Study project:

CEO Characteristics and firm performance: evidence from Fortune 1000 companies

Group-4 “Milton-Friedman”

Mikhail Mironov u211361

Vera Garmanova

Introduction

Research goals

- Find patterns in CEO dataset and determine how CEO's characteristics affect company's performance
- Cluster CEOs into groups based on their prior work experience
- Study how CEOs differ in various sectors

Analysis methods applied

- Principal Component Dimension reduction for plotting clusters and analysis with correlations
- Multidimensional Scaling using various distances: mainly focused on Euclidean and Manhattan
- Hierarchical clustering with various distances and linkages using SciPy implementation. Scree plot of merged clusters
- KMeans with Euclidean and Manhattan distances. Silhouette analysis and WSS Scree plot. Bootstrap for stability.

Dataset. Data collection

Financial data on Fortune 1000 companies such as revenue and profits and basic data on CEOs like name

Collected data on followers and prior work experience as a CEO at current company and other related work experience

Data on total compensation and salaries of CEOs and their age

507 observations
22 variables

company	sector	revenue	profit	months_ceo_company	months_company	months_company_others	age	num_comp	num_followers	total_compensation
Walmart	Retailing	572754.0	13673.0	166	166	42	55.0	2	1136161	25670672.0
Amazon	Retailing	469822.0	33364.0	308	308	0	54.0	1	354740	212701170.0
CVS Health	Health Care	292111.0	7910.0	22	49	258	59.0	6	285180	7045167.0

Source:

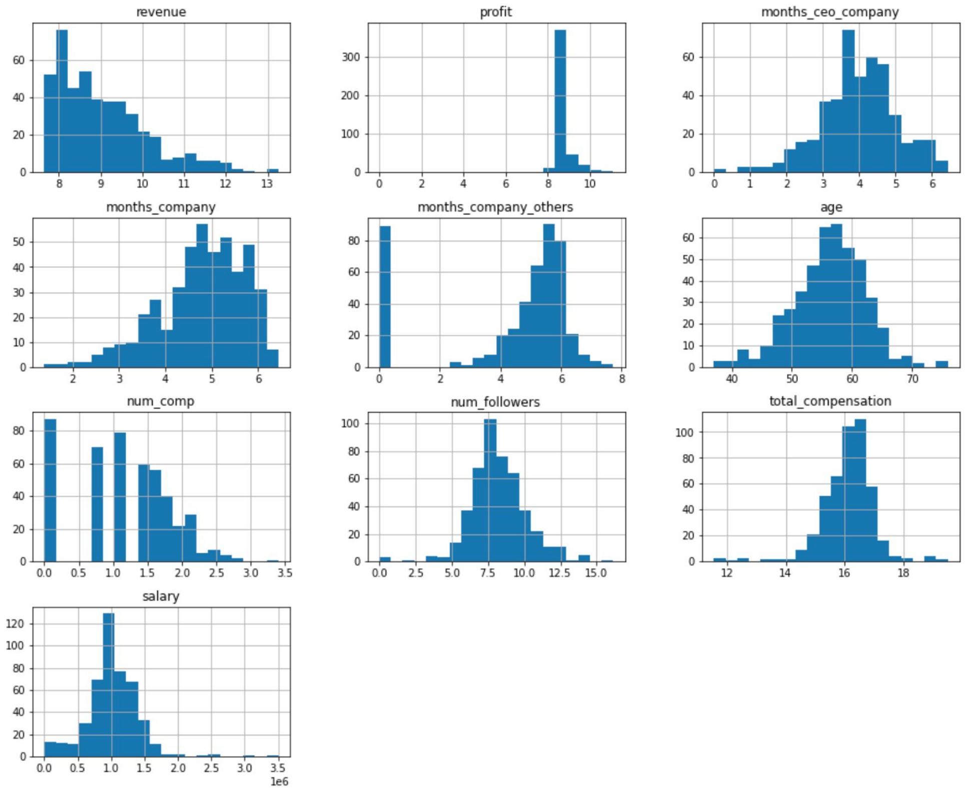
Dataset

After initially seeing that the majority of variables are distributed log-normally, we decided to take the log of such columns.

Before taking the log we eliminated all negatives and zeros by using either +1 to the column or the following:

