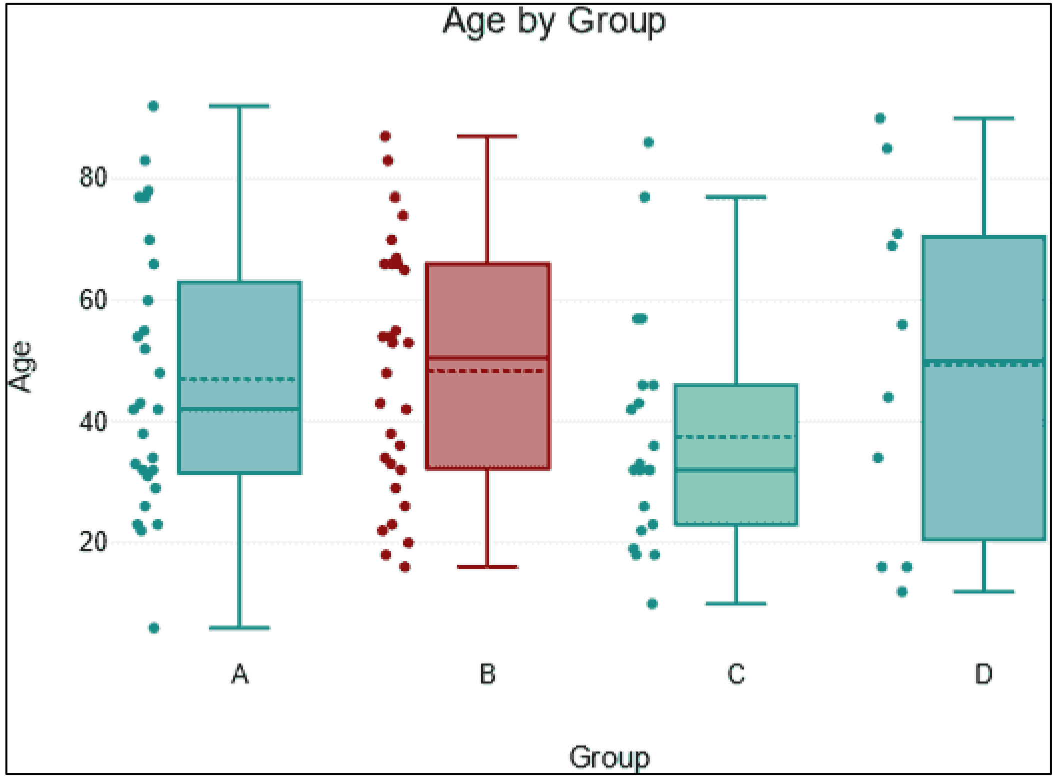
Data Cleaning and Pre-processing Laboratory Activity

IT 116 – Information Management 2

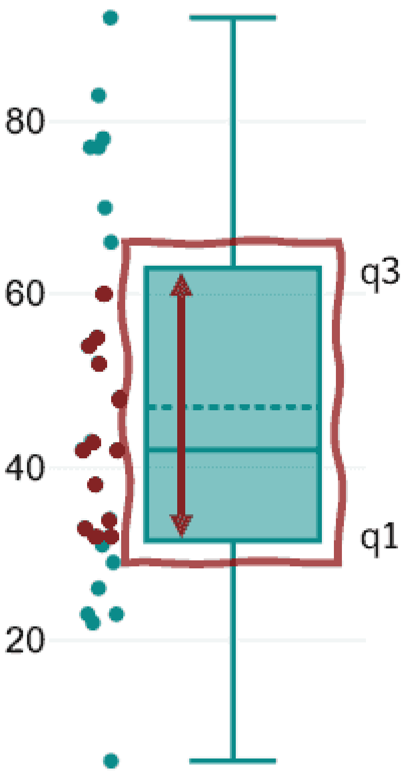
Data Cleaning on Loans Data

1.Importing libraries
2.Importing datasets
3.Missing Values detection and treatment
4.Outliers detection and treatment
5.Transformation of Variables
6.Scaling the Numerical Variables
7.Encoding the Categorical Variables
8.Creation of New Variables
9.Splitting the data into training and test set

