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

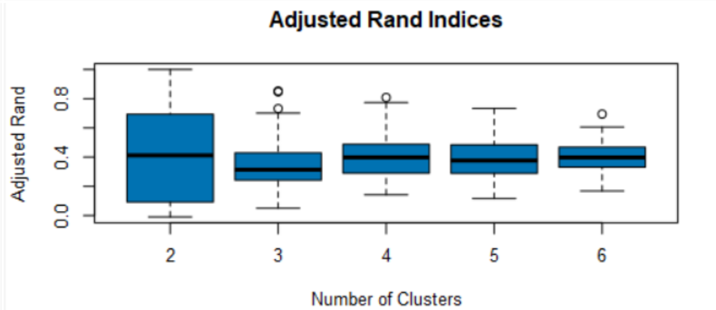
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.009603	0.049806	0.141931	0.116126	0.168334
1st Quartile	0.099849	0.24398	0.290669	0.28925	0.331267
Median	0.411974	0.313552	0.397509	0.376582	0.397359
Mean	0.397078	0.366329	0.395003	0.398799	0.397648
3rd Quartile	0.684312	0.422888	0.487787	0.482849	0.466522
Maximum	1	0.85296	0.808194	0.73384	0.694723

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	8.793333	9.518351	10.1893	10.76523	9.907697
1st Quartile	17.287672	15.068863	13.67247	12.47522	11.988572
Median	19.893032	16.341659	14.514	13.44709	12.565255
Mean	18.435697	16.093783	14.40569	13.39308	12.730988
3rd Quartile	20.893268	17.545663	15.33504	14.1877	13.645308
Maximum	21.992646	19.113662	16.8206	16.72381	15.424901



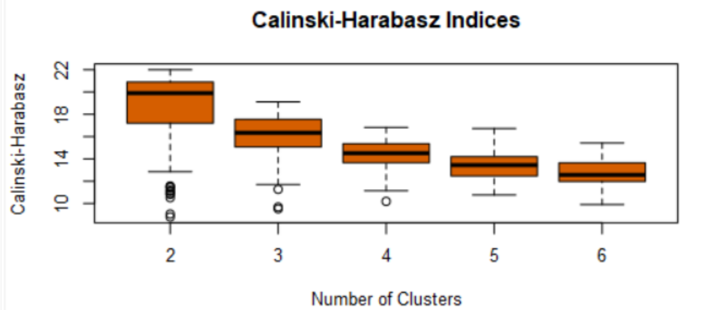


Figure 1 K-Means Cluster Report

Based on the K-means report, the optimal number of store formats is 2 when both the indices registered the highest median value. However, when 2 clusters were selected, there were more than 40 stores in each cluster. Therefore, the optimal number of store formats is 3 as it is stated in the supporting material that cluster must not have less than 20 and not over 40 stores.

2. **How many stores fall into each store format?**

Cluster 1 has 23 stores, cluster 2 has 29 stores 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	2.540086	4.475133	2.118708
3	33	2.115045	4.9262	1.702843

Figure 2 Cluster Information

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

Summary Report of the K-Means Clustering Solution Cluster

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + Sum_Dry_Grocery + Sum_Dairy + Sum_Frozen_Food + Sum_Meat + Sum_Produce + Sum_Floral + Sum_Deli + Sum_Bakery + Sum_General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475133	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Sum_Dry_Grocery	Sum_Dairy	Sum_Frozen_Food	Sum_Meat	Sum_Produce	Sum_Floral	Sum_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Sum_Bakery	Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

Figure 3 Summary Report of the K-Means Clustering

Based on the result shown above, cluster 1 sells a lot of General Merchandise compared to the other two clusters. Cluster 2 sells a lot of Produce compared to the other two clusters.

