Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Consiglio Nazionale delle Ricerche

Data Analysis

Training FOSSR

"Data lifecycle and architectures in the social sciences"

FOSSR
Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

Nicolò Marchesini (CNR-IRPPS)

5 October 2023

Data analysis

- Overview
- Data management
- Descriptive analysis
- Likert scales + FA (PCA)

- R

Data analysis

- Data analysis is the **process** of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. [1]

[1] M.S.Brown (2014), Transforming Unstructured Data into Useful Information, in S. Kudyba (eds.) *Big Data, Mining, and Analytics*, Auerbach Publications, pp. 227–246, doi:10.1201/b16666-14

R + RStudio

- The R Project for Statistical Computing
  - Focusing on statistics
  - Opensource
  - https://www.r-project.org
- RStudio integrated development environment (IDE)
  - R and Python
  - User-friendly
  - https://posit.co/download/rstudio-desktop/

R packages

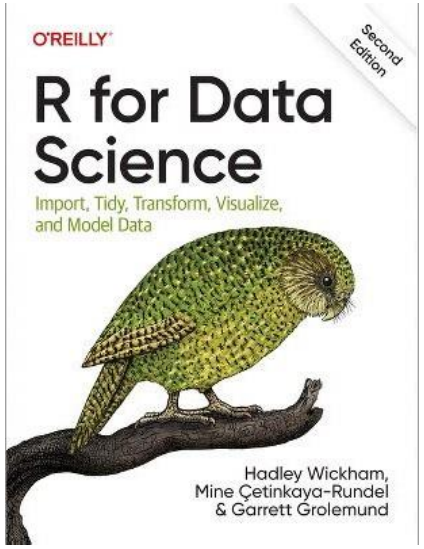
- tidyverse
- foreign
- summarytools
- gtsummary
- openxlsx
- likert
- FactoMineR
- factoextra

Descriptives

- **Descriptive statistics**: properly describe, represent and synthesise a set or sample of data
- 3 main aspects:
  1. distribution;
  2. central tendency;
  3. variability.

- Wickham H., Çetinkaya-Rundel M. & Grolemund G., *R for Data Science*, 2nd ed., O'Reilly, 2023

https://r4ds.hadley.nz



Descriptives/2

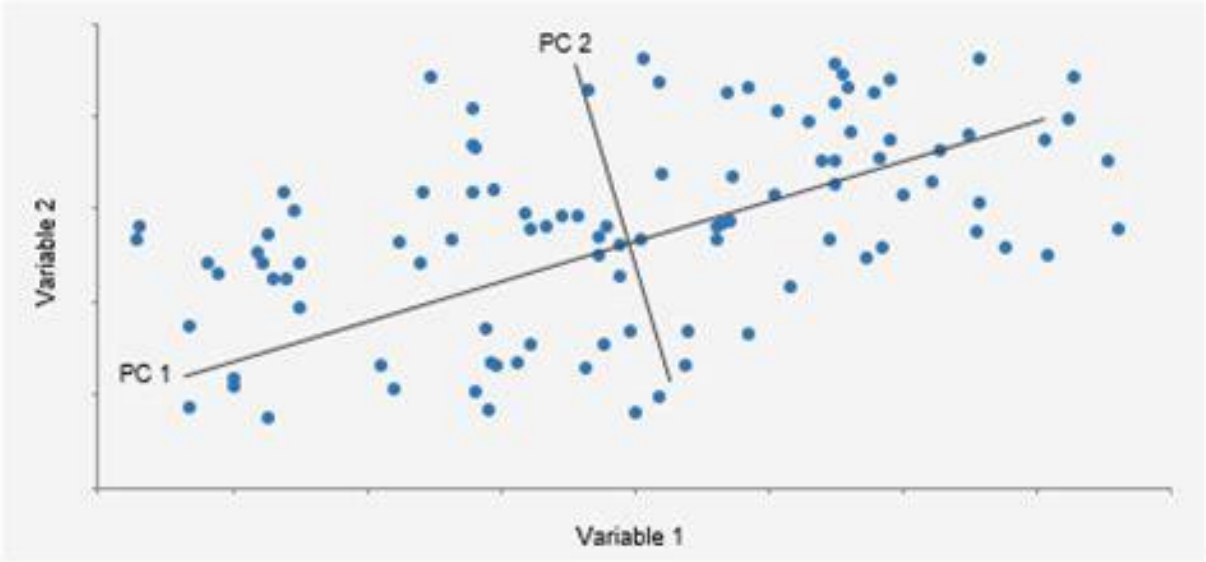
- Two kind of data… → Quantitative (numeric characteristics); Qualitative (qualities)
- …different types of vars → Numeric (discrete or continues): e.g. age; Categorical (dichotomus): e.g. sex, job; Ordinal: e.g. age class, educational level; Text

Likert Scale

- A Likert scale is a psychometric scale named after its inventor, American social psychologist Rensis Likert, which is commonly used in research questionnaires.
- **Collection of responses to a set of items** or the format in which responses are scored along a range.
- Range: steps, ordinality, distance between choices, balance.
- Items reduction: validation thanks to factor analysis (PCA)

Factor Analysis – Principal Component Analysis (PCA)

- **Exploratory** data analysis
- **Reducing the dimensionality** of a dataset while preserving as much of the data's variation as possible:
  n -> a few dimensions (e.g., <3)
- **Quantitative** & **scaled**
  (i.e., standardized) data



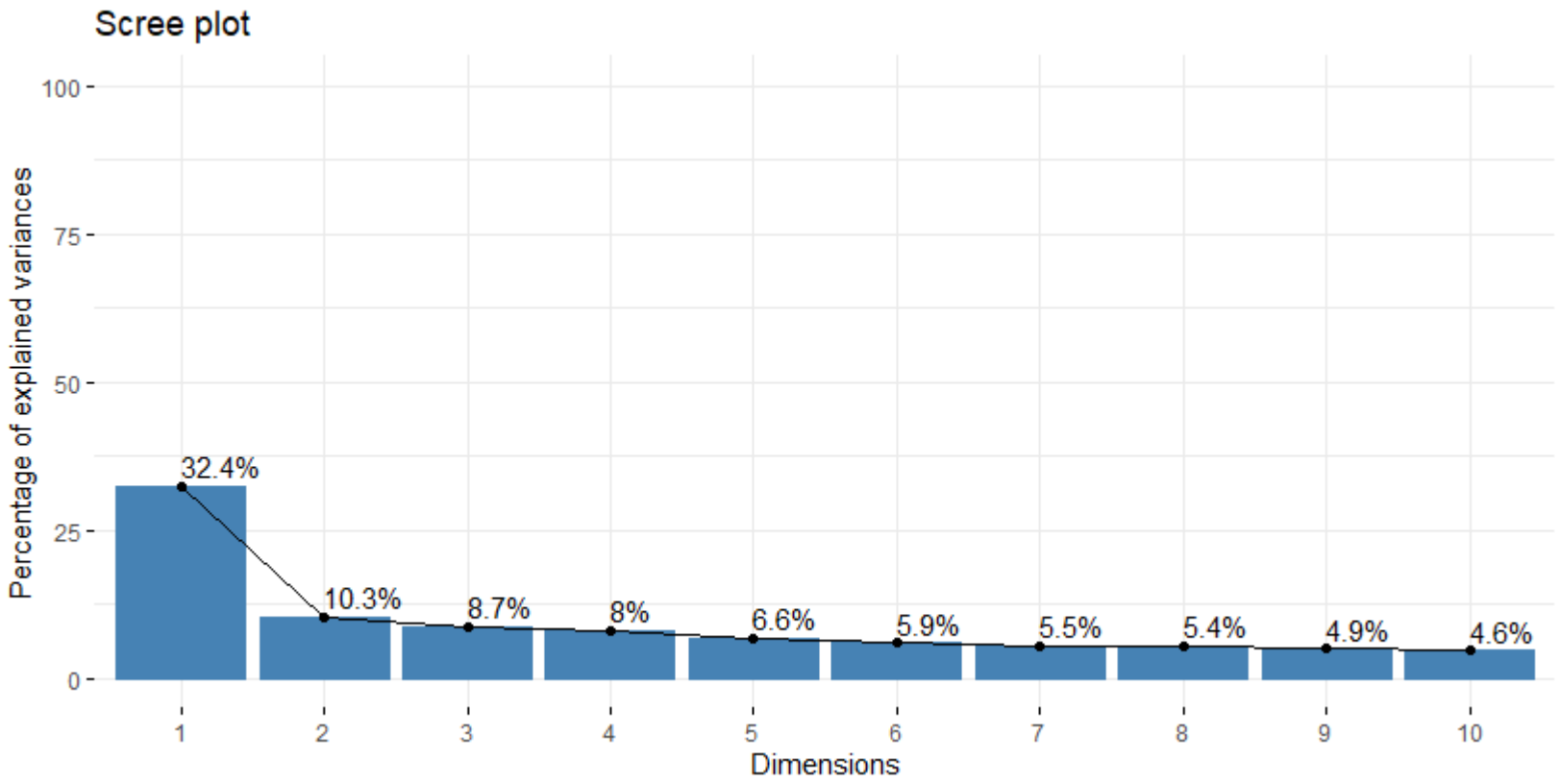
Factor Analysis – Principal Component Analysis (PCA)/2

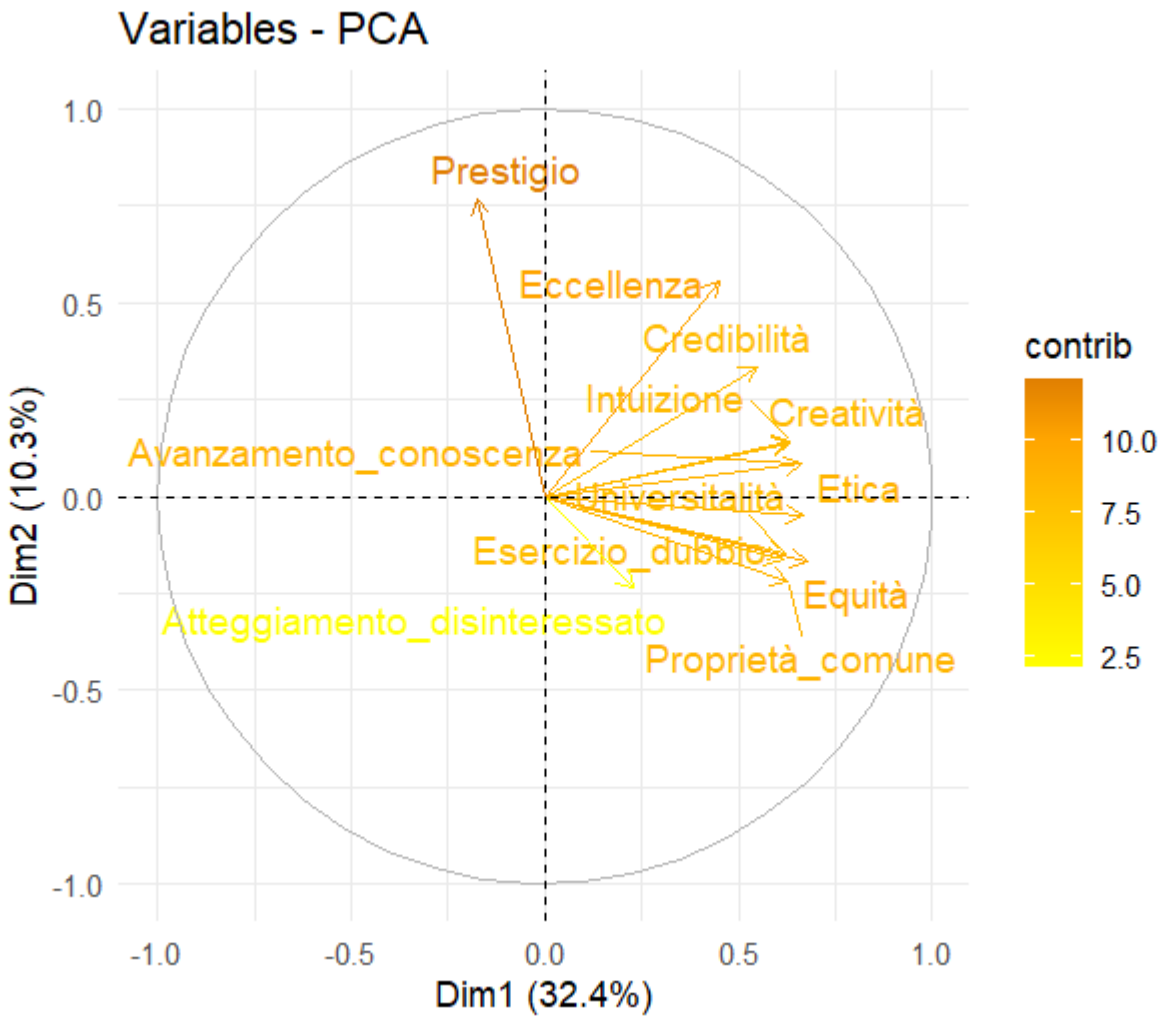
- Eigenvalues -> measure the amount of variation retained by each principal component. Large for the first PCs and small for the subsequent PCs. The first PCs corresponds to the directions with the maximum amount of variation in the data set.

```
> eig.val
       eigenvalue variance.percent cumulative.variance.percent
Dim.1   3.8886036        32.405030                    32.40503
Dim.2   1.2413539        10.344616                    42.74965
Dim.3   1.0487977         8.739981                    51.48963
Dim.4   0.9600009         8.000008                    59.48963
Dim.5   0.7960805         6.634004                    66.12364
Dim.6   0.7055162         5.879302                    72.00294
Dim.7   0.6599922         5.499935                    77.50287
Dim.8   0.6434543         5.362119                    82.86499
Dim.9   0.5920440         4.933700                    87.79869
Dim.10  0.5505189         4.587658                    92.38635
Dim.11  0.4703979         3.919982                    96.30633
Dim.12  0.4432399         3.693666                   100.00000
>
```

- An eigenvalue > 1 indicates that PCs account for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cutoff point for which PCs are retained. This holds true only when the data are standardized.

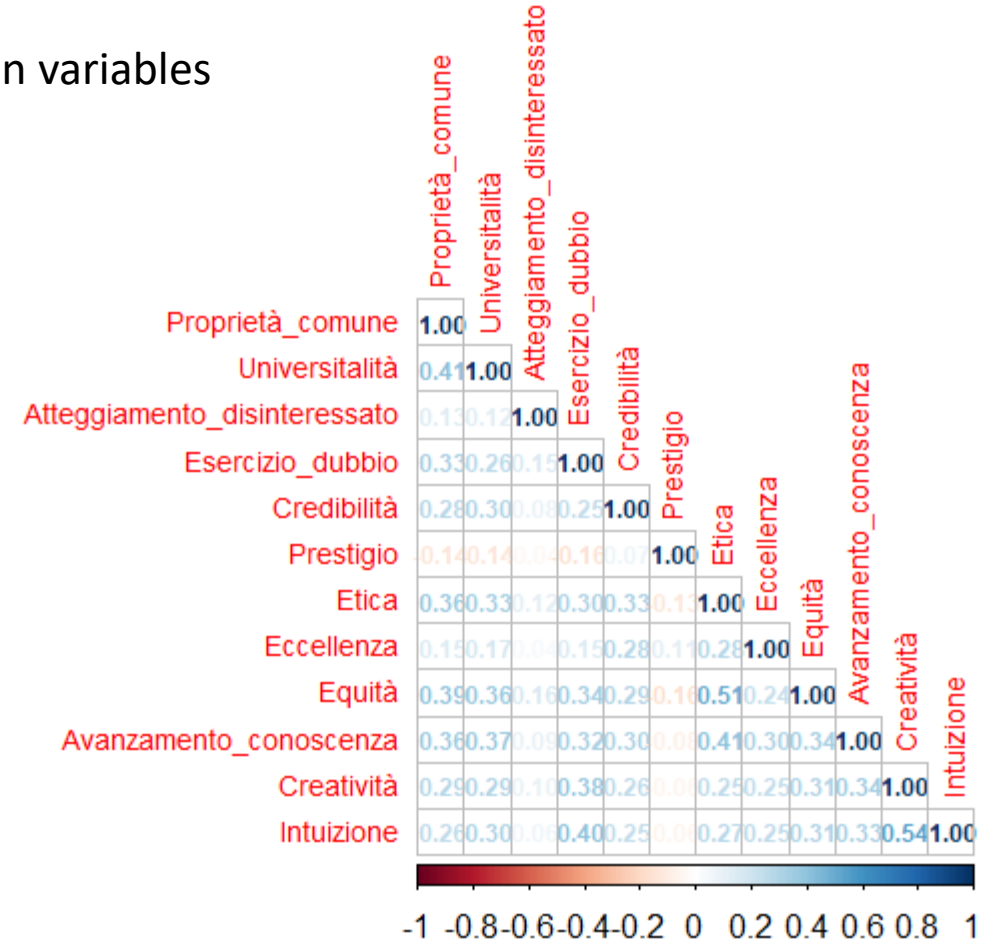
Factor Analysis – Principal Component Analysis (PCA)/3




Variables - PCA
Prestigio
Eccellenza
Credibilità
Intuizione
Creatività
Avanzamento_conoscenza
Universalità
Etica
Esercizio_dubbio
Equità
Atteggiamento_disinteressato
Proprietà_comune
contrib
10.0
7.5
5.0
2.5
Dim2 (10.3%)
Dim1 (32.4%)
1.0
0.5
0.0
-0.5
-1.0

Correlation

- Relationship between variables
- From -1 to 1



Factor Analysis – Principal Component Analysis (PCA)/6

Cronbach's Alpha

- is a reliability coefficient
- measure of the internal consistency of tests and measures
- From 0 to 1
- Alpha > 0,7 -> A high value of Cronbach's alpha indicates internal consistency

```
> CA

Cronbach's alpha for the 'df_scales_valid_num[, -1]' data-set

Items: 12
Sample units: 895
alpha: 0.77

Bootstrap 95% CI based on 1000 samples
 2.5% 97.5%
0.743 0.794
```

Item reduction

- inclusion or exclusion of items
- sum, mean, …
- description