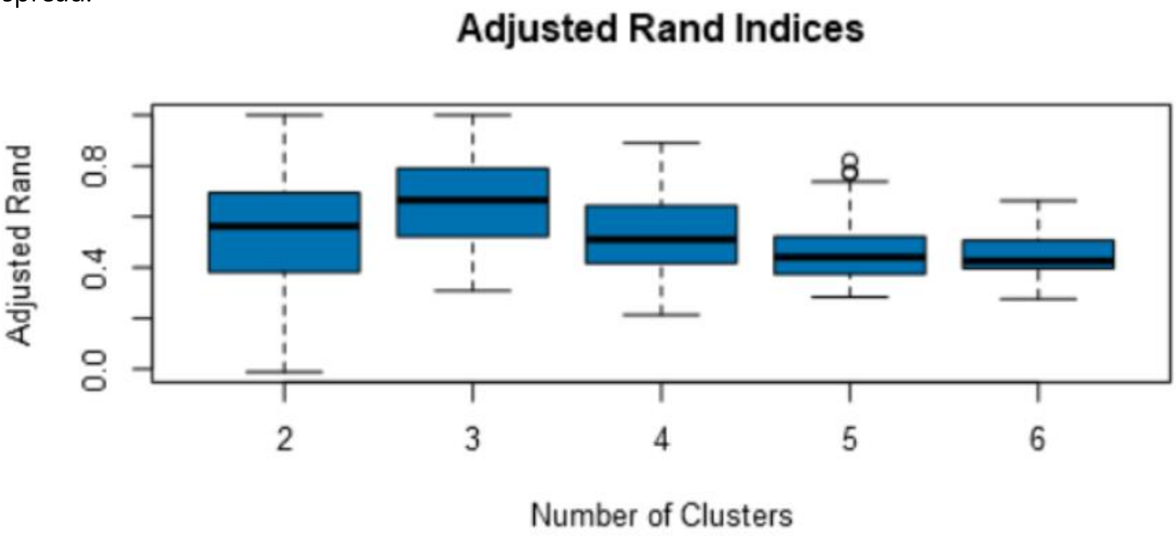
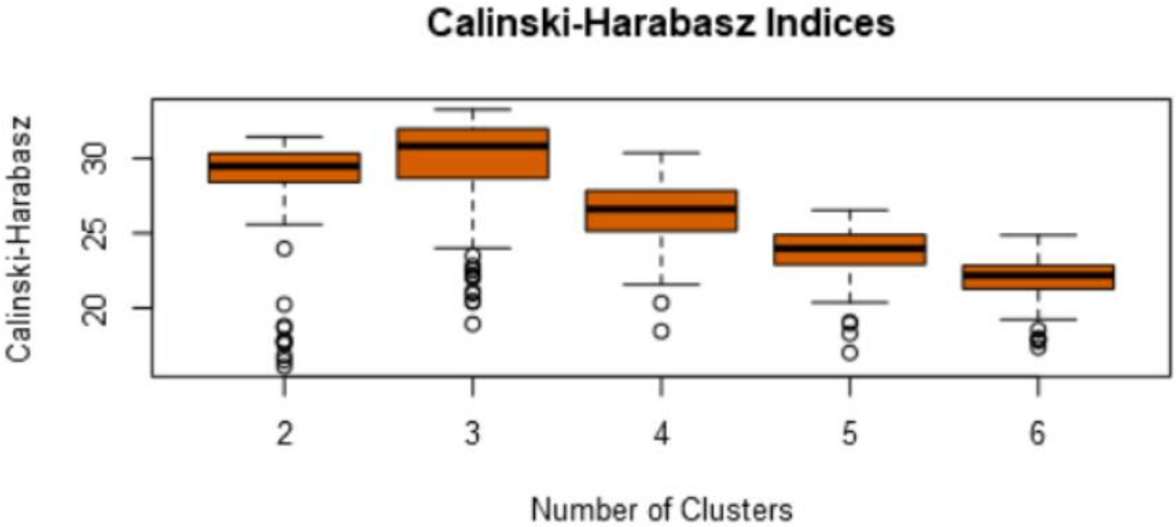
Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

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 stores is 3 based on the k means report , Adjusted Rand and Calinski-Harabasz as they both showed highest median value and smallest variation in spread.

