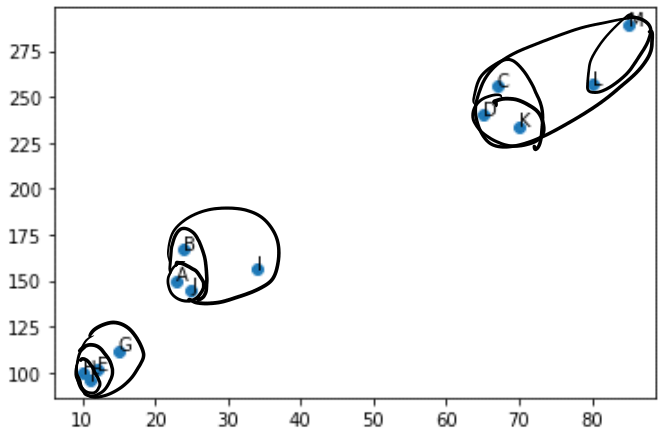
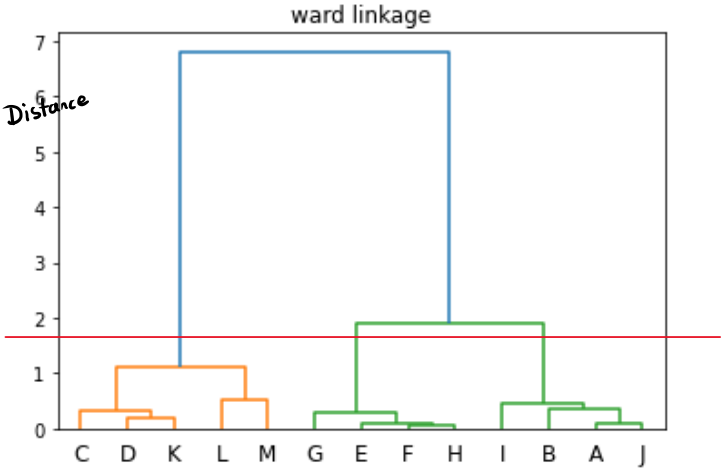
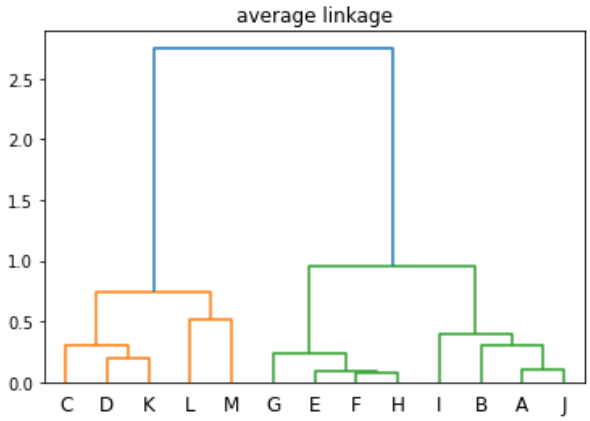
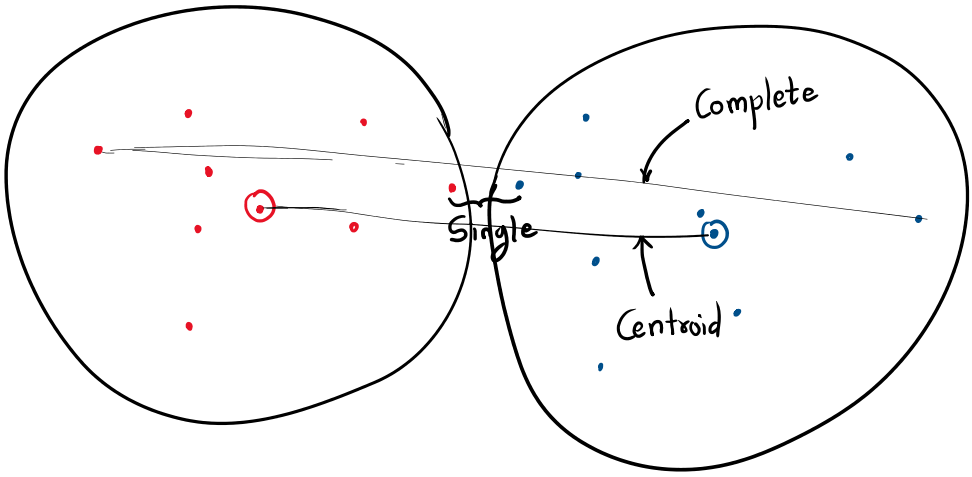
Clustering

