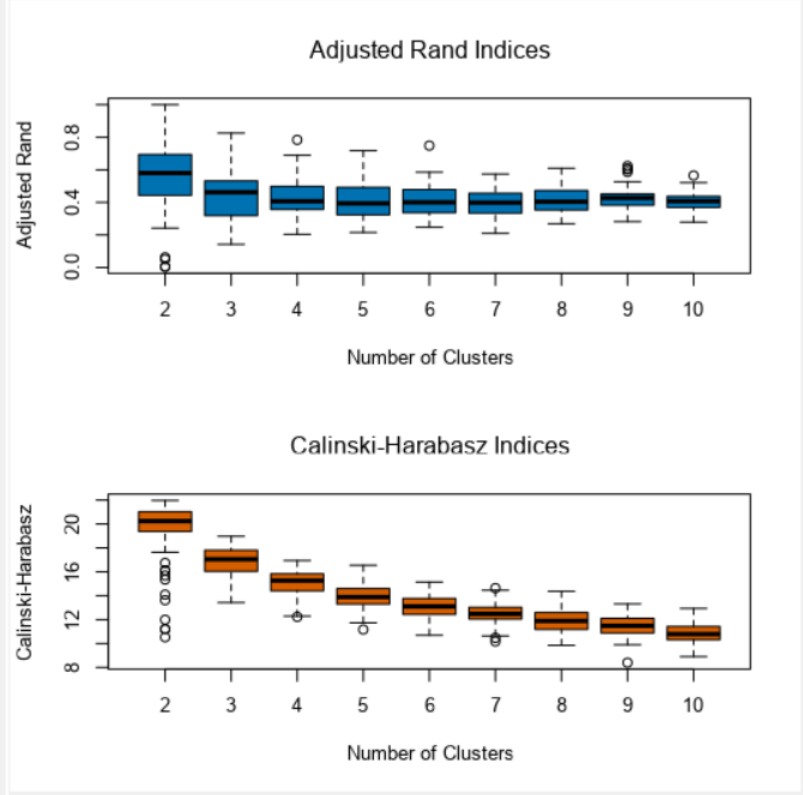
Project: Predictive Analytics Capstone

Task 1: Determine Store Formats for Existing Stores

1. What is the optimal number of store formats? How did you arrive at that number?
    - It was decided by using Neural Gas clustering method.
    - As for 2 formats the median for Adjusted Rand indices is 0.57 with IQR of 0.25.
    - As for 2 formats the median for Calinski Harabasz Indices is 20.24 with IQR of 1.6.
    - The nearest neighbor was 3 formats which had a median of 0.46 and 17.04. IQR of 0.21 and 1.8.
    - The optimal number of store formats would be 3.
    - 

    - As the number of outliers are more in 2 formats. So, we would choose 3 format as optimal number.

2. How many stores fall into each store format?

Formats	Size
1	25
2	35
3	25

Please refer to this link as I have done all the steps that were asked but still, I got these

results. As this link suggests it is because of version of Alteryx. https://knowledge.udacity.com/questions/246950.

