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. See the following analysis result from K-Centroids Diagnostics:

K-Means Cluster Assessment Report

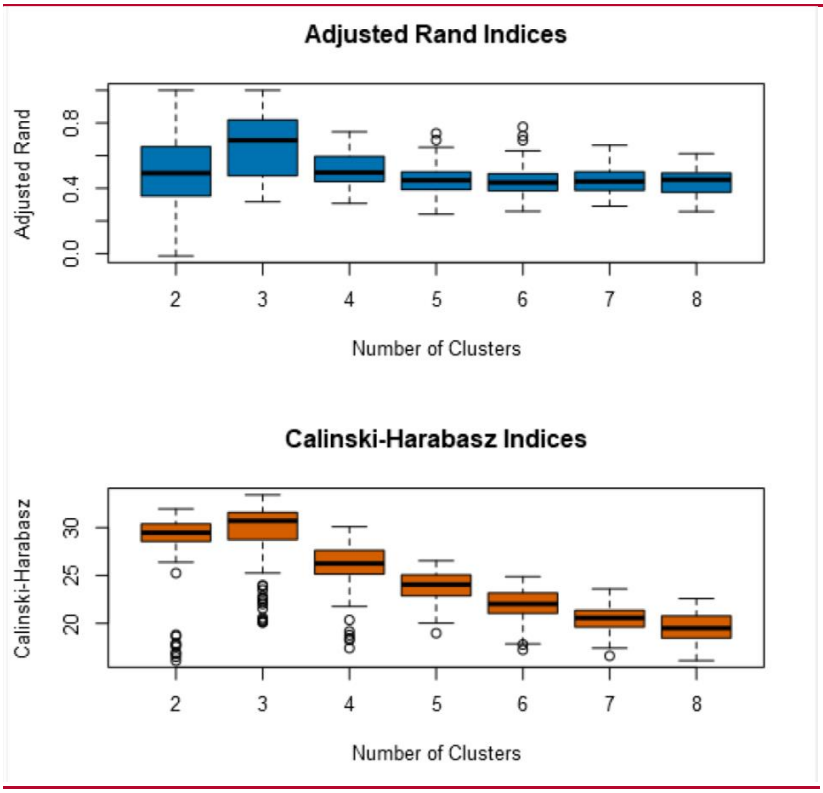
Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7	8
Minimum	-0.0152	0.3171	0.3072	0.2412	0.2586	0.2903	0.2568
1st Quartile	0.352	0.4819	0.4431	0.3943	0.3896	0.3877	0.377
Median	0.4926	0.6936	0.4964	0.4487	0.4348	0.4417	0.4526
Mean	0.484	0.6575	0.5125	0.4623	0.4532	0.4498	0.4411
3rd Quartile	0.655	0.816	0.5913	0.4982	0.489	0.4997	0.491
Maximum	1	1	0.7458	0.7366	0.7762	0.6637	0.6118

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	20.09	17.41	18.98	17.24	16.61	16.11
1st Quartile	28.61	28.76	25.16	22.91	21.05	19.61	18.46
Median	29.47	30.7	26.25	24.05	22.02	20.56	19.5
Mean	28.41	29.47	25.99	23.88	21.96	20.48	19.62
3rd Quartile	30.39	31.58	27.62	25.06	23.14	21.35	20.77
Maximum	31.95	33.41	30.09	26.53	24.87	23.6	22.59



As we can see, 3 clusters has the highest adjusted rand index and Calinski-Harabasz index since the quartiles and the mean are highest among different number of clusters. So we will choose 3 clusters.

2. How many stores fall into each store format?

By choosing 3 clusters, percentages of sales for each category, z-score standardization and K-means, we get the following result:

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

~~2.~~

So sizes of the three clusters are 23, 29, and 33.

