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

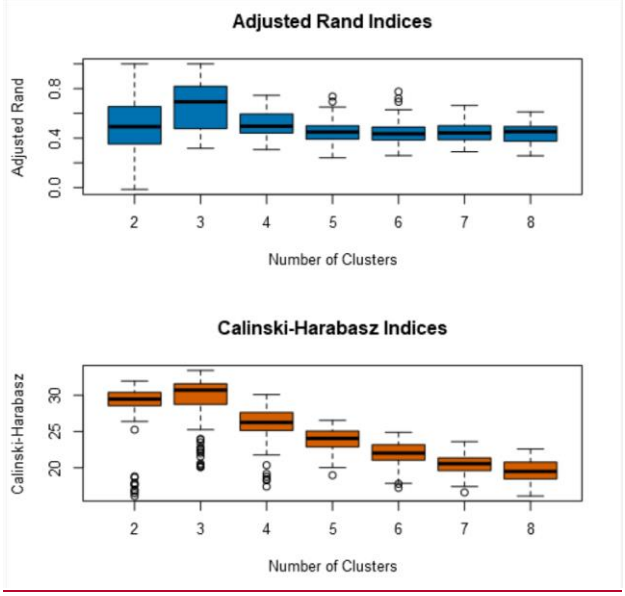
Summary Statistics

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



2. How many stores fall into each store format?

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	33	0.413544	0.933737	0.346518
2	32	0.485158	0.884516	0.358749
3	20	0.432155	0.746268	0.523589

~~2.~~

So sizes of the three clusters are 33, 32, and 20.

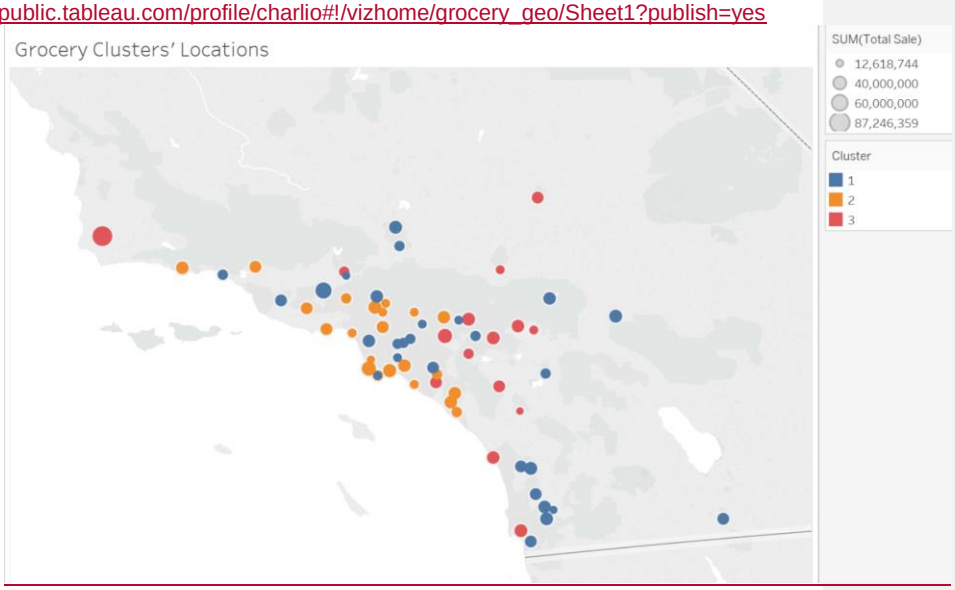
3. Based on the results of the clustering model, what is one way that the clusters differ from one another?

	Per_Dry_Grocery	Per_Dairy	Per_Frozen_Food	Per_Meat	Per_Produce	Per_Floral	Per_Deli
1	0.53963	0.335986	0.491454	0.418523	0.374605	0.258971	0.67632
2	0.2828	0.491688	0.553873	0.283702	0.665884	0.503774	0.43862
3	0.478427	0.230222	0.433789	0.317406	0.364014	0.311984	0.498401

	Per_Bakery	Per_General_Merchandise
1	0.426433	0.253448
2	0.453345	0.322944
3	0.208087	0.739116

Cluster 2 differs from clusters 1 and 3 by focusing more on Dairy, Produce, and Floral. Clusters 1 and 3 differ in Bakery and General_Merchanidse.