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.**

https://public.tableau.com/profile/emily6902#!/vizhome/P9_Combining_Predictive_TechniquesTask1-MapVisualization/Sheet3?publish=yes

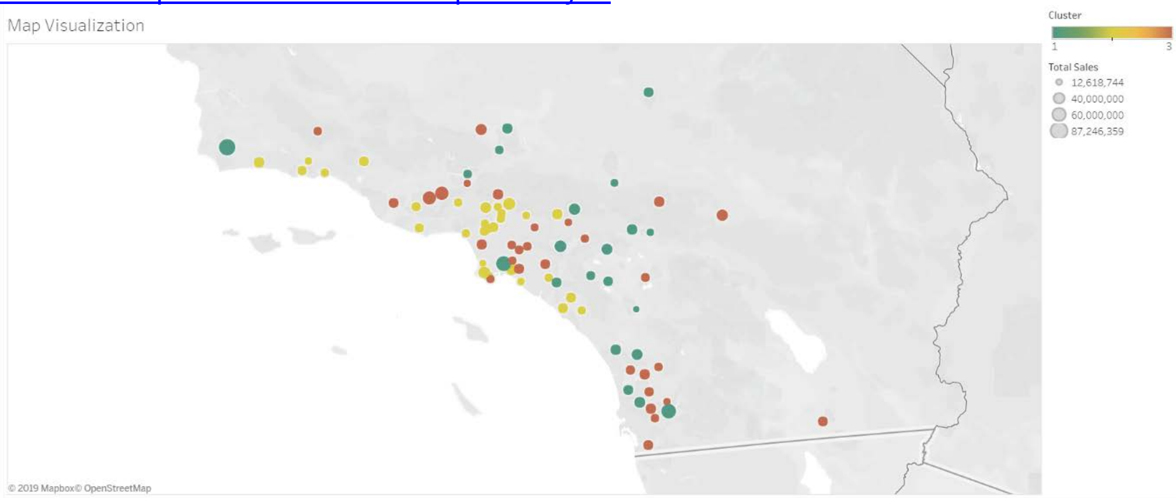


Figure 4 Location of the stores

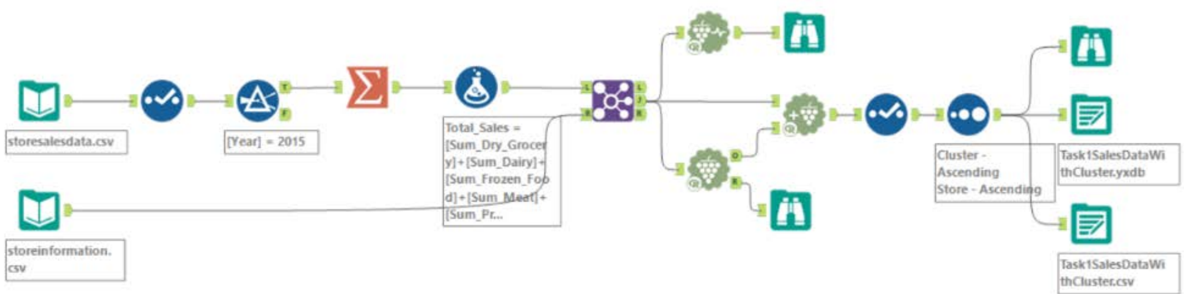


Figure 5 Alteryx Flow (Task 1)

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?**

The model comparison report shows the comparison between Forest Model, Decision Tree and Boosted Model. All models have the same accuracy but Boosted Model is chosen due to higher F1 value of 0.8889.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
store_FM	0.8235	0.8426	0.7500	1.0000	0.7778
store_DT	0.8235	0.8426	0.7500	1.0000	0.7778
store_Boosted	0.8235	0.8889	1.0000	1.0000	0.6667

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

Confusion matrix of store_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 store_DT

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

Confusion matrix of store_FM

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

Figure 6 Model Comparison Report

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

The 3 most important variables are Age0to9, HVal750KPlus and EdHSGrad.