06 April 2024 14:06



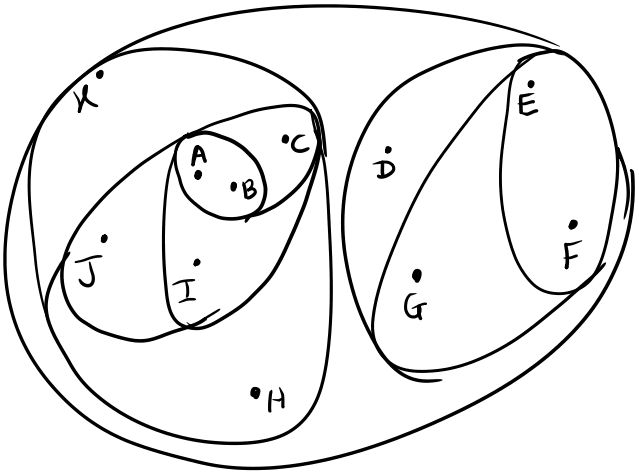


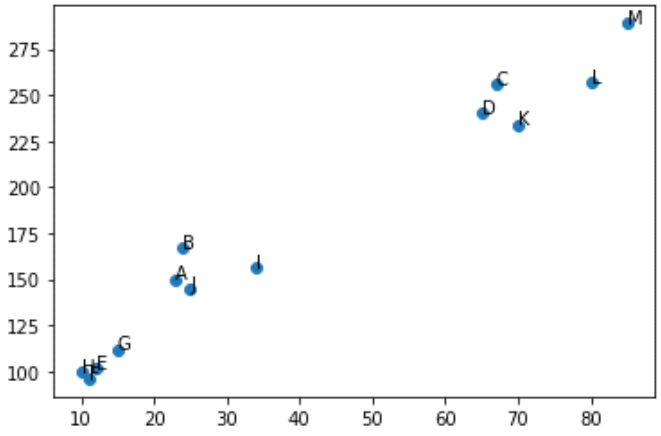




Ward : based on ANOVA

Average



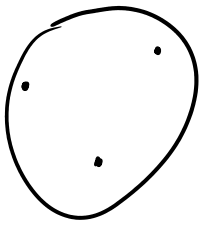


$$\text{Sil coef} = \frac{b-a}{\max(a,b)}$$

b: avg nearest cluster distance

a: avg intra-cluster distance

Sil score = avg (Sil coef of all point)



