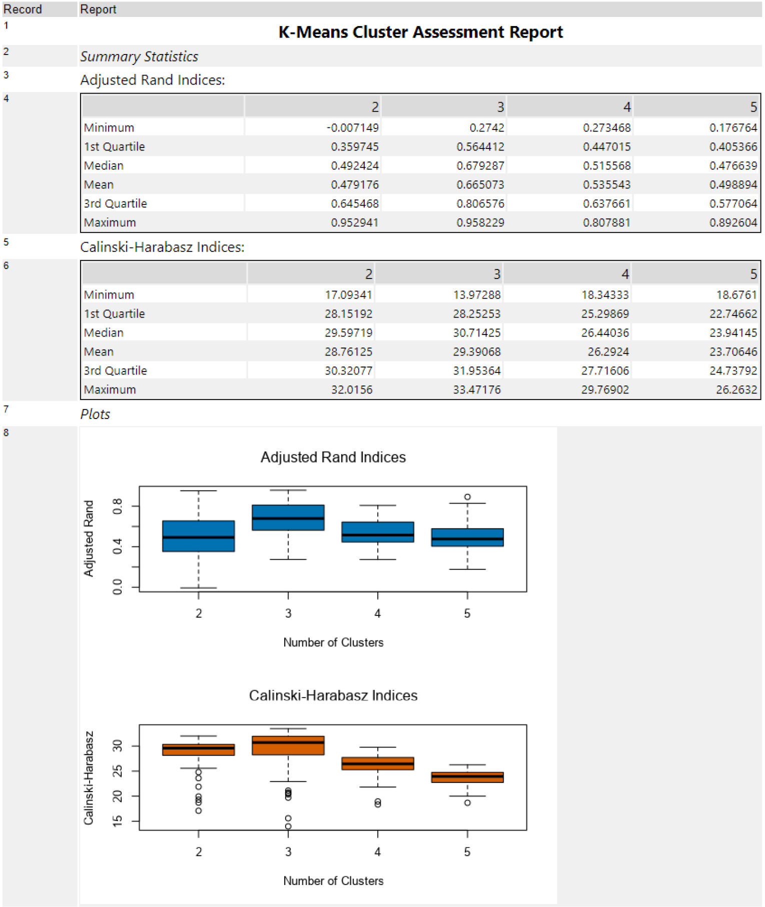
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   Three is the optimal number of store formats according to the indices of the K-Means Cluster Assessment Report. The Adjusted Rand index and Calinski-Harabasz index are at their highest median with 3 clusters.



Record | Report

1 K-Means Cluster Assessment Report

2 *Summary Statistics*

3 Adjusted Rand Indices:

4

	2	3	4	5
Minimum	-0.007149	0.2742	0.273468	0.176764
1st Quartile	0.359745	0.564412	0.447015	0.405366
Median	0.492424	0.679287	0.515568	0.476639
Mean	0.479176	0.665073	0.535543	0.498894
3rd Quartile	0.645468	0.806576	0.637661	0.577064
Maximum	0.952941	0.958229	0.807881	0.892604

5 Calinski-Harabasz Indices:

6

	2	3	4	5
Minimum	17.09341	13.97288	18.34333	18.6761
1st Quartile	28.15192	28.25253	25.29869	22.74662
Median	29.59719	30.71425	26.44036	23.94145
Mean	28.76125	29.39068	26.2924	23.70646
3rd Quartile	30.32077	31.95364	27.71606	24.73792
Maximum	32.0156	33.47176	29.76902	26.2632

