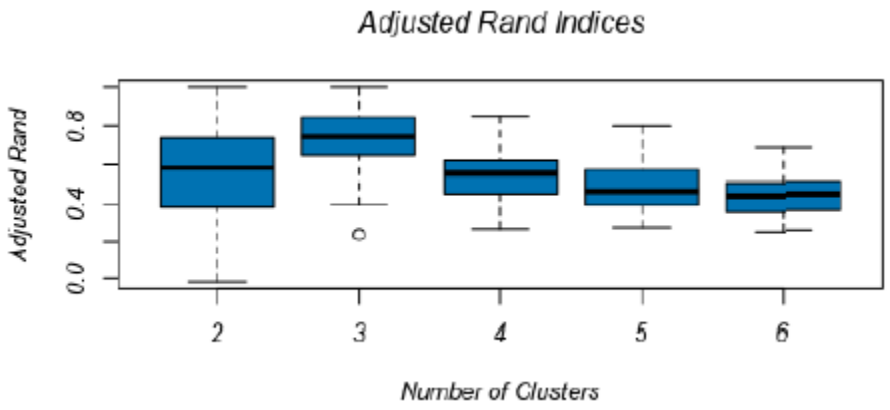
Project: Predictive Analytics Capstone

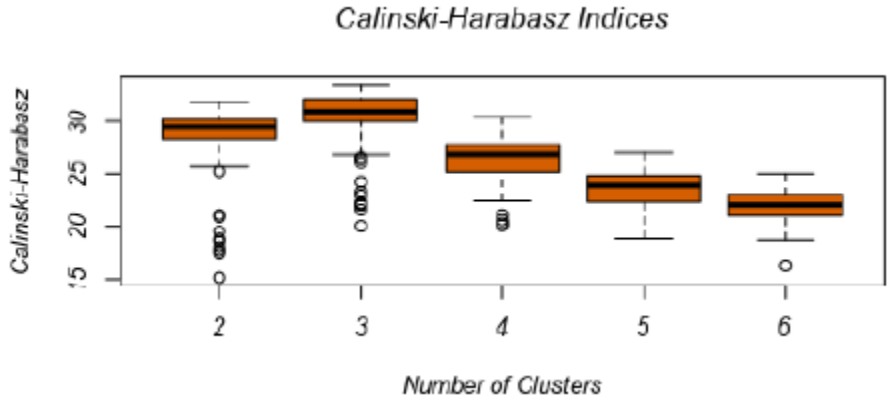
Complete each section. When you are ready, save your file as a PDF document and submit it here: https://coco.udacity.com/nanodegrees/nd008/locale/en-us/versions/1.0.0/parts/7271/project

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 stores formats is 3. I used a K-Means Clustering method and I've attached below the graphs that show that the ideal number is three.





2. How many stores fall into each store format?

Cluster	Size
1	23
2	29
3	33

23 stores on the Cluster 1
29 stores on the Cluster 2
33 stores on the Cluster 3

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

✓ Below I've attached the K-Centroids cluster analysis built in Alteryx. Note that the more positive the number is, the higher the sales for that particular product are.