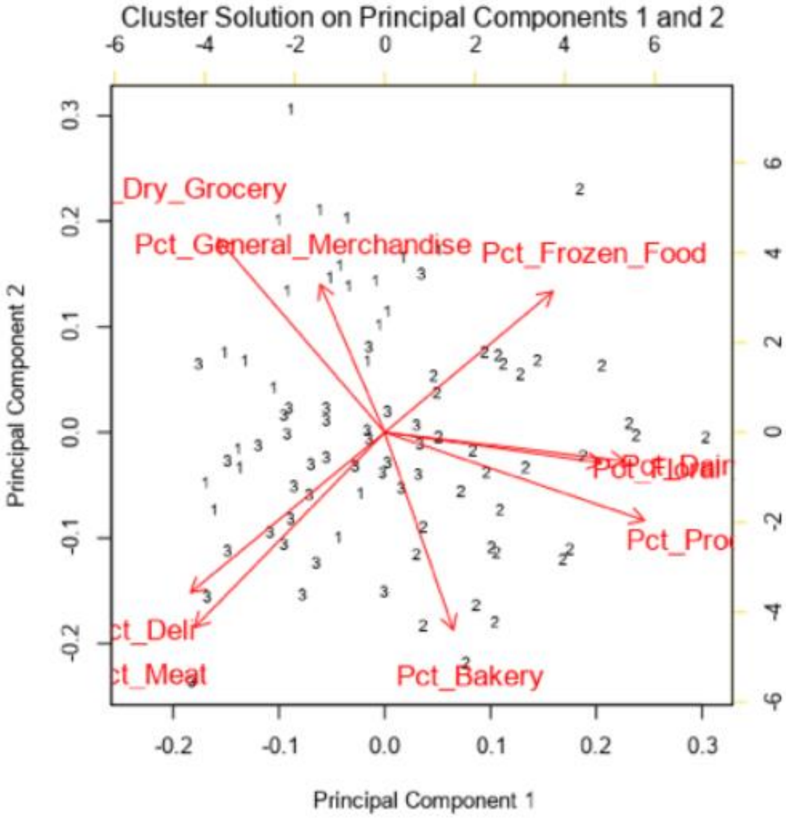




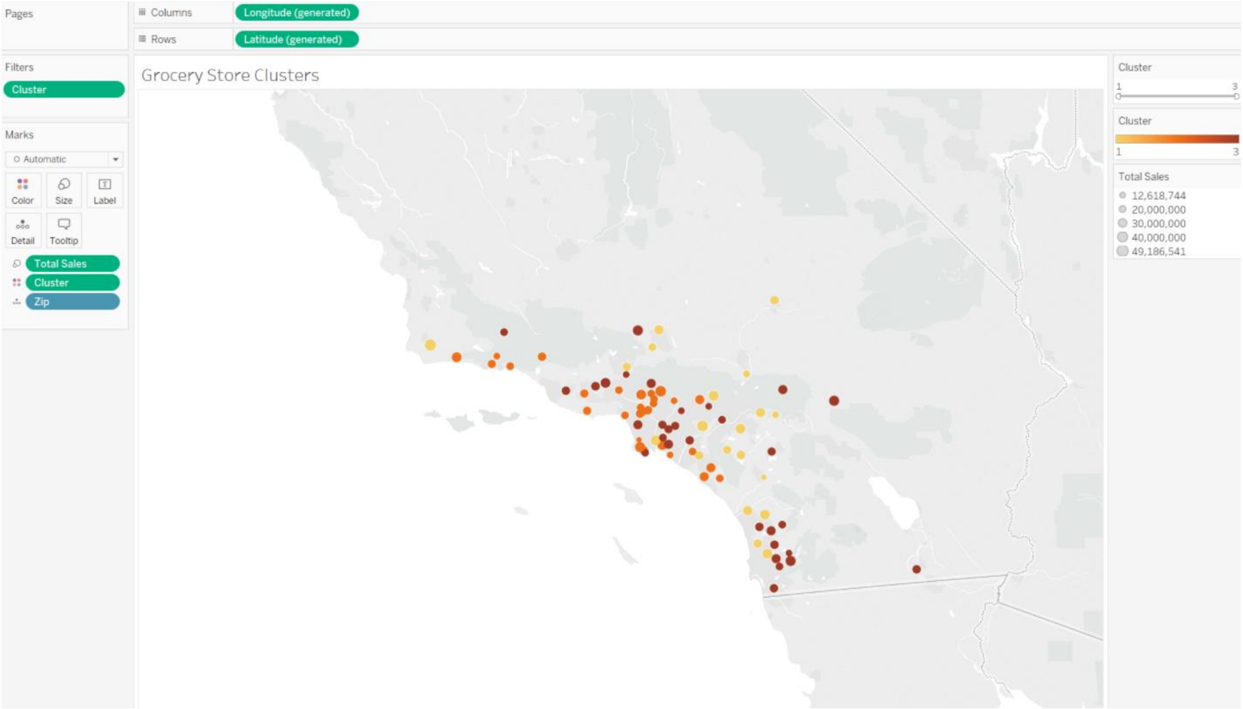
2. How many stores fall into each store format?