7 *Plots*

8

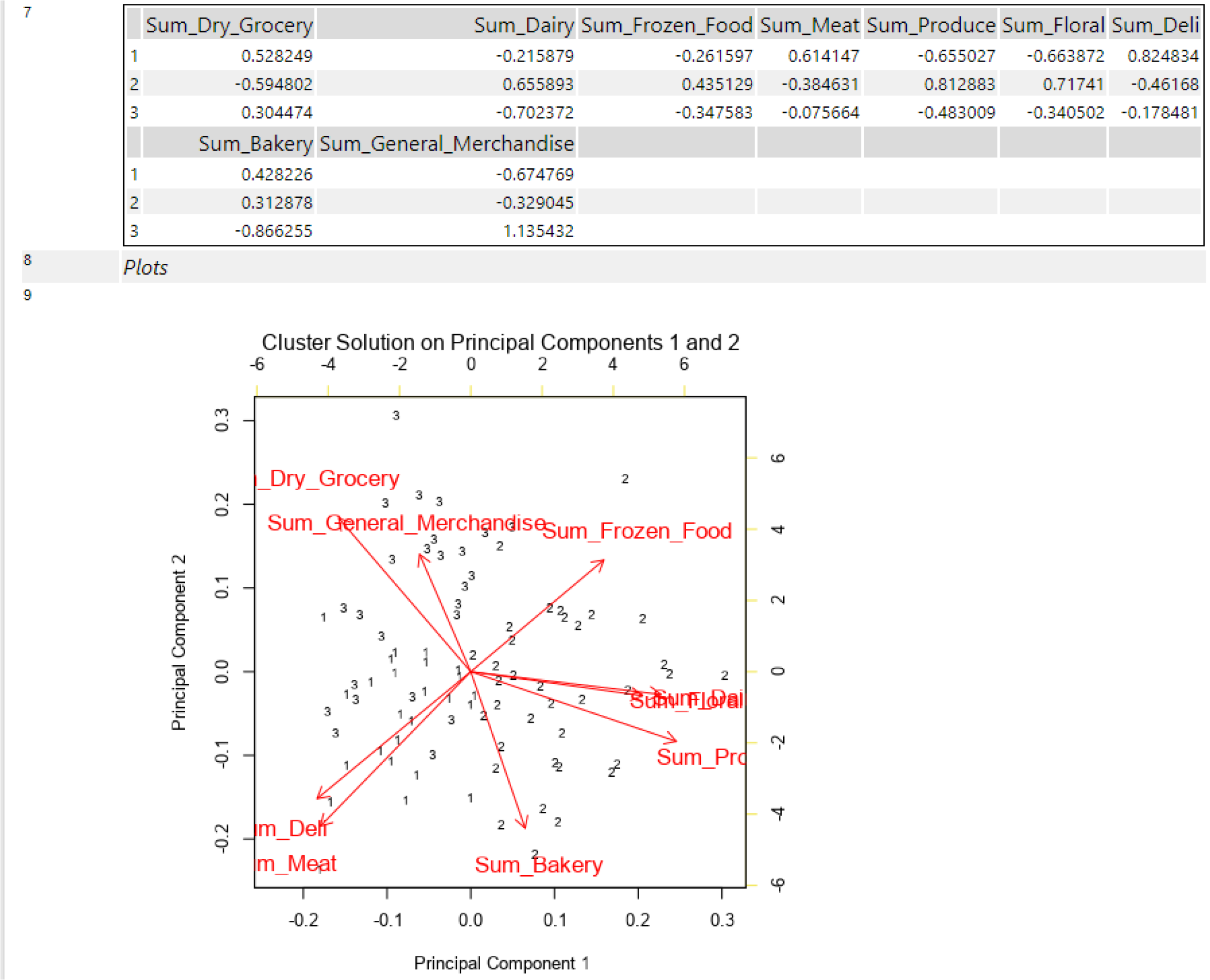
K-Means Cluster Assessment Report

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

   Cluster 1: 25 stores
   Cluster 2: 35 stores
   Cluster 3: 25 stores

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

   Cluster 3 is oriented more into the sales of General Merchandise which might indicate that there might not be many alternative or specialty stores near those locations. Cluster 1 sells a lot more meat than the other clusters. Cluster 1 and 2 are total opposites in floral sales.