$$X_{transformed} = X + 1 - \min(X)$$

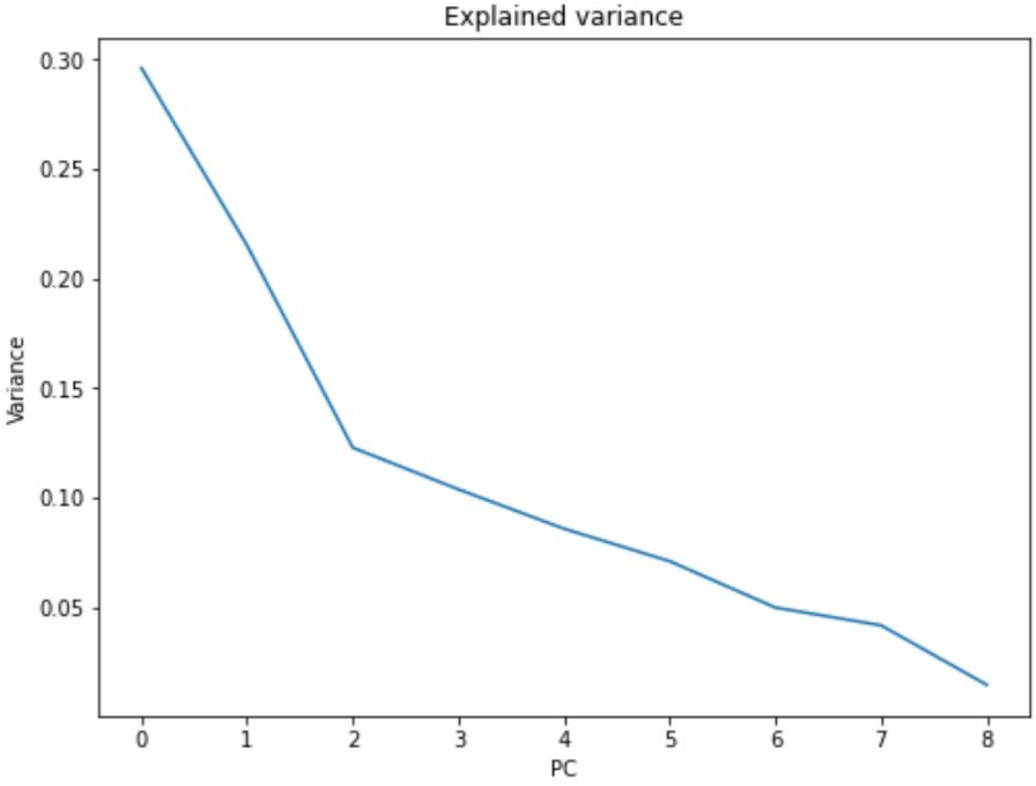
These log columns will be used further in the modelling for clusters and hierarchical clustering. But for interpretation we will use initial data preserving scale



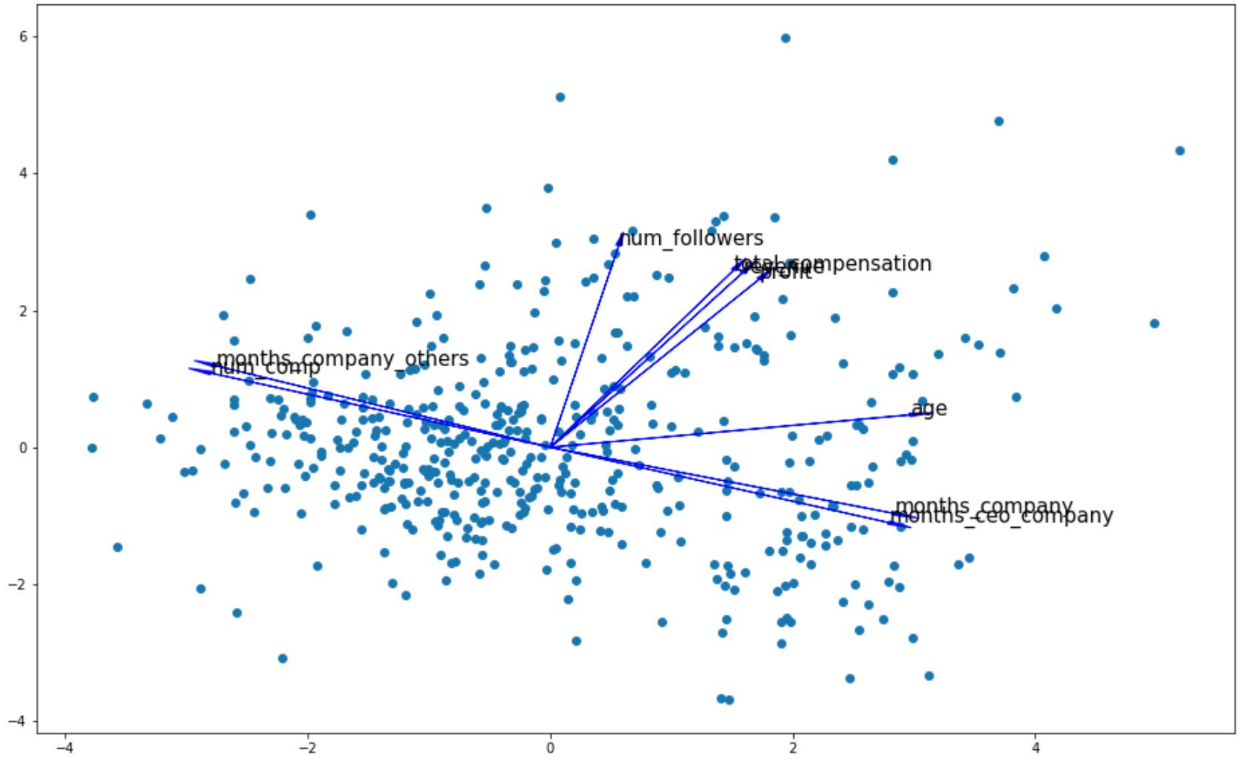
Principal Component Analysis

- We have done PCA with multiple dimensions, we were not satisfied with 2 components plots which explain 55% of total variance, so we add the 3rd component to our analysis.

