**Team Assignment: Pharmaceutical Industry**

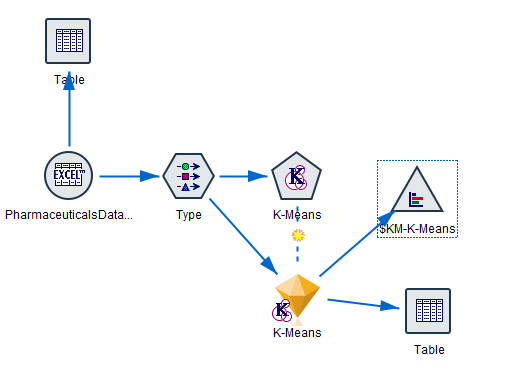
Use cluster analysis to explore and analyze the given dataset as follows:

1. Use the quantitative variables (1 to 9) to cluster the 21 firms. Justify the various choices made when conducting the cluster analysis

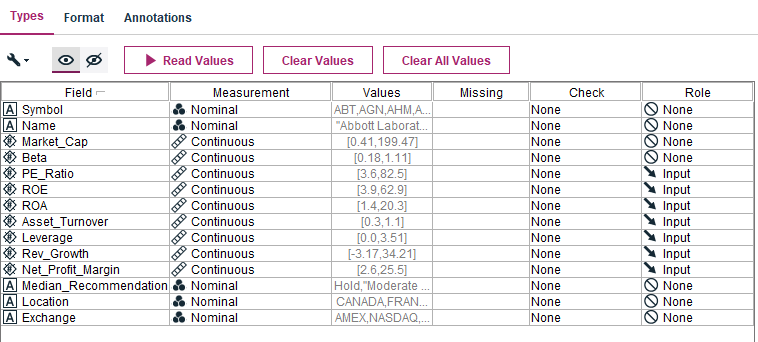
Tip: In this case, to understand the structure of the industry in financial terms we might use basic domain knowledge to consider that size and financial performance are fundamental dimensions into which our various metrics fall. To make sense of our analysis, we will need at least two clusters, but more than 3 or 4 will probably defeat the purpose of clustering

**NOTE: In order to have a uniform criterion for comparison and although you can run several algorithms and test them with different number of clusters, I would like you to focus on 4 clusters**

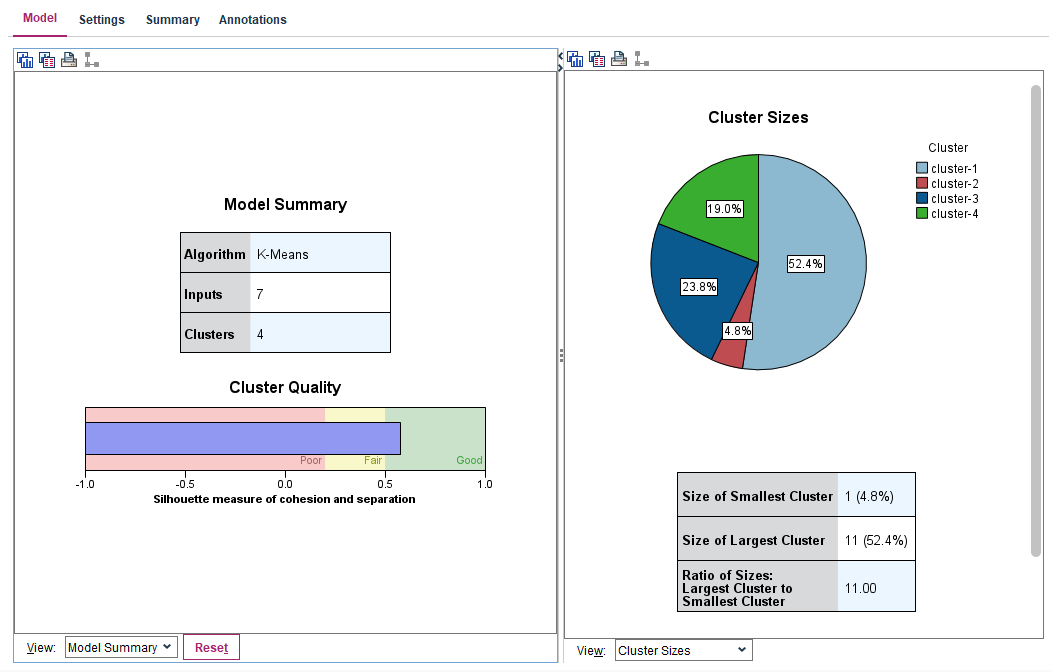
The following stream is created in SPSS modeler for clustering 21 firms in the given pharmaceuticalsdata.xlsx sheet



