
# Personas Creation

**PyLab #03 (Oct 25th 2024)**

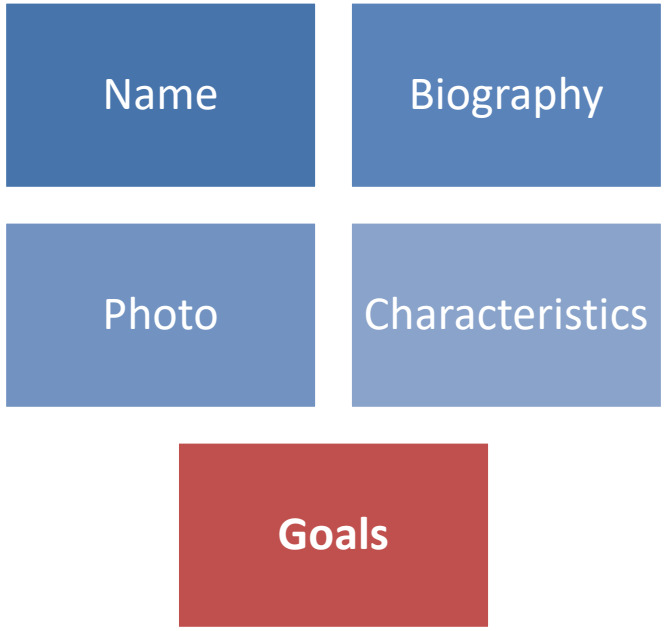
**Emanuele Tauro:** emanuele.tauro@polimi.it

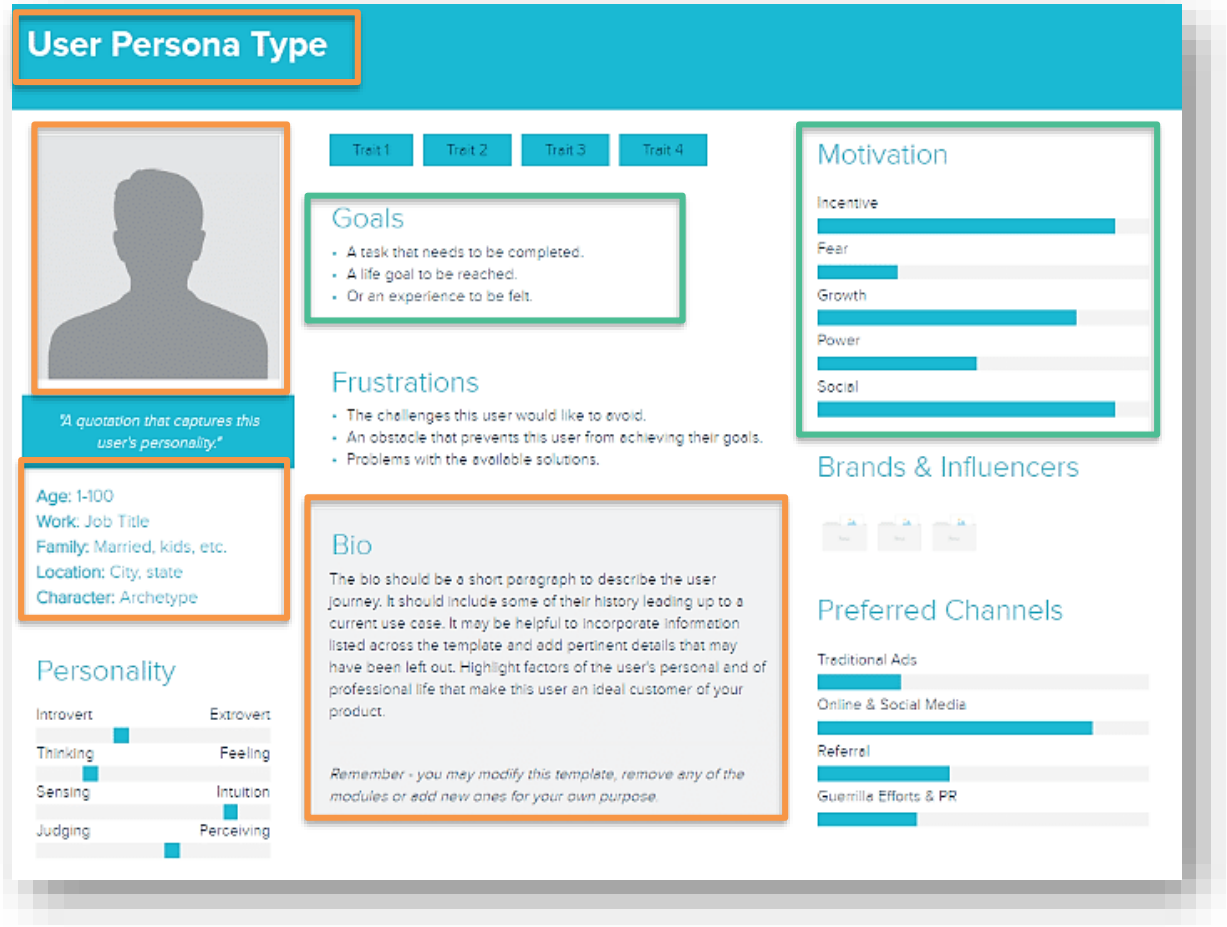
# Schedule

- **Personas Creation Workflow**
  - Data preparation
  - Preprocessing
  - Clustering
  - **How many Personas?**
  - Statistical Analysis
  - Personification

# Personas

*"Personas are not real people, but they represent them throughout the design process. Personas are defined by their **goals**."*
- Alan Cooper



# Personas Creation Workflow

- Goal Definition
- Survey creation
- Survey dissemination

**Data Collection**

- Data preparation
- Clustering
- Statistical Analysis

**Data Analysis**

- Persona Cards development
- Qualitative data collection

**Personification**

# Personas Creation Workflow

- Goal Definition
- Survey creation
- Survey dissemination

**Data Collection**

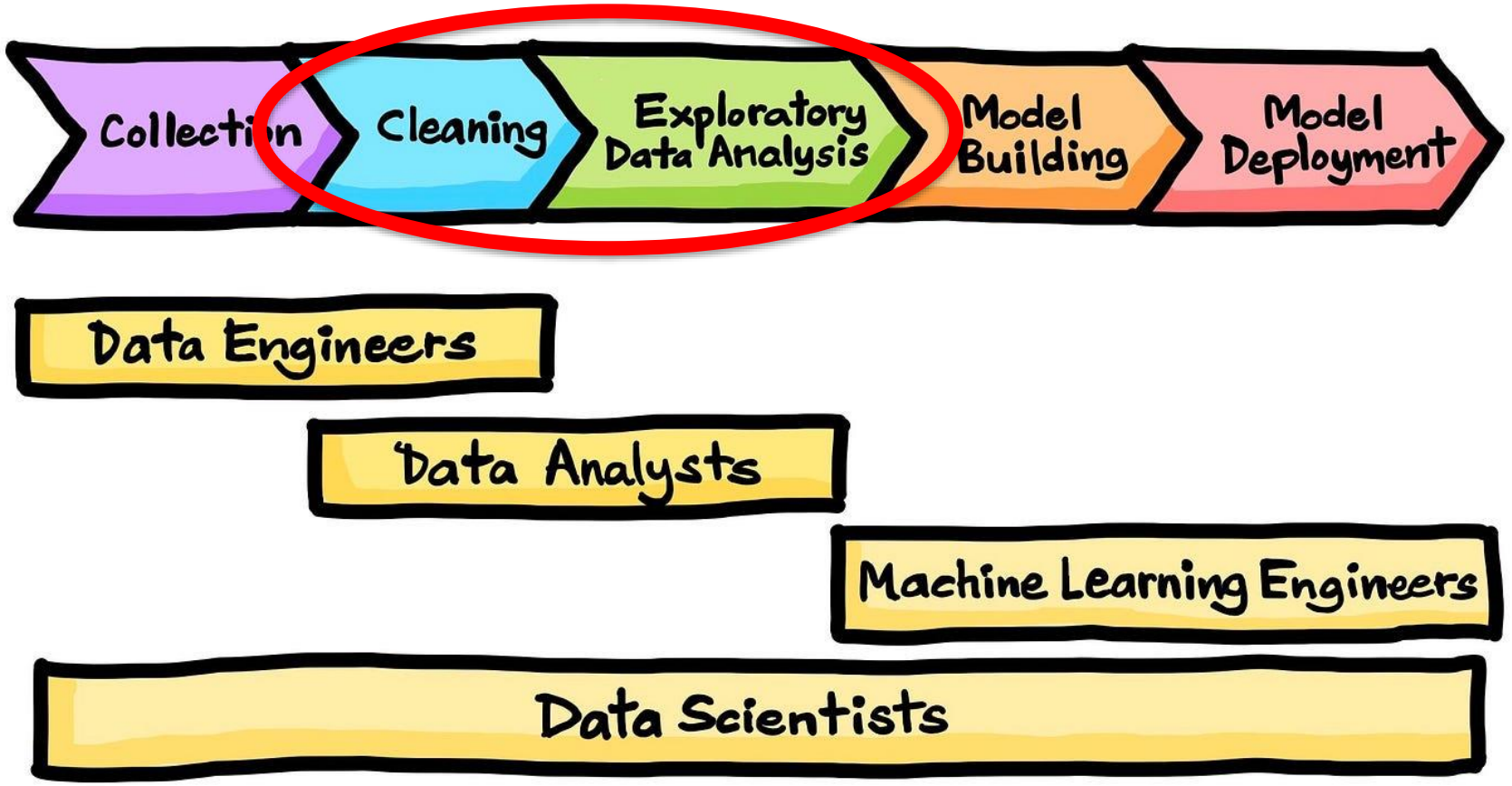
- Data preparation
- Clustering
- Statistical Analysis

**Data Analysis**

- Persona Cards development
- Qualitative data collection

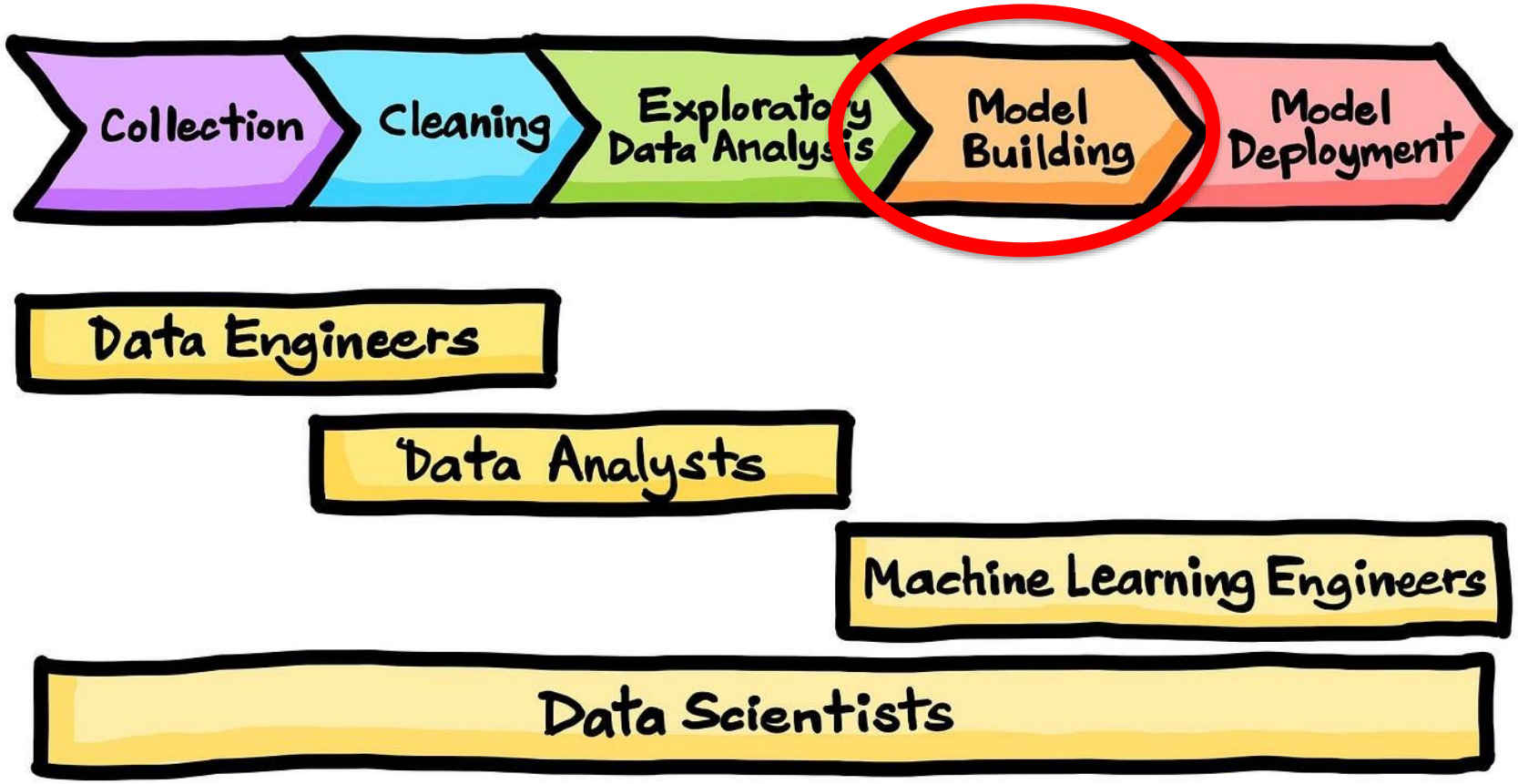
**Personification**

# Data Analysis: Data preparation




https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

# Data Analysis: Data preparation




https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

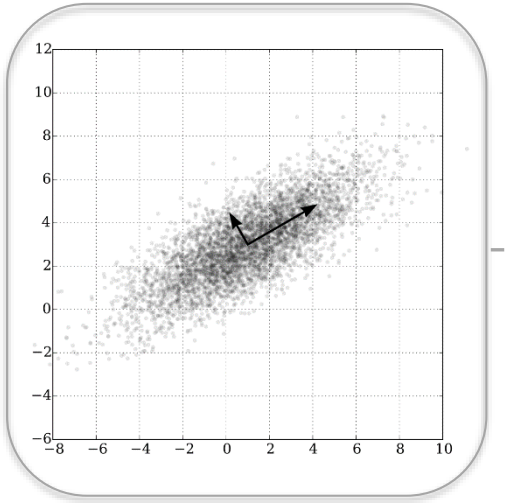
# Data Analysis: Data Preparation

Quantitative

Ordinal — Ordered values

| Age | Marital Status | Education |
|---|---|---|
| 35 | Married | Elementary |
| 48 | Engaged | Master Degree |
| 66 | Married | Master Degree |
| 51 | Single | High School |

Nominal — One Hot Encoding



Dimensionality Reduction Methods

**n x m**

↓

**n x p**

**p << m**

- Avoid the curse of dimensionality
- Find the most interesting variations in analyzed data
- Possibly plot data

# Data Analysis: Data Preparation

- **Principal Component Analysis (*PCA*):** most commonly used data reduction method. Works on numerical data and binary data only. Finds the principal components (linear combinations of variables) that explain most of the variance within the dataset. Implemented in **sklearn.**

- **Multiple Correspondence Analysis (*MCA*):** identical to PCA, but only for categorical variables. Implemented in **mca**.

- **Principal Component Analysis of Mixed Data (*PCAMIX*)**: a combination of PCA and MCA. It performs PCA on quantitative data and MCA on categorical data. Can assign weights to each variable. Implementation is exclusively available in R.

- **Factor Analysis of Mixed Data (*FAMD*)**: identical to PCAMIX. Has an implementation in python with the **prince** library. In R, it allows to divide the columns of the dataset into groups, differentiating it from PCAMIX. No weight is allowed.

- **Distance metrics**: when the data presents fewer records than features, it is possible to reduce its dimensionality by calculating a distance matrix. The distance metric used is of fundamental importance. For mixed datasets, with numerical and categorical data, one of the most used distance metrics is **Gower (*GW*).** Implemented in **gower.**

# Data Analysis: Data Preparation

*How many components do we have to use when performing PCA, PCAMIX, FAMD or MCA on our dataset?*
*It depends on the dataset you are analyzing.*

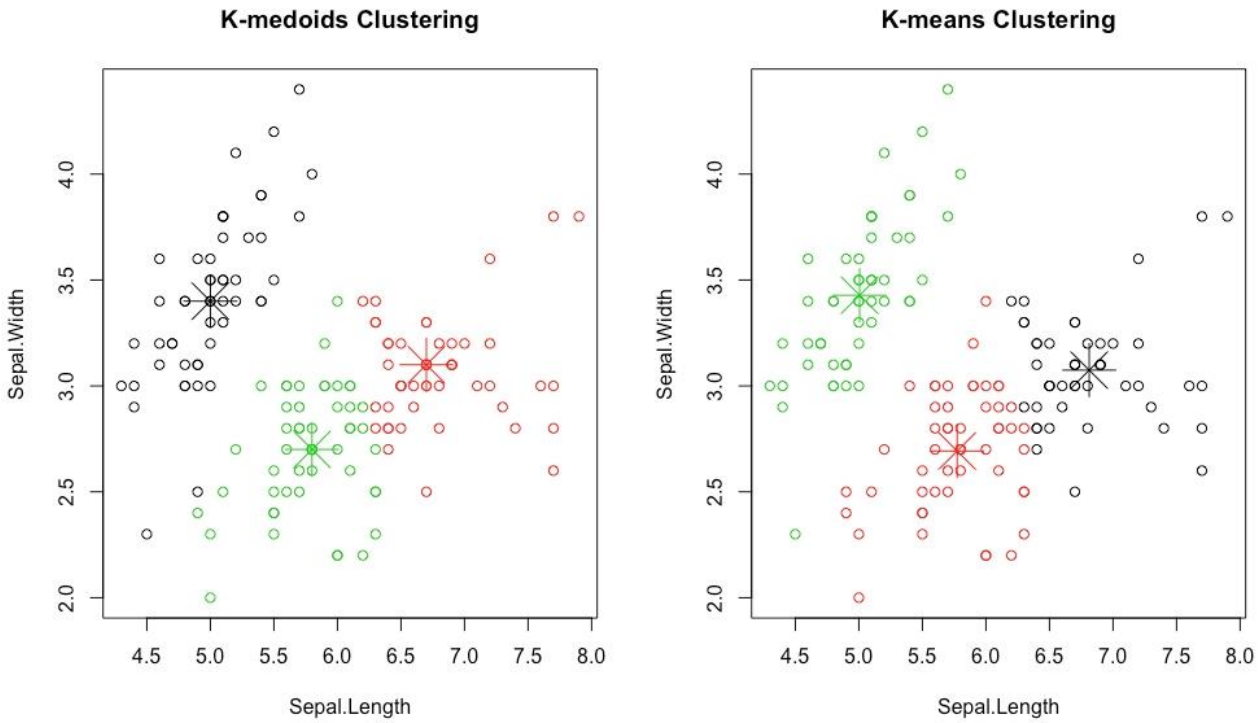
Each principal component obtained through these methods explains a certain **percentage** of variance within the dataset. Most times the focus is not on the number of principal components, but on the amount of variance explained by them. Reduced datasets should explain a percentage of variance from **70% to 90%** of the original dataset.

There is, however, no golden standard available. It is a **Trial and Error process.**

In other situations, you might be prompted to favour the number of components with respect to the percentage of variance. This might happen if the number of components that explain around 70% of the dataset variance is 4 or 5. With 3 components it is possible to plot data, gaining more visual insights, but at the same time information, in the form of variance explained, is lost.

# Data Analysis: Clustering

## Partition Clustering




### K-Medoids
### (Partitioning Around Medoids)

<u>Build</u>

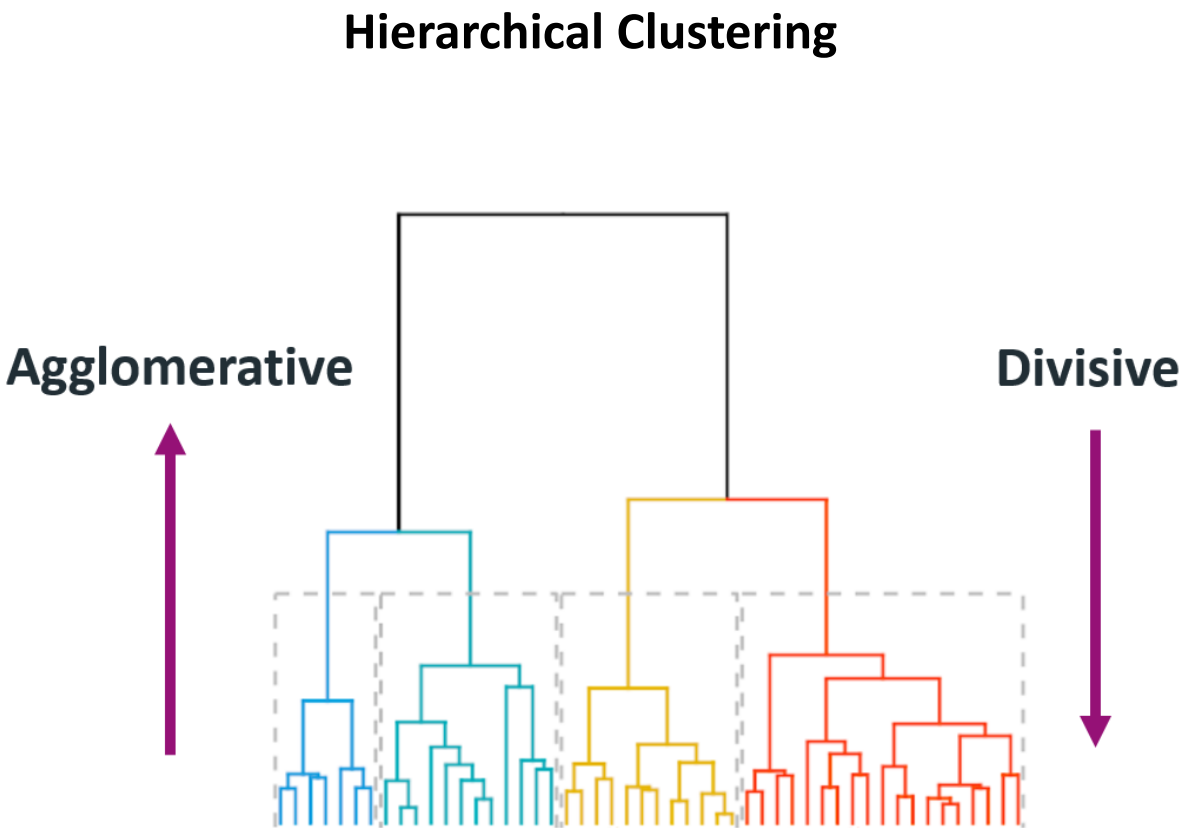
Assign each cluster a potential medoid.

<u>Swap</u>

Test each point as a medoid.
Assign each point to its closest medoid.

- The number of clusters **k** must be known a priori!
- Prone to find local minimum.
- Randomness counts.

# Data Analysis: Clustering




## Agglomerative Clustering

Calculate
Measure the distance between clusters.

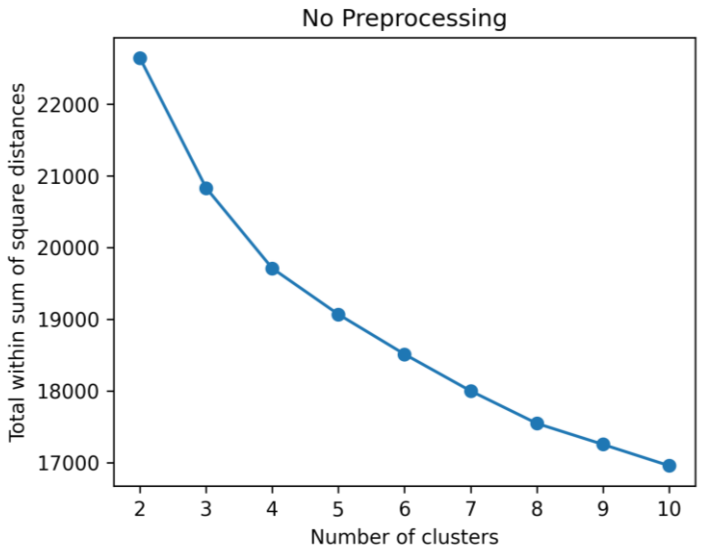
Merge
Join the two closest clusters into one.

- The **linkage** (method to calculate distance) must be decided a priori.
- Does not work well on mixed data types.
- Must decide where to cut the dendrogram.

# Defining the optimal number of clusters

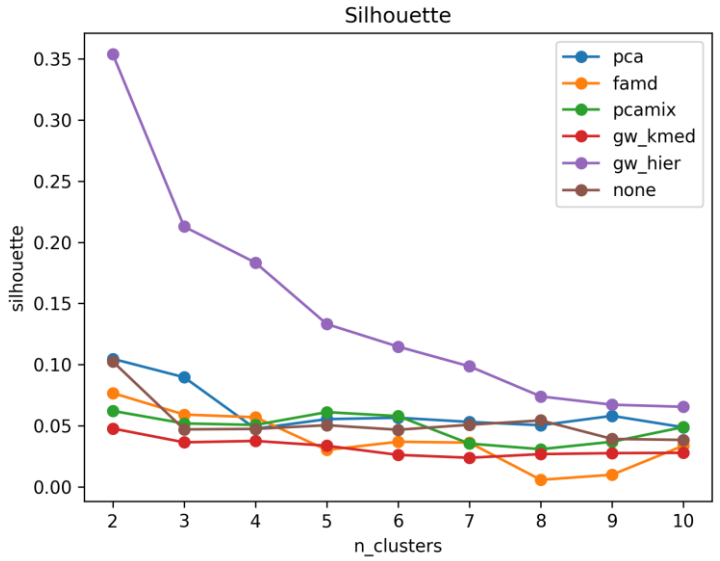
## Graphical Method

Calculating the total within sum of square distances (inertia) for a varying number of clusters, iti s possible to graphically identify the **elbow** in the figure. The number of clusters where the elbow is found is the optimal number of clusters to be used.




## Analytical Method

The **average silhouette** is a measure of how similar an object is to its cluster, compared to other clusters in the same partition. The higher the value, the better. However, this information has to be supported by other information (number of samples in each cluster, statistical difference between clusters).



# Data Analysis: Clustering

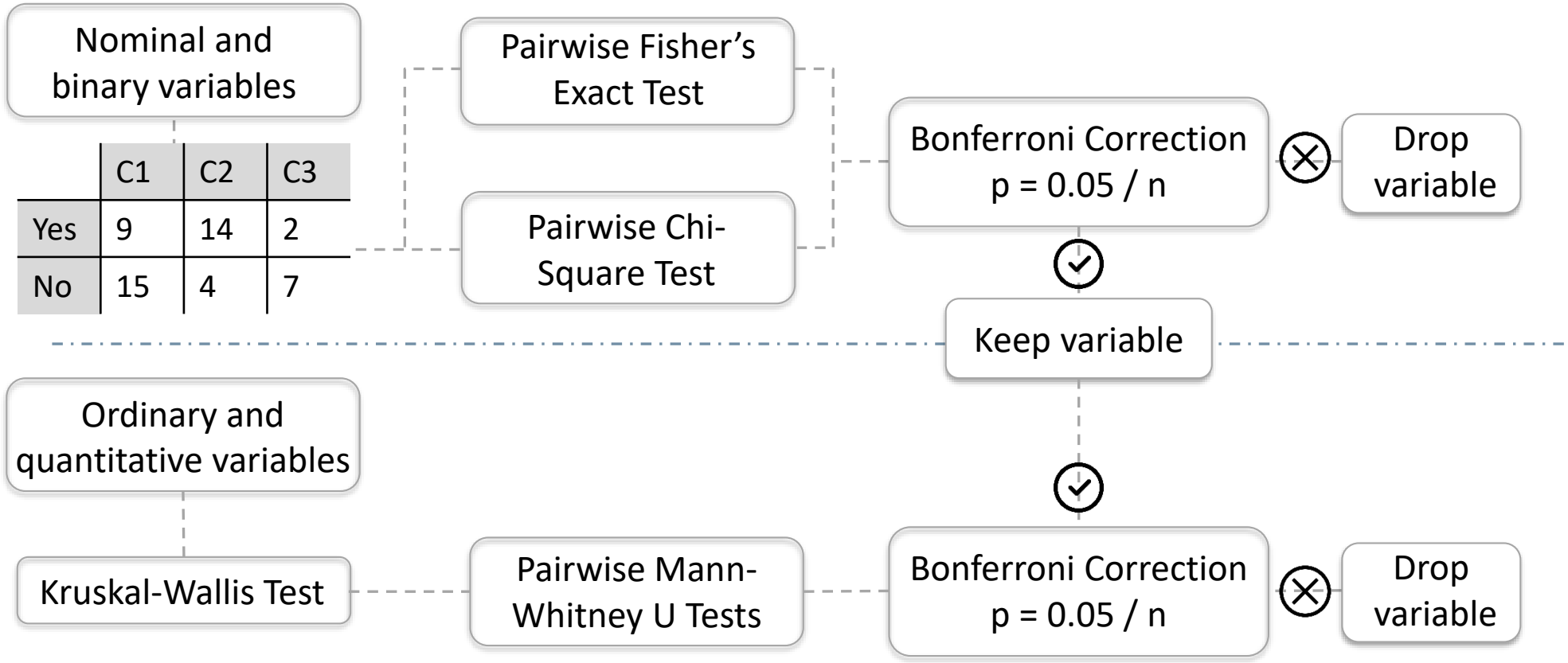
| Name | Type 1 | Type 2 | Total | HP | Attack | Defense | Sp. Atk | Sp. Def | Speed | Generation | Legendary | LABEL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bulbasaur | Grass | Poison | 318 | 45 | 49 | 49 | 65 | 65 | 45 | 1 | FALSO | 1 |
| Ivysaur | Grass | Poison | 405 | 60 | 62 | 63 | 80 | 80 | 60 | 1 | FALSO | 0 |
| Venusaur | Grass | Poison | 525 | 80 | 82 | 83 | 100 | 100 | 80 | 1 | FALSO | 1 |
| VenusaurMega Venusaur | Grass | Poison | 625 | 80 | 100 | 123 | 122 | 120 | 80 | 1 | FALSO | 1 |
| Charmander | Fire | | 309 | 39 | 52 | 43 | 60 | 50 | 65 | 1 | FALSO | 0 |
| Charmeleon | Fire | | 405 | 58 | 64 | 58 | 80 | 65 | 80 | 1 | FALSO | 0 |
| Charizard | Fire | Flying | 534 | 78 | 84 | 78 | 109 | 85 | 100 | 1 | FALSO | 1 |
| CharizardMega Charizard X | Fire | Dragon | 634 | 78 | 130 | 111 | 130 | 85 | 100 | 1 | FALSO | 1 |
| CharizardMega Charizard Y | Fire | Flying | 634 | 78 | 104 | 78 | 159 | 115 | 100 | 1 | FALSO | 0 |
| Squirtle | Water | | 314 | 44 | 48 | 65 | 50 | 64 | 43 | 1 | FALSO | 0 |
| Wartortle | Water | | 405 | 59 | 63 | 80 | 65 | 80 | 58 | 1 | FALSO | 1 |
| Blastoise | Water | | 530 | 79 | 83 | 100 | 85 | 105 | 78 | 1 | FALSO | 1 |
| BlastoiseMega Blastoise | Water | | 630 | 79 | 103 | 120 | 135 | 115 | 78 | 1 | FALSO | 1 |
| Caterpie | Bug | | 195 | 45 | 30 | 35 | 20 | 20 | 45 | 1 | FALSO | 1 |
| Metapod | Bug | | 205 | 50 | 20 | 55 | 25 | 25 | 30 | 1 | FALSO | 1 |
| Butterfree | Bug | Flying | 395 | 60 | 45 | 50 | 90 | 80 | 70 | 1 | FALSO | 1 |
| Weedle | Bug | Poison | 195 | 40 | 35 | 30 | 20 | 20 | 50 | 1 | FALSO | 0 |
| Kakuna | Bug | Poison | 205 | 45 | 25 | 50 | 25 | 25 | 35 | 1 | FALSO | 0 |
| Beedrill | Bug | Poison | 395 | 65 | 90 | 40 | 45 | 80 | 75 | 1 | FALSO | 0 |
| BeedrillMega Beedrill | Bug | Poison | 495 | 65 | 150 | 40 | 15 | 80 | 145 | 1 | FALSO | 1 |
| Pidgey | Normal | Flying | 251 | 40 | 45 | 40 | 35 | 35 | 56 | 1 | FALSO | 0 |
| Pidgeotto | Normal | Flying | 349 | 63 | 60 | 55 | 50 | 50 | 71 | 1 | FALSO | 1 |
| Pidgeot | Normal | Flying | 479 | 83 | 80 | 75 | 70 | 70 | 101 | 1 | FALSO | 0 |
| PidgeotMega Pidgeot | Normal | Flying | 579 | 83 | 80 | 80 | 135 | 80 | 121 | 1 | FALSO | 1 |
| Rattata | Normal | | 253 | 30 | 56 | 35 | 25 | 35 | 72 | 1 | FALSO | 1 |

# Data Analysis: Statistical Analysis

Statistical analysis is performed separately for each attribute in the dataset.

Depending on the type of the data, a different technique is applied.

In this case, the data in example is supposed to have a non-normal distribution.



# Data Analysis: Statistical Analysis

| | Cluster 1 (n = 67) | Cluster 2 (n = 38) | Cluster 3 (n = 70) | P value |
|---|---|---|---|---|
| **Children** | yes, two (42%) | no (76%) * | yes, two (50%) # | < 0.001 |
| **Marital status** | married (58%) | engaged (39%) * | married (50%) # | < 0.001 |
| **Lives with** | spouse + children (31%) | alone (47%) * | spouse (56%) # | < 0.001 |
| **Protective measures at home** | PPE (52%) | no need (55%) * | PPE (27%) # | < 0.001 |
| **COVID-19 fear for family** | 83.5 (70; 95) | 0 (0; 0) * | 55 (46; 71.5)* # | < 0.001 |
| **COVID-19 fear for self** | 80 (70.38; 90.5) | 64 (50; 76) * | 52.5 (40; 70.5)* | < 0.001 |
| **COVID-19 cases in ward** | yes (76%) | yes (76%) | yes (60%) * # | 0.015 |
| **Workload impact** | 77 (70; 82) | 69 (58; 77) * | 51 (43; 58) * # | < 0.001 |
| **Stress** | 80 (65; 95) | 70 (60; 85) | 60 (52.5; 75) * # | < 0.001 |
| **MBI** | 16 (12; 20) | 12 (8; 18) * | 8 (5; 13) * # | < 0.001 |
| **IES** | 38 (27; 50) | 25 (15;44) * | 20 (12; 31) * | < 0.001 |
| **PHQ-4** | 5 (4; 9) | 4 (2; 6) * | 3 (2; 5) * | < 0.001 |

# Personas Creation Workflow

- Goal Definition
- Survey creation
- Survey dissemination

**Data Collection**

- Data preparation
- Clustering
- Statistical Analysis

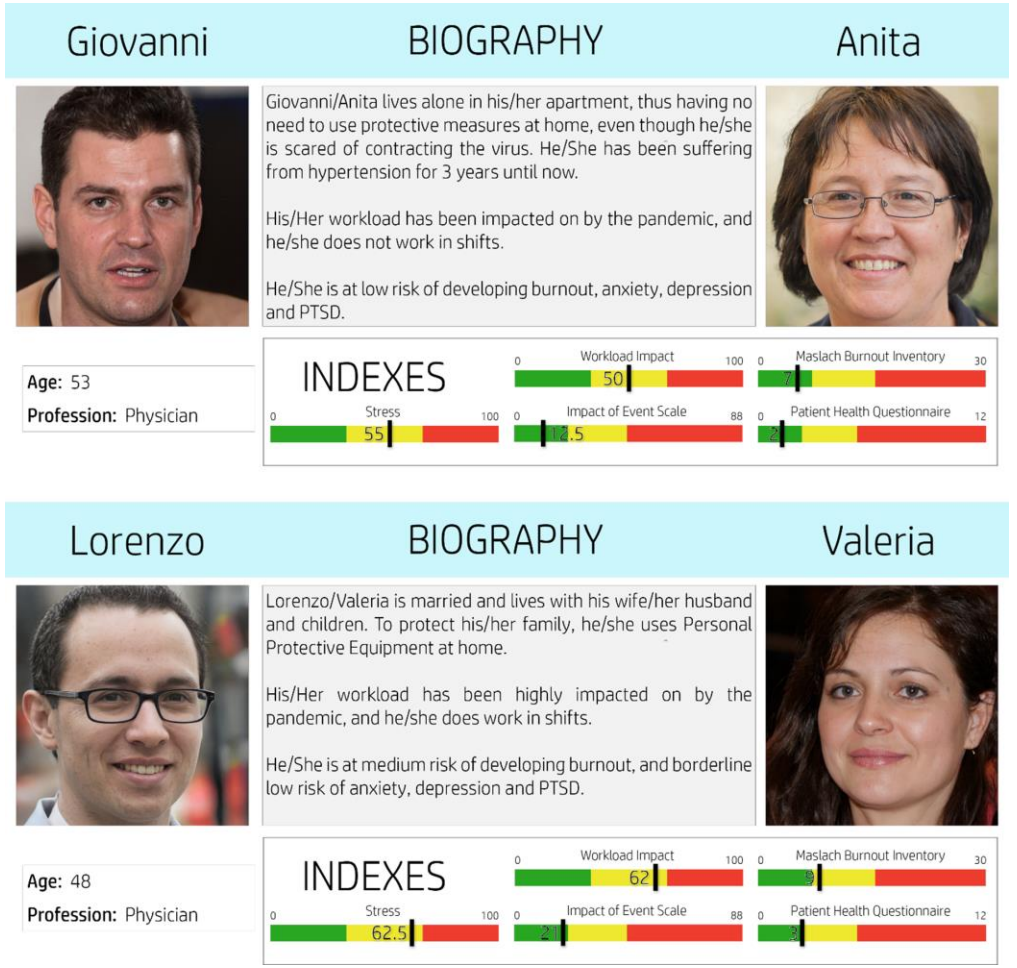
**Data Analysis**

- Persona Cards development
- Qualitative data collection

**Personification**

# Personification: Persona cards development

| | Cluster 1 (n = 67) | Cluster 2 (n = 38) | Cluster 3 (n = 70) | P value |
|---|---|---|---|---|
| Children | yes, two (42%) | no (76%) * | yes, two (50%) # | < 0.001 |
| Marital status | married (58%) | engaged (39%) * | married (50%) # | < 0.001 |
| Lives with | spouse + children (31%) | alone (47%) * | spouse (56%) # | < 0.001 |
| Protective measures at home | PPE (52%) | no need (55%) * | PPE (27%) # | < 0.001 |
| COVID-19 fear for family | 83.5 (70; 95) | 0 (0; 0) * | 55 (46; 71.5)* # | < 0.001 |
| COVID-19 fear for self | 80 (70.38; 90.5) | 64 (50; 76) * | 52.5 (40; 70.5)* | < 0.001 |
| COVID-19 cases in ward | yes (76%) | yes (76%) | yes (60%) * # | 0.015 |
| Workload impact | 77 (70; 82) | 69 (58; 77) * | 51 (43; 58) * # | < 0.001 |
| Stress | 80 (65; 95) | 70 (60; 85) | 60 (52.5; 75) * # | < 0.001 |
| MBI | 16 (12; 20) | 12 (8; 18) * | 8 (5; 13) * # | < 0.001 |
| IES | 38 (27; 50) | 25 (15;44) * | 20 (12; 31) * | < 0.001 |
| PHQ-4 | 5 (4; 9) | 4 (2; 6) * | 3 (2; 5) * | < 0.001 |

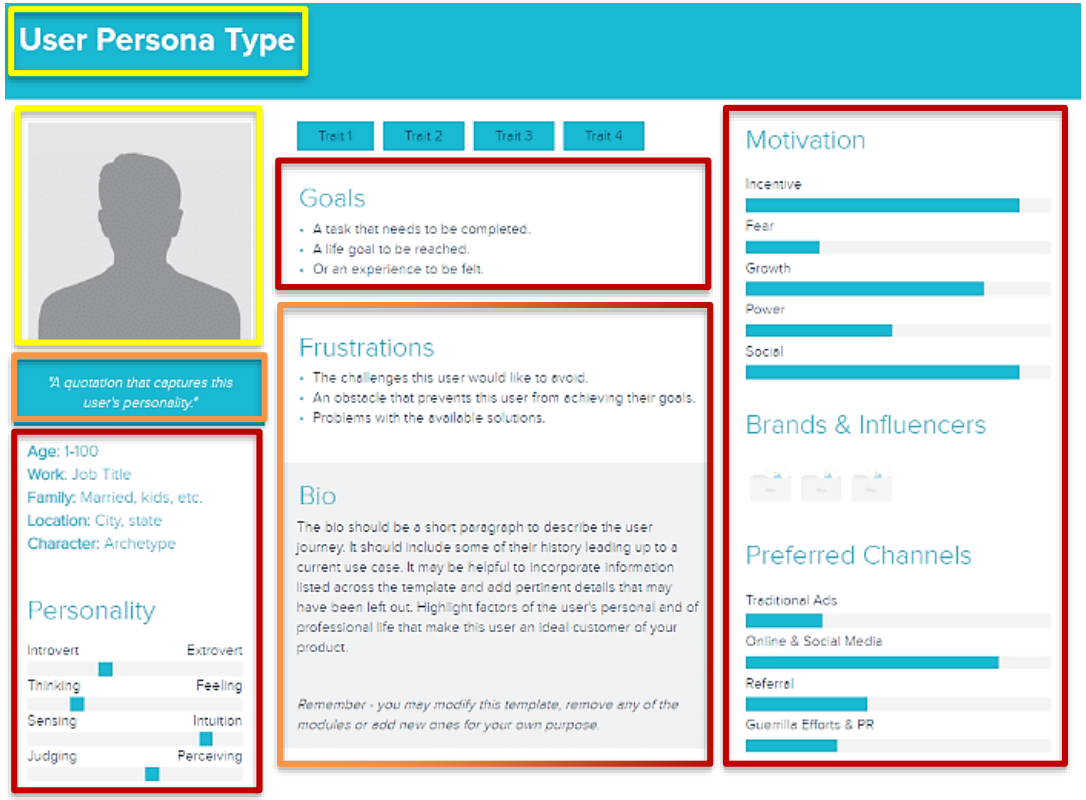


# Personification: Persona cards development

With the obtained information from the attributes that differentiate between clusters it is possible to create Persona Cards.

Information in a Persona card can be of three types.

- **Inferred information:** usually name and photo. Even though there are no attributes directly related to these attributes they can be inferred from sex and age.
- **Quantitative information**: obtained by the data analysis. Goals are part of the quantitative information and must be expressed clearly in the Persona card. Usage of color-coding bars and indicators is highly encouraged.
- **Qualitative information**: obtained by the data analysis and other sources.



# Project Work

## You can start working!