Project Overview

With so many companies around the world how can we find similarities or dissimilarities between different companies? Moreover, on what aspects should we look when we are comparing different companies.

if we will have smart and not so straightforward segmentation on company levels it could bring a lot of value to the table, for example acquisition and merging decision, trends, competitors and potential collaboration. In this project I will take as a case study the fortune 500 dataset and create cluster analysis.

The Fortune 500 is an annual list compiled and published by Fortune magazine that ranks 500 of the largest United States corporations by total revenue for their respective fiscal years. The list includes publicly held companies, along with privately held companies for which revenues are publicly available. The concept of the Fortune 500 was created by Edgar P. Smith, a Fortune editor, and the first list was published in 1955. (wiki)

The Fortune 500 can be used to gauge the health of the overall U.S. economy. When many companies of a sector are removed from the list, it may signal weakness in that sector.

Problem statement

The goal is to create clustering analysis for this dataset. the tasks involved are the following:

1. Import the fortune 500 dataset
2. Enrich the dataset
3. Analyze the data
4. Preprocess the data
5. Create a clustering analysis using unsupervised method.
6. Validate the results and estimate the feature importance
7. Get insights from the clustering results

The final clusters labels should serve as input to analyze companies on variety of aspects

Metrics

The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters

a(i) average distance between (i) and all other data within the same cluster

b(i) smallest average distance of (i) to all points in any other cluster

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Data description

The fortune 500 original dataset contains different attributes on company level:

- Company Name - Name of the company (string)
- Website address - Website address of the company (string)
- Sector - (string)
- Industry (string)
- Hqlocation- headquarter location (string)
- Hqaddr- headquarter address (string)
- Hqcity - headquarter city (string)
- Hqstate - headquarter state (string)
- Hqtel- headquarter telephone (string)
- Hqzip - headquarter zip code (string)
- CEO – CEO full name (string)
- CEO Title - (string)
- Ticker – ticker of the equity (string)
- Full name (string)- Full name of the company
- Address - full address of the company (string)
- Number of Employees - Total number of employees in the company (int)
- Revenues - Revenue of the company for the year 2016-17 in $ millions. (float)
- Revenue Change - Percentage of Revenue change from last year (float)
- Profits - Profits of the company in $ million (float)
- Profit Change - Change in the percentage of profit from previous year. (float)
- Assets - Value of assets in $ millions (float)

Data enrichment

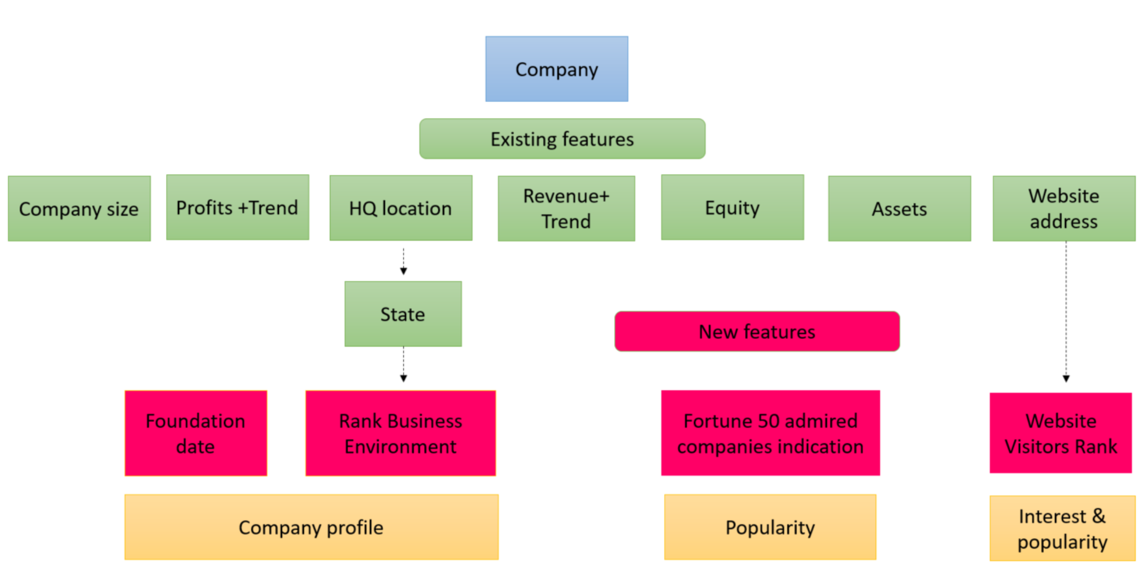
In addition to this original dataset I decided to put more spicy features:

1. Company foundation date
2. Head quarter location- I ranked each state according to Business environment measures:

Business Environment Rank	State	Entrepreneurship	Low Tax Burden	Patent Creation	Top Company Headquarters	Venture Capital
#1	California	4	46	1	24	1
#2	Massachusetts	13	38	2	13	2
#3	Colorado	8	18	9	12	6
#4	Washington	31	16	3	30	5
#5	Utah	7	28	13	41	4

*https://www.usnews.com/news/best-states/rankings/economy/business-environment

3. Popularity
   a. Indication if company was on list of most admired companies
   b. Website visitors – SimilarWeb
4. CEO popularity- Fortune CEO list
5. Profit/Revenue ratio