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

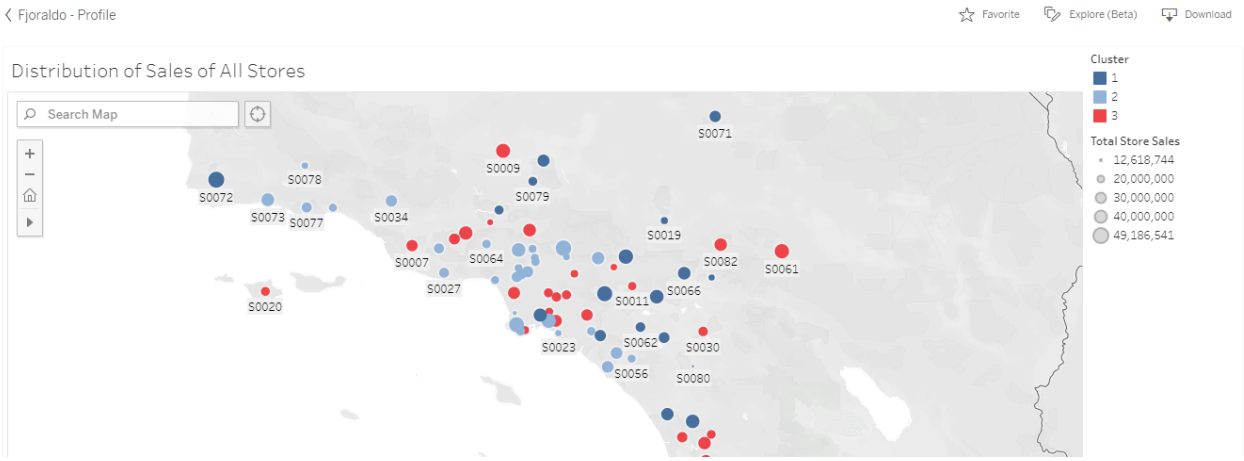
Convergence after 12 iterations.
Sum of within cluster distances: 196.83135.

	Perc_Dry_Grocery	Perc_Diary	Perc_Sum_Frozen_Food	Perc_Sum_Meat	Perc_Sum_Produce	Perc_Sum_Floral	Perc_Sum_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	0.46698	-0.53665	-0.538327	0.64952

	Perc_Sum_Bakery	Perc_Sum_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

✓ For Cluster 1 the driver is: General Merchandise
✓ For Cluster 2 the driver is: Production
✓ For Cluster 3 the driver is: Meat and Deli

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.



LINK TO VIZ:
https://public.tableau.com/profile/fjoraldo#!/vizhome/Project08_CombiningPredictiveTechniques_15988804675310/stores_map

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 Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
DT_Cluster	0.7059	0.7685	0.7500	1.0000	0.5556
FM_Cluster	0.8235	0.8426	0.7500	1.0000	0.7778
BM_Cluster	0.8235	0.8889	1.0000	1.0000	0.6667

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BM_Cluster

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	1
Predicted_2	0	4	2
Predicted_3	0	0	6

Confusion matrix of DT_Cluster

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	2
Predicted_2	0	4	2
Predicted_3	1	0	5

Confusion matrix of FM_Cluster

	Actual_1	Actual_2	Actual_3
Predicted_1	3	0	1
Predicted_2	0	4	1
Predicted_3	1	0	7

- ✓ I used the Boosted Model. As shown in the screenshot above, while the accuracy of the three models is equivalent, the F1 measure of the Boosted Model is higher.

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	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?

I used the original dataset to compare the performance of ETS and ARIMA model and check which one is better for forecasting the store's performance.

I used ETS(M, N, M) without dampening for the ETS model.

The error plot shows variance over the years. It is fluctuating with different seizes; this means we need to use the error multiplicatively(M).

Decomposition Plot – Data Graph

We can’t say for sure if there is a pattern in the below data, that is why I have applied neutral trend(N).

Decomposition Plot – Trend Graph

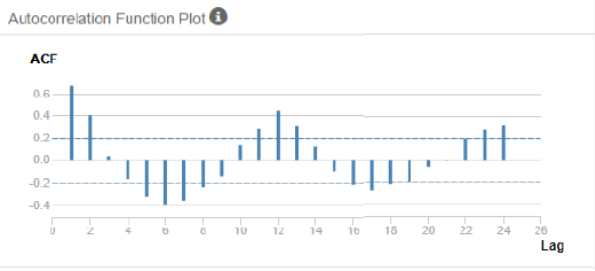
The seasonal plot shows seasonality in similar periods. That is why I have applied seasonality in the multiplicative method(M).

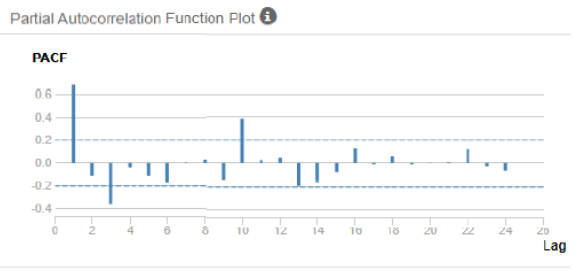
Decomposition Plot – Seasonal Graph

Using a time series plot, we can identify that the plot isn’t stationary, and will need to apply some changes to it to use the ARIMA model effectively.