   The first cluster has 23 stores and the second has 29 stores and the third 33 stores

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   Based on the cluster plot below one way that could differ one cluster from another is the percentage of sales by category as shown below cluster 1 is highly affected by dry and General Mechanism and cluster 2 sells more in produce and floral while cluster 3 sells more meat and deli



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.



Pages

Columns: Longitude (generated)

Rows: Latitude (generated)

Filters: Cluster

Marks: Automatic

Color Size Label Detail Tooltip

Total Sales

Cluster

Zip

Grocery Store Clusters

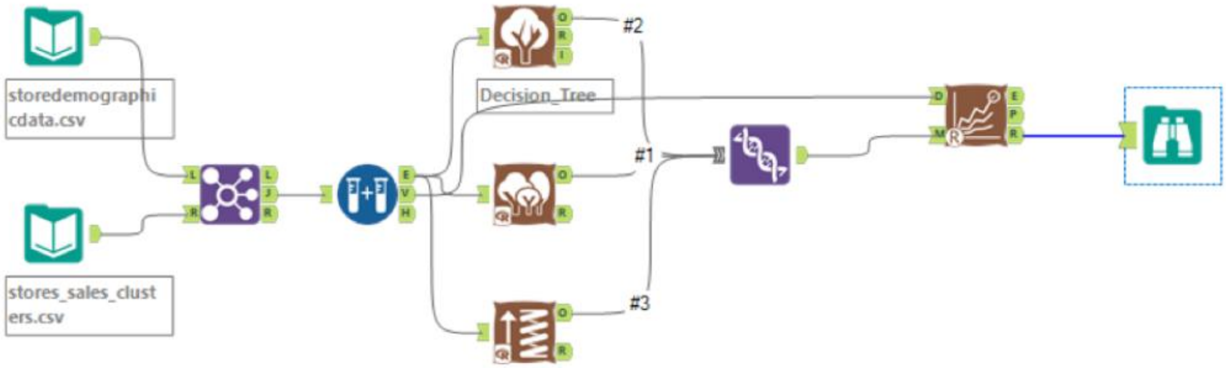
Cluster 1 3

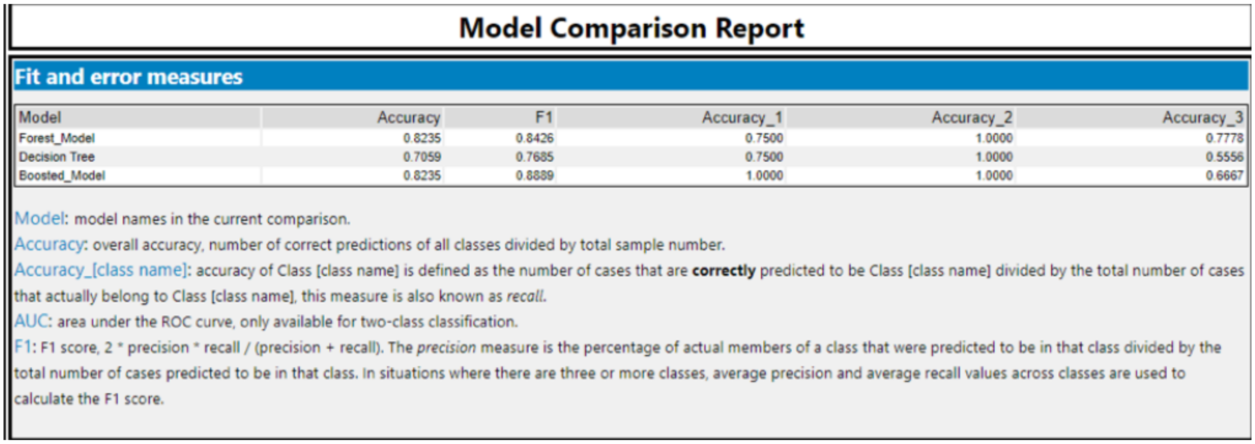
Cluster 1 3

Total Sales
- 12,618,744
- 20,000,000
- 30,000,000
- 40,000,000
- 49,186,541

