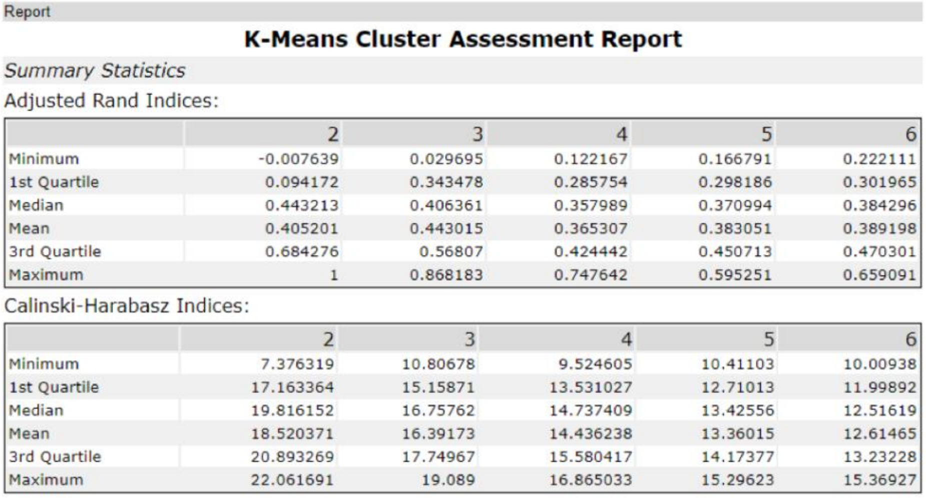
Project: Predictive Analytics Capstone

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 3. From the K-Means Cluster assessment report, Adjusted Rand and Calinski-Harabasz indices are used to determine the median and spread by each cluster. Based on that, we can see that cluster 3 in Adjusted Rand and Calinski-Harabasz indices registered the highest median value show a strong indication. This means that it is most stable, and the clusters are more distinct and compact.



Report

K-Means Cluster Assessment Report

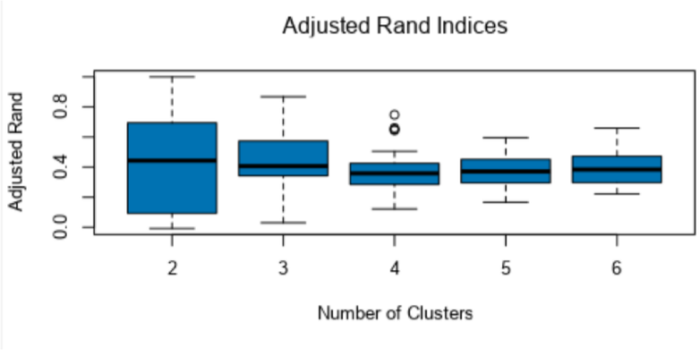
Summary Statistics

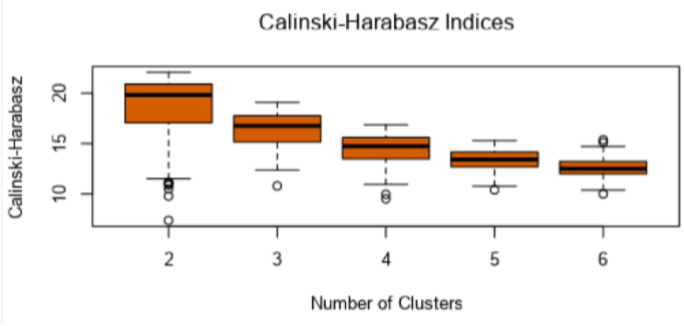
Adjusted Rand Indices:

	2	3	4	5	6
Minimum	-0.007639	0.029695	0.122167	0.166791	0.222111
1st Quartile	0.094172	0.343478	0.285754	0.298186	0.301965
Median	0.443213	0.406361	0.357989	0.370994	0.384296
Mean	0.405201	0.443015	0.365307	0.383051	0.389198
3rd Quartile	0.684276	0.56807	0.424442	0.450713	0.470301
Maximum	1	0.868183	0.747642	0.595251	0.659091

Calinski-Harabasz Indices:

