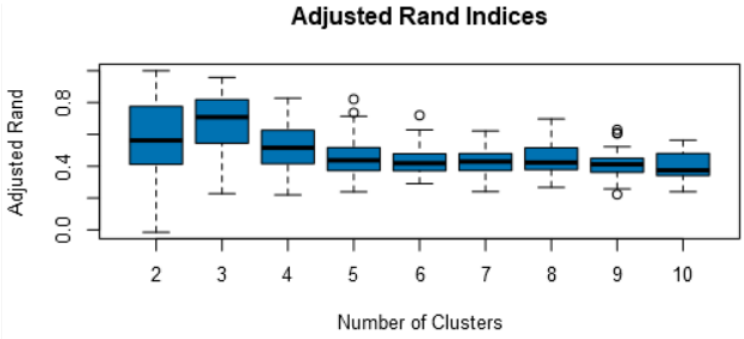
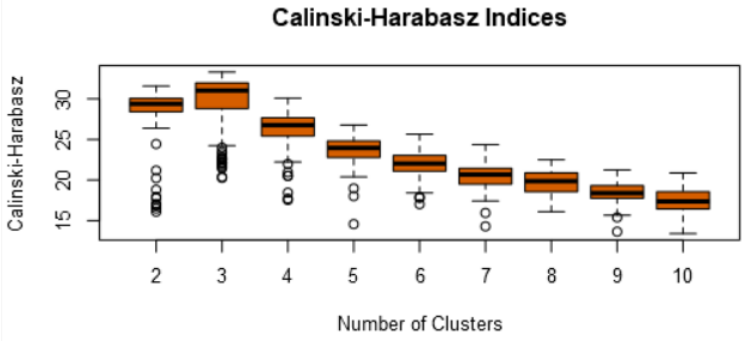
Task 1: Determine Store Formats for Existing Stores

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

   I have determined the optimal number of store formats is 3. I used K-Centroids Diagnostics tool with Clustering method = K-Means, Standardize fields = z-score, Minimum number of clusters = 2, and Maximum number of clusters = 10.





   By looking at AR and CH indices, I picked 3 clusters because it has the highest median of index.

2. How many stores fall into each store format?

Record #	Cluster	Count
1	1	23
2	2	29
3	3	33

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

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

The cluster 1 is bit smaller in size compare to cluster 2 and 3. The cluster 3 is the most compact cluster because it has the smallest average distance. The max distance shows how far out the farthest from the centroid. The cluster 3 has the farthest outlier. The cluster 2 has the largest Separation value. There is more separation between the cluster 2 and all of the other clusters. Since there is no huge difference between values among 3 clusters for Size, Average Distance, and Separation, I would say that 3 is optimal number of clusters.

	Perc_Dry_Grocery	Perc_Dairy	Perc_Frozen_Food	Perc_Meat	Perc_Produce	Perc_Floral	Perc_Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	Perc_Bakery	Perc_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