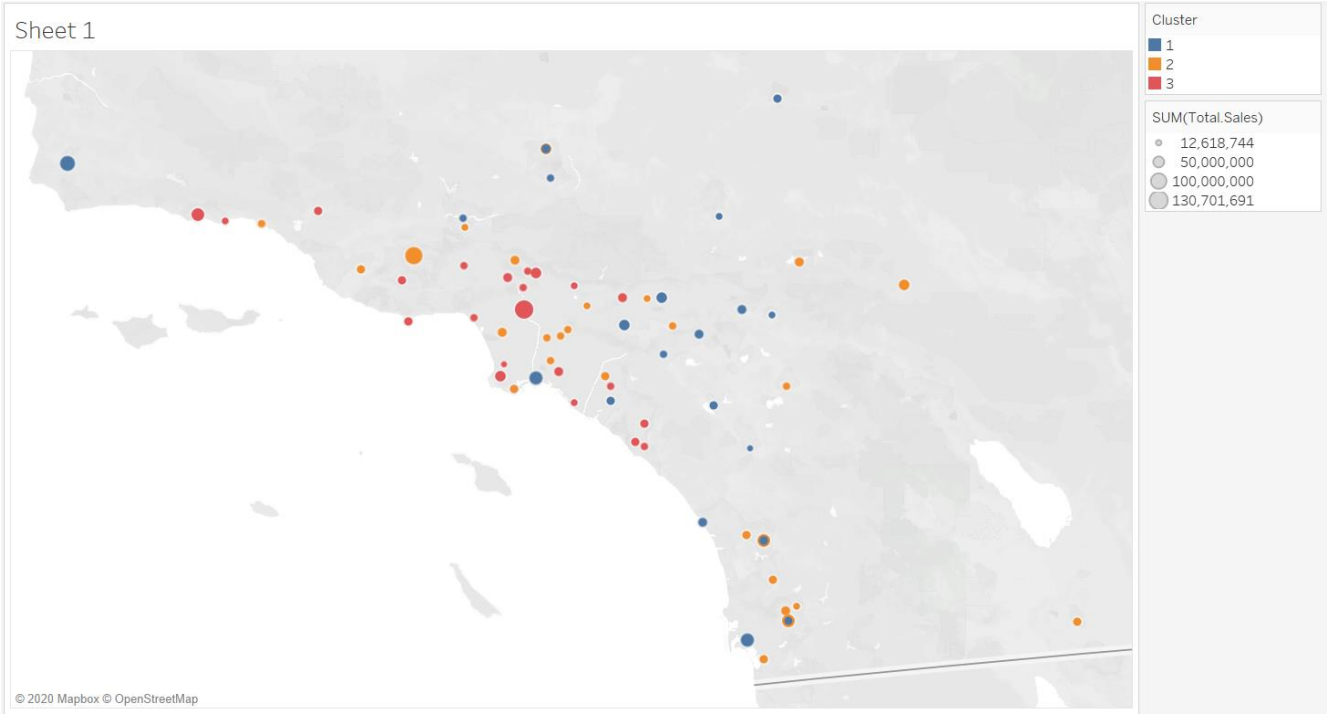
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

	Sum_Dry_Grocery	Sum_Dairy	Sum_Frozen_Food	Sum_Meat	Sum_Produce	Sum_Floral	Sum_Deli
1	0.528249	-0.215879	-0.261597	0.614147	-0.655028	-0.663872	0.824834
2	-0.594802	0.655893	0.435129	-0.384631	0.812883	0.71741	-0.46168
3	0.304474	-0.702372	-0.347583	-0.075664	-0.483009	-0.340502	-0.178482

	Sum_Bakery	Sum_General_Merchandise
1	0.428226	-0.674769
2	0.312878	-0.329045
3	-0.866255	1.135432

- Here the opposite signs in the clusters indicate that in that category the clusters with opposite signs are different or reflect opposite behavior.

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.