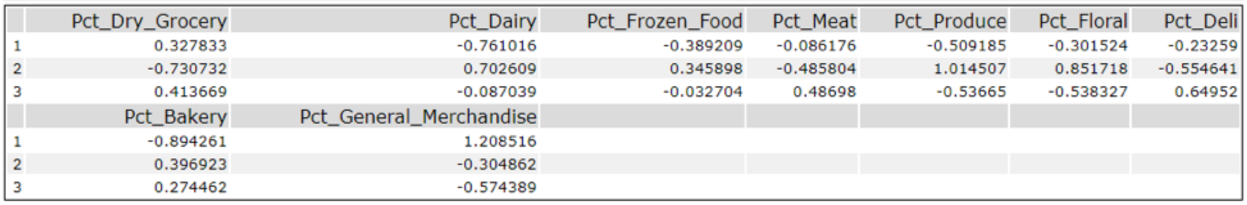
	2	3	4	5	6
Minimum	7.376319	10.80678	9.524605	10.41103	10.00938
1st Quartile	17.163364	15.15871	13.531027	12.71013	11.99892
Median	19.816152	16.75762	14.737409	13.42556	12.51619
Mean	18.520371	16.39173	14.436238	13.36015	12.61465
3rd Quartile	20.893269	17.74967	15.580417	14.17377	13.23228
Maximum	22.061691	19.089	16.865033	15.29623	15.36927





2. How many stores fall into each store format?
   - Cluster 1: 23 stores.
   - Cluster 2: 29 stores.
   - Cluster 3: 33 stores.

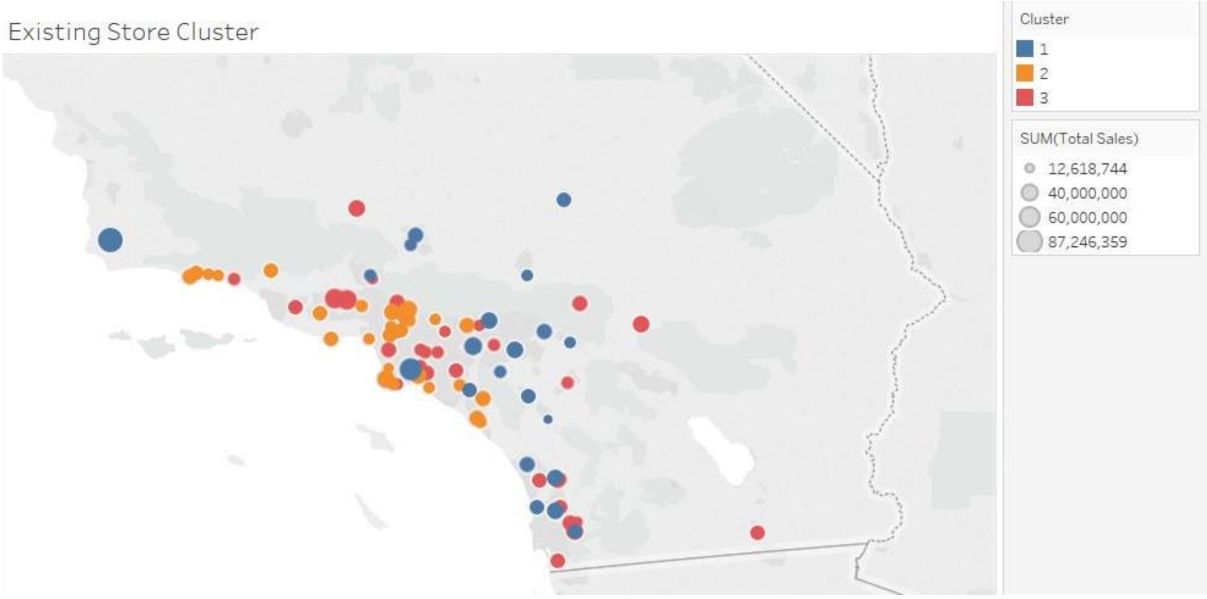
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   One way that the clusters differ from one another could be through considering the percentage of sales by category of each store. For example, cluster 1 has the most positive percentage sales in general merchandise comparing to cluster 3 with has the most negative percentage. cluster 2 has the most positive percentage sales in in frozen food, produce product, and dairy product compared to cluster 1 and 3.