Time Series Plot

The same is observed on the ACF and PACF function plots.





Using the TS plot, I have discovered that I should use the models with these parameters: (0,1,2)(0,1,0).

After the two models are completed, we can compare how good their predictions are.

Accuracy Measures:

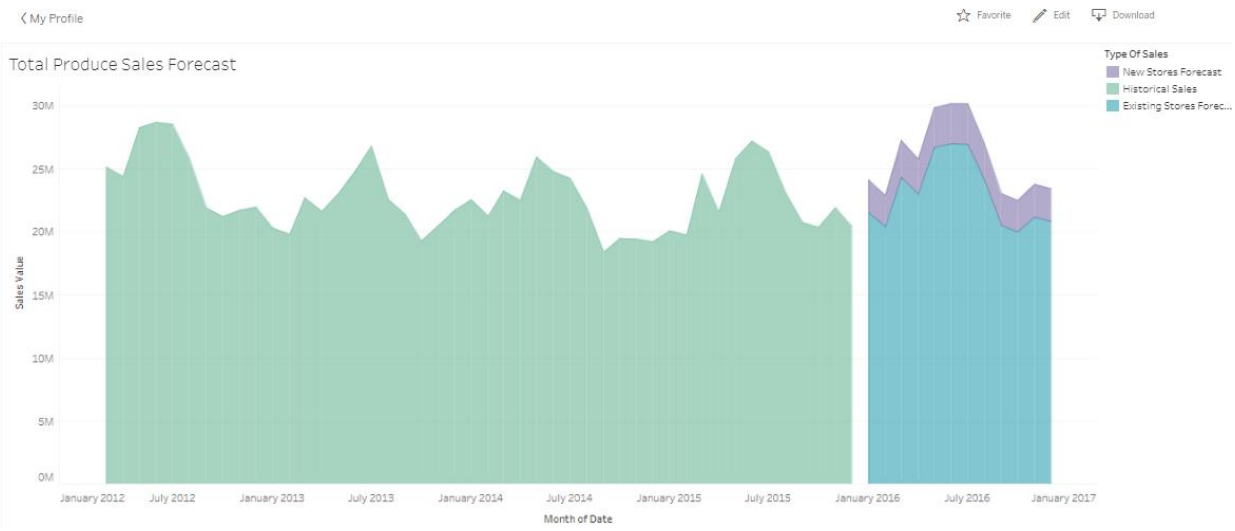
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ARIMA	584382.4	846863.9	664382.6	2.5998	2.9927	0.3909

Accuracy between ETS and ARIMA models

Based on table above, which was obtained from running the two time-series models against the holdout sample of 6 months data, the ETS model's accuracy is higher when compared to ARIMA model . ETS model has lower RMSE value and lower MASE value.

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.

Year	Month	New Store Sales	Existing Store Sales
2016	1	$ 2,587,450.85	$ 21,539,936.01
2016	2	$ 2,477,352.89	$ 20,413,770.60
2016	3	$ 2,913,185.24	$ 24,325,953.10
2016	4	$ 2,775,745.61	$ 22,993,466.35
2016	5	$ 3,150,866.84	$ 26,691,951.42
2016	6	$ 3,188,922.00	$ 26,989,964.01
2016	7	$ 3,214,745.65	$ 26,948,630.76
2016	8	$ 2,866,348.66	$ 24,091,579.35
2016	9	$ 2,538,726.85	$ 20,523,492.41
2016	10	$ 2,488,148.29	$ 20,011,748.67
2016	11	$ 2,595,270.39	$ 21,177,435.49
2016	12	$ 2,573,396.63	$ 20,855,799.11



LINK TO VIZ:
https://public.tableau.com/profile/fjoraldo#!/vizhome/TotalProduceSalesForecast_15988822195880/Sheet1

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.