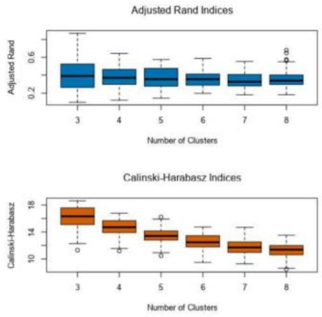
Project: Predictive Analytics Capstone

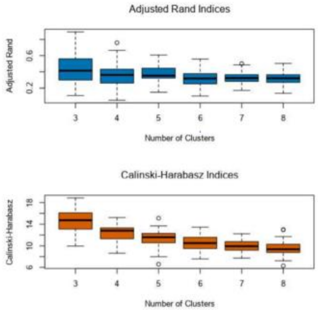
Task 1: Determine Store Formats for Existing Stores

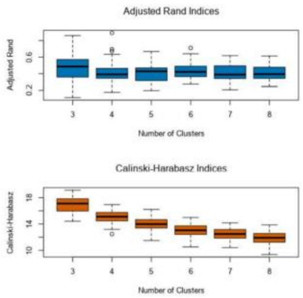
1. What is the optimal number of store formats? How did you arrive at that number?

   I have run k-centroids diagnostics tool for all three clustering methods for an internal validation. Below are the indices for each:

   K-MEANS ***K-MEDIANS*** ***NEURAL GAS***



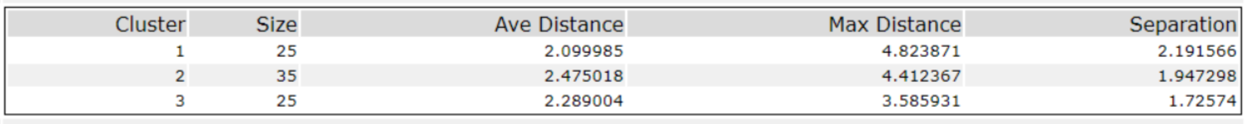




   ***Clearly **3 clusters** have the highest mean-median in all methods with a fairly small range of 1st to 3rd quartiles.*

2. How many stores fall into each store format?

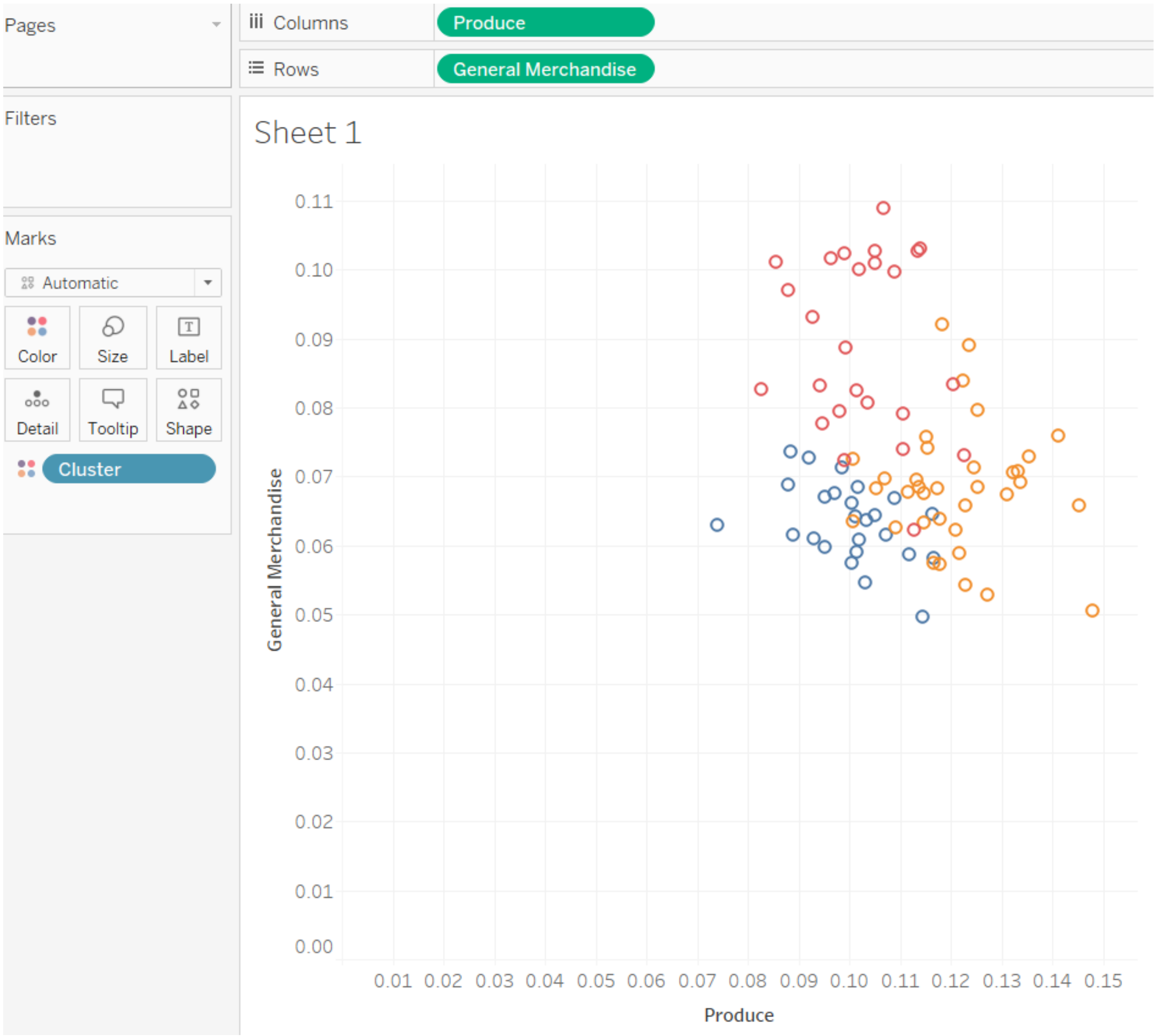
Below is the clustering solution of K-Means (with z standardization applied, 3 clusters and 10 starting seeds)