Datasets / milk.csv

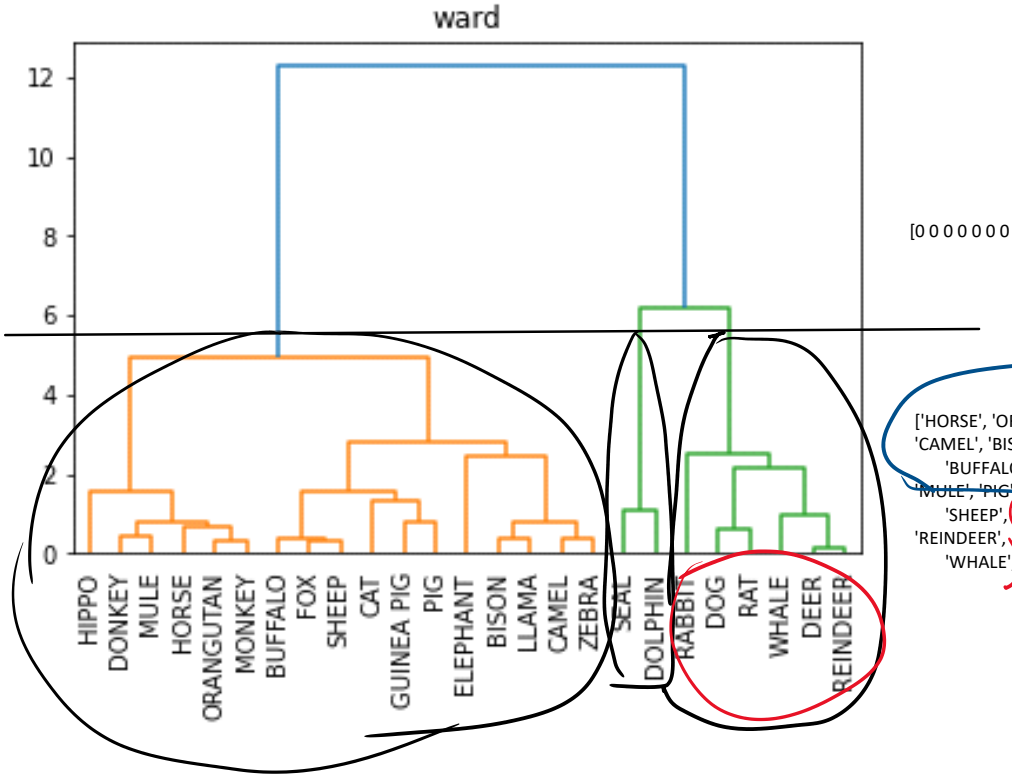
index_col = 0

1) Dendrogram

2) To... clusters = 2 3 4 5 6

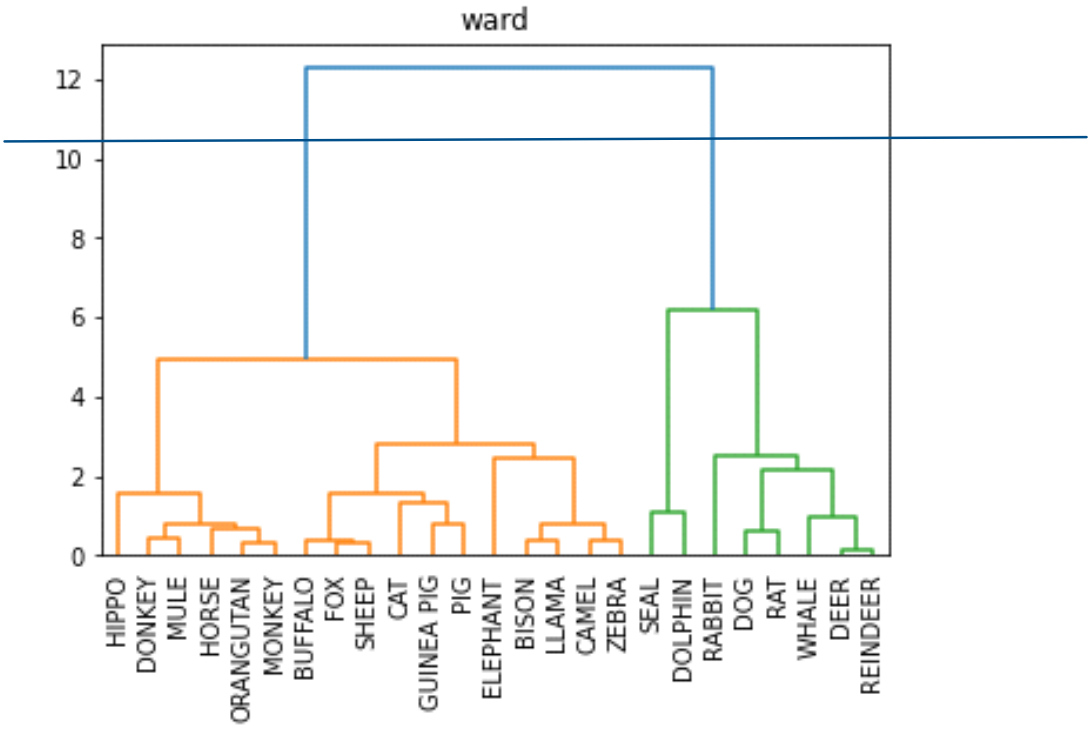
1) Dendrogram

2) Try clusters = 2, 3, 4, 5, 6



[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 2 2]

['HORSE', 'ORANGUTAN', 'MONKEY', 'DONKEY', 'HIPPO', 'CAMEL', 'BISON',
'BUFFALO', 'GUINEA PIG', 'CAT', 'FOX', 'LLAMA', 'MULE', 'PIG', 'ZEBRA',
'SHEEP', 'DOG', 'ELEPHANT', 'RABBIT', 'RAT', 'DEER', 'REINDEER',
'WHALE', 'SEAL', 'DOLPHIN']