- We will use plots for both projections onto PC1 & PC2 and PC1 & PC3.



Principal Component Analysis. Component correlations



Hierarchical Cluster Analysis

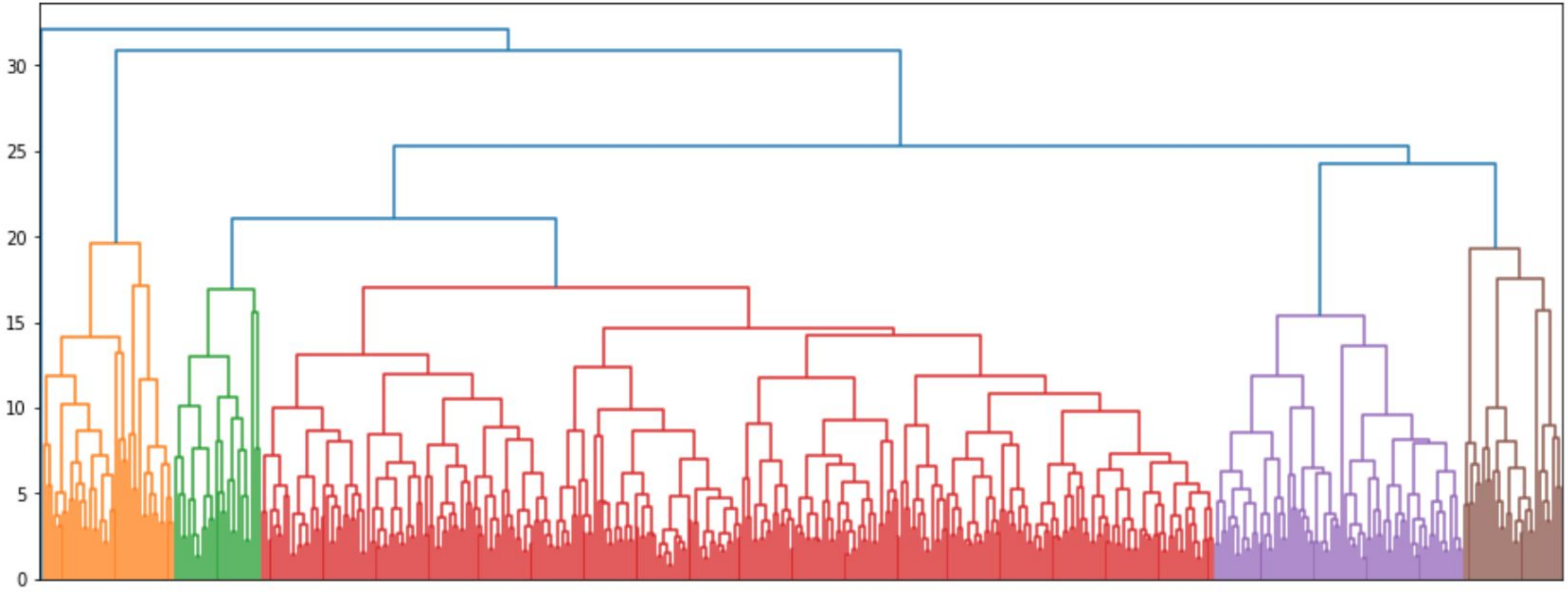
We have tested multiple combinations of distance metrics and linkage methods. As a result we have found out the best combinations