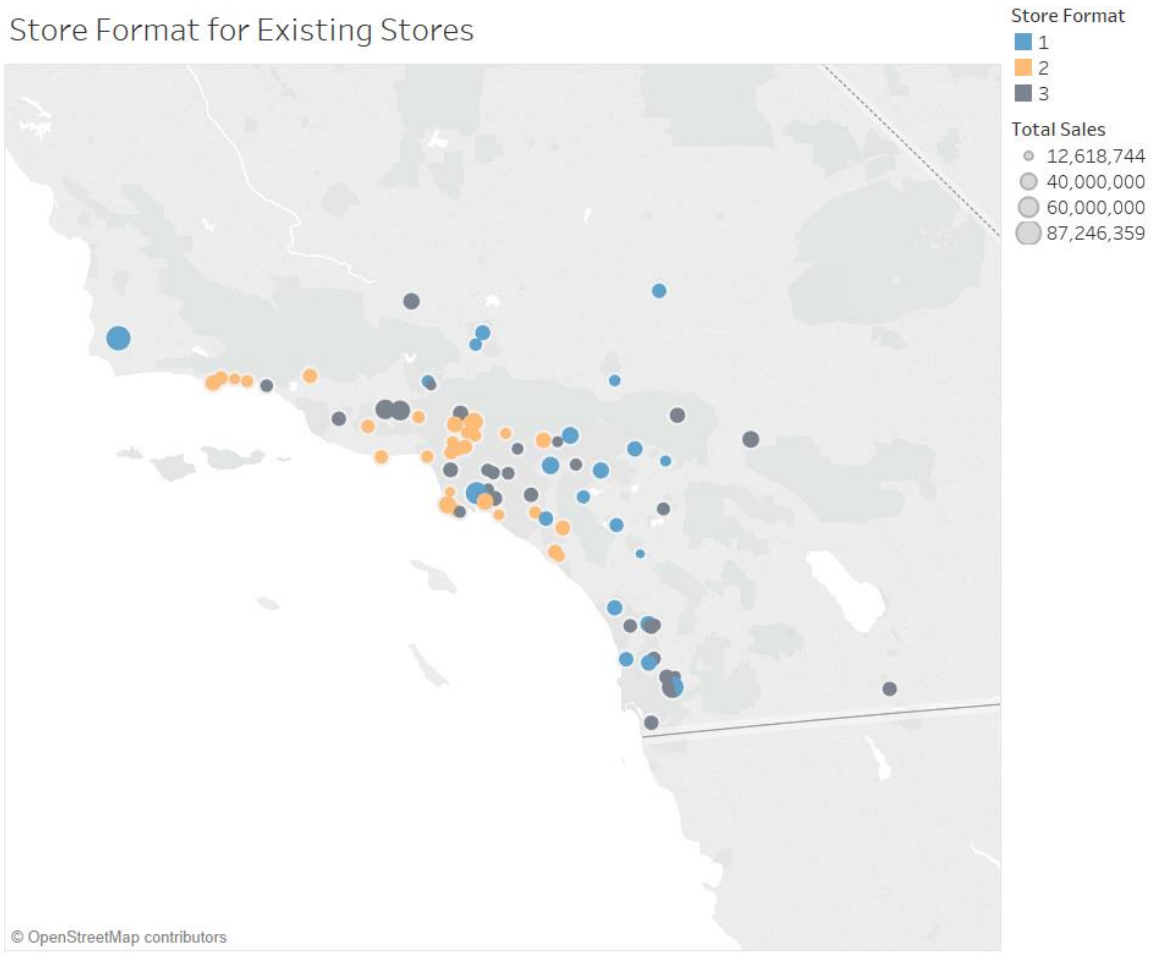
Clusters have the largest positive values and the largest negative values are related in a category.

For example. The cluster 1 and the cluster 2 are extreme clusters related to Frozen Food and Bakery. Stores in one cluster potentially sells a lot of Frozen Food and Bakery and stores in the other clusters with small amount of Frozen Food and Bakery sales.

The same can be said true to the following categories and clusters.

Frozen Food, Bakery, and General Merchandise – Cluster 1 & 2
Produce Floral, Dry Grocery, Dairy, Meat, and Deli – Cluster 2 & 3

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.



https://public.tableau.com/profile/ayumi.ohashi#!/vizhome/StoreFormat/StoreFormats

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.)

   What we need to solve here is to determine a cluster for the new stores. Since the cluster is considered categorical and there will be 3 clusters (non-binary), I need to use non-binary classification models. Possible choices of non-binary classification models are Decision Tree Model, Forest Model, and Boosted Model.

   Here is the result of validating my model against the validation set.

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
FM_NewStores	0.8235	0.8251	0.7500	0.8000	0.8750
BM_NewStores	0.8235	0.8543	0.8000	0.6667	1.0000
DT_NewStores	0.7059	0.7327	0.6000	0.6667	0.8333

   Both Forest Model and Boosted Model have the same accuracy as 0.8235 and higher

than the Decision Tree Model. But the Boosted Model has slightly higher F1-Score than the Forest Model. Thus, the Boosted Model is the best model for determining a store format for 10 new stores.

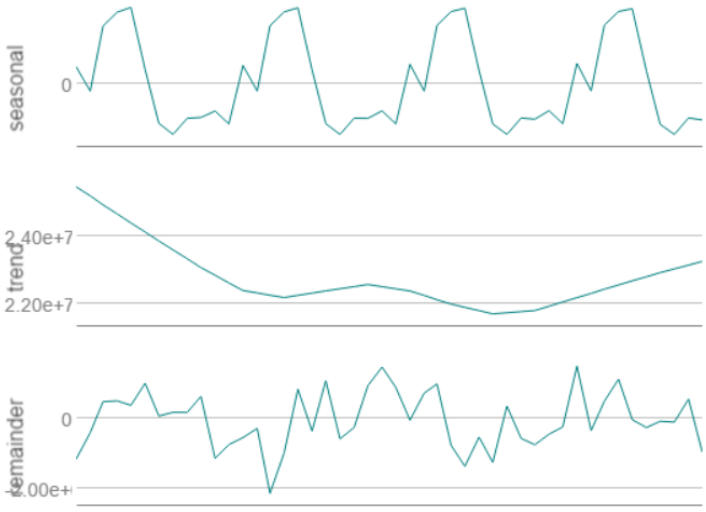
2. What format do each of the 10 new stores fall into? Please fill in the table below.