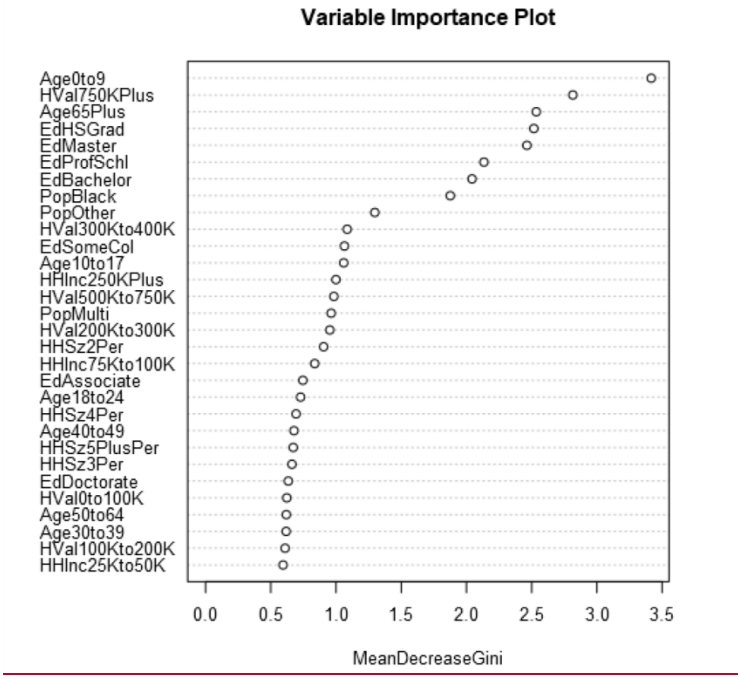
4. Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.
Tableau Public file link:
https://public.tableau.com/profile/charlio#!/vizhome/grocery_geo/Sheet1?publish=yes



~~4.~~

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 forest and boosted have the same result on accuracy and confusion matrix, they both perform better than decision tree. So I used the forest model. The three most important variables are Age0to9, HVal750KPlus and Age65Plus.



~~1.~~

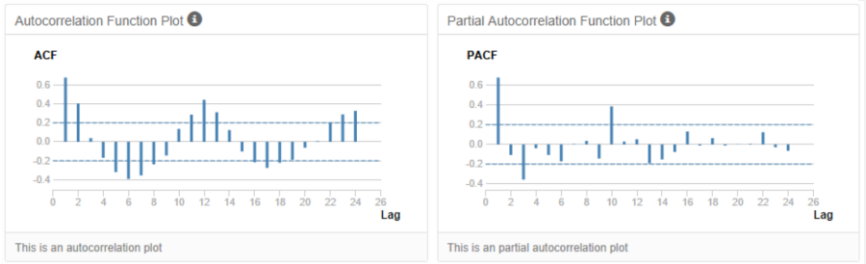
2. What format do each of the 10 new stores fall into? Please fill in the table below.

Store Number	Segment
S0086	1
S0087	2
S0088	1
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2

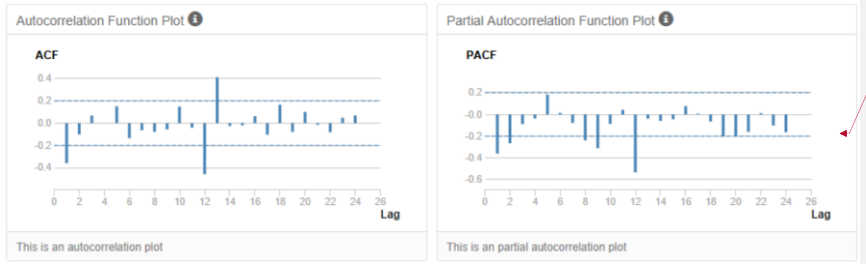
S0095	2

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?
   I used ARIMA(ar=0, i=1, ma=2)(P=0, D=1, Q=2).
   Below is the ACF and PACF for the raw total sales for all existing stores