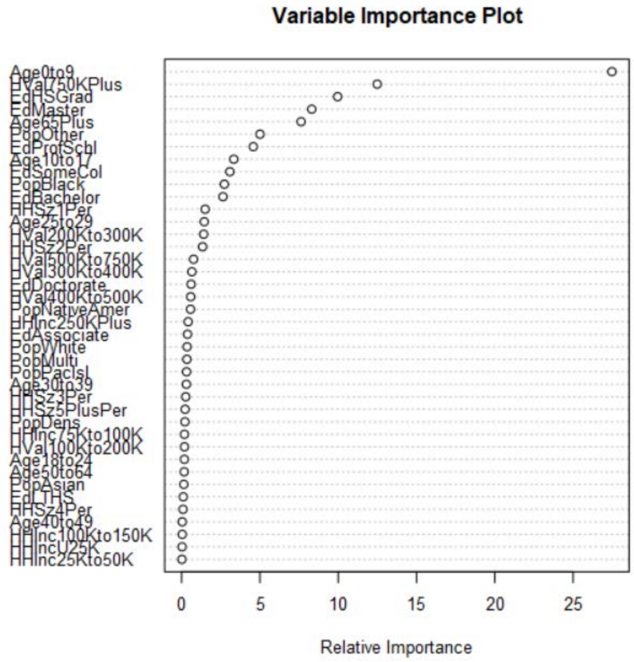


Figure 7 Variable Importance Plot

3. **What format do each of the 10 new stores fall into?**

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

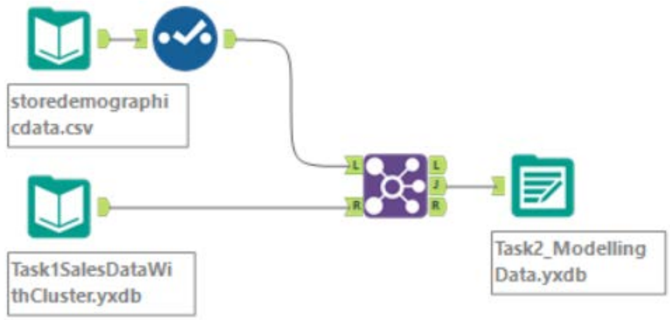


Figure 8 Alteryx Flow (Data Preparation)

