Project: Predictive Analytics Capstone

Business Questions:

Overview

The company has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all grocery stores use the same store format. Our tasks for this project are to find out the optimal store format for each New store, and forecasting sales for all stores.

Steps

Task 1: Determine Store Formats for Existing Stores.
Find out the optimal number of store formats through Adjusted Rand Indices and Calinski-Harabasz Indices. Built clustering model based on the existing stores information and assigned Existing stores to each cluster.

Task 2: Formats for New Stores
Developed a predictive model to predict optimal store formats for New Store. Compared results of Decision Tree, Forest Model and Boosted Model. Find out the cluster for each New store based on the clustering model developed in Task 1 and their demographic information.

Task 3: Predicting Produce Sales
Conducted ETS and ARIMA for time series analysis. Prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores.

Dataset

StoreSalesData.csv - This file contains sales by product category for all existing stores for 2012, 2013, and 2014.
StoreInformation.csv - This file contains location data for each of the stores.
StoreDemographicData.csv - This file contains demographic data for the areas surrounding each of the existing stores and locations for new stores.

Technical Support

Alteryx, Tableau

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?

   Three store formats would be the optimal choice.

To find out what is the optimal number of clusters in this question, I performed two steps for:

First, I prepared data cleaning for data aggregation, generating percentages of total sales for each product category. The new dataset would be used for cluster analysis.

Second, I performed K-centroid diagnostics for assessment, using K-means and Neural Gas as clustering methods. From results below, we can see that three clusters method would be the optimal choice since it has smaller variation comparing to others.

Cluster Assessment Results:

K-Means Cluster Assessment Report

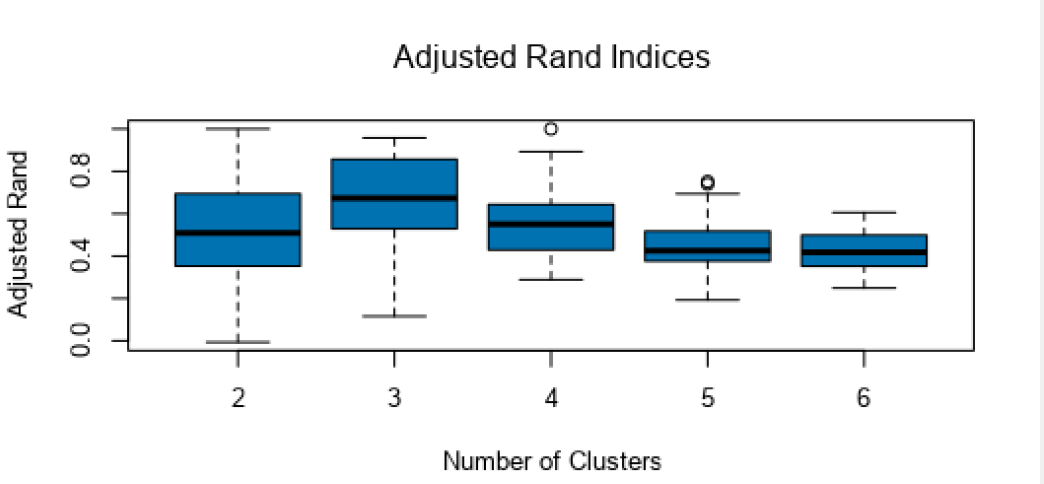
Summary Statistics

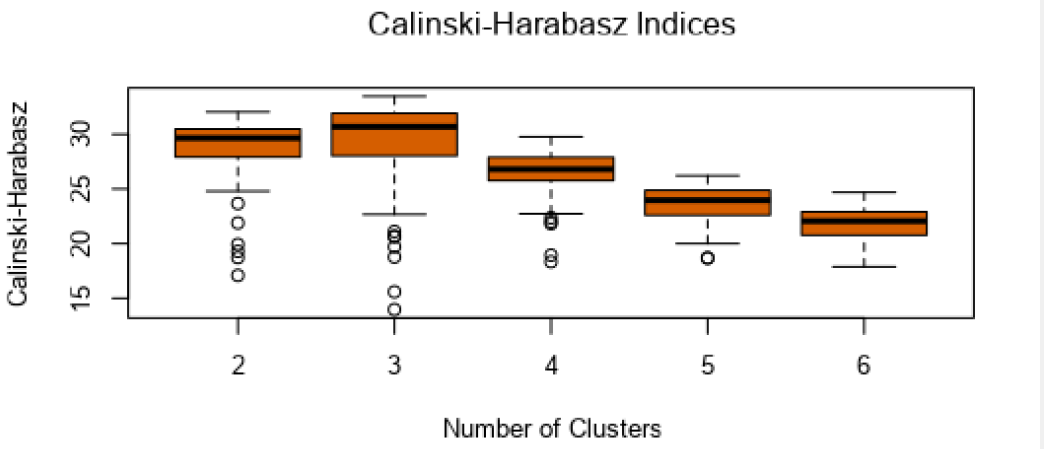
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.006314	0.115601	0.288813	0.192634	0.248942
1st Quartile	0.360019	0.529207	0.42999	0.376951	0.353886
Median	0.509255	0.673301	0.550645	0.425295	0.417472
Mean	0.51421	0.666412	0.553738	0.453703	0.429691
3rd Quartile	0.694205	0.852621	0.641424	0.513797	0.498759
Maximum	1	0.958229	1	0.751743	0.605414

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	17.0933	13.97291	18.34327	18.67605	17.88536
1st Quartile	27.96141	28.07868	25.79998	22.60516	20.79349
Median	29.65043	30.68169	26.80484	23.97152	22.05294
Mean	28.82137	29.32231	26.58932	23.66465	21.82463
3rd Quartile	30.48105	31.9193	27.88506	24.88107	22.90327
Maximum	32.04783	33.47162	29.76894	26.22175	24.68578





