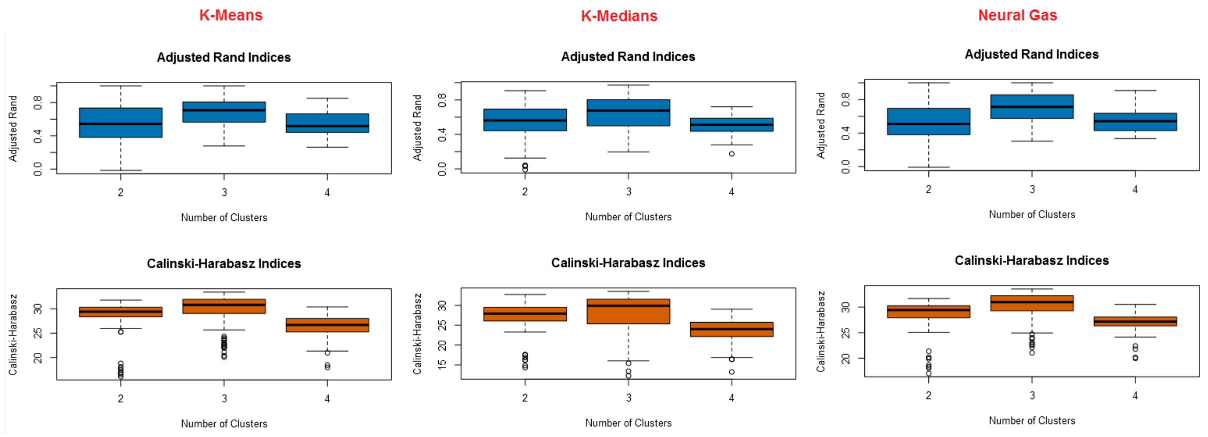
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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

   Optimal number of store formats is 3.

   I made K-Centroid Diagnosis by using 3 clustering methods. Here is the result.



Adjusted Rand Indices			
Feature	**K-Means**	**K-Medians**	**Neural Gas**
Median	0.71	0.68	0.71
Q1	0.57	0.51	0.58
Q3	0.81	0.80	0.85
IQR	0.24	0.51	0.27

Calinski-Harabasz Indices			
Feature	**K-Means**	**K-Medians**	**Neural Gas**
Median	30.80	29.91	30.98
Q1	29.14	25.41	29.28
Q3	31.93	31.55	32.25
IQR	2.79	6.14	2.97

   In adjusted Rand indices K-means and Neural Gas has high median values than that of K-Medians.
   And the spread of interquartile range is a little bit more spread in Neural Gas method.
   In Calinski-Harabasz Indices, K-media of Neural Gas is higher than that of K-Means and K-Medians.
   And the spread of interquartile range is narrowest in K-Means method.

   So, I decided to choose K-Means method for clustering store sales data since, it has fair median values with fair interquartile spread in both adjusted rand indices and Calinski-Harabasz Indices.

2. How many stores fall into each store format?

   Cluster 1 contains 23 stores.
   Cluster 2 contains 29 stores.
   Cluster 3 contains 33 stores.

