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?

Based on the report results the optimal number of cluster is 3, because AR and CH indices show the highest median and the smallest variation in its spread.

	2	3	4
Minimum	-0.008598	0.047321	0.190877
1st Quartile	0.21411	0.311458	0.260379
Median	0.427746	0.425431	0.393611
Mean	0.426051	0.438655	0.37657
3rd Quartile	0.60704	0.577371	0.443479
Maximum	0.862177	0.806806	0.728735

Calinski-Harabasz Indices:

	2	3	4
Minimum	10.84432	10.18405	10.90095
1st Quartile	18.29771	15.23665	13.71761
Median	20.0721	16.6871	14.68046
Mean	19.04128	16.26252	14.49592
3rd Quartile	20.98638	17.42509	15.44396
Maximum	22.44228	18.75042	16.86351

Fig. 1.1. Adjusted Rand and Calinski-Harabasz indices report.

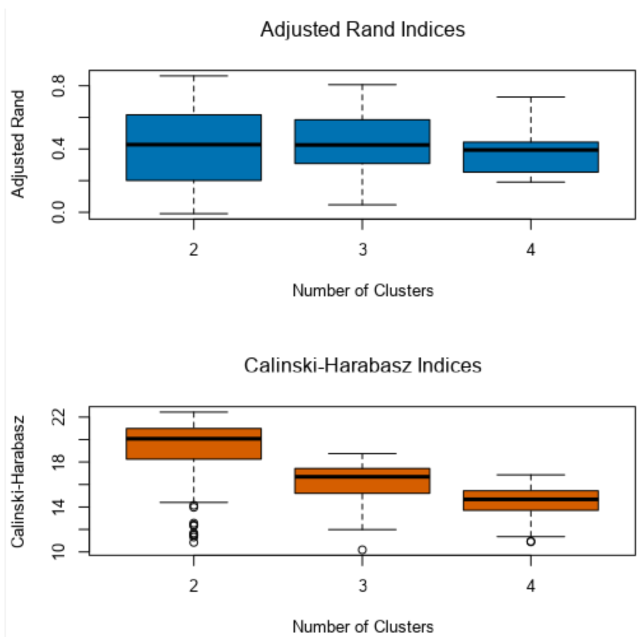

Fig. 1.2. Adjusted Rand and Calinski-Harabasz plots.

2. How many stores fall into each store format?
    a. Cluster1 – 25 stores;
    b. Cluster2 – 35 stores;
    c. Cluster3 – 25 stores.

Summary Report of the K-Means Clustering Solution K_Centroids_Cluster_Analysis

Solution Summary

Call:
stepFlexclust(scale(model.matrix(~-1 + X._Dry_Grocery + X._Dairy + X._Frozen_Food + X._Meat + X._Produce + X._Floral + X._Deli + X._Bakery + X._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

Fig. 1.3. Cluster information report.

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

Based on the clustering model, stores that fall in cluster 2 appear to have an increase in inventory overall higher than cluster 1 and 2, especially on the dry grocery segment.

