



UNIVERSITÀ  
DEGLI STUDI  
DI MILANO

Scientific Visualization  
A.A. – 2025/2026



# DATA IMPUTATION

based on PTB-XL dataset (PhysioNet)

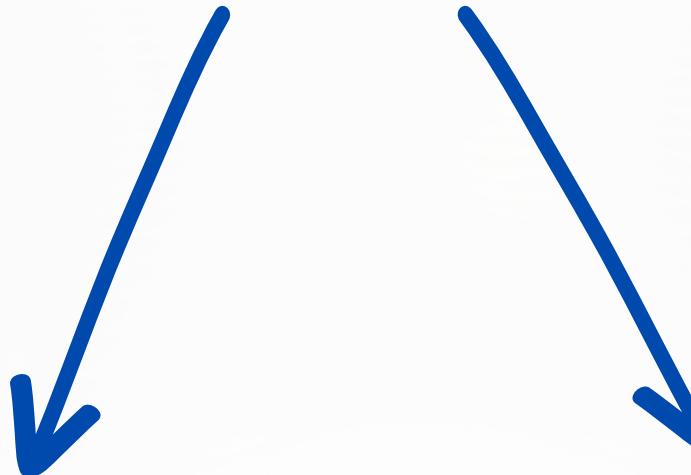
Bertoletti Matteo – Invernizzi Carlo – Maggioni Sofia – Zombini Cecilia

# OUR GOAL

Study how the **Body Mass Index (BMI)** of patients in the PTBXL dataset is related to the appearance of particular **cardiovascular pathologies**.



identify **patterns** on the patient's physical condition



**probability** of developing certain diseases



**Initial data exploration  
and cleaning**

**Data Imputation**

**BMI calculation**

**Analysis of BMI  
linked to pathologies**

**1**

**2**

**3**

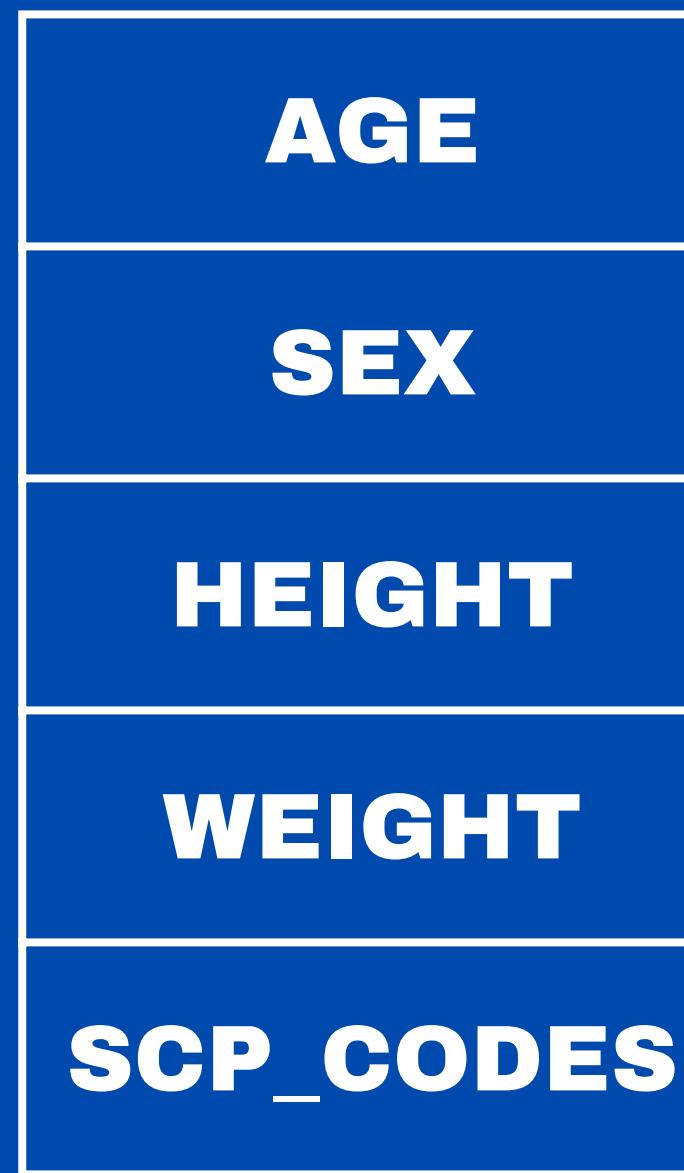
**4**

# OUR **WORKFLOW**

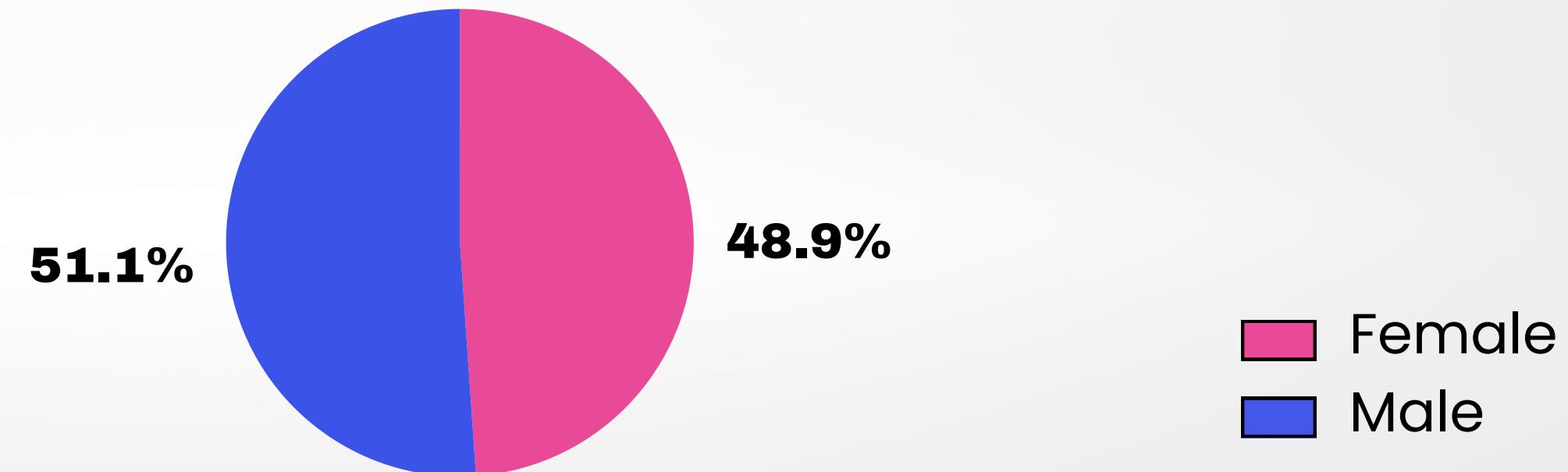
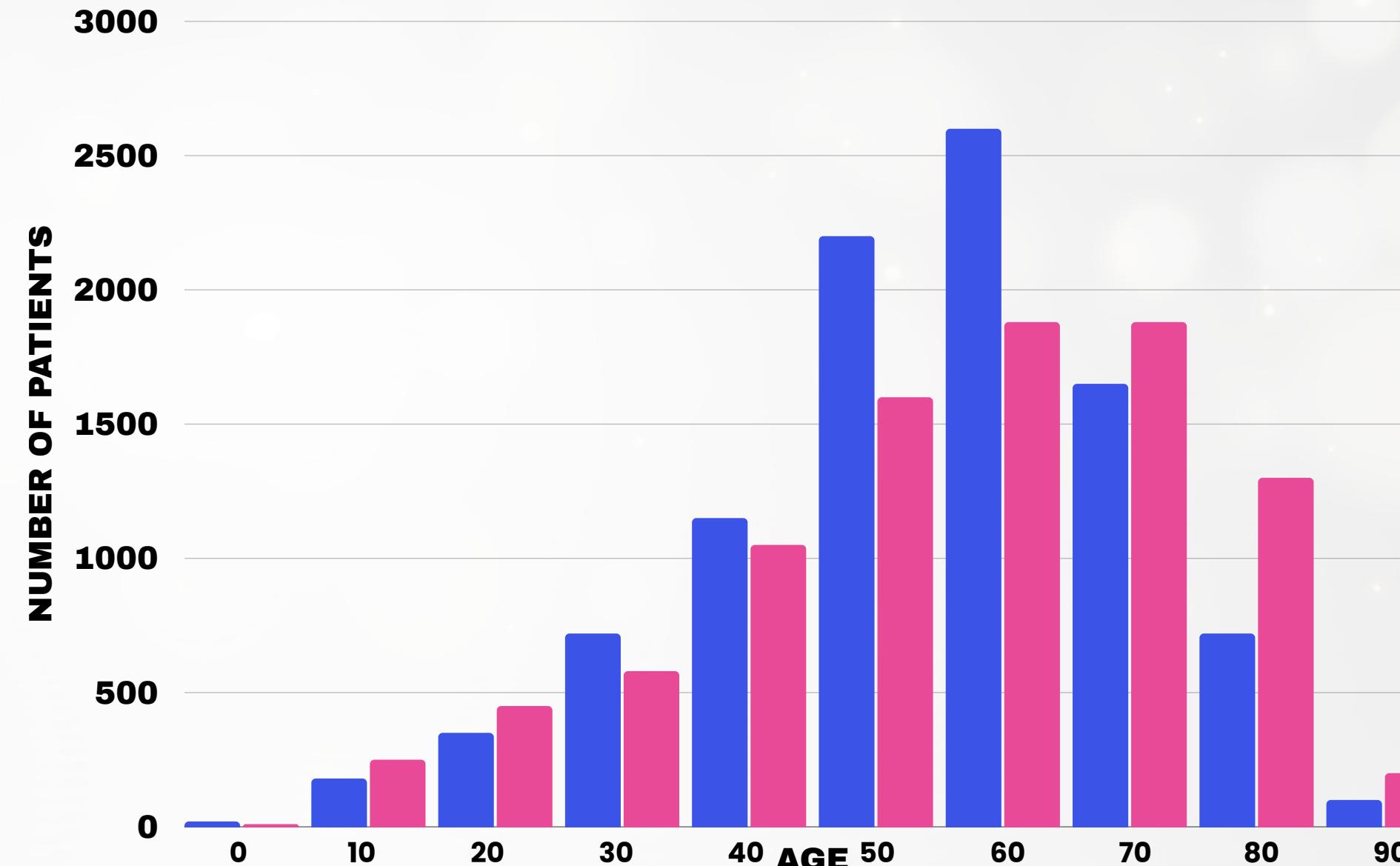
# 1