Ward linkage with Euclidean distances. As a result we get 4-5-ish clusters

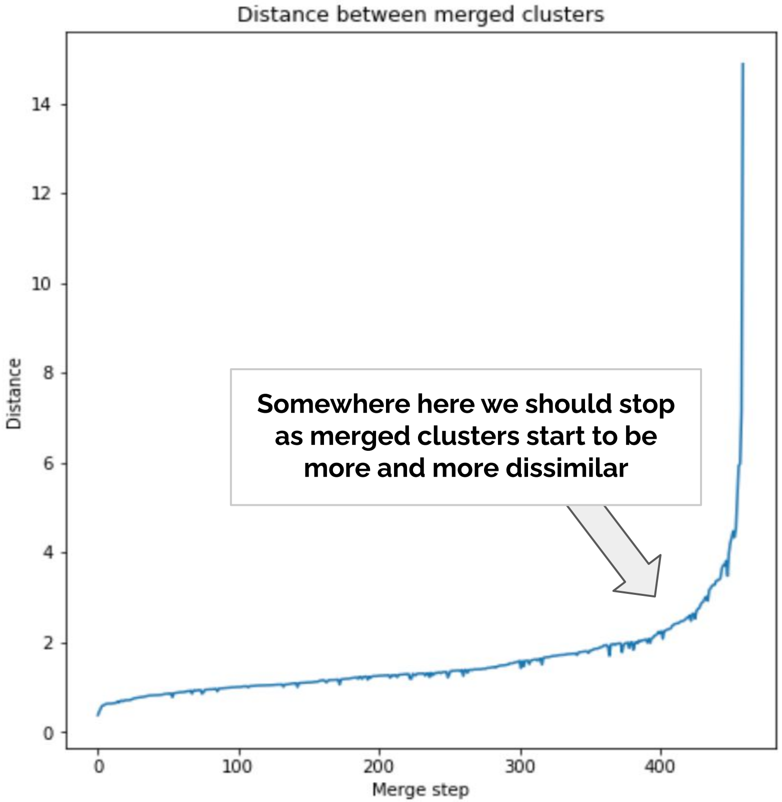
Hierarchical Cluster Analysis

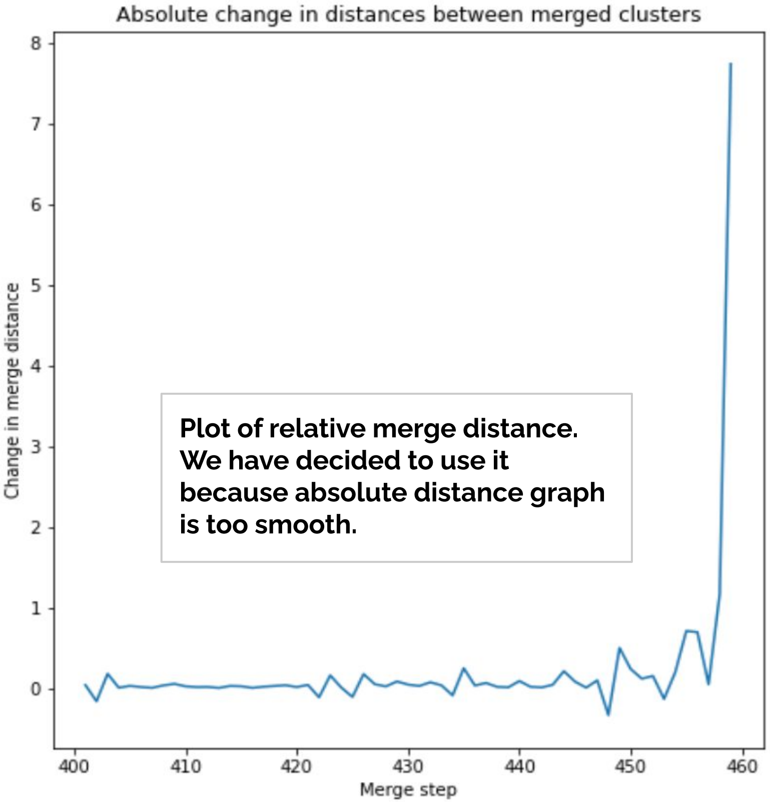
Another promising result was obtained by using Manhattan distances, other combinations resulted in 2 or 3 disproportional clusters.

