

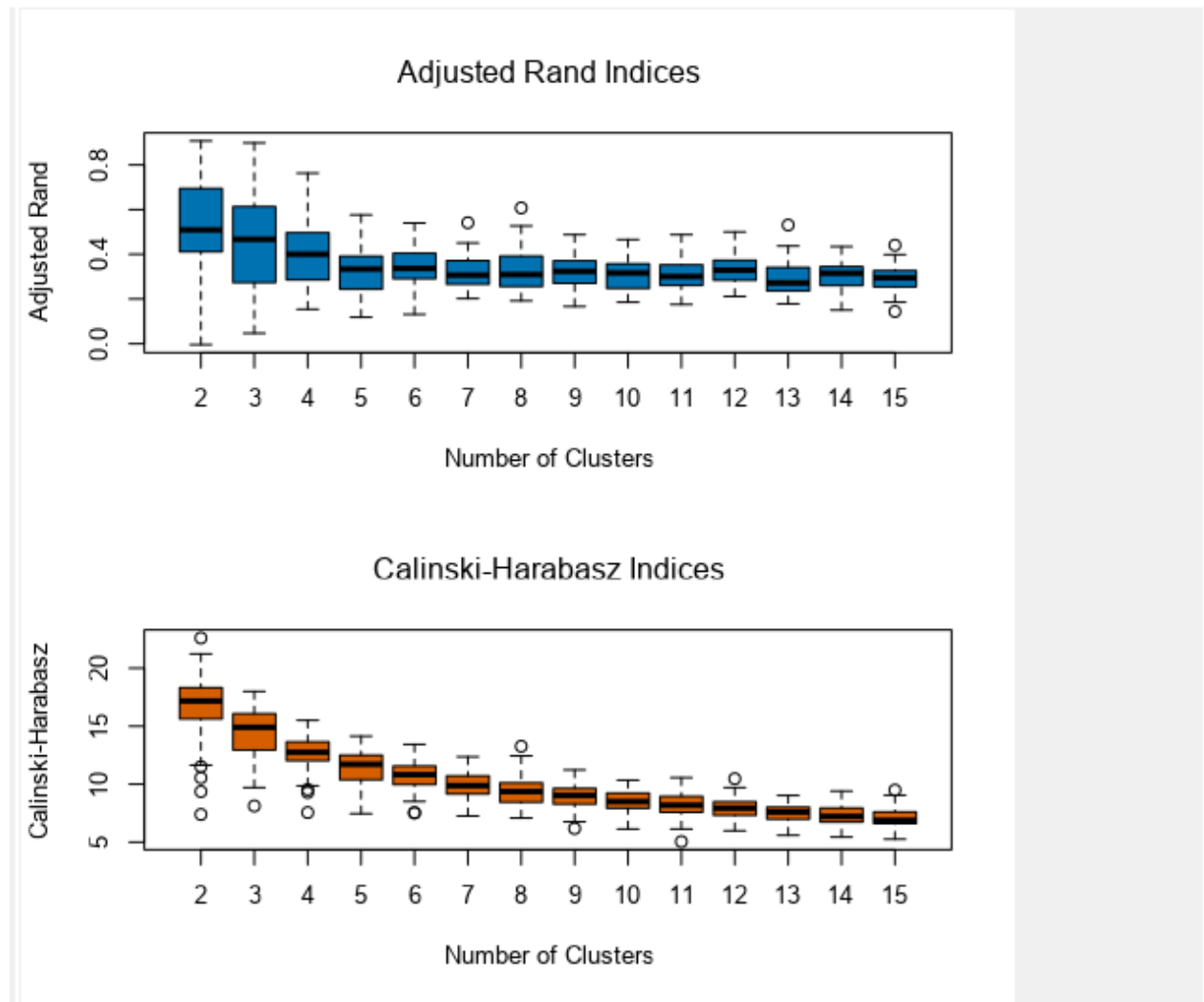
Project: Predictive Analytics Capstone

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?