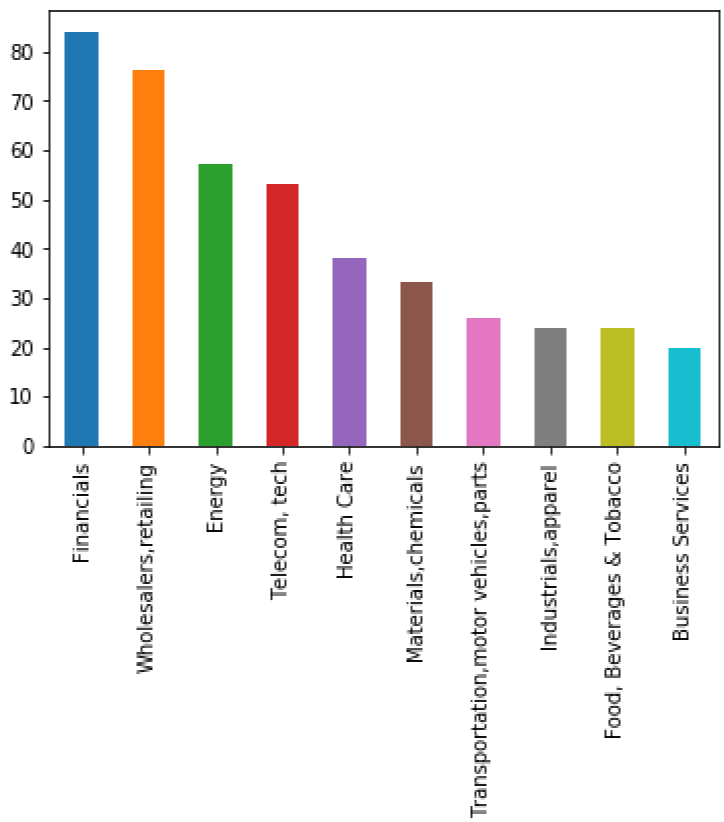
Data Analysis

With the new dataset I preformed deep analysis:

Sector:

The graph below shows the top 10 sectors.

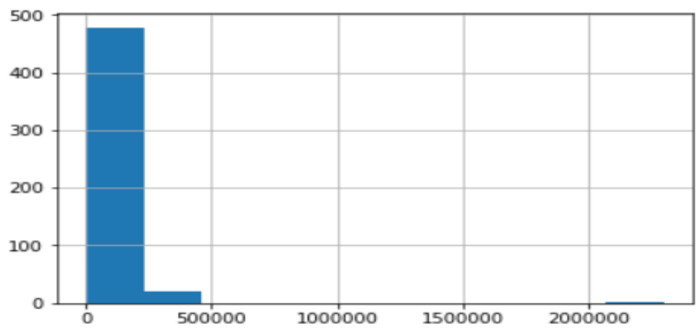
- We can see that most of the companies are in the finanical sector.
- We can also say that approximately 60% of the companies are in these 5 sectors: Financial, Energy, Retailing, Tech and Healthcare.



Number of employees:

- We can see that the average number of employees are 56K
- There is 1 outlier with 2.3M employees which belong to the number 1 rank in Fortune 500 -Walmart

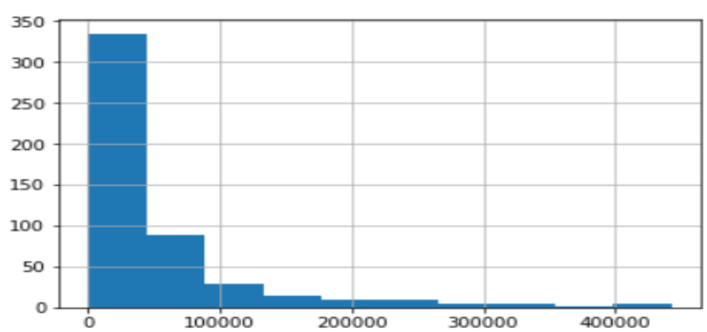
```
: count          500.000000
  mean         56350.132000
  std         123452.025921
  min             83.000000
  25%          11900.000000
  50%          25000.000000
  75%          56825.250000
  max        2300000.000000
  Name: Employees, dtype: object
```



If we are ignoring this outlier in terms of employees, we can see the following distribution:

Most of the companies are between 0-5000 employees

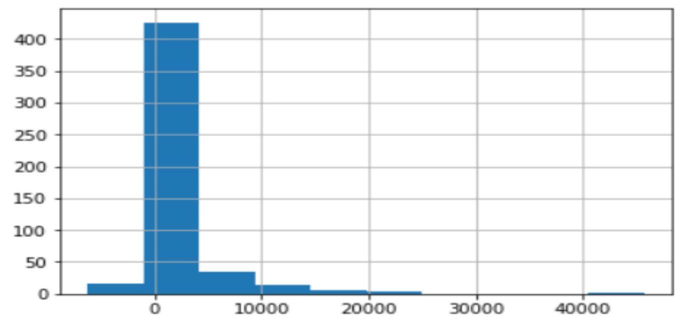
```
count         499.000000
mean        51853.839679
std         71710.429995
min            83.000000
25%         11900.000000
50%         25000.000000
75%         56583.500000
max        443000.000000
Name: Employees, dtype: object
```



Profits: small amount of companies are with negative profits

The most profitable company 45,687 is Apple which ranked in the third place

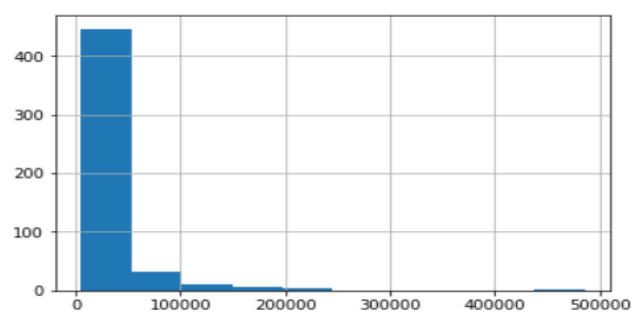
```
count      500.000000
mean      1779.479800
std       3937.558721
min      -6177.000000
25%        235.725000
50%        683.600000
75%       1770.775000
max      45687.000000
Name: Profits, dtype: float64
```



Revenue:

The ranking of the fortune 500 is according to Revenue we can see the big gap between the number 1 place and the second one - 260K

```
count       500.000000
mean      24111.748000
std       38337.353337
min        5145.000000
25%        7245.000000
50%       11384.000000
75%       22605.250000
max      485873.000000
Name: Revenues, dtype: float64
```

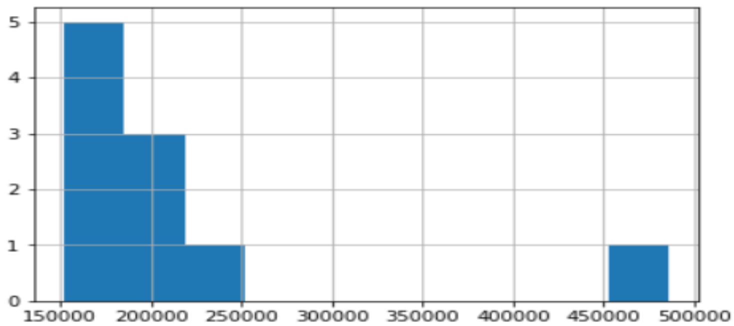


Top 10 vs other:

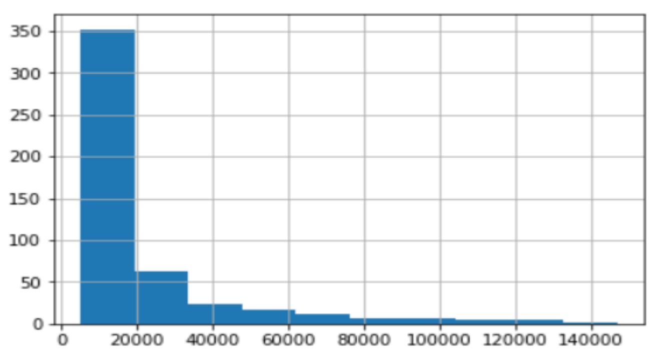
We can see that the average in the top 10 companies is around 216K Vs 23K in all the others.

We can also see that most of companies ~ 350 have revenues between 0-20K

```
count        10.000000
mean     216693.900000
std       97365.078129
min      151800.000000
25%      169166.500000
50%      188663.500000
75%      212980.250000
max      485873.000000
Name: Revenues, dtype: float64
```

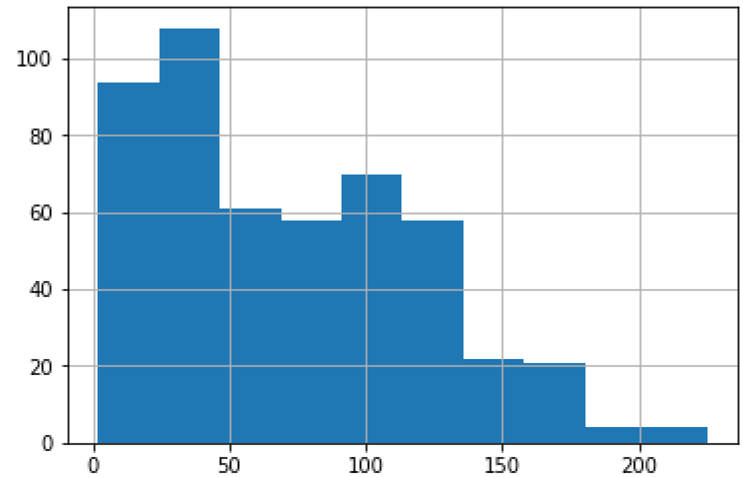



```
count       490.000000
mean      20181.500000
std       23482.105746
min        5145.000000
25%        7144.250000
50%       11131.000000
75%       21076.750000
max      146850.000000
Name: Revenues, dtype: float64
```



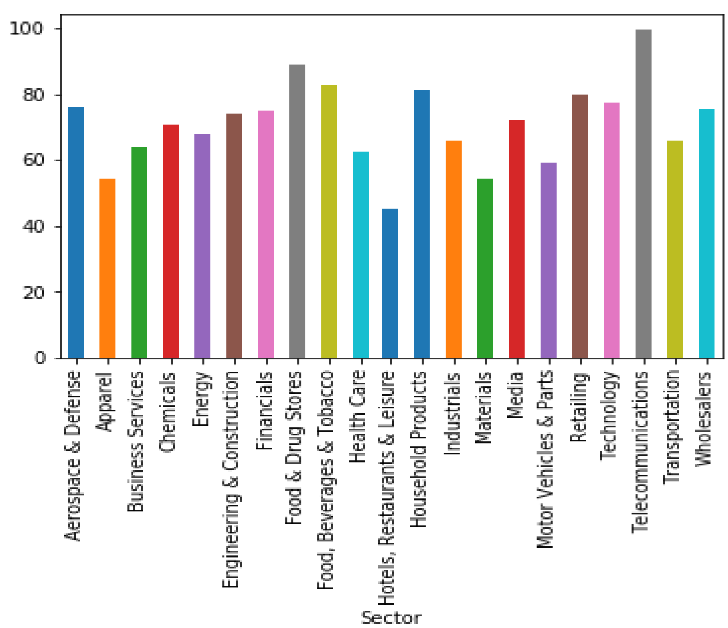
Foundation date:

- Decent amount of companies are between [0-25] years old
- The average age is around 72 years



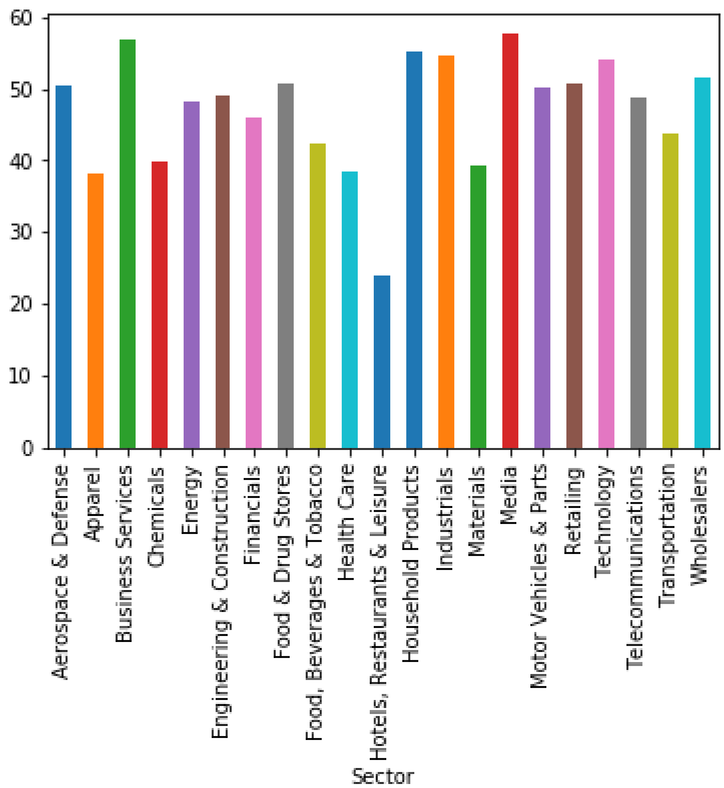
In the table below, we can see the average companies age of each Sector