## DATASET

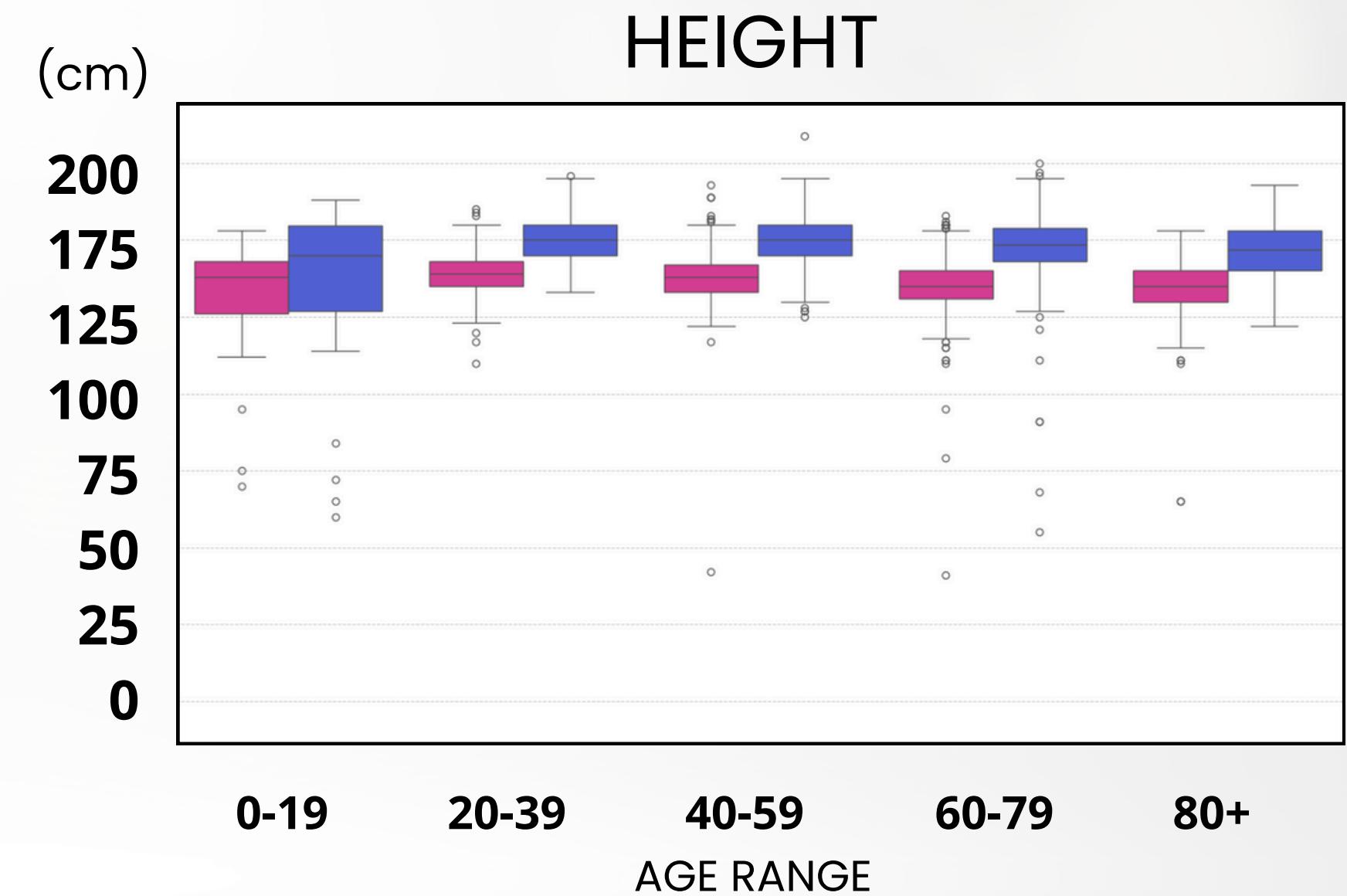
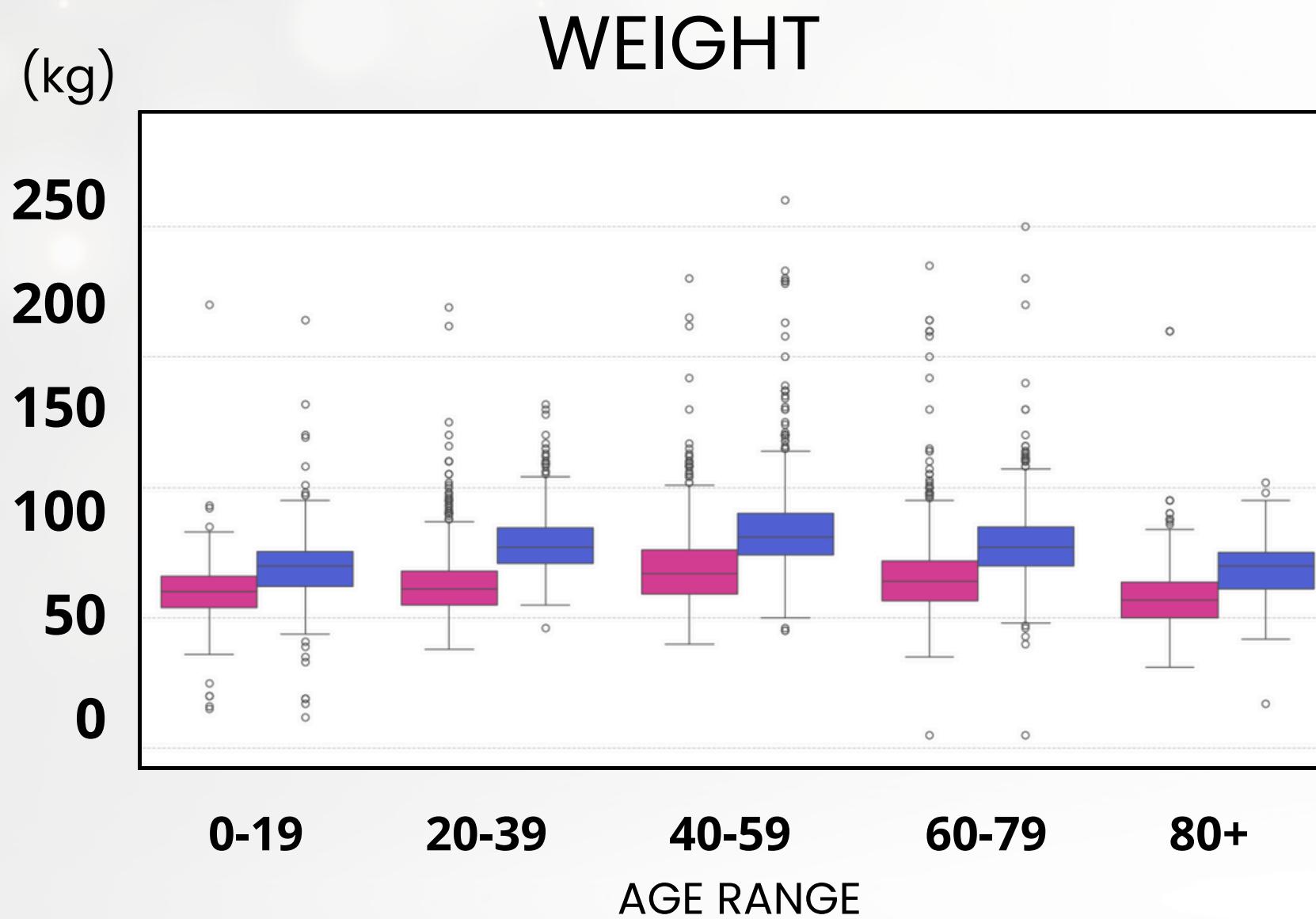
18.870 patients



