Score_2] > [Score_1] AND [Score_2] > [Score_3] THEN "2" ELSE "3" ENDIF

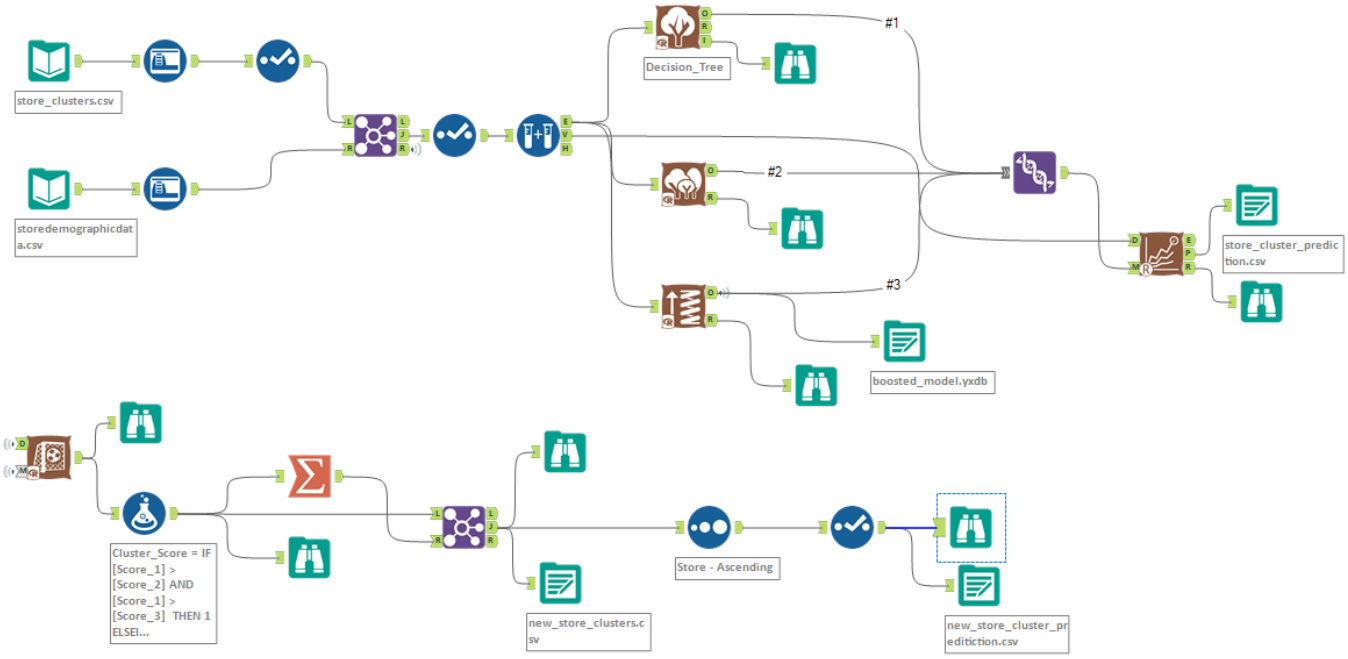
New Stores range from S0086 – S0095

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

- 10 New Stores Split into 3 Clusters
- Cluster 1 – 10%
- Cluster 2 – 60%
- Cluster 3 – 30%