Neural Gas Cluster Assessment Report

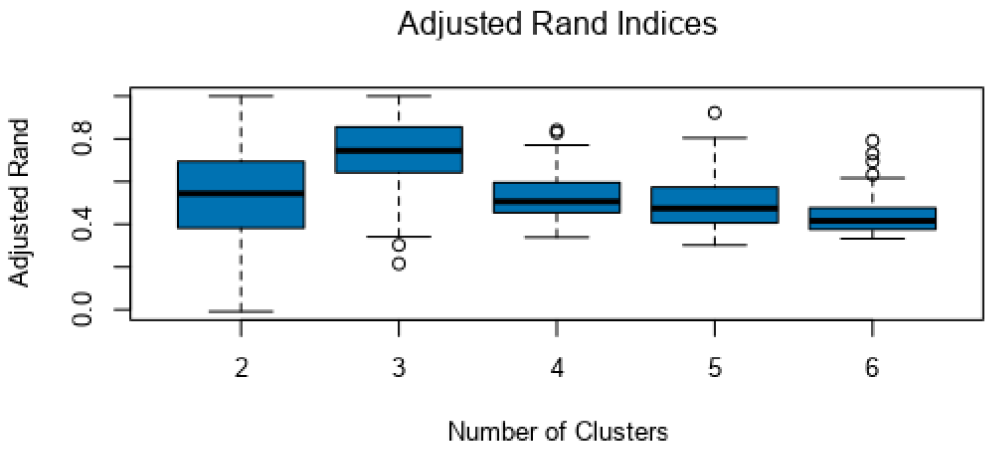
Summary Statistics

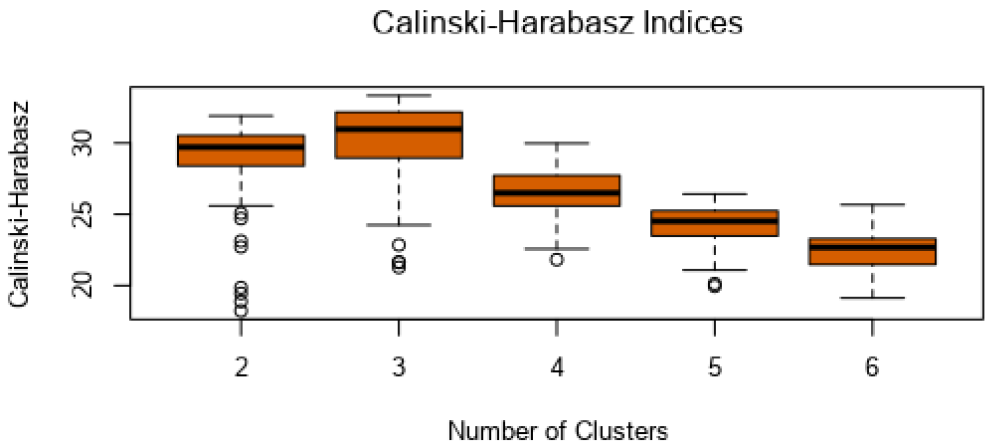
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.008749	0.214927	0.33858	0.301738	0.332499
1st Quartile	0.38973	0.642476	0.454727	0.409925	0.376388
Median	0.544111	0.744402	0.506347	0.474403	0.415745
Mean	0.51218	0.711116	0.534271	0.502127	0.444699
3rd Quartile	0.674809	0.854407	0.594615	0.56737	0.476265
Maximum	1	1	0.840953	0.922591	0.791475

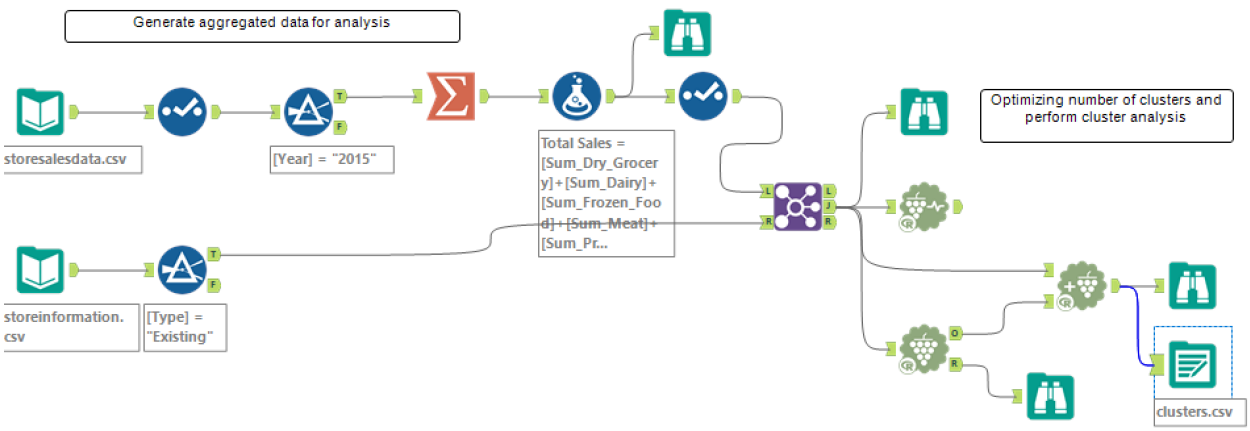
Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	18.22826	21.28358	21.82172	19.95657	19.12966
1st Quartile	28.41053	29.02334	25.58974	23.48208	21.4881
Median	29.70938	30.96412	26.49571	24.51724	22.68135
Mean	29.01329	30.18398	26.60983	24.38448	22.48538
3rd Quartile	30.52814	32.13774	27.73179	25.24584	23.29039
Maximum	31.90318	33.32816	29.97984	26.4221	25.66814





Alteryx Workflow:



2. <u>How many stores fall into each store format?</u>

The below report showed that there were 25 stores in cluster 1, 35 stores in cluster 2 and 25 stores in cluster 3.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.100007	4.82387	2.191553
2	35	2.475008	4.412365	1.947284
3	25	2.289014	3.586002	1.725751

