Exploratory Data Analysis

Data Science & Business Analytics

Duarte Gomes

Test

Which of these is most likely to have a roughly symmetric distribution?

(a) Salaries of a random sample of people from Portugal

(b) Weights of adult females

(c) Last digits of phone numbers

How do the **mean** and **median** of the following two datasets compare?

Dataset 1: 30, 50, 70, 90

Dataset 2: 30, 50, 70, 1000

- (a) $\bar{x}1 = \bar{x}2$, median1 = median2
- (b) $\bar{x}1 < \bar{x}2$, median1 = median2
- (c) $\bar{x}1 < \bar{x}2$, median1 < median2
- (d) $\bar{x}1 > \bar{x}2$, median1 < median2
- (e) $\bar{x}1 > \bar{x}2$, median1 = median2

The **range** is always at least as large as the **IQR** for a given dataset?

- (a) Yes
- (b) No

Is the **range** or the **IQR** more robust to outliers?

- (a) Range
- (b) IQR

In Python perform the following exercises, using EDA techniques.

- **Exercise 1 :** To check minimum and maximum of ‘year’ column
- **Exercise 2 :** To find out total number of fires in ‘**Acre**’ state and visualizing data based on each ‘year’
- **Exercise 3 :** To find out total number of fires in all states
- **Exercise 4 :** To find out total number of fires in 2017 and visualizing data based on each ‘month’
- **Exercise 5 :** To find out average number of fires occurred
- **Exercise 6 :** To find out the state names where fires occurred in ‘December’ month