Complete linkage with Manhattan distances. Similarly we get 5-ish clusters



Hierarchical Cluster Analysis. Scree plot criteria

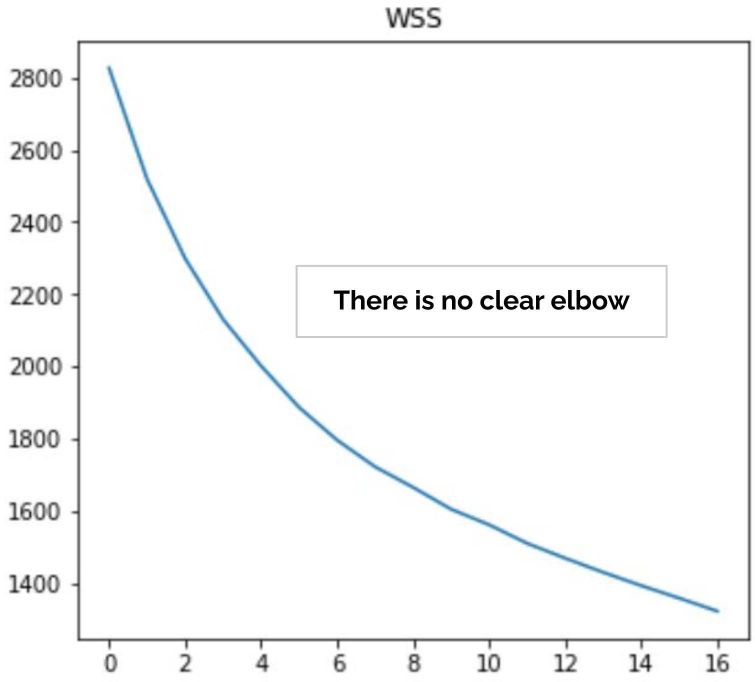


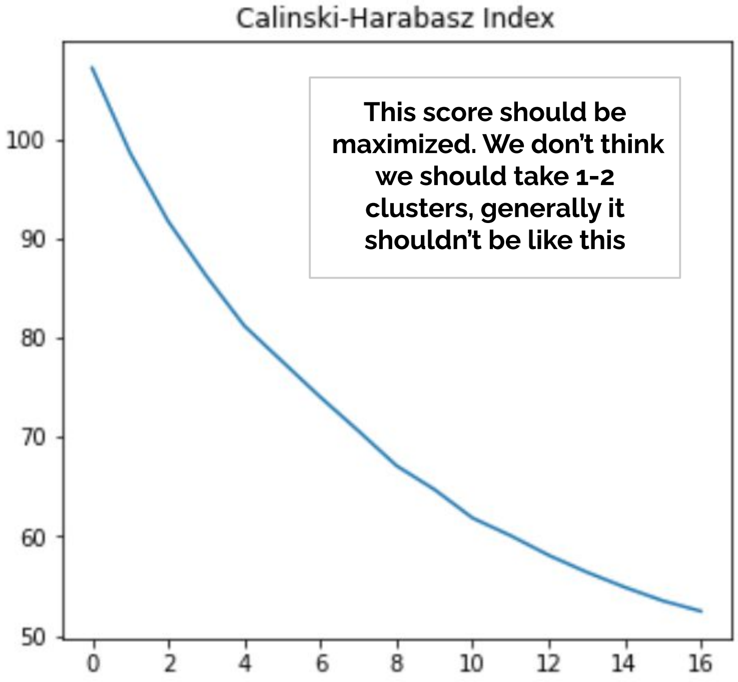


Source:

KMeans clustering. Number of clusters. WSS, CH index.

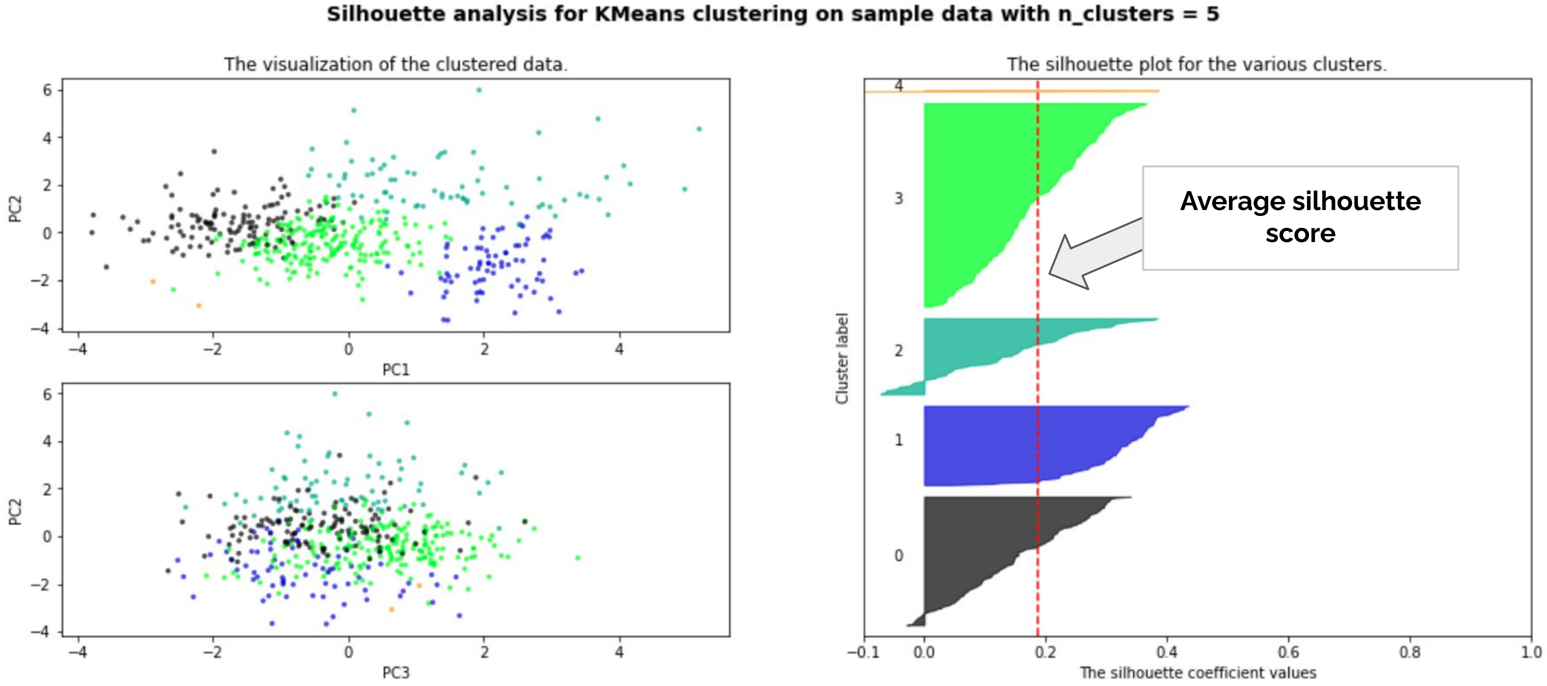
No clear answer as to how many clusters we should try to find. We will assume it is 4-5ish based on the results obtained by hierarchical cluster analysis.





