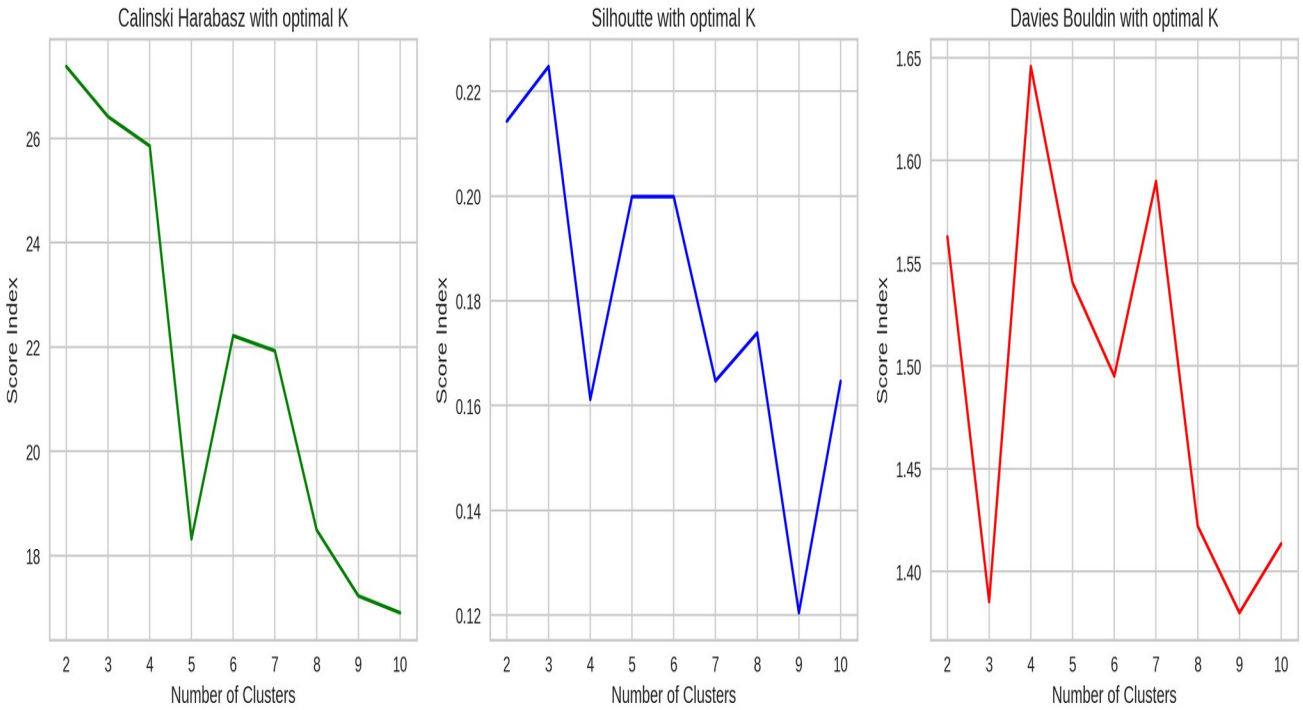
Project: Predictive Analytics Capstone

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

Task 1: Determine Store Formats for Existing Stores

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

   The optimal number of store formats is 3. I arrived at this number by doing the following steps:
   1. Aggregating the sum of stores data by StoreID and Year
   2. Finding the percentage of sales data per category for clustering
   3. Filtering only 2015 sales data.
   4. Clustering the sales data by using K-means clustering analysis and testing the score using Calinski Harabasz Score, Silhoutte Score, Davies Bouldin Score methods and comparing the results of these methods as below:



Therefore, from the above plot, the score result shows that:

- Calinski Harabasz Score:-the higher the score is the better: 2 Cluster
- Silhoutte Score:- the higher the score is the better: 3 Cluster
- Davies Bouldin Score:- the lower the score is the better: 3 Cluster

2. How many stores fall into each store format?

[125]:

	Segment	Size
0	0	23
1	1	29
2	2	33

N.B: I used Python hard coding to implement and analysis forecasting new store format and the result may vary depending on the random state setting, but I checked that it is very close to the result of alteryx software.

I don't use Alteryx due to license reason.

I would like to thank project reviewer. Their feedback really helped me. I learned a lot and improved my work and got the right store format above.

