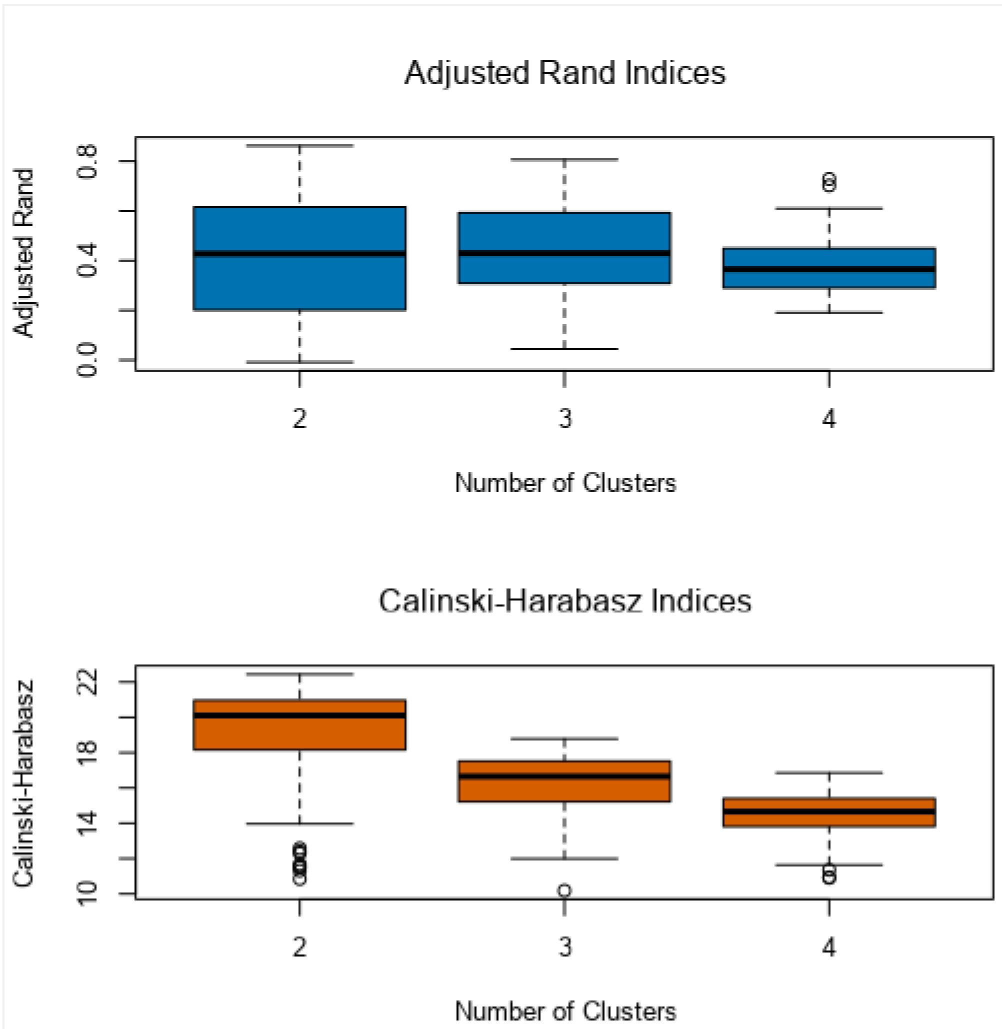
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

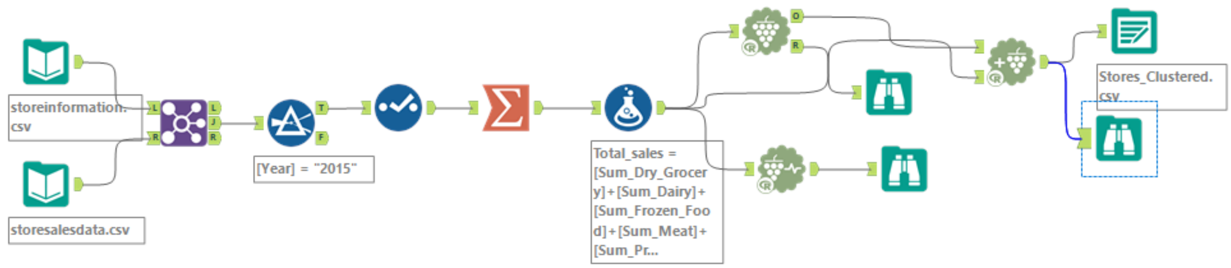
1. What is the optimal number of store formats? How did you arrive at that number?