### AGE DISTRIBUTION PER SEX

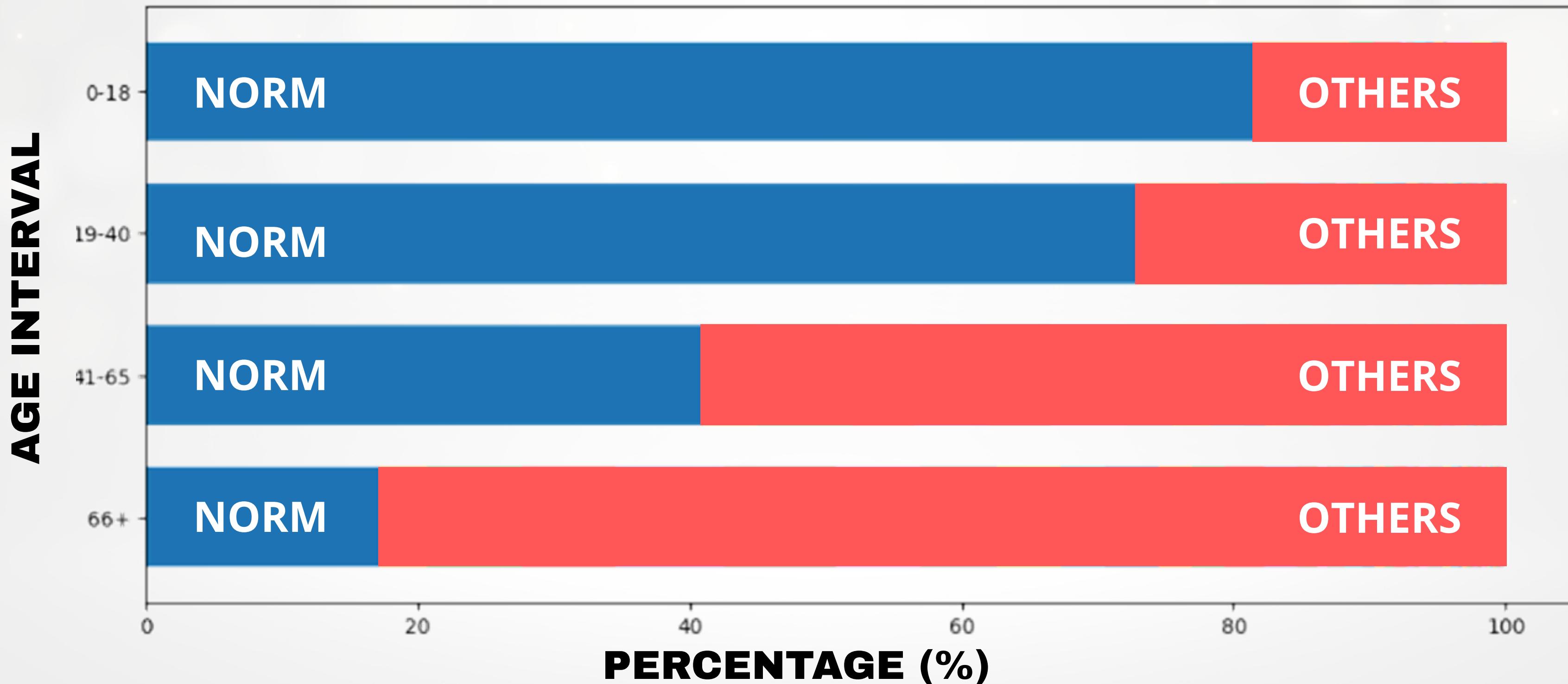


# BOXPLOT

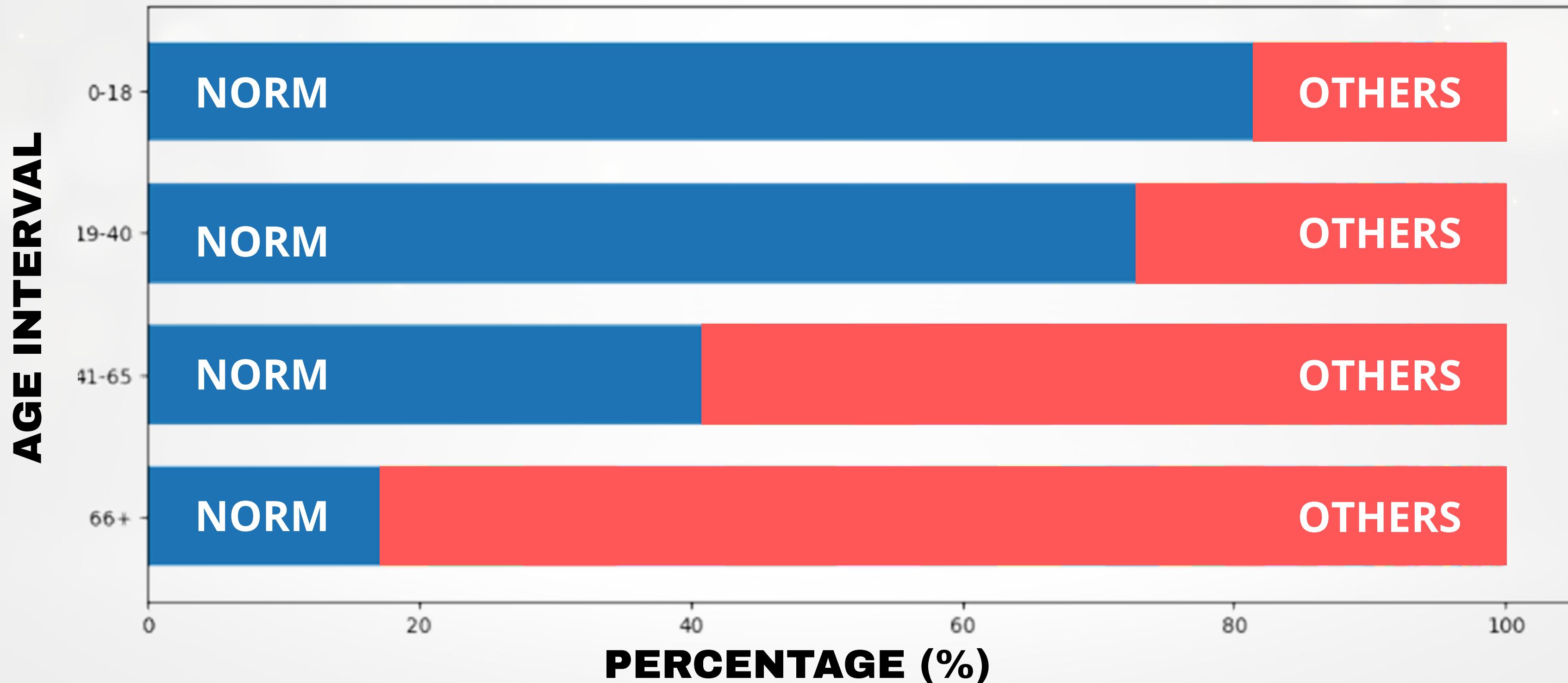


■ Female  
■ Male

TYPES OF  
**PATHOLOGIES (scp\_codes)**



TYPES OF  
**PATHOLOGIES (scp\_codes)**

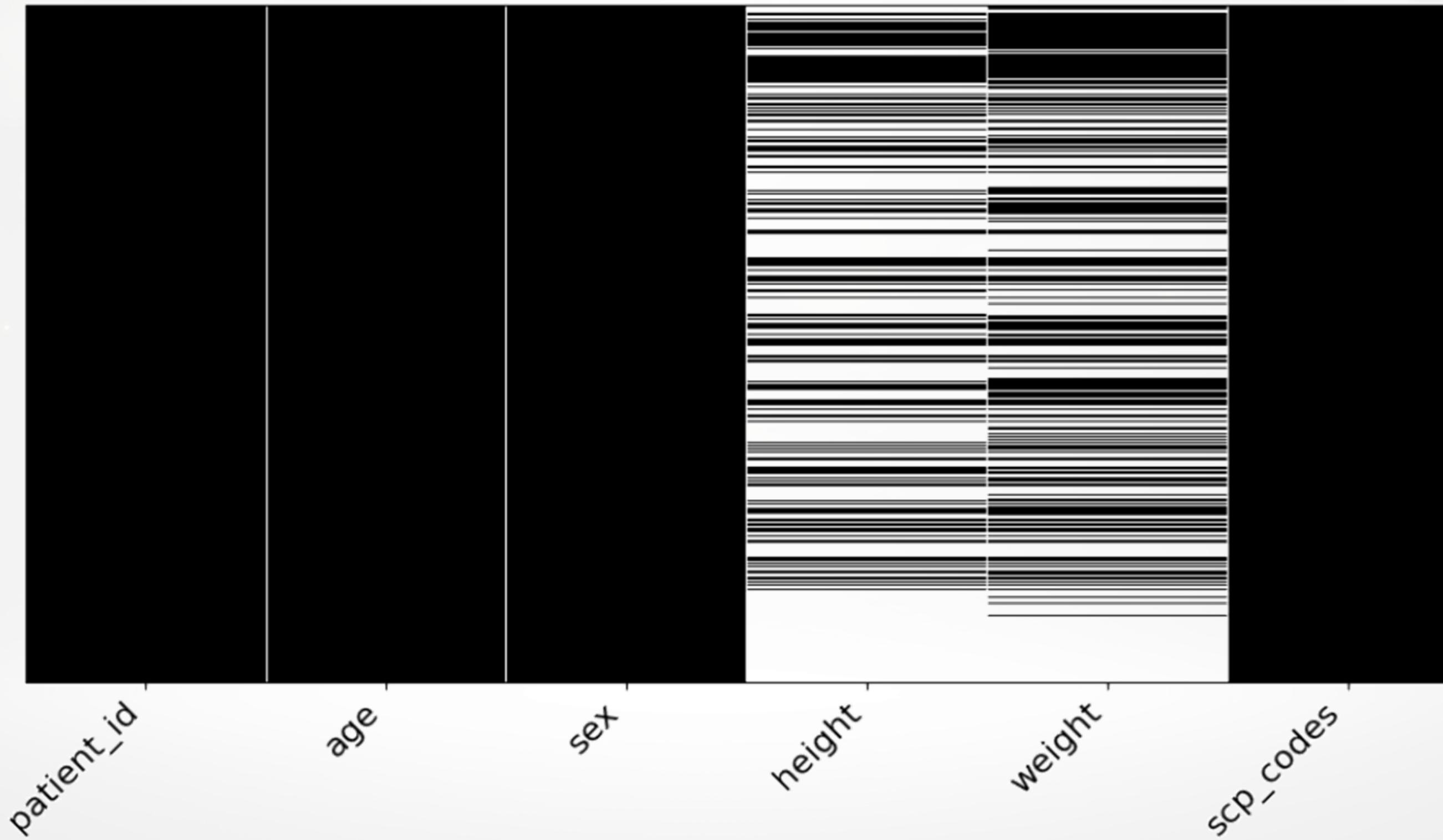


# STUDY OF CHI<sup>2</sup>

The Chi-Squared Test is a statistical hypothesis test used to examine the differences between **observed data (O)** and **expected data (E)** under a specific hypothesis.

$$\chi^2 = \sum \frac{(O - E)^2}{E} \longrightarrow 2619,3798$$

# MISSING DATA PATTERN VISUALIZATION



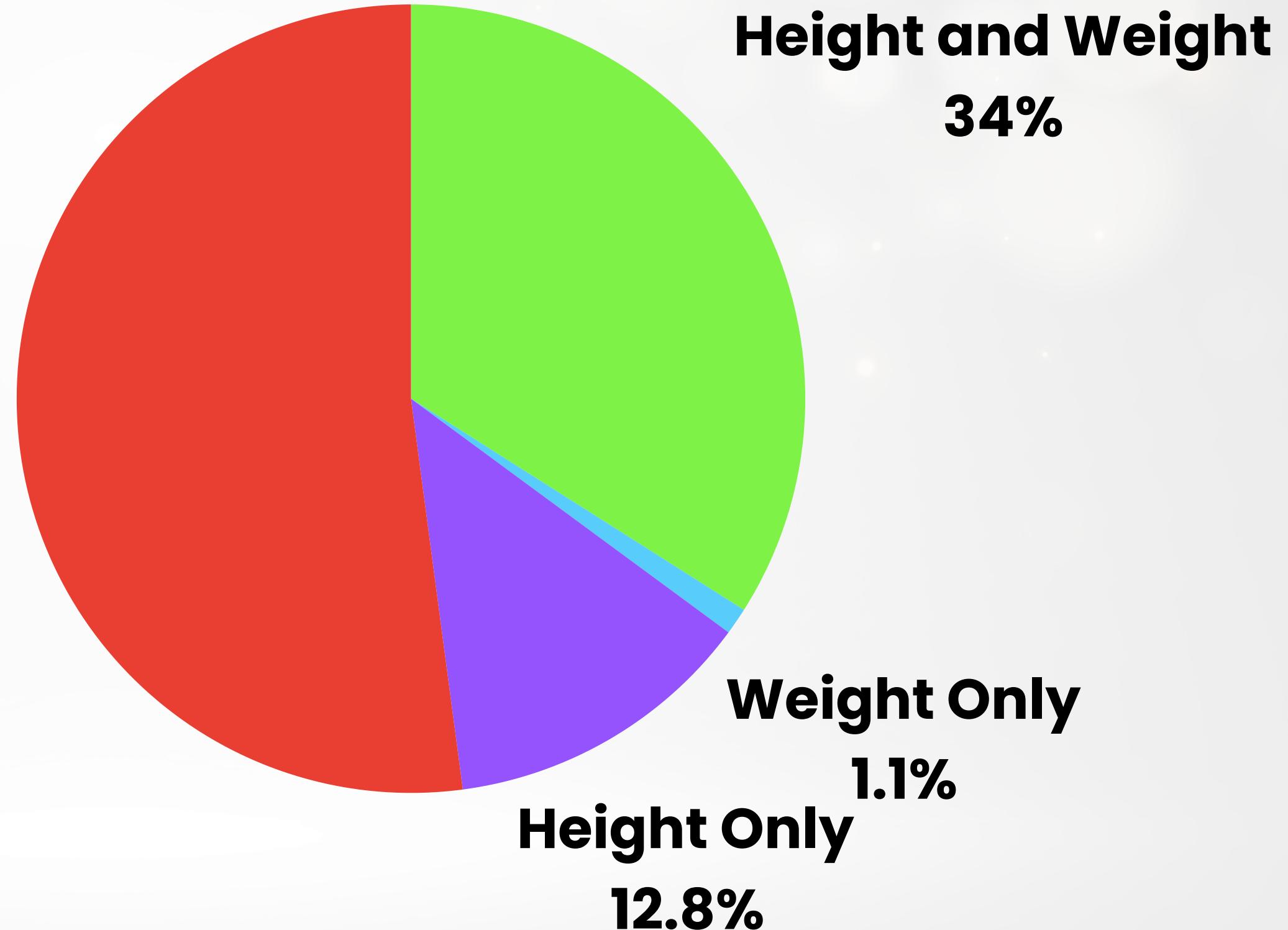
# 2

## DATA IMPUTATION

Missing data in  
**WEIGHT** and **HEIGHT**

**Data imputation** is the process of replacing missing data in a data set with estimated or predicted values.

**None**  
52.1%



# TYPES OF MISSING DATA

**MCAR**

**Missing Completely  
At Random**

The probability that a value is missing is unrelated to any observed or unobserved data.

**MAR**

**Missing At Random**

The probability of missingness depends only on observed data, not on the missing values themselves.

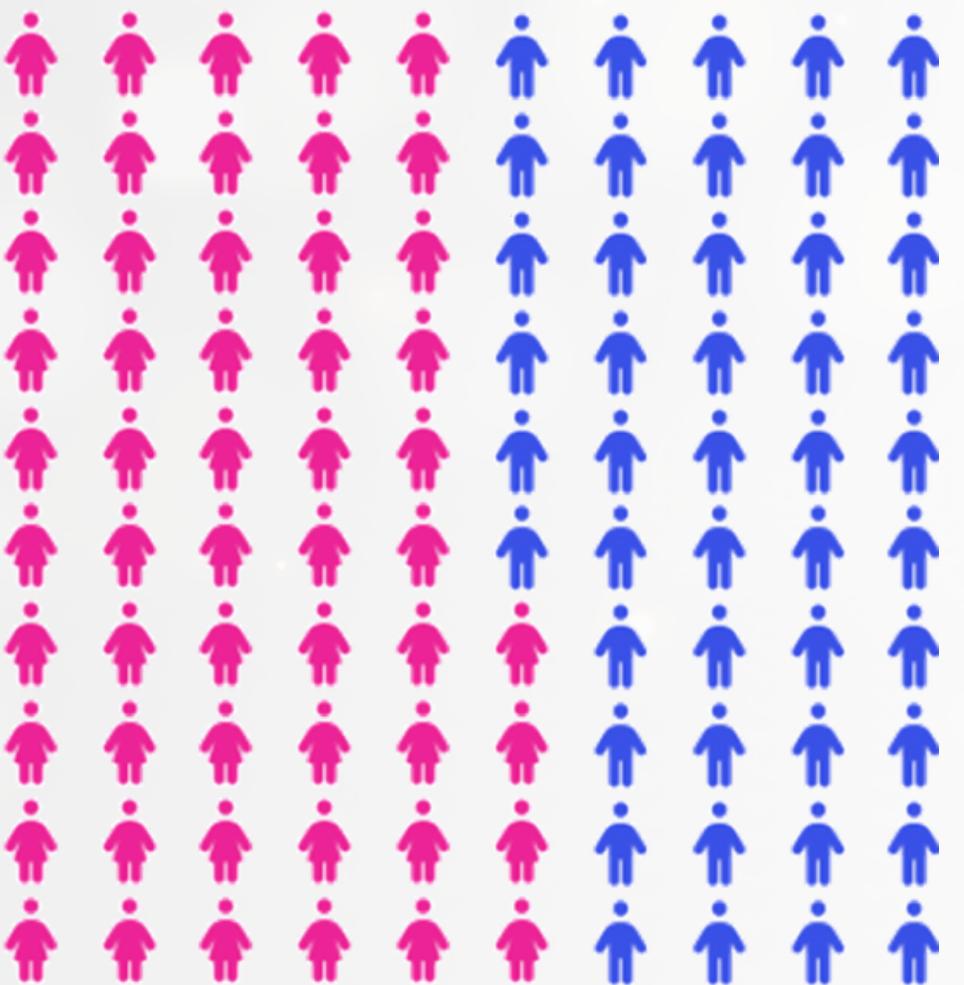
**MNAR**

**Missing Not At Random**

The probability that data are missing depends on the missing values themselves, even after controlling for other observed data.