Alteryx Workflow

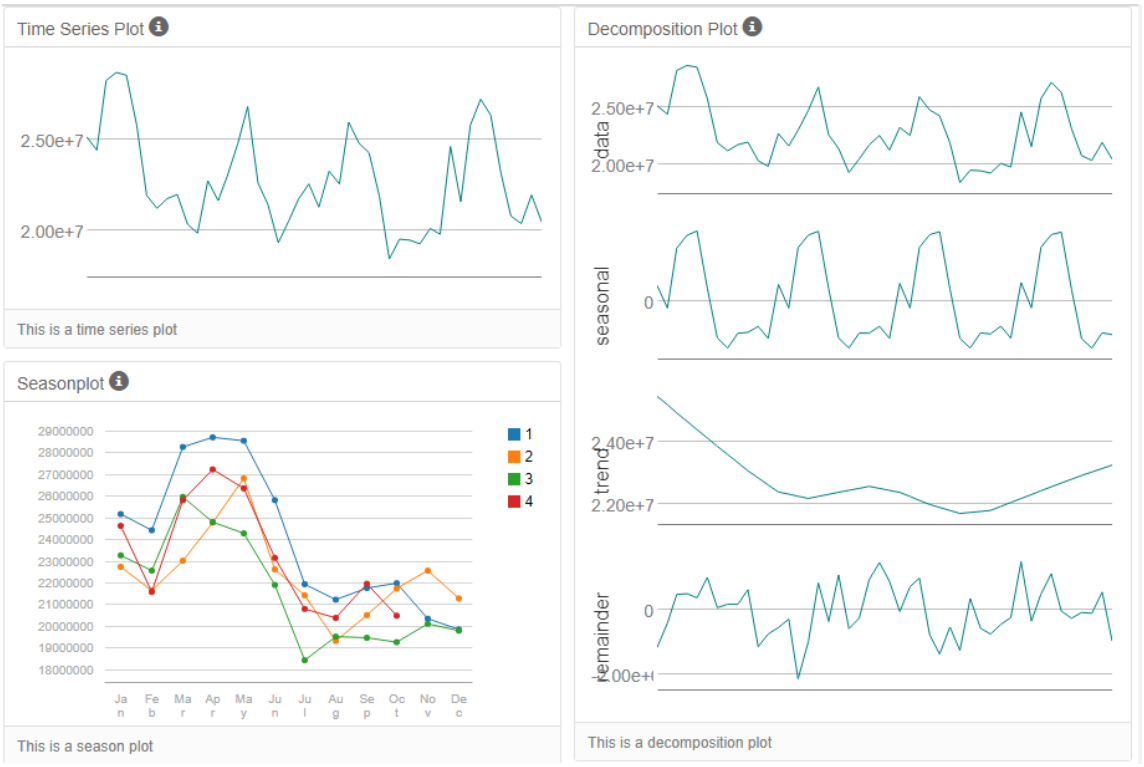


Predicting Produce Sales

For predicting the Produce Sales will use Time Series Forecasting (ETS and ARIMA Models), for new and existing stores. Will be using the available Store Sales Data, the predicted Clustering for the new Stores, and then combine the time for existing and new stores.

To Create the models, for need to group by year and month and aggregate the produce sales, next from the 46 records, 6 records are used as a holdout.

ETS(MNM) Plots, non-dampened.

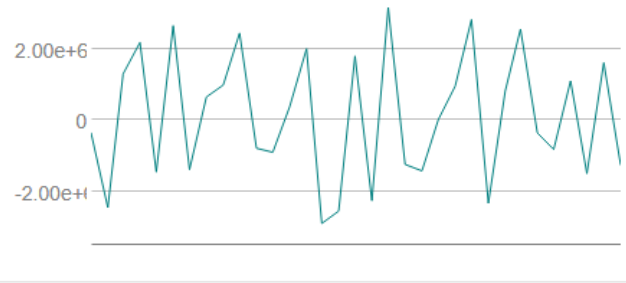


We can see that there is not trend in the sales, and that that there is multiplicative in the seasonality and the error is irregular.

