Exploratory analysis of innovation data in 4.0 industry in two Italian provinces

Topics in Statistical Learning

Ekaterina Kirillova
Kevin Pirazzi

CONTENTS

01

Introduction

Presenting the data

Industry 4.0

New phase in the Industrial Revolution that focuses heavily on interconnectivity, automation, machine learning, and real-time data

RESEARCH TARGET:

Assessment of the implementation of the technologies from Industry 4.0 in Italy

HYPOTHESIS:

The company is likely to be investing into the technologies from Industry 4.0, if there is evidence of increase in assets and the high return on assets.

INITIAL DATASET

224 mechatronics enterprises

Data from Balance Sheets, 2013 - 2019

VARIABLES

1. Revenues
2. Yearly results
3. Return on Assets (ROA)
4. Intangibles Assets
5. Total Assets
6. High ROA (0 or 1)

Data Descriptive Statistics

Removed from dataset for the analysis

Only utilized in logistic regression
Variable assumes values: 0,1

Regular Data

Firm	Province	Revenue	Results	ROA	Immobilizzazioni_Immateriali	Assets	HIGH_ROA
Length:202	Length:202	Min. : 76	Min. :-14893.0	Min. :-0.25000	Min. : 9.0	Min. : 832	Min. :0.000
Class :character	Class :character	1st Qu.: 1961	1st Qu.: 31.5	1st Qu.: 0.03000	1st Qu.: 264.2	1st Qu.: 25714	1st Qu.:0.000
Mode :character	Mode :character	Median : 4464	Median : 131.5	Median : 0.06000	Median : 773.5	Median : 66252	Median :1.000
		Mean : 21541	Mean : 956.6	Mean : 0.06614	Mean : 4267.4	Mean : 1144106	Mean :0.505
		3rd Qu.: 10266	3rd Qu.: 367.0	3rd Qu.: 0.10000	3rd Qu.: 2514.8	3rd Qu.: 209223	3rd Qu.:1.000
		Max. :877983	Max. : 52558.0	Max. : 0.30000	Max. :234946.0	Max. :132094717	Max. :1.000

Normalized Data

Revenue	Results	ROA	Immobilizzazioni_Immateriali	Assets
Min. :-0.2702	Min. :-3.0299	Min. :-4.18537	Min. :-0.2396	Min. :-0.12163
1st Qu.:-0.2464	1st Qu.:-0.1768	1st Qu.:-0.47844	1st Qu.:-0.2252	1st Qu.:-0.11899
Median :-0.2149	Median :-0.1577	Median :-0.08127	Median :-0.1966	Median :-0.11467
Mean : 0.0000	Mean : 0.0000	Mean : 0.00000	Mean : 0.0000	Mean : 0.00000
3rd Qu.:-0.1419	3rd Qu.:-0.1127	3rd Qu.: 0.44829	3rd Qu.:-0.0986	3rd Qu.:-0.09946
Max. :10.7794	Max. : 9.8643	Max. : 3.09610	Max. :12.9778	Max. :13.93194

Data Modifications

- Averages were computed for the six years period in order to obtain a unique value for each one of the variables utilized as part of our analysis.
- In XLS file (prior to input): we removed the companies with no observations for at least one year for each one of the variables.
- Input obtained → 202 lines or companies which we analyzed further.

02

METHODS

Methodologies Utilized

METHODOLOGIES

1

Clustering

2

PCA

3

Logistic Regression

Clustering

Cluster Analysis is conducted in order to group observations into similar groups based on their similarity

- Based on different types of distances
- Other unsupervised methods → K-Clustering
  - Each observation is part of k-numbers of cluster centroid
  - Minimizing within cluster euclidean distance (Euclidean Distances). Below the objective function of k-means:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

PCA

- In order to reduce the dimensions (into uncorrelated PCs) → PCA analysis.
- PCA maximizes variance and reduces dimensions in an attempt to minimize information loss
  - Under a new re-centered coordinate system
  - Can include a rotation in the data
- The principal components have correlations to the variables within the data

Logistic Regression

Binary Logistic Regression Model → outcomes (0,1)

The log-odds are the linear combination of one or more independent variables (predictors). Binary logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non-instance.