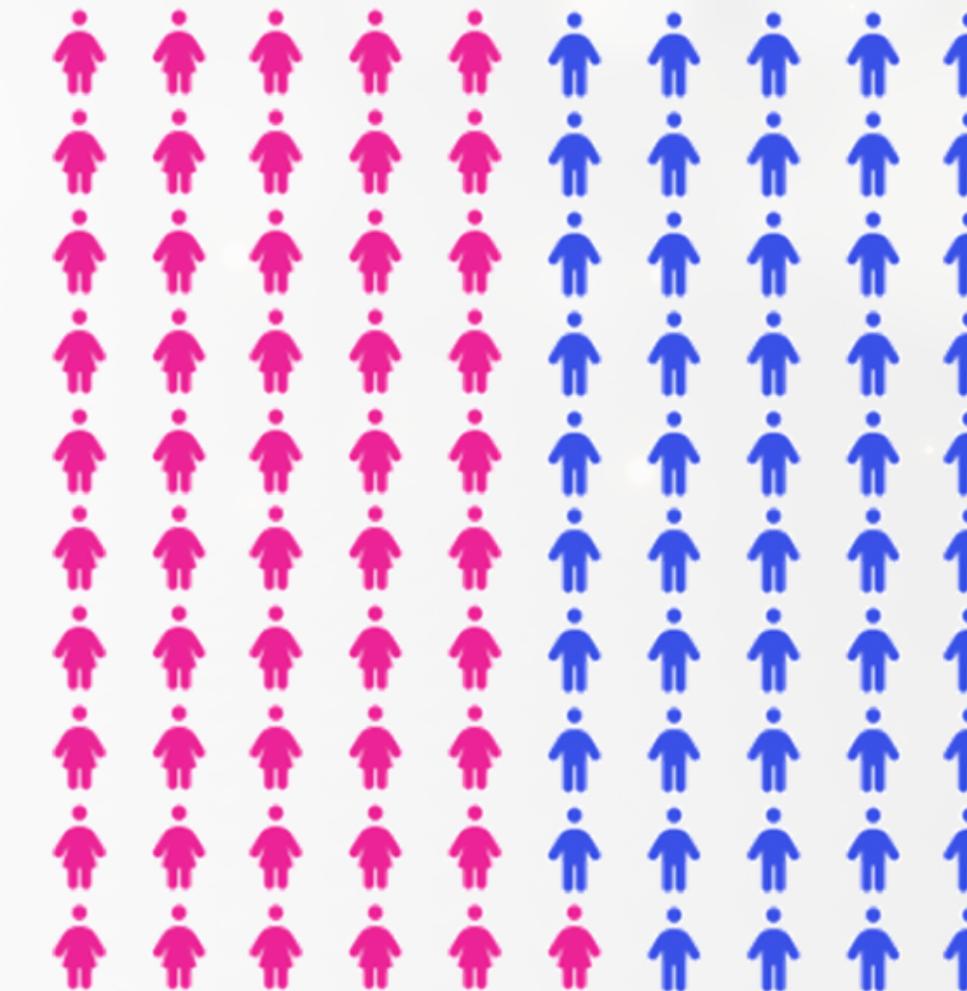
# OUR DATA

Missing HEIGHT



Female - 54%  
Male - 46%

Missing WEIGHT

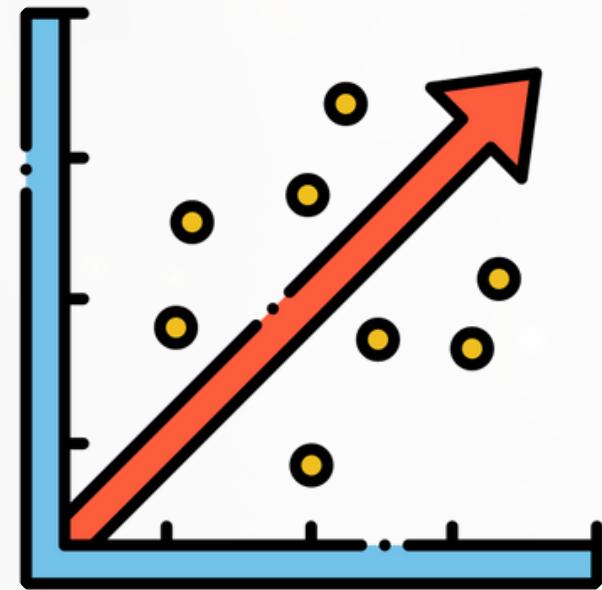


Female - 51%  
Male - 49%

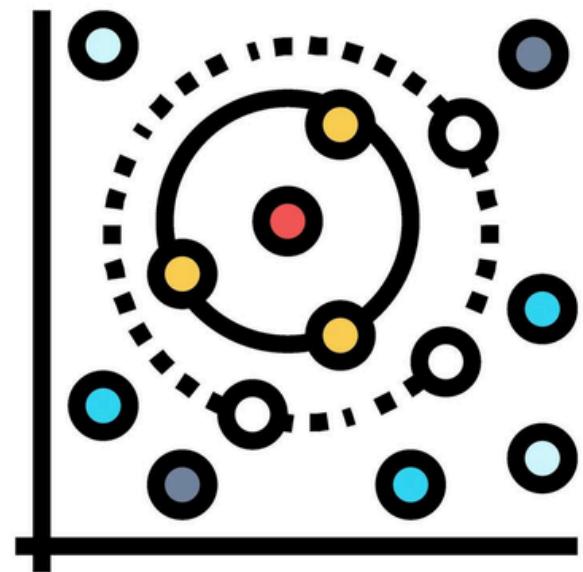
MCAR

Missing Completely At Random

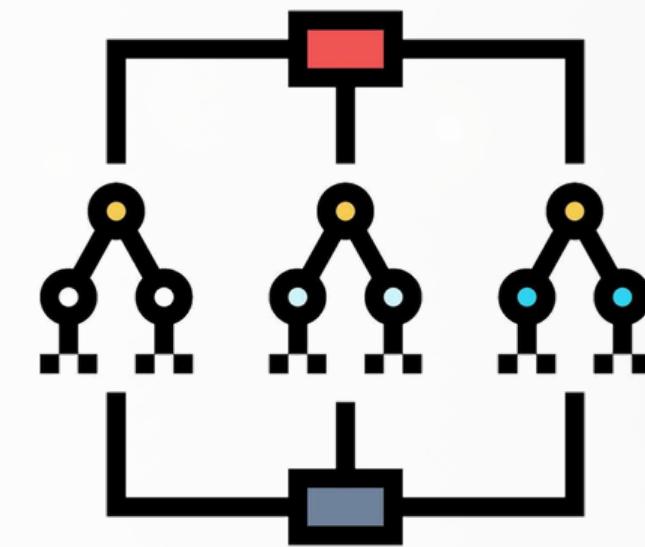
# TYPES OF DATA IMPUTATION



Linear Regression



K-Nearest  
Neighbors  
(KNN)



Miss Forest

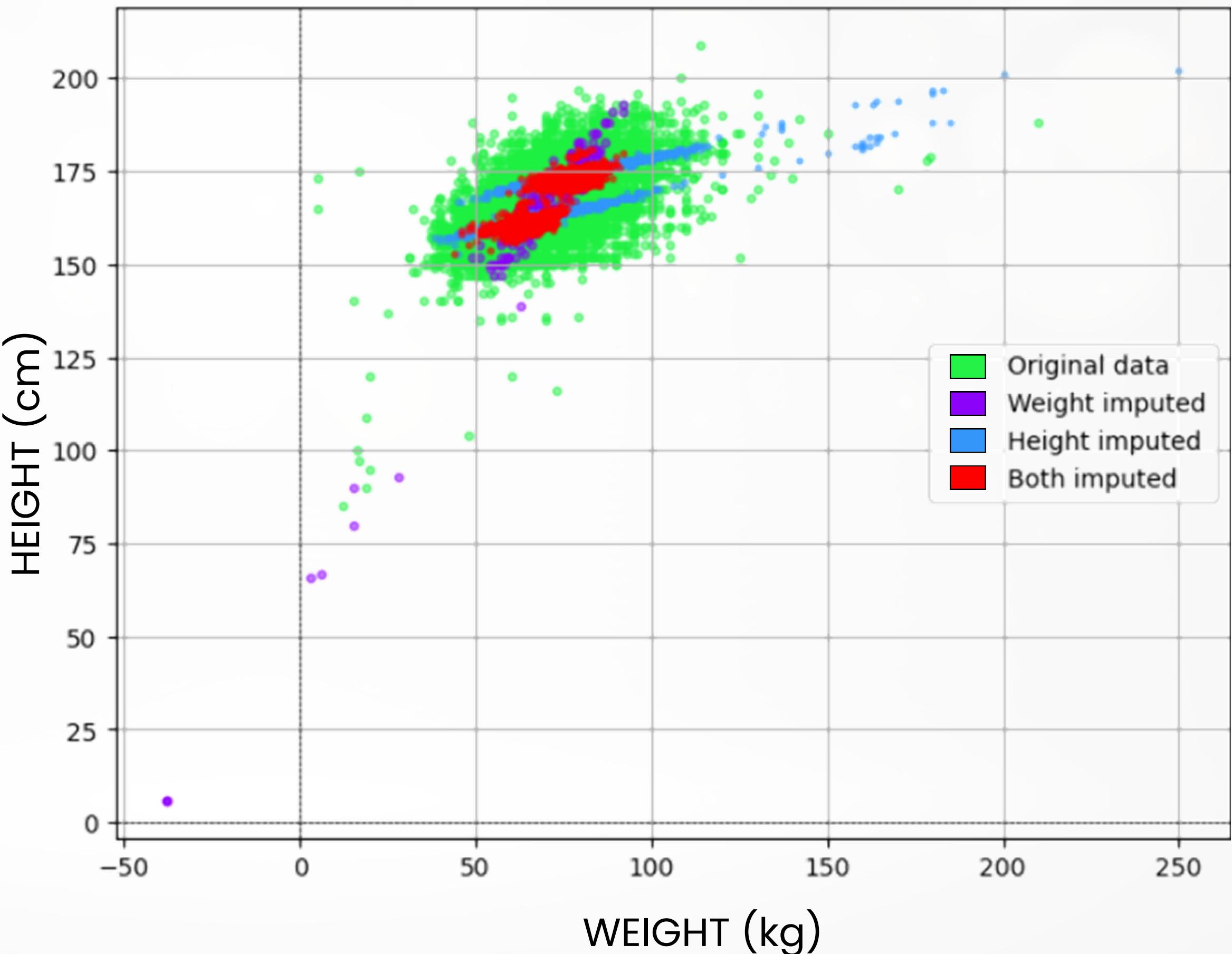
# LINEAR REGRESSION

Linear regression estimates and replaces missing values by modeling the relationships between variables.

Specifically, it involves fitting a **linear equation** to **observed data**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

## LINEAR REGRESSION



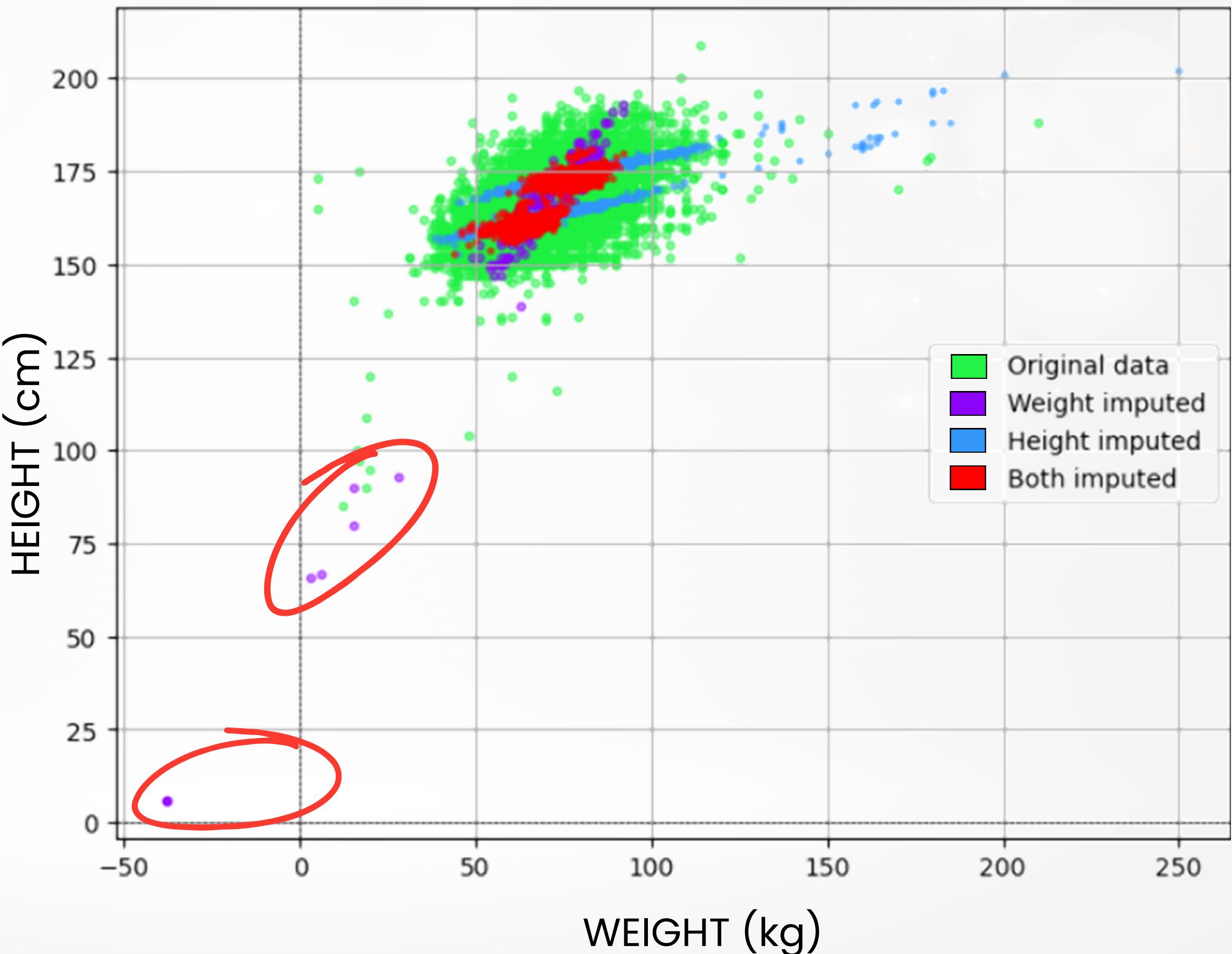
# LINEAR REGRESSION

Linear regression estimates and replaces missing values by modeling the relationships between variables.