Trend, Season and Error

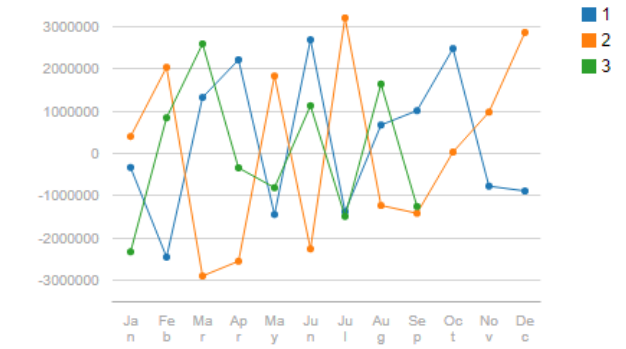
ARIMA is (1,0,0)(1,1,0)[12] Model,

Time Series Plot



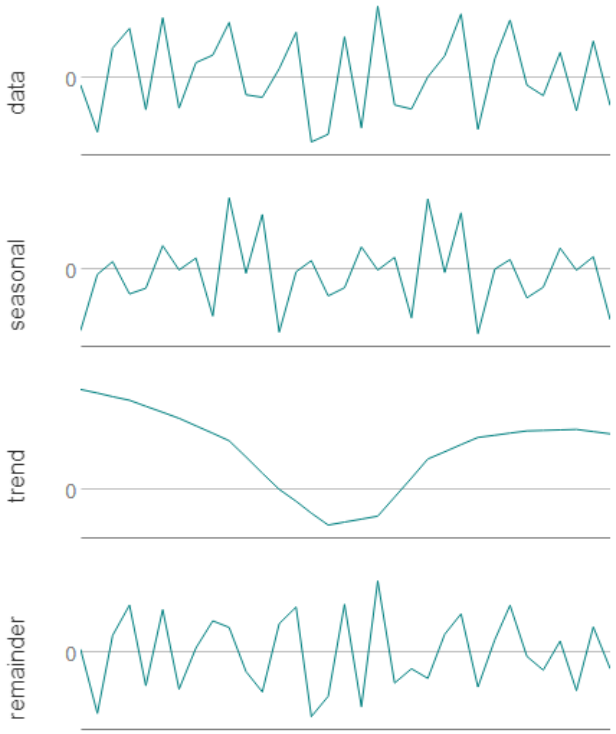
This is a time series plot

Seasonplot



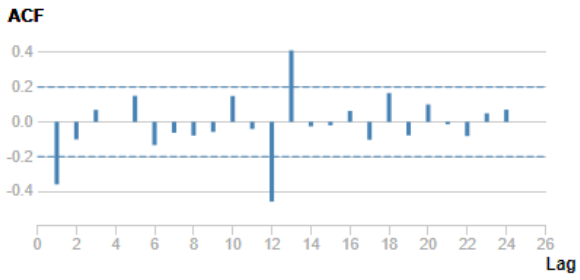
This is a season plot

Decomposition Plot



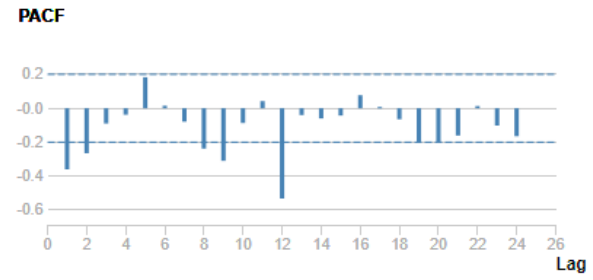
This is a decomposition plot

Autocorrelation Function Plot



This is an autocorrelation plot

Partial Autocorrelation Function Plot



This is an partial autocorrelation plot

Model Matrices Comparisons

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

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8325034	1042209.8528363	738087.5530941	-0.5465069	3.3006311	0.4120218	-0.1854462

Ljung-Box test of the model residuals:
Chi-squared = 15.0973, df = 12, p-value = 0.23616

Method: ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
3502.9443415	969051.6076376	787577.7006835	-0.1381187	3.4677635	0.4396486	0.0077488

Looking at the two models we can see that ETS as the lower RMSE, and the MASE values are equal.

Next to compare the holdouts.