The optimal number of store formats is three. The Adjusted Rand Indices shows that three clusters has a high median with compact values and less spread.
The Calinski-Harabasz Indices shows that three clusters has a lower median than two clusters but has less outliers.

K-Means Clustering Model



The following workflow is used to cluster the existing stores.



2. How many stores fall into each store format?

Cluster 1 has 25 stores, cluster 2 has 35 stores and cluster 3; has 25 stores.

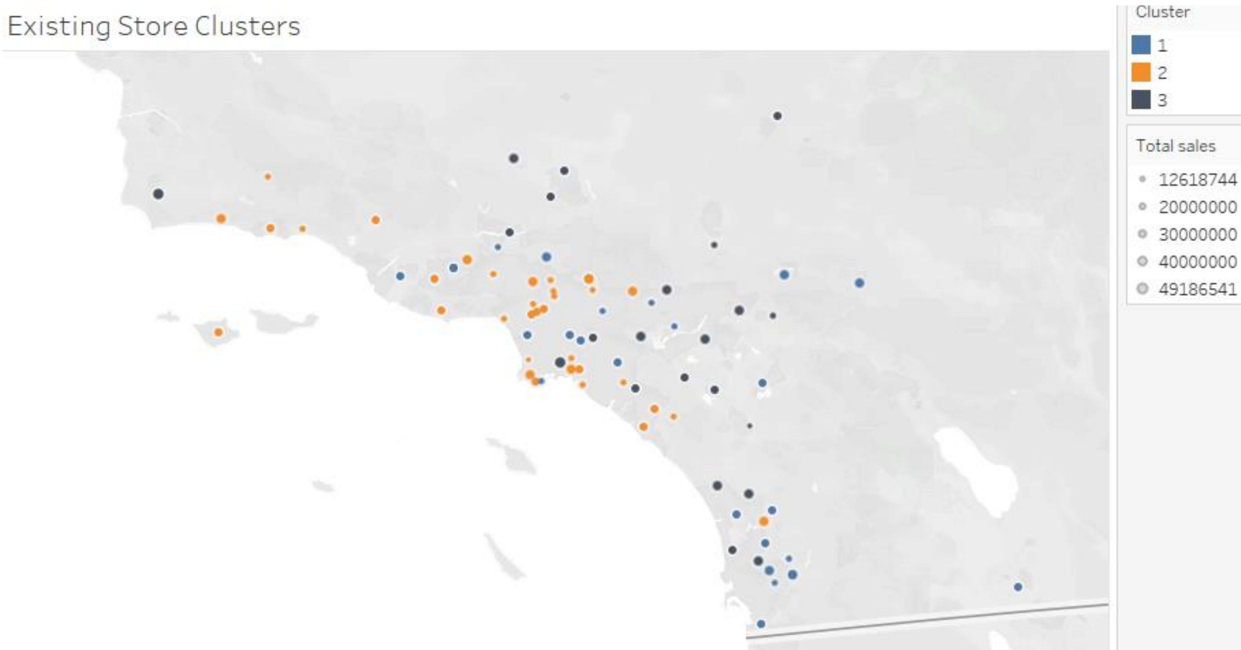
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

The Average Distance of cluster 2 is the least compact which might show more variability. Cluster 1 shows the most separation from the other clusters which is good for the model.

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	25	2.100598	4.823985	2.193986
2	35	2.475232	4.410756	1.9441
3	25	2.287649	3.582763	1.723182

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.

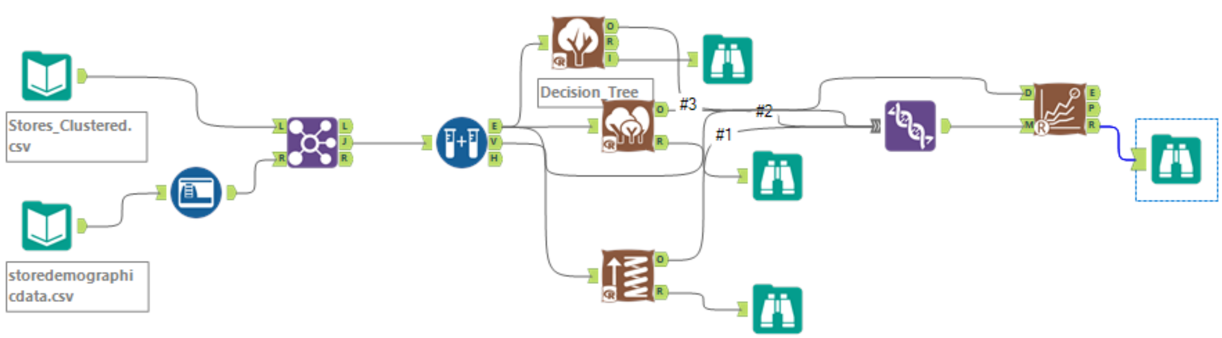


Existing Store Clusters | Tableau Public

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.)

