Data Journalism

Non-Visual Exploratory Data Analysis

Angelica Lo Duca
angelica.loduca@iit.cnr.it

Exploratory Data Analysis

1 **Extract** important information from data,such as pattern, trends and relationships among data.
2 **Understand** which questions data can answer and extract

EDA Workflow

Problem Setting

Define which questions the dataset can answer.

Typical questions:

- typical values of a column, its ranges, uncertainties, and distributions
- most important/influencing column
- Is there a correlation among columns?
- The best value for a given column
- Presence of outliers
- Missing values

Preliminary Data Analysis

Discover:

- hidden patterns
- relationships between the data
- recurring trends
- extract important variables
- detect outliers and anomalies
- …

Provide an initial answer to the questions we asked previously.

Preliminary Data Analysis (cont.)

Two types of preliminary data analysis:

- **univariate analysis** – focus only on a single variable at a time.
- **multivariate analysis** – when we focus on multiple variables at a time.

Preliminary Results

- We choose only relevant data, i.e. data able to answer our questions.

EDA Main techniques

- **Non-visual EDA** – calculate statistics or metrics to extract insights from data.
- **Visual EDA** – use graphs to extract insights from data.

Non-visual EDA

Non-visual EDA

Non-visual EDA calculates **some statistics on data**.

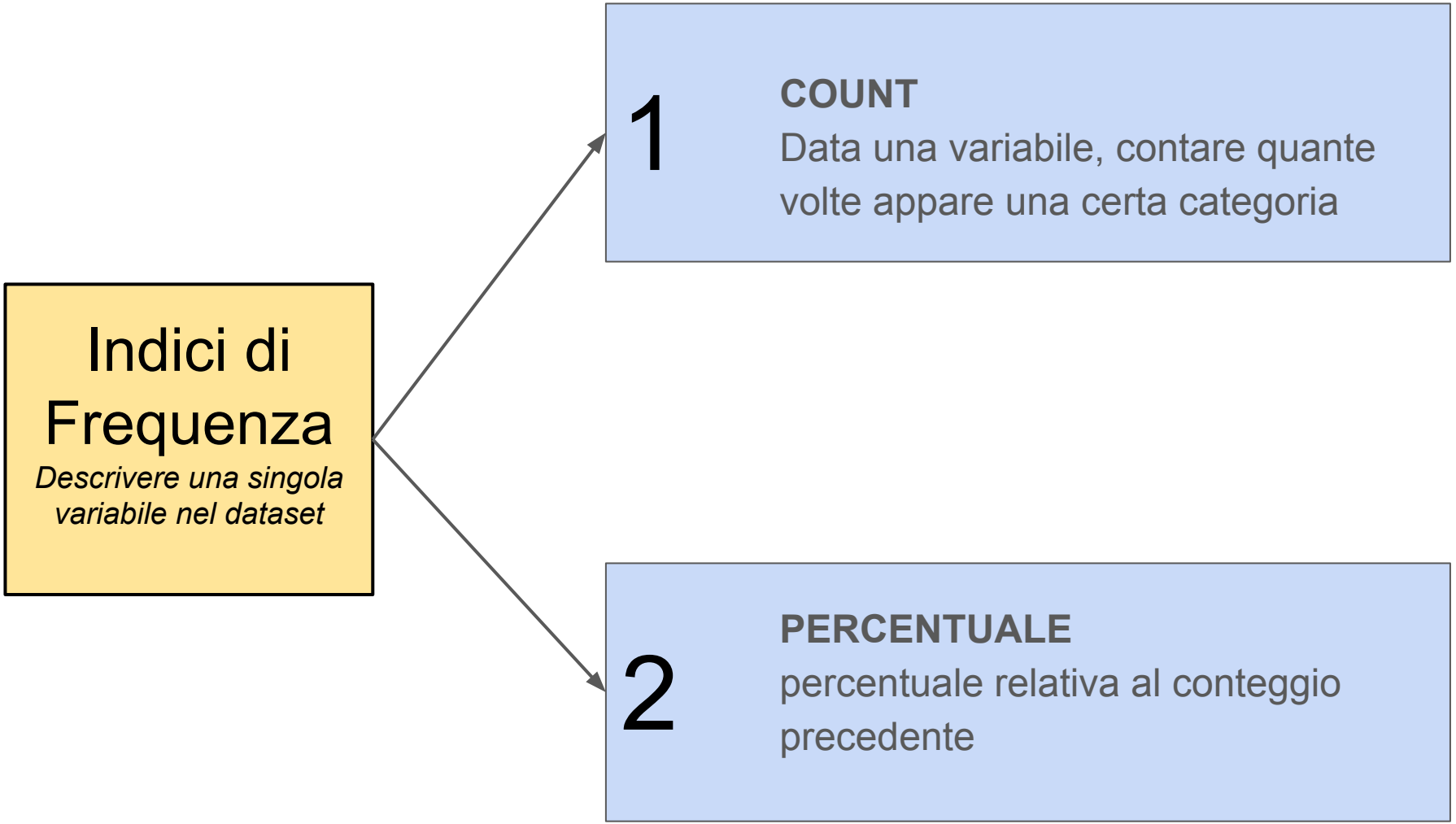
univariate analysis → calculate descriptive statistics (for numerical data), missing values, negative and distinct values, and memory size.

multivariate analysis → calculate the correlation among columns

Statistica descrittiva

Si basa sul calcolo di alcune **metriche** o **indici**

- Indici di frequenza
- Indici di tendenza centrale
- Indici di variabilità


Indici di Frequenza
Descrivere una singola variabile nel dataset
1
COUNT
Data una variabile, contare quante volte appare una certa categoria
2
PERCENTUALE
percentuale relativa al conteggio precedente

Indici di frequenza

Presentazione grafica

grafici a barre, linee, diagrammi a torta, distribuzione della frequenza

Indici di Tendenza Centrale

Descrivere i dati con un solo valore

1

MEDIA ARITMETICA

Somma dei dati

Numero dei dati

2

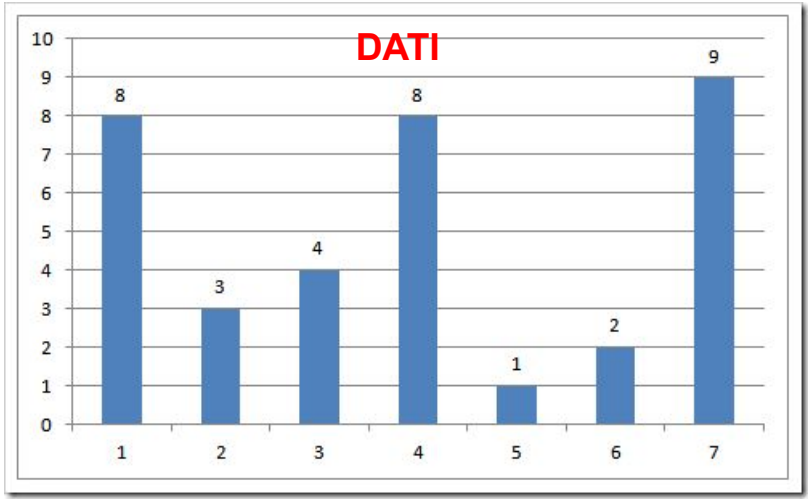
MEDIANA (o 50° percentile)

