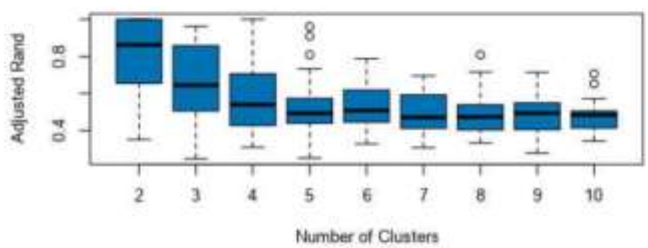
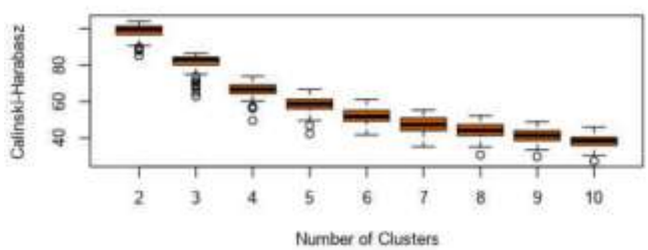
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

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





Adjusted Rand Indices:

	2	3	4
Minimum	0.352178	0.24812	0.309898
1st Quartile	0.664805	0.506795	0.428687
Median	0.862177	0.644809	0.53975
Mean	0.805606	0.6644	0.572697
3rd Quartile	0.988235	0.85824	0.702268
Maximum	1	0.962601	1

Calinski-Harabasz Indices:

	2	3	4
Minimum	85.07799	62.72109	49.56539
1st Quartile	96.78399	80.21352	64.42923
Median	99.54629	83.06892	66.63066
Mean	98.56119	81.26833	66.33271
3rd Quartile	101.29084	84.25545	69.04128
Maximum	103.99264	86.45017	73.94085

Three is the most optimal number of store formats with high median and relative minimized spread of Adjusted Rand Indices and Calinski-Harabasz Indices.

2. How many stores fall into each store format?
The number of stores in each store format (cluster):

```
stepFlexclust(scale(model.matrix(~-1 + Per_Dry_Grocery + Per_Frozen_Food + Per_Meat + Per_Produce
+ Per_Floral + Per_Deli + Per_Bakery + Per_General_Merchandise + Per_Dairy, the.data)), k = 3, nrep =
10, FUN = kcca, family = kccaFamily("kmeans"))
```

Cluster Information:

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

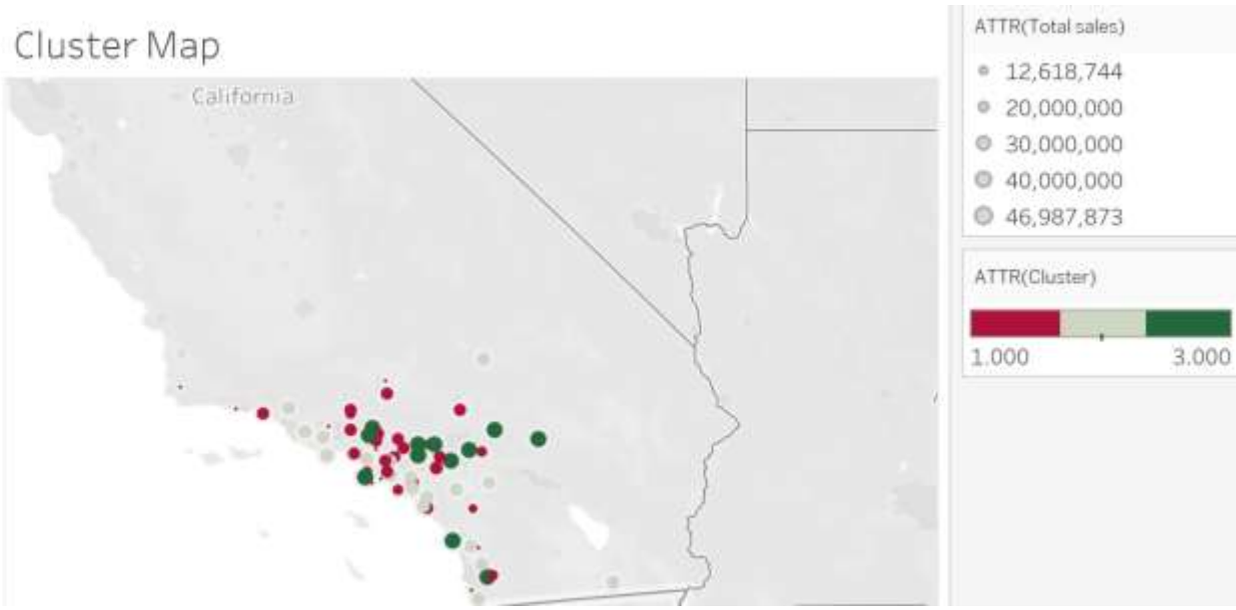
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