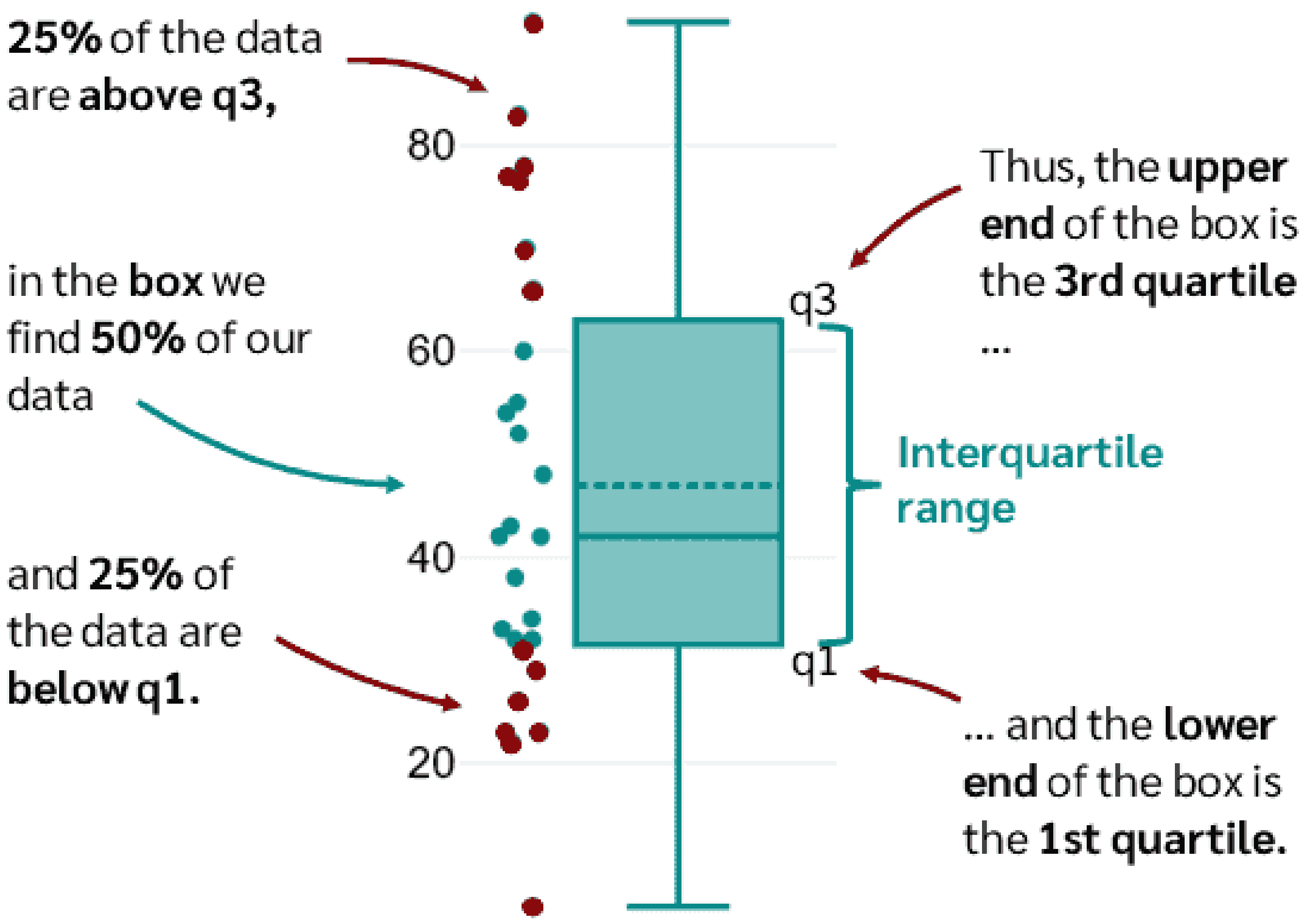
Interpreting Boxplots



The box itself indicates the range in which the middle 50% of all values lie. Thus, the lower end of the box is the 1st quartile and the upper end is the 3rd quartile.