Task 2 Workflow to Predict Store Format



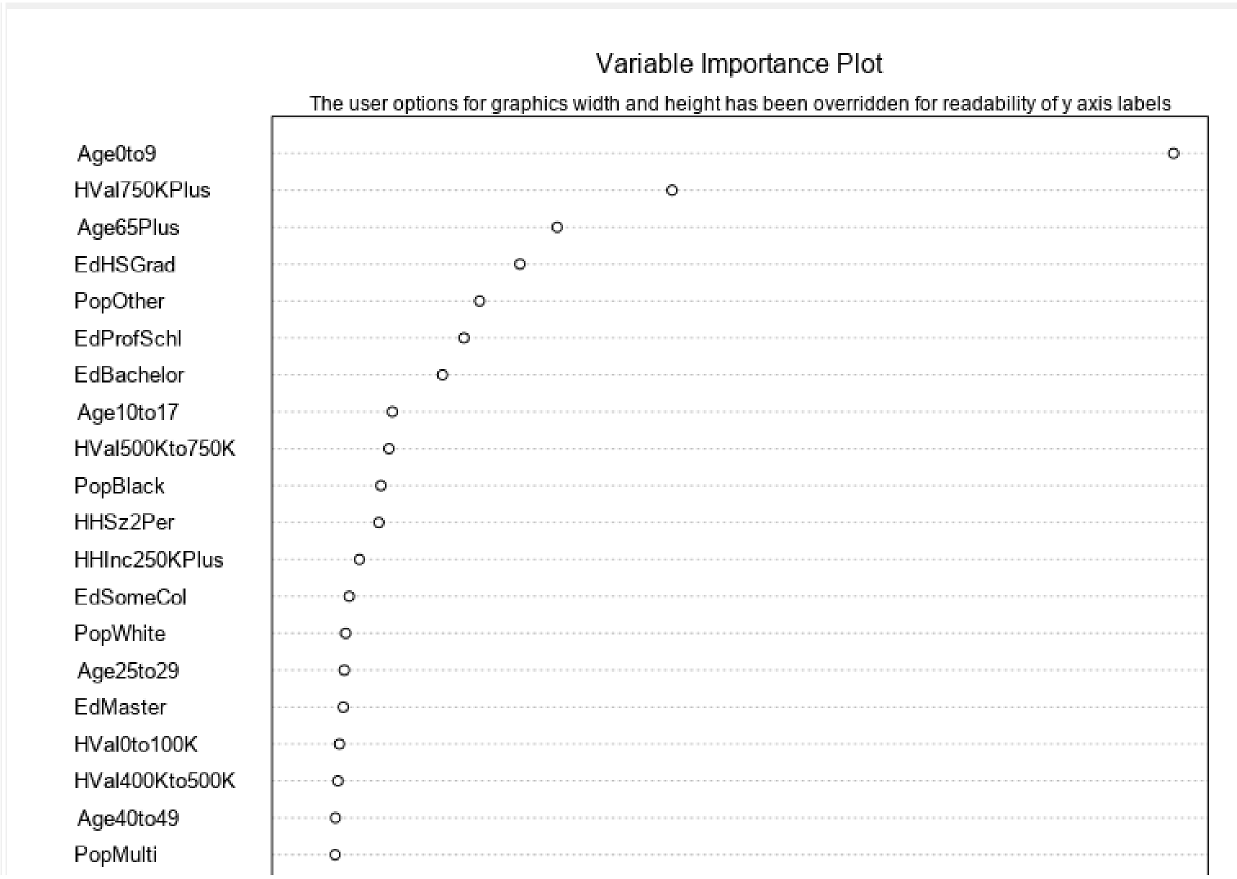
Based on the Model Comparison Report the Boosted Model was selected to predict the store formats.