Datasets / nutrient.csv

index_col = 0

1) Dendrogram

2) Try clusters = 2, 3, 4, 5, 6

Datasets / Protein.csv

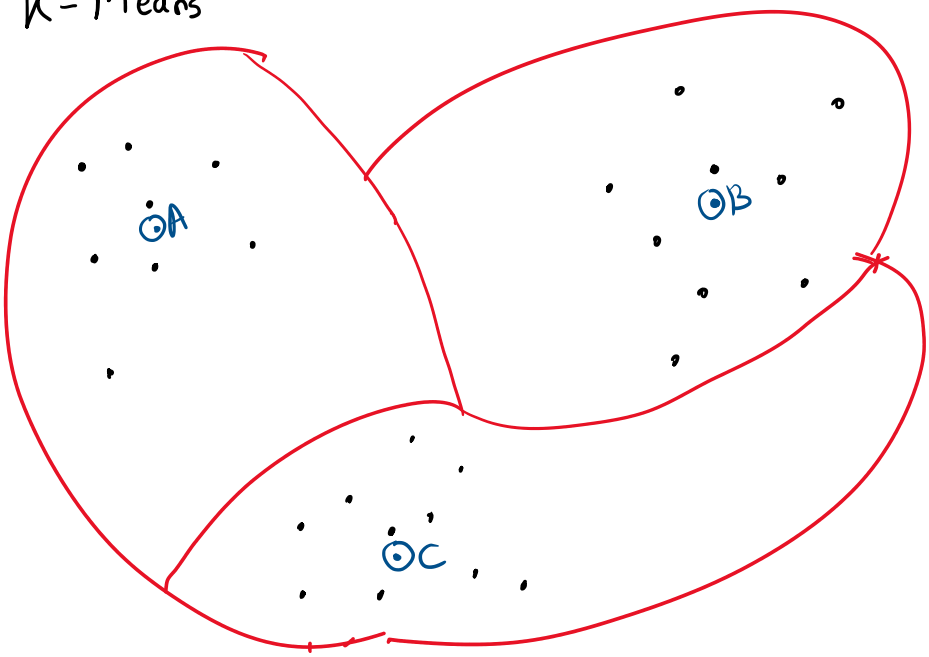
index_col = 0

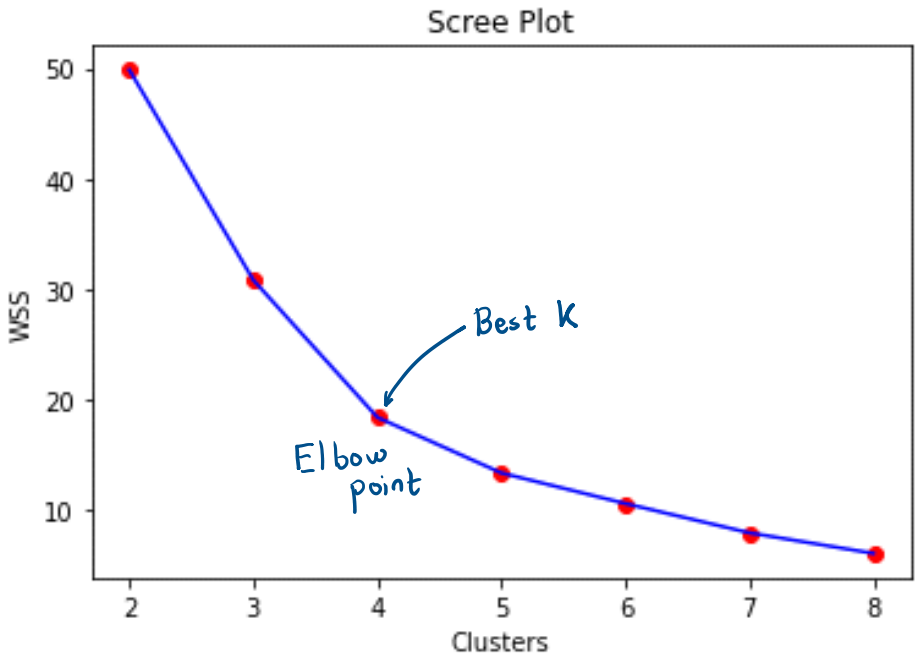
1) Dendrogram

2) Try clusters = 2, 3, 4, 5, 6

$x_1\ y_1$ $x_2\ y_2$ $x_4\ y_4$ $x_3\ y_3$ C $\left(\frac{x_1+x_2+x_3+x_4}{4}, \frac{y_1+y_2+y_3+y_4}{4}\right)$