Below picture shows the quantitative variables chosen for clustering. 7 attributes-PE-Ratio, ROE, ROA, Asset\_Turnover, Leverage, Rev\_Growth and Net\_Profit\_Margin were considered in clustering as they appear to be significant.

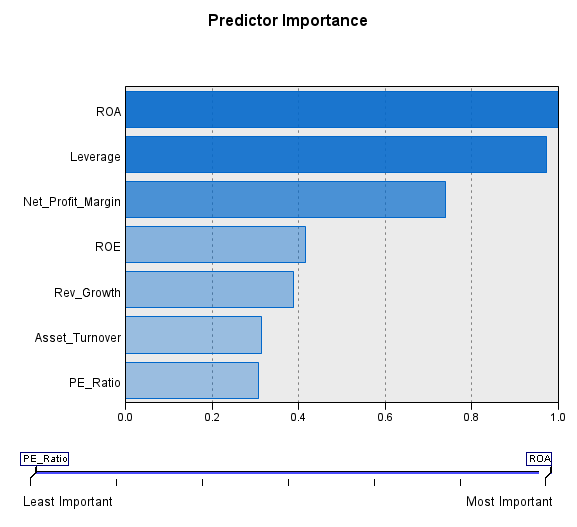


The following is the output when the stream is executed under K-means node.



The 4 clusters induced by 7 inputs gave relatively good cluster quality 0.6, which is silhouette measure of cohesion and separation. The pie chart shows the cluster distribution with 1 record as smallest cluster and 11 records as largest cluster bearing 52.4% of total records.

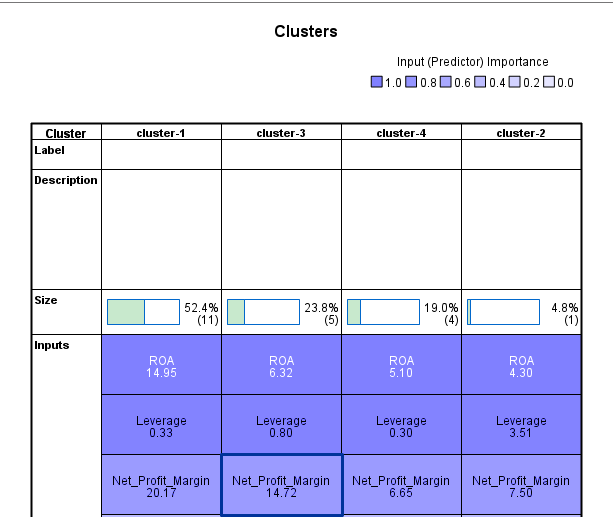
Below is the predictor importance histogram, shows that ROA, leverage and Net\_profit\_margin appear to be most relevant attributes.

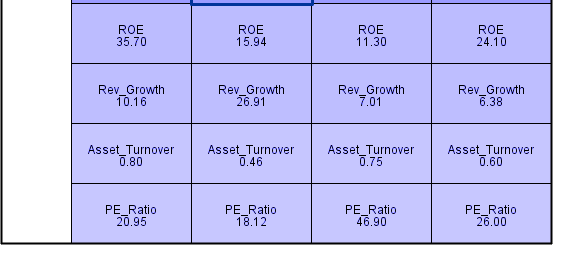


1. Interpret the clusters with respect to the quantitative variables that were used in forming the clusters

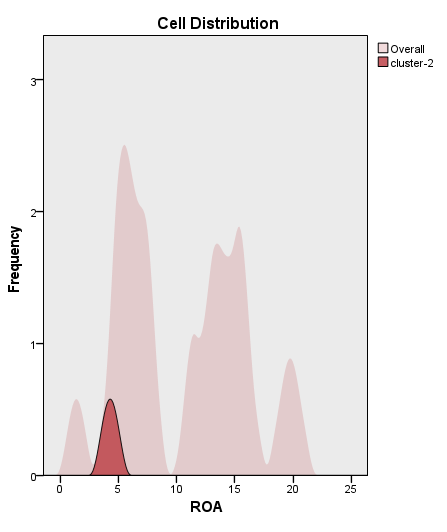
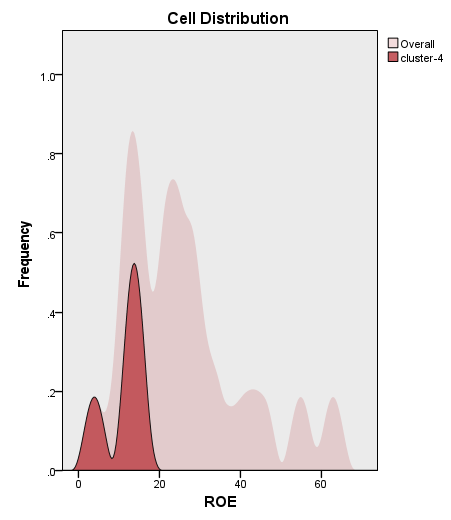
The figure shows 4 generated clusters with ROA, leverage and Net\_profit\_margin to be most relevant attributes.

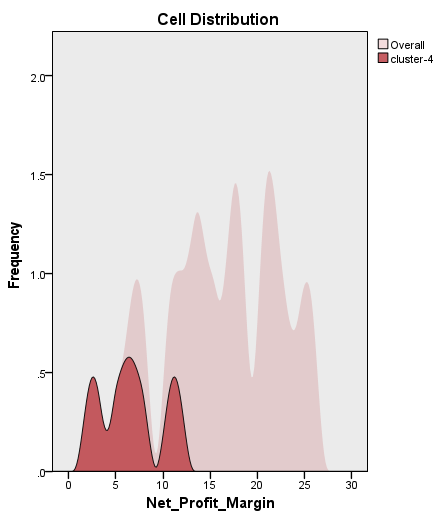
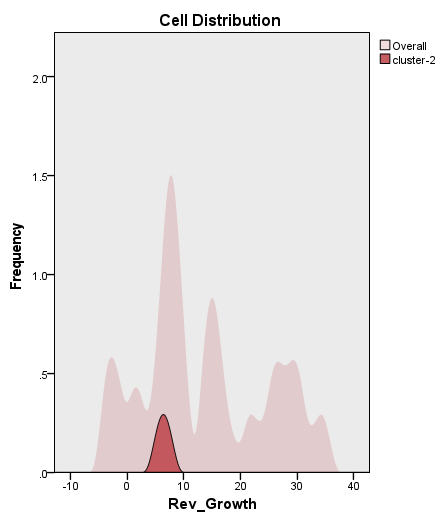
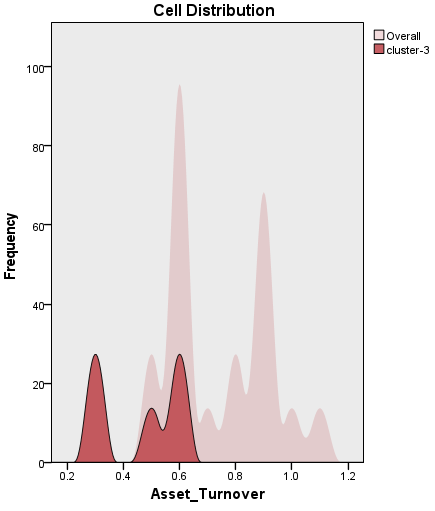
There seems to be a significant variation (not too much) of all the attributes except asset\_turover which is between 0.8 as highest and 0.46 as lowest value cluster. This significant variation shows that there is no too fitting in between the clusters and they are mostly well defined different from each other.



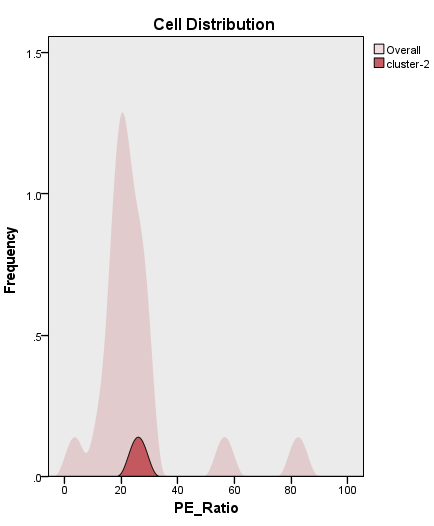
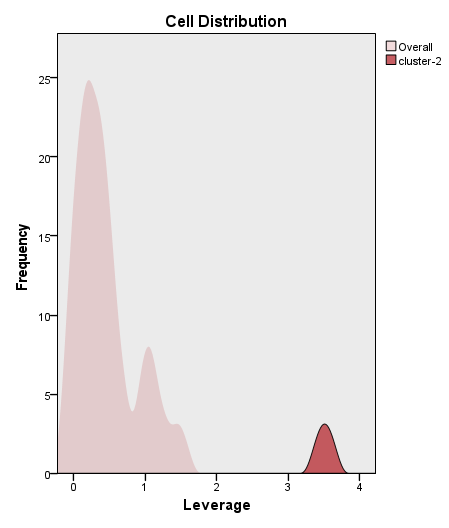


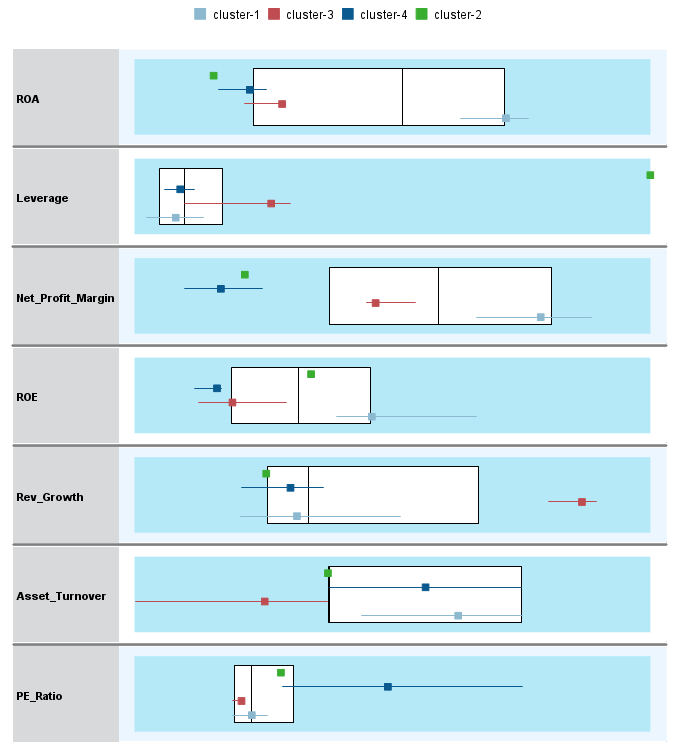
Let’s compare the lowest values against cell distribution. The ROA (with importance as 1) shows below graph of **minimum**, cluster 2 against frequency cell distribution of remaining clusters. Similarly, the remaining figures show the frequency distribution of all the attributes against **minimum** cluster size.

The below figures show the Leverage frequency distribution of maximum leverage (0.97 importance) of **maximum** of 3.51 of cluster 2 and PE\_ratio with **maximum** of 26.00.



The above figure shows the cluster comparison. We could see that there is no similarity among one attribute to another.

1. Is there a pattern in the clusters with respect to the qualitative variables?

There is no defined pattern in the clusters as from the comparison figure we could see that there is no particular tend followed by the clusters with respect to any attributes.

Also, the Cluster table, in answer b shows that the minimums and maximums are varied among the clusters and there is no particular order or trend in the clusters.

1. Provide an appropriate name for each cluster using any or all of the variables in the dataset

Renaming the cluster as – Cluster 1: 1\_ROA

Cluster 2: 4\_Leverage

Cluster 3: 2\_Rev\_Growth

Cluster 4: 3\_PE\_Ratio

The first part in the name denotes the order of size of the cluster and second part denotes the high attribute among other qualitative variabes.