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

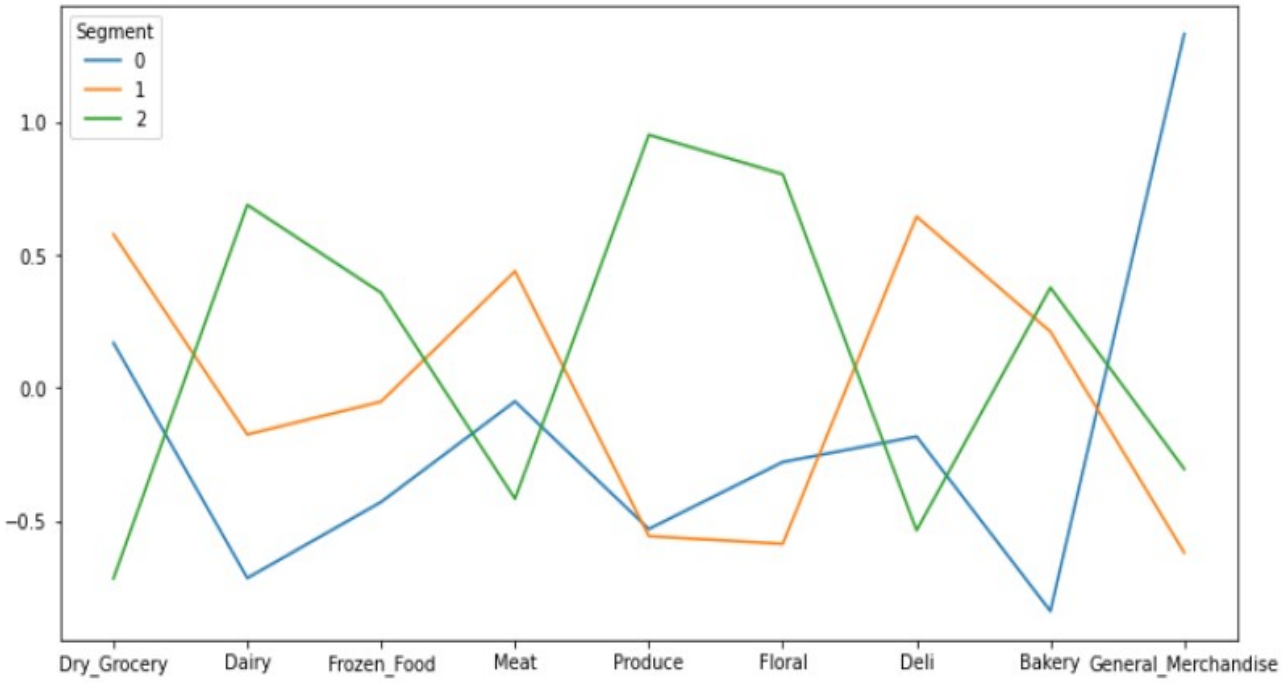
I compared the percentage sales of each cluster and identified the differences of each cluster as below table

[121]:

Segment	Dry_Grocery	Dairy	Frozen_Food	Meat	Produce	Floral	Deli	Bakery	General_Merchandise
0	0.169648	-0.715919	-0.429498	-0.050894	-0.531226	-0.279179	-0.183078	-0.840332	1.332182
1	0.578442	-0.175279	-0.051571	0.439251	-0.557843	-0.587300	0.645243	0.212032	-0.620086
2	-0.717496	0.689004	0.358039	-0.417302	0.952837	0.804372	-0.536132	0.377492	-0.305331