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

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

Comparing the percentage sales, cluster 1 has higher percent of General Merchandise sold, whereas Cluster 3 has a higher for Produce Sales.



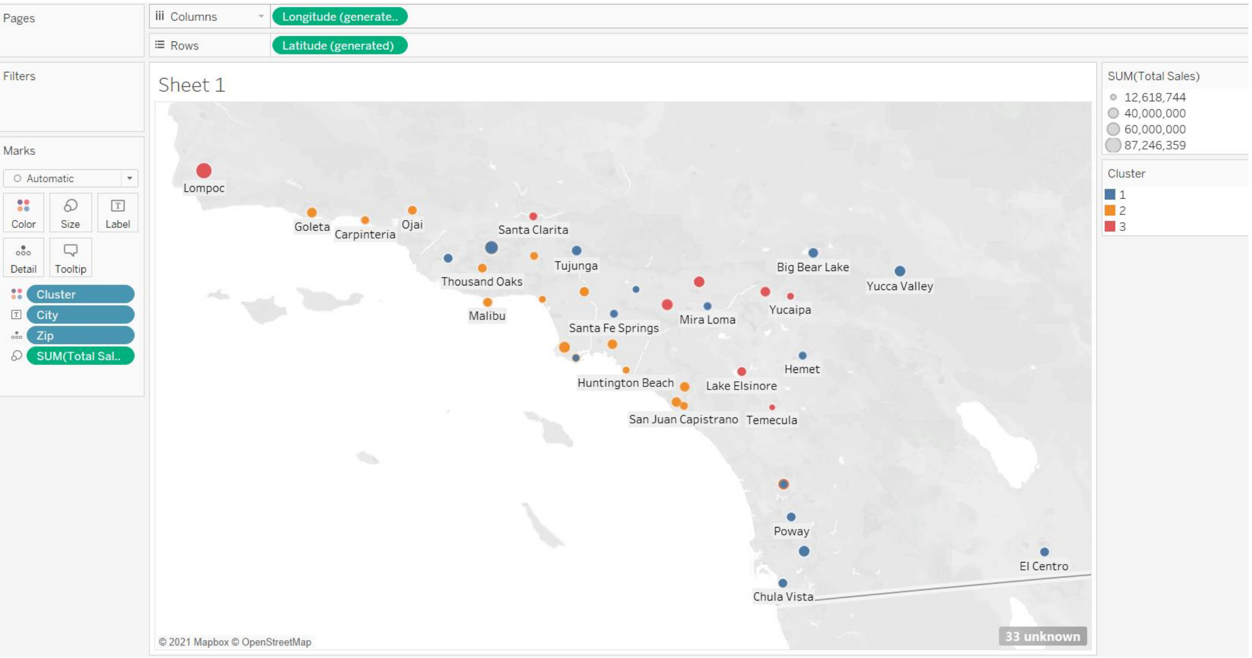
Columns: Produce

Rows: General Merchandise

Sheet 1

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.

Below is the final map of stores, clusters are displayed with color&label, sales is shown with size. 33 cities remain unknown,could not resolve that issue with Tableau.



Columns: Longitude (generate..

Rows: Latitude (generated)

Marks: Automatic

Cluster
City
Zip
SUM(Total Sal..