Specifically, it involves fitting a **linear equation** to **observed data**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

## LINEAR REGRESSION



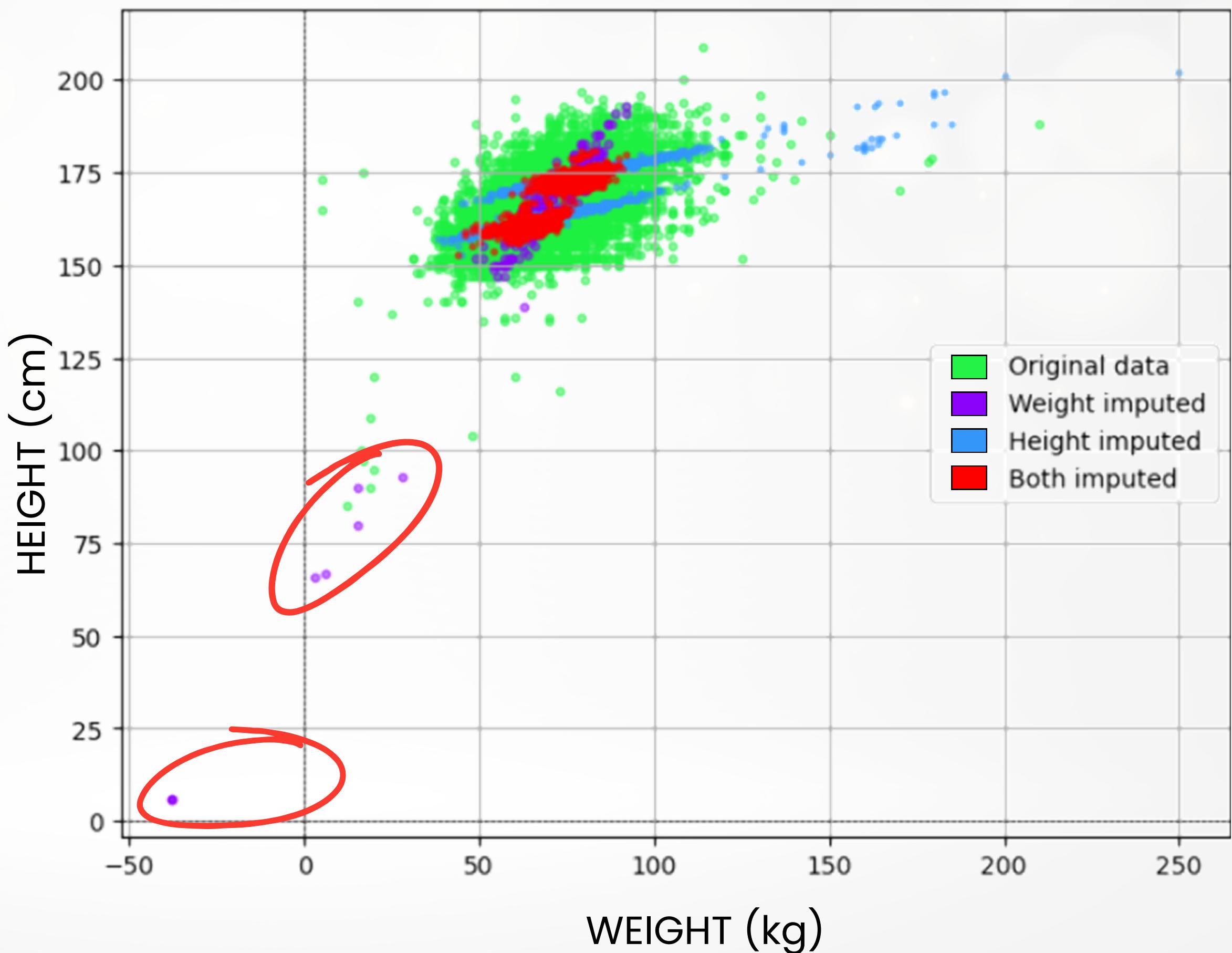
## LINEAR REGRESSION

# LINEAR REGRESSION

Linear regression estimates and replaces missing values by modeling the relationships between variables.

Specifically, it involves fitting a **linear equation** to **observed data**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

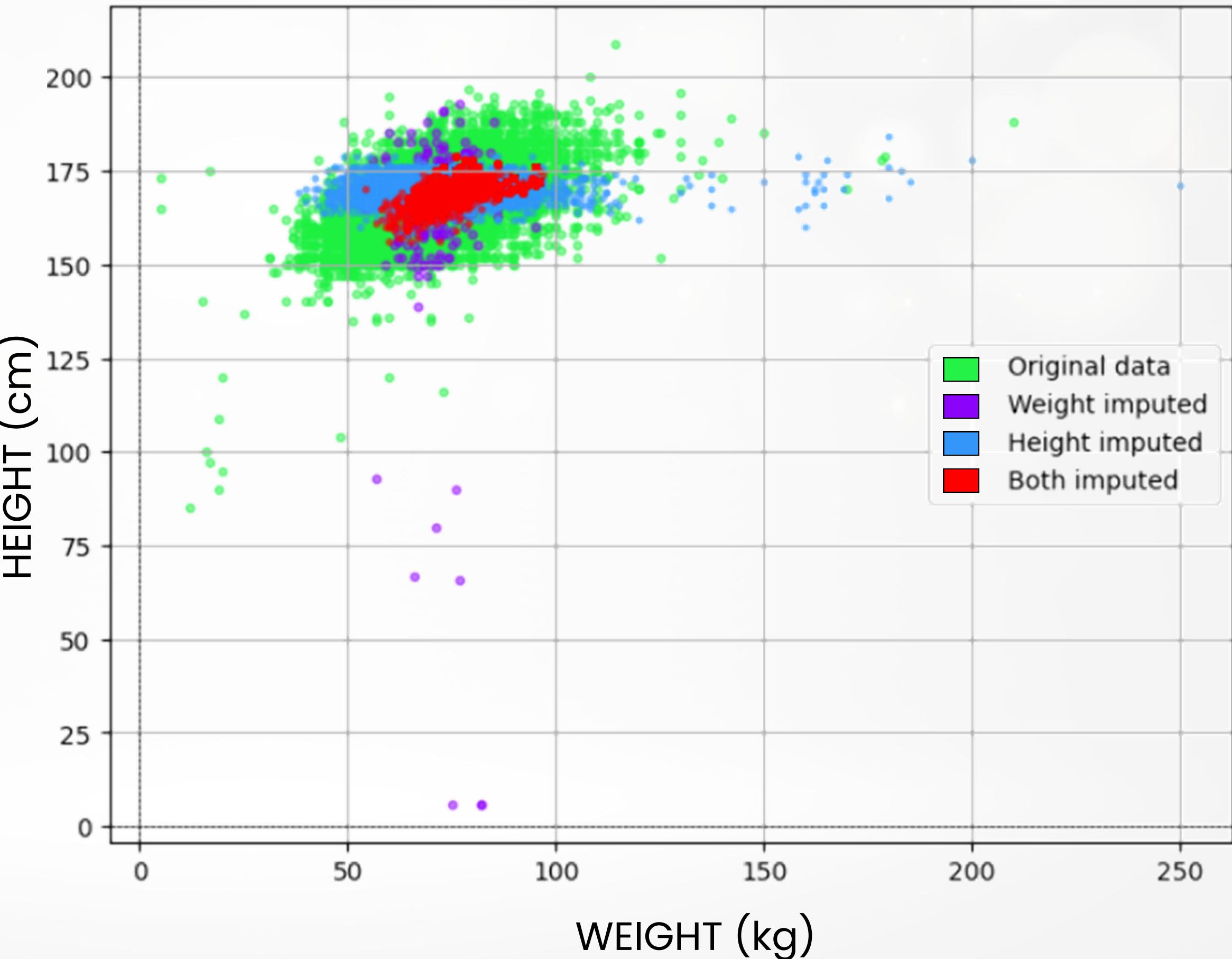


# KNN

K-Nearest Neighbors fills in missing values by using the values of the **k most similar** instances in the dataset, based on the values computed **using the available** (non-missing) **features**.

$$d(p, q) = \sqrt{\sum_i (q_i - p_i)^2}$$

## KNN

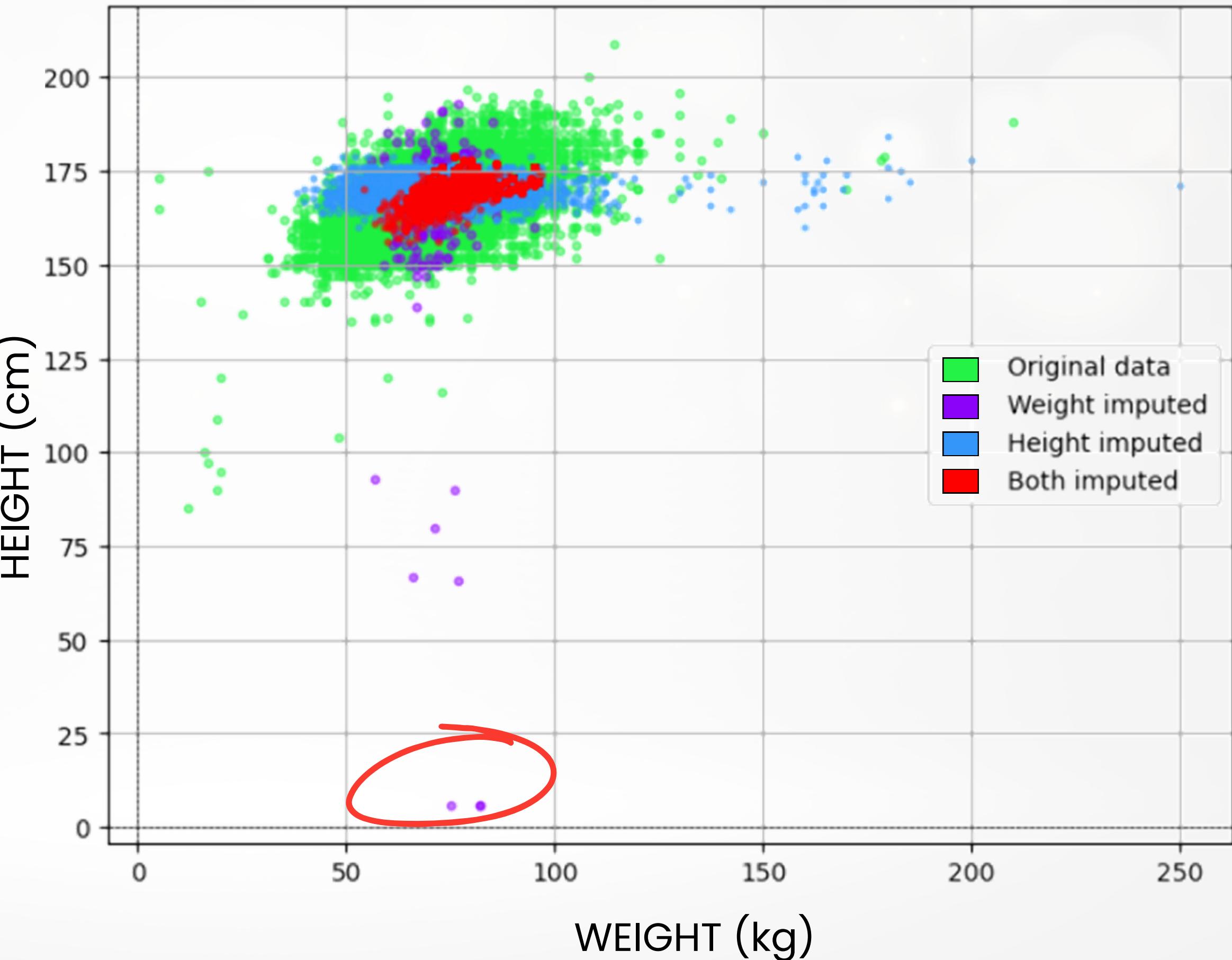


# KNN

K-Nearest Neighbors fills in missing values by using the values of the **k most similar** instances in the dataset, based on the values computed **using the available** (non-missing) **features**.