Interpreting Boxplots



25% of the data are **above q3,**

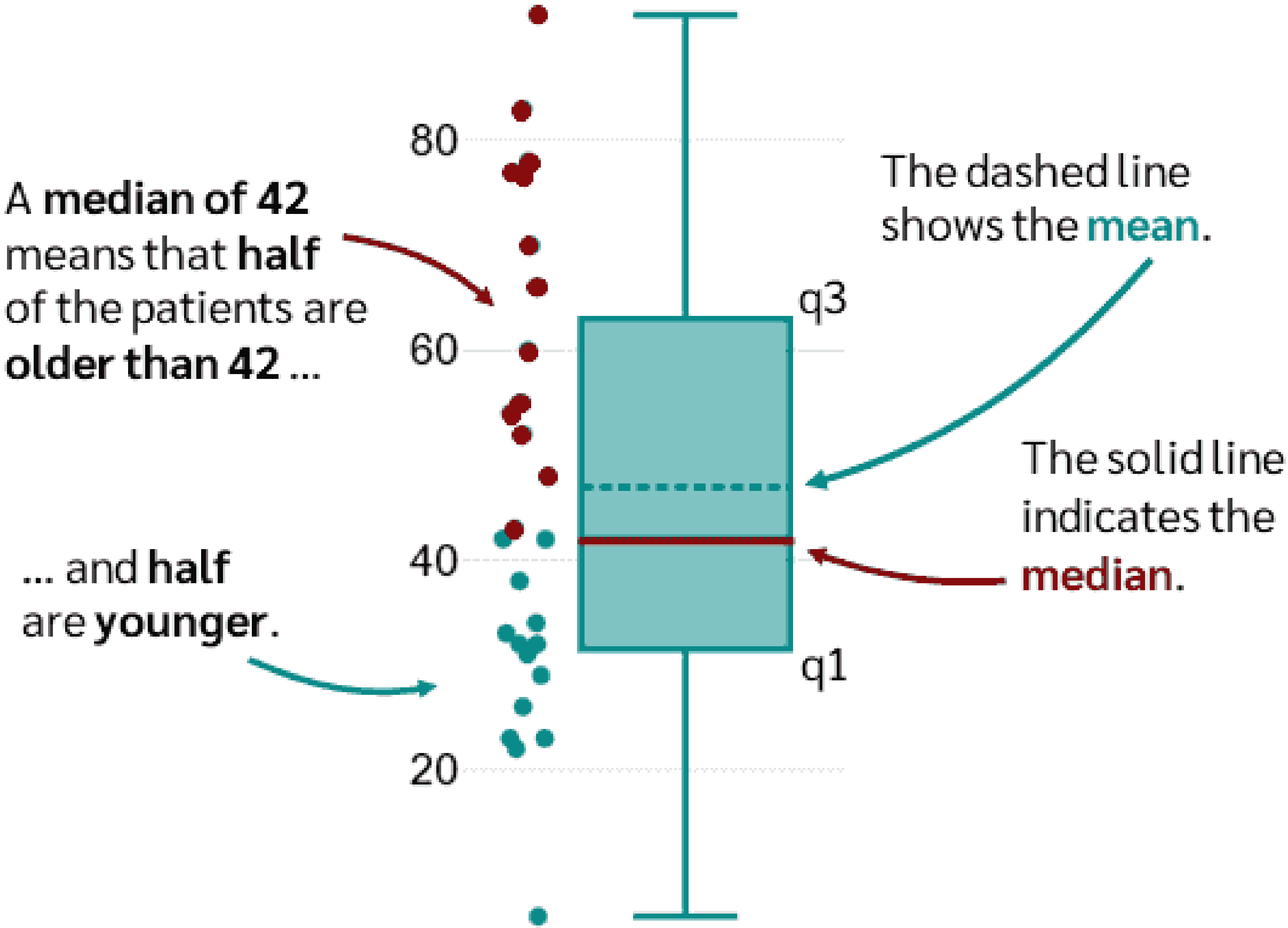
in the **box** we find **50%** of our data