SUM(Total Sales)
○ 12,618,744
○ 40,000,000
○ 60,000,000
○ 87,246,359

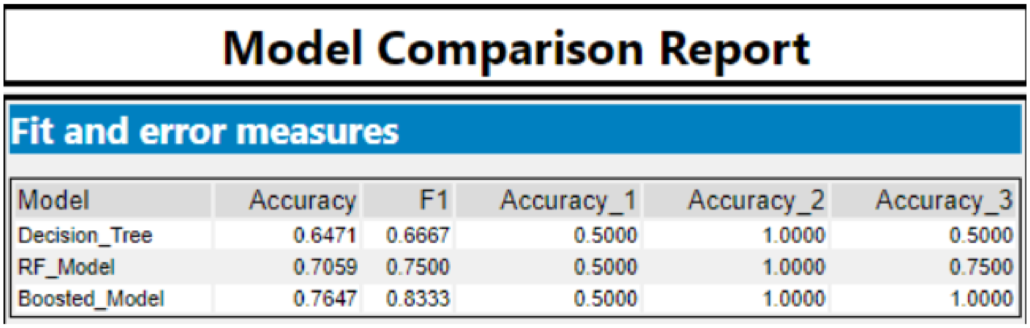
Cluster
🟦 1
🟧 2
🟥 3

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.)

   After data preprocessing and sampling, I applied Decision Tree, Forest Model and Boosted Model. Then compared all models with the Model Comparison tool using the validation data.

   Boosted Model had the highest accuracy and F1 score in predictions.



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_Model	0.7059	0.7500	0.5000	1.0000	0.7500
Boosted_Model	0.7647	0.8333	0.5000	1.0000	1.0000

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

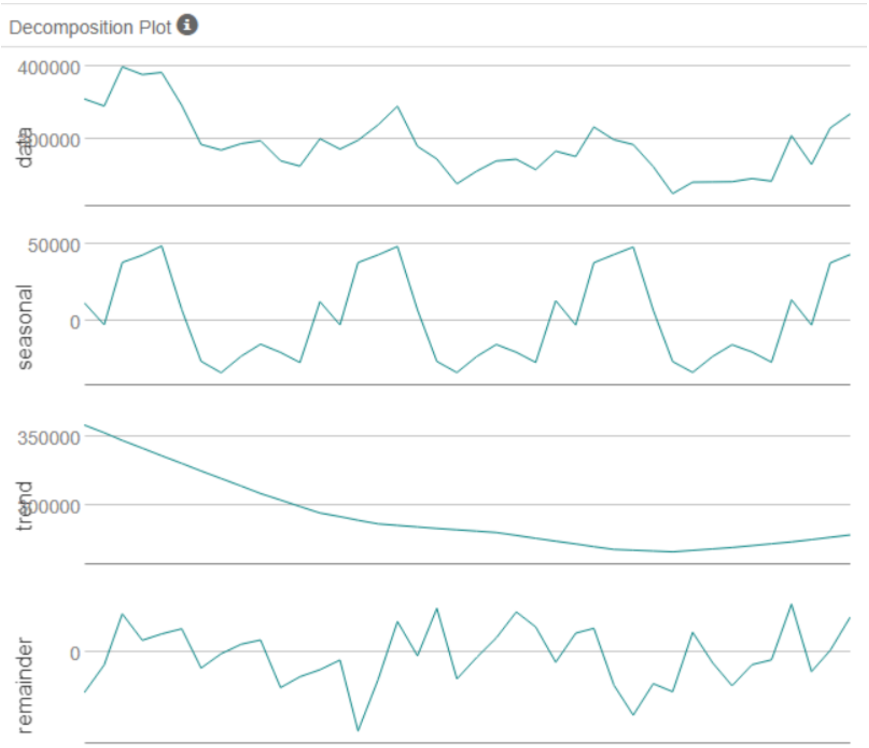
Store Number	Segment
S0086	1
S0087	2
S0088	1
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

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?

ETS MODEL

Applying Ts Plot to our monthy produce data, I observed yearly seasonality, with peaks decreasing every year. Seasonality component can be used multiplicatively. Variance of errors is changing. I decided to apply error component multiplicatively.
Therefore my ETS Model is M-N-M



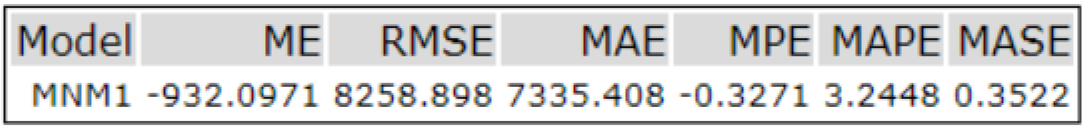
ARIMA MODEL

With the help of Ts Plot tool, I ended up with a stationary series after 1 seasonal differencing + 1 monthly differencing. Same with all 3 clusters.

With the use of ACF & PACF plots, I decided to use `Method: ARIMA(0,1,2)(1,1,0)[12]` for all clusters.