- The youngest Sector is hotels, restaurants and leisure which make sense
- The oldest companies are in the telecommunication Sector

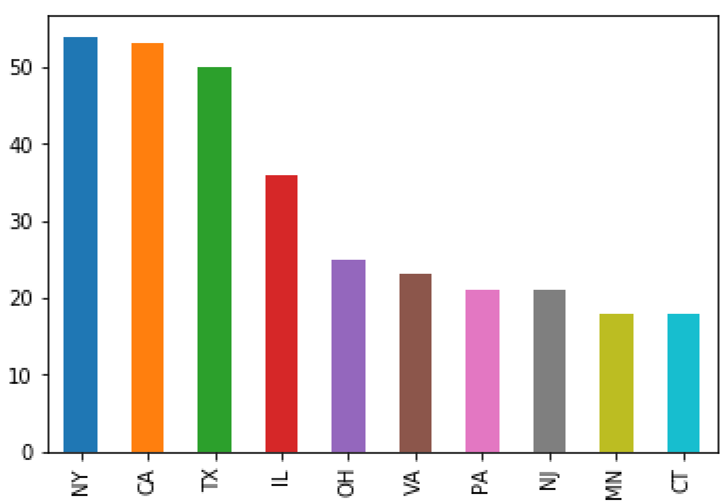


In the graph below we can see the company age STD of each sector

- Lower STD for hotels, restaurants and leisure.
- Highest STD in Media and Business Services

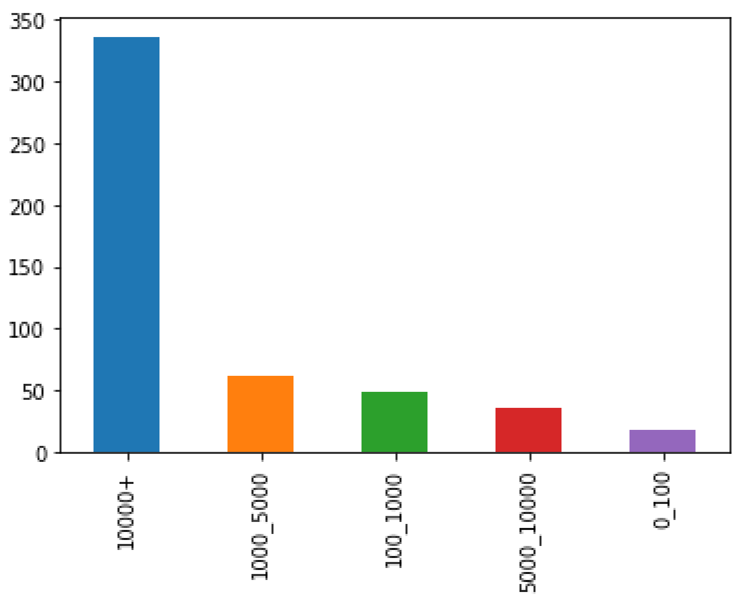


<u>Headquarter State</u>: Top 10 states



- Around 30% from the companies located in California, New York and Texas
  Which is reasonable, Silicon Valley for example located in California

Popularity – Website Rank: I divided the ranks to 5 different levels



- It's interesting to see that most of the companies doesn't have popular websites

  Probably its related to the Sector of the company
- In the table we can see the companies that have the most popular website

  We can say that there is correlation between rank and website rank

  As most of the top company's websites are also in the top 100 list
- Most of the companies came from Technology and Retailing Sectors

Rank	Title	website_rank_in_US	Sector
1	Walmart	20.0	Retailing
3	Apple	68.0	Technology
9	AT&T	58.0	Telecommunications
12	Amazon.com	4.0	Technology
23	Home Depot	52.0	Retailing
25	Wells Fargo	39.0	Financials
26	Bank of America Corp.	46.0	Financials
27	Alphabet	1.0	Technology
28	Microsoft	53.0	Technology
38	Target	62.0	Retailing
40	Lowe's	87.0	Retailing
72	Best Buy	73.0	Retailing
98	Facebook	2.0	Technology
100	Capital One Financial	63.0	Financials
264	PayPal Holdings	28.0	Business Services
310	eBay	10.0	Technology
314	Netflix	18.0	Technology
498	Yahoo	5.0	Technology

Correlation Matrix:

- There is low correlation between revenues and assets
- High correlation between number of employees and revenue
- Low correlation between employees and assets

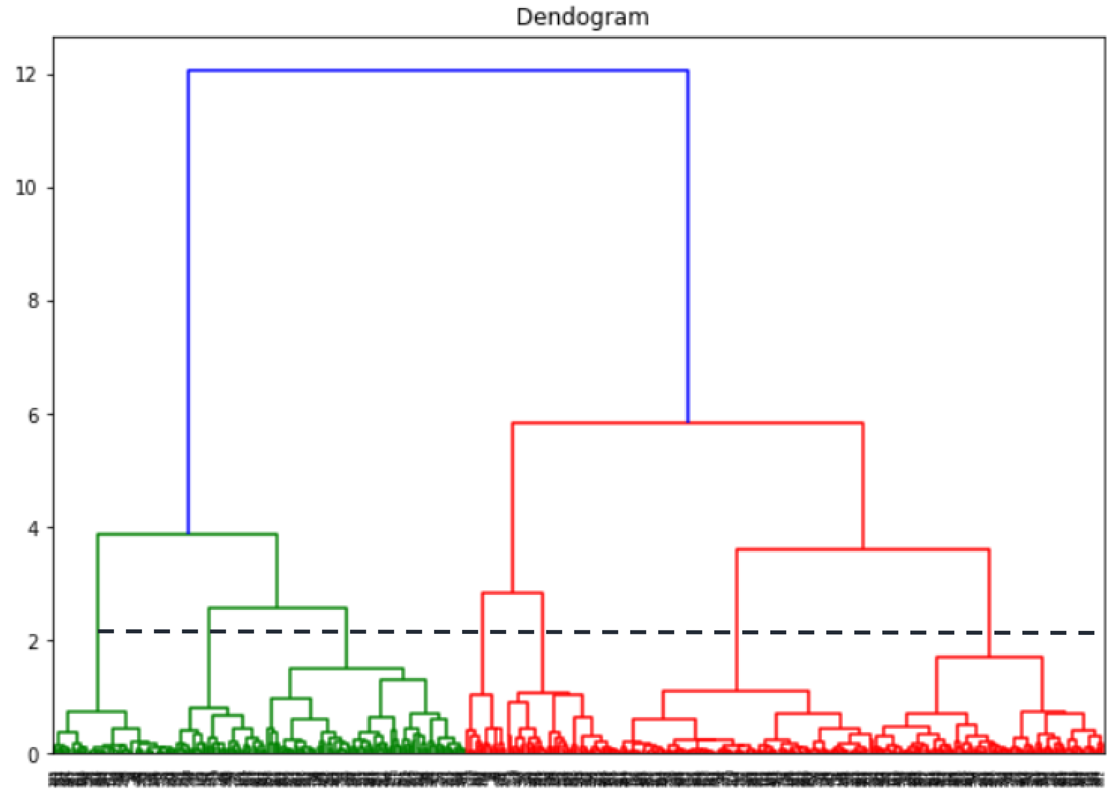
Preprocessing Step:

in the dataset we have both categorical and numerical features, so I will create dummy Variables for the Sector Category.

And then I will normalize all the features, so they will be in the range [0,1].

Benchmark

As a benchmark model we will use simple hierarchical clustering.



We can see in the graph above that the number of clusters is 7.

The silhouette score for this model is 0.378

Algorithms and techniques

I decided to try few algorithms from the clustering family:

- K-means
- DBSCAN
- GMM

DBSCAN:

The silhouette score for this algorithm is 0.179, the number of clustered estimated is 4.

With this poor result we should find other clustering algorithm.

GMM:

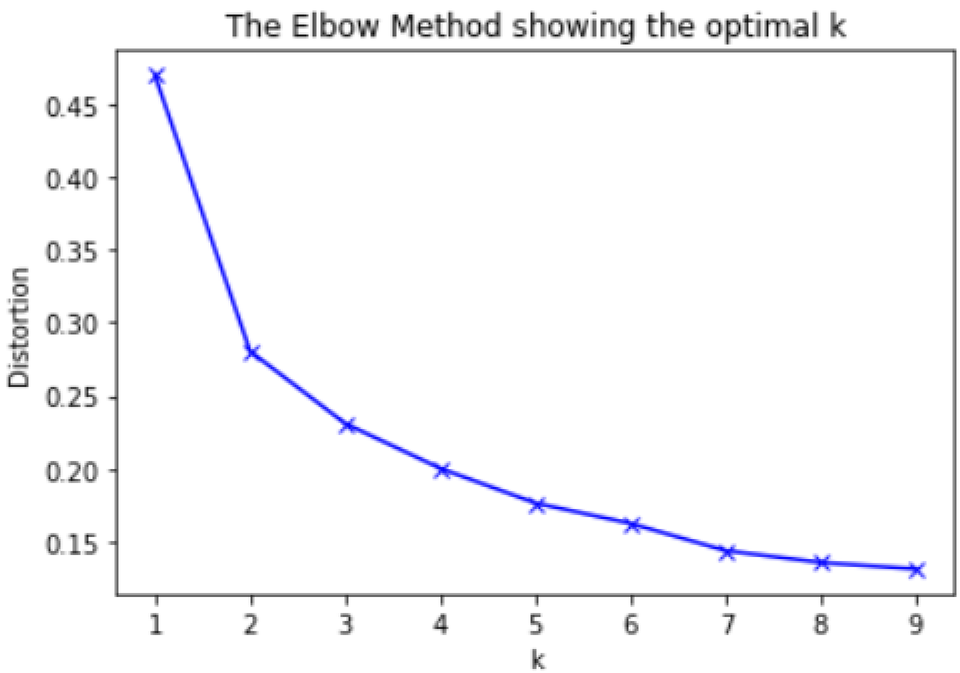
For GMM with 4 components we are getting silhouette score of 0.419

Looks much better than DBSCAN and even slightly better than the benchmark model.

K-MEANS:

First, we will check the optimal k for k-means.

Elbow method:



We can consider 4 clusters as good K for k means.

Silhouette score:

```
For n_clusters = 2, silhouette score is 0.5319082538685731)
For n_clusters = 3, silhouette score is 0.45516975600456777)
For n_clusters = 4, silhouette score is 0.46680172424712685)
For n_clusters = 5, silhouette score is 0.39782490461504755)
For n_clusters = 6, silhouette score is 0.4164819109976355)
For n_clusters = 7, silhouette score is 0.39642904813305097)
For n_clusters = 8, silhouette score is 0.37646778679318604)
For n_clusters = 9, silhouette score is 0.37764176221662)
For n_clusters = 10, silhouette score is 0.36122207869270795)
For n_clusters = 11, silhouette score is 0.37324634906929294)
For n_clusters = 12, silhouette score is 0.36889134356625564)
For n_clusters = 13, silhouette score is 0.3720169481156756)
For n_clusters = 14, silhouette score is 0.32599913463857866)
```

Although for 2 clusters we are getting the highest silhouette score, in business perspective 2 clusters is not enough to get powerful and interesting insights.

Again, we can proceed with 4 clusters which yield a decent silhouette score of 0.466.

Feature importance:

I must say that for unsupervised algorithms there is no enough information on this topic.

I performed a research and found few ways to tackle this.

I will use one dimensional analysis between each variable and the clusters labels and then get the ANOVA table.

Coming from the ANOVA framework, the information we are really after in this table it the F-statistic and its corresponding p-value. This tells us if we explained a significant amount of the overall variance.

1. F-statistics

An F-statistic is the ratio of two variances, or technically, two mean squares. Mean squares are simply variances that account for the degrees of freedom (DF) used to estimate the variance. A big F-statistic which yield small p-value mean that the variance explained by this variable is significant.