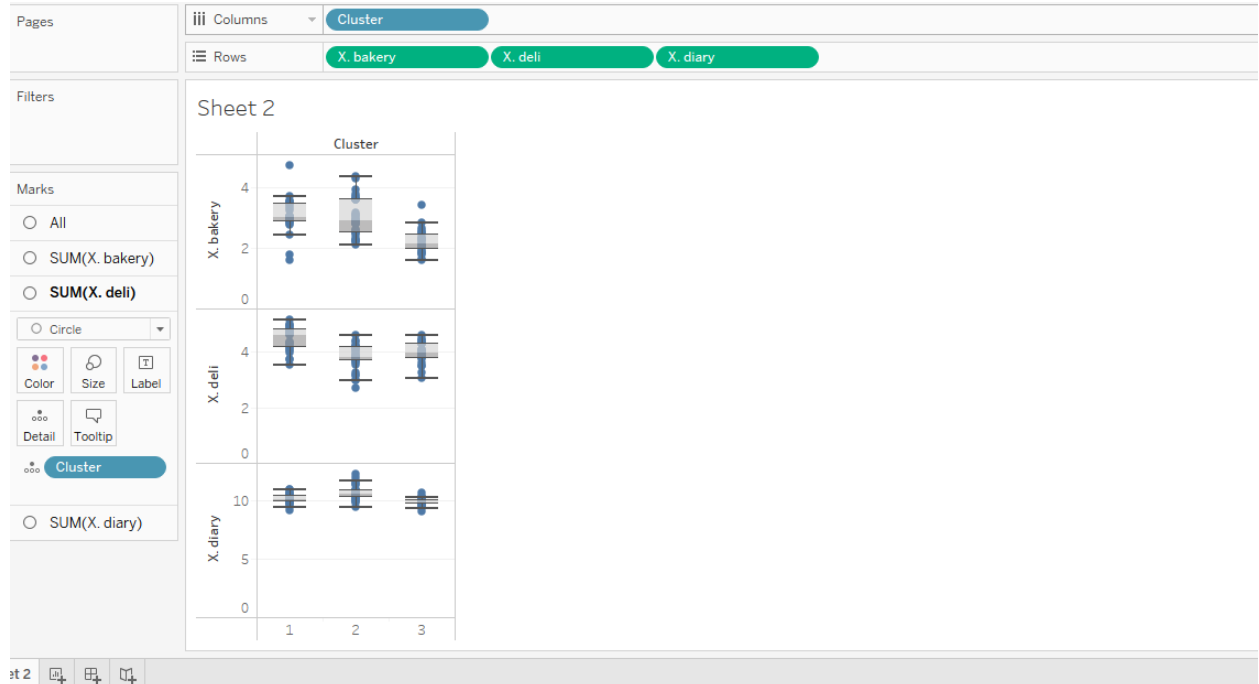
Answer: 3 is the optimal number of store formats. I arrived at the number after using the k-centroids diagnostics tool in Alteryx to check out the Adjusted Rand indices and the Calinski - harabasz indices. The k-means clustering method stands out the most with the highest points at 2 clusters.



2. How many stores fall into each store format?

Cluster	Number of stores
Cluster 1	25
Cluster 2	35

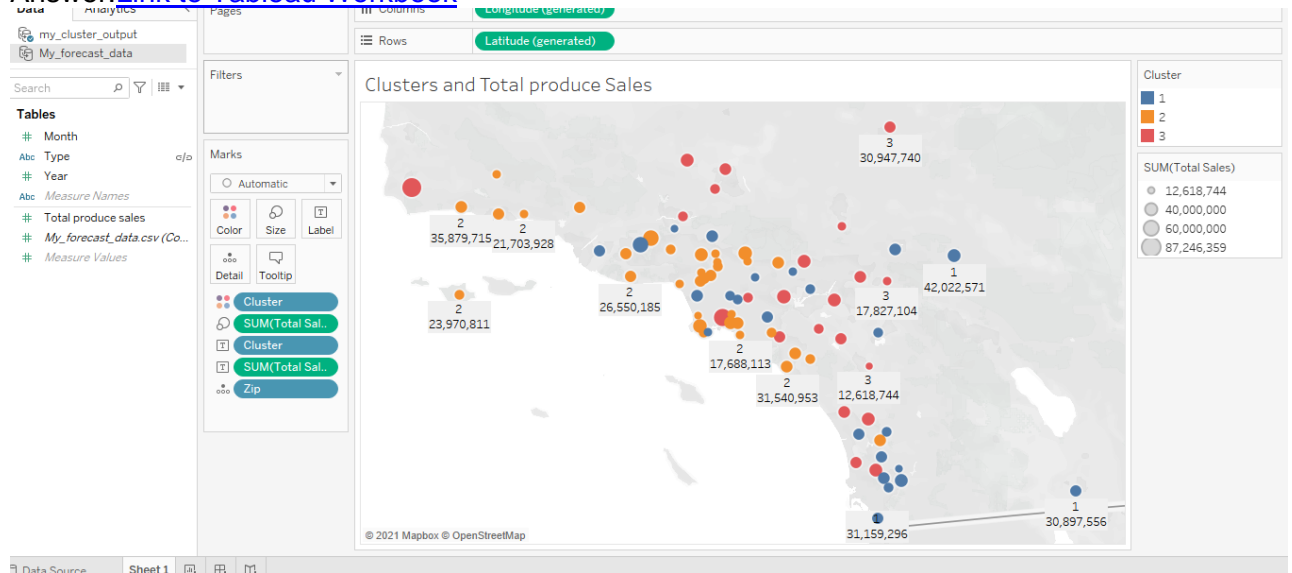
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?



From the above box and whiskers plot of clusters and sales category, it can be seen that each cluster is different in terms of sales for each of the category for instance in terms of deli, cluster 2 is made up of stores that their percentage of sales from deli is the lowest.

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.