Convergence after 8 iterations.
Sum of within cluster distances: 196.35082.

	percent_dry_grocery	percent_diary	percent_frozen_food	percent_meat	percent_produce	percent_floral	percent_deli
1	0.528251	-0.215869	-0.261593	0.614147	-0.655031	-0.663893	0.824836
2	-0.594802	0.655881	0.435121	-0.384636	0.812883	0.717413	-0.461671
3	0.304472	-0.702364	-0.347576	-0.075657	-0.483005	-0.340485	-0.178496

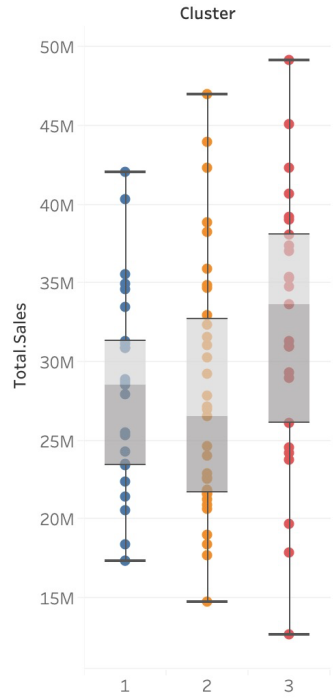
	percent_bakery	percent_general_mechandise
1	0.42823	-0.674768
2	0.312867	-0.329046
3	-0.866245	1.135433

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

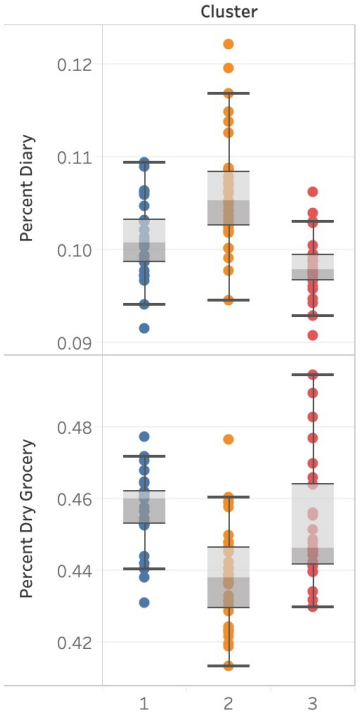
   From the report shown above, we can see that cluster 1 group sold more Deli, Meat and Dry Grocery than other clusters; the cluster 2 group sold more Produce, Floral and Diary than other clusters; the cluster 3 group sold more General Merchandise than other groups.

4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show clusters, and size to show total sales.