Holdout Model Comparison

The comparison against the holdout samples us the TS compare tool. Can See that the ETS model has a better predictive quality, have lower values in most metrices. With a lower RMSE and MASE value falling below the generic 1.00 at .36

Report

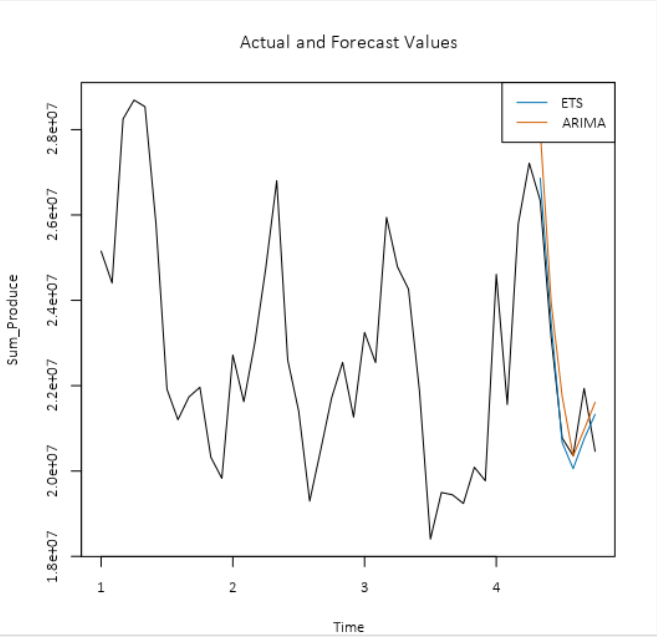
Comparison of Time Series Models

Actual and Forecast Values:

Actual	ETS	ARIMA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

Accuracy Measures:

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



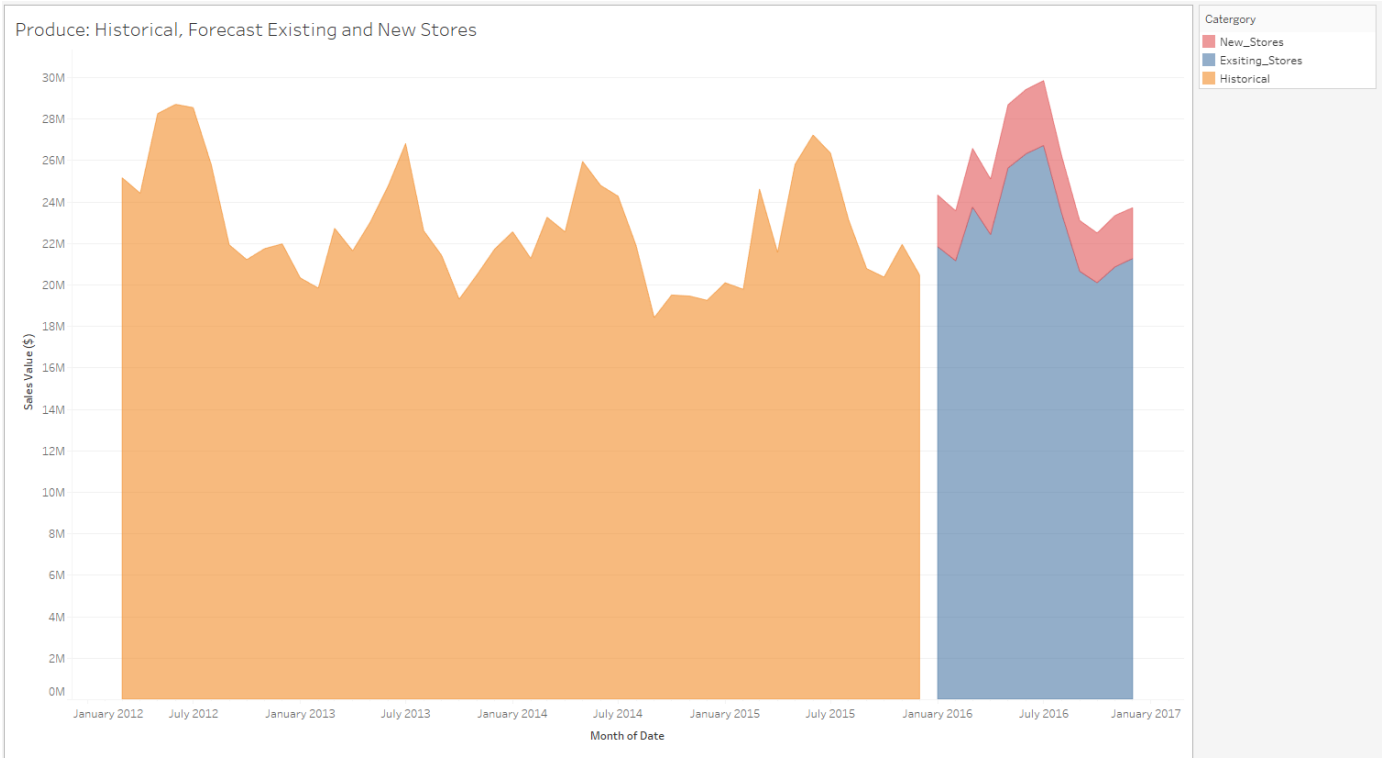
Takin into consideration the two models and the holdout validation The ETS(MNM) is a better model to predict the sales forecast.