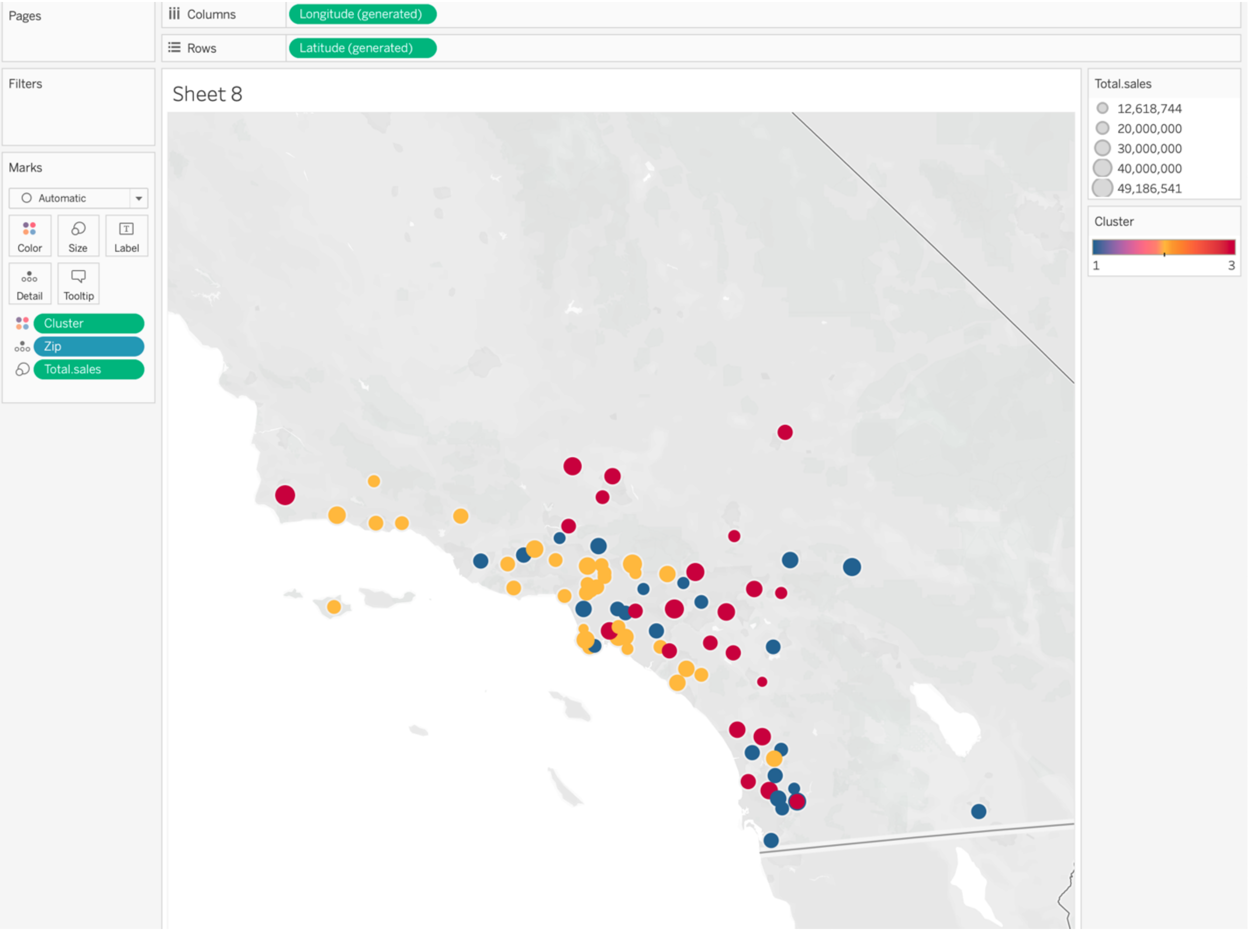


	Sum_Dry_Grocery	Sum_Dairy	Sum_Frozen_Food	Sum_Meat	Sum_Produce	Sum_Floral	Sum_Deli	Sum_Bakery	Sum_General_Merchandise
1	0.528249	-0.215879	-0.261597	0.614147	-0.655027	-0.663872	0.824834	0.428226	-0.674769
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168	0.312878	-0.329045
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178481	-0.866255	1.135432

Plots

Cluster differences

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.*



Columns: Longitude (generated)

Rows: Latitude (generated)