2. R^2

   $R^2 =$ SSM/SST= How much variance is explained by the model

   SST= the total variance

   SSM = total sum of squares

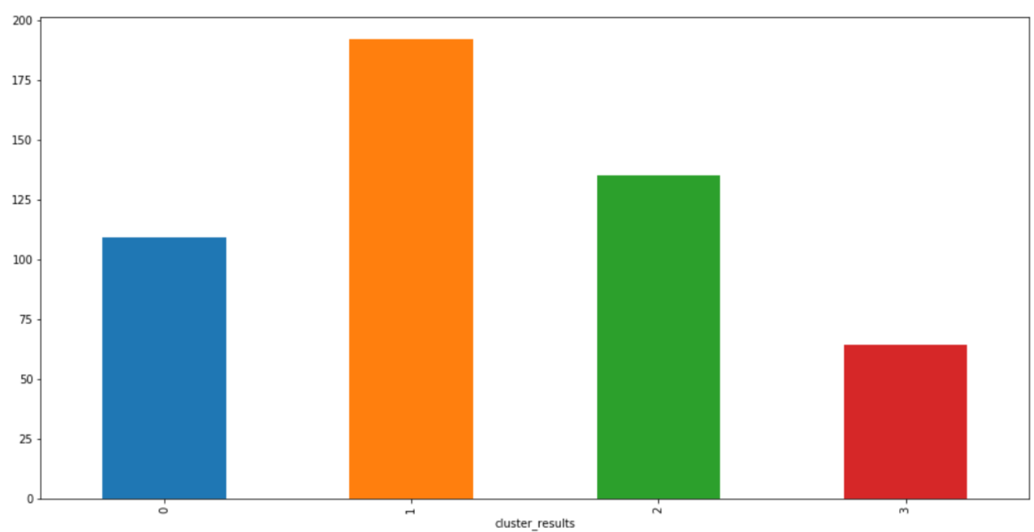
The idea is to measure the proportion of the variance (of the variable) explained by the group membership.

Feature	F-statistics	P-value	R^2
Assets	1188	0.00	88%
Employees	1005	0.00	86%
Revenues	347	0.00	68%
Totshequity	85	0.00	34%
company_age	24	0.00	13%
rank_business_state	17	0.00	10%
website_rank	14	0.00	8%
Profits	7	0.00	4%
profit_revenue_ratio	5	0.00	3%
ceo_in_100_fortune_list	4	0.00	3%
Prftchange	3	0.03	2%
Revchange	2	0.18	1%
admired_list_indication	0	0.76	0%

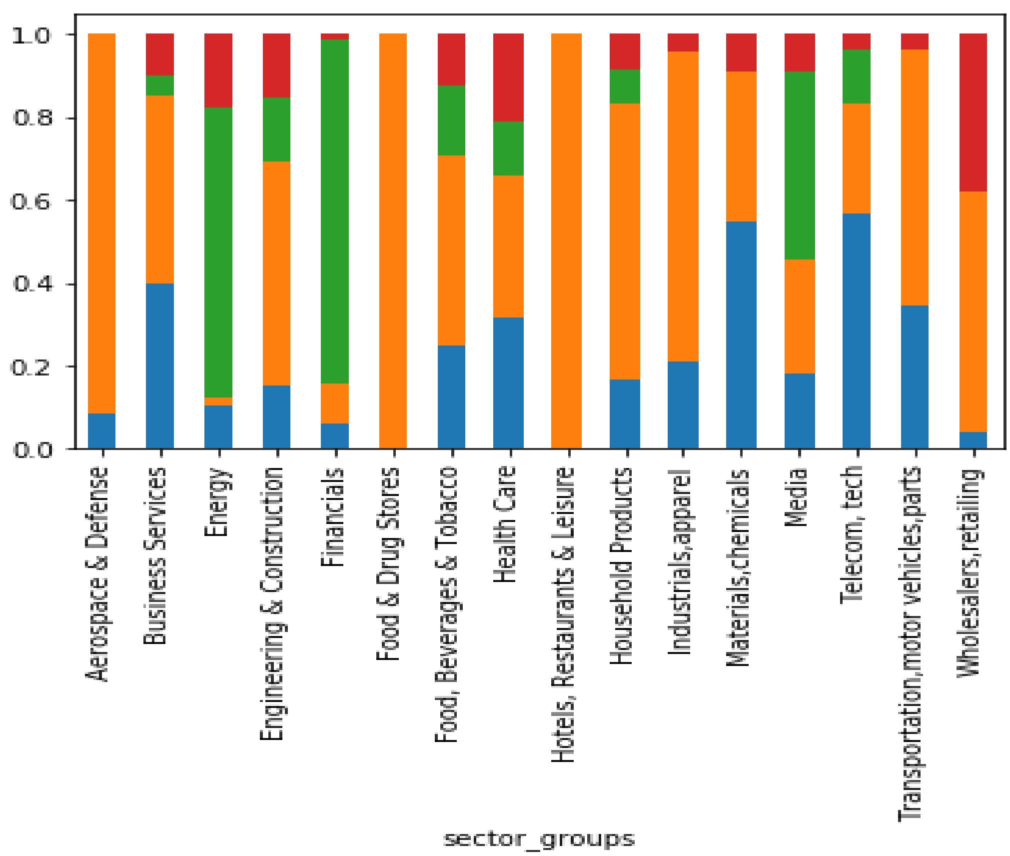
- From the table above, we can get an idea on the top features which are Assets, Employees and Revenues

Clustering insights

- First, we can see the size of each cluster, cluster number 1 have the highest number of companies close to 200 which is 40% from all the companies
  Controversy the lowest number of companies is in cluster number 3 which reflect around 10% from the 500 fortune companies.