	Per_Dry_Grocery	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_Deli	Per_Bakery
1	0.528249	-0.261597	0.614147	-0.655028	-0.663872	0.824834	0.428226
2	-0.594802	0.435129	-0.384631	0.812883	0.71741	-0.46168	0.312878
3	0.304474	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482	-0.866255

	Per_General_Merchandise	Per_Dairy
1	-0.674769	-0.215879
2	-0.329045	0.655893
3	1.135432	-0.702372

Based on these results, it can be seen that:
- Cluster 1 is oriented to Dry Grocery, Meat, Deli, and Bakery, Cluster 2 is oriented to Frozen Food, Produce & Floral, and Dairy, while Cluster 3 is oriented toward General Merchandise.
- Cluster 1 has the lowest sales on Produce, Floral, and General Merchandise. Cluster 2 has the lowest sales on Dry Grocery, Meat, and Deli. Cluster 3 has the lowest sales on Frozen Food, Dairy and Bakery.
- The average distance is highest for Cluster 2. There is a difference in their max distance value with Cluster 1 and Cluster 3 have the highest and the lowest values, respectively. Cluster 1 has the highest separation within itself.

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.
I used the color palette of red-green. Red, light green, and dark green denote for cluster 1, cluster 2, and cluster 3, respectively.



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.)

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree	0.2353	0.2286	0.2857	0.2000	0.2000
Boosted_Tree	0.2353	0.2476	0.1429	0.2000	0.4000
Random_Forest	0.3529	0.3810	0.1429	0.6000	0.4000

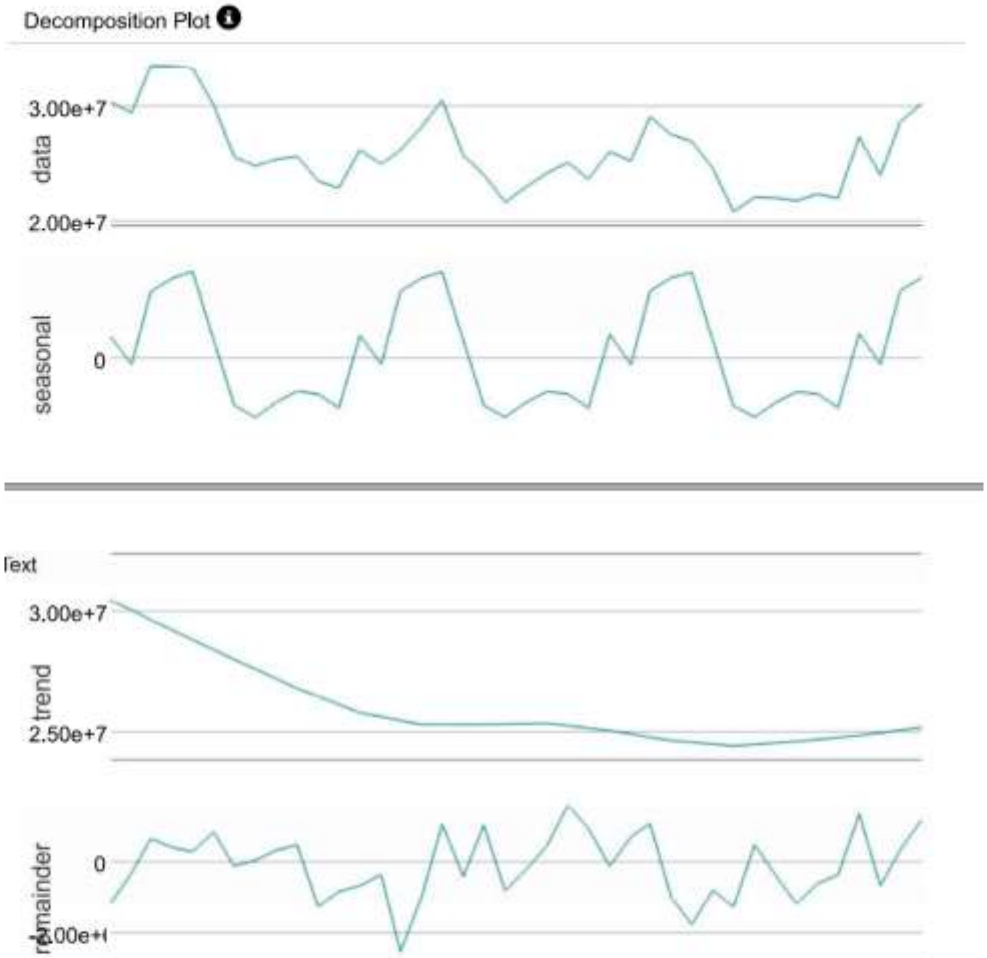
Random Forest has the highest Accuracy and F1 score among the three methods. Therefore, I used Random Forest to predict the best store format.

