DATA MINING – GROUP PROJECT

PATENT DATA SET
CBA Batch 6

TEAM MEMBERS:

SHANKHA SUBHRA GHOSH -71610075
SUBBA REDDY YERUVA-71610085
SURAJIT DHAR -71610091
VANEESH NARAYANAN -71610099
ARJUN TEWARI -71610008
MANISH KUMAR -71610036

BUSINESS PROBLEM/EXECUTIVE SUMMARY

A leading tech-focused investment firm has undertaken a research project through which it wishes to identity potential tech companies in which it can invest either through private equity investments or by purchasing publicly traded shares.

The company wants to invest in companies which are high on innovation, has good financial health and has high growth potential. The company is looking to maximize its ROI, however, the company is also cautious in picking the right portfolio which minimizes risk that is associated with equity investments.

DATA DESCRIPTION

The investment firm has curated a dataset which has information such as sales revenue, net income, total asset value and R&D expenditure for about 1645 companies across various industries for consecutive years (2003, 2004 and 2005).

Additionally, in the same dataset, the investment firm has also captured information on patents that were granted to these 1645 companies in year of 2004. The dataset in all has 42499 rows of data capturing the above information.

This dataset is handed over to a team of analysts who are expected analyse the data and come up with insights based on which the management team of the firm can make their investment decision.

ANALYSIS PROCESS

As a first step, a quick analysis was done to identity top performing industries by average net income. It was found that petroleum refining as an industry had the highest average net income while computer programming”-data processing as an industry was ranked 5th by average net income. Packaged-Software industry was ranked 7th in the list by average net income.

INDUSTRY	INDUSTRY CODE	AVG. NI - 2005
PETROLEUM REFINING	2911	15664
CONGLOMERATES	9997	11622
COMMERCIAL BANKS	6020	11095
RETAIL-VARIETY STORES	5331	10267
SERVICES-COMPUTER PROGRAMMING, DATA PROCESSING, ETC	7370	8165
SOAP, DETERGENTS, CLEANG PREPARATIONS, PERFUMES, COSMETICS	2840	6615
SERVICES – PACKAGED SOFTWARE	7372	5616
BEVERAGES	2080	4472
PHARMACEUTICAL PREPARATIONS	2834	3575
FINANCE SERVICES	6199	3445

TABLE1: LIST OF TOP 10 INDUSTRIES BY AVERAGE NET INCOME - 2005

IDENTIFICATION OF DATA SUBSET AND INITIAL EXPLORATION

As the investment firm is tech-focused, packaged-software industry (Sic – 7372) was chosen for the next level of analysis. As a next step, data was isolated for the packaged-software industry which had 88 unique companies and 1200 rows capturing various patent related attributes of these companies.

To evaluate whether there is a correlation between R&D Expenses and Total Assets of a company, a scatter plot was drawn with "Total Assets" in X axis and "R&D Expenses" in Y axis. On zooming-in on the plot, as clear linear relationship could be seen (graph below).

While, the graph showed a positive correlation between R&D Expenses and Total Assets of a company, the causation could not be established. While it could have been assumed that more R&D expenditure means more patent acquisition and thus higher asset value. However, that would have been a wrong assumption as the R&D expenditure and Total Assets data are for the same year and it is known it usually take 3-4 years for patents to be granted.