   Below is the corresponding seasonal second difference plot:



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~~

Formatted: Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.5" + Indent at: 0.75"

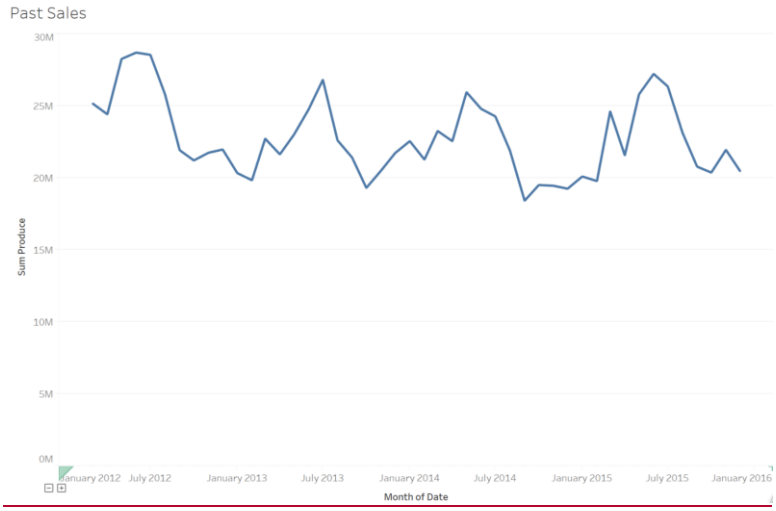
Formatted: Indent: Left: 0.75"

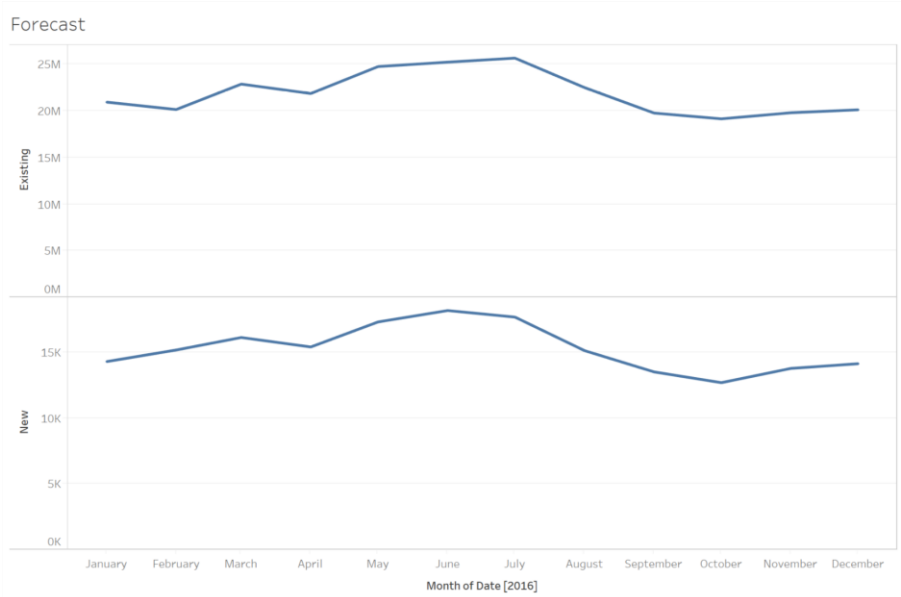
Formatted: Font: (Default) inherit, 11.5 pt, Font color: Custom Color(RGB(79,79,79))

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.5" + Indent at: 0.75"

~~new, and one for all stores. Please name the tab in the Tableau file "Task 3".~~

Year	Month	Existing	New
2016	1	20953969.819059	14314.100171
2016	2	20158134.528995	15196.189186
2016	3	22889569.570794	16148.153034
2016	4	21878375.827751	15426.965215
2016	5	24774819.201672	17333.691219
2016	6	25246762.815218	18199.012921
2016	7	25681136.538454	17701.479078
2016	8	22518098.675437	15149.43436
2016	9	19780347.340634	13529.192196
2016	10	19160357.36375	12709.525821
2016	11	19813266.232422	13795.51934
2016	12	20131473.405634	14152.763131





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.