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 made a comparison between the results of training three models which are the decision tree ,forest model and the boosted model and after using the comparison tool I choose the boosted model as it has better f1 score of 88%





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

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.

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

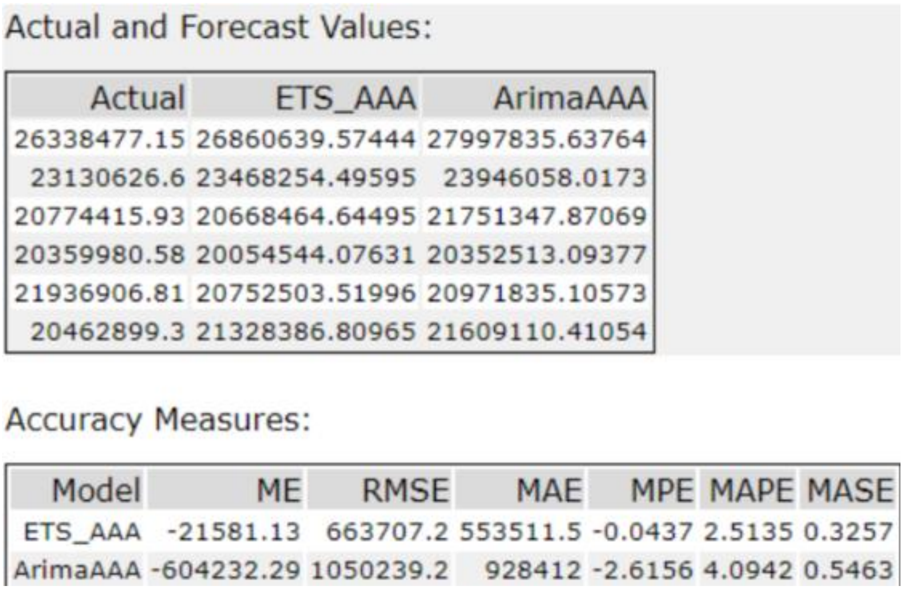
Store Number	Segment
S0086	3
S0087	2
S0088	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

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?

The decision was to chose the ETS(M,N,M) model for forecasting as after using the auto selection of type for AR ,MA and I values in ARIMA the model gave an output of (1,0,0)(1,1,0) configuration and after using the auto selection type for error seasonality and trend the model gave and output of MNM configuration for the ETS model.

After combining the results using union using the TS compare the ETS model showed more accurate prediction to the hold off samples .