Marks: Automatic

Cluster

Zip

Total.sales

Sheet 8

Total.sales

○ 12,618,744

○ 20,000,000

○ 30,000,000

○ 40,000,000

○ 49,186,541

Cluster

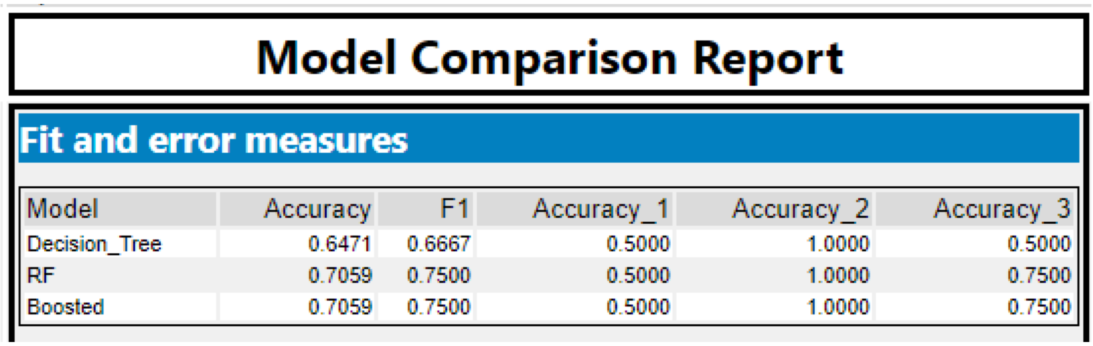
1

3

Tableau visualization of the locations, colors showing the store's format and size showing the total 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.)*



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
RF	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted	0.7059	0.7500	0.5000	1.0000	0.7500

Model comparison report

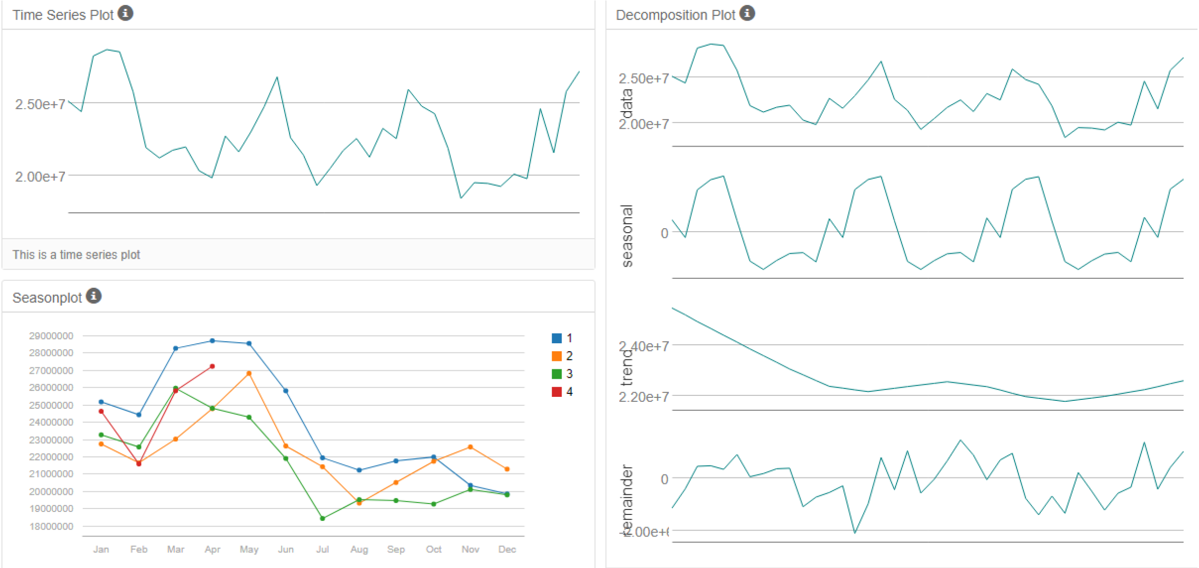
I decided to choose the Random Forest model for predicting the new stores even though the Boosted model had the same, equally high accuracy.

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	1
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