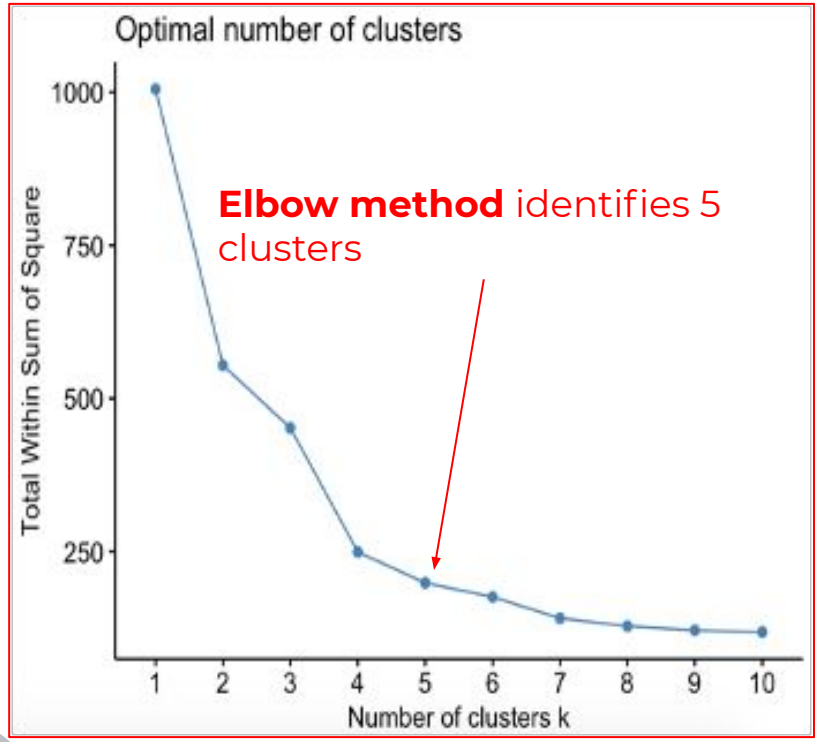
$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

03

RESULTS

Analysis is based on the first part of the course

Clustering - Choice of Cluster

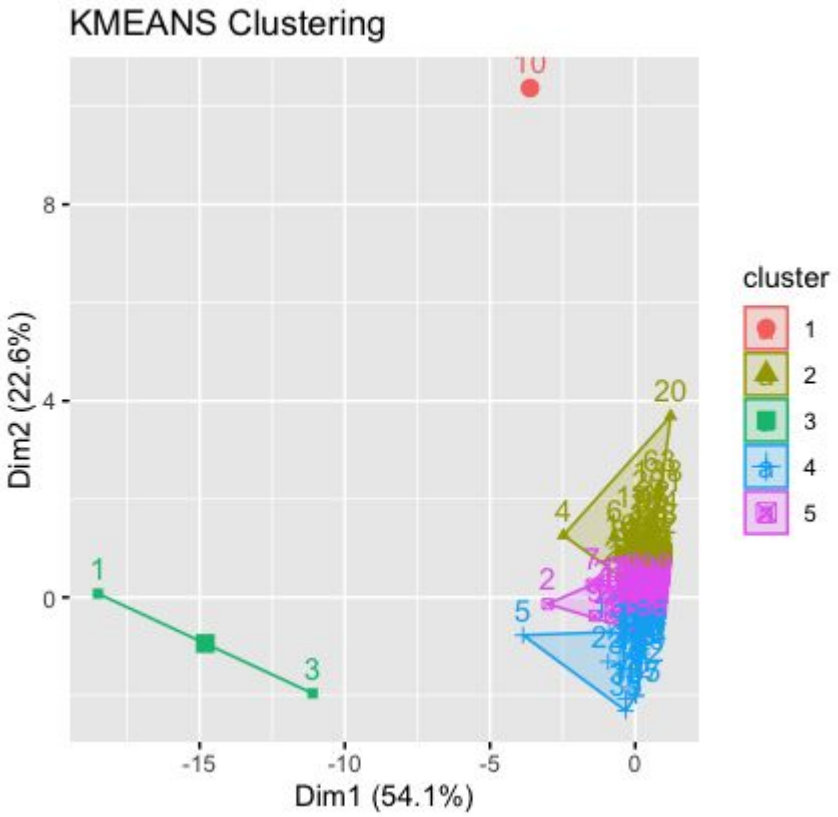



```
> NbClust(Book5, distance = "euclidean", method = "complete", index='hartigan')
$All.index
     2       3       4       5       6       7
116.4681  43.3892  72.6613  38.9763  17.2499 120.2485
     8       9      10      11      12      13
 15.9914  20.4244  98.8327  67.7616  16.3543   9.6411
    14      15
 28.4875   8.7063

$Best.nc
Number_clusters    Value_Index
      8.0000       104.2571

$Best.partition
  [1] 1 2 3 2 2 4 4 4 4 5 4 4 4 4 4 4 4 4 4 6 4 7 4 4 4
 [26] 4 4 4 4 4 4 8 4 4 7 7 7 7 4 4 4 4 4 8 4 8 4 4 4
 [51] 4 4 4 4 8 4 4 4 4 4 4 7 8 4 4 4 4 4 7 4 4 4 4 4
 [76] 4 4 4 4 8 4 8 4 4 7 4 4 4 4 4 4 4 4 4 4 4 7 4
[101] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 7 4 4
[126] 4 4 8 4 4 7 4 4 4 4 4 4 8 4 7 4 4 4 4 4 4 4 4 4
[151] 4 7 4 4 4 4 7 4 4 7 4 4 4 8 7 4 4 4 4 4 4 4 4 4
[176] 4 4 4 4 4 4 4 4 4 4 4 8 4 7 4 4 7 4 4 4 4 4 4 4
[201] 8 8
```