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

	Per_Dry_Grocery	Per_Dairy	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_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

	Per_Bakery	Per_General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

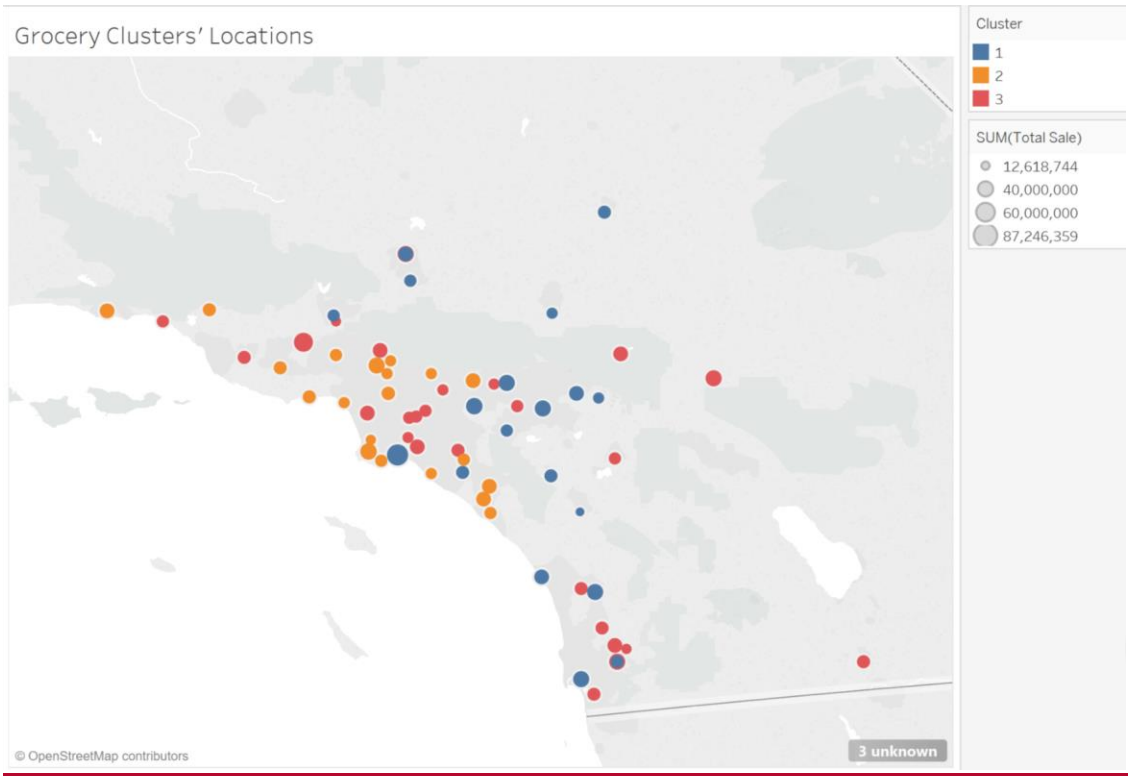
Cluster 1 seems to contain grocery stores by focusing on dry grocery and general merchandise.

Cluster 2 seems to contain fresh markets by focusing on dairy, frozen food, produce, floral and bakery.

Cluster 3 seems to contain Deli stores by focusing on grocery, meat, especially Deli and bakery.

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.

Tableau Public file link:
https://public.tableau.com/profile/charlio#!/vizhome/grocery_geo/Sheet1?publish=yes

5. 