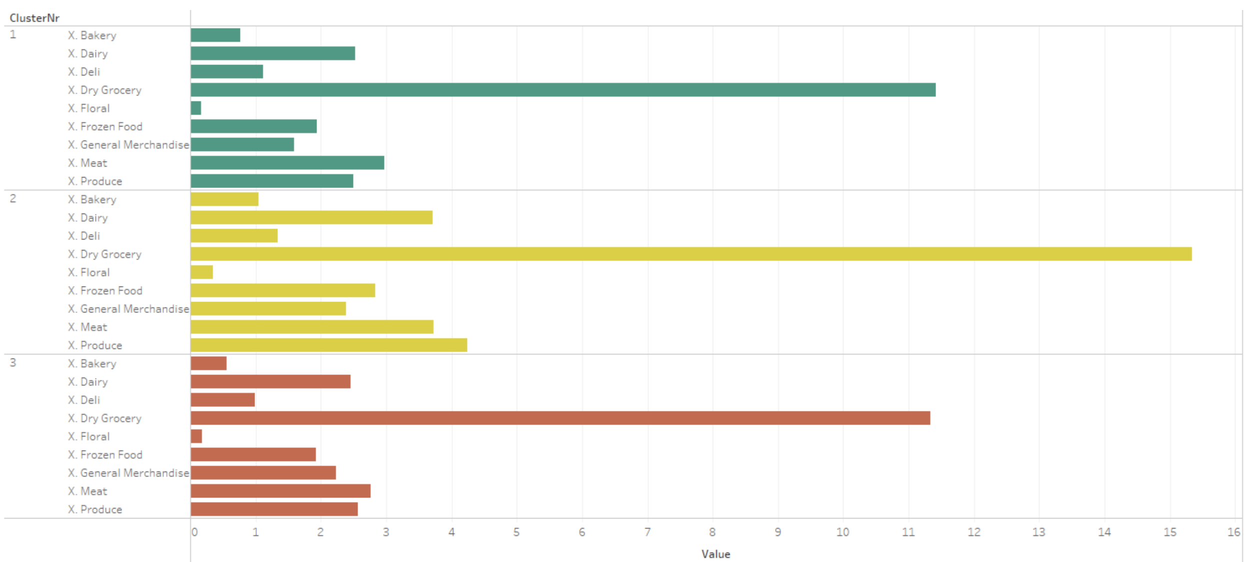

Fig. 1.4. Clusters segmentation overview.

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.

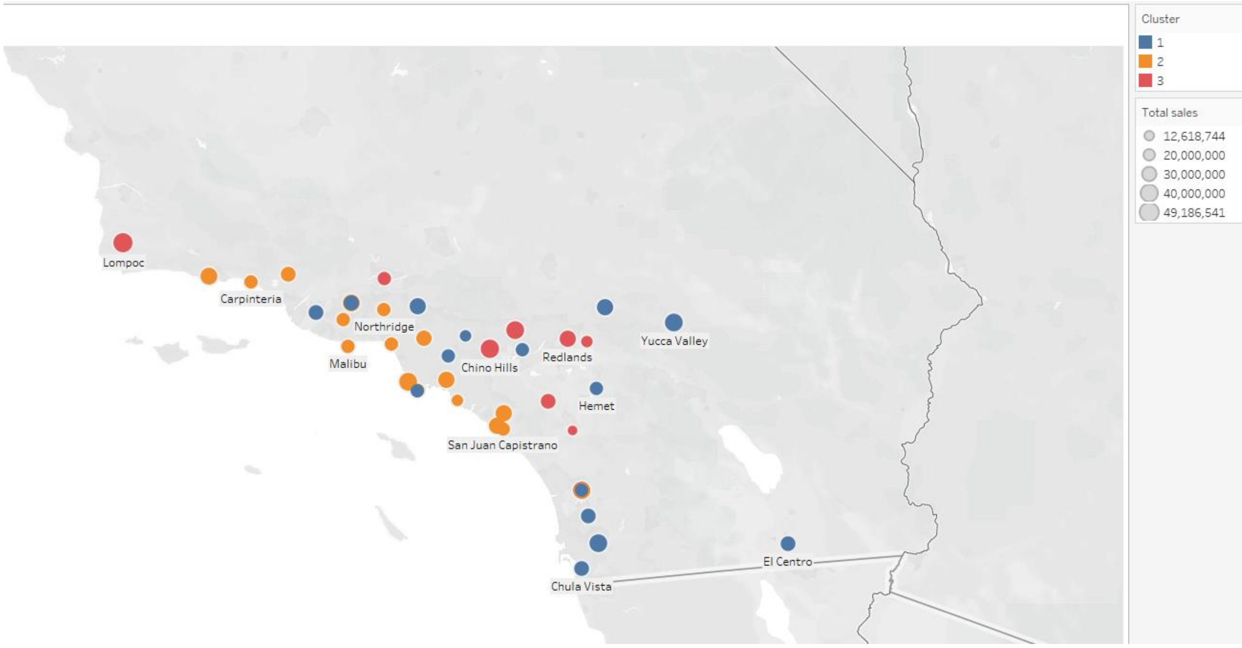

Fig. 1.5. Store location and size by sales.

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.)

I have used the Boosted_Model in order to predict the best store format for the new stores. Based on the model comparison report this is the best suited for this business problem considering that has the highest accuracy and F1 score.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.7059	0.7083	0.6250	1.0000	0.5000
Forest	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_Model	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_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	5	0	2
Predicted_2	2	5	0
Predicted_3	1	0	2

Confusion matrix of Forest

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

Fig. 2.1. Model comparison report.

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?

Based on the time series report, the decomposition plot shows there is a bit of seasonality, the trend slightly turns up at the end so it should not be applied and remainder changes in magnitude, meaning we should apply it multiplicatively.