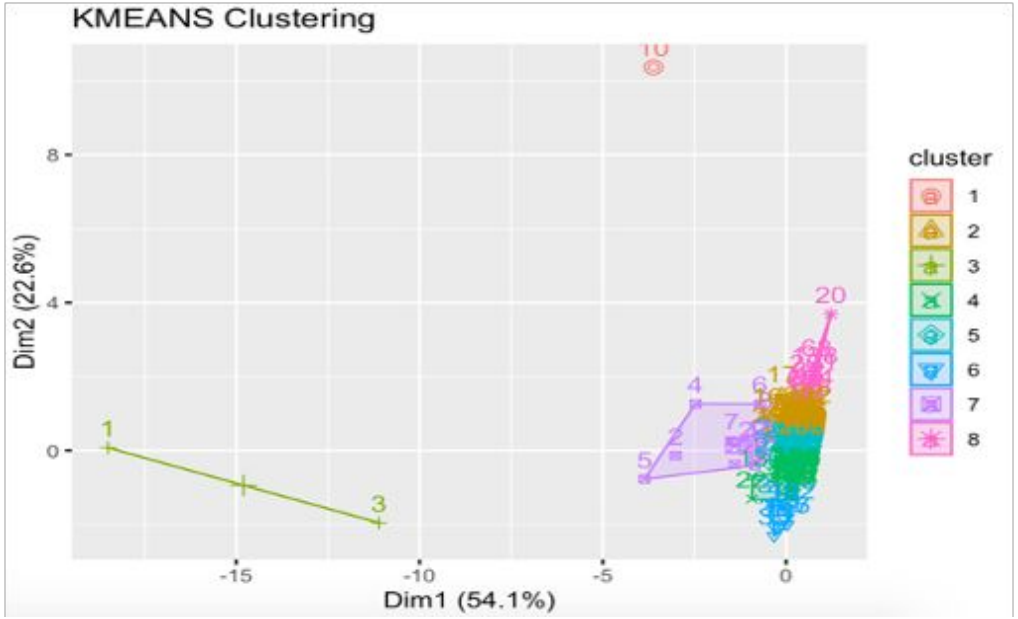
Hartigan Index identifies 8 clusters

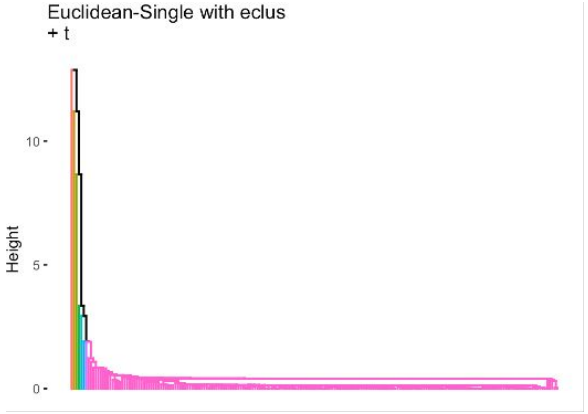
Clustering (5) - Visual Representation

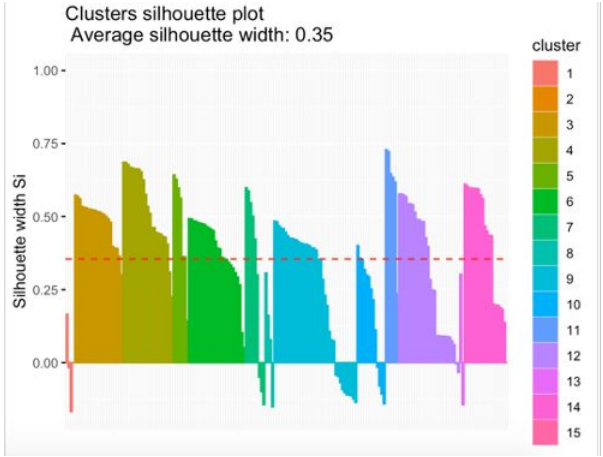
- In this slide the visual representation of how a 5(k-clusters) visualization looks like