Task 2: Formats for New Stores

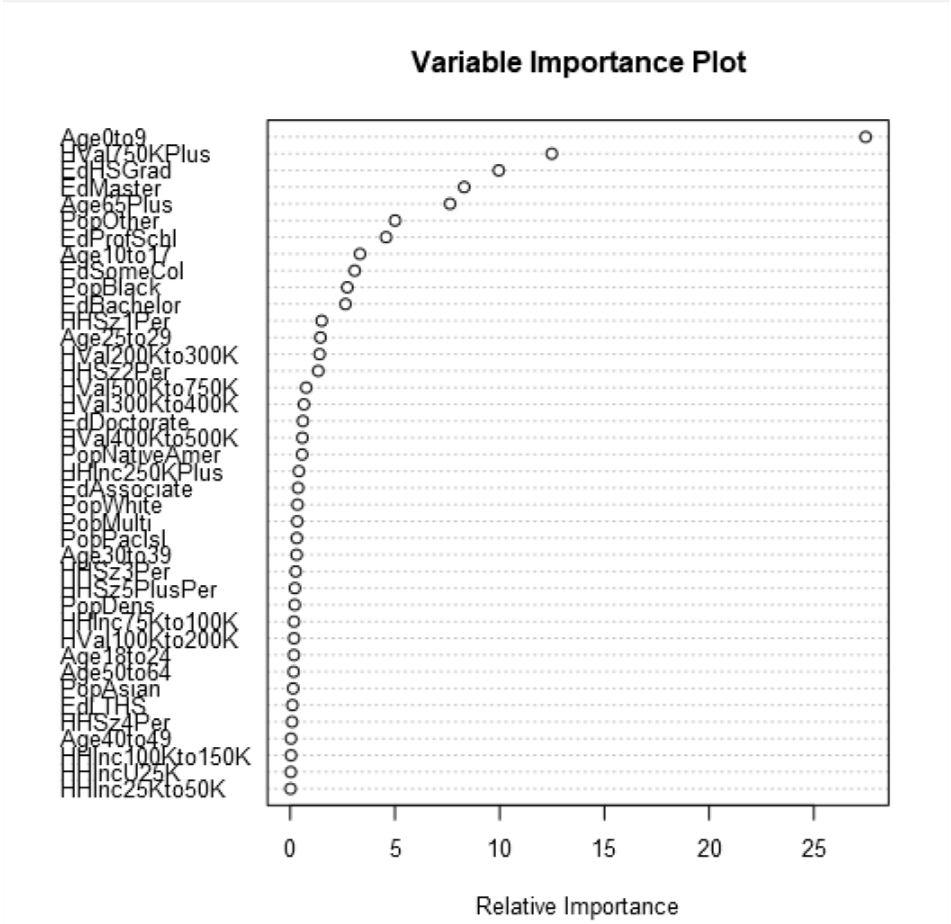
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)

   We need to predict the cluster field for the new stores. Cluster includes 3 values. So this is a non-binary classification problem. Thus we shall use decision tree, forest or boosted models. After running all three models, it turns out boosted model has the highest accuracy and F1 score. So I used the boosted model.

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
forest	0.8235	0.8251	0.7500	0.8000	0.8750
decision_tree	0.7059	0.7327	0.6000	0.6667	0.8333
boosted	0.8235	0.8543	0.8000	0.6667	1.0000

   The three most important variables are Age0to9, HVal750KPlus and EdHSGrad.