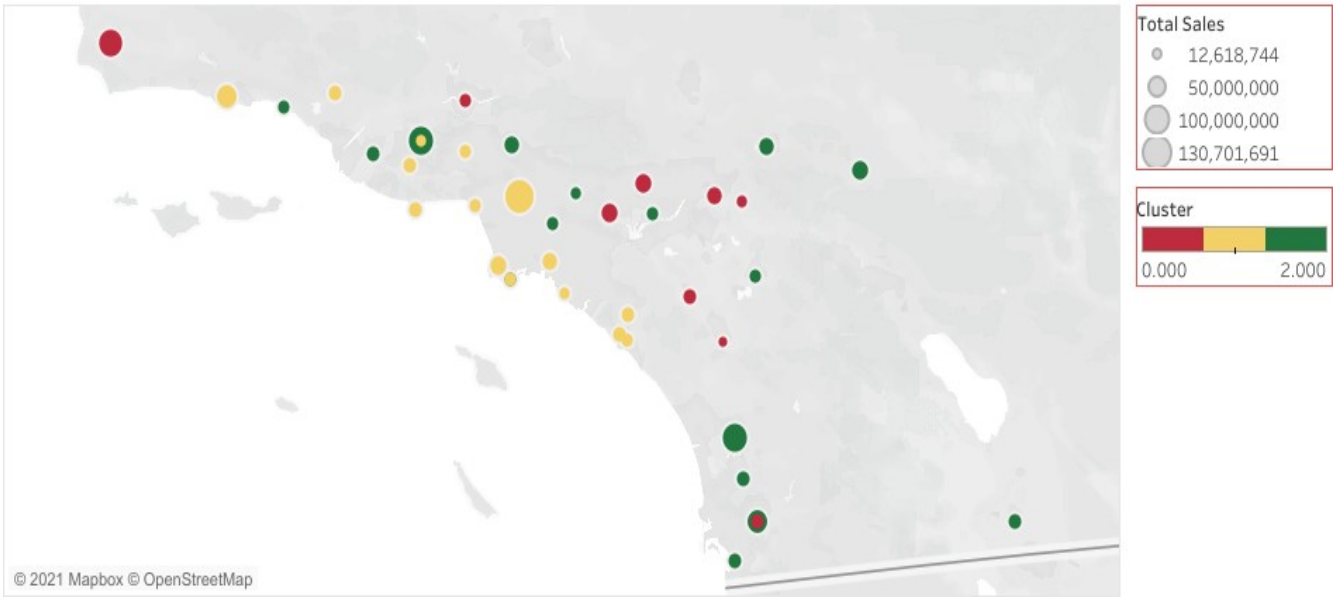
Based on the results of the clustering model above, we can see that General_Merchandise has a higher value in Cluster 0 when comparing the other two clusters which mean that the products in general merchandise category are more likely to be sold to wards **cluster 0.** Unlike **Cluster 0** and **Cluster 2,** the stores in **Cluster 1** have relatively high sales on "Produce, Dairy, and Floral" product categories.

Moreover, I visualized the average percentage sales in each cluster that shows the differences between clusters.



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.

The Cluster of Existing Store by Total Sales



Note that Cluster number started from 0.

Task 2: Formats for New Stores

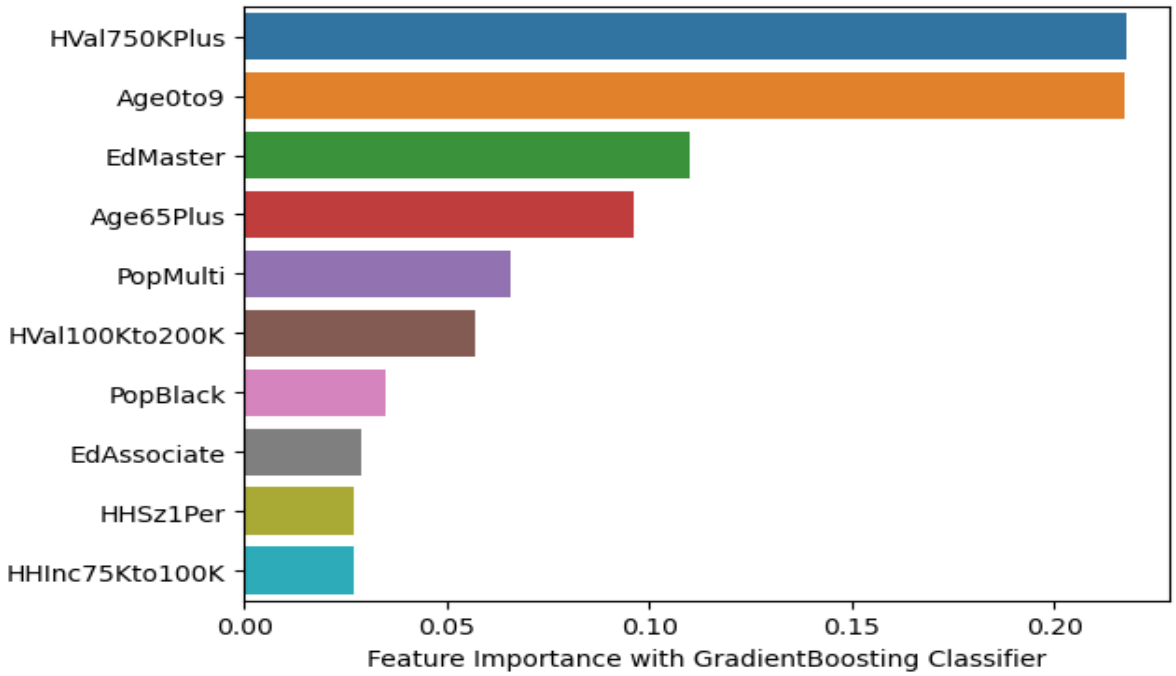
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

Model Comparison with Accuracy

	Overall_Accuracy	Cluster 0	Cluster 1	Cluster 2
Random Forest	0.71	0.5	0.83	0.71
Decision Tree	0.71	1.0	0.83	0.43
GradientBoosting	0.88	1.0	1.00	0.71

I tested all the three models based on the criteria given. From the above comparison report, we can see that the boosted model is better performed with 88% of overall accuracy and also predict ed cluster 0 and cluster 1 with 100% accuracy and cluster 2 is 71%.