and **25%** of the data are **below q1.**

Thus, the **upper end** of the box is the **3rd quartile** …

… and the **lower end** of the box is the **1st quartile.**

Interpreting Boxplots



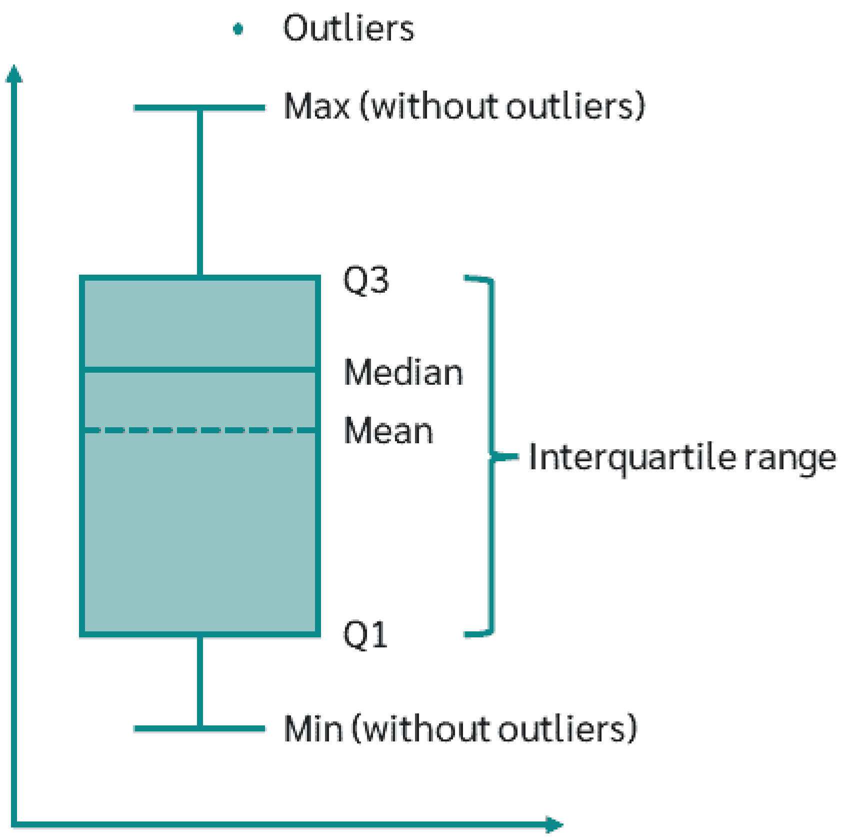
A **median of 42** means that **half** of the patients are **older than 42** ...

... and **half** are **younger**.

The dashed line shows the **mean**.

The solid line indicates the **median**.

Interpreting Boxplots



The box indicates the range in which the middle 50% of all data lies

Thus, the lower end of the box is the 1st quartile and the upper end is the 3rd quartile

Between Q1 and Q3, is the interquartile range

In the boxplot, the solid line indicates the median and the dashed line indicates the mean.

The T-shaped whiskers go to the last point, which is still within 1.5 times the interquartile range.

Points that are further away are considered extreme values (outliers).

Treat Outliers : Z-score Approach

- All the values above 3 standard deviation and below -3 standard deviation are outliers and can be removed

Lab Activity

- Download the Jupyter notebooks and datasets from this link and try running the different data cleaning and preprocessing approaches on your computers/laptops.

https://github.com/TrilokiDA/Data-pre-processing/tree/master

- **Groupwork**:

In a Jupyter notebook, perform exploratory data analysis, descriptive analysis and data cleaning on your own dataset.