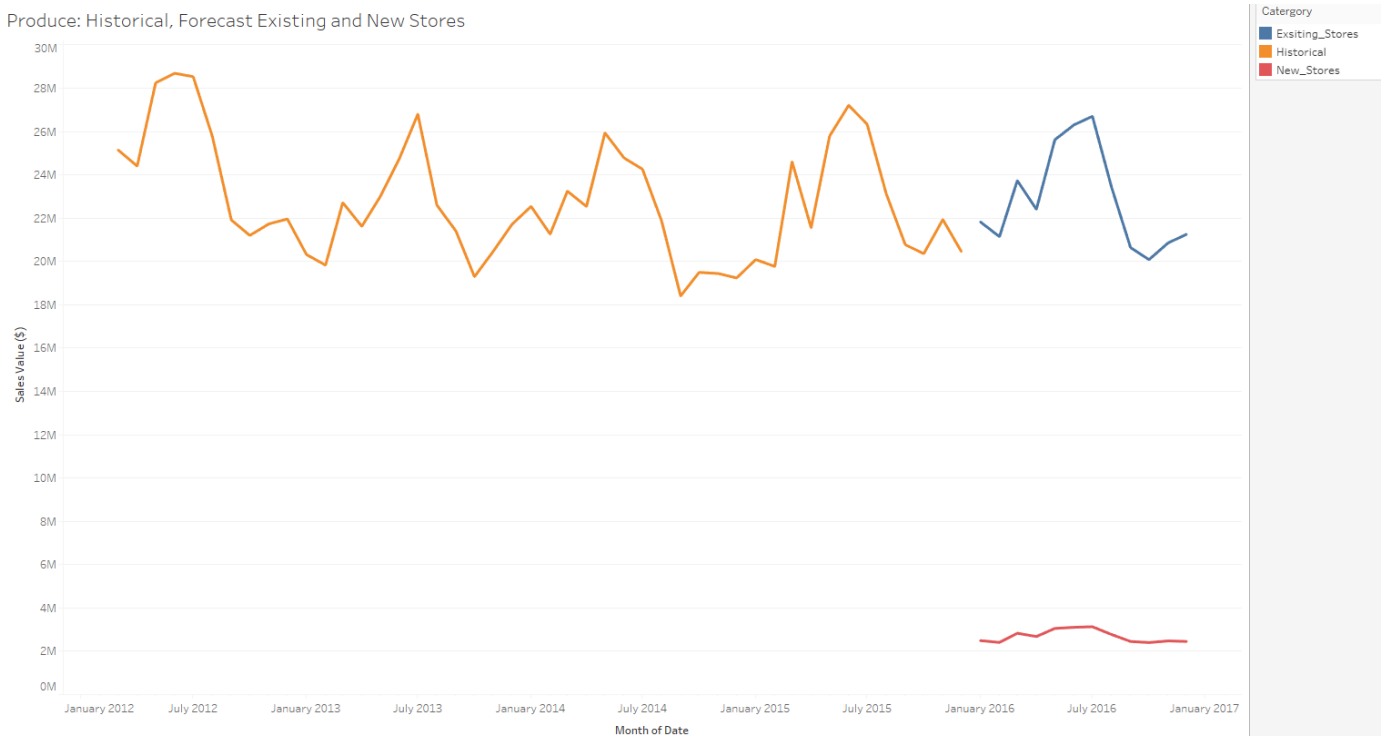
Sales Forecasts

The forecasted sales value for both the existing stores and new stores

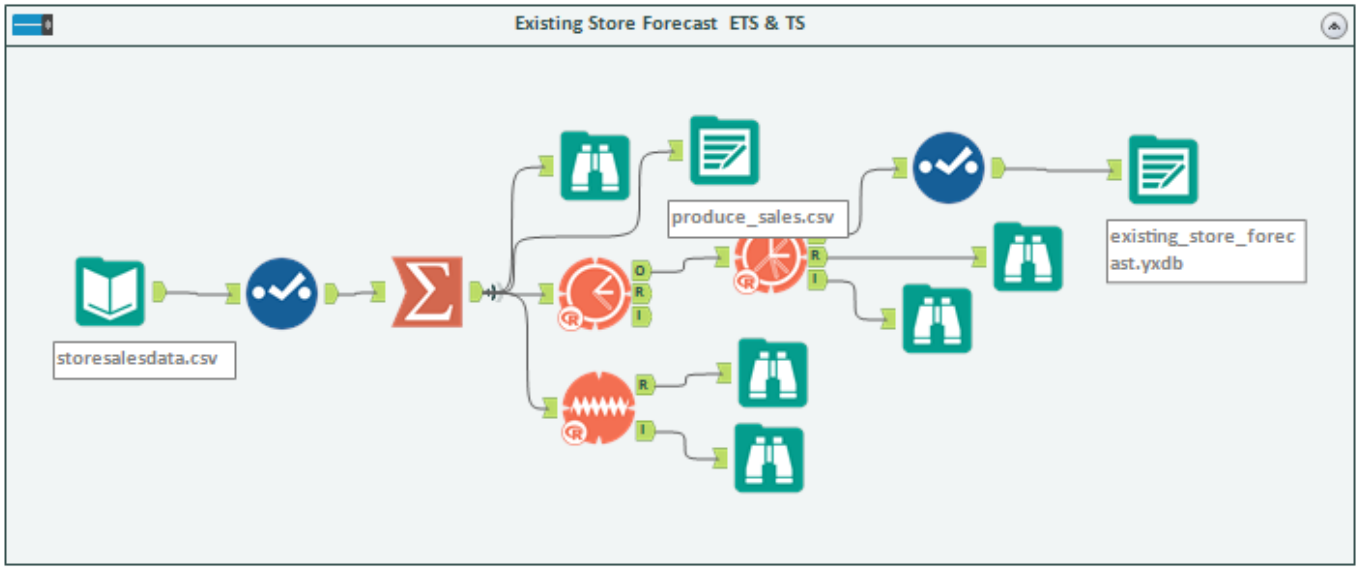
Month	New Stores	Existing Stores
Jan-16	2,491,319	21,829,060
Feb-16	2,408,385	21,146,330
Mar-16	2,833,157	23,735,687
Apr-16	2,679,433	22,409,515
May-16	3,054,886	25,621,829
Jun-16	3,106,152	26,307,858
Jul-16	3,132,699	26,705,093
Aug-16	2,776,154	23,440,761
Sep-16	2,451,566	20,640,047
Oct-16	2,401,772	20,086,270
Nov-16	2,477,302	20,858,120
Dec-16	2,452,170	21,255,190