Therefore, I choose that boosted model is the best model to predict the best store format for the new stores.



From the above feature importance plot using boosted model, the top three variables are **HVal750Kplus, Age0to9, and EdMaster.**

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

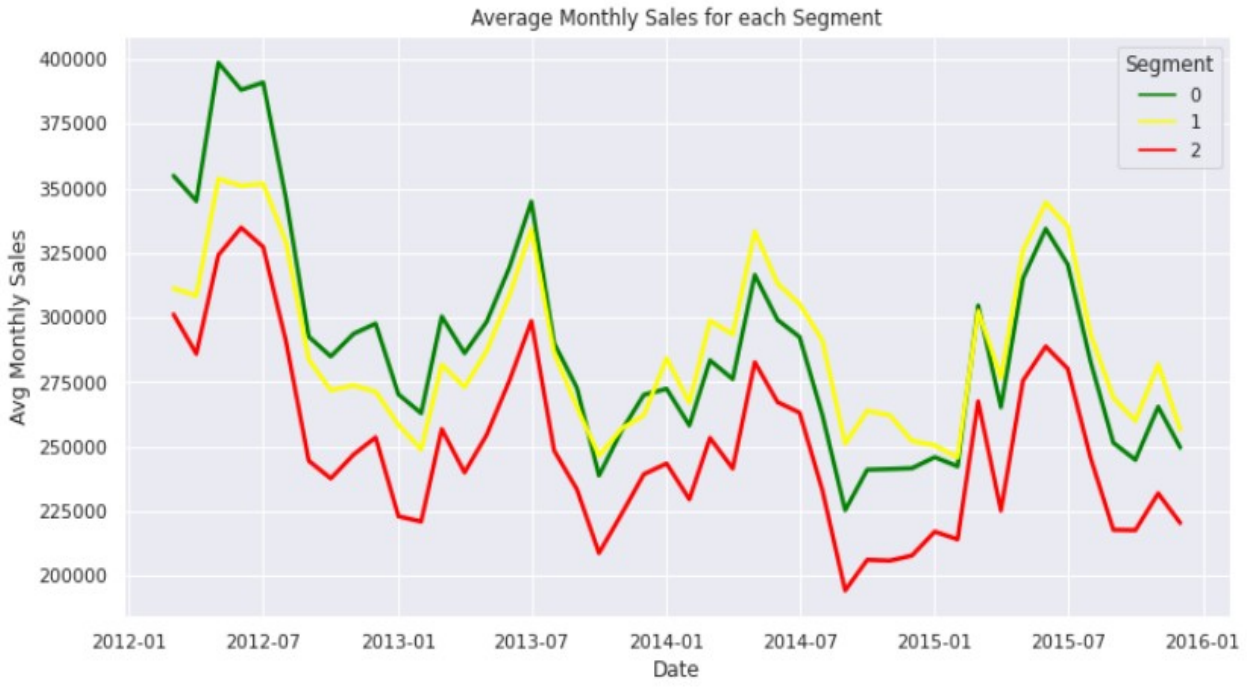
[174]:

	Store	Segment
0	S0086	2
1	S0087	1
2	S0088	2
3	S0089	1
4	S0090	1
5	S0091	0
6	S0092	1
7	S0093	0
8	S0094	1
9	S0095	1