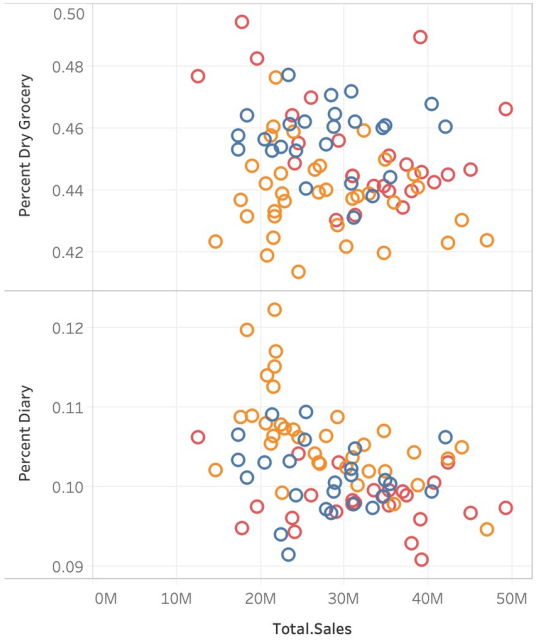
Total Sales by Clusters



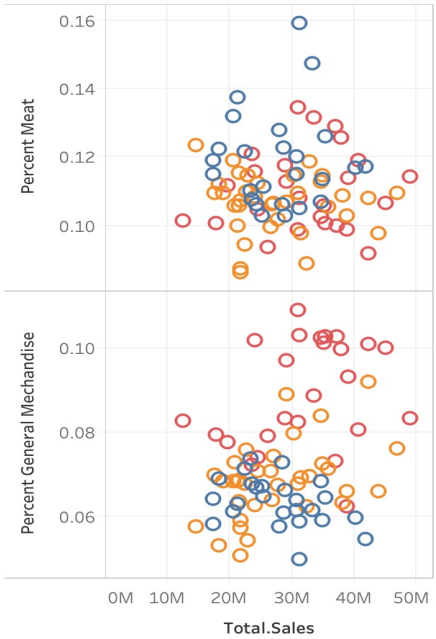
Dry Grocery and Diary in box

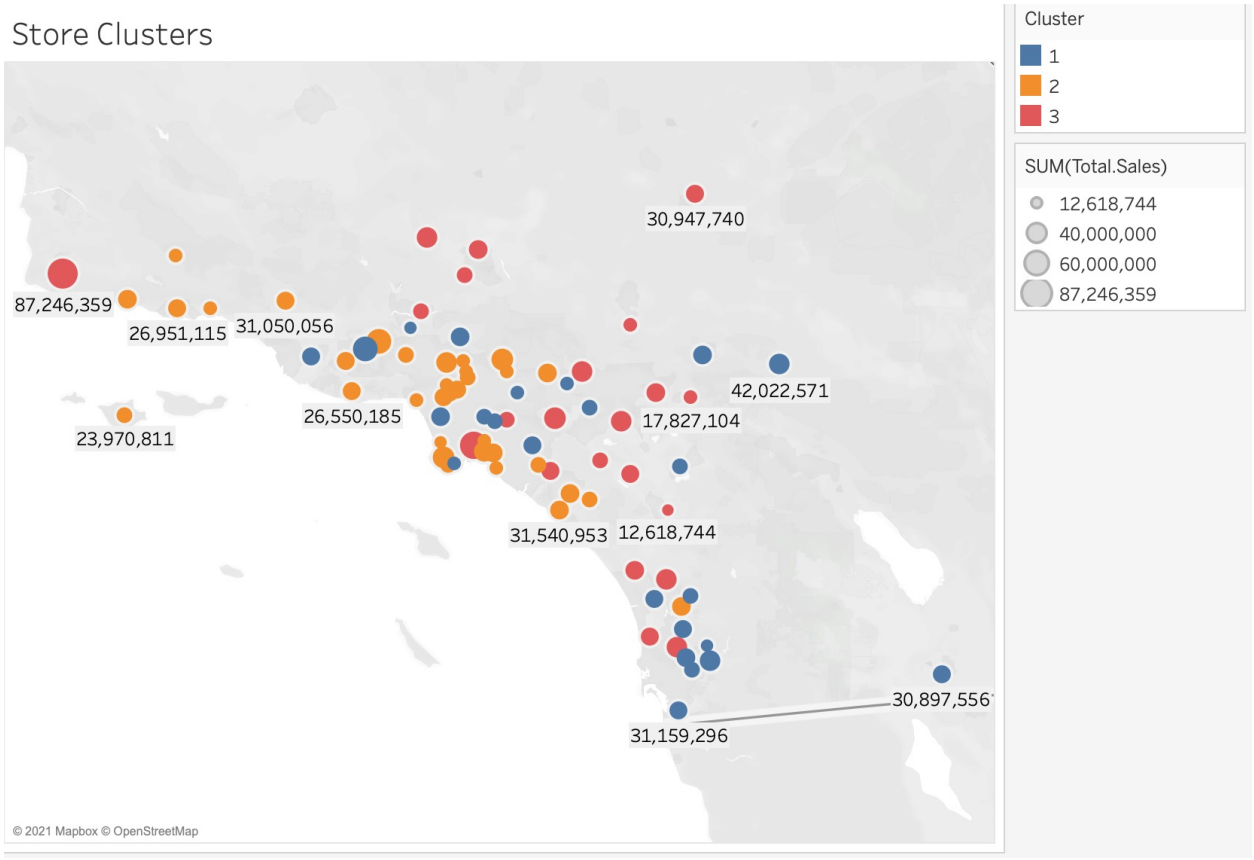


Dry Grocery and Diary Clusters VS Sales

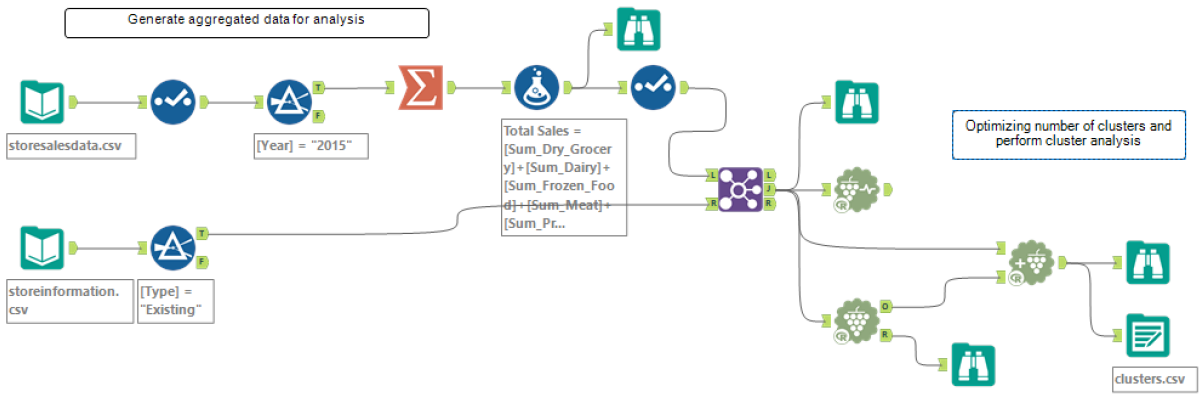


Meat and General Mechandise VS Sales





Alteryx Workflow:



Task 2: Formats for New Stores

1. <u>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.)</u>

   Boosted Model would be the best choice for this problem.

Because we wanted to predict which format a store falls into based on demographic information, I built the Decision Tree model, Forest model and Boosted model to find out the most appropriate model. From the model comparison report, we can see that the Boosted model has the highest accuracy score.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Clusters	0.7059	0.7500	0.5000	1.0000	0.7500
Decision_Tree_cluster	0.6471	0.6667	0.5000	1.0000	0.5000
Boosted_Clusters	0.7647	0.8333	0.5000	1.0000	1.0000

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 Boosted_Clusters

	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_cluster

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

Confusion matrix of Forest_Clusters

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

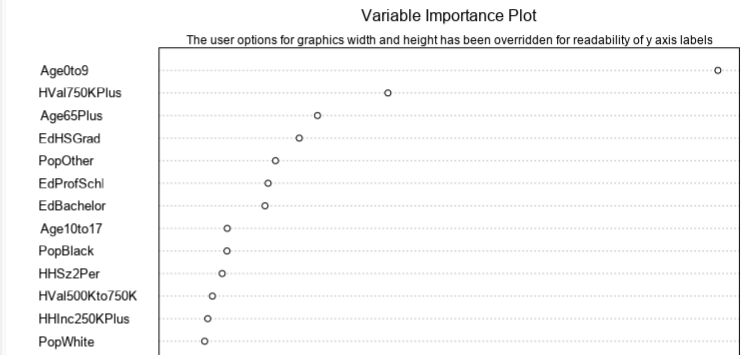
Boosted Model Result:

Report for Boosted Model Boosted_Clusters

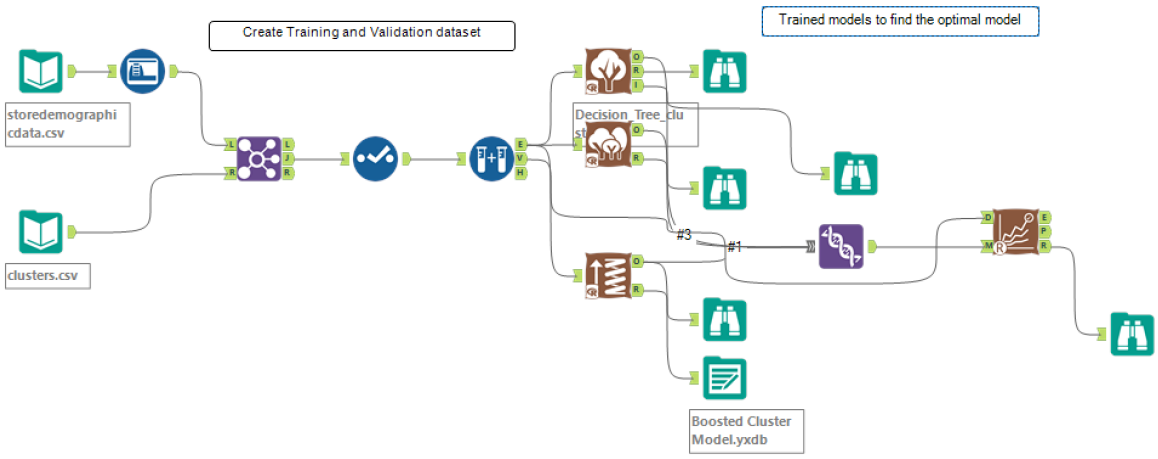
Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1829

Plots:



Alteryx Workflow:



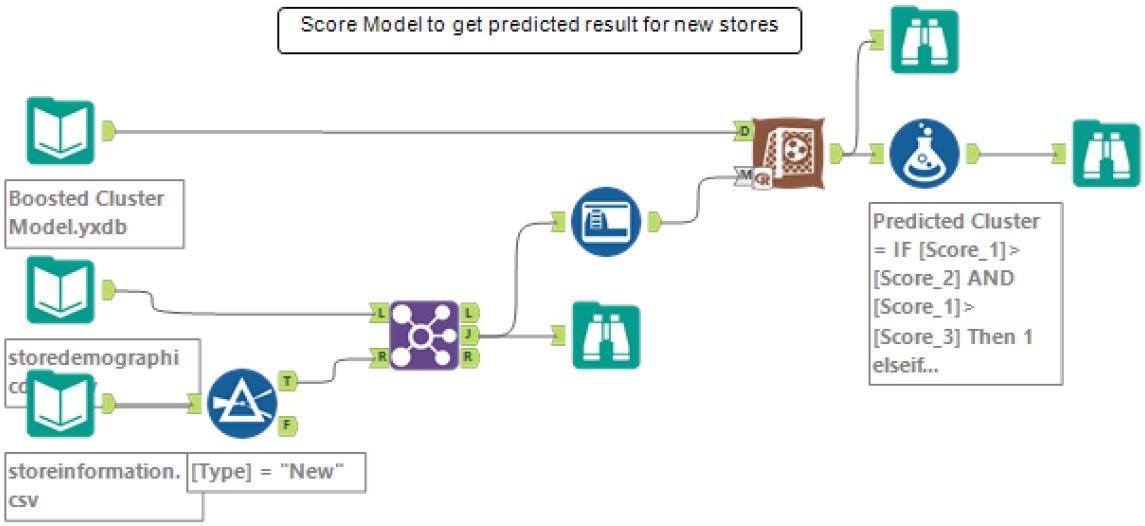
2. <u>What format do each of the 10 new stores fall into? Please fill in the table below.</u>

I implemented the Boosted model we generated in the previous step to get the prediction clusters for new stores. Below is the result:

Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2

S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

Alteryx Workflow:

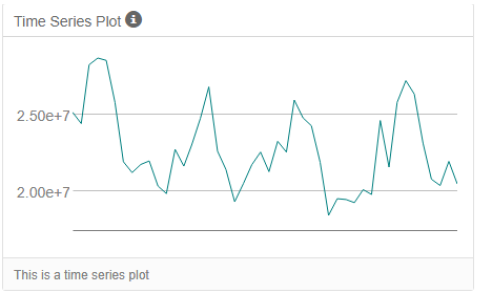


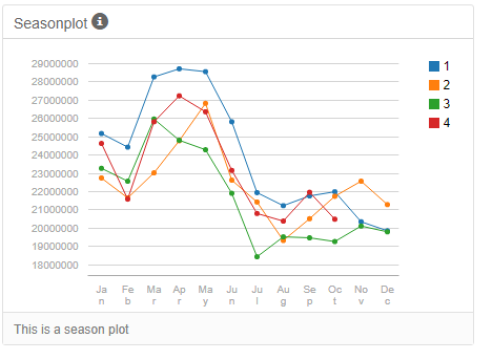
Task 3: Predicting Produce Sales

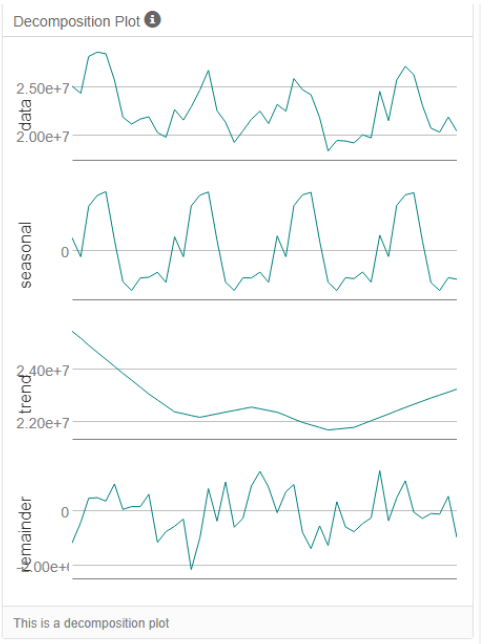
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

ETS

As the Time Series plot generated below, I used ETS(M,N,M) without dampening for forecasting:
M: The error term showed growing and shrinking over time, so in this case it should select Multiplicative.
N: There is no significant trend in the Decomposition plot, so in this case it should select No trend.
M: The seasonal component of the plot showed growing and shrinking over time, so in this case it should select Multiplicative.