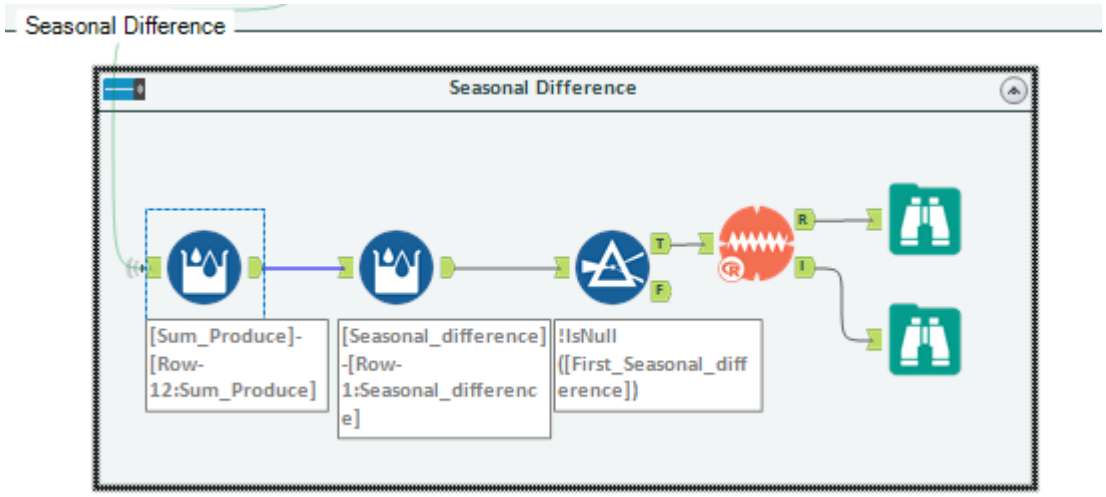
Graphic Representation of Historical and Forecast sales.

