Assignment_myusuf2_4_Summary

Fundamentals of Machine Learning-Assignment-4

Clustering- Financial data on 21 firms in the pharmaceutical industry

Problem Statement

An equities analyst is studying the pharmaceutical industry and would like my help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data was gathered on 21 firms in the pharmaceutical industry. The 12 variables are;

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. Of these variables, 9 are numerical and 5 are categorical.

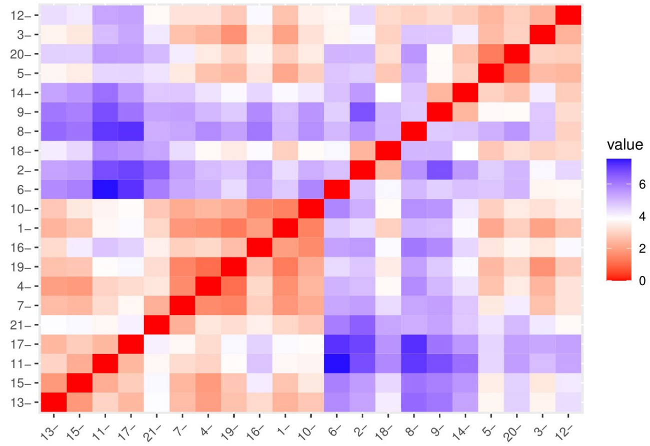
Data Preparation

I loaded my dataset to R data frame in the form of Pharmaceuticals.csv

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

A1. Action-Solution:

I first tidy the dataset by pivoting the financial longer under one column as "financial", then scaled the select numerical variables in the data frame as depicted in the knitted R;

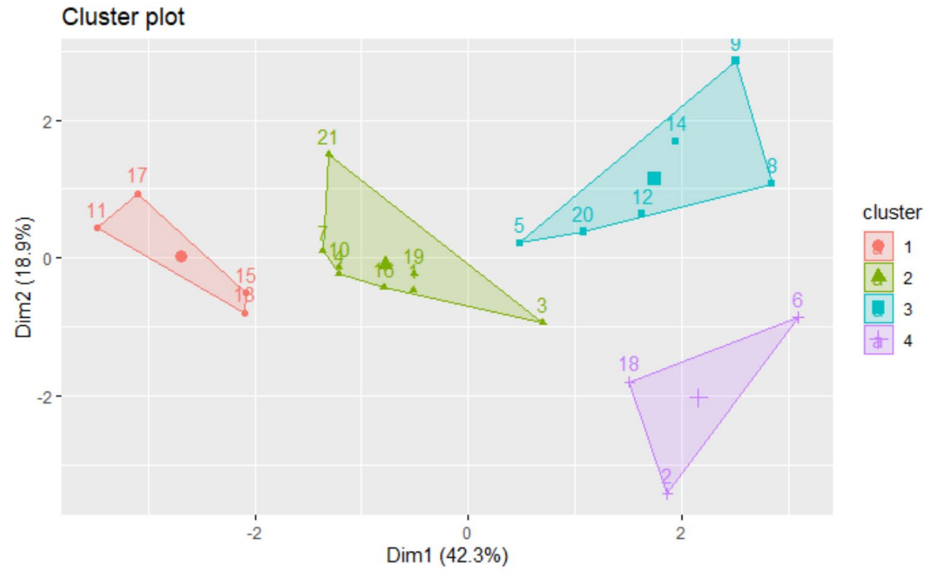


K-means (KNN) was employed in clustering the variables in a way that one iteration converged when we have no group center. The center changed many times because we ran the algorithms at least 25 times which enable us to view where it converged. I generated several averages of the 9 select groups (centroid).

QB. Interpret the clusters with respect to the numerical variables used in forming the clusters.

Action-Solution:

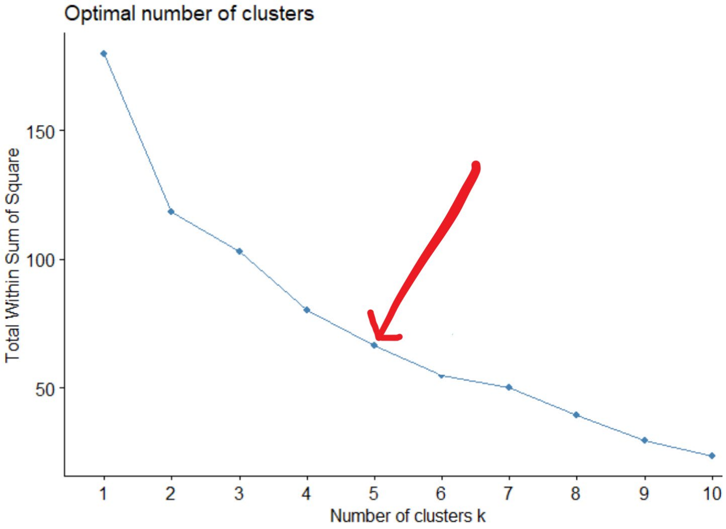
Clustering, which is unsupervised learning, is similar to KNN algorithms is used to measure the distance to identify the nearest neighbor. To do that, I used the select numerical values "3,4,5,6,7,8,9,10,11" to form 4 clusters of 21 firms, visualized as;