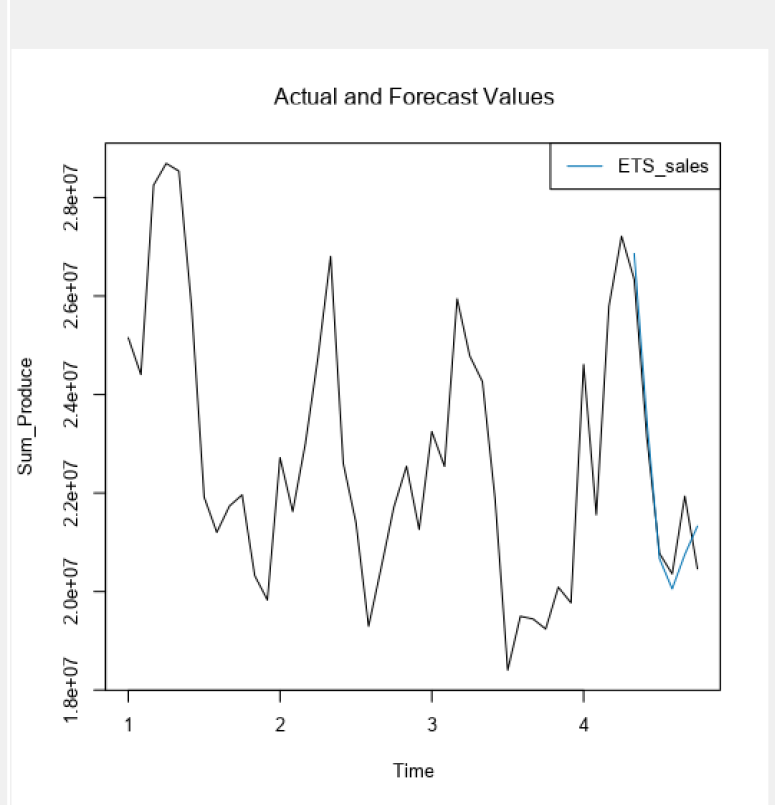






Accuracy Measures:

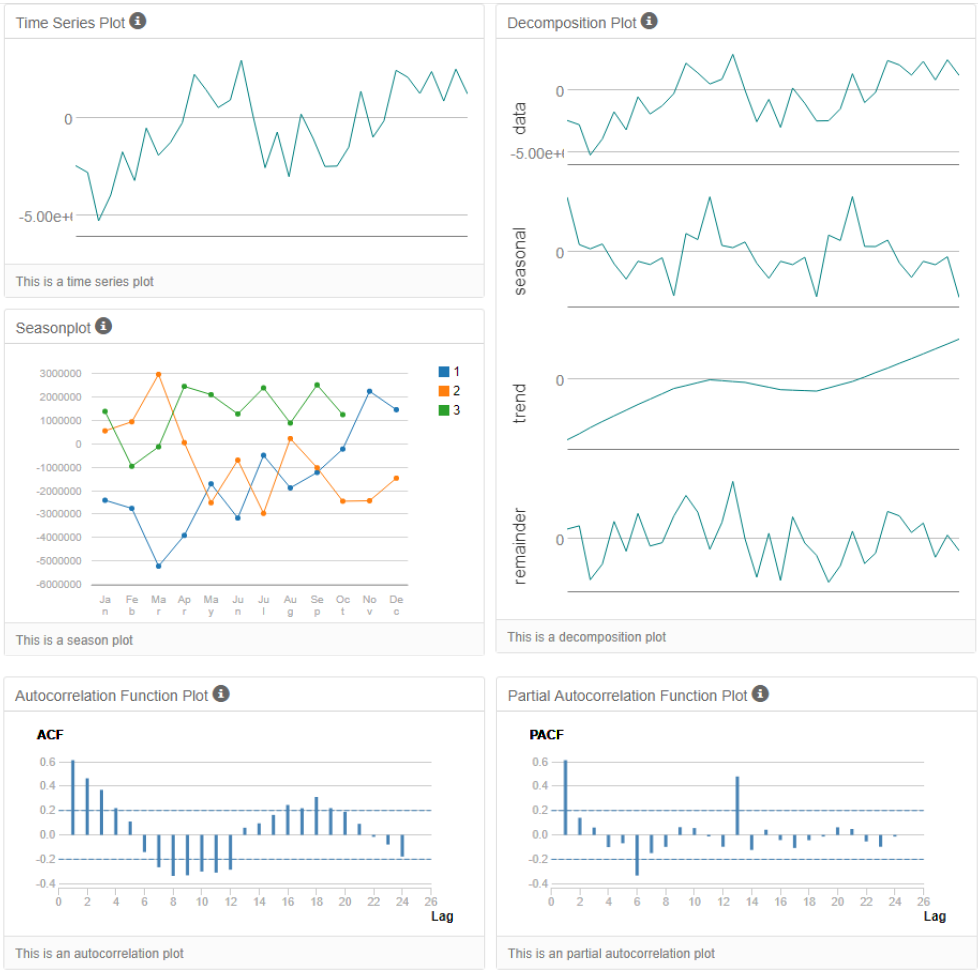
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_sales	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257