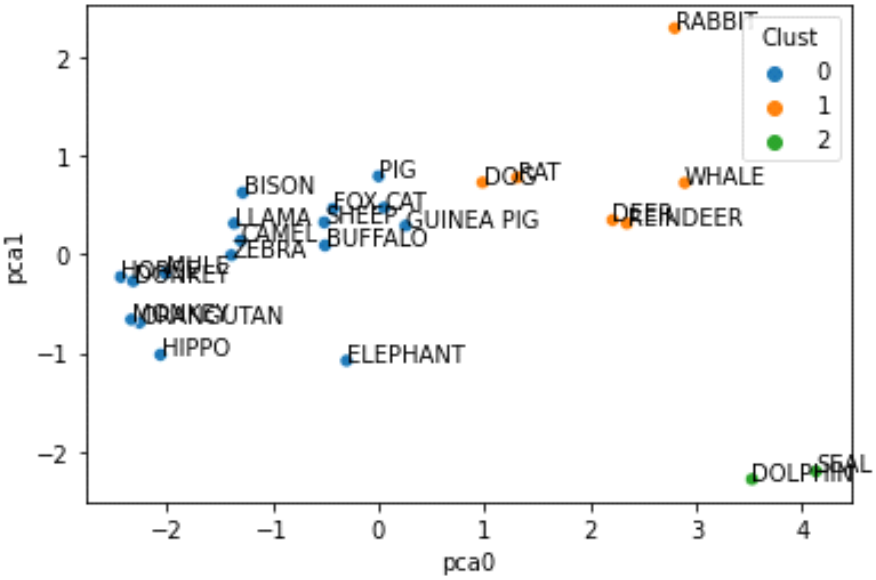
K-Means





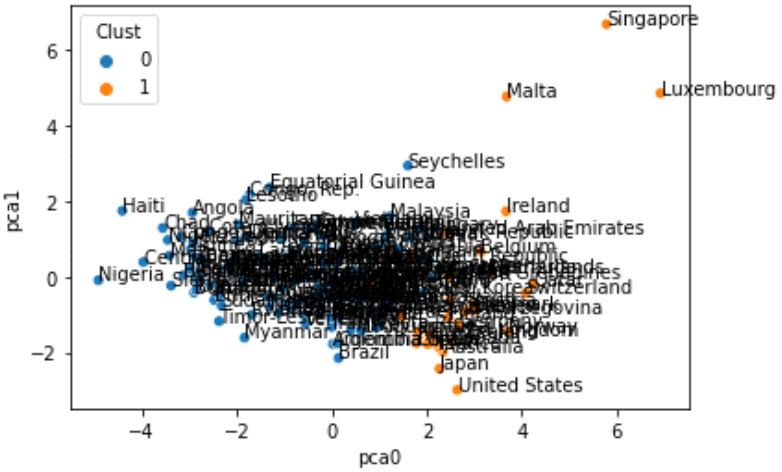
Name	X1	X2
A	-0.620863	-0.411108
B	-0.584506	-0.154456
C	0.978838	1.189194
D	0.906124	0.947639
E	-1.020788	-1.135773
F	-1.057145	-1.226356
G	-0.911718	-0.984801
H	-1.093502	-1.165967
I	-0.220938	-0.320525
J	-0.548149	-0.486594
K	1.087909	0.857056
L	1.451477	1.204291
M	1.633261	1.687401

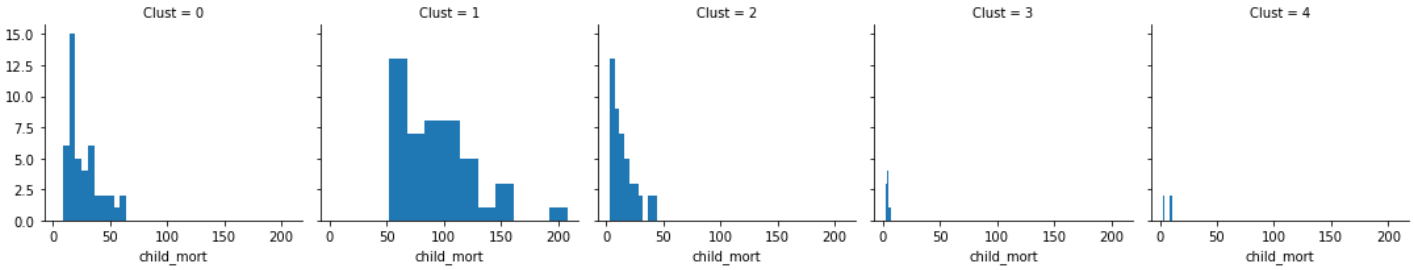
```
In [15]: print(clust.labels_)
[0 0 1 1 0 0 0 0 0 0 1 1 1]
```