$$d(p, q) = \sqrt{\sum_i (q_i - p_i)^2}$$

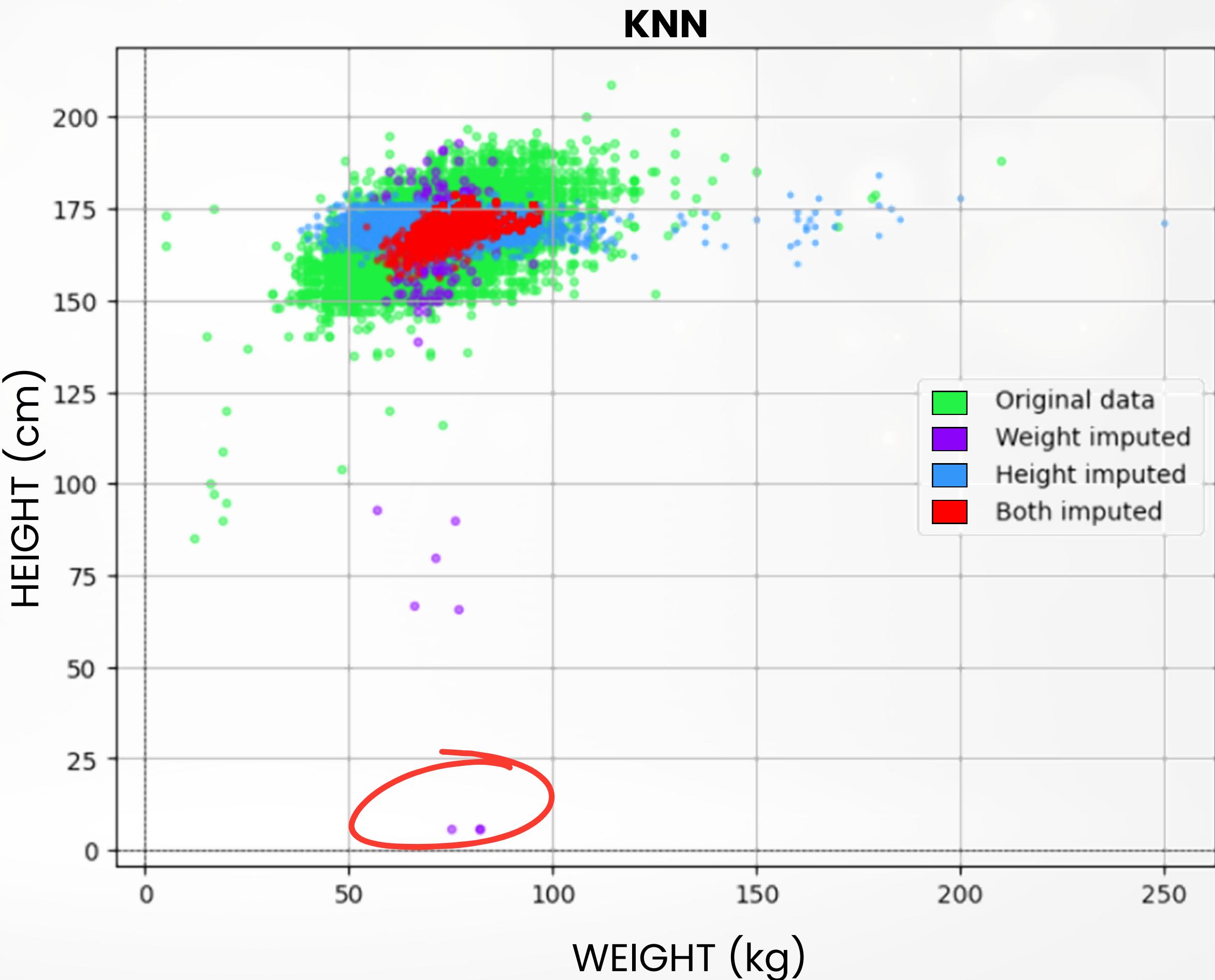
## KNN



~~KNN~~

K-Nearest Neighbors fills in missing values by using the values of the **k most similar** instances in the dataset, based on the values computed **using the available** (non-missing) **features**.

$$d(p, q) = \sqrt{\sum_i (q_i - p_i)^2}$$



CONCEPT OF  
**MISS FOREST**

	<b>age</b>	<b>weight</b>	<b>height</b>
<b>patient_1</b>	<b>45</b>	<b>75</b>	<b>nan</b>
<b>patient_2</b>	<b>30</b>	<b>nan</b>	<b>175</b>
<b>patient_3</b>	<b>65</b>	<b>66</b>	<b>nan</b>

# CONCEPT OF **MISS FOREST**

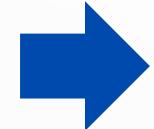
■ data missing

	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan

# CONCEPT OF **MISS FOREST**

■ data observed  
■ data missing

	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan

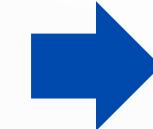


	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan

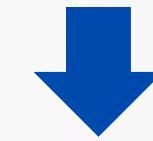
# CONCEPT OF MISS FOREST

- [green square] data observed
- [red square] data missing
- [light blue square] result

	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan



	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan

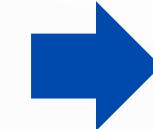


	age	weight	height
patient_1	45	75	nan
patient_2	30	70.5	175
patient_3	65	66	nan

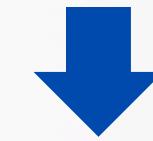
# CONCEPT OF MISS FOREST

■ data observed  
■ data missing

	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan



	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan

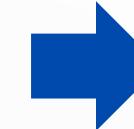


	age	weight	height
patient_1	45	75	nan
patient_2	30	70.5	175
patient_3	65	66	nan

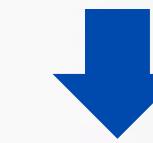
# CONCEPT OF MISS FOREST

- [green square] data observed
- [red square] data missing
- [light blue square] result

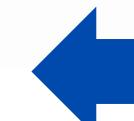
	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan



	age	weight	height
patient_1	45	75	nan
patient_2	30	nan	175
patient_3	65	66	nan



	age	weight	height
patient_1	45	75	170
patient_2	30	70.5	175
patient_3	65	66	169

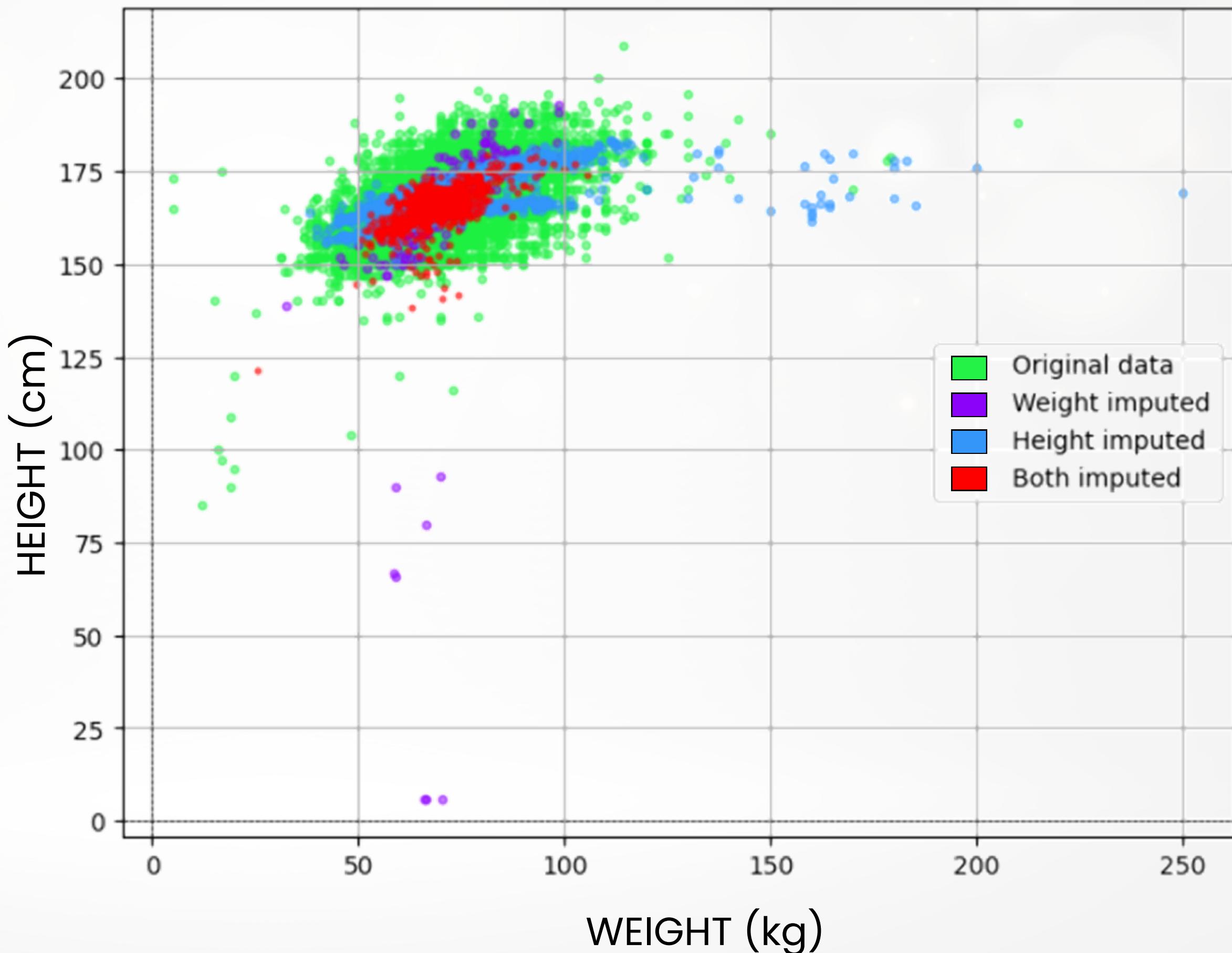


	age	weight	height
patient_1	45	75	nan
patient_2	30	70.5	175
patient_3	65	66	nan

# MISS FOREST

MissForest fills in missing data by iteratively **training Random Forest** models to predict missing values based on the observed relationships among all other variables (both continuous and categorical variables).

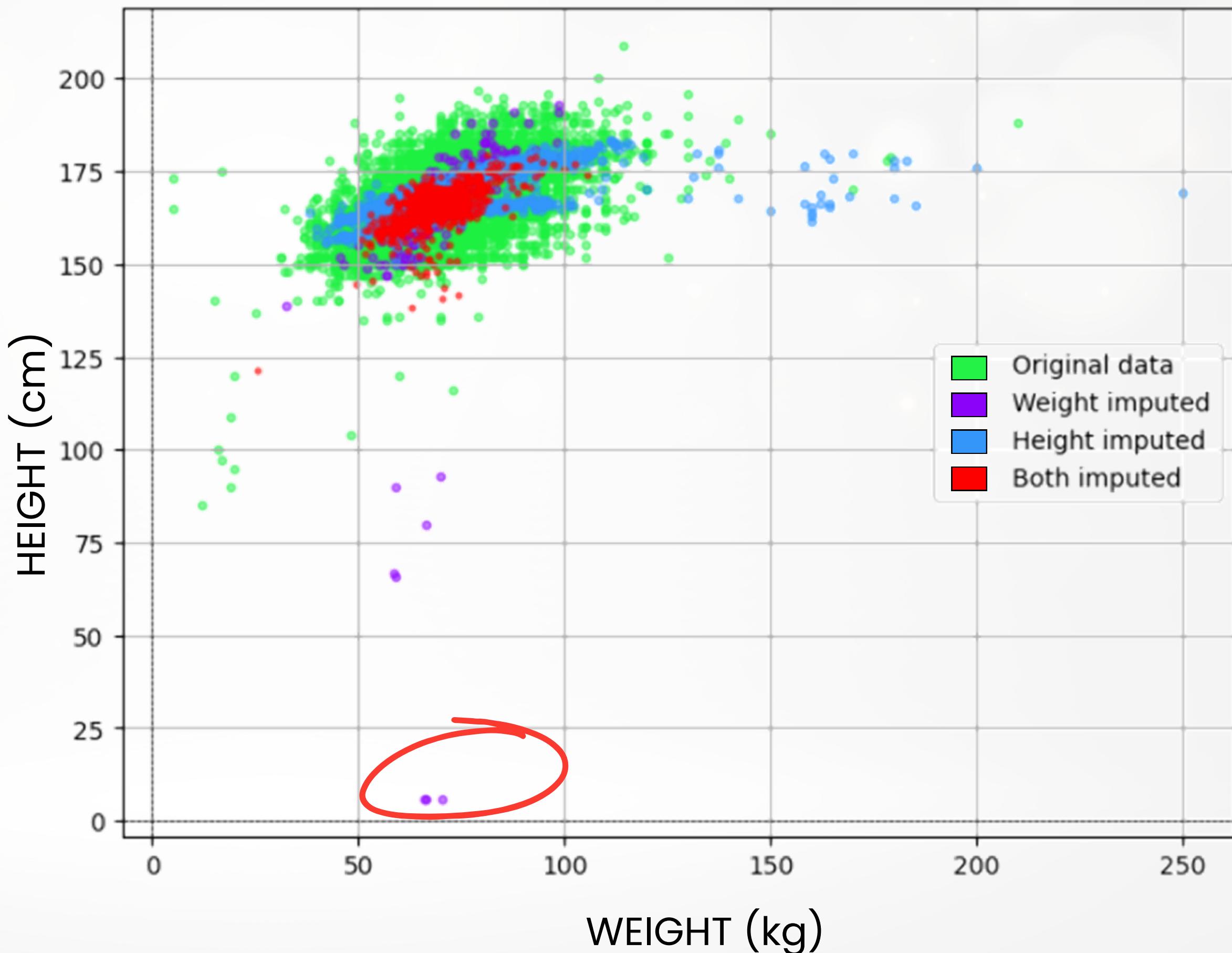
## MISS FOREST



# MISS FOREST

MissForest fills in missing data by iteratively **training Random Forest** models to predict missing values based on the observed relationships among all other variables (both continuous and categorical variables).