Model Comparison Report

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Decision_Tree	0.7059	0.7083	0.6250	1.0000	0.5000
Boosted	0.7059	0.7500	0.5000	1.0000	0.7500

From the Variable Importance Plot below it can be seen that the three most important variables that help explain the relationship between demographic indicators and store formats are;

Age0to9, HVal750KPlus and Age65Plus.



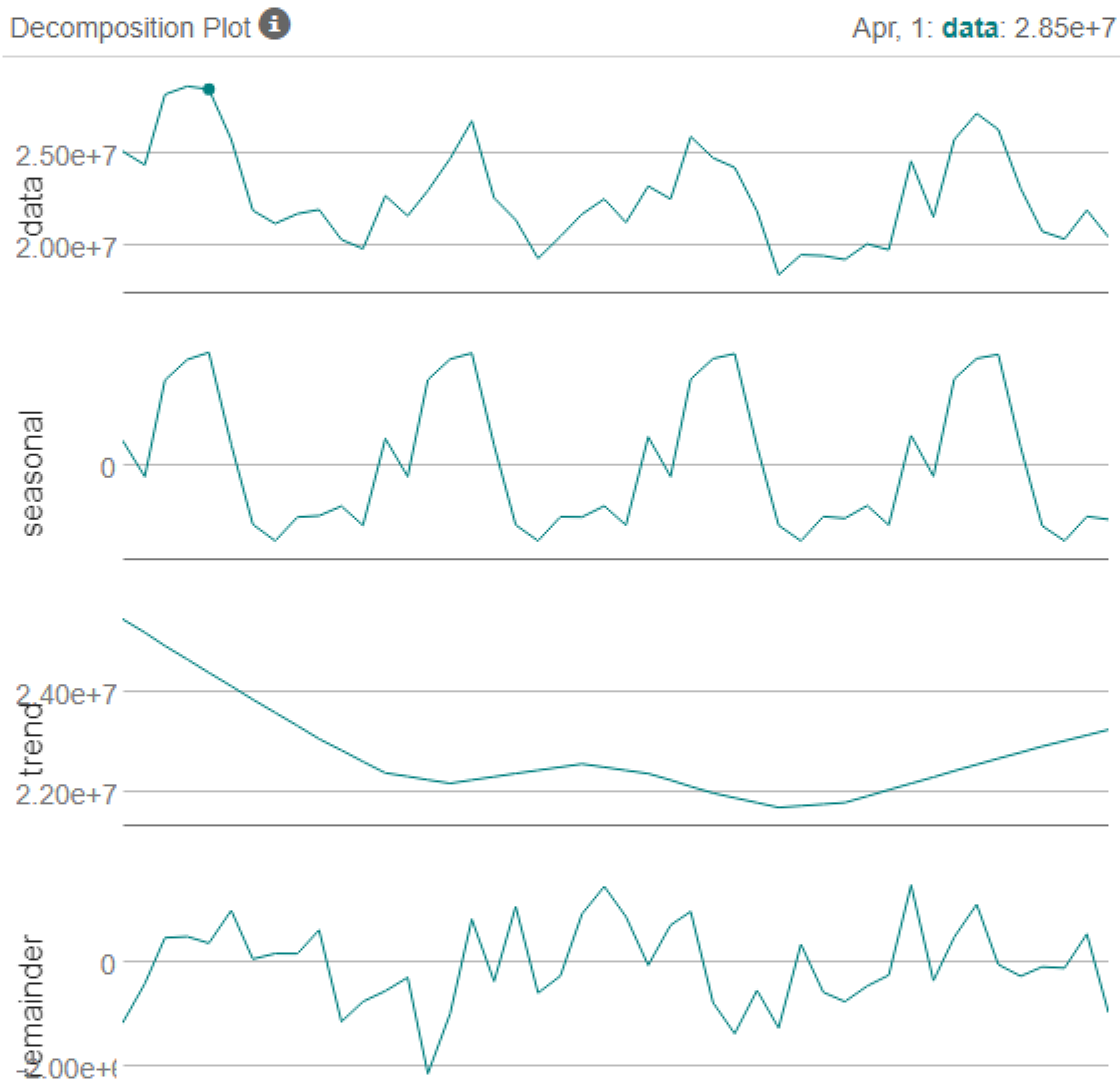
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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