Clustering (8) - Visual Representation





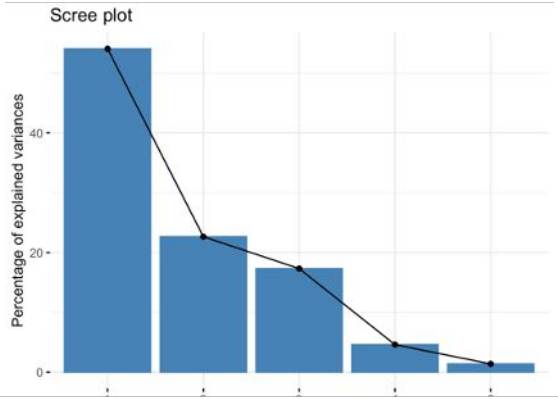


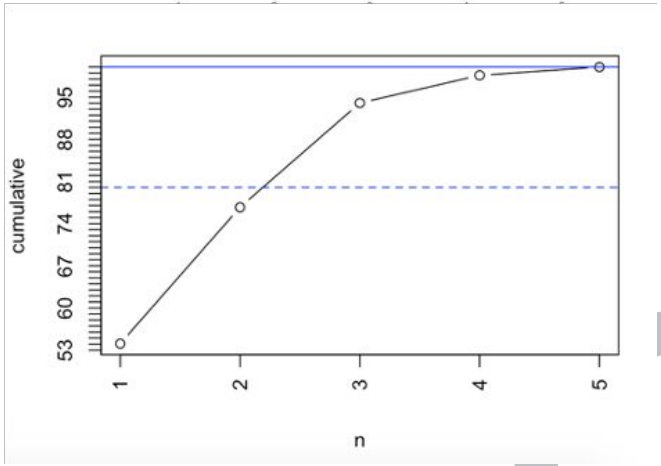
Silhouette = Method of matching ranging -1 to +1 → The higher the better match of clusters