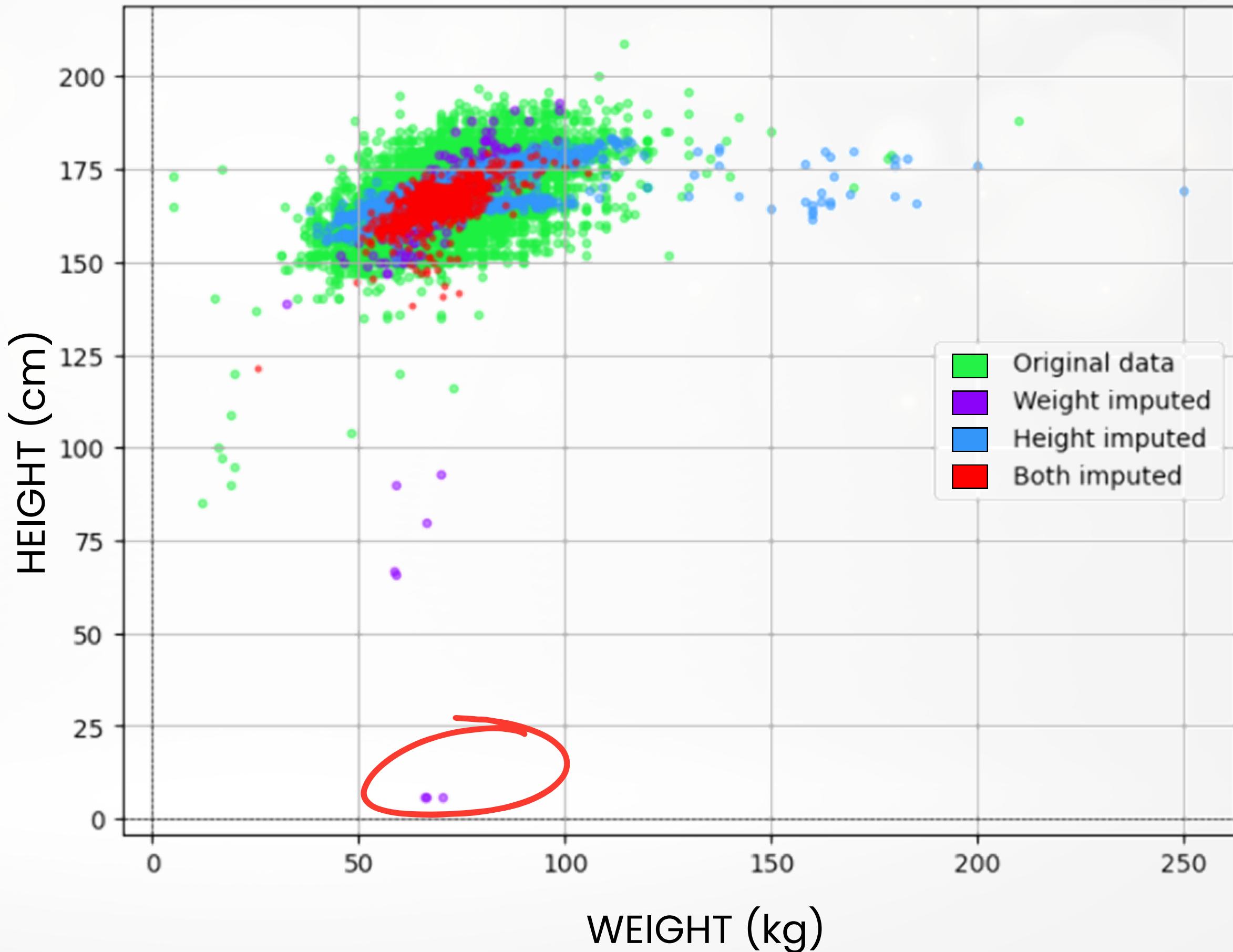
## MISS FOREST



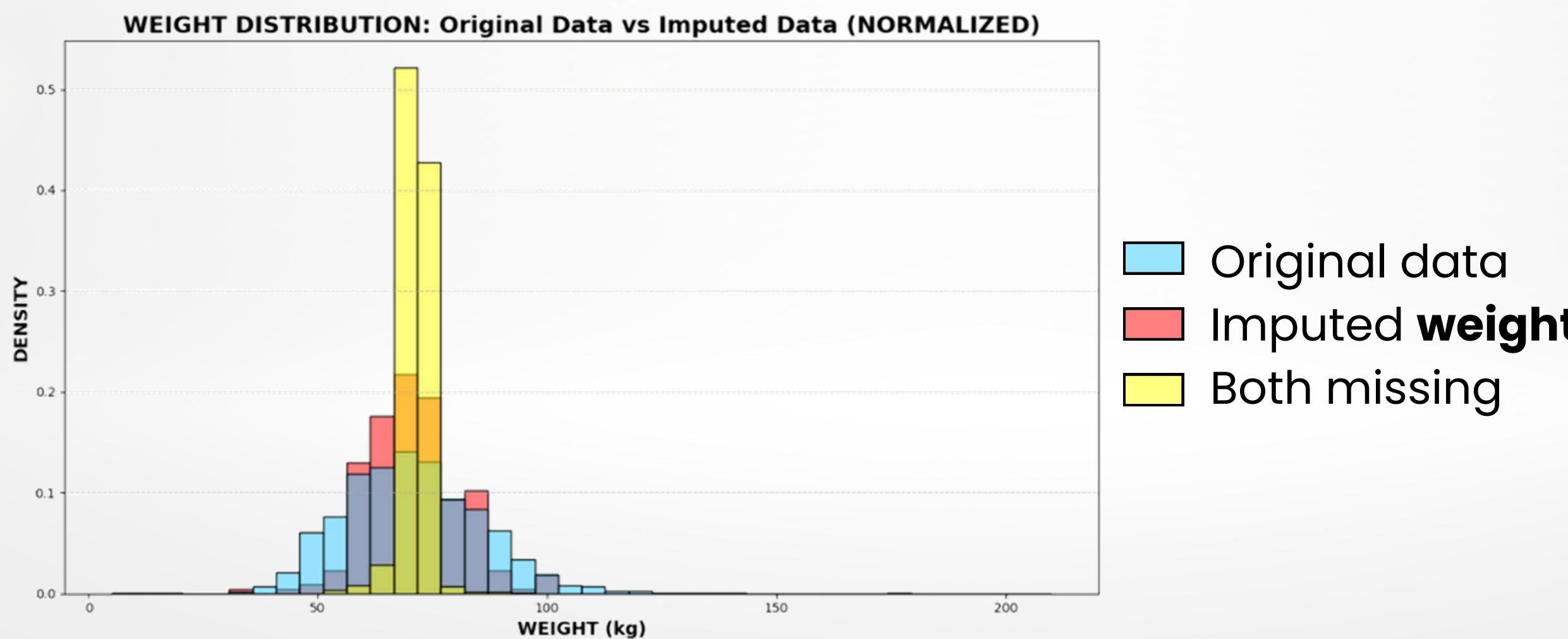
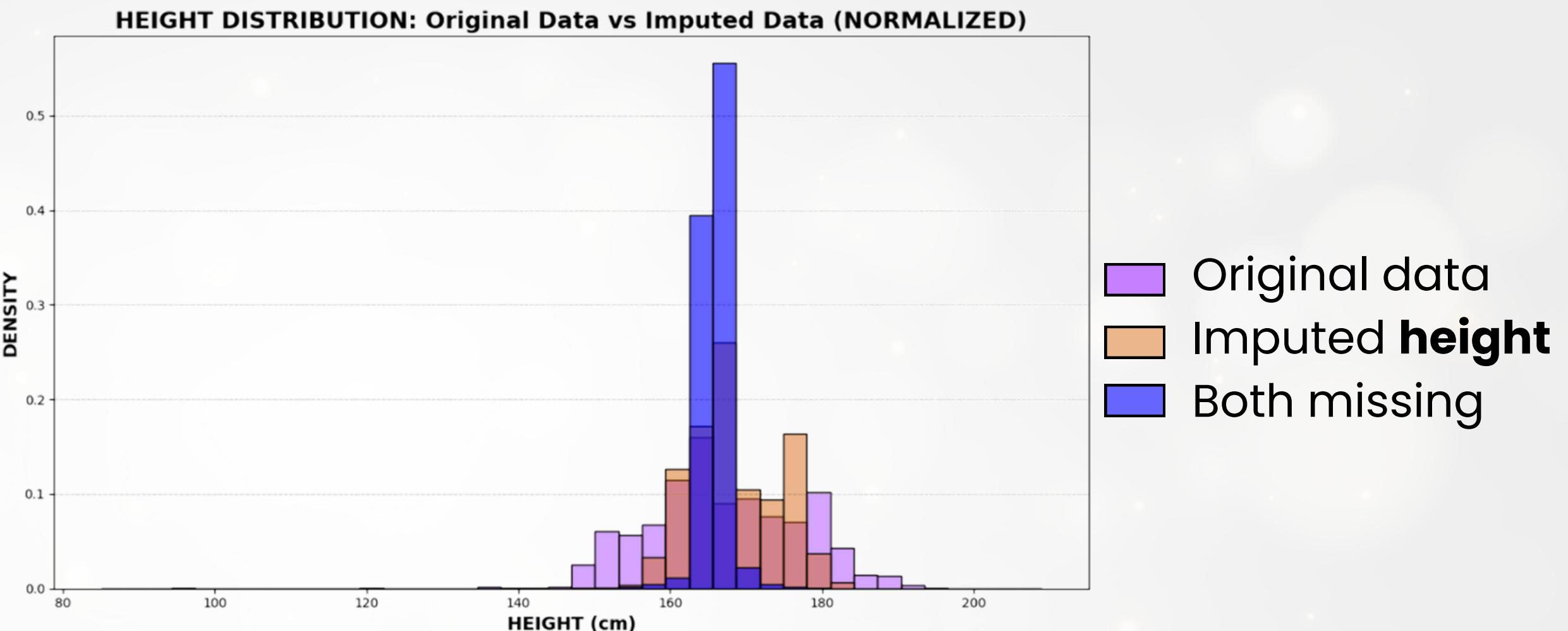
# MISS FOREST

MissForest fills in missing data by iteratively **training Random Forest** models to predict missing values based on the observed relationships among all other variables (both continuous and categorical variables).