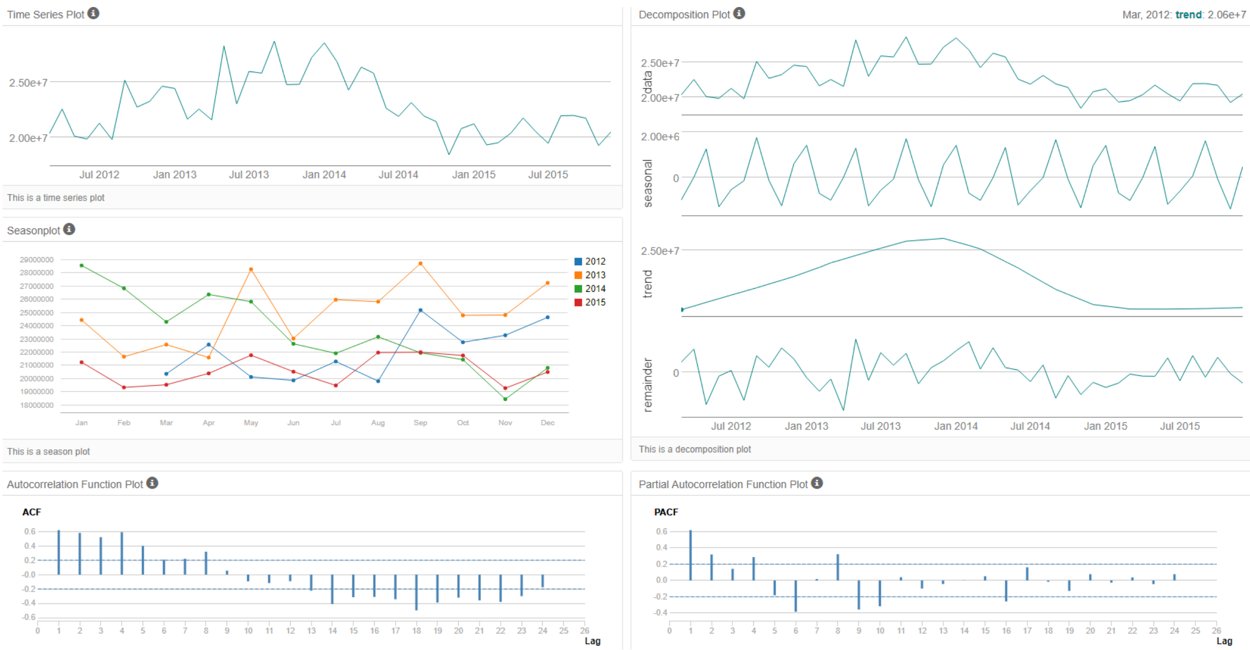

Fig. 3.1. Time series report.

Afterwards, I have setup both models: ETS(M,N,M) and ARIMA(1,0,0)(1,1,0)[12] and compared the time series models.

a. ETS(M,N,M) model results:

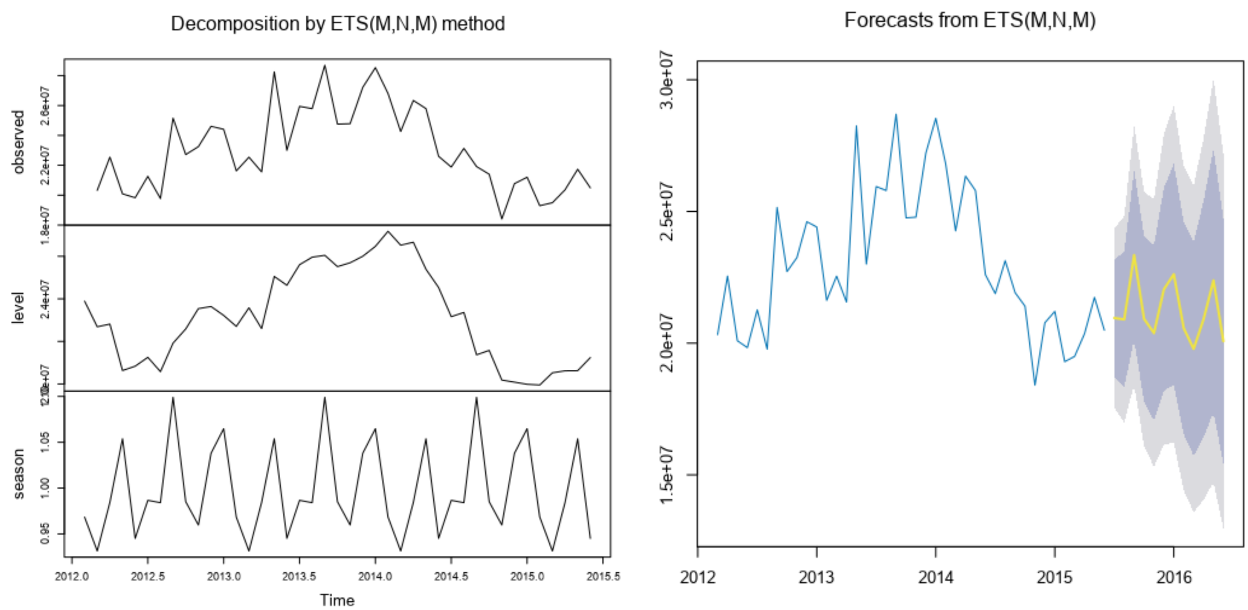

Fig. 3.2. Decomposition and forecast of the ETS(M,N,M) model.

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-115442.8071963	1575585.5998785	1154198.0139137	-0.8328272	5.0600506	0.2809384	-0.1305869

Information criteria:

AIC	AICc	BIC
1317.0842	1337.0842	1342.4174

Smoothing parameters:

Parameter	Value
alpha	0.582236
gamma	1e-04

Initial states:

State	Value
l	23891334.885217
s0	0.968322
s1	1.064817
s2	1.037938
s3	0.959677
s4	0.985047
s5	1.099031
s6	0.984043
s7	0.986617
s8	0.945085
s9	1.053767
s10	0.984319

Fig. 3.3. Summary of Time Series Exponential Smoothing Model ETS(M,N,M) method.

b. ARIMA(1,0,0)(1,1,0)[12] model results:

Call:
Arima(Sum_Produce, order = c(1, 0, 0), seasonal = list(order = c(1, 1, 0), period = 12))

Coefficients:

	ar1	sar1
Value	0.737388	-0.593345
Std Err	0.125868	0.160497

sigma^2 estimated as 6157021842218.35: log likelihood = -453.9537

Information Criteria:

AIC	AICc	BIC
913.9074	914.9074	917.904

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
9490.0265316	2000515.9828009	1283261.1342739	-0.4628167	5.4942828	0.3123532	-0.3361839

Ljung-Box test of the model residuals:
Chi-squared = 21.1164, df = 12, p-value = 0.048702

Fig. 4.4. Summary of the ARIMA(1,0,0)(1,1,0)[12] method.

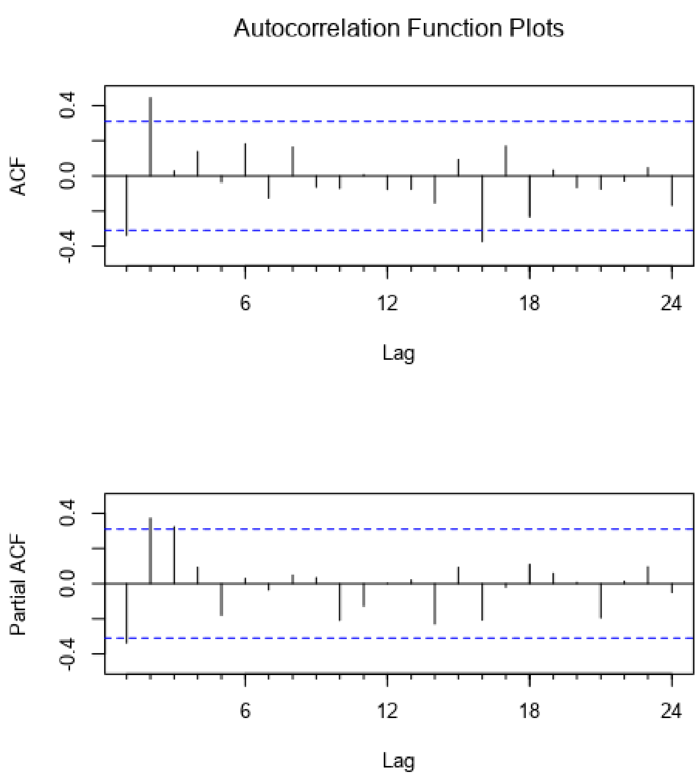

Fig. 4.5. Autocorrelation function plots.