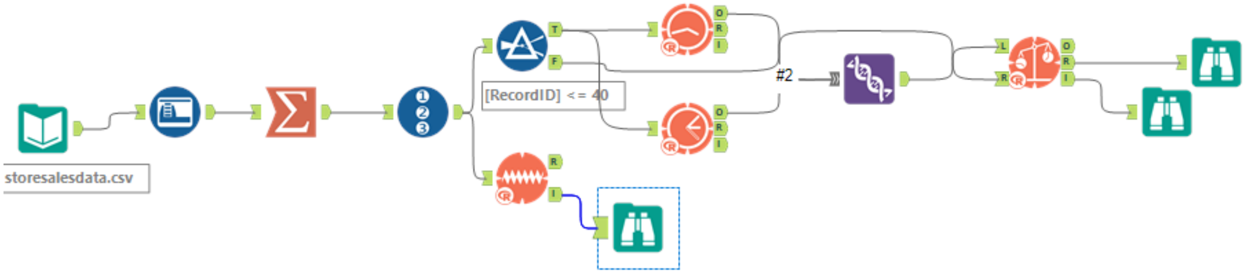
Task 3: Predicting Produce Sales

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?

The Decomposition Plot shows an inconsistent Error so is applied Multiplicatively. The Trend will be applied as None. The Seasonality show an increasing trend so is applied Multiplicatively.



ETS and ARIMA Workflow



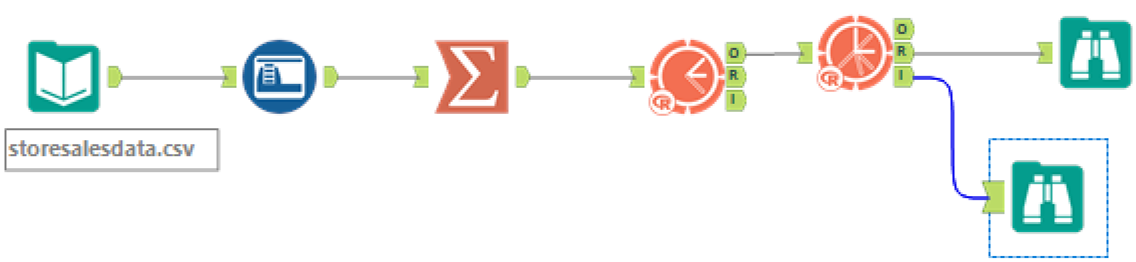
The ETS(M,N,M) model was selected to forecast. The ETS model shows an RMSE of 663707.2 and a MASE value of 0.32. The ARIMA model has a RMSE of 1050239.2 and a MASE value of 0.54.

Accuracy Measures:

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

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.

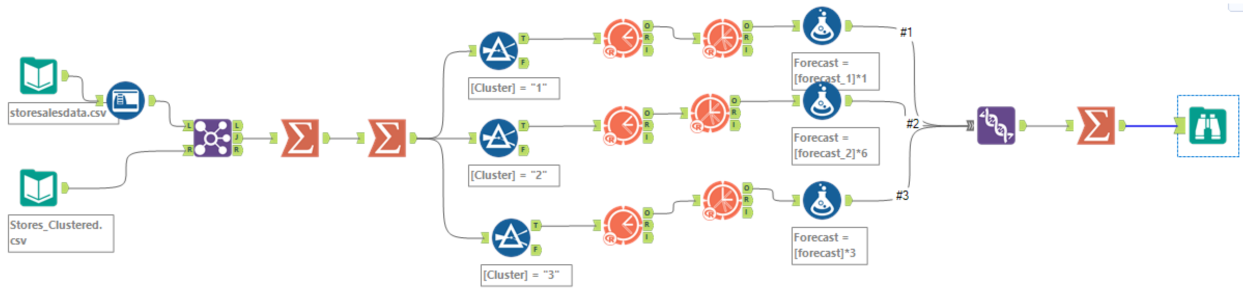
The workflow below is used to forecast sales for the existing stores for the year 2016.