Store Number	Segment
S0086	3
S0087	2
S0088	3
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?

ETS Model : ETS(M,N,M)



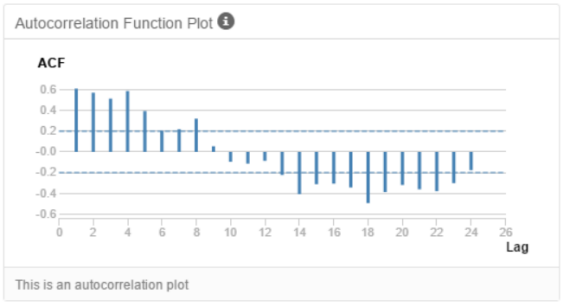
The error is shrinking over time. => (**M**)ultiplicatively

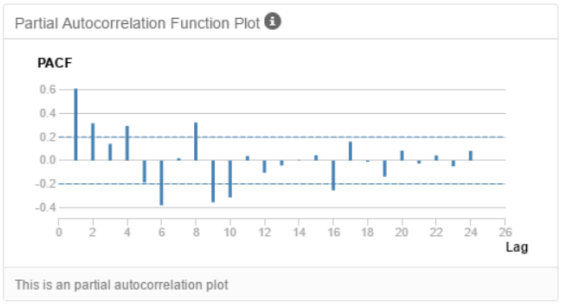
The trend line changes direction towards the end of the period and goes back up. This line as a none trend line. => (**N**)one

It contains seasonality. It is decreasing over time => (**M**)ultiplicatively

ARIMA Model : ARIMA(0,1,1)(0,1,1)12

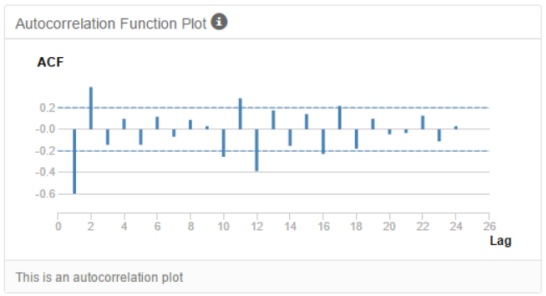
Looking at ACF for Time Series plot, seasonal decreases can be observed at 12 and 24 lags. The seasonal differencing and seasonal first differencing were taken to make the data set stationary.

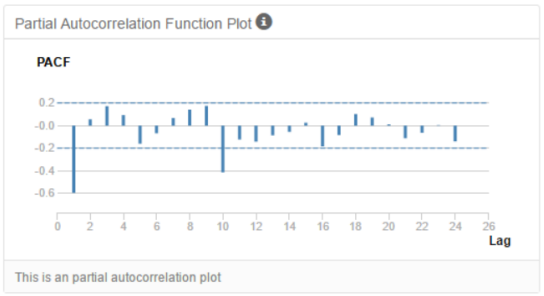




By observing at Seasonal First Difference plots, I came up with

d = 1 & D = 1 : Seasonal difference and Seasonal First Difference were taken.
q = 1 : Negative autocorrelation at lag 1 in ACF and PACF.
Q = 1 : Significant autocorrelation at lag 12 and 24 in ACF.
m = 12 : The data set is monthly data.





Here are the forecast error measurements against the holdout samples.
ETS:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
MNM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

ARIMA:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
MNM	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822	NA

The ETS model has smaller RMSE and MASE. This means the error is smaller, and the ETS model is better model than the ARIMA model. The actual vs forecast value also

looks better for ETS. I chose ETS(M,N,A) to do the forecasting.