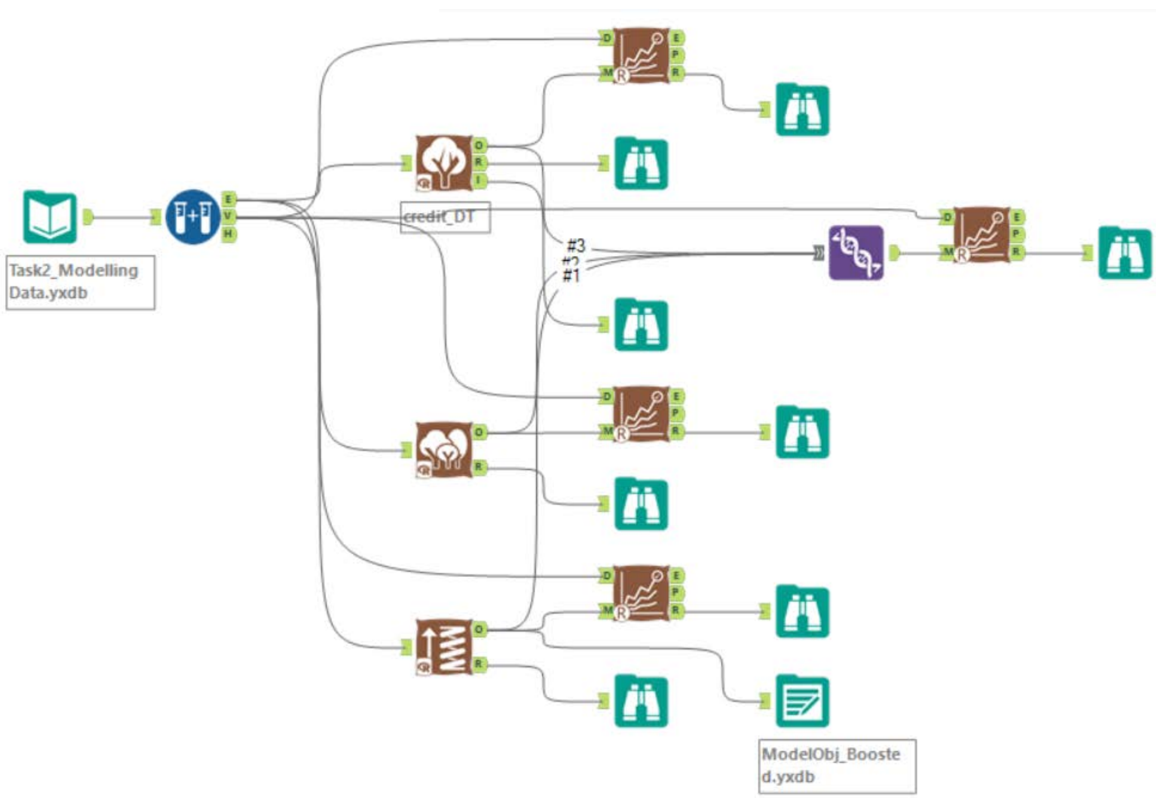


Figure 9 Alteryx Flow (Model Comparison)

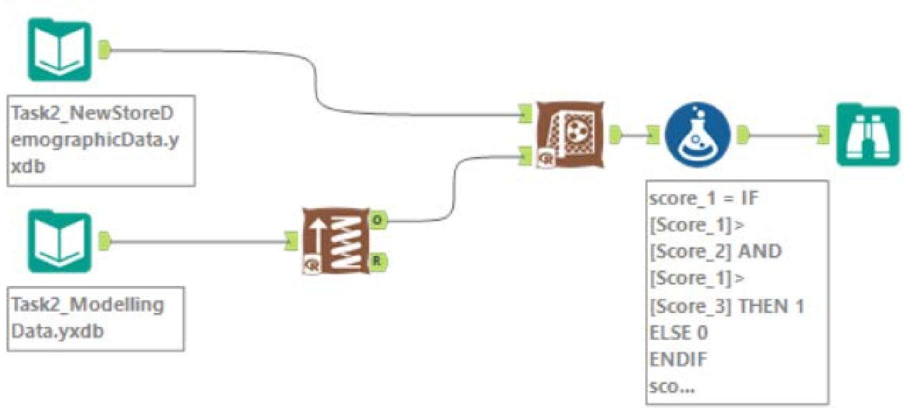


Figure 10 Alteryx Flow (Boosted)

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?**

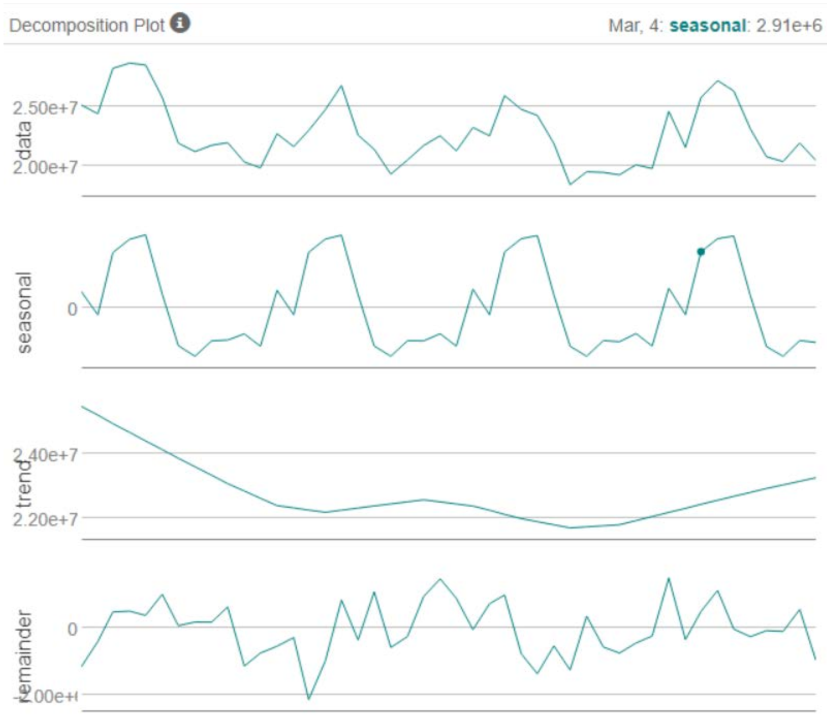


Figure 11 Decomposition Plot

The time series decomposition plot shown above allows us to observe the seasonality, trend and error/remainder terms of a time series. There is no clear trend so no trend component is included (N). The size of the seasonal fluctuations tends to increase or decrease with the level of time series so we apply it multiplicatively (M). The error plot is fluctuating between large and small errors over time, we apply it multiplicatively (M).