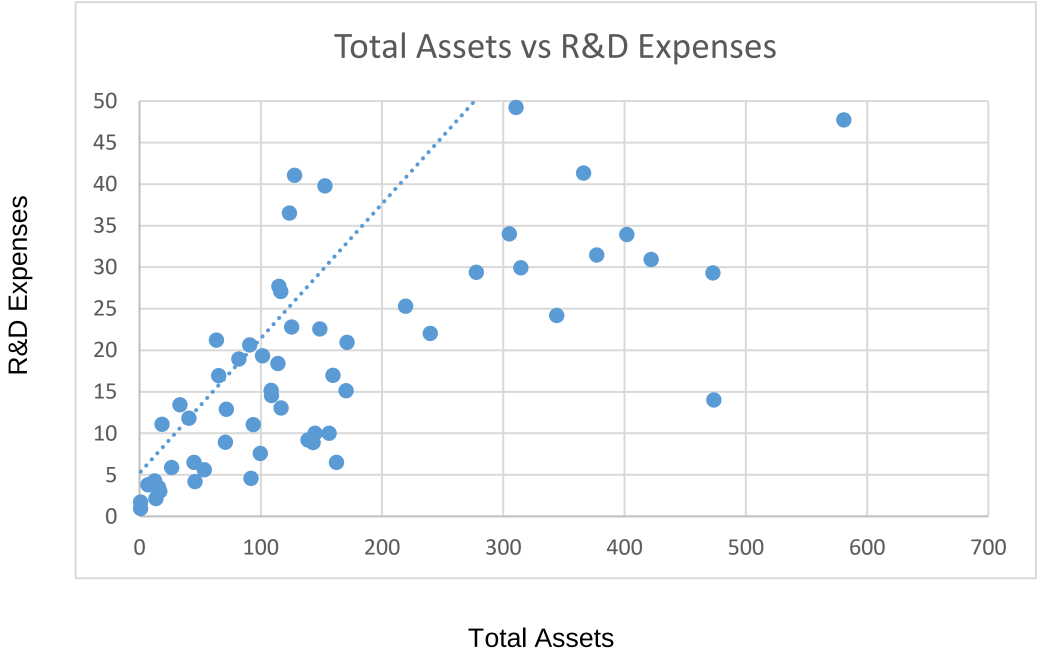


FIG 1: R&D EXPENDITURE VS. TOTAL ASSETS – STRONG LINEAR CORRELATION

Patents granted in 2004 vs R&D expenditure for (Gyear – 1) were also plotted against each other. However, no correlation was seen, further strengthening the insight that R&D expenditure will not generate immediate patents (graph below).

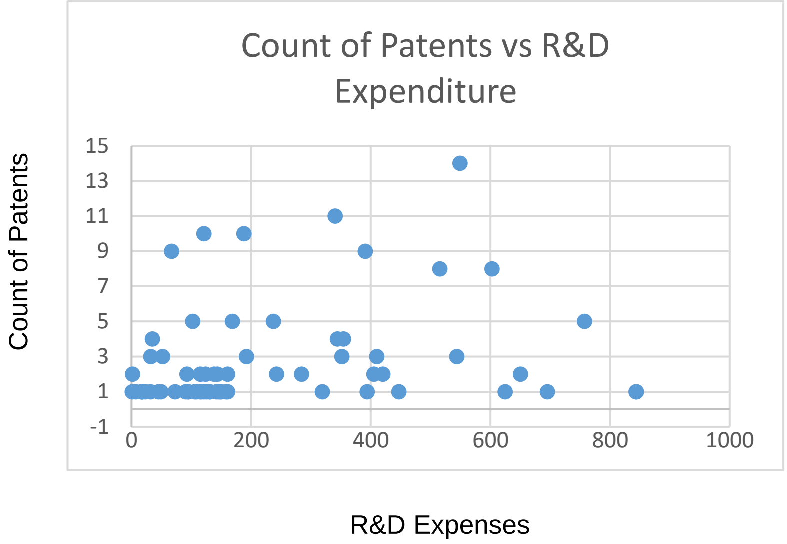


FIG 2: R&D EXPENDITURE VS. COUNT OF PATENTS – NO CORRELATION

CLASSIFICATION OF COMPANIES BY K-MEANS CLUSTERING

As a next step, K-Means clustering was run on the data set with K=4. Following were the outcome of the clustering.

Cluster Centers

Cluster	at_1	at0	at1	ni_1	ni0	ni1	sale_1	sale0	sale1	xrd_1	xrd0	xrd1
Cluster-1	79,571	92,389	70,815	9,993	8,168	11,513	32,187	36,835	39,253	4,659	7,779	6,184
Cluster-2	2,306	2,729	2,909	(1)	102	187	1,272	1,284	1,435	318	317	336
Cluster-3	11,064	12,763	20,607	2,307	2,681	2,884	9,475	10,156	11,782	1,180	1,278	1,537
Cluster-4	337	353	383	(114)	(7)	26	241	235	268	50	47	49

Distance Between Centers	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Cluster-1	0	150571.8	125603.3	154947.8
Cluster-2	150571.8	0	27677.56	4438.687
Cluster-3	125603.3	27677.56	0	31947.05
Cluster-4	154947.8	4438.687	31947.05	0