Task 3: Predicting Produce Sales

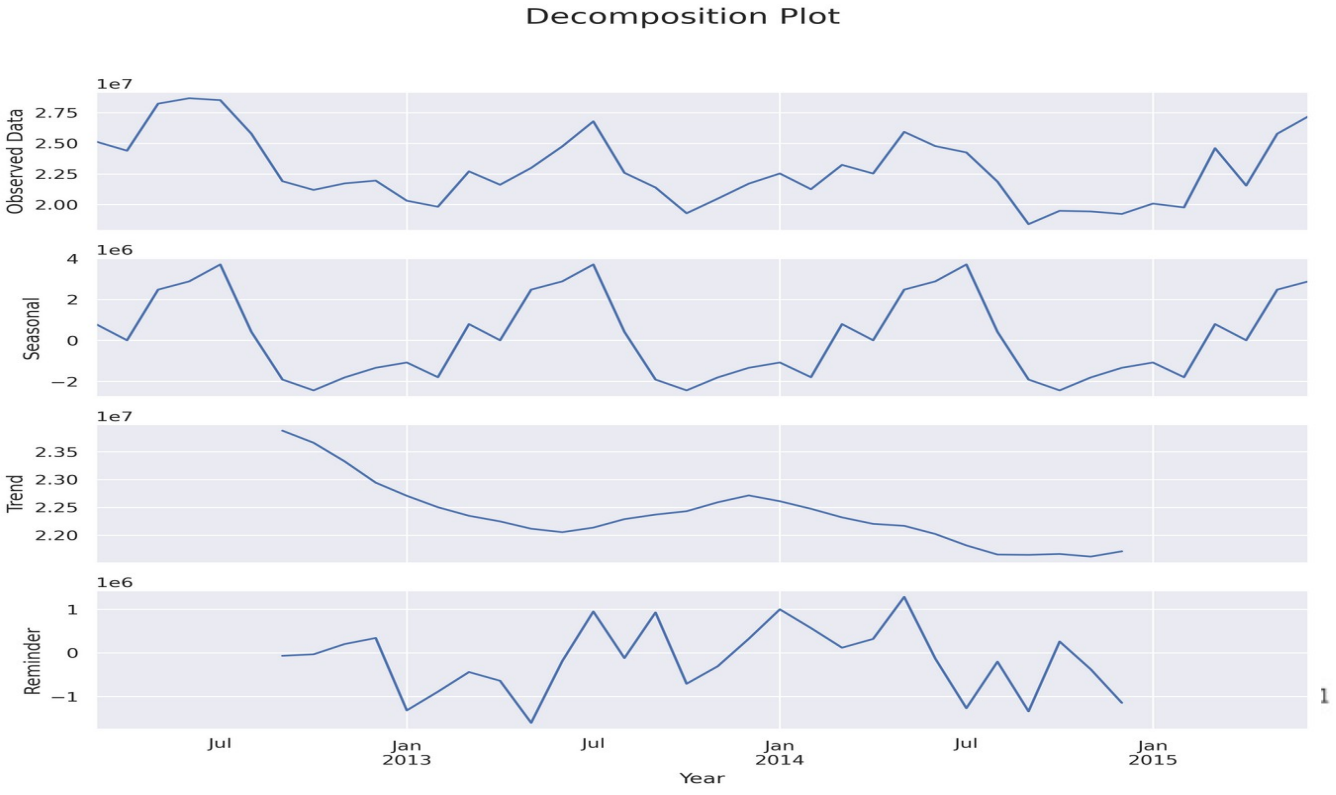
1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?

Before I came to the decision which model is appropriate for forecasting new store formats, I have calculated and visualized the average monthly sales for each cluster.



I have trained both ETS and ARIMA models using statsmodels library with the support of TS visualization of existing stores sales data.

Time Series Decomposition Plot



As we can see the above decomposition plot, there are three components: trend, seasonal and the error component. Each of these components makes up the time series and helps us to confirm what we saw in the previous time series plot of existing sales data.

From this we can clearly see that, there is a seasonality in the data and we can also recognize that the magnitude of seasonality is slightly decreasing. The error shows changing variance as the time series moves along. Therefore, the most appropriate ETS model is ETS(m, n, m) where as it has the lowest AIC value, while comparing to other possible ETS models.

From the comparison plot for both ARIMA and ETS above, we can see the forecast values by ETS(m, n, m) model is most near to the actual values than the forecast values by ARIMA(0,1,1) model. Therefore, I choose that ETS is the best model for forecasting the 2016 stores sales data of each segment.

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.

[302]:

	New Stores	Existing Stores
2016-01	21129077.0	2552112.0
2016-02	20408753.0	2477090.0
2016-03	23638763.0	2570738.0
2016-04	22375530.0	2561306.0
2016-05	25576825.0	2551866.0
2016-06	26222939.0	2542430.0
2016-07	26426500.0	2532990.0
2016-08	23258223.0	2523558.0
2016-09	20592937.0	2514122.0
2016-10	20117865.0	2504682.0
2016-11	20909303.0	2495248.0
2016-12	20887023.0	2485808.0