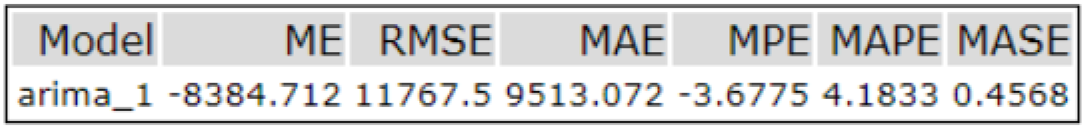
Using the TS Compare tool and the holdout set, below are the performances of both models.

ETS Model



Model	ME	RMSE	MAE	MPE	MAPE	MASE
MNM1	-932.0971	8258.898	7335.408	-0.3271	3.2448	0.3522

Arima Model

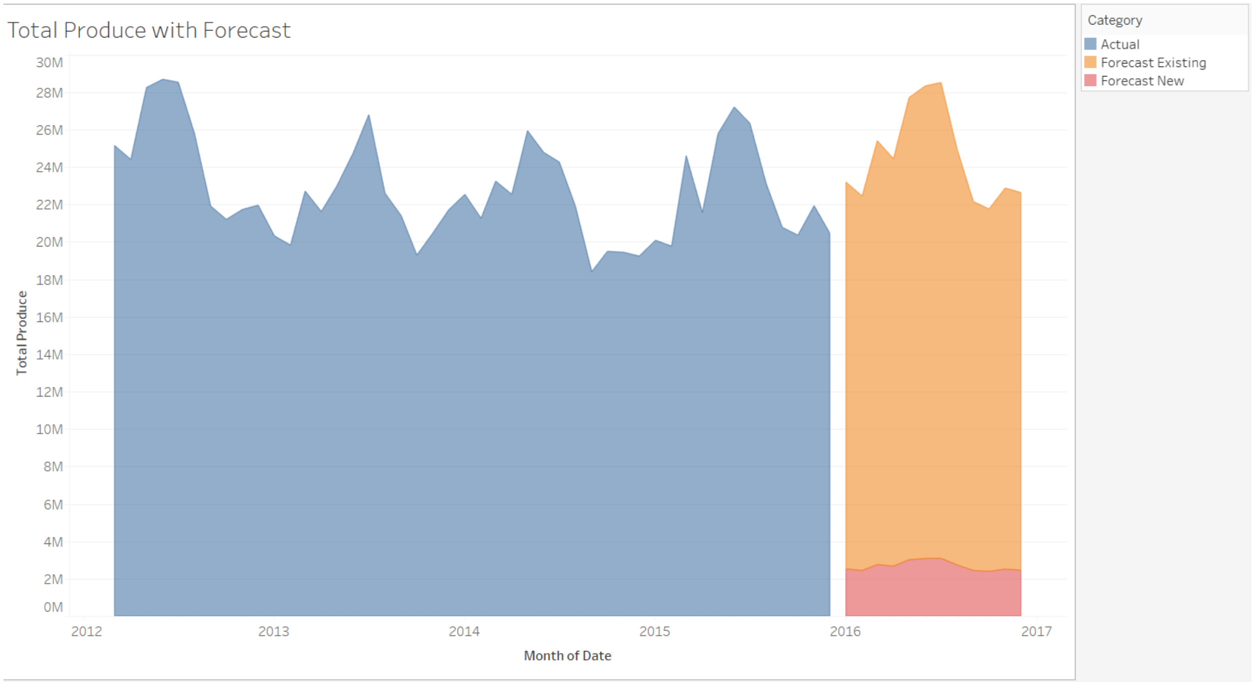


Model	ME	RMSE	MAE	MPE	MAPE	MASE
arima_1	-8384.712	11767.5	9513.072	-3.6775	4.1833	0.4568

Since RMSE is smaller with ETS Model, forecasted values are closer to predictions in this model. Besides, MASE is smaller compared to Arima model, which means ETS model is better at reducing error.

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.

Month	New Stores	Existing Stores
16-Jan	2,527,339	21,057,161
16-Feb	2,446,155	20,415,892
16-Mar	2,872,051	24,078,058
16-Apr	2,722,158	22,670,736
16-May	3,098,096	25,858,188
16-Jun	3,150,603	26,288,437
16-Jul	3,172,545	26,501,401
16-Aug	2,814,270	23,303,548
16-Sep	2,486,632	20,583,812
16-Oct	2,434,261	20,160,032
16-Nov	2,517,523	20,888,455
16-Dec	2,491,340	20,891,395


Total Produce with Forecast
Category
Actual
Forecast Existing
Forecast New
30M
28M
26M
24M
22M
20M
18M
16M
14M
12M
10M
8M
6M
4M
2M
0M
Total Produce
2012
2013
2014
2015
2016
2017
Month of Date