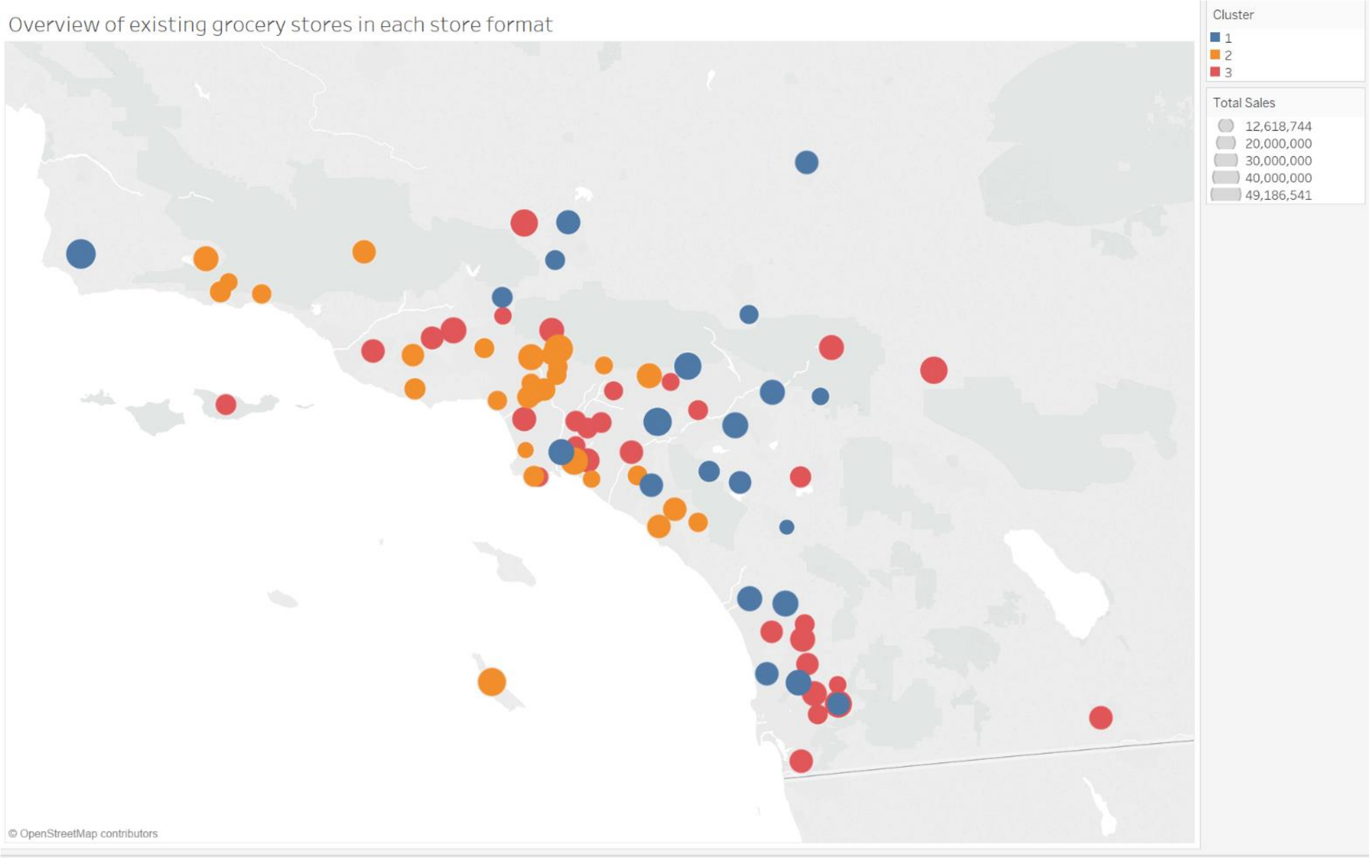
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

Cluster	Average Sales
1	32253842
2	27472964
3	28356955

I compared the average sales of stores in each clusters and found that stores in cluster 1 made more income than that of cluster 3 and 2 respectively.
(Average sales of Cluster 1 > Average sales of Cluster 3 > Average Sales of Cluster 2)

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.

Link to my tableau public


Overview of existing grocery stores in each store format
Cluster
1
2
3
Total Sales
12,618,744
20,000,000
30,000,000
40,000,000
49,186,541
© OpenStreetMap contributors

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 used Boosted Model to predict the best store format for the new stores.
   After training data using three different non-binary decision model, I compared the accuracy of all these three models.

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision Tree	0.7059	0.7327	0.6000	0.6667	0.8333
Forest Model	0.8235	0.8251	0.7500	0.8000	0.8750
Boosted Model	0.8235	0.8235	0.8000	0.6667	1.0000

   Forest Model and Boosted Model has high accuracy than Decision Tree. Overall accuracy of Boosted Model and Forest Model is the same, so I checked the accuracy of predicting target variables. Boosted model can predict 100% correct for cluster 3 while it can predict only 67% for cluster 2. Boosted Model has higher F1 value than Forest Model. So, boosted model is the best fit to predict store format.