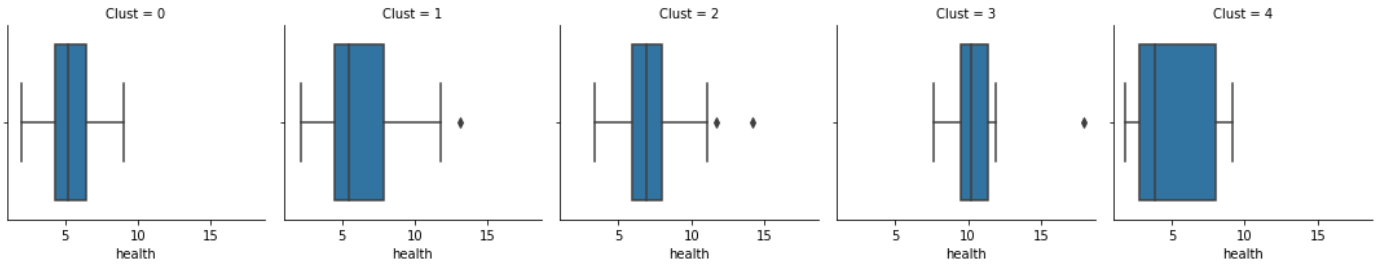
Ward's Distance

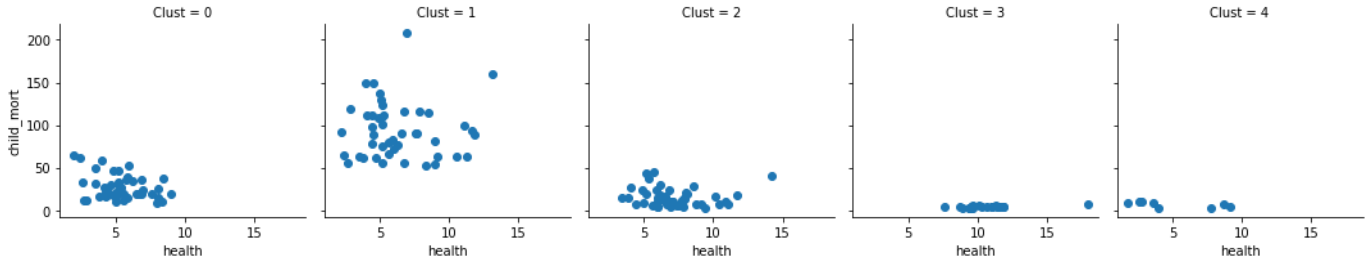
https://www.statisticshowto.com/wards-method/#:~:text=Calculate%20the%20distance%20between%20each,of%20squares%20from%20Step%204.