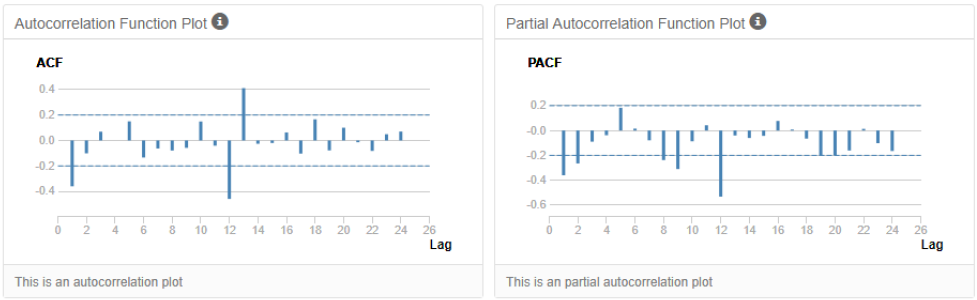
ARIMA

Due to the seasonality in our dataset, I performed ACF and PACF analysis for seasonal difference and first difference after seasonal difference. From the plot, we can see that the data is stationary after the first seasonal difference. Here, for optimal results, I implemented the Alteryx automatic ARIMA model to find optimal parameters.
Below are the results for the ARIMA (1,0,0)(1,1,0)[12]

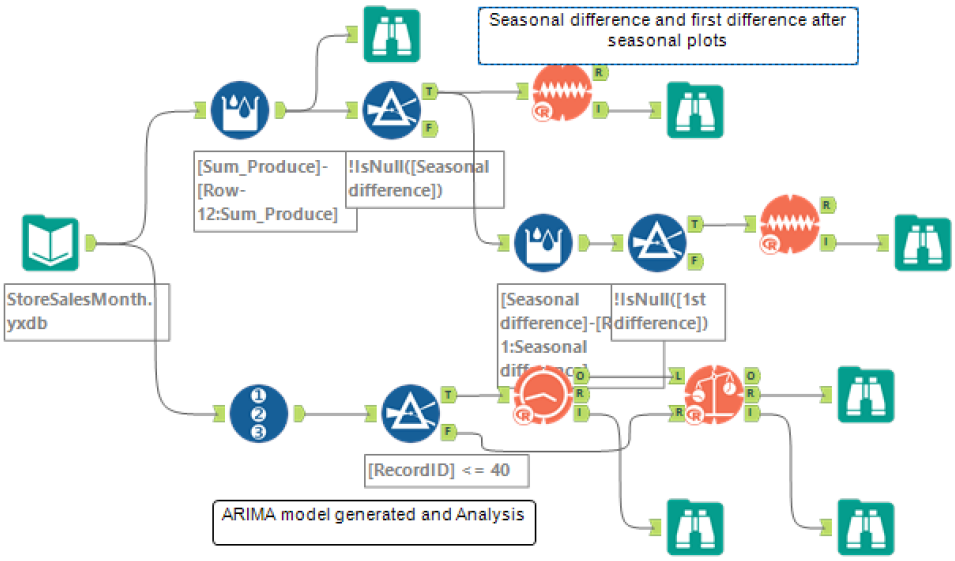
Seasonal Difference:



First difference after seasonal differencing:



Alteryx Workflow:



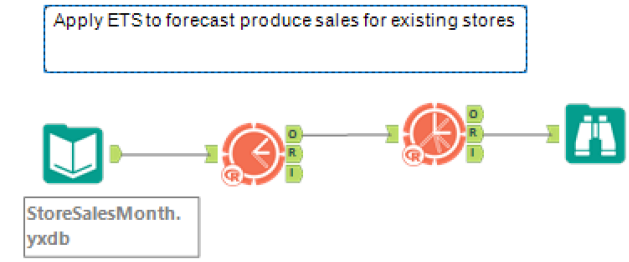
Accuracy Measures:

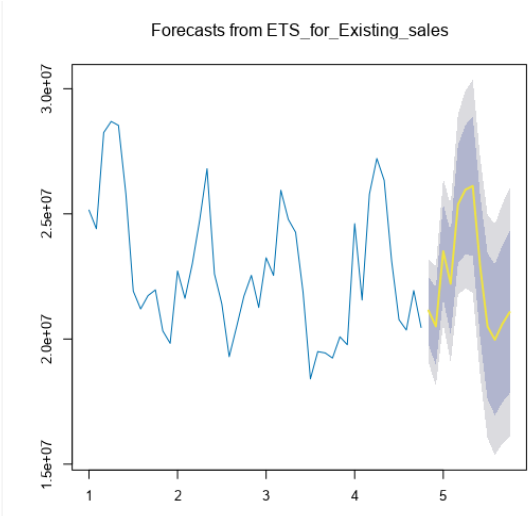
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_sales	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

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.

For this problem, I would use the ETS model to conduct time series analysis and forecast future sales based on the results.

For existing stores:

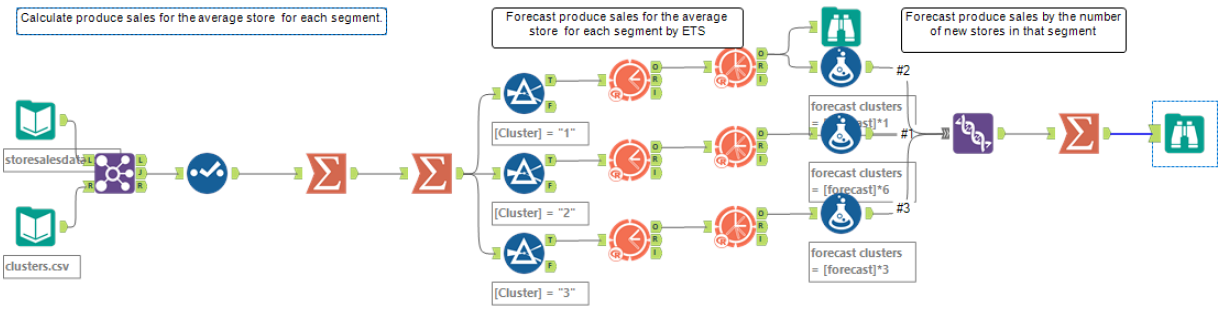




Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
4	11	21136641.781775	23208185.028684	22491151.105472	19782132.458079	19065098.534867
4	12	20507039.12384	22880476.575432	22058946.457752	18955131.789929	18133601.672248
5	1	23506565.982355	26405361.061884	25401986.205209	21611145.7595	20607770.902825
5	2	22208405.755153	25340024.343149	24256061.087491	20160750.422815	19076787.167158
5	3	25380147.771963	28988615.88472	27739598.24942	23020697.294506	21771679.659206
5	4	25966799.465113	29918337.661734	28550571.413033	23383027.517192	22015261.268491
5	5	26113792.565116	30368443.282705	28895759.137452	23331825.992781	21859141.847528
5	6	22899285.769116	27261865.623116	25751823.410525	20046748.127707	18536705.915117
5	7	20499583.908226	24962248.504876	23417563.445324	17581604.371129	16036919.311577
5	8	19971242.820704	24578175.8382	22983554.407806	16958931.233603	15364309.803208
5	9	20602665.916965	25388988.317688	23732273.917026	17473057.916904	15816343.516242
5	10	21073222.081854	26036030.707662	24318228.221798	17828215.941909	16110413.456046

For New stores, below Alteryx flow exhibited rationale for generating forecast for New stores:

- Forecast produce sales (not total sales) for the average store (rather than the aggregate) for each segment.
- Multiply the average store produce sales forecast by the number of new stores in that segment.
- For example, if the forecasted average store produce sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
- Sum the new stores produce sales forecasts for each of the segments to get the forecast for all new stores.



Combined the forecast for Existing stores and New stores, we have the table below:

Month	New Stores	Existing Stores
Jan 16	2,634,750	21,136,642
Feb 16	2,490,131	20,507,039
Mar 16	2,437,935	23,506,566
Apr 16	2,455,687	22,208,406
May 16	2,580,550	25,380,148
Jun 16	2,460,094	25,966,799
Jul 16	2,461,334	26,113,793
Aug 16	2,482,276	22,899,286
Sep 16	2,640,179	20,499,584
Oct 16	2,527,918	19,971,243
Nov 16	2,472,497	20,602,666
Dec 16	2,531,870	21,073,222

Visualization:

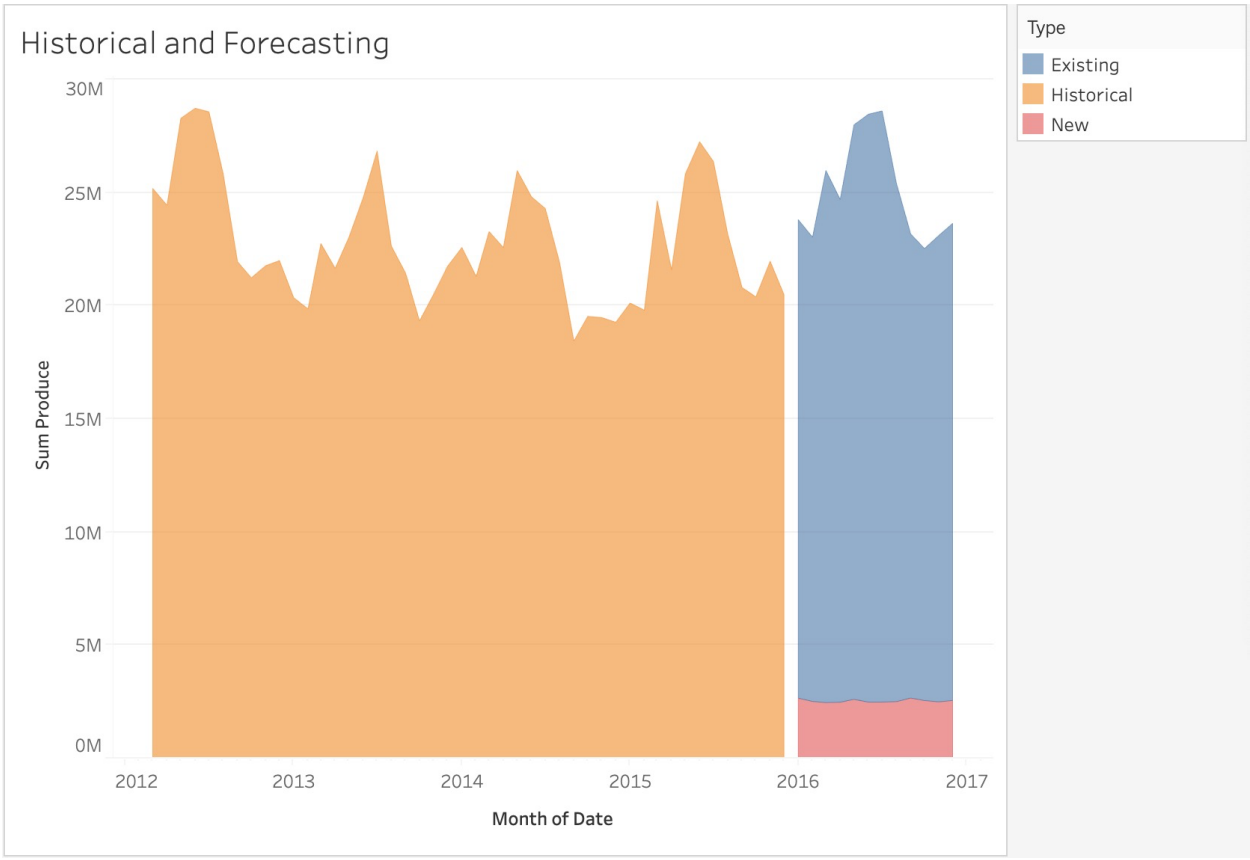


Tableau Public:
https://public.tableau.com/views/UdacityPredictiveAnalyticsCapstoneforecastingvisualization/Sheet2?:language=en&:display_count=y&publish=yes&:origin=viz_share_link