Title

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

Team

Ekaterina Kirillova

Kevin Pirazzi

Appendix

Abstract

The purpose of this study was to identify the group of firms investing in the technologies from Industry 4.0 in Trento and Belluno. It was assumed that if a firm has an increase of assets and high return on assets, most likely it invests in the technologies from Industry 4.0. For this study the cluster analysis and PCA analysis were conducted.However, the clusters obtained did not help to identify the target group of firms. That could have happened due to the choice of wrong variables for the analysis, or the wrong hypothesis. Therefore, further readings are needed in order to change the hypothesis and/or better identify the variables that can capture the investments in Industry 4.0.

Introduction

Industry 4.0 refers to a new phase in the Industrial Revolution that focuses heavily on interconnectivity, automation, machine learning, and real-time data. Industry 4.0, also sometimes referred to as IIoT or smart manufacturing, marries physical production and operations with smart digital technology, machine learning, and big data to create a more holistic and better connected ecosystem for companies that focus on manufacturing and supply chain management.

The implementation of Industry 4 technologies promises significant benefits for companies, being improved productivity, improved efficiency, cost reduction, creation of new innovative opportunities and others. Nevertheless, many manufacturing enterprises are still struggling to understand what Industry 4.0 implementation really means to them, and there is a lack of information on the adoption rate and on the benefits gained by the adopters.

This paper represents the initial phase of the research aiming at assessing the implementation of the technologies from Industry 4.0 in Italy. The territories in focus of the research are Trento and Belluno.

In order to identify companies investing into technologies from Industry 4.0, the following hypothesis was created:

H1: 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.

The analysis presented in this paper is focused on identification of these companies.

The dataset:

The initial dataset contains the information on the 224 mechatronics enterprises located in Trento and Belluno. The data is taken from the companies Annual Statements of Financial Performance (Balance Sheets) from 2013 to 2019.

Among the variables in our dataset we decided to select the most relevant variables for the study, being: 1) Revenue, 2) Yearly results, 3) Return on Assets (ROA), 4) Intangible assets (Immobilizzazioni_Immateriali), and 5) Total assets. For the logistic regressiona analysis we also utilized the binary variable High_ROA identified as two distinct ranges of ROAs within the data. Then the 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. Directly in the excel file prior to the input in R, we decided to remove the companies for which we had no observations for at least one year for each one of the variables. Applying this selection criteria we obtained 202 lines or companies which we analyzed further.

Here a summary of the variables included:

- Revenue à Firm Total Revenue in Thousands
- Results à Firm Net Income (in Thousands)
- ROA à Return on Assets (%)
- Immobilizzazioni_Immateriali (in Thousands)
- Assets (in Thousands)
- High_ROA (only used for logistic regression) à Binary variable 1 or 0

The variables contained in the data set were of different units of measured, some being in thousands of euros (financial) and some being percentages (the indicators, such as the ROA). In order to account for this we standardized the data and scaled it with mean zero and one in one standard deviation.

Methodologies:

For the purpose of this assignment we underwent the following statistical procedures:

- Clustering,
- Principal Component Analysis (also referred to as “PCA”) and,
- Logistic Regression

Please see the next pages for a summary of the analysis we underwent.

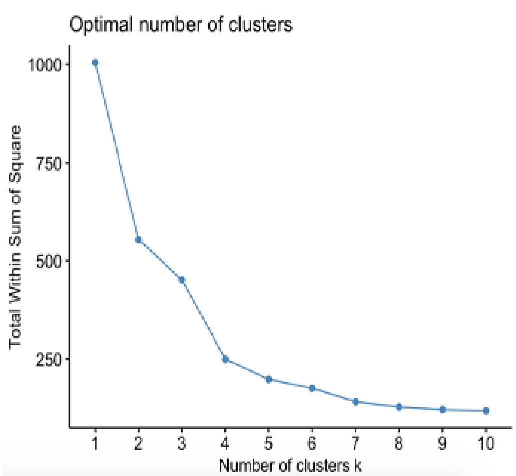
First Statistical Procedure Applied - Clustering

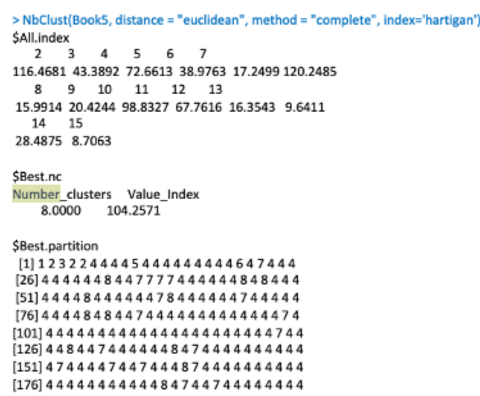
After normalizing the dataset as described in the previous page we conducted cluster analysis.

In order to identify the number of clusters the Elbow Method was used and the hartigan index (please see below the R output screenshots). The goal is to minimize the within the cluster sum of squares and maximize the distance between clusters. Here we can see the drop in the sum of squared distance which starts to slow down after k = 5-8.

To check the result, Silhouette scores and Calinski-Harabasz scores were computed.

Number of clusters	Silhouette score	Calinski-Harabasz score
5	0.4042018356126361	215.69947205535104
8	0.4040726333291286	323.56926308557064

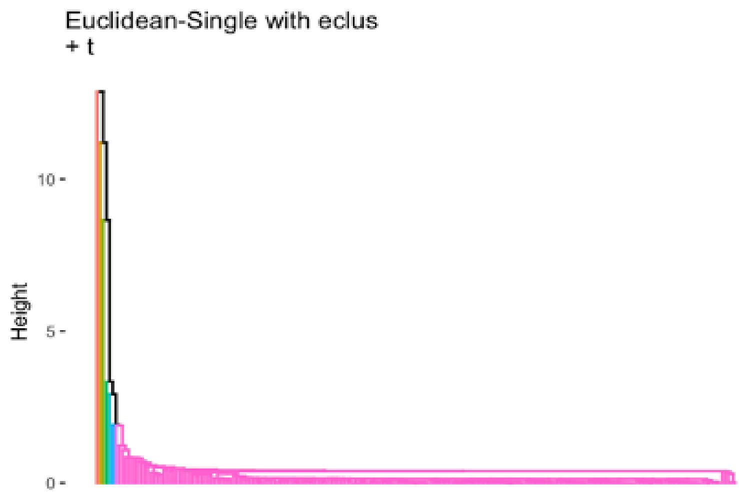




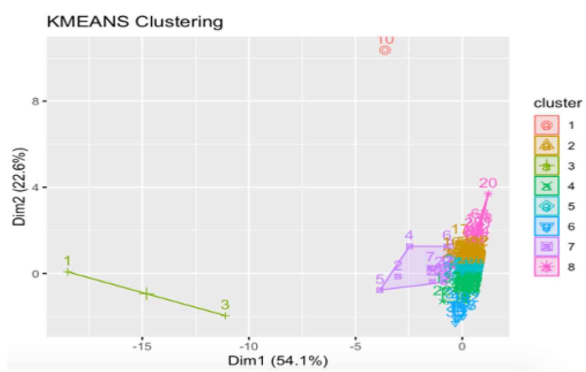
```
> NbClust(Book5, distance = "euclidean", method = "complete", index='hartigan')
$All.index
     2        3        4        5        6        7
116.4681  43.3892  72.6613  38.9763  17.2499 120.2485
     8        9       10       11       12       13
 15.9914  20.4244  98.8327  67.7616  16.3543   9.6411
    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 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 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 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 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 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 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 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 4
[201] 8 8
```

For the purpose of this exercise we decided to take 8 of clusters as shown by the Hartigan Index and then to visualize the data utilizing the k-clustering technique.

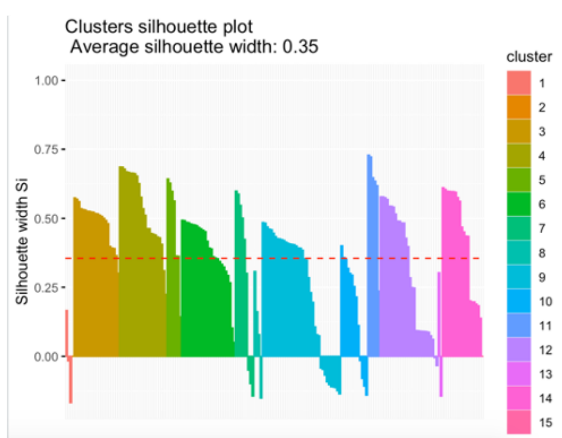
Below you can see the 8 clusters in terms of euclidean distances represented in a dendogram view:



Below you can see the clusters in term of k-clusters for representation:



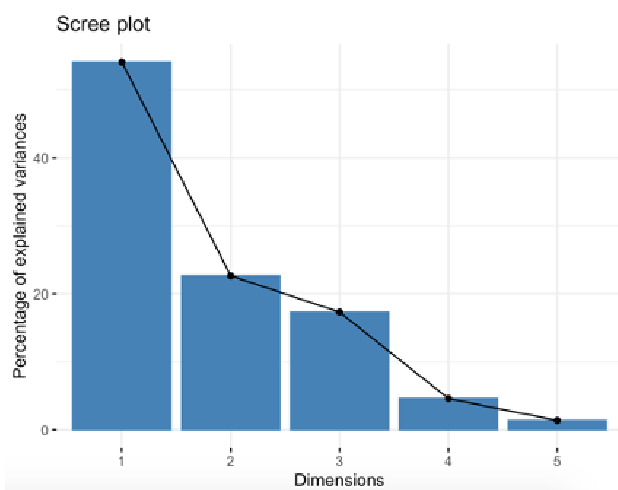
Below you can see the silhouette for the 8 clusters. The value seem optimal as all 8 scores are above the average silhouette score.



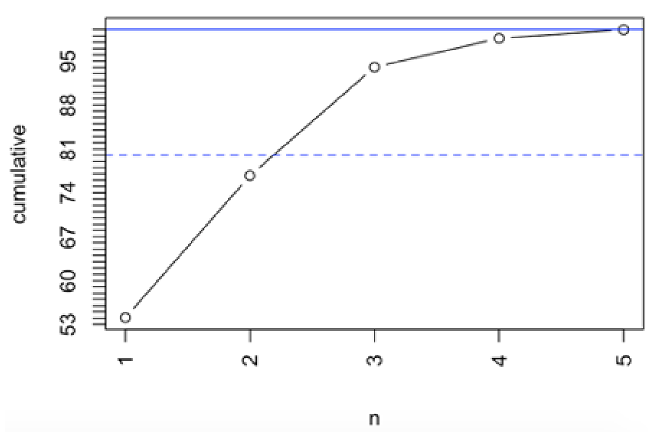
Second Statistical Procedure Applied - Principal Component Analysis

In order to dimensionally reduce the dimensions of the dataset we utilized the PCA analysis in an attempt to capture a large part of the variation in the data. In practice PCA maximizes variance and reduces dimensions in an attempt to minimize information loss (under a new coordinate system and includes a rotation in the data).

Please see below for the elbow method. We selected 2 Principal components as the helbow method highlights 2 components.



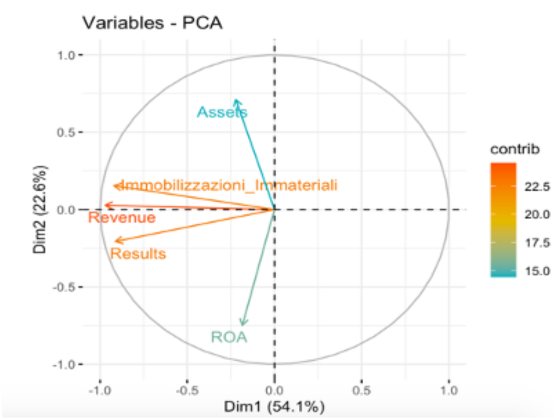
As confirmation of the elbow method we ran the below. Please see below for a selection based on 80% (>75% - not exactly 80 %) explanatory power of the data (around 2 PCs).

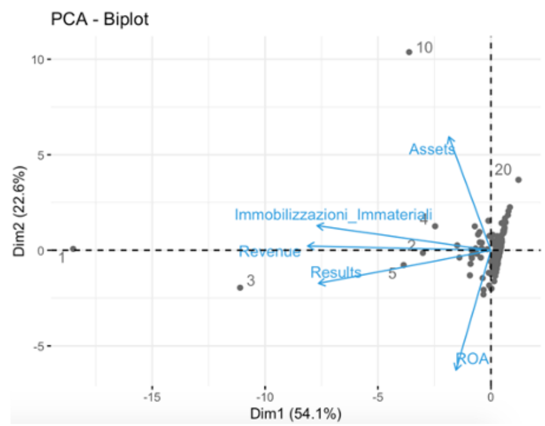


After re-centering and rotating the axis of the data we obtained the following two dimensional plot plotting the principal components

> loadings <- res$rotation

> loadings





Please see below for the factor loadings for the two PCs

	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

Third Statistical Procedure Applied - Logistic Regression

In this case with a Binary Logistic Regression Model where HIGH_ROA represents (1: instance, 0: non-instance). 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.

Sample function:

$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ **--> 1.43 e-01 - 2.65e-07**

In this case we did not proceed further as p-value >0.05 is not statistically significant. These columns provide the z-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0. If you use a 2-tailed test, then you would compare each p-value to your preselected value of alpha. Coefficients having p-values less than alpha are statistically significant.

Conclusions

The cluster analysis presented above indicates the presence of 8 groups. The dimensions were reduced to 2, as the elbow method highlighted 2 principal components that were ultimately used. We did not proceed with the regression analysis, as the p-value was not statistically significant.

Based on the conducted analysis, we did not manage to identify the target group of firms. Further, we tried to exclude the outliers and re-conduct the analysis, however, the obtained result has not significantly changed. Therefore, the hypothesis should be reconsidered as well as the choice of the variables. Further readings are necessary in order to better understand Industry 4.0 and the mechatronics sector to find out what parameters can help to identify whether the firm is investing in the technologies from Industry 4.0 or not.

In the next months the cluster analysis will be conducted again. When the target group of firms will be identified, we will extract the names of these companies and will try to reach the company's representatives for participation in the survey.

References

https://www.epicor.com/en/resource-center/articles/what-is-industry-4-0/

https://slcontrols.com/benefits-of-industry-4-0/

Appendix - This section contains the code utilized and ran in R

```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(readxl)
> Book2 <- read_excel("~/Desktop/Book2.xlsx")

> View(Book2)
> head(Book2)
# A tibble: 6 x 8
  Firm          Province Revenue Results   ROA
  <chr>         <chr>      <dbl>   <dbl> <dbl>
1 LUXOTTICA S.… Belluno   877983   45101  0.15
2 LUXOTTICA IT… Belluno   307562    6847  0.09
3 DANA ITALIA … Trento    561730   52558  0.15
4 MARCOLIN S.P… Belluno   217173    4258  0.02
5 ADIGE S.P.A.  Trento    130285   17542  0.12
6 DE RIGO VISI… Belluno   165461   -3873  0.02
# … with 3 more variables:
#   Immobilizzazioni_Immateriali <dbl>,
#   Assets <dbl>, HIGH_ROA <dbl>
> library(cluster)
> library(factoextra)
Loading required package: ggplot2
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
> library(NbClust)
> Book5 <- Book2[,3:7]
> Book5 <- scale(Book5)
> head(Book5)
       Revenue    Results        ROA
[1,] 10.779417  8.4388173  1.1102441
[2,]  3.599940  1.1260388  0.3159018
[3,]  6.798969  9.8643255  1.1102441
[4,]  2.462279  0.6311158 -0.6108309
[5,]  1.368682  3.1705352  0.7130730
[6,]  1.811417 -0.9232367 -0.6108309
     Immobilizzazioni_Immateriali
[1,]                   12.9778405
[2,]                    0.3501911
[3,]                    2.6106379
[4,]                    1.0446010
[5,]                    2.1759215
[6,]                    0.1921022
           Assets
[1,]  0.503865830
[2,]  0.241507368
[3,]  0.281302728
[4,]  1.109148438
[5,] -0.004539155
[6,]  0.859366642
> pairs(Book5)

> help(hclust)
> eu_Book <- dist(Book5, method='euclidean')
> hc_single <- hclust(eu_Book, method='single')
> hc_complete <- hclust(eu_Book, method='complete')
> hc_average <- hclust(eu_Book, method='average')
> hc_centroid <- hclust(eu_Book, method='centroid')
> str(hc_single)
List of 7
 $ merge      : int [1:201, 1:2] -172 -171 -149 -191 -176 -137 -182 -179 -143 -183 ...
 $ height     : num [1:201] 0.00192 0.00197 0.00365 0.00367 0.00519 ...
 $ order      : int [1:202] 10 1 3 20 5 2 4 6 17 7 ...
 $ labels     : NULL
 $ method     : chr "single"
 $ call       : language hclust(d = eu_Book, method = "single")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

```
> head(hc_single$merge)
     [,1] [,2]
[1,] -172 -180
[2,] -171 -184
[3,] -149 -162
[4,] -191    1
[5,] -176 -185
[6,] -137 -153
> fviz_dend(hc_single, as.ggplot = TRUE, show_labels = FALSE, main='Euclidean-Single')
> fviz_dend(hc_complete, as.ggplot = TRUE, show_labels = FALSE, main='Euclidean-Complete')
> fviz_dend(hc_centroid, as.ggplot = TRUE, show_labels = FALSE, main='Euclidean-Centroid')
> cluster_k <- cutree(hc_complete, k = 2) #identifying 2 groups
> fviz_dend(hc_complete, k = 2, k_colors = "jco", as.ggplot = TRUE, show_labels = FALSE, main='Euclidean-Distance')
> cluster_k
  [1] 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [26] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [51] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [76] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[101] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[126] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[151] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[176] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[201] 2 2
> pairs(Book5, col=cluster_k)
> cluster_h <- cutree(hc_complete, h = 4.1)
> fviz_dend(hc_complete, h = 4.1, k_colors = "jco", as.ggplot = TRUE, show_labels = FALSE, main='Euclidean-Complete')
> cluster_h
  [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 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 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 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 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 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 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 4
[201] 8 8
> set.seed(123)
> fviz_nbclust(Book5, kmeans, method = "wss")
> set.seed(123)
> fviz_nbclust(Book5, hcut, method = "wss")
> library(NbClust)
> 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 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 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 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 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 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 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 4
[201] 8 8

> library(NbClust)
> NbClust(Book5, distance = "euclidean", method = "kmeans", index='hartigan')
$All.index
        2         3         4         5         6 
  45.6984   31.4105  170.2708   37.4412   48.0076 
        7         8         9        10        11 
   8.6958    9.4272   19.7120    8.6794  238.1804 
       12        13        14        15 
  61.4901 -115.1471   12.2313  422.6439 

$Best.nc
Number_clusters     Value_Index 
        15.0000        410.4126 

$Best.partition
  [1] 14  6 14  6  6  5 15  5 15 10 15 15 15 15 15  5  5
 [18] 15 13  7 15  2 12 15 15 15 15  3  1  4 13  8 13  1
 [35]  9  2  2  9  1 12 12 13 13 13  8  4 11  1  3  3 13
 [52]  1 11  3 11 13  3  1  1 13 12  2  8 13 12 11 12  1
 [69] 13  9  3 12  1 13 13  4  1  1 13 11 13  8 12  4  2
 [86]  1 13 12 12 12 12  1 12  1  3  3  1 13  2  4 12 13
[103] 12 13 12 13  1  4  4 13 13 13 12  4  1 12 12  4 12
[120]  4 13 12  2  1  3  3  3 11  4  1  2 12 13  1  1 13
```

```
[137] 12  8 12  2 12  1  1  4 12 12 13 12  4 12  2  2 12
[154] 13  4 13  9 13  1  2  3  4  4  8  9  1  2 11 13  1
[171]  1 12 12 13 13  1  1 13 12 12 13 12 12  1  1  3 11
[188]  3  2 13 12  9 12  2  3  3  1 13 13 12  8 11

> fviz_nbclust(Book5, kmeans, method = "silhouette")+
+ labs(subtitle = "Silhouette method")
> res <- kmeans(Book5, 15)
> pairs(Book5, col=res$cluster)
> res <- kmeans(Book5, 15)
> str(res)
List of 9
 $ cluster     : int [1:202] 13 1 13 1 1 8 10 8 10 2 ...
 $ centers     : num [1:15, 1:5] 2.477 0.988 -0.2 -0.218 -0.196 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:15] "1" "2" "3" "4" ...
  .. ..$ : chr [1:5] "Revenue" "Results" "ROA" "Immobilizzazioni_Immateriali" ...
 $ totss       : num 1005
 $ withinss    : num [1:15] 9.42 0 0.304 0.212 1.343 ...
 $ tot.withinss: num 90.8
 $ betweenss   : num 914
 $ size        : int [1:15] 3 1 22 23 7 26 9 4 38 13 ...
 $ iter        : int 5
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
> pairs(Book5, col=res$cluster)
> hc_res <- eclust(Book5, "hclust", k = 8, hc_metric = "euclidean", hc_method = "single")
> str(hc_res)
List of 12
 $ merge      : int [1:201, 1:2] -172 -171 -149 -191 -176 -137 -182 -179 -143 -183 ...
 $ height     : num [1:201] 0.00192 0.00197 0.00365 0.00367 0.00519 ...
 $ order      : int [1:202] 10 1 3 20 5 2 4 6 17 7 ...
 $ labels     : NULL
 $ method     : chr "single"
 $ call       : language stats::hclust(d = x, method = hc_method)
 $ dist.method: chr "euclidean"
 $ cluster    : int [1:202] 1 2 3 4 5 6 6 6 6 7 ...
 $ nbclust    : num 8
 $ silinfo    :List of 3
  ..$ widths         :'data.frame':           202 obs. of  3 variables:
  .. ..$ cluster  : Factor w/ 8 levels "1","2","3","4",..: 1 2 3 4 5 6 6 6 6 6 ...
  .. ..$ neighbor : num [1:202] 3 4 5 2 4 4 4 4 4 4 ...
  .. ..$ sil_width: num [1:202] 0 0 0 0 0 ...
  ..$ clus.avg.widths: num [1:8] 0 0 0 0 0 ...
  ..$ avg.width      : num 0.619
 $ size       : int [1:8] 1 1 1 1 1 195 1 1
 $ data       : num [1:202, 1:5] 10.78 3.6 6.8 2.46 1.37 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:5] "Revenue" "Results" "ROA" "Immobilizzazioni_Immateriali" ...
  ..- attr(*, "scaled:center")= Named num [1:5] 2.15e+04 9.57e+02 6.61e-02 4.27e+03 1.14e+06
  .. ..- attr(*, "names")= chr [1:5] "Revenue" "Results" "ROA" "Immobilizzazioni_Immateriali" ...
  ..- attr(*, "scaled:scale")= Named num [1:5] 7.95e+04 5.23e+03 7.55e-02 1.78e+04 9.40e+06
  .. ..- attr(*, "names")= chr [1:5] "Revenue" "Results" "ROA" "Immobilizzazioni_Immateriali" ...
 - attr(*, "class")= chr [1:3] "hclust" "hcut" "eclust"
> hc_res$cluster
  [1] 1 2 3 4 5 6 6 6 6 7 6 6 6 6 6 6 6 6 6 8 6 6 6 6 6
 [26] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
 [51] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
 [76] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
[101] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
[126] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
[151] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
[176] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
[201] 6 6
> fviz_dend(hc_res, as.ggplot = TRUE, show_labels = FALSE, main='Euclidean-Single with eclus
+ + t')
> km_res <- eclust(Book5, "kmeans", k = 8, hc_metric = "euclidean")
> distance <- dist(Book5, method="euclidean")
> sil <- silhouette(x = res$cluster, dist = distance)
> sil[1:5,]
     cluster neighbor   sil_width
[1,]      13        1  0.30258988
[2,]       1       10  0.16596083
[3,]      13        1 -0.14425485
[4,]       1        8 -0.16794284
[5,]       1       10 -0.01651362
> fviz_silhouette(sil)
   cluster size ave.sil.width
1        1    3         -0.01
2        2    1          0.00
3        3   22          0.49
4        4   23          0.54
5        5    7          0.47
6        6   26          0.38
7        7    9          0.30
8        8    4          0.10
9        9   38          0.24
```

```
10      10   13          0.17
11      11    6          0.60
12      12   28          0.31
13      13    2          0.08
14      14   19          0.43
15      15    1          0.00
```

PCA

```
> library(readxl)
> Book2 <- read_excel("~/Desktop/Book2.xlsx")
> View(Book2)
> head(Book2)
# A tibble: 6 x 8
  Firm   Province Revenue Results   ROA Immobilizzazion…
  <chr>  <chr>      <dbl>   <dbl> <dbl>            <dbl>
1 LUXOT… Belluno   877983   45101  0.15           234946
2 LUXOT… Belluno   307562    6847  0.09            10492
3 DANA … Trento    561730   52558  0.15            50671
4 MARCO… Belluno   217173    4258  0.02            22835
5 ADIGE… Trento    130285   17542  0.12            42944
6 DE RI… Belluno   165461   -3873  0.02             7682
# … with 2 more variables: Assets <dbl>, HIGH_ROA <dbl>
> library(mvtnorm)
> library(NbClust)
> library(factoextra)
> Book2<- Book2[, 3:7]
> head(Book2)
# A tibble: 6 x 5
  Revenue Results   ROA Immobilizzazioni_Immate…   Assets
    <dbl>   <dbl> <dbl>                    <dbl>    <dbl>
1  877983   45101  0.15                   234946  5.88e6
2  307562    6847  0.09                    10492  3.41e6
3  561730   52558  0.15                    50671  3.79e6
4  217173    4258  0.02                    22835  1.16e7
5  130285   17542  0.12                    42944  1.10e6
6  165461   -3873  0.02                     7682  9.22e6
> Book2 <- scale(Book2)
> head(Book2)
       Revenue    Results        ROA
[1,] 10.779417  8.4388173  1.1102441
[2,]  3.599940  1.1260388  0.3159018
[3,]  6.798969  9.8643255  1.1102441
[4,]  2.462279  0.6311158 -0.6108309
[5,]  1.368682  3.1705352  0.7130730
[6,]  1.811417 -0.9232367 -0.6108309
     Immobilizzazioni_Immateriali       Assets
[1,]                   12.9778405  0.503865830
[2,]                    0.3501911  0.241507368
[3,]                    2.6106379  0.281302728
[4,]                    1.0446010  1.109148438
[5,]                    2.1759215 -0.004539155
[6,]                    0.1921022  0.859366642
> help(prcomp)
> res <- prcomp(Book2, scale = TRUE)
> str(res)
List of 5
 $ sdev    : num [1:5] 1.644 1.064 0.93 0.479 0.262
 $ rotation: num [1:5, 1:5] -0.59 -0.555 -0.112 -0.56 -0.135 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:5] "Revenue" "Results" "ROA" "Immobilizzazioni_Immateriali" ...
  .. ..$ : chr [1:5] "PC1" "PC2" "PC3" "PC4" ...
 $ center  : Named num [1:5] 2.99e-18 1.24e-17 -2.33e-17 3.01e-18 -1.46e-17
  ..- attr(*, "names")= chr [1:5] "Revenue" "Results" "ROA" "Immobilizzazioni_Immateriali" ...
 $ scale   : Named num [1:5] 1 1 1 1 1
  ..- attr(*, "names")= chr [1:5] "Revenue" "Results" "ROA" "Immobilizzazioni_Immateriali" ...
 $ x       : num [1:202, 1:5] -18.5 -3.01 -11.11 -2.47 -3.86 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:5] "PC1" "PC2" "PC3" "PC4" ...
 - attr(*, "class")= chr "prcomp"
> get_eig(res)
      eigenvalue variance.percent
Dim.1 2.70369706        54.073941
Dim.2 1.13227575        22.645515
Dim.3 0.86545342        17.309068
Dim.4 0.22982153         4.596431
Dim.5 0.06875224         1.375045
      cumulative.variance.percent
Dim.1                    54.07394
Dim.2                    76.71946
Dim.3                    94.02852
Dim.4                    98.62496
```

```
Dim.5               100.00000
> fviz_eig(res)
> plot(get_eig(res)$cumulative.variance.percent, type='b', axes=F, xlab='n', ylab='cumulative')
> abline(h=100, col='blue')
> abline(h=80, lty=2, col='blue')
> box()
> axis(2, at=0:100,labels=0:100)
> axis(1,at=1:ncol(Book2),labels=1:ncol(Book2),las=2)
> loadings <- res$rotation
> loadings
                                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
                                PC3        PC4
Revenue                    -0.12857260 -0.0159003
Results                    -0.08947094 -0.6778357
ROA                         0.67903329  0.1700096
Immobilizzazioni_Immateriali -0.08396496  0.6950584
Assets                      0.71226942 -0.1681563
                                PC5
Revenue                    -0.79649857
Results                     0.43259879
ROA                        -0.05268523
Immobilizzazioni_Immateriali  0.41900937
Assets                      0.01018471
> fviz_pca_var(res,
+ col.var = "contrib",
+ gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
+ repel = TRUE )
> fviz_pca_biplot(res, repel = TRUE,
+ col.var = "#2E9FDF",
+ col.ind = "#696969"
+ )

Warning message:
ggrepel: 190 unlabeled data points (too many overlaps). Consider increasing max.overlaps
> plot.new()
Warning messages:
1: ggrepel: 195 unlabeled data points (too many overlaps). Consider increasing max.overlaps
2: ggrepel: 195 unlabeled data points (too many overlaps). Consider increasing max.overlaps
> par(mar = c(1,4,0,2), mfrow = c(4,1))
> for(i in 1:4)
+ {}
> barplot(loadings[,i], ylim = c(-1, 1))
> abline(h=0)
> }
```

LOGISTIC REGRESSION

```
Error: unexpected '}' in "}"
> library(readxl)
> library(caret)
Loading required package: lattice
> library(MASS)
> library(ggplot2)
> library(klaR)
> df <- read_excel("~/Desktop/Book2.xlsx")

> head(df)
# A tibble: 6 x 8
  Firm   Province Revenue Results   ROA Immobilizzazion…
  <chr>  <chr>      <dbl>   <dbl> <dbl>            <dbl>
1 LUXOT… Belluno   877983   45101  0.15           234946
2 LUXOT… Belluno   307562    6847  0.09            10492
3 DANA … Trento    561730   52558  0.15            50671
4 MARCO… Belluno   217173    4258  0.02            22835
5 ADIGE… Trento    130285   17542  0.12            42944
6 DE RI… Belluno   165461   -3873  0.02             7682
# … with 2 more variables: Assets <dbl>, HIGH_ROA <dbl>
> df <- df[,3:8]
> head(df)
# A tibble: 6 x 6
  Revenue Results   ROA Immobilizzazion… Assets HIGH_ROA
    <dbl>   <dbl> <dbl>            <dbl>  <dbl>    <dbl>
1  877983   45101  0.15           234946 5.88e6        1
2  307562    6847  0.09            10492 3.41e6        1
3  561730   52558  0.15            50671 3.79e6        1
4  217173    4258  0.02            22835 1.16e7        0
5  130285   17542  0.12            42944 1.10e6        1
6  165461   -3873  0.02             7682 9.22e6        0
> str(df)
tibble [202 × 6] (S3: tbl_df/tbl/data.frame)
 $ Revenue                     : num [1:202] 877983 307562 561730 217173 130285 ...
 $ Results                     : num [1:202] 45101 6847 52558 4258 17542 ...
 $ ROA                         : num [1:202] 0.15 0.09 0.15 0.02 0.12 0.02 0.05 0.02 0.1 0 ...
 $ Immobilizzazioni_Immateriali: num [1:202] 234946 10492 50671 22835 42944 ...
```

```
 $ Assets                     : num [1:202] 5880098 3414109 3788158 11569338 1101441 ...
 $ HIGH_ROA                   : num [1:202] 1 1 1 0 1 0 0 0 1 0 ...
> df$HIGH_ROA <- as.factor(df$HIGH_ROA)
> summary(df)
    Revenue          Results              ROA
 Min.   :    76   Min.   :-14893.0   Min.   :-0.25000
 1st Qu.:  1961   1st Qu.:    31.5   1st Qu.: 0.03000
 Median :  4464   Median :   131.5   Median : 0.06000
 Mean   : 21541   Mean   :   956.6   Mean   : 0.06614
 3rd Qu.: 10266   3rd Qu.:   367.0   3rd Qu.: 0.10000
 Max.   :877983   Max.   : 52558.0   Max.   : 0.30000
 Immobilizzazioni_Immateriali     Assets
 Min.   :     9.0             Min.   :      832
 1st Qu.:   264.2             1st Qu.:    25714
 Median :   773.5             Median :    66252
 Mean   :  4267.4             Mean   :  1144106
 3rd Qu.:  2514.8             3rd Qu.:   209223
 Max.   :234946.0             Max.   :132094717
 HIGH_ROA
 0:100
 1:102
```

```
> set.seed(123)
> library(magrittr)
> training_samples <- df$HIGH_ROA %>% createDataPartition(p = 0.75, list = FALSE)
> train <- df[training_samples, ]
> test <- df[-training_samples, ]
> simple_glm <- glm(HIGH_ROA ~ Assets, data = train, family = 'binomial')
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(simple_glm)

Call:
glm(formula = HIGH_ROA ~ Assets, family = "binomial", data = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.239  -1.223   1.118   1.122   1.808

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.431e-01  1.752e-01   0.817    0.414
Assets      -2.655e-07  1.893e-07  -1.403    0.161

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 210.69  on 151  degrees of freedom
Residual deviance: 206.20  on 150  degrees of freedom
AIC: 210.2

Number of Fisher Scoring iterations: 7

> simple_glm$coefficients
  (Intercept)        Assets
 1.430625e-01 -2.655324e-07
```