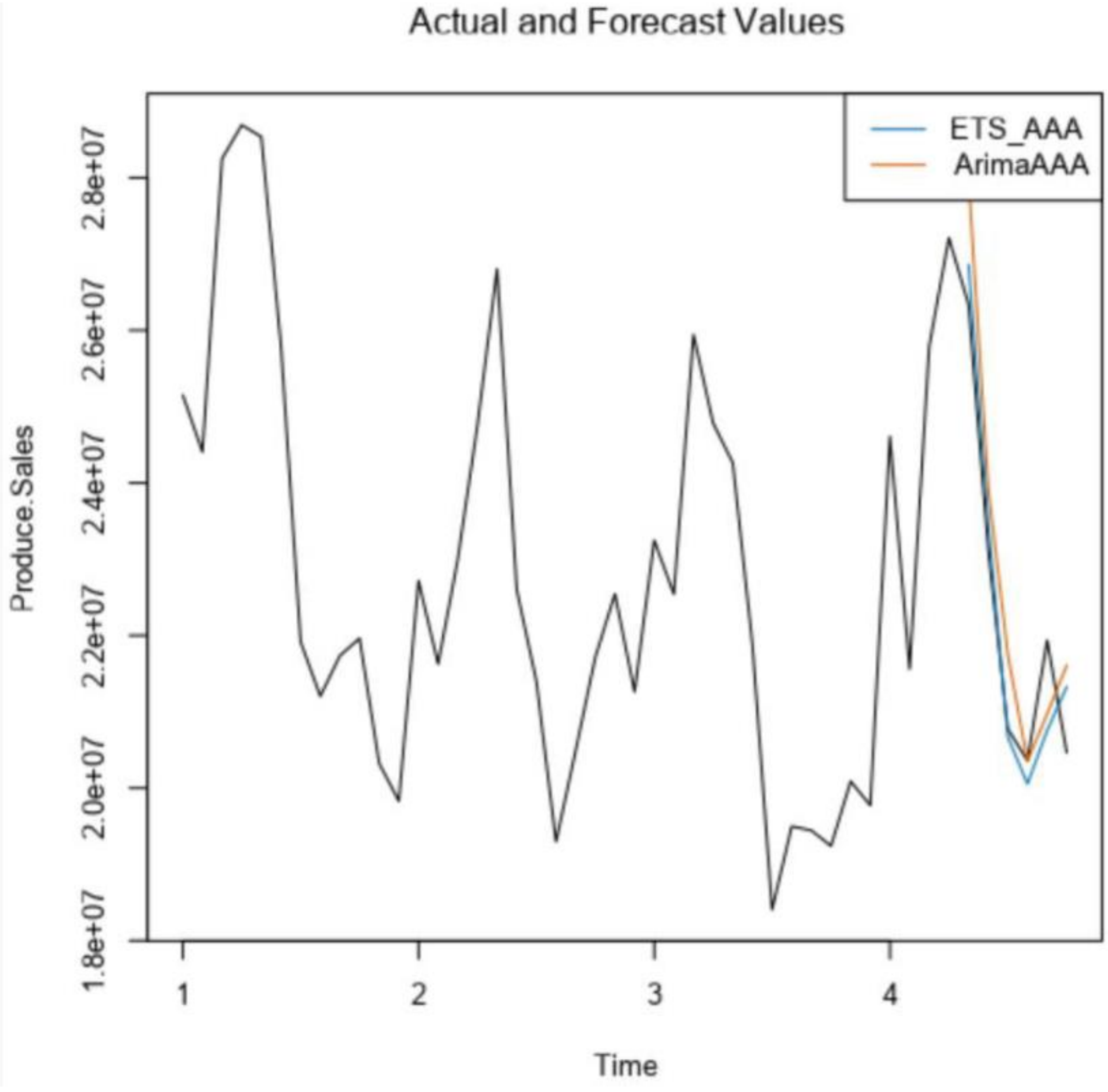


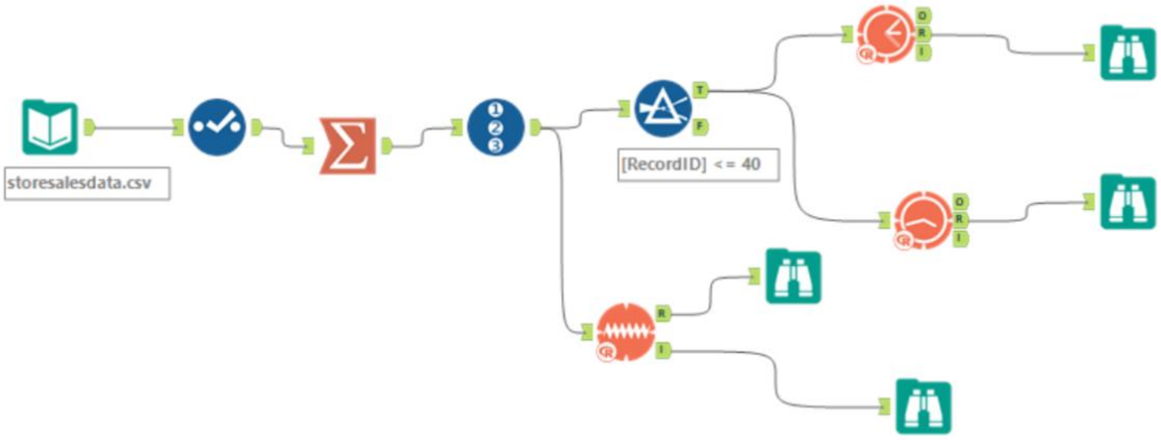
Actual and Forecast Values:

Actual	ETS_AAA	ArimaAAA
26338477.15	26860639.57444	27997835.63764
23130626.6	23468254.49595	23946058.0173
20774415.93	20668464.64495	21751347.87069
20359980.58	20054544.07631	20352513.09377
21936906.81	20752503.51996	20971835.10573
20462899.3	21328386.80965	21609110.41054

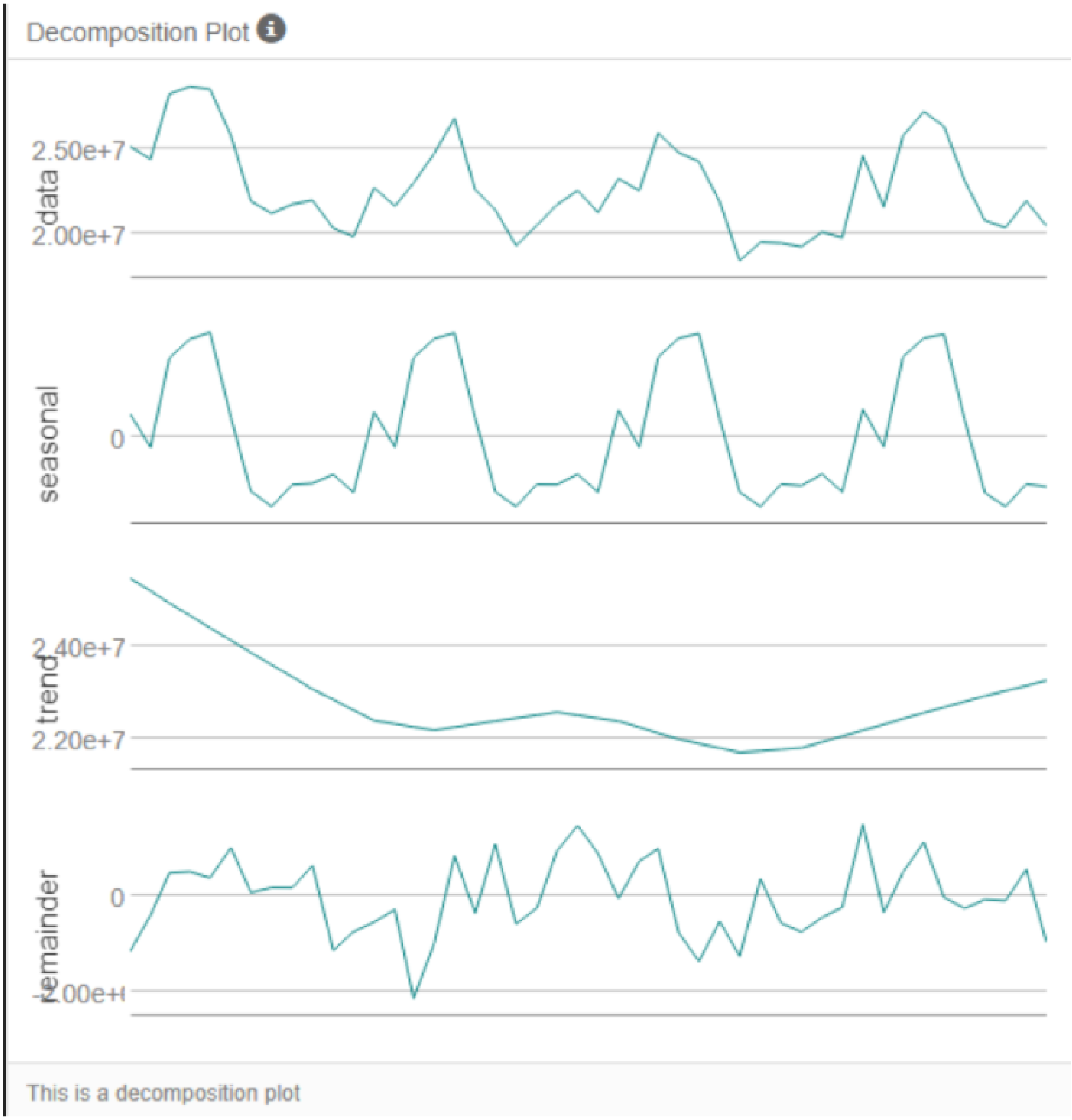
Accuracy Measures:

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





And as shown in the decomposition plot below seasonal and error are multiplicative and there is no trend



Decomposition Plot

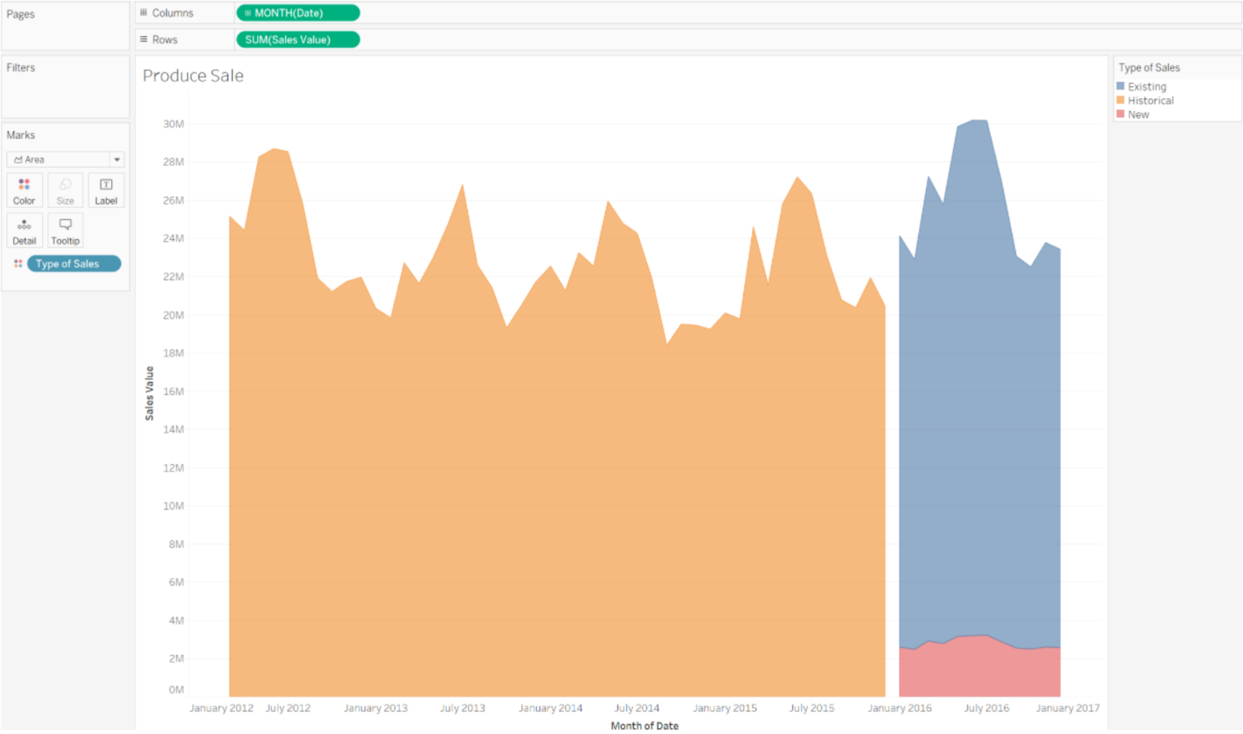
This is a decomposition plot

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.

Date	New stores sales	Existing stores sales
January 2016	2587451	21539936
February 2016	2477353	20413771
March 2016	2477353	24325953
April 2016	2775746	22993466
May 2016	3150867	26691951
June 2016	3188922	26989964
July 2016	3214746	26948631
August 2016	2866346	24091579

September 2016	2538727	20523492
October 2016	2488148	20011749
November 2016	2595270	21177435
December 2016	2573397	20855799

The plot below shows the historical data in orange and the existing stores forcasts blue and the new stores forcast in pink



Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.