	Pct_Dry_Grocery	Pct_Dairy	Pct_Frozen_Food	Pct_Meat	Pct_Produce	Pct_Floral	Pct_Deli	Pct_Bakery	Pct_General_Merchandise
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259	-0.894261	1.208516
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641	0.396923	-0.304862
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952	0.274462	-0.574389

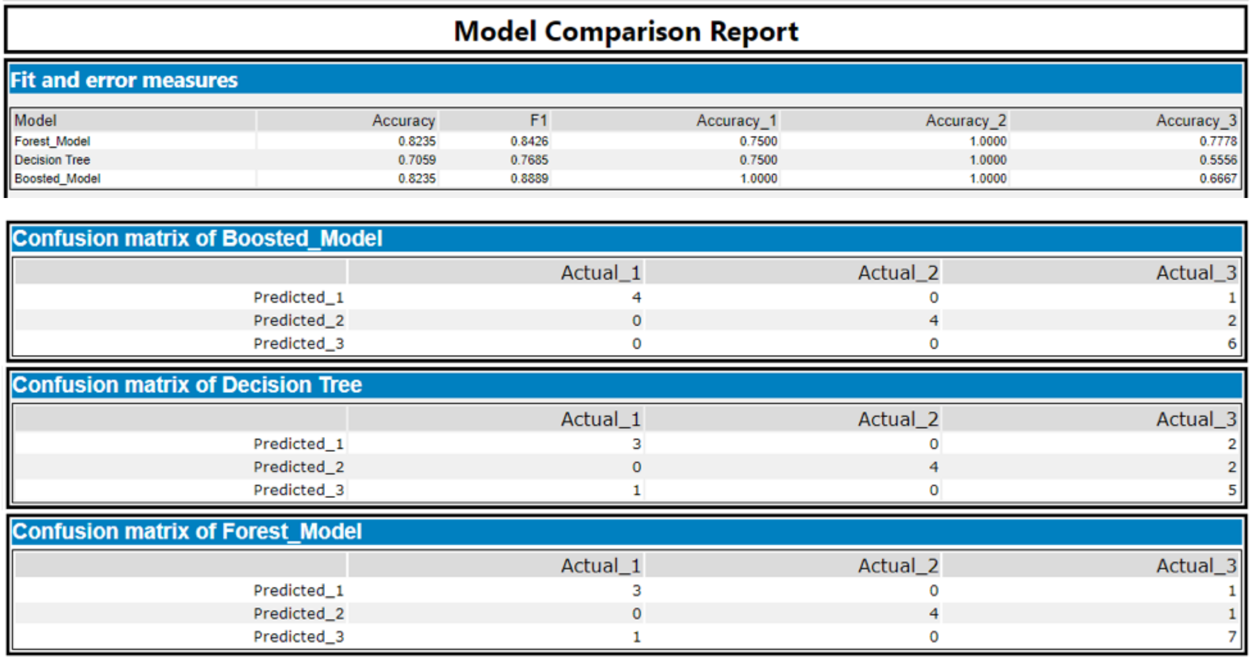
   Another way is by comparing the physical distance between these stores. For example, cluster 2 has the closest range of physical distance between stores where all stores are in same area which is the opposite with cluster 3 where spread randomly through southern California.

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.



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.)
   In order to find the most suitable methodology to predict the best store format for the new stores, model comparison report is used to compare the Decision Tree, Forest Model and Boosted Model. Although it has the same accuracy as Forest Model, the Boosted Model is chosen to predict the best store format for the new stores since result shows it has the higher F1 value.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Forest_Model	0.8235	0.8426	0.7500	1.0000	0.7778
Decision Tree	0.7059	0.7685	0.7500	1.0000	0.5556
Boosted_Model	0.8235	0.8889	1.0000	1.0000	0.6667