c. Time series model comparison:

Actual and Forecast Values:

Actual	ETS	ARIMA
19444753.17	20954549.498	22559587.46935
21936906.81	20899853.78763	23433220.41289
21962976.75	23342005.09054	24994661.55192
21715706.67	20921264.24179	22697309.26284
19240384.75	20382324.73577	21677953.16035
20462899.3	22044587.46326	24216979.5775

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-630159.6	1271062	1240658	-3.2204	6.0156	0.6604
ARIMA	-2469347.3	2649864	2469347	-12.0298	12.0298	1.3145

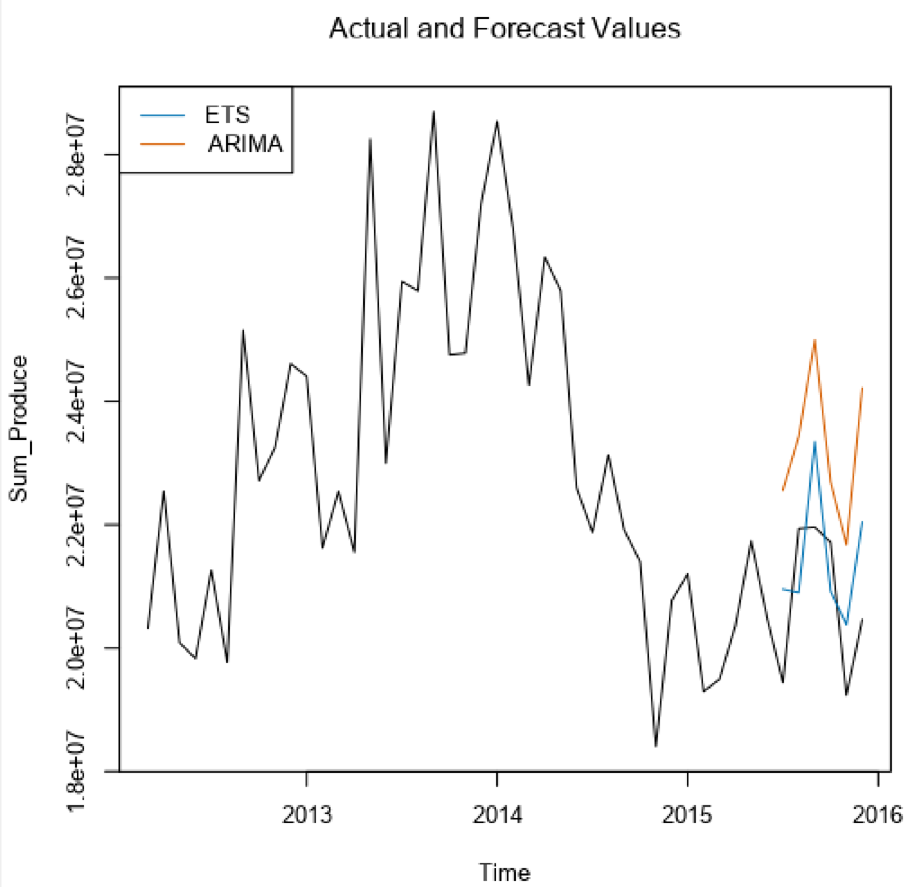

Fig. 4.6. Comparison of Time Series Models.

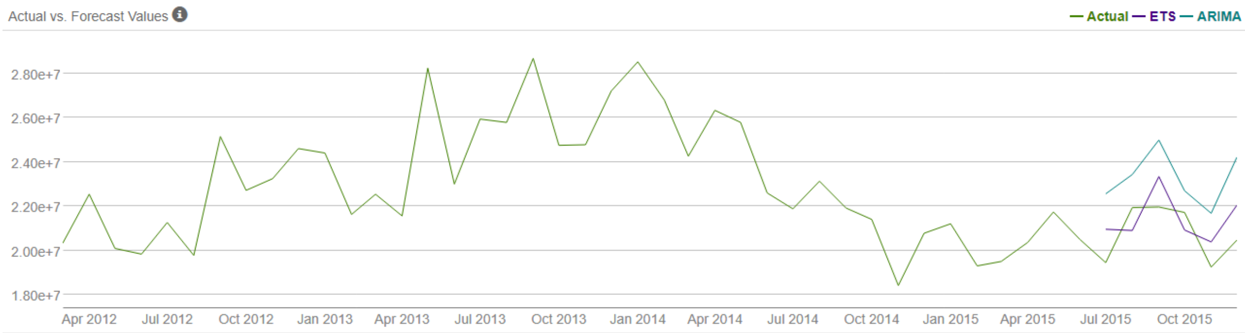

Fig. 4.7. Actual vs. Forecast values graph.