Data Summary

Cluster	#Obs	Avg. Dist
Cluster-1	629	1496.205
Cluster-2	175	1675.099
Cluster-3	99	161.8862
Cluster-4	297	703.5194
Overall	1200	1216.023

Ideal K-Means, elbow curve:

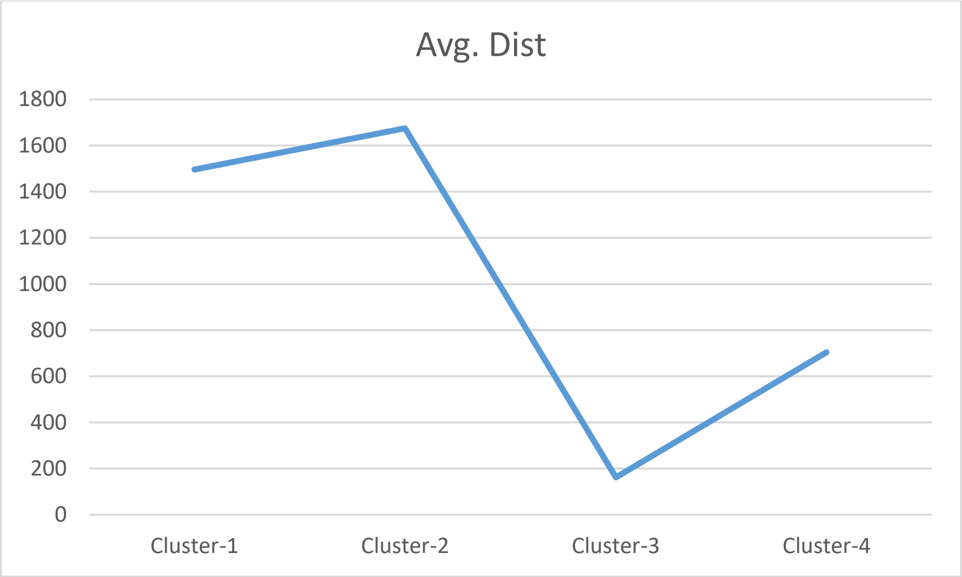


FIG 3: ELBOW CURVE

Going by the elbow curve, ideal K seems to be 2, i.e. 2 clusters.

However, on running K-means with K=2, we get one cluster which is Microsoft with 629 records. All other companies are clubbed in 2nd cluster. Thus 2 clusters are not being able to capture the entire characteristics of the companies.