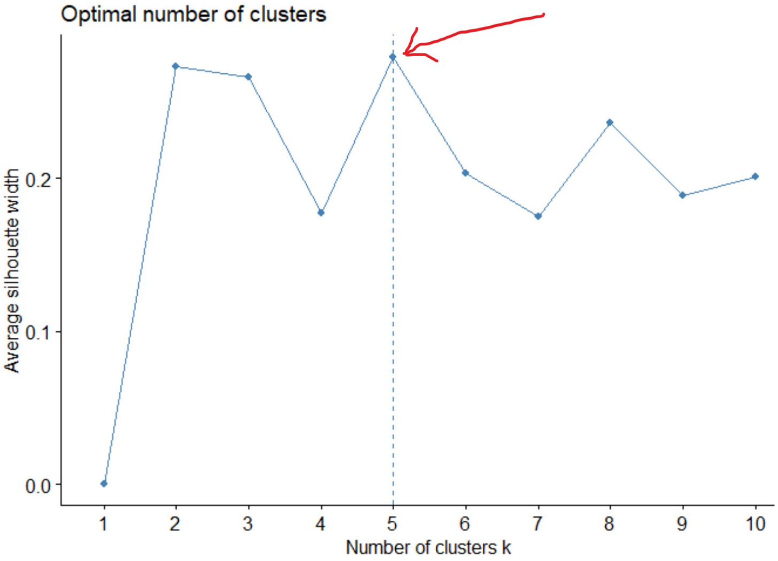
I employed these relevant R libraries that enabled the system to generate the output; libraries (tidyr), (dplyr), (tidyverse), (factoextra), and (flexclust). The output indicates the sizes of the cluster as 4, 8, 6, and 3 respectively. Cluster of 4 indicates that the data in rows (11, 17, 15, and 13) are the closest, rows (21, 7, 10, 16, 19, and 3) are the closest neighbors.

Using the C-bind, I was able to add a column to the original dataset which comprises assigned clusters 1 to 4 to each of the pharmaceutical industry.

Furthermore, I applied the Elbow method to ascertain the total intra-cluster known as the total within-cluster sum of the square to ensure that the variation is minimized. The decline in heterogeneity as we add cluster indicates cluster k at point number 5 in X-axis as the knee point. This means that k=5 gives us the best value between bias and overfitting.



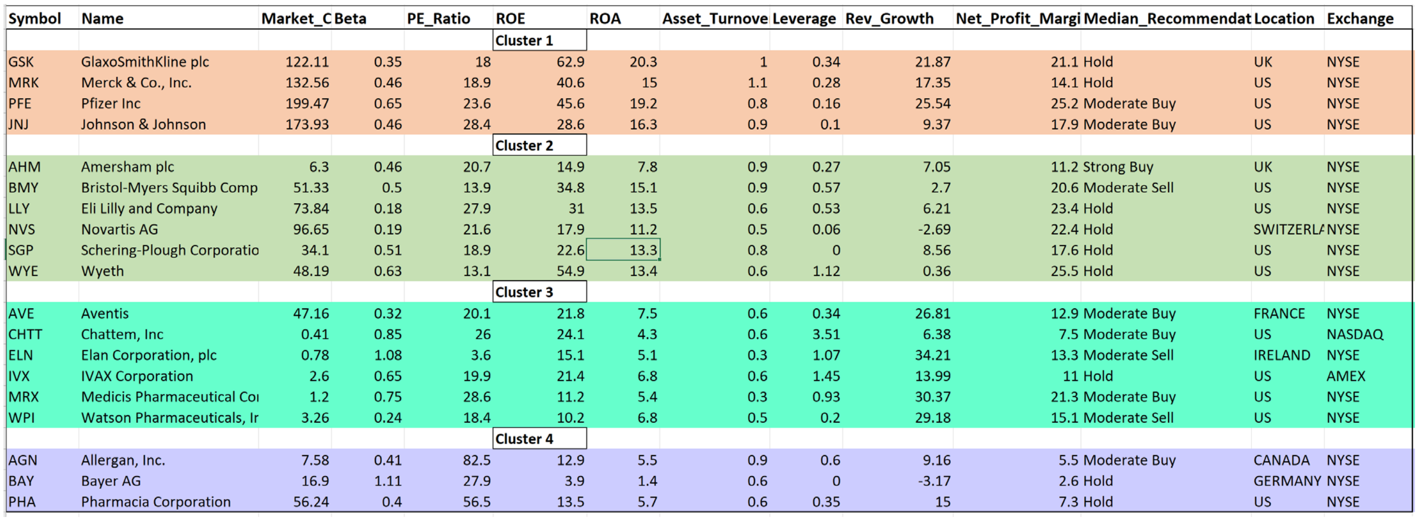
I then used an Average Silhouette Method to ascertain that the assignment of records to the cluster is valid and consistent. The output confirms that 5 is the most ideal number to cluster for it has the largest value of Silhouette Width on the Y-axis.