By comparing the two methods, the forecast for the ETS(M,N,M) method is closer to the actual values than the ARIMA(1,0,0)(1,1,0)[12] method, therefore, the ETS(M,N,M) model will be used for the forecast.

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.

Tab. 3.1. New and existing forecasted stores sales.

Date	New stores sales	Existing stores sales
Jan-16	2,532,275.58	21,829,060.03
Feb-16	2,534,886.22	21,146,329.63
Mar-16	2,538,156.82	23,735,686.94
Apr-16	2,538,122.84	22,409,515.28
May-16	2,543,397.94	25,621,828.73
Jun-16	2,543,039.84	26,307,858.04
Jul-16	2,546,073.44	26,705,092.56
Aug-16	2,548,503.29	23,440,761.33
Sep-16	2,557,980.73	20,640,047.32
Oct-16	2,558,588.30	20,086,270.46
Nov-16	2,561,618.79	20,858,119.96
Dec-16	2,554,737.81	21,255,190.24

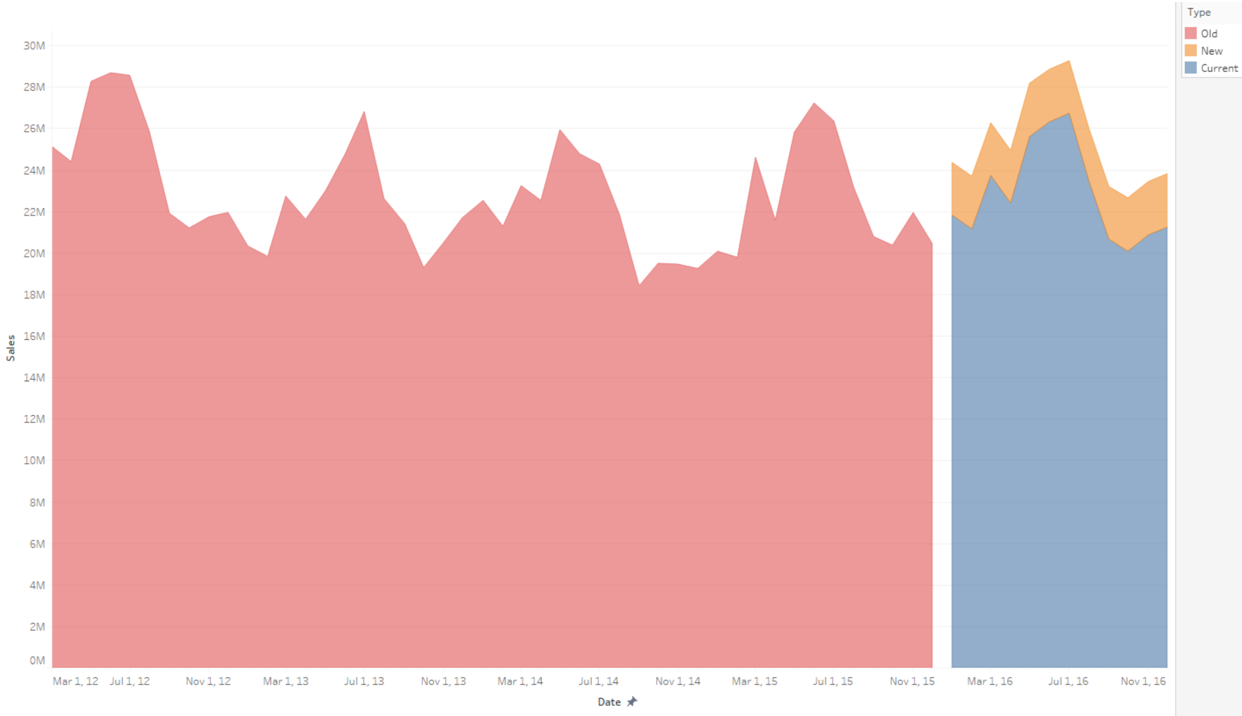

Fig. 3.2. Sales forecast.