ETS:

Actual	MNM
26338477.15	26907095.61191
23130626.6	22916903.07434
20774415.93	20342618.32222
20359980.58	19883092.31778
21936906.81	20479210.4317
20462899.3	21211420.14022

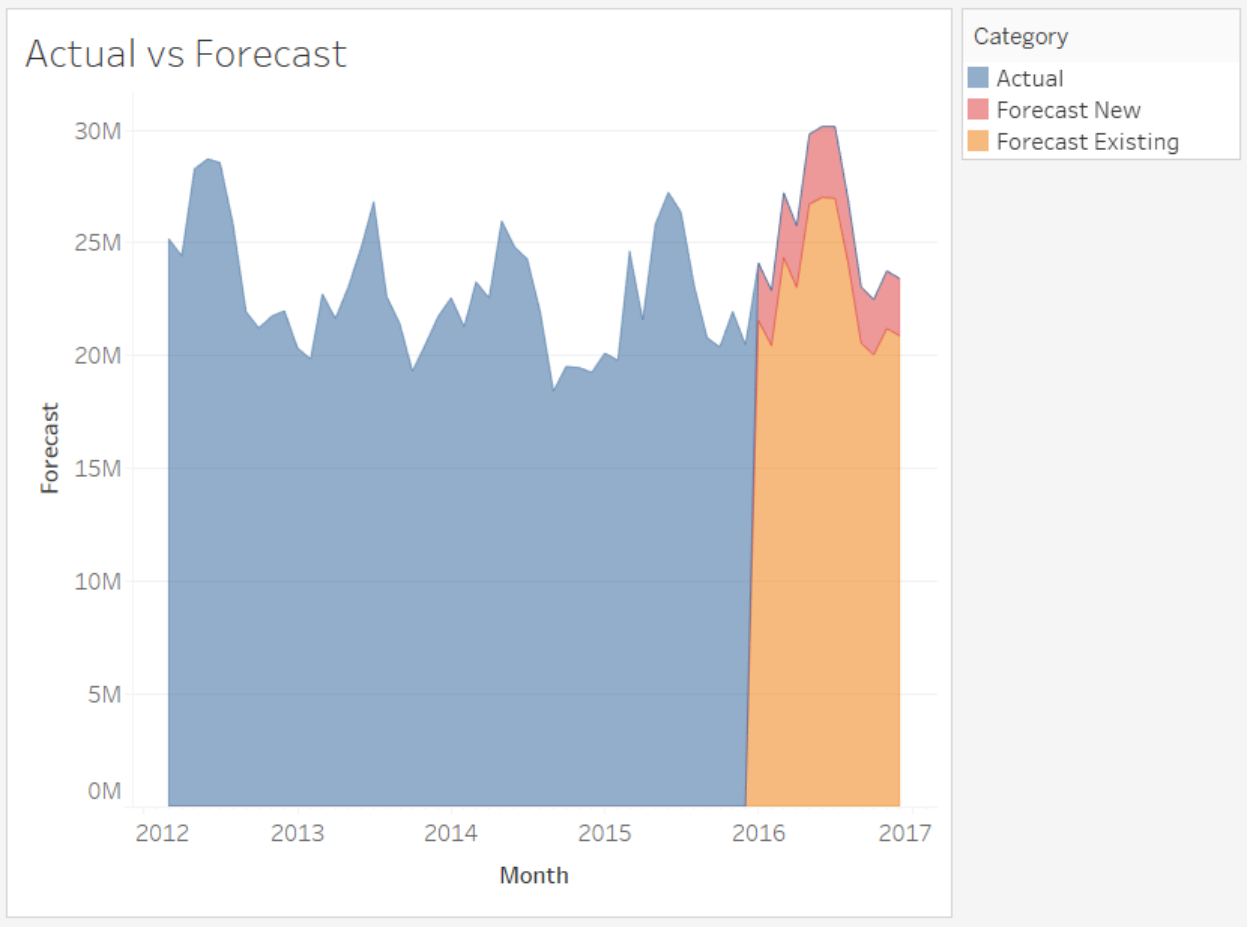
ARIMA:

Actual and Forecast Values:

Actual	ARIMA_MA
26338477.15	27182961.16627
23130626.6	24073582.27177
20774415.93	21223756.4441
20359980.58	20648299.23319
21936906.81	21205988.81004
20462899.3	21622151.40814

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	2,557,108.48	21,539,936.01
Feb-16	2,456,921.75	20,413,770.60
Mar-16	2,880,690.97	24,325,953.10
Apr-16	2,737,423.76	22,993,466.35
May-16	3,111,087.09	26,691,951.42
Jun-16	3,154,942.45	26,989,964.01
Jul-16	3,178,114.92	26,948,630.76
Aug-16	2,837,587.43	24,091,579.35
Sep-16	2,505,155.30	20,523,492.41
Oct-16	2,459,401.49	20,011,748.67
Nov-16	2,560,806.53	21,177,435.49
Dec-16	2,545,760.49	20,855,799.11



https://public.tableau.com/profile/ayumi.ohashi#!/vizhome/Forecast_153/Dashboard1