KMeans clustering. Silhouette analysis

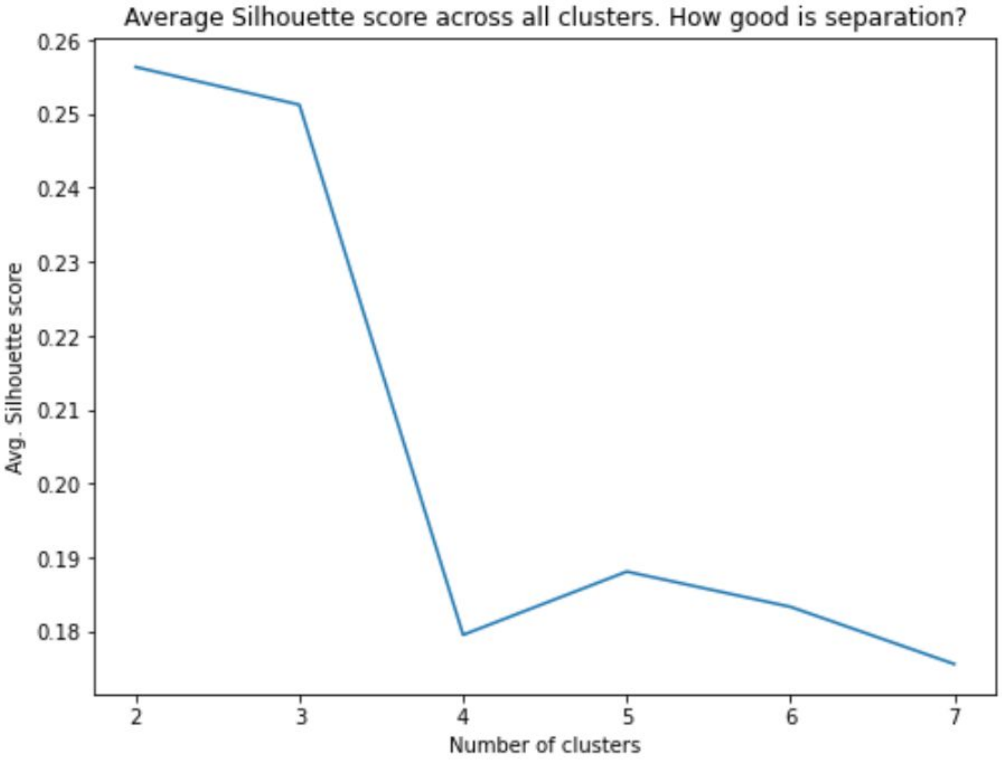
Additionally, we decided to use Silhouette scores which is a quite popular way to evaluate how good the separation is. This method is implemented and well documented in both Python and R



KMeans clustering. Silhouette analysis

- We can see clearly that as we increase the number of clusters, Silhouette score goes down which indicates more overlaps and overall worse separation. But still we don't want to choose 2 or 3 clusters since we believe KMeans should produce results in line with Hierarchical clustering.

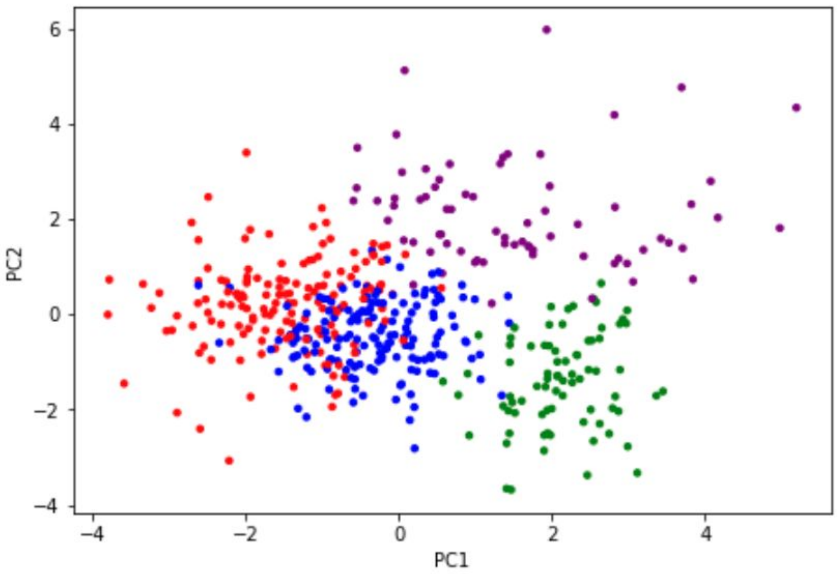
- Further, we will use 4-5 clusters for analysis with KMeans

