**Histogram plots for each column in the dataset**

Logo

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**A picture containing icon

Description automatically generated**

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing shape

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing chart

Description automatically generated

Chart, bar chart

Description automatically generated

Icon

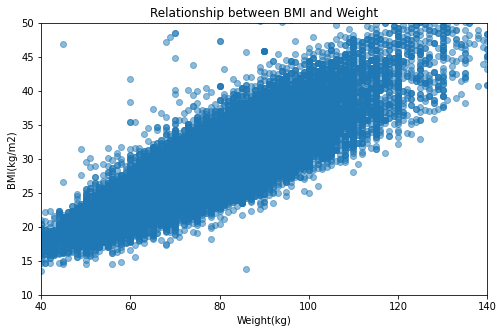
Description automatically generated

**Correlation matrix plot between the columns in the dataset**

A picture containing square

Description automatically generated

**Scatter plot between the BMI and weight columns (the most correlated ones)**



**Target label distribution the dataset**

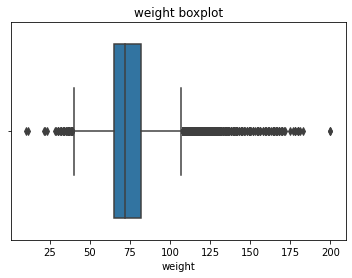
Icon

Description automatically generated

**Boxplots for numerical features in the dataset for outlier detection**

Chart, diagram, box and whisker chart

Description automatically generated



A picture containing shape

Description automatically generated

Shape, rectangle

Description automatically generated

Chart, box and whisker chart

Description automatically generated

A picture containing box and whisker chart

Description automatically generated

**Scatter plot between features in the dataset for outlier detection**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Shape

Description automatically generated

Chart

Description automatically generated with low confidence

**Outliers’ detection: 1) visually**

height>=250

weight>=200 and weight<=25

ap\_hi>=10000

ap\_lo>=5000

age\_years<=30

Size of original dataset = 70,000

Size of dataset after outliers’ removal = 69,955 (45 outliers)

**Outliers’ detection: 2) IQR score (interquartile range)**

1. We calculate the 25% quartile of the data (Q1), the 75% quartile of the data (Q3), and the difference between them (IQR = Q3-Q1)
2. We remove any data point x that fits this condition:

x< (Q1 - 1.5 \* IQR) and x> (Q3 + 1.5 \* IQR)

Size of original dataset = 70,000

Size of dataset after outliers’ removal = 37,602 (32,398 outliers)

**Outliers’ detection: 3) Z-score (interquartile range)**

1. We calculate the z score of the data distribution (z = [data – mean]/standard deviation)
2. Set a threshold = 3 and find the data points that have a z-score < threshold to be considered an outlier

Size of original dataset = 70,000

Size of dataset after outliers’ removal = 55,604 (14,396 outliers)

**Features selection**

**Feature Selection with Univariate Statistical Tests**

* Method 1: Feature Extraction with RFE (Logistic regression)

🡪 Features ranking:

[('ap\_hi', 1),

('cholesterol', 1),

('gluc', 1),

('alco', 1),

('active', 1),

('bmi', 1),

('gender', 2),

('smoke', 3),

('height', 4),

('weight', 5),

('age\_years', 6),

('ap\_lo', 7)]

Get features having features ranking = 1 (best estimated features):

['ap\_hi', 'cholesterol', 'gluc', 'alco', 'active', 'bmi']

* Method 2: Features Importance with Extra Trees Classifier

🡪 Features importance:

[('ap\_hi', 0.1868844524597645),

('bmi', 0.16711438339373774),

('weight', 0.15461972304489624),

('age\_years', 0.14770610031360512),

('height', 0.1459426629263903),

('ap\_lo', 0.10923315694269002),

('cholesterol', 0.044139116106184685),

('gluc', 0.01447066508142241),

('gender', 0.009778959534314121),

('active', 0.007701173530480231),

('smoke', 0.006744981215201378),

('alco', 0.005664625451313029)]

Get features having features importance > 0.01 (threshold for best estimated features):

['ap\_hi', 'bmi', 'weight', 'age\_years', 'height', 'ap\_lo', 'cholesterol', 'gluc']

* Method 3: Removing correlated features

🡪 Correlated features with score > 0.7:

[‘bmi’]

Finally, select features that exits in the intersection of the three selected features:

['height', 'weight', 'ap\_hi', 'ap\_lo', 'cholesterol', 'gluc', 'alco', 'active', 'age\_years']