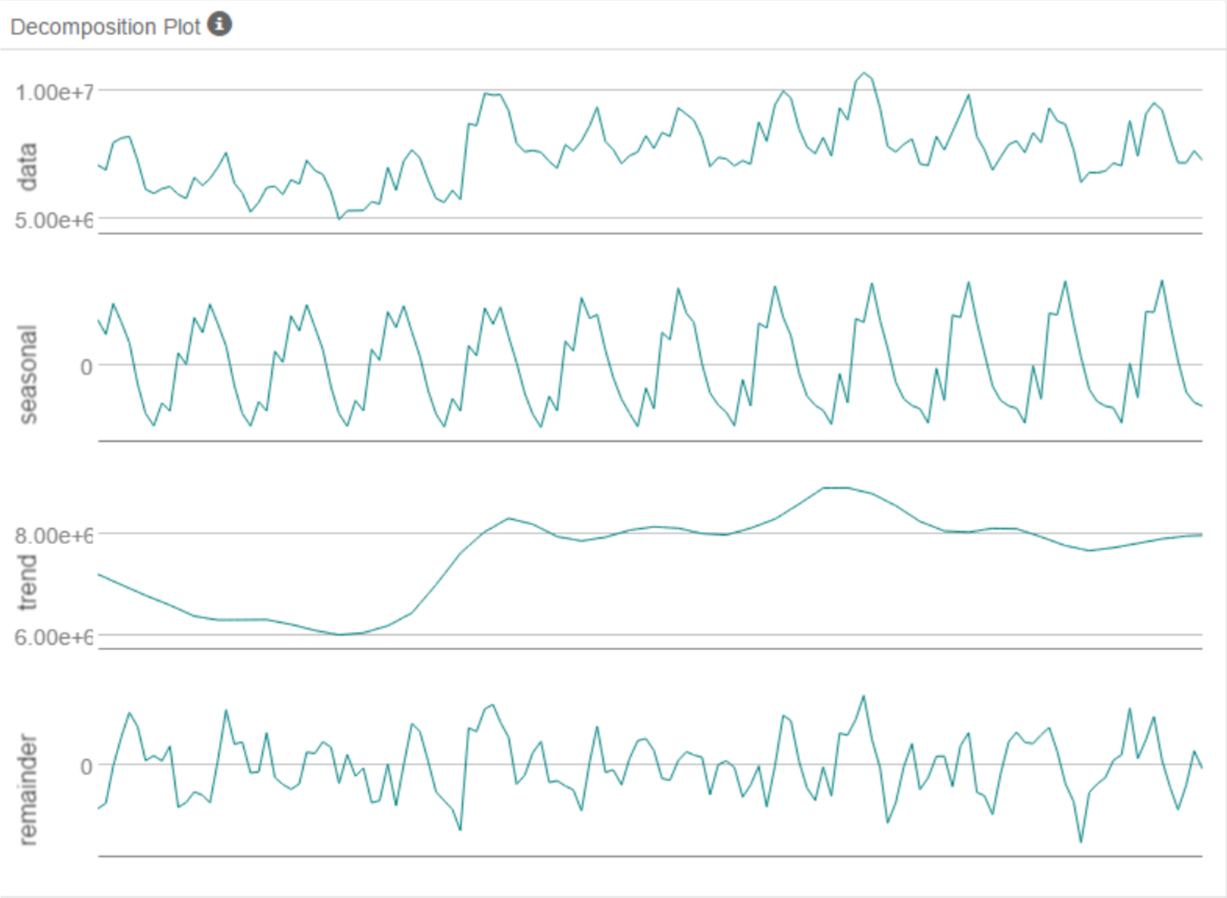
2. What format do each of the 10 new stores fall into? Please fill in the table below.

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?

I made two prediction models, ETS and ARIMA, to predict ‘produce sales’ in each store format and compare the model accuracy measures and found that ETS (M, N, M) model is the best fit to predict future sales.


Decomposition Plot

Format 1
- ETS (M, N, M)
- ARIMA (1, 0, 0) (1, 1, 0) [12]

Format 2
- ETS (M, N, M)
- ARIMA (1, 0, 0) (1, 1, 0) [12]

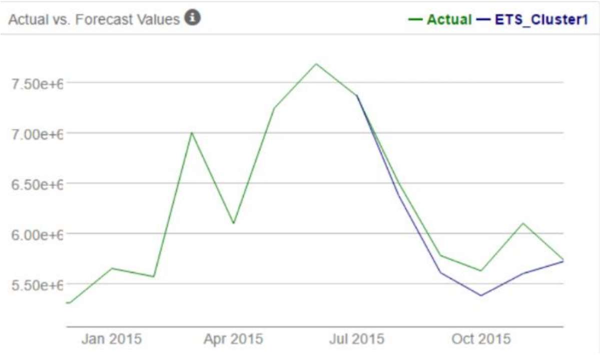
Format 3
- ETS (M, N, M)
- ARIMA (1, 0, 0) (1, 1, 0) [12]