Answer: [Link to Tableau Workbook](#)



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.)

Answer: I chose the boosted model for classifying the new stores because of all the models it has the best accuracy metrics that is in terms of the general accuracy, the precision and in-class accuracy as shown below:

Model Comparison Report

Fit and error measures					
Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_8	0.6471	0.6667	0.5000	1.0000	0.5000
forest_project6	0.7059	0.7500	0.5000	1.0000	0.7500
PT_boosted	0.7647	0.8333	0.5000	1.0000	1.0000

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	3
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?

Answer: I used the ETS (M, N, M) for forecasting. This is because overall, it has the better accuracy measures in terms of the Mean error, the Root mean squared error and the MASE when compared with the ARIMA model as shown below:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_MNM	-21581.13	663707.2	553511.5	-0.0437	2.5135	0.3257
ARIMA	-604232.29	1050239.2	928412	-2.6156	4.0942	0.5463

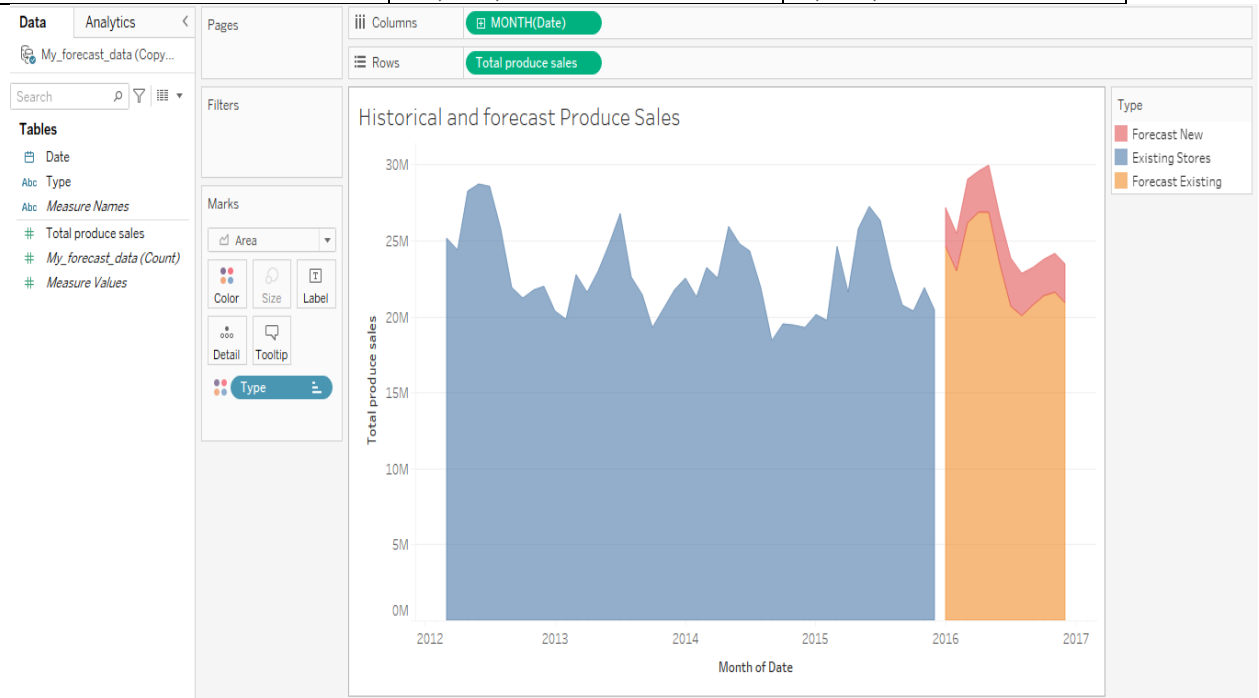
Although the AIC of the ARIMA model is lower than that of the ETS, I cannot use the AIC as a metric because one cannot compare the AICs of two different models. Also the ETS is applied in M,N,M because as can be seen from the decomposition plot below, the error is fluctuating over time which shows that it should be applied multiplicatively, there is no clear pattern in the trend(Neutral) and the seasonality also shows highs and lows suggesting multiplicity.



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	Existing Stores	New Stores
Jan	21,611,877.980495	2,542,607.196008
Feb	20,931,380.132725	2,453,949.95867
March	24,588,621.430699	2,852,466.037767
April	22,974,656.794772	2,707,887.118146
May	26,185,910.648663	3,063,031.113953
June	26,879,542.76363	3,120,920.161619
July	26,860,649.680752	3,143,616.534095

August	23,468,263.325883	2,792,396.757633
September	20,668,472.421463	2,481,441.001115
October	20,054,551.621834	2,425,385.938578
November	20,752,511.328094	2,517,119.713135
December	21,328,394.834456	2,490,642.445392



[Link to Tableau Public](#)

Before you submit

Please check your answers against the requirements of the project dictated by the rubric. Reviewers will use this rubric to grade your project.