Confusion matrix of Boosted_Model

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

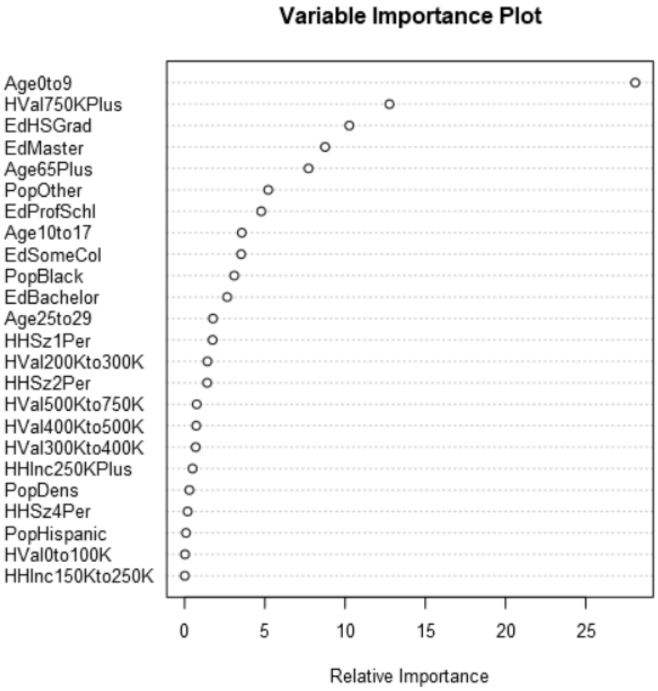
Confusion matrix of Decision Tree

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

Confusion matrix of Forest_Model

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

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.
   The three most important variables that help explain the relationship between demographic indicators and store formats are Age0to9, HVal750KPlus, and EdHSGrad.



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

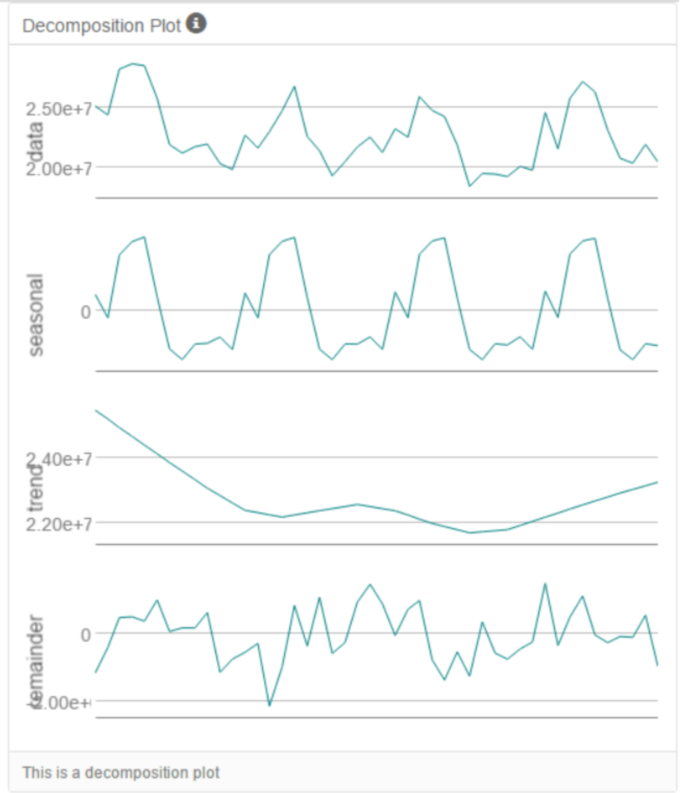
Store Number	Segment
S0086	**3**
S0087	**2**
S0088	**1**
S0089	**2**
S0090	**2**
S0091	**1**
S0092	**2**
S0093	**1**
S0094	**2**
S0095	**2**

Task 3: Predicting Produce Sales

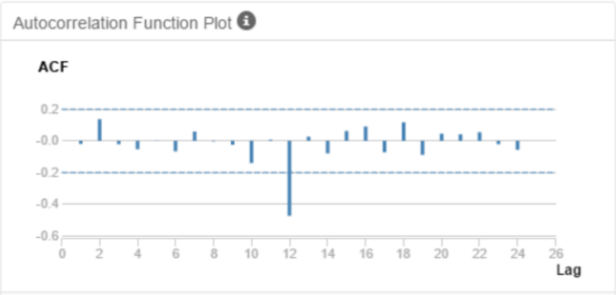
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

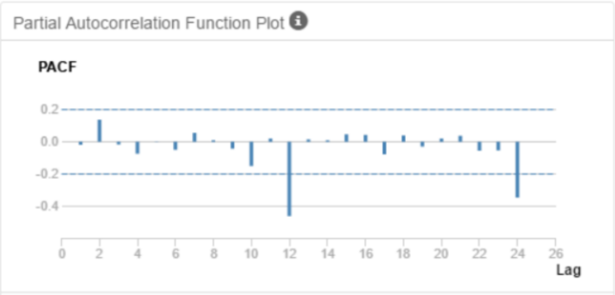
To find the suitable type of model to be used for each forecast, I compared the ETS(M,N,M) and the ARIMA (1,0,0)(1,1,0)12.