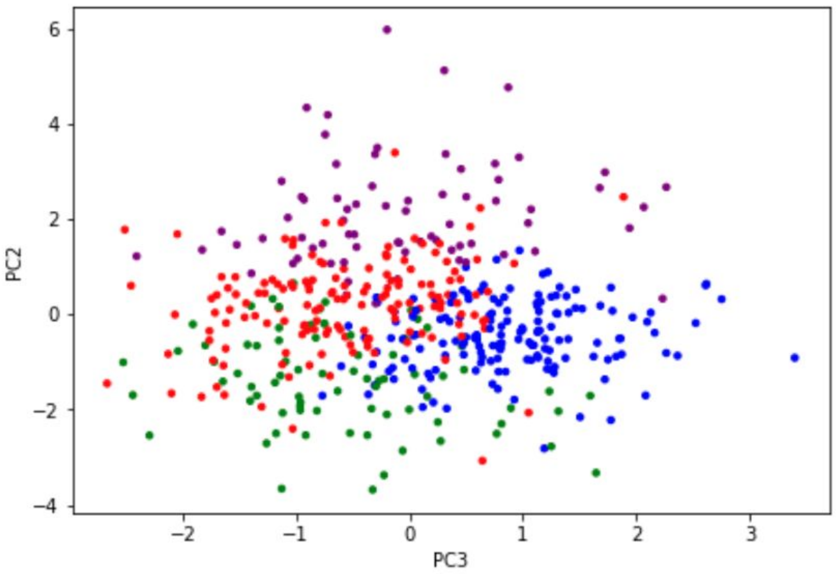


KMeans clustering. Results

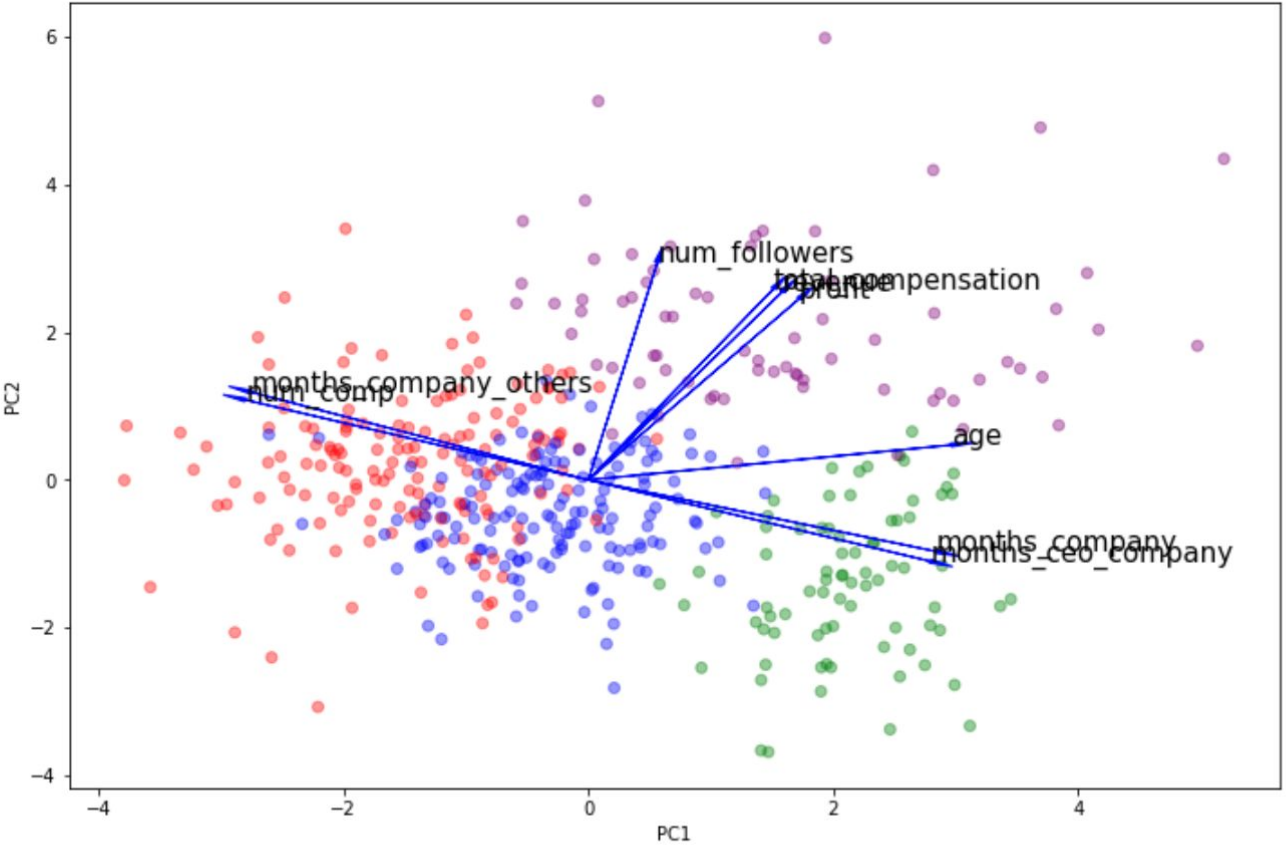
Algorithm has been run 10000 times with random starting centers.