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?

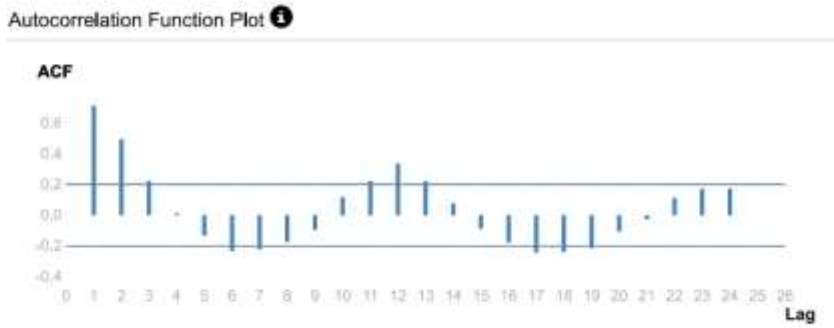


The decomposition plot shows the repeated patterns by year with no change in magnitude for seasonal and error portion. The trend line is relatively gradually downward. That suggests ETS model with additive method in three components and ARIMA model with seasonal differencing.

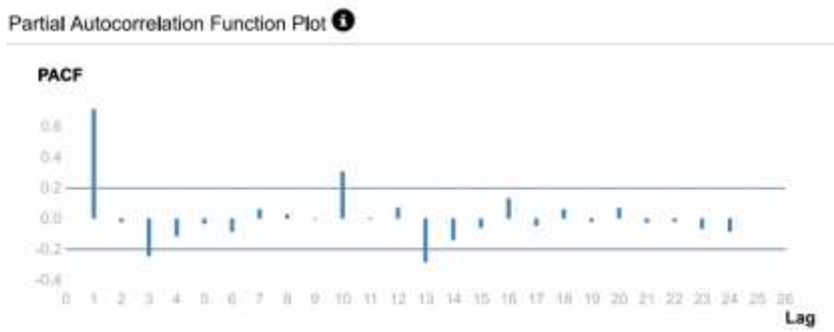
Configuring ETS model: ETS(a,a,a)

Configuring ARIMA model: Since there are seasonal components found in the time series I will use an ARIMA(p, d, q)(P, D,Q)S model for forecasting. The time series data was explored further.

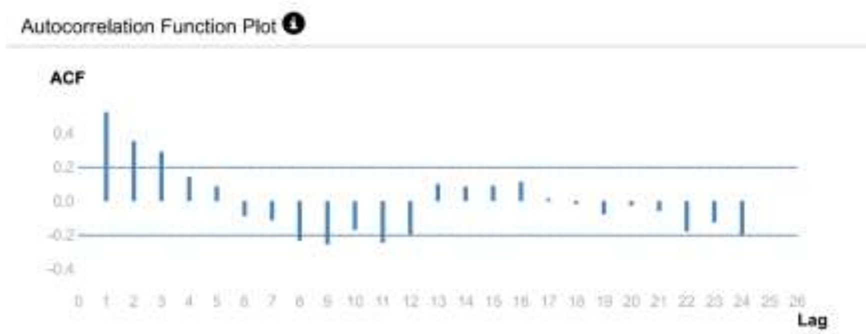
- The initial ACF and PACF plots confirm the high seasonal effect. Therefore, I will difference seasonal data.