For ETS model, the decomposition plot shows that the error is multiplicative, no trend, and the seasonality is multiplicative. For that the (M,N,M) is chosen for ETS model.



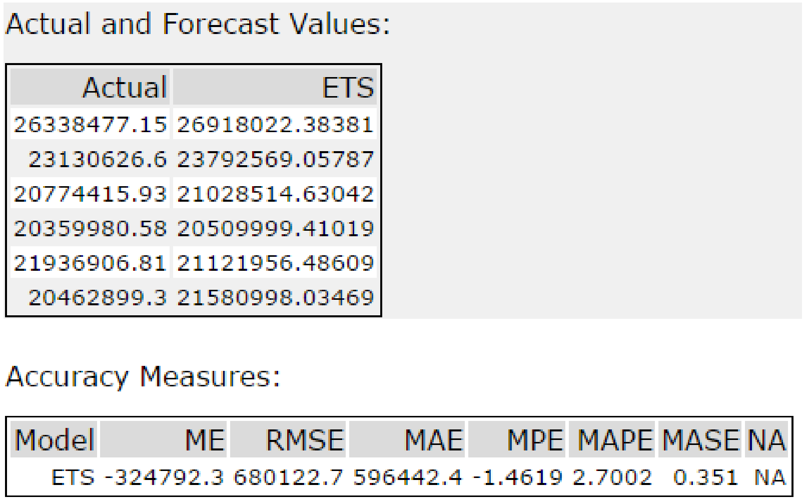
For ARIMA model, the ACF and PACF graphs of autocorrelation function plot shows that it has a negative correlation as AR 1 and MA 0.





ETS(M,N,M) and ARIMA (1,0,0)(1,1,0)12 models compared where a holdout period of 6 periods used to validate the models based on forecasted values compared to the actual and the accuracy measures.

For ETS(M,N,M):



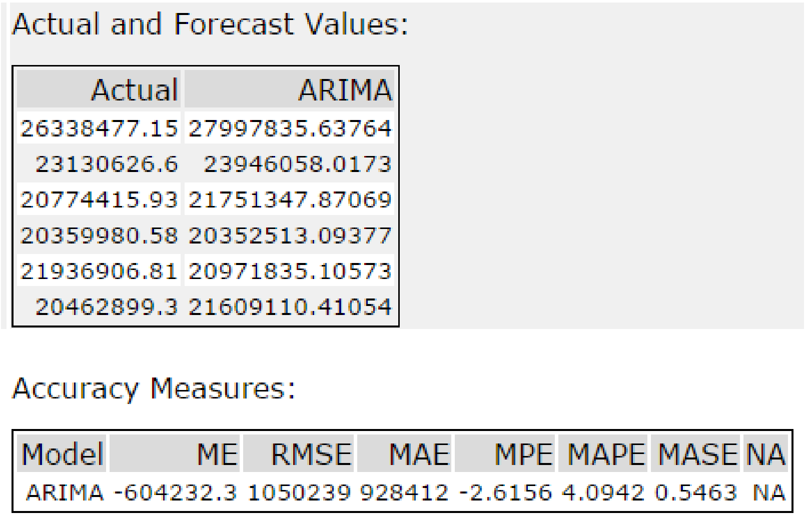
Actual and Forecast Values:

Actual	ETS
26338477.15	26918022.38381
23130626.6	23792569.05787
20774415.93	21028514.63042
20359980.58	20509999.41019
21936906.81	21121956.48609
20462899.3	21580998.03469

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ETS	-324792.3	680122.7	596442.4	-1.4619	2.7002	0.351	NA

For 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:

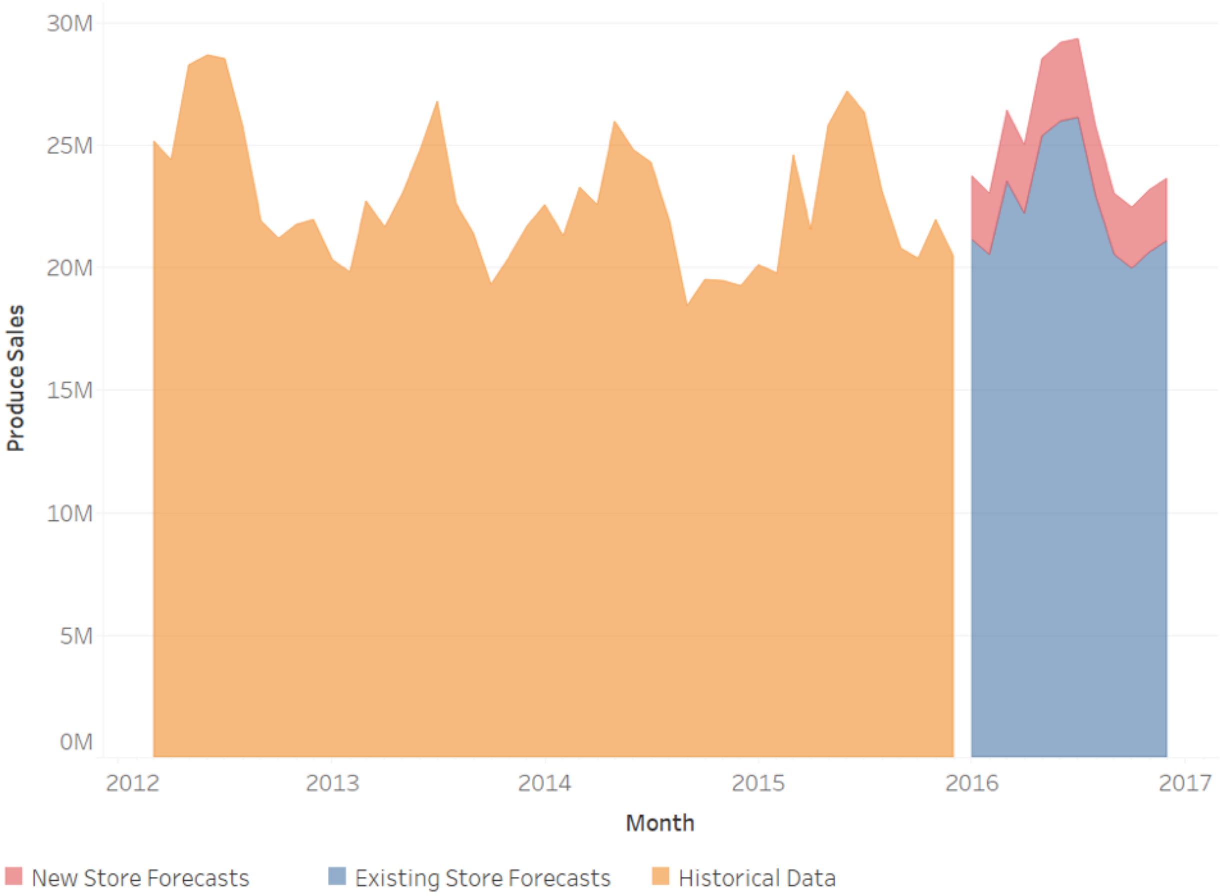
Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
ARIMA	-604232.3	1050239	928412	-2.6156	4.0942	0.5463	NA

After comparing the ETS(M,N,M) and ARIMA (1,0,0)(1,1,0)12 models, I chose the ETS(M,N,M) to be used for each forecast. This is because the ETS model have higher accuracy and lower error value compared to ARIMA model.

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.

Year	Month	New Stores	New Stores
2016	1	2,588,250	21,136,642
2016	2	2,499,159	20,507,039
2016	3	2,916,908	23,506,566
2016	4	2,791,560	22,208,406
2016	5	3,156,890	25,380,148
2016	6	3,200,940	25,966,799
2016	7	3,224,858	26,113,793
2016	8	2,861,958	22,899,286
2016	9	2,534,353	20,499,584
2016	10	2,481,117	19,971,243
2016	11	2,578,336	20,602,666
2016	12	2,561,917	21,073,222

Produce Sales Forecast


30M
25M
20M
15M
10M
5M
0M
Produce Sales
Month
New Store Forecasts
Existing Store Forecasts
Historical Data