Kmeans Euclidean distances 4 clusters. Plotted with 3 PCA projections



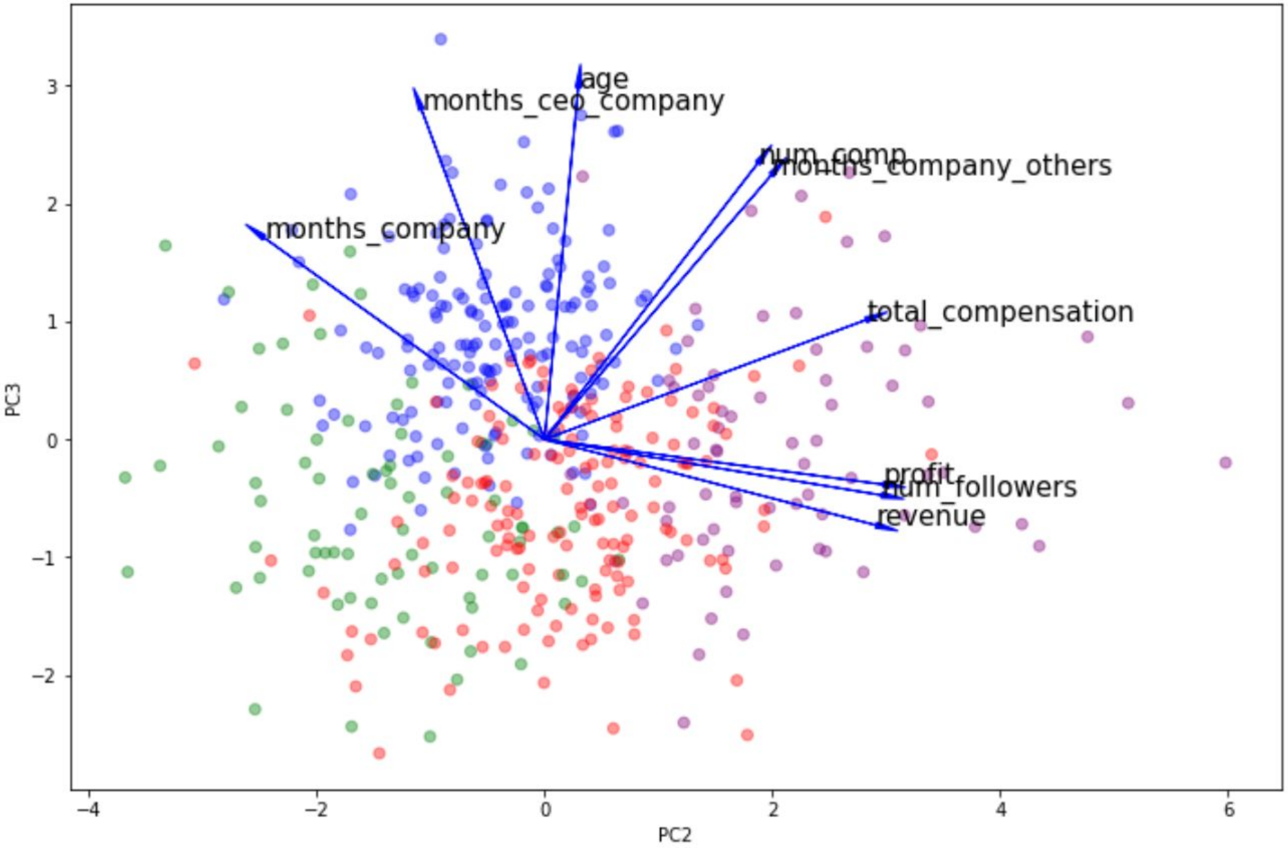


PCA plot with correlations and KMeans colormap. PC1 & PC2



Source:

PCA plot with correlations and KMeans colormap. PC2 & PC3



Source:

Descriptive statistics of obtained with KMeans clusters

	kmeans_4	0	1	2	3
revenue	mean	11159.799342	7061.549375	11122.348718	74149.400000
profit	mean	503.422368	661.605625	1156.326923	9570.152113
months_ceo_company	mean	33.427632	108.356250	145.371795	98.042254
months_company	mean	77.671053	176.918750	270.589744	227.084507
months_company_others	mean	283.072368	300.381250	0.782051	184.281690
age	mean	52.000000	58.325000	57.987179	58.591549
num_comp	mean	5.190789	4.718750	1.076923	3.309859
num_followers	mean	13224.763158	3657.618750	5200.153846	308568.169014
total_compensation	mean	14606362.157895	10217382.812500	11647091.782051	29992770.704225
salary	mean	902855.960526	1017724.112500	1031356.256410	1323759.492958
profit_ratio	mean	0.066522	0.101285	0.115886	0.190222
kmeans_manhattan	mean	0.105263	0.068750	1.115385	1.253521

Source:

Conclusions

Source: