**Pharmaceuticals**

**Tasks:**

Financial data gathered on 21 ﬁrms in the pharmaceutical industry are available in the ﬁle Pharmaceuticals.xls. For each ﬁrm, the following variables are recorded:

1. Market capitalization (in billions of dollars)

2. Beta

3. Price/earnings ratio

4. Return on equity

5. Return on assets

6. Asset turnover

7. Leverage

8. Estimated revenue growth

9. Net proﬁt margin

10. Median recommendation (across major brokerages)

11. Location of ﬁrm’s headquarters

12. Stock exchange on which the ﬁrm is listed

Preprocessing: We have preprocessed the data and found out that there are no missing values. This can be seen in the Imputation sheet.

**a.** Use only the quantitative variables (1 to 9) to cluster the 21 ﬁrms. Justify the various choices made in conducting the cluster analysis, such as weights accorded different variables, the speciﬁc clustering algorithm(s) used, the number of clusters formed, and so on.

We used K-means clustering algorithm for the dataset. We analyzed the data by performing different numbers of clusters in the algorithm, ended up by deciding to choose 4 clusters Random start was used instead of fixed start. We initially ran K-means for 3,4,5 and 6 clusters and observed the results of all the clusters. From the results we have found out that, optimal number of clusters for the data is 4. We can say this from the sheets KMC\_Output\_3, KMC\_Output\_4, KMC\_Output\_5, KMC\_Output\_6. When we change the number of clusters from 3 to 4 we can see that there is meaningful division of clusters, whereas when we change from 4 to 5 or 6 we can see that there is no meaningful division (just one or two clusters are divided into another cluster). We can also see that when we change from 4 to 5 the inter-cluster distances between 1 and 4, 5 and 4(5 clusters sheet) is very low which further can be combined into a single one. So, we have chosen the number of clusters as 4.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

We can see from the above pictures that there is meaning division of clusters from 3 clusters to 4 clusters. Whereas we do not see any meaningful division from 4 to 5 and 5 to 6. From 4 to 5 clusters we can see that only 2 clusters are affected by addition of new cluster and also in 6 clusters we can see that the sixth cluster has 0 records, so we have 6th cluster which is of no use. So we have chosen 4 clusters which is best for this data.

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

Cluster 1- 1,4,7,10,16,18,19,21

Cluster 2- 11,15

Cluster 3- 13,17

Cluster 4- 2,3,6,8,9,12,14,20

We can see this in the sheet KMC\_Clusters\_4.

A screenshot of a cell phone

Description automatically generated

**c.** Is there a pattern in the clusters with respect to the qualitative variables (10 to 12) (those not used in forming the clusters)?

Following are the median recommendation that we have for all the clusters. We can get the median recommendations from the sheet data in the workbook.

**Cluster 1**

Moderate buy - ABT

Moderate Sell - AZN, BMY

Hold - LLY, NVS, PHA, SGP, WYE

**Cluster 2**

Hold - GSK, MRK

**Cluster 3**

Moderate Buy - JNJ, PFE

**Cluster 4**

Moderate Buy - AGN, CHTT, MRX

Strong Buy - AHM

Hold - BAY, IVX,

Moderate Sell - ELN, WPI

We see that there are patterns in the clusters with respect to the media recommendations. Cluster 2 which is highest at ROE, Assest\_Turnover and Rev\_Growth has all Hold recommendation. Cluster 3 which has the highest Market\_Cap, ROA and Net\_Profit\_Margin has only Moderate\_Buy recommendation.Cluster 4 has the highest Beta, PE\_Ratio and Leverage, it has mixed recommendation but the highest for Moderate Buy. Cluster 1 which is a moderate cluster in all, mostly has Hold recommendation.

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

Cluster 3 is considered to be High Scale as it has high Market Cap value, low beta(risk) value and high net profit margin. Cluster 4 is considered as under performing cluster as it has the least market cap value, highest beta value and low net profit margin. Cluster 2 has the highest ROE, Asset\_Turnover and Rev\_Growth.

Cluster 1 has every variable moderate as compared to the other clusters.

A screenshot of a social media post

Description automatically generated