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

Action-Solution:

variables 10 to 12 are the Median_Recommendation and Location in the dataset. Here is my logical analysis;



Symbol	Name	Market_C	Beta	PE_Ratio	ROE	ROA	Asset_Turnove	Leverage	Rev_Growth	Net_Profit_Margi	Median_Recommendat	Location	Exchange
					Cluster 1								
GSK	GlaxoSmithKline plc	122.11	0.35	18	62.9	20.3	1	0.34	21.87	21.1	Hold	UK	NYSE
MRK	Merck & Co., Inc.	132.56	0.46	18.9	40.6	15	1.1	0.28	17.35	14.1	Hold	US	NYSE
PFE	Pfizer Inc	199.47	0.65	23.6	45.6	19.2	0.8	0.16	25.54	25.2	Moderate Buy	US	NYSE
JNJ	Johnson & Johnson	173.93	0.46	28.4	28.6	16.3	0.9	0.1	9.37	17.9	Moderate Buy	US	NYSE
					Cluster 2								
AHM	Amersham plc	6.3	0.46	20.7	14.9	7.8	0.9	0.27	7.05	11.2	Strong Buy	UK	NYSE
BMY	Bristol-Myers Squibb Comp	51.33	0.5	13.9	34.8	15.1	0.9	0.57	2.7	20.6	Moderate Sell	US	NYSE
LLY	Eli Lilly and Company	73.84	0.18	27.9	31	13.5	0.6	0.53	6.21	23.4	Hold	US	NYSE
NVS	Novartis AG	96.65	0.19	21.6	17.9	11.2	0.5	0.06	-2.69	22.4	Hold	SWITZERL/	NYSE
SGP	Schering-Plough Corporatio	34.1	0.51	18.9	22.6	13.3	0.8	0	8.56	17.6	Hold	US	NYSE
WYE	Wyeth	48.19	0.63	13.1	54.9	13.4	0.6	1.12	0.36	25.5	Hold	US	NYSE
					Cluster 3								
AVE	Aventis	47.16	0.32	20.1	21.8	7.5	0.6	0.34	26.81	12.9	Moderate Buy	FRANCE	NYSE
CHTT	Chattem, Inc	0.41	0.85	26	24.1	4.3	0.6	3.51	6.38	7.5	Moderate Buy	US	NASDAQ
ELN	Elan Corporation, plc	0.78	1.08	3.6	15.1	5.1	0.3	1.07	34.21	13.3	Moderate Sell	IRELAND	NYSE
IVX	IVAX Corporation	2.6	0.65	19.9	21.4	6.8	0.6	1.45	13.99	11	Hold	US	AMEX
MRX	Medicis Pharmaceutical Coı	1.2	0.75	28.6	11.2	5.4	0.3	0.93	30.37	21.3	Moderate Buy	US	NYSE
WPI	Watson Pharmaceuticals, Ir	3.26	0.24	18.4	10.2	6.8	0.5	0.2	29.18	15.1	Moderate Sell	US	NYSE
					Cluster 4								
AGN	Allergan, Inc.	7.58	0.41	82.5	12.9	5.5	0.9	0.6	9.16	5.5	Moderate Buy	CANADA	NYSE
BAY	Bayer AG	16.9	1.11	27.9	3.9	1.4	0.6	0	-3.17	2.6	Hold	GERMANY	NYSE
PHA	Pharmacia Corporation	56.24	0.4	56.5	13.5	5.7	0.6	0.35	15	7.3	Hold	US	NYSE

Cluster 1 comprises related Median_Recommendation that are mostly categorized as Hold with an insignificant number of Strong Buy and Moderate Sell with 75% located in the US. This indicates a high correlation between those companies with their Headquarters. located in the US and Cluster 1.

Cluster 2 basically comprises of related Median_Recommendation of equally divided Hold and Moderate buy with 67% of their Hqts. located in the US and the rest in UK and Switzerland

Cluster 3 basically comprises related Median_Recommendation that are mostly Moderate Buy and Sell with equally divided companies with their Hqts. located in the US and the three different countries in the UK.

Cluster 4 although relatively small in size, comprises of related Median_Recommendation that are mostly Hold whose Hqts. are located in three different countries namely; Canada, Germany, and the US, respectively.

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

Action-Solution:

based on the problem statement and the pivot_longer column that I named in my tidy data, I would assign each cluster as follows;

Cluster1: financial_structure_1

Cluster1: financial_structure_2

Cluster1: financial_structure_3

Cluster1: financial_structure_3