Forecast for the existing stores

Period	Sub_Period	forecast	forecast_high_95	forecast_high_80	forecast_low_80	forecast_low_95
2016	1	21829060.031666	24149899.115321	23346575.14138	20311544.921952	19508220.948011
2016	2	21146329.631982	23512577.365832	22693535.862148	19599123.401815	18780081.898131
2016	3	23735686.93879	26517865.796798	25554855.912929	21916517.964651	20953508.080782
2016	4	22409515.284474	25150243.401256	24201581.075733	20617449.493214	19668787.167691
2016	5	25621828.725097	28880596.484529	27752622.431914	23491035.018279	22363060.965665
2016	6	26307858.040046	29777680.067343	28576652.715009	24039063.365084	22838036.01275
2016	7	26705092.556349	30348682.320364	29087507.847195	24322677.265503	23061502.792334
2016	8	23440761.329527	26742106.733295	25599395.061562	21282127.597491	20139415.925758
2016	9	20640047.319971	23635033.372194	22598363.439189	18681731.200753	17645061.267747
2016	10	20086270.462075	23084199.797487	22046511.090727	18126029.833423	17088341.126662
2016	11	20858119.95754	24055437.105831	22948733.269445	18767506.645635	17660802.809249
2016	12	21255190.244976	24596988.126893	23440274.43075	19070106.059202	17913392.363058

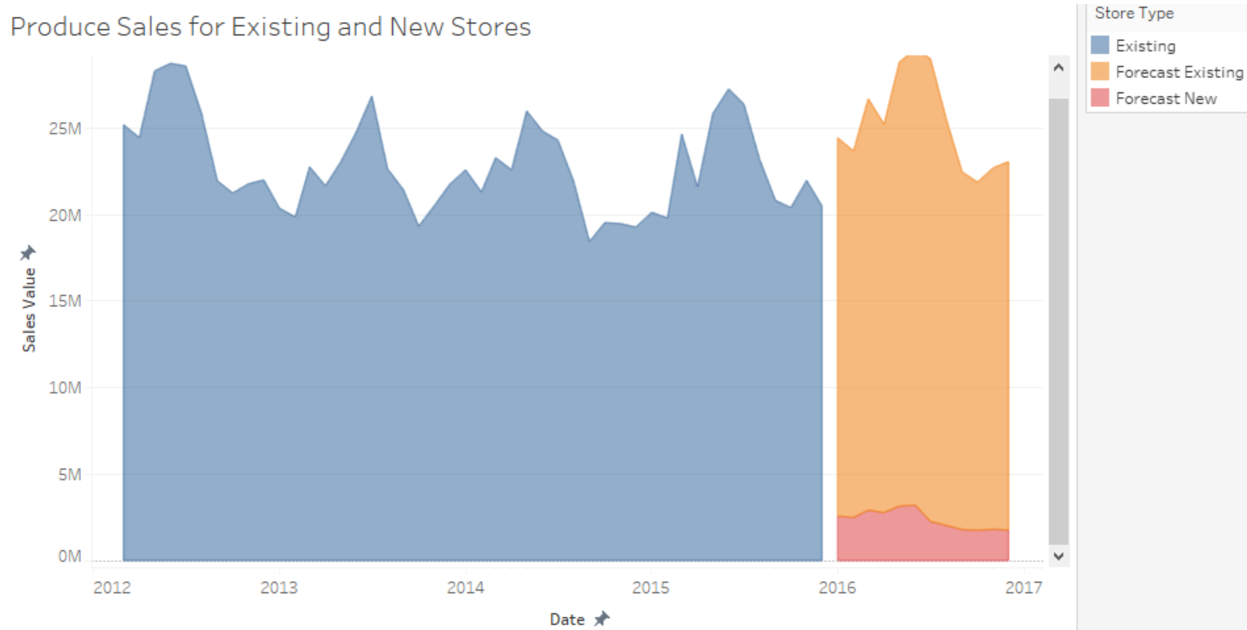
The below workflow is used to forecast sales for the new stores for the year 2016.



The table below shows sales forecasts for new and existing stores for the year 2016.

Month-Year	New Stores	Existing Stores
Jan-16	2,563,357.91	21,829,060.03
Feb-16	2,483,924.73	21,146,329.63
Mar-16	2,910,944.15	23,735,686.94
Apr-16	2,764,881.87	22,409,515.28
May-16	3,141,305.87	25,621,828.73
Jun-16	3,195,054.20	26,307,858.04
Jul-16	2,253,197.18	26,705,092.56
Aug-16	2,020,427.56	23,440,761.33
Sep-16	1,785,271.85	20,640,047.32
Oct-16	1,747,496.79	20,086,270.46
Nov-16	1,799,586.00	20,858,119.96
Dec-16	1,756,415.45	21,255,190.24

The visualization below shows historical and forecast sales for existing and new stores for the years 2012 to 2016.



[Produce Sales for Existing and New Stores | Tableau Public]()