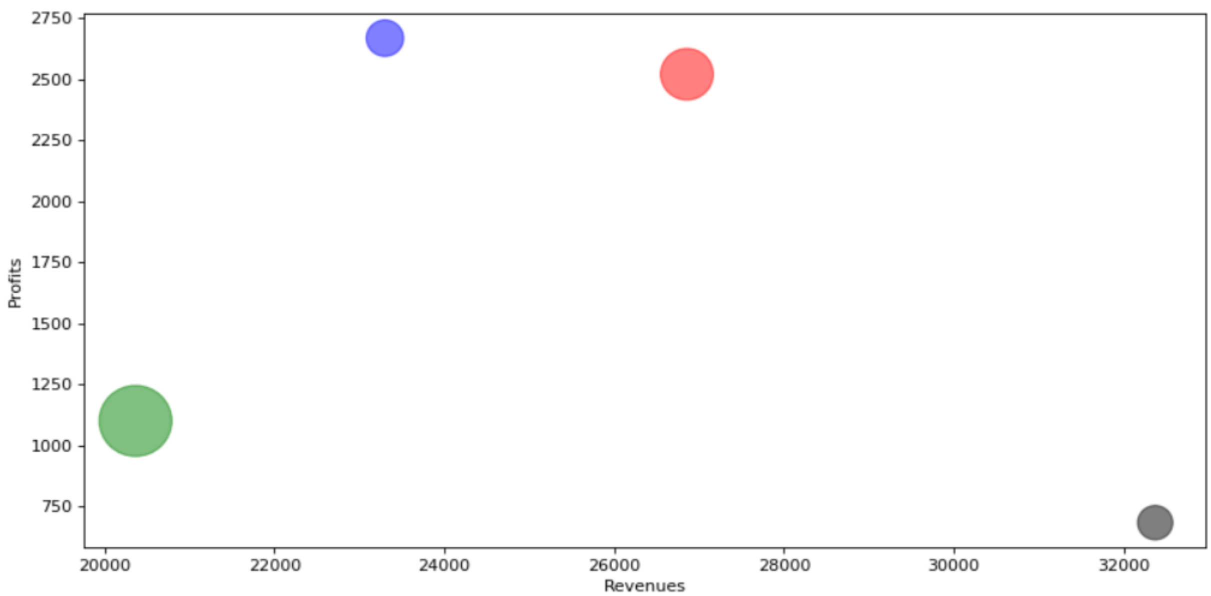


- In terms of sectors we can see that for cluster number 2 most of the companies came from Financial and Energy.
- For cluster number 0 most of the companies coming from Technology Health Care And Chemicals
- Most of the retailing companies came from cluster number 1
- Wholesalers are almost 50% from the companies in cluster number 3



- We can see in the scatter chart below the Average Profits & Revenue for each cluster, the size of the bubbles shows the Average number of employees
  the most profitable companies are in cluster number 0 and 2
  The highest Revenues companies are in cluster number 3 but their profit is relatively low, cluster number 1 is the weakest cluster in terms of Revenues & Profits.

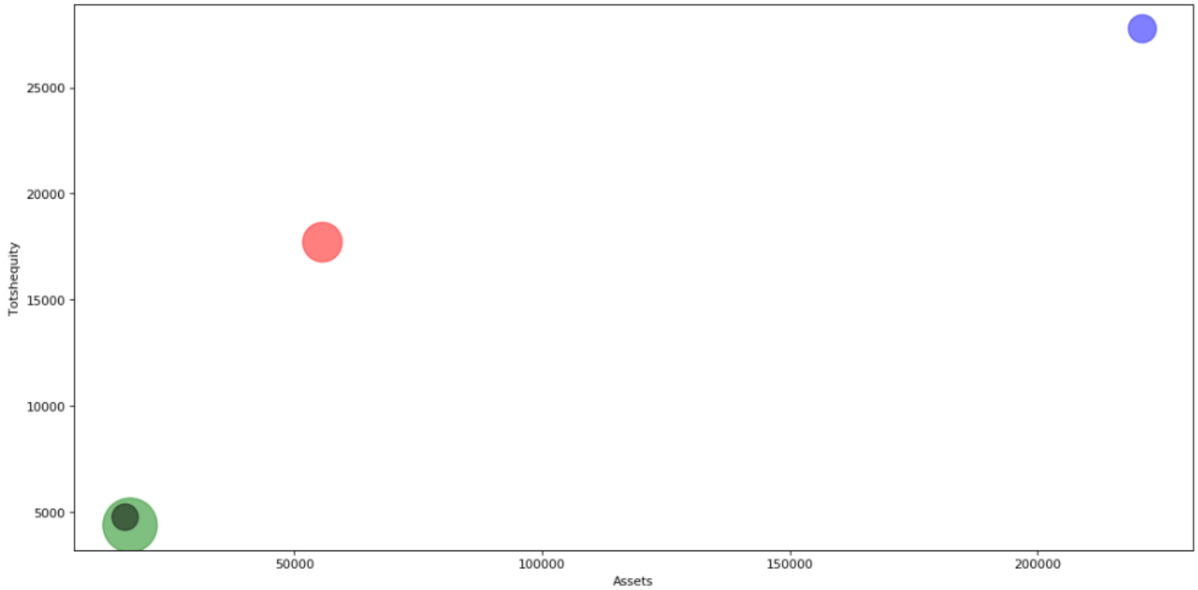
Cluster 0- Red	Cluster 1- Green	Cluster 2 – Blue	Cluster 3 – Black



Cluster	Revenues	Profits	Employees
0	26859.908257	2520.589908	49450.027523
1	20365.609375	1099.370313	94006.348958
2	23303.362963	2668.484444	24680.896296
3	32374.890625	682.360937	21935.515625

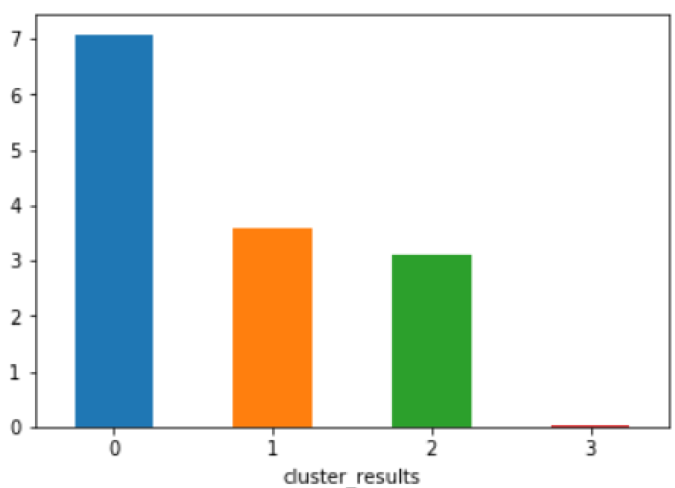
In the next scatter we can see the same view but on Assets & Total equity