Task 2: Formats for New Stores

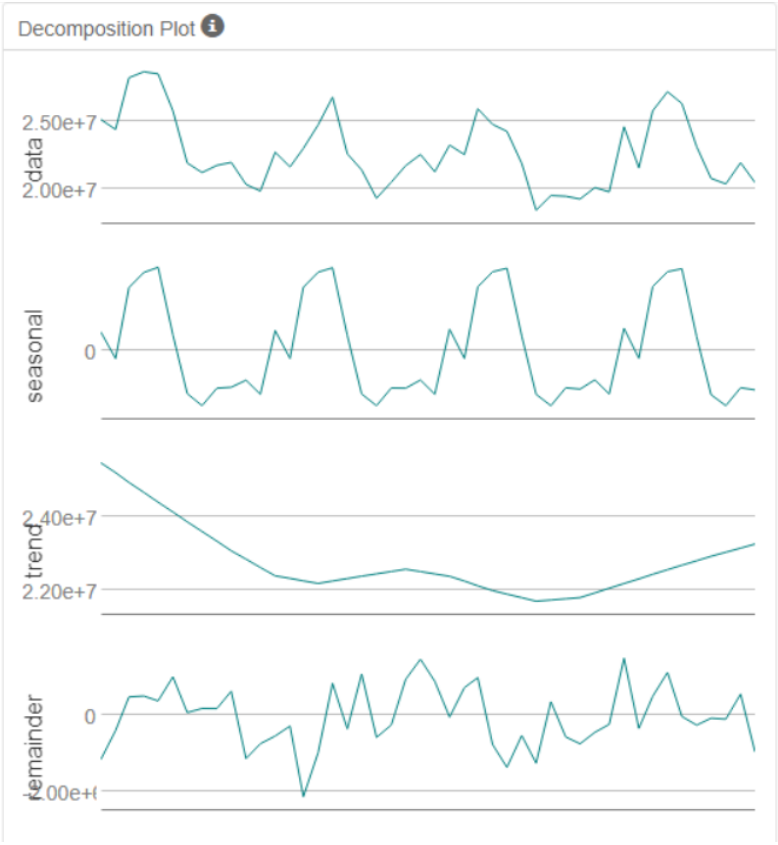
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)
    - The methodology used to predict the best store format for the new store is Random Forest Model.
    - Accuracy for all the models:
        - Random Forest → 82.35%
        - Decision Tree → 70.59%
        - Boosted Model → 58.82%
    - As we can clearly see accuracy on validation data is highest for random forest. So, we will choose this model for prediction of store format for 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	1
S0088	3
S0089	1
S0090	1
S0091	2
S0092	1
S0093	2
S0094	1
S0095	1

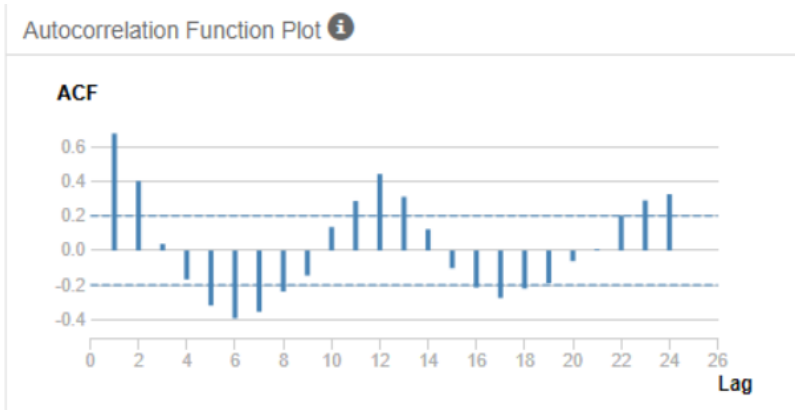
Task 3: Predicting Produce Sales

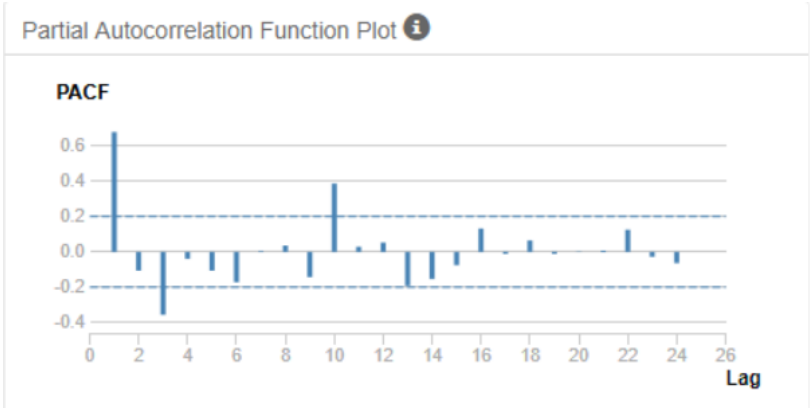
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