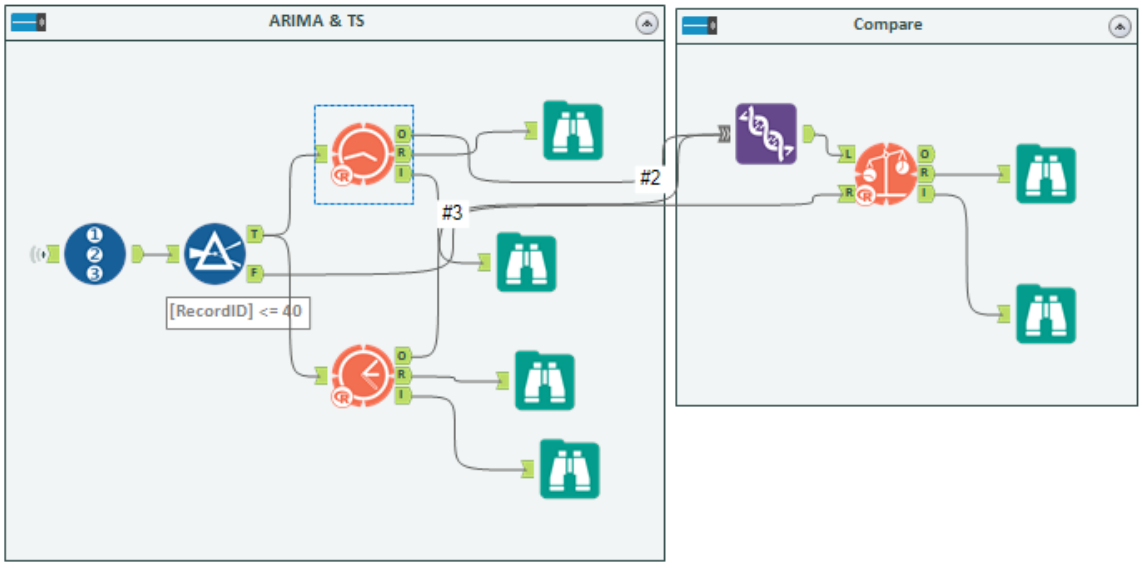


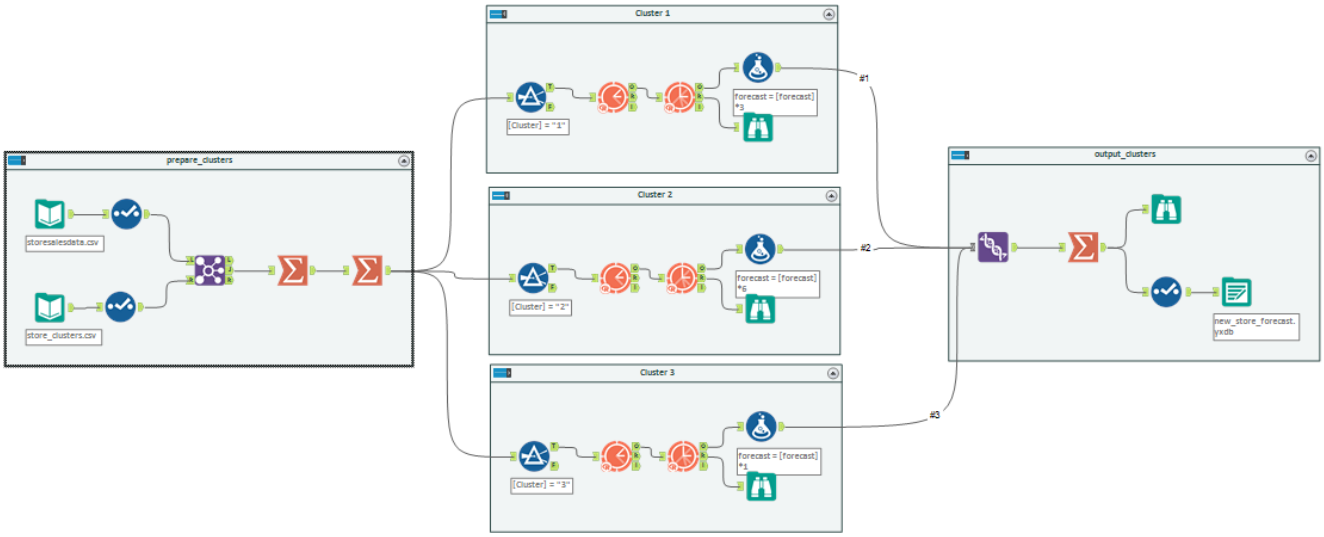

Produce: Historical, Forecast Existing and New Stores
Catergory
Exsiting_Stores
Historical
New_Stores
30M
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
Sales Value ($)
January 2012
July 2012
January 2013
July 2013
January 2014
July 2014
January 2015
July 2015
January 2016
July 2016
January 2017
Month of Date

Workflows


Existing Store Forecast ETS & TS
produce_sales.csv
existing_store_forecast.yxdb
storesalesdata.csv


Seasonal Difference
Seasonal Difference
[Sum_Produce]-[Row-12:Sum_Produce]
[Seasonal_difference]-[Row-1:Seasonal_difference]
!IsNull([First_Seasonal_difference])


ARIMA & TS
Compare
#2
#3
[RecordID] <= 40



Resources

https://knowledge.udacity.com/questions/507954

https://knowledge.udacity.com/questions/535025

https://knowledge.udacity.com/questions/202805

forecast_video_game_sales_answer_key