- Clusters 1 and 3 are the same and are quite low comparing to the 2 other clusters. cluster number 2 preforming much better comparing to all other clusters



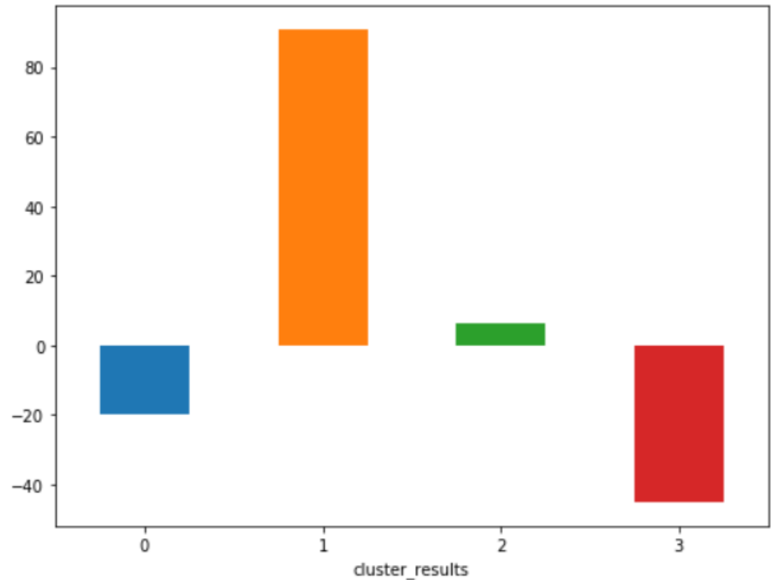
Revenue trend -In terms of trend we can see in the graph below the average revenue change for each cluster.

- There is no change in cluster number 3
- Around 3.5% change in cluster number 1 & 2
- 7% change in cluster number 0 (which is logic -this is the cluster with low revenue)



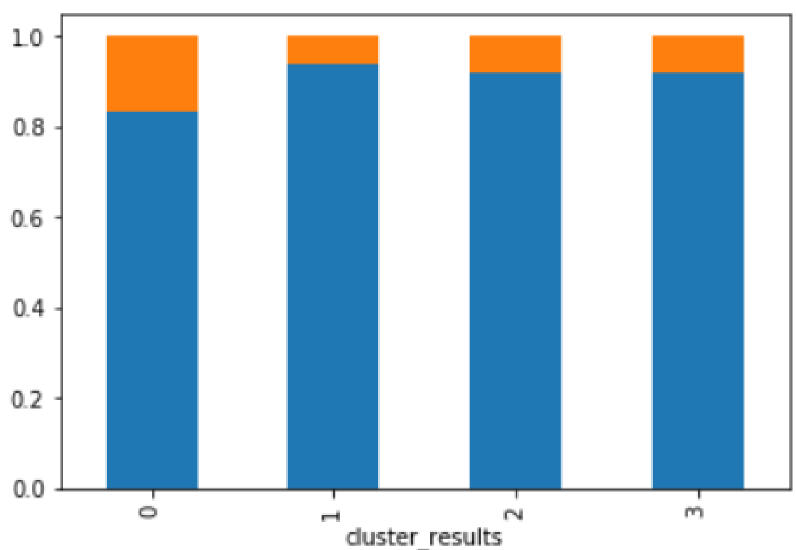
Profit trend –

- Negative trend in cluster 0 & 3
- Highly positvly trend in cluster number 1 – its quite logic consideing this is the cluster with the companies with lowes average profit.

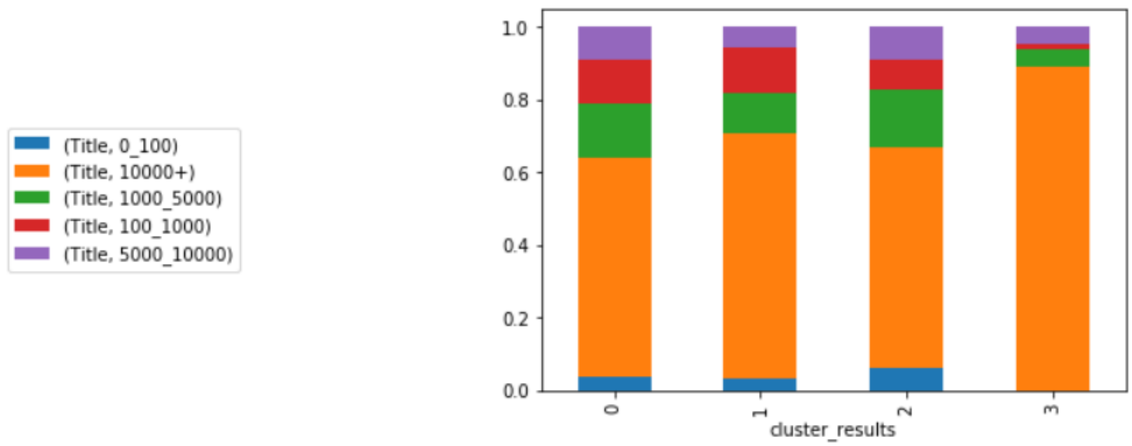


CEO-

- Most of the allstar CEO are in cluster number 0



- Most of the popular website companies are in cluster number 2
- Most of the unpopular websites are in cluster number 3 – looks like correlation!



Insights summary

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Dominate Sector	Technology & Chemicals	Retailing	Financial & Energy	Wholesalers
Revenue	High	Low	High	High
Profit	High	Low	High	Low
Number of Employees	Medium	High	Small	Small
Assets	Medium	Low	High	Low
Equity	Medium	Low	High	Low
Revenue change	High Positive	Medium Positive	Medium Positive	No change
Profit change	Small Negative	High Positive	Small Positive	High Negative
Website Rank	Medium Popularity	Medium Popularity	High Popularity	Low Popularity
Allstar CEO	High number	Medium number	Medium number	Medium number

End note

We saw quite interesting insights on the clustering results. I think that with more interesting data like employee satisfaction- glassdoor reviews, followers on Facebook, mentioned in news, google search volumes and more we can even get more valuable insights.

Moreover, I think that with more companies (even worldwide) we can get more powerful results.