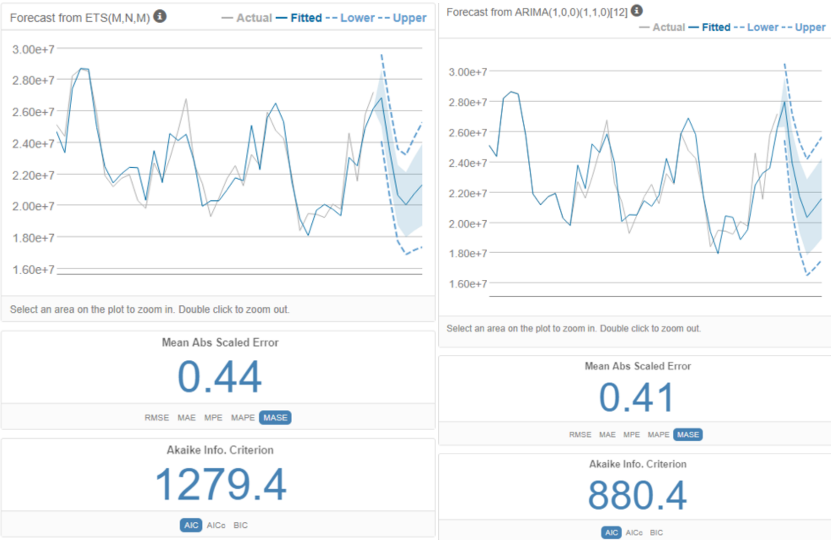
Task 3: Predicting Produce Sales

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?



Plotting the existing sales data

First, I decomposed the sales data and analyzed the seasonality, trend and error plots. For the ARIMA model I also had to make the data stationary with seasonal differencing. After this, I was going to use the ARIMA(1,0,0)(1,1,0)[12] model for the forecast since it had a lower AIC value and lower MASE compared to the ETS(M,N,M) model.



Forecast from ETS(M,N,M)
— Actual — Fitted -- Lower -- Upper
3.00e+7
2.80e+7
2.60e+7
2.40e+7
2.20e+7
2.00e+7
1.80e+7
1.60e+7
Select an area on the plot to zoom in. Double click to zoom out.
Forecast from ARIMA(1,0,0)(1,1,0)[12]
— Actual — Fitted -- Lower -- Upper
3.00e+7
2.80e+7
2.60e+7
2.40e+7
2.20e+7
2.00e+7
1.80e+7
1.60e+7
Select an area on the plot to zoom in. Double click to zoom out.
RMSE
MAE
MPE
MAPE
MASE
AIC
AICc
BIC

Mean Abs Scaled Error 0.44

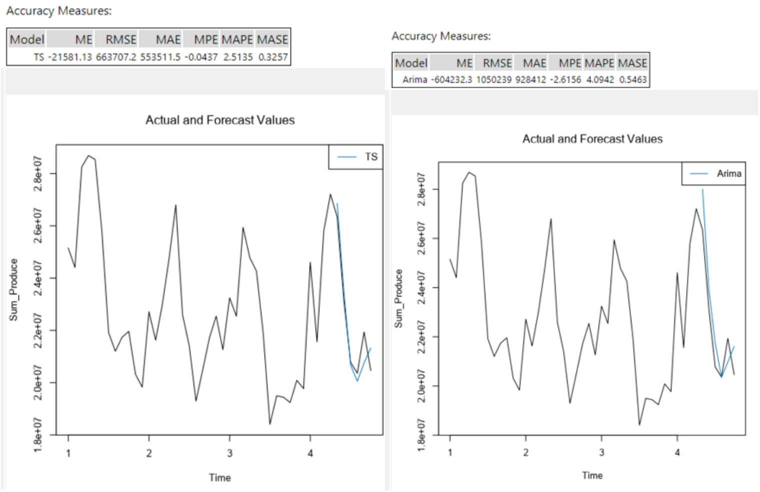
Akaike Info. Criterion 1279.4

Mean Abs Scaled Error 0.41

Akaike Info. Criterion 880.4

ETS and ARIMA forecasts

But then I compared the ETS and ARIMA models against the 6-month holdout sample and the ETS(M,N,M) delivered more accurate results with lower RMSE and MASE. I decided to go with the ETS model after this comparison.



Accuracy Measures:

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

Accuracy Measures:

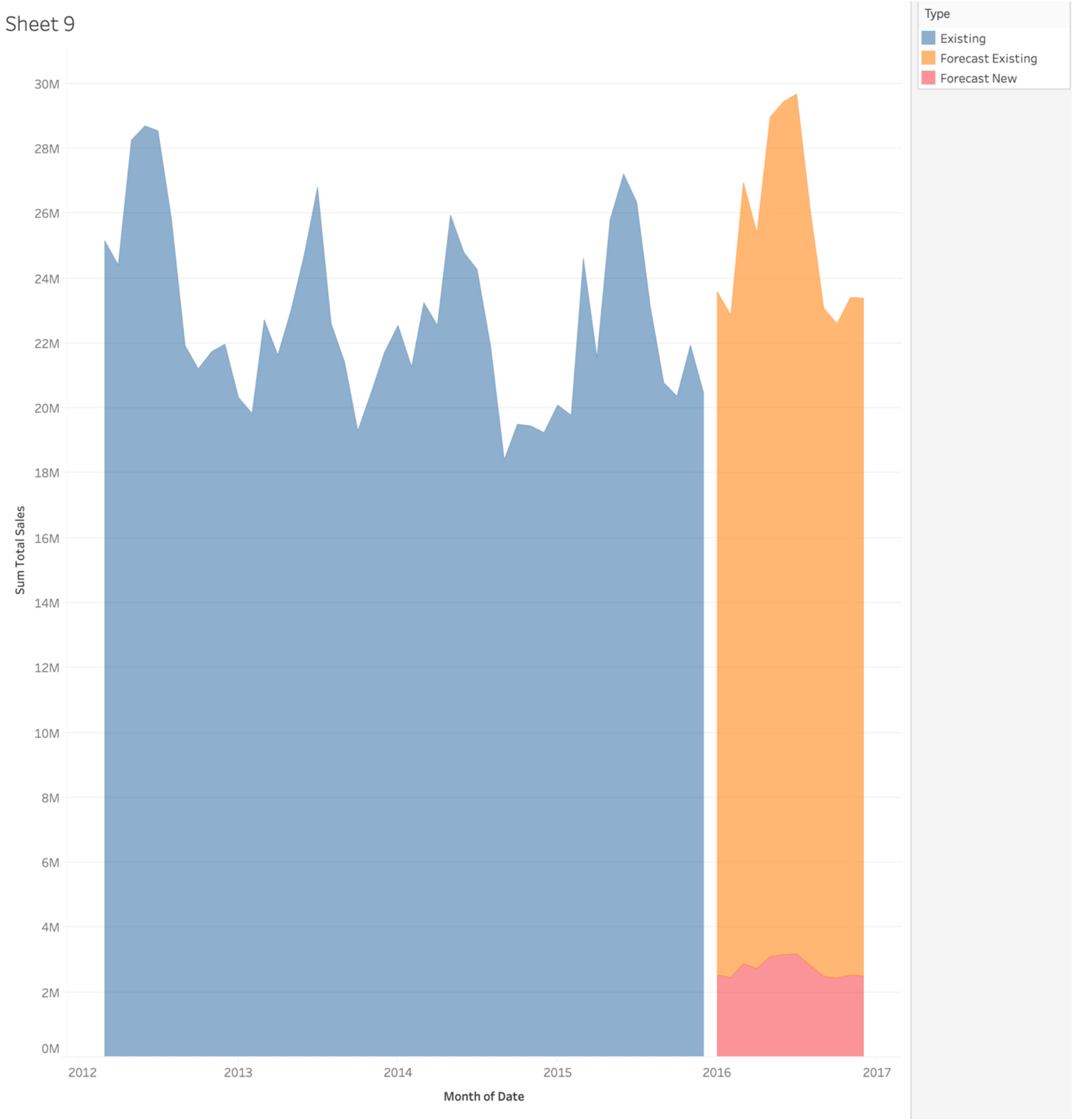
Model	ME	RMSE	MAE	MPE	MAPE	MASE
Arima	-604232.3	1050239	928412	-2.6156	4.0942	0.5463

Accuracy of the models against the 6-month holdout sample

3. *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.*

Here’s the table for the forecasted produce sales data of the new and existing stores.

Date	New stores	Existing stores
JAN-16	$ 2 527 338,50	$ 21 057 160,62
FEB-16	$ 2 446 154,76	$ 20 415 891,84
MAR-16	$ 2 872 050,73	$ 24 078 058,16
APR-16	$ 2 722 157,62	$ 22 670 735,53
MAY-16	$ 3 098 095,87	$ 25 858 187,53
JUN-16	$ 3 150 602,99	$ 26 288 436,90
JUL-16	$ 3 172 545,05	$ 26 501 400,91
AUG-16	$ 2 814 269,98	$ 23 303 548,46
SEP-16	$ 2 486 631,56	$ 20 583 812,16
OCT-16	$ 2 434 261,23	$ 20 160 031,58
NOV-16	$ 2 517 523,25	$ 20 888 455,26
DEC-16	$ 2 491 340,44	$ 20 891 395,24



Visualization of the forecasts and existing sales data