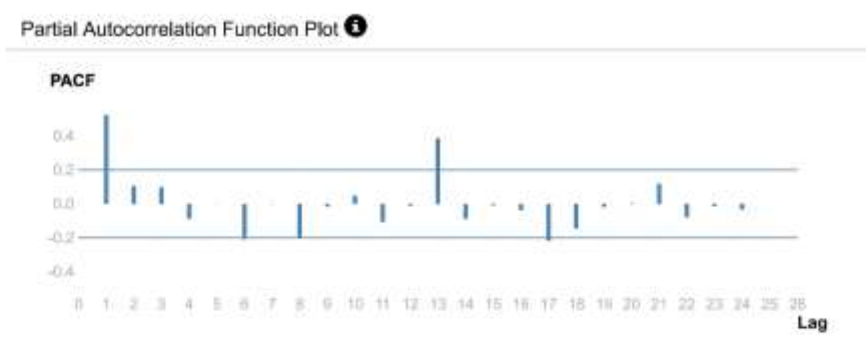
This is an autocorrelation plot



- Seasonal difference ACF and PACF: the data is less correlated but need to be differenced further.

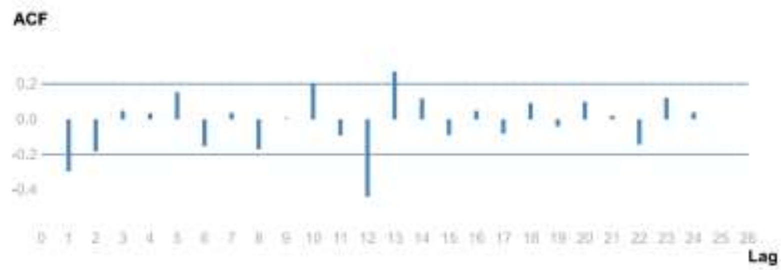


This is an autocorrelation plot



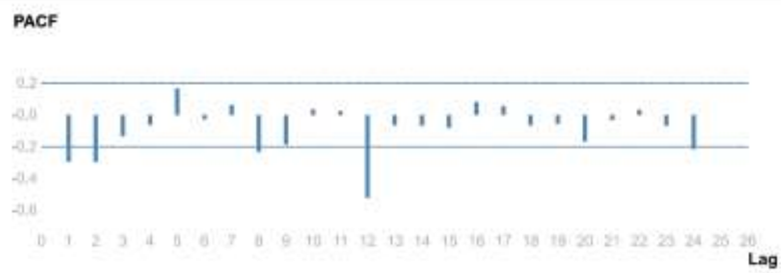
- Seasonal first difference ACF and PACF: almost the significant lags were removed so there is no need to difference further. The differencing terms will be d(1) and D(1). The ACF plot shows negative correlation at lag 1 and lag 12, a seasonal lag. Therefore, moving average terms are q(1) and Q(1).

Autocorrelation Function Plot



This is an autocorrelation plot

Partial Autocorrelation Function Plot



Therefore, the model terms for ARIMA are: ARIMA(0,1,1)(0,1,1)[12] with 12 periods (months) in each season (year).

Error terms: the ACF and PACF plots resulted from the ARIMA(0,1,1)(0,1,1)[12] model show nearly no significant lags suggesting no need for adding additional AR or MA term.

Autocorrelation Function

Partial Autocorrelation Function

When looking at the performance of ETS and ARIMA model on the holdout sample:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	-269611.8	710319.3	621599.5	-1.06	2.5013	0.3262
ARIMA	-307792.6	787349.7	675481.1	-1.1873	2.6929	0.3544

ETS model has better predictive qualities in almost the metrics. Therefore, I chose ETS for forecasting the sales.

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.

Month	**New Stores**	**Existing Stores**
Jan-16	2349445.27	20716705.28
Feb-16	2267446.9	20012792.5
Mar-16	2582867.99	22813230.66
Apr-16	2451062.93	21650320.2
May-16	2782454.1	24536431.62
Jun-16	2839761.08	25068890.25
Jul-16	2855984.67	25210750.27
Aug-16	2497888.49	22057092.53
Sep-16	2188291.87	19310514.74
Oct-16	2117890.16	18723695.71
Nov-16	2220559.57	19643855.06
Dec-16	2223041.09	19669065.38


Sales of Existing, Forecast Existing, and Forecast New stores
Date
2015
2016
Sum Total Sales
20M
10M
0M
March
April
May
June
July
August
September
October
November
December
January
February
March
April
May
June
July
August
September
October
November
December
Type
Existing
Forecast Existing
Forecast New