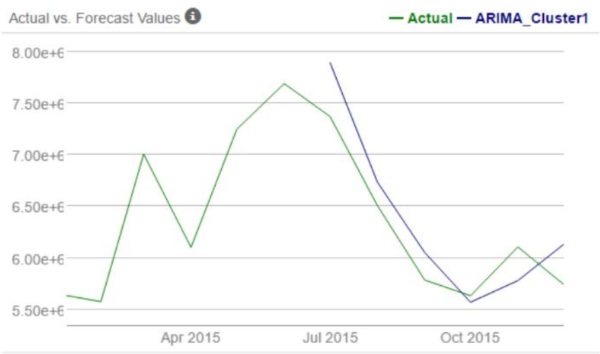
In-sample error measures

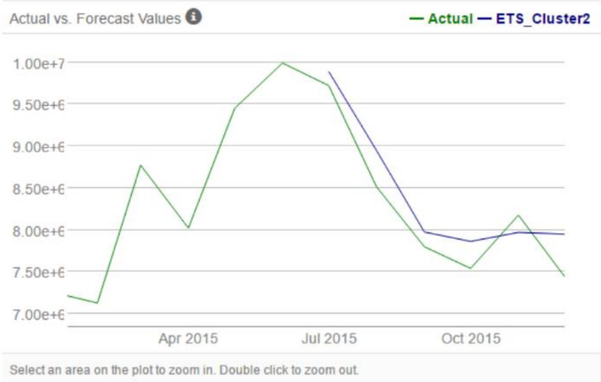
Measures	Format 1		Format 2		Format 3	
	ETS	ARIMA	ETS	ARIMA	ETS	ARIMA
ME	174375.30	-169533.90	-232953.00	-226003.80	68198.30	-198642.20
RMSE	243265.90	331113.30	328700.70	353315.20	220683.30	368046.00
MAE	177878.20	299192.70	301499.10	341233.70	185714.40	33340.40
MPE	2.92	-2.59	-2.93	-2.71	0.98	-2.41
MAPE	2.97	4.74	3.77	4.12	2.45	4.20
MASE	0.36	0.61	0.57	0.64	0.27	0.48

By comparing in-sample errors, we can see that mean errors (ME) and mean percent error (MPE) of ARIMA is smaller than that of ETS in Format1 and Format 2, but in Format 3, ME of ETS is much smaller than ARIMA. ETS models has smaller RMSE and MAPE values than that of ARIMA models.

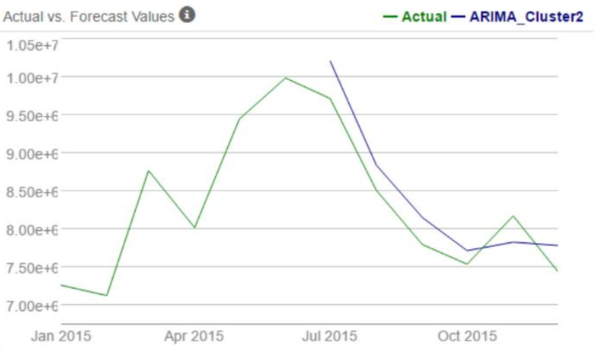
MAPE value of ETH model is much smaller than that of ARIMA model in all three store formats. The MASE values of both models in all store formats are smaller than 1, that means that we can use these models to predict future.

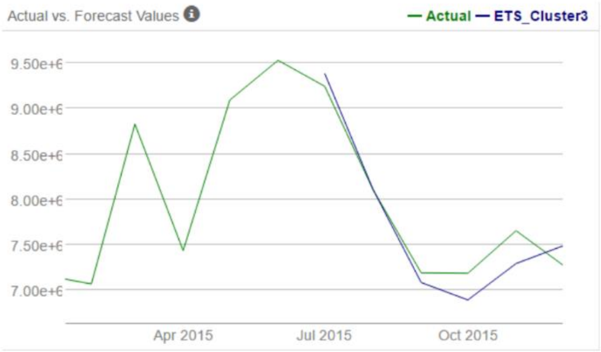


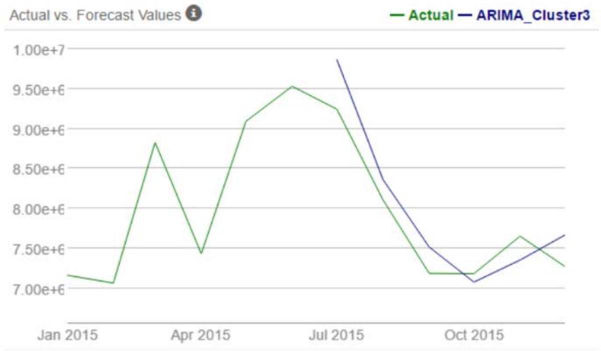




Select an area on the plot to zoom in. Double click to zoom out.