5 : 0.4042018356126361
8: 0.4040726333291286

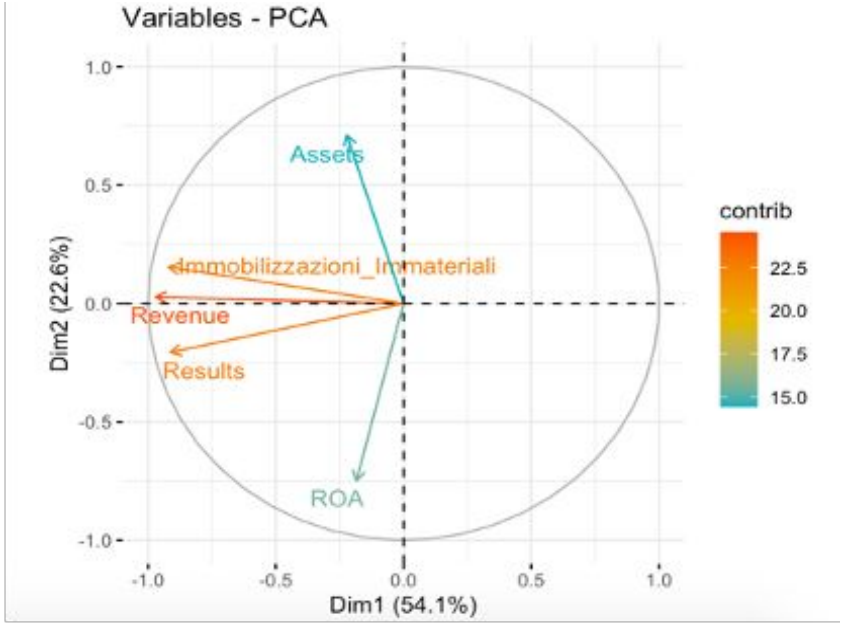
PCA - Choice of PCA

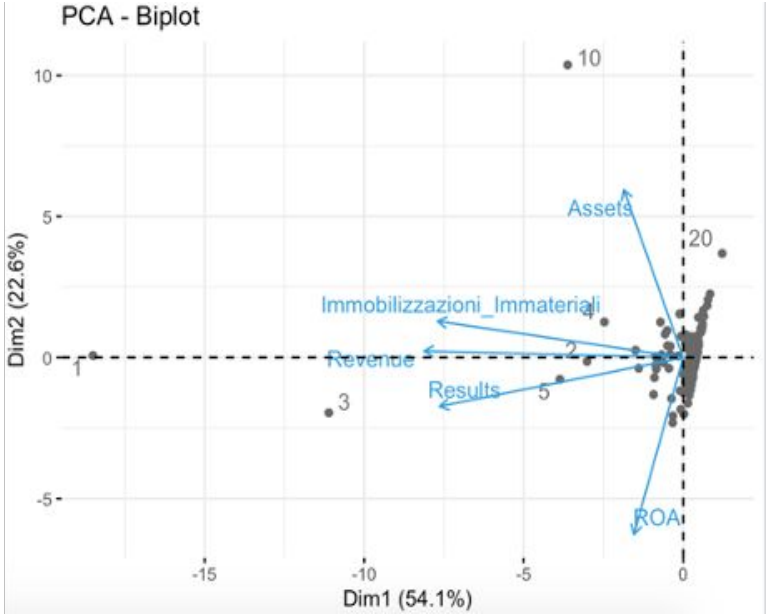
- Utilizing the elbow method we obtain 2 PCs (top right)
- Utilizing cumulative explained variance (>80%) we obtain just above 2 PCs (bottom right)
- For the purpose of this exercise we decided to utilize **2 PCs**





PCA - Visual Representation





PCA - Factor Loadings

In this slide the factor loadings are presented

- **What is it?** Correlation between Factors and PCs
- **Why is it used?** To understand which variables are most strongly correlated with each component

Factors	PC1	PC2
Revenue	**-0.5900331**	0.02583083
Results	**-0.5547107**	-0.19413409
ROA	-0.1122168	**-0.70330804**
Immobilizzazioni_Immateriali	**-0.5597218**	0.14486705
Assets	-0.1351886	**0.66784434**

* above are highlighted the most strongly correlated variables to each component

04 CONCLUSIONS

Conclusions

- **We identified 5-8 clusters based on similarity**
  - Different combinations of company size, results and innovative capability

Cluster Number	Revenue	Results	ROA	Immobilizz	Assets
2	719.857	48.830	0,15	142.809	4.834.128
4	100.015	- 583	-	49.261	132.094.717
3	8.500	- 285	- 0,05	1.836	132.577
5	5.007	619	0,15	1.075	28.964
1	3.659	112	0,06	611	75.796

*based on cluster medians

Cluster Num	Revenue	Results	ROA	Immobilizzaz	Assets
2	Very High	Very high	Very High	Very High	High
4	High	Very Low	Medium	High	Very High
3	Medium	Low	Low	Medium	Medium
5	Low	High	Very High	Low	Very Low
1	Very Low	Medium	Medium	Very Low	Low

- **We identified 2 principal components,** which are most strongly correlated to (please see below):
  - Dim 1: Assets, ROA
  - Dim 2: Revenue, Imm_Immateriali, Results
- **We also ran an analysis of Logistic regression** but this was not statistically significant due to p-value not statistically significant at alpha 0.05

Thank you for your attention