## MISS FOREST



# ACTUAL IMPUTED DATA DISTRIBUTION

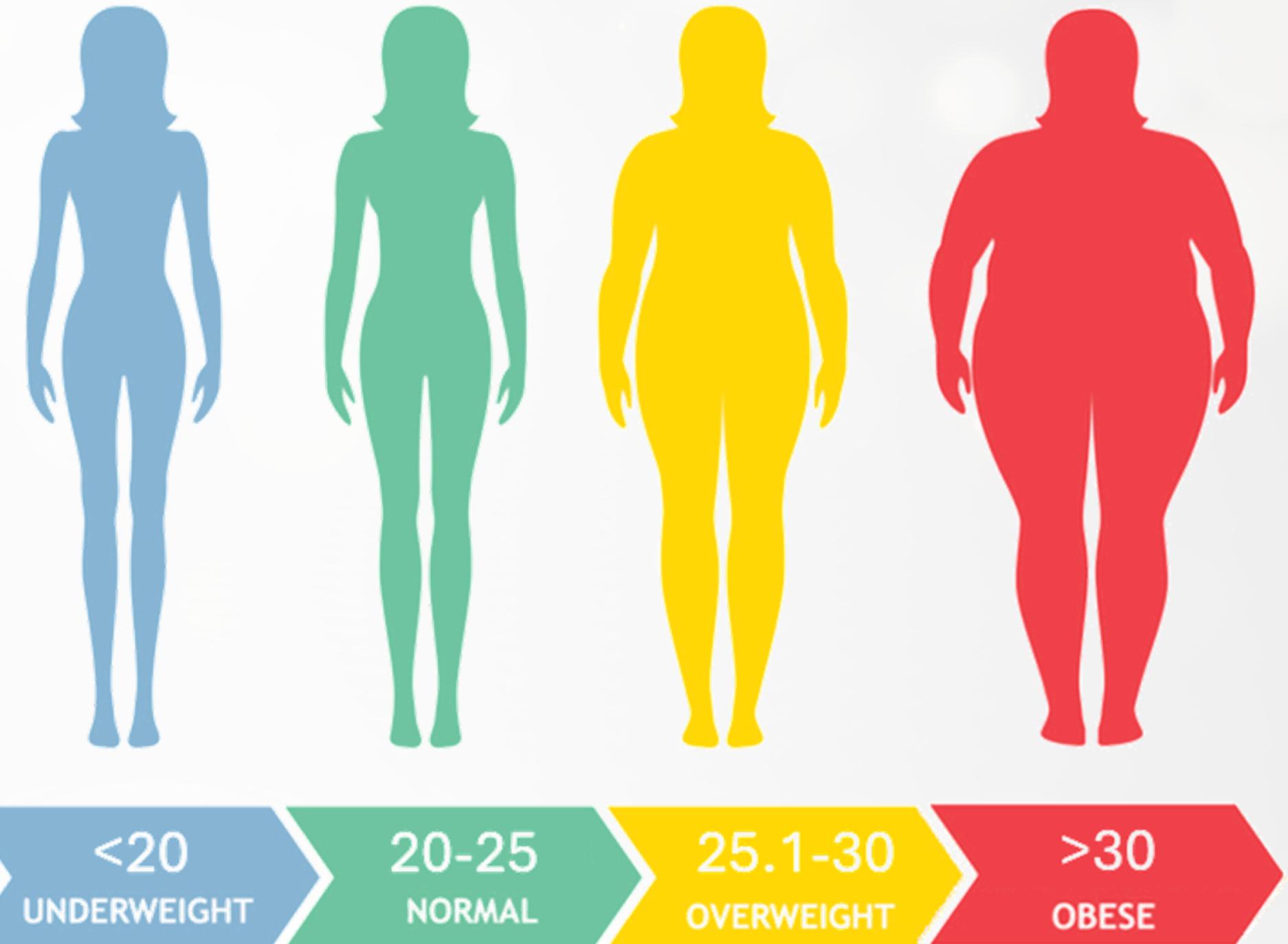


# 3

## BMI

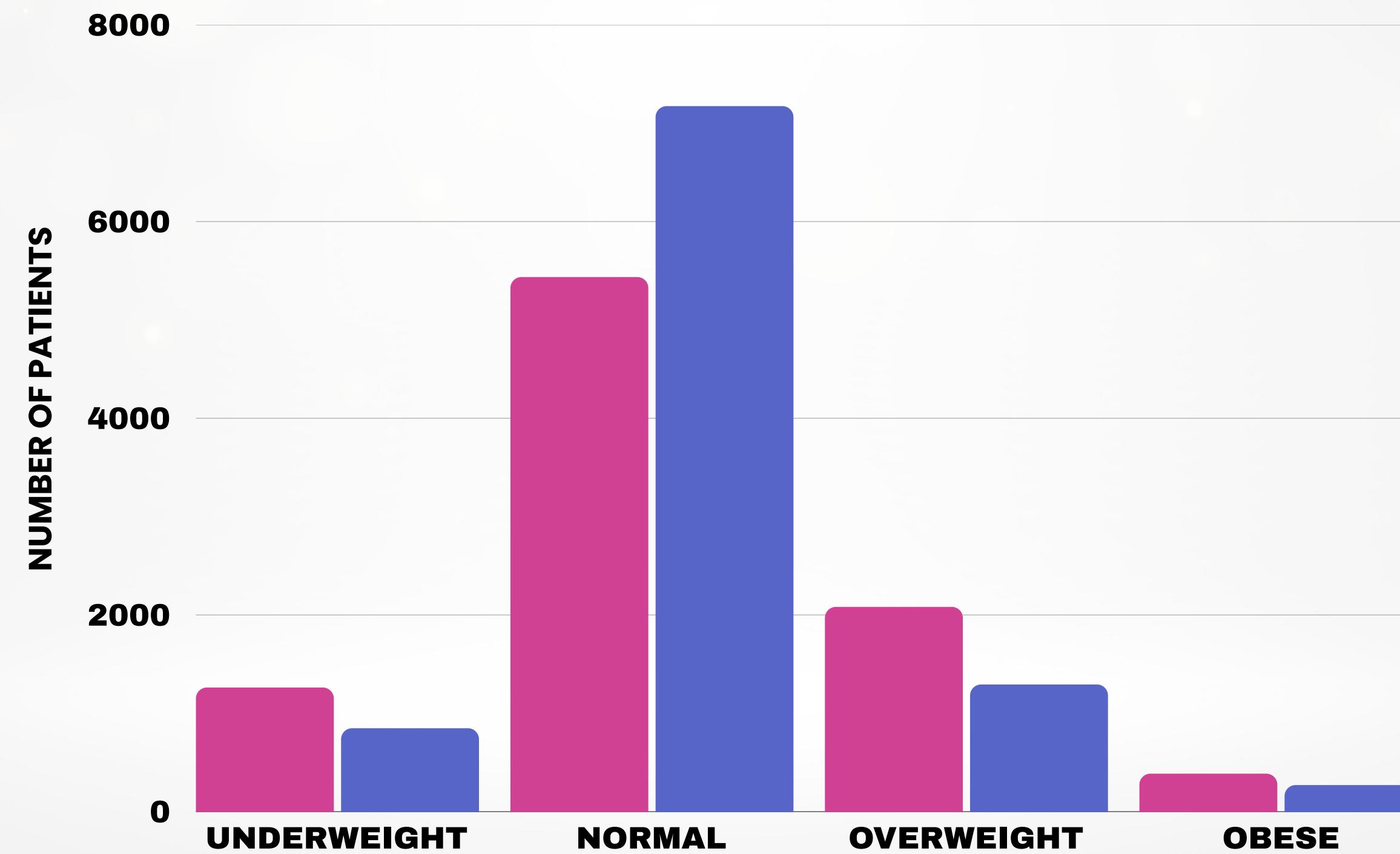
Body Mass Index (BMI) is a numerical value that estimates a **person's body fat** based on their weight relative to height.

$$BMI = \frac{weight(kg)}{(height(m))^2}$$



# BMI CATEGORIES BY SEX

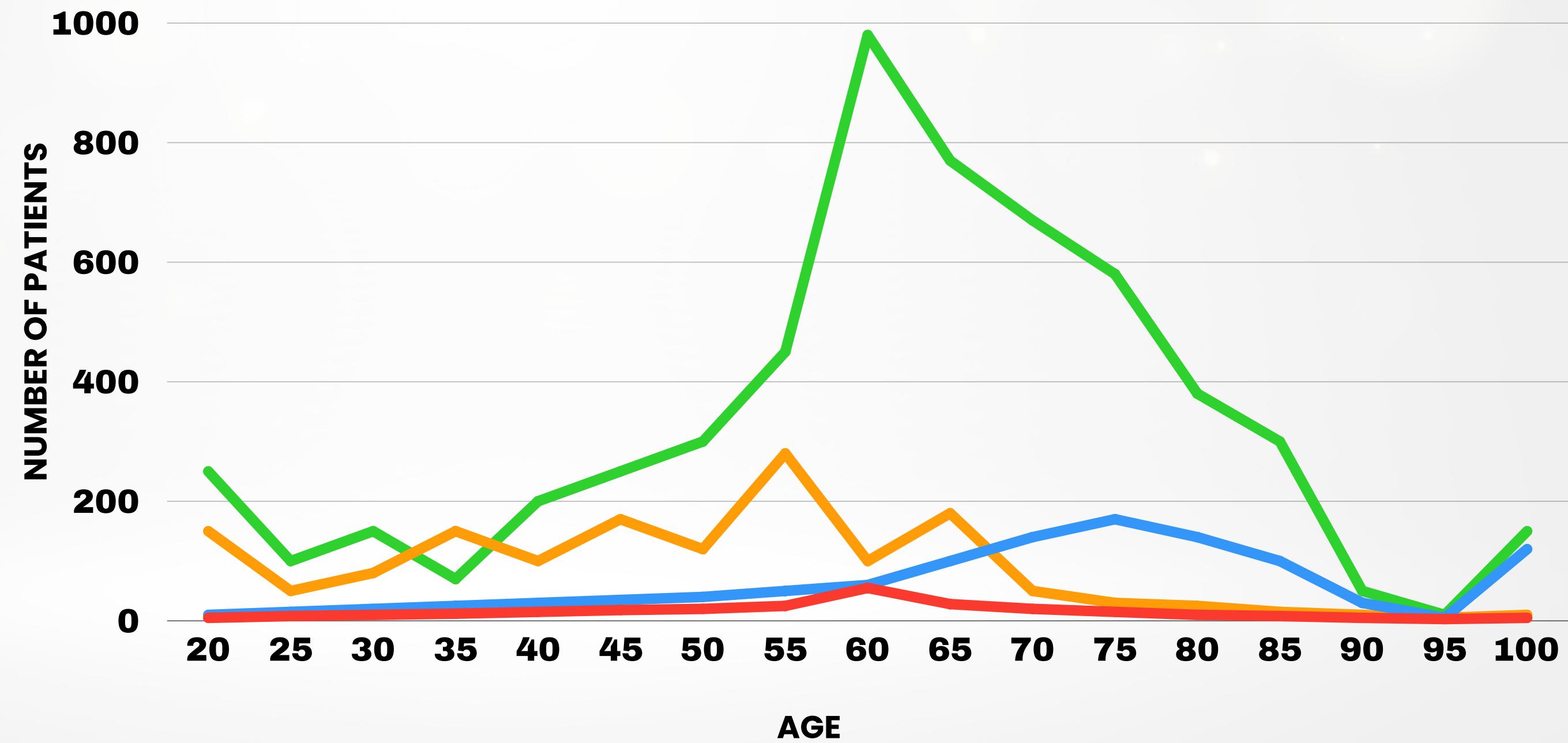
Female  
Male



# BMI CATEGORIES BY AGE



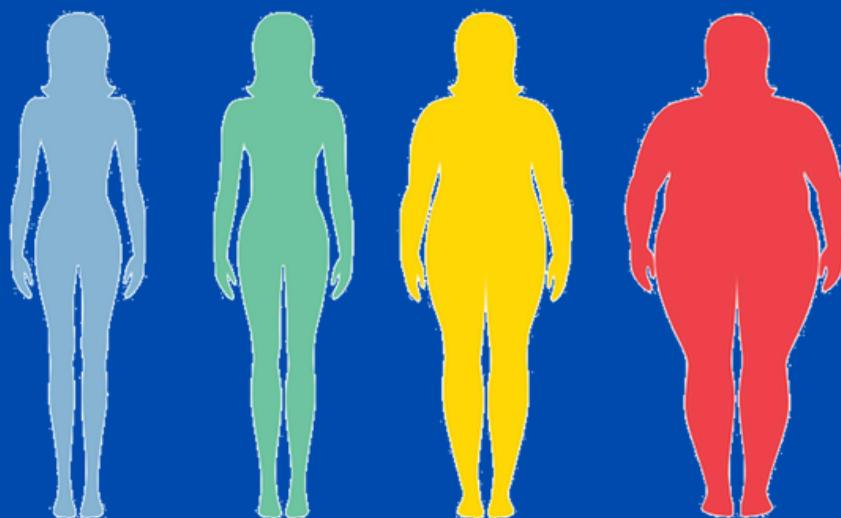
- Underweight
- Normal Weight
- Overweight
- Obese



# 4

## BMI & PATHOLOGIES

In which BMI category  
is it most common to  
**develop** a specific  
heart **disease**?



$$\text{Enrichment} = \frac{P_{\text{disease in BMI group}}}{P_{\text{disease in total population}}}$$



Mathematical  
formula

- $E = \frac{P(\text{disease}|\text{BMI group})}{P(\text{disease})}$
- $P(\text{disease}) = \frac{\text{total number of patients with the disease}}{\text{total number of patients}}$

# OUR CONCLUSIONS



Is the Body Mass Index (**BMI**) related to the appearance of particular **cardiovascular diseases**?

We discovered that some heart diseases are more linked to the extreme BMI categories: **underweight** and **obese**.

For example:

- Sinus tachycardia (STACH) - **increased heart rate** that occurs when the heart has to work harder, even with minimal effort.
- Atrial Overload/Enlargement (AO/AE) - **atrium muscle** becomes **thicker** because it works harder than normal.



Other pathologies, like Myocardial Infarction (MI), become more frequent as we **age**.

**THANK YOU**