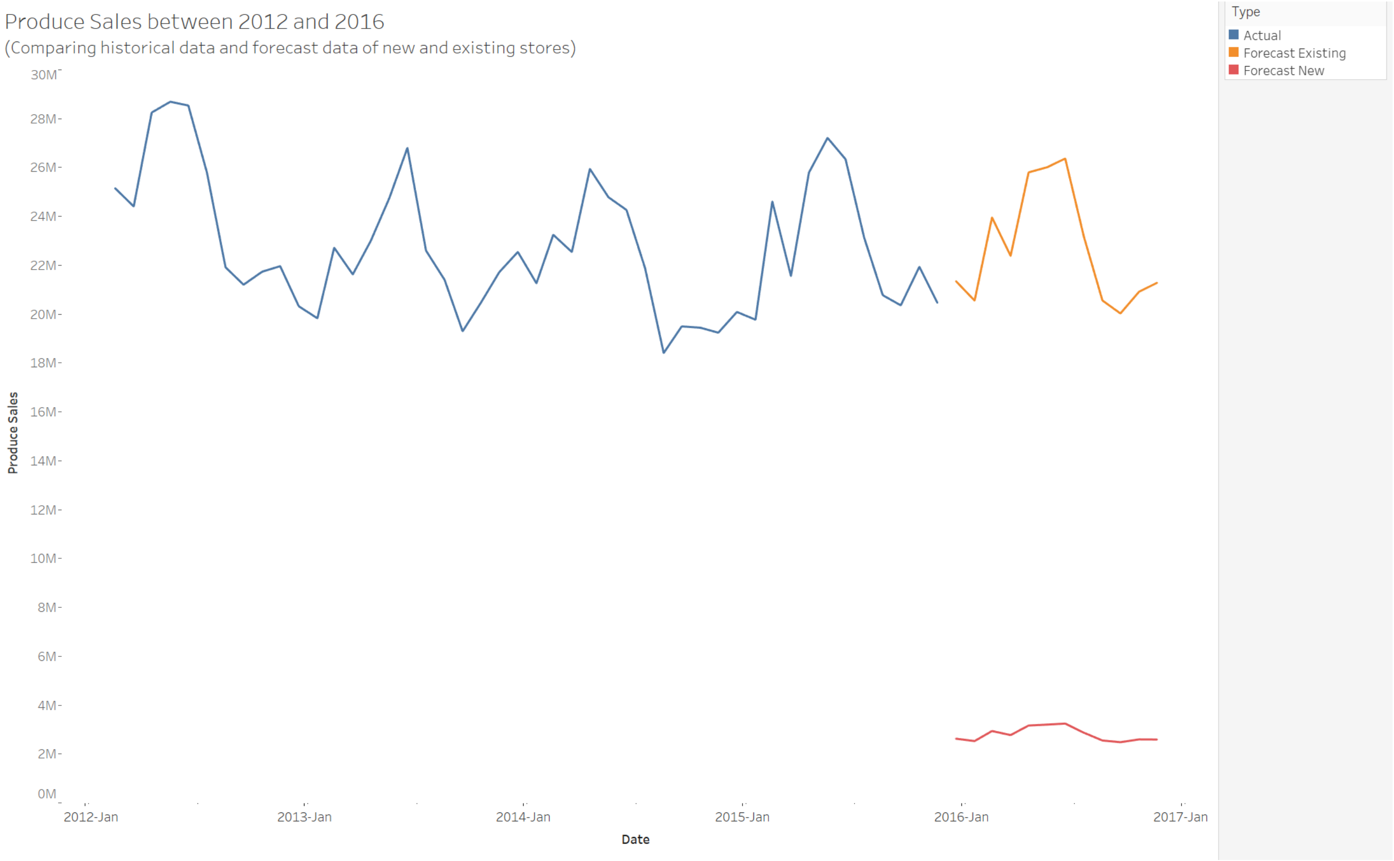
By comparing forecast quality on holdout sample, we can see that ETS model is a better choice.

2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".

Data tables of forecasted 'Produce Sales' for the year 2016

Month	New Stores	Existing Stores
Jan-16	2,626,198.009	21,347,066.79
Feb-16	2,529,185.842	20,553,550.63
Mar-16	2,940,263.647	23,955,683.51
Apr-16	2,774,134.767	22,383,629.66
May-16	3,165,320.367	25,800,158.07
Jun-16	3,203,285.506	26,011,176.59
Jul-16	3,244,463.964	26,367,263.12
Aug-16	2,871,487.558	23,175,507.30
Sep-16	2,552,417.752	20,557,092.95
Oct-16	2,482,836.849	20,028,039.70
Nov-16	2,597,779.886	20,915,736.96
Dec-16	2,591,814.624	21,281,965.96

Link to my tableau public



Produce Sales between 2012 and 2016

(Comparing historical data and forecast data of new and existing stores)