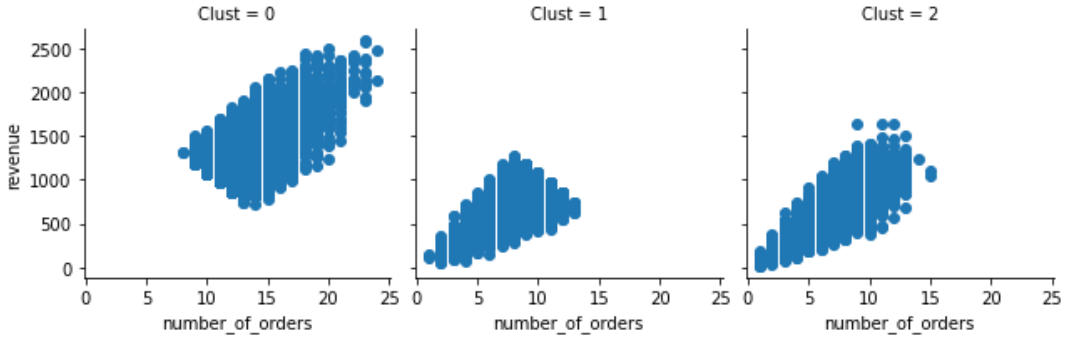


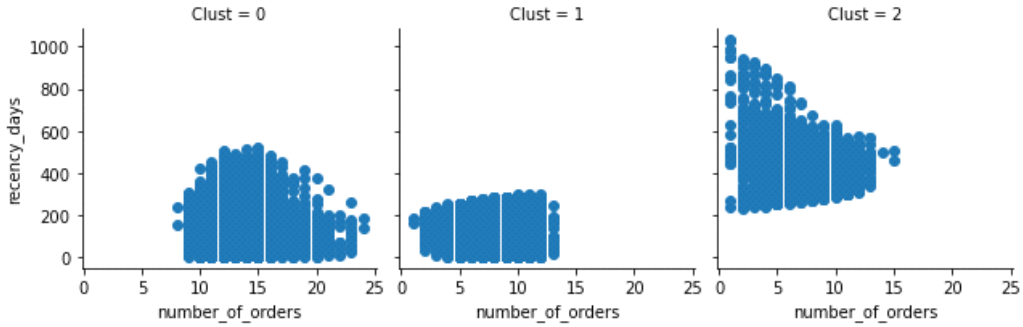


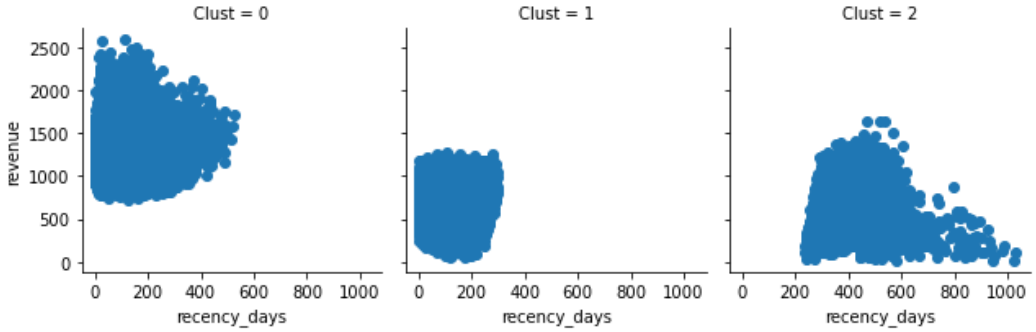

```
In [97]: clust_data.groupby('Clust')['child_mort'].mean()
Out[97]:
Clust
0    26.917778
1    94.180435
2    15.160870
3     4.300000
4     6.937500
```

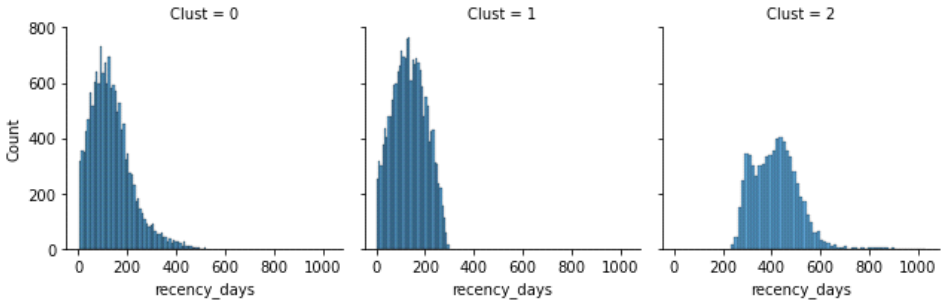




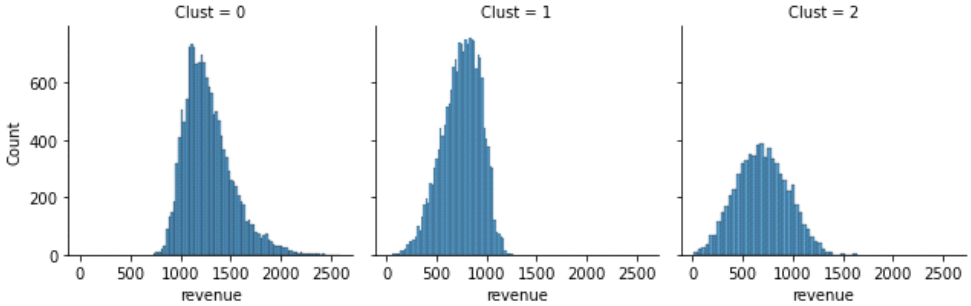

Clust = 0
Clust = 1
Clust = 2
revenue
number_of_orders


Clust = 0
Clust = 1
Clust = 2
recency_days
number_of_orders


Clust = 0
Clust = 1
Clust = 2
revenue
recency_days


Clust = 0
Clust = 1
Clust = 2
Count
recency_days


Clust = 0
Clust = 1
Clust = 2
Count
number_of_orders


Clust = 0
Clust = 1
Clust = 2
Count
0
200
400
600
0
500
1000
1500
2000
2500
revenue