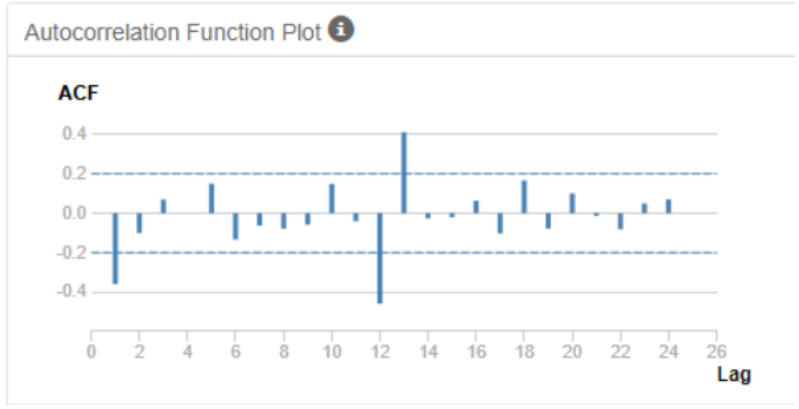
- ETS Model used is ETS (M, N, M)
  - For ETS model the time series plot looks like:
  - 

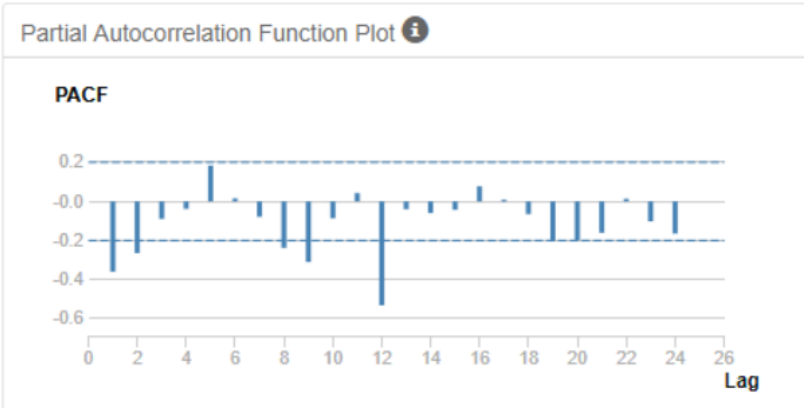
  - From the plot above we can clearly see that:
  - Trend → The is no trend for the data as the line first decreases then increases. So, no trend in data
  - Seasonality → There is seasonality in the data, but the magnitude does increase so we will use multiplicative for seasonal component.
  - Error → Error is increasing it does not look constant so we will use multiplicative for it.

- ARIMA Model
  - The model I choose is ARIMA (0,1,1) x (0,1,1) [12]
  - The ACF and PACF plots for data are:

- 

- 

- As we can see data is not stationary, there is a clear pattern in ACF plot we will do differencing to make data stationary.
- Plots after seasonal and non-seasonal differencing
- 



- Now we can clearly see that the data is stationary as there is no pattern in ACF plots. and using these plots we will choose the model components.
- As we have done seasonal and non-seasonal differencing, we will keep d and D as 1
- There is negative correlation at lag1, we will keep q and Q as 1.

Now we will compare between ETS and ARIMA model using AIC score and whichever score is lower we will use that model in forecasting.
The model we choose will be ARIMA (0,1,1) x (0,1,1) [12].

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.

Table of Forecasts:

Time Period Year	Existing Stores Forecast	New Stores Forecast
Jan 2016	23473777.97	2886242.48
Feb 2016	22076027.22	2739993.57
Mar 2016	25288949.18	3148415.55
Apr 2016	25903108.36	3113252.43
May 2016	26027868.52	3045256.75
Jun 2016	22893869.47	2760204.63
July 2016	20167444.82	2414140.31
Aug 2016	19632240.18	2455689.32
Sep 2016	20444735.61	2520270.80
Oct 2016	20388351.82	2436199.51
Nov 2016	20543930.93	2469553.74
Dec 2016	19846862.42	2403985.01

Tableau Visualization