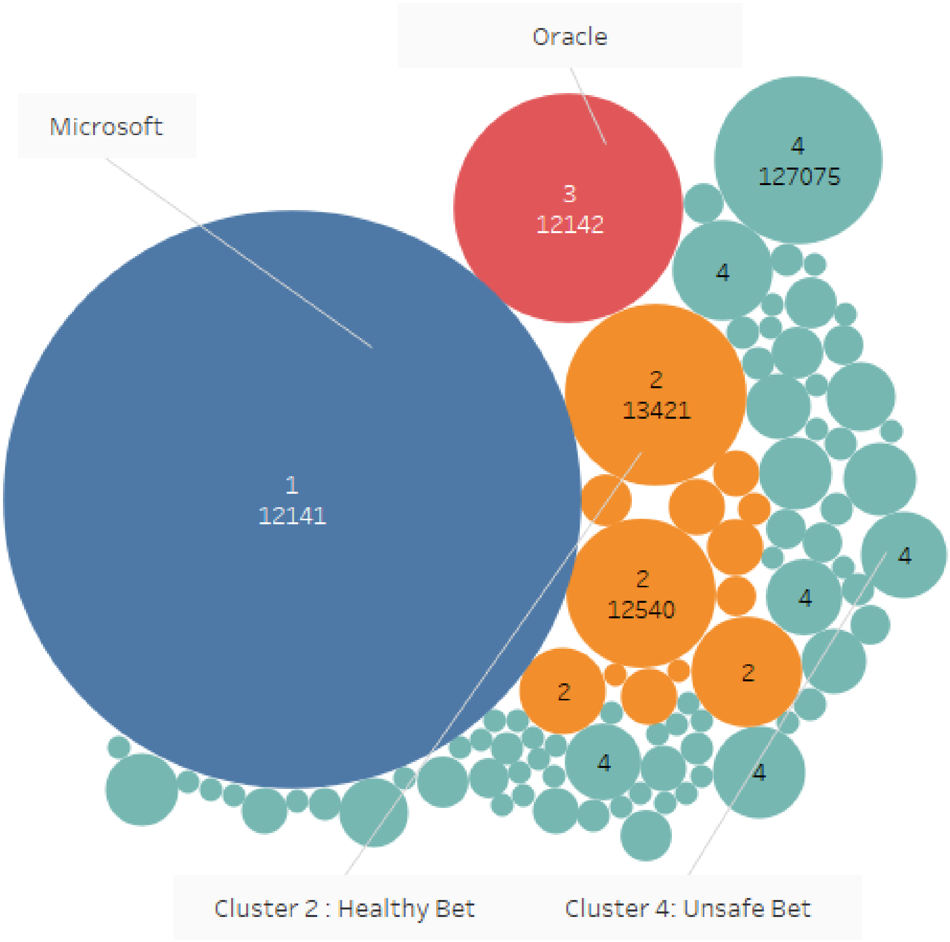
Data Summary – K=2

Original coordinates		
Cluster	#Obs	Avg. Dist
Cluster-1	629	1496.205
Cluster-2	571	9023.339
Overall	1200	5077.866

Normalized coordinates		
Cluster	#Obs	Avg. Dist
Cluster-1	629	0.220393
Cluster-2	571	0.512515
Overall	1200	0.359394

Visual Representation of 4 Clusters:

7372(SIC) - 4 clusters



Cluster ID and Gvkey. Color shows details about Cluster ID. Size shows sum of Number of Records. The marks are labeled by Cluster ID and Gvkey.

FIG 4: PICTORIAL REPRESENTATION OF CLUSTERS

	RND Intensity		Innovation Efficiency	
	Min	Max	Min	Max
Cluster 1	17%	17%	10%	
Cluster 2	5%	35%	0%	17%
Cluster 3	13%	13%	7%	7%
Cluster 4	0%	99%	1%	697%

R&D Intensity = R&D spend/Sales

Innovation Efficiency = R&D spend/ Total No. of Patents

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) was run on the 12 variables i.e. all operational metrics (sales revenue, net income, total asset value and R&D expenditure for 3 years).

We found that PC1 component explained 99.7% of the variability.

Variances												
	1	2	3	4	5	6	7	8	9	10	11	12
Variance	11.94	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Variance Percentage	99.52	0.32	0.12	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cumulative Variance %	99.52	99.84	99.96	99.99	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Also, we ran K-means clustering with the PC1 values. We got 3 clusters which was similar to our findings without using PCA.

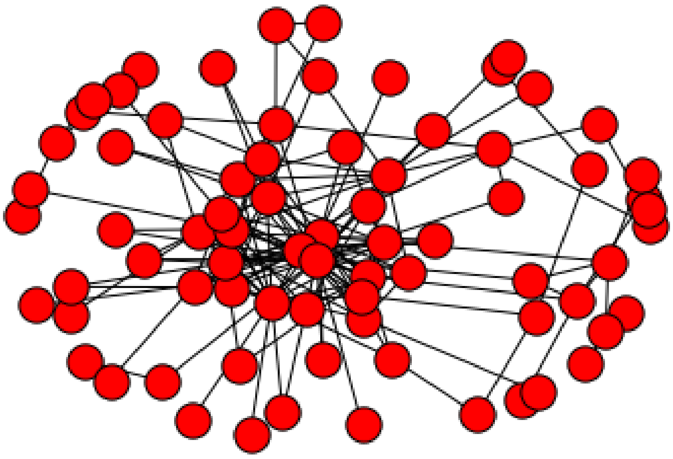
Cluster 1: Microsoft
Cluster 2: Other companies
Cluster 3: Oracle
Cluster 4: no records

Data Summary:

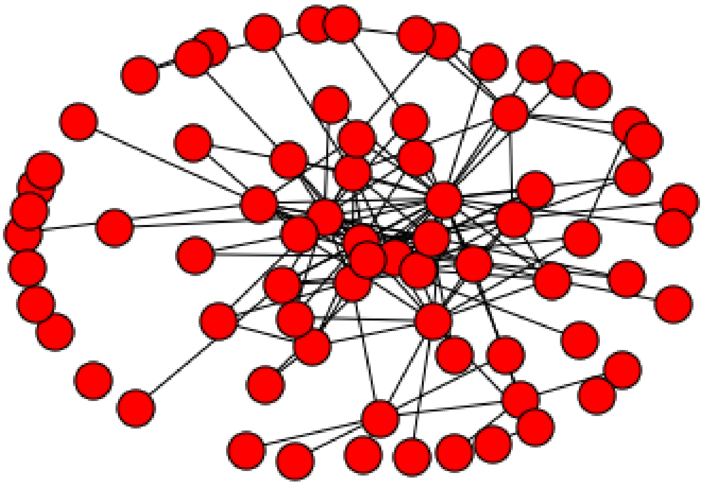
Cluster	#Obs	Avg. Dist
Cluster-1	629	0.074894
Cluster-2	472	0.107259
Cluster-3	99	0.002089
Cluster-4	0	N/A
Overall	1200	0.081618

Since PC1 explained 99.5% of variability, plotting PC1 and PC2 in 2 dimensions also was not expected to give significant insights.

NETWORK ANALYSIS



The above is a network connectivity of patents for all industries.



The above network is without Packaged-Software industry, which has one of the highest patents – the network graph looks slightly dis-jointed in comparison to the 1st network graph

Degree of Centrality of Nodes (All SICs):

Top 10 degree of centrality

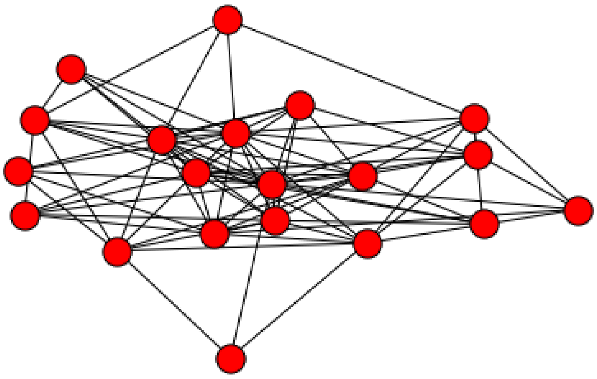
```
[(3812, 0.14864864864864866),
(3825, 0.14864864864864866),
(3576, 0.14864864864864866),
(3714, 0.17567567567567569),
(3663, 0.1891891891891892),
(9997, 0.20270270270270271),
(3570, 0.22972972972972974),
(7372, 0.22972972972972974),
(7370, 0.2567567567567568),
(3674, 0.35135135135135137)]
```

Between-ness Centrality:

```
[(3570, 0.0542880368827186),
(9997, 0.06872130913226807),
(3572, 0.07037692406346717),
(4813, 0.07238201375993476),
(3825, 0.08323871143170096),
(1389, 0.0839309989793471),
(7370, 0.09011209339573563),
(7372, 0.09705365813343252),
(3714, 0.11466604647507146),
(3674, 0.30268244802894273)]
```

Of the all the industries, Packaged-Software industry has 3rd highest degree of centrality and 3rd highest between-ness centrality.

Top 20 Degree Centralities:



```
[6172,
3600,
3572,
1389,
3661,
3711,
3571,
7373,
3728,
4813,
3812,
3825,
3576,
3714,
3663,
9997,
3570,
7372,
7370,
3674]
```

In top 20 degree centralities, packaged-software in no. 3.

CONCLUSION

In the analysis of Packaged-Software industry, two companies' that standout distinctively – Microsoft and Oracle (Cluster-1 & Clsuter-3 respectively) due to their patents as well operational metrics.

These two companies are the safest options for any investment opportunity.

However, there are also other companies who has high potential for great ROI. Companies in cluster-2 such as Adobe, BMC, Symantec, CA Inc., Electronic Arts, Intuit, etc. has good R&D intensity and Innovation efficiency. These could be potential for higher returns and are moderate risk.

Cluster-4, however, has high variance in both innovation efficiency and R&D intensity and thus companies in this cluster would be risky investment. Most companies in cluster-4 are found to have un-impressive sales figures YOY.

Network analysis supports that Packaged-Software industry has high degree of centrality, between-ness, closeness and thus is a key industry and should be explored for investments.

This is a preliminary analysis based on limited data and further study is required with more recent data covering longer duration.