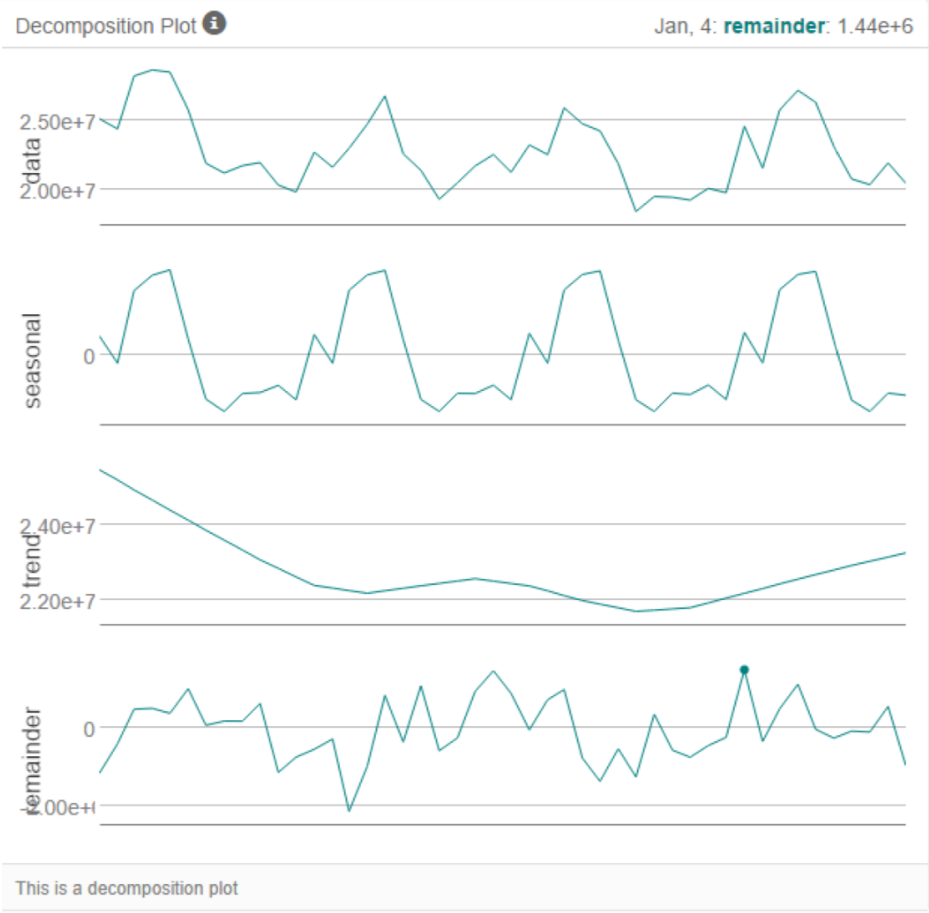
1.

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	1
S0089	2
S0090	2
S0091	1
S0092	2
S0093	1
S0094	2
S0095	2

Task 3: Predicting Produce Sales

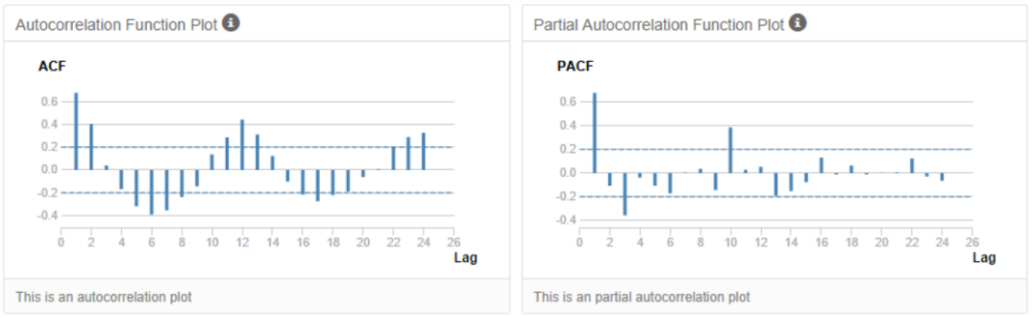
1. 1. What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?



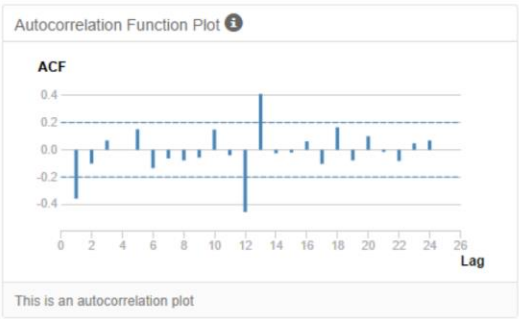
From the above decomposition plot, for ETS model, we shall choose multiplicative error, non trend, and multiplicative seasonality (slightly changing in the above plot), So ETS(M, N, M)

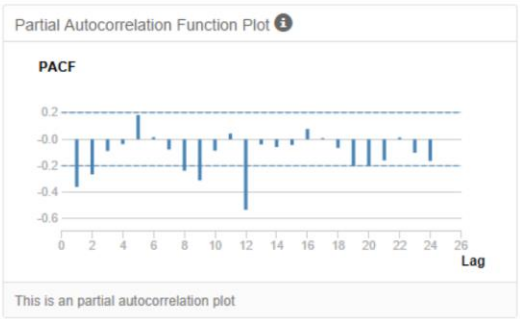
For ARIMA model, after seasonal differencing, and further first and second order differencing, I finally chose ARIMA(ar=0, i=1, ma=2)(P=0, D=1, Q=2). See the following ACF and PACF plots.