Auto options are chosen to train one ETS and one ARIMA model which gave us the optimal options as shown below.

TS Comparison:

ETS(M,N,M):

Actual and Forecast Values:

Actual	MAM
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
MAM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822

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

Actual and Forecast Values:

Actual	ARIMA
26338477.15	27997835.63764
23130626.6	23946058.0173
20774415.93	21751347.87069
20359980.58	20352513.09377
21936906.81	20971835.10573
20462899.3	21609110.41054

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

By comparing the forecast and actual results, we can see that ETS model's accuracy is higher with overall lower errors across all variable. The ETS model's RMSE (760,267.3) and MASE (0.3822) are lower.

Based on the above, ETS(M,N,M) is chosen as our forecasting mode.

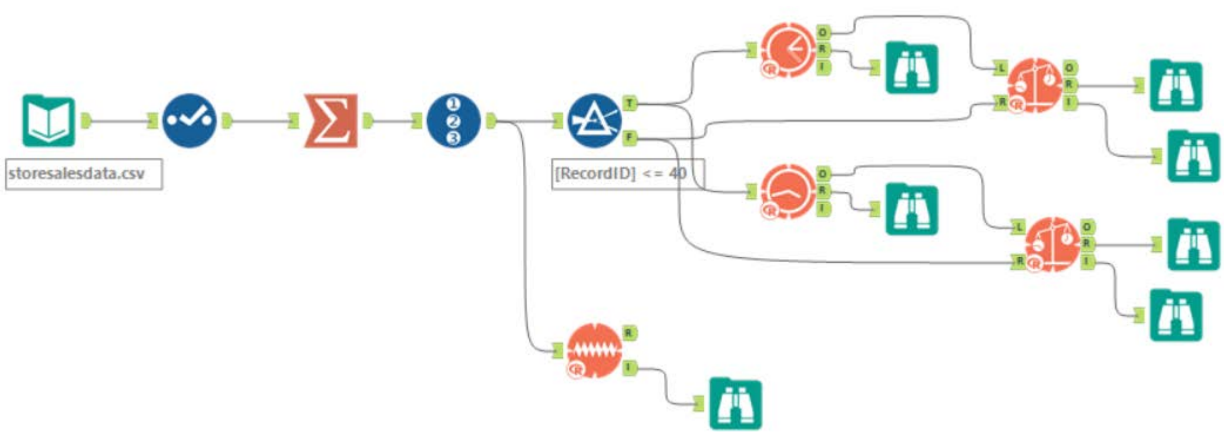


Figure 12 Alteryx (ETS & ARIMA)

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.**

Month	New Stores	Existing Stores
Jan-16	2,587,451	21,539,936
Feb-16	2,477,353	20,413,771
Mar-16	2,913,185	24,325,953
Apr-16	2,775,746	22,993,466
May-16	3,150,867	26,691,951
Jun-16	3,188,922	26,989,964
Jul-16	3,214,746	26,948,631
Aug-16	2,866,349	24,091,579
Sep-16	2,538,727	20,523,492
Oct-16	2,488,148	20,011,749
Nov-16	2,595,270	21,177,435
Dec-16	2,573,397	20,855,799

https://public.tableau.com/profile/emily6902#!/vizhome/P9_Combining_Predictive_TechniquesTask_3-Forecast/Sheet1

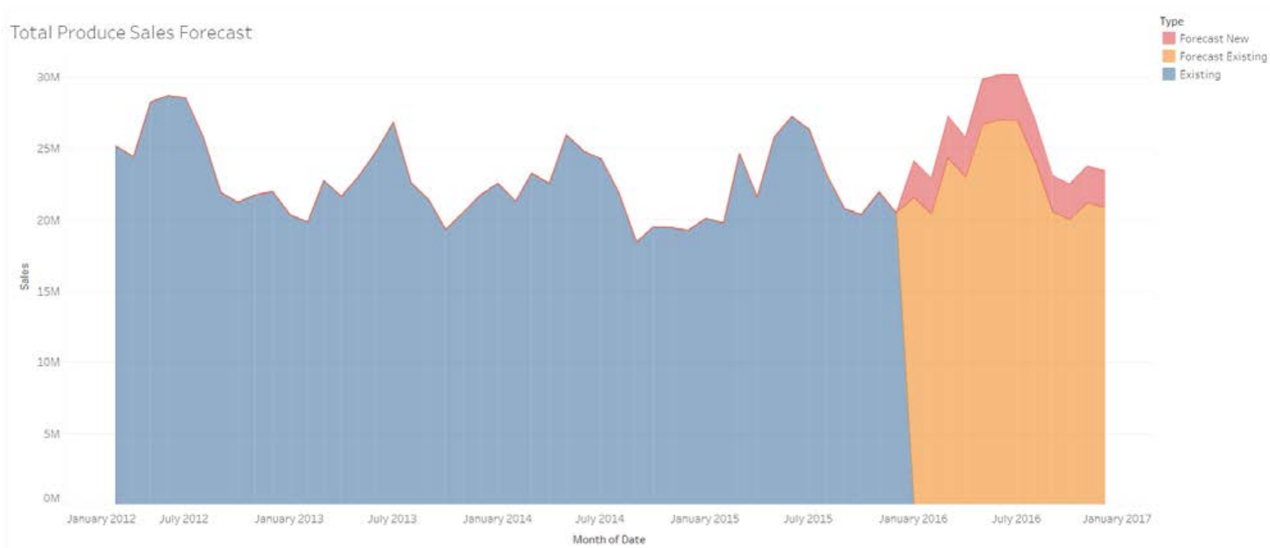


Figure 13 Total Sales Forecast

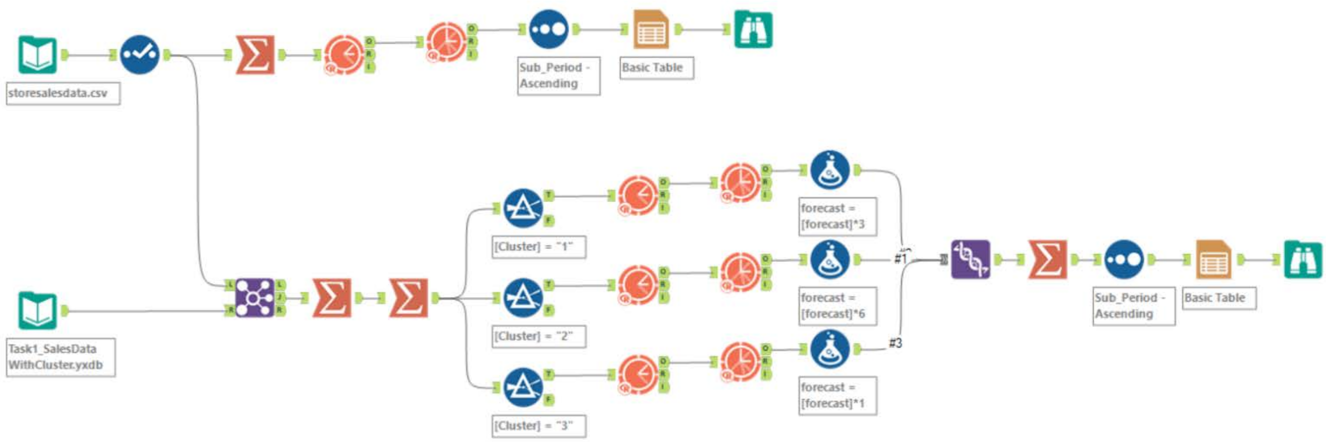


Figure 14 Alteryx (Forecast)