Below is the ACF and PACF for the raw total sales for all existing stores



Below is the corresponding seasonal second difference plot:





From the following accuracy measures on 12 holdout samples, we can see the ETS model has much smaller errors (RMSE, MASE) than the ARIMA model. So we will use ETS(M, N, M) model for forecasting.

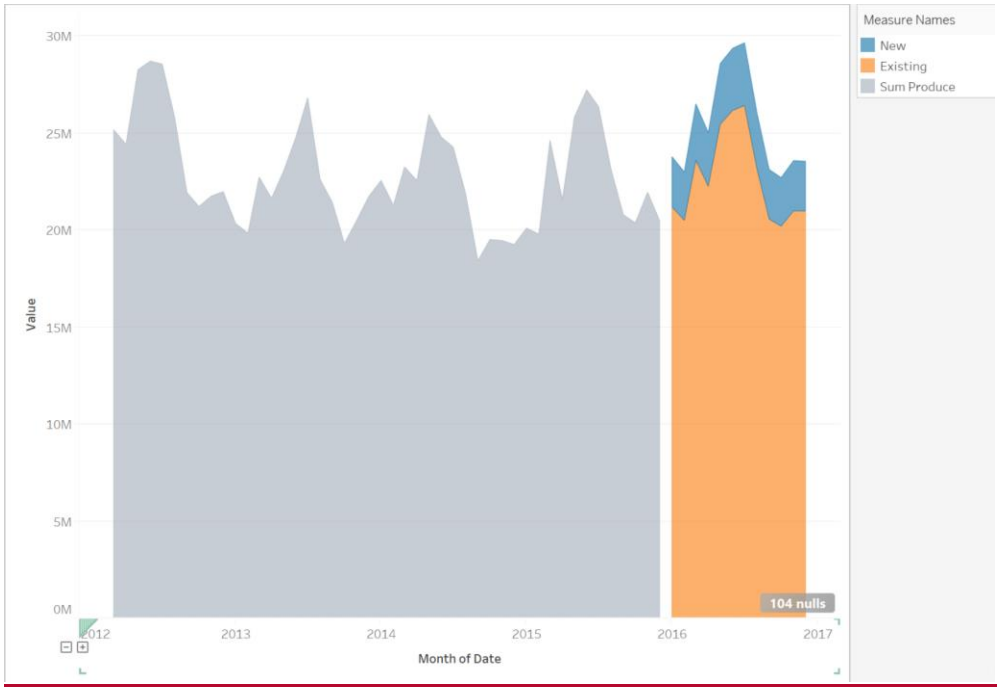
Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE	NA
arima	2868947	3035191	2868947	12.5754	12.5754	1.8355	NA
ets	1978789	2200153	1978789	8.4769	8.4769	1.266	NA

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.

~~2. Please provide a Tableau Dashboard (saved as a Tableau Public file) that includes a table and a plot of the three monthly forecasts; one for existing, one for new, and one for all stores. Please name the tab in the Tableau file "Task 3".~~

Year	Month	Existing	New
2016	1	21174989.40366	2590566.585695
2016	2	20479354.577583	2503135.097223
2016	3	23580340.680392	2910154.07951
2016	4	22236546.234701	2772193.191798
2016	5	25427255.457066	3142262.475899
2016	6	26143967.404048	3203694.414631
2016	7	26399993.267031	3233436.116193
2016	8	23172393.880014	2884618.003153
2016	9	20544268.638821	2562088.683447
2016	10	20182471.085707	2506670.539657
2016	11	20966876.352467	2598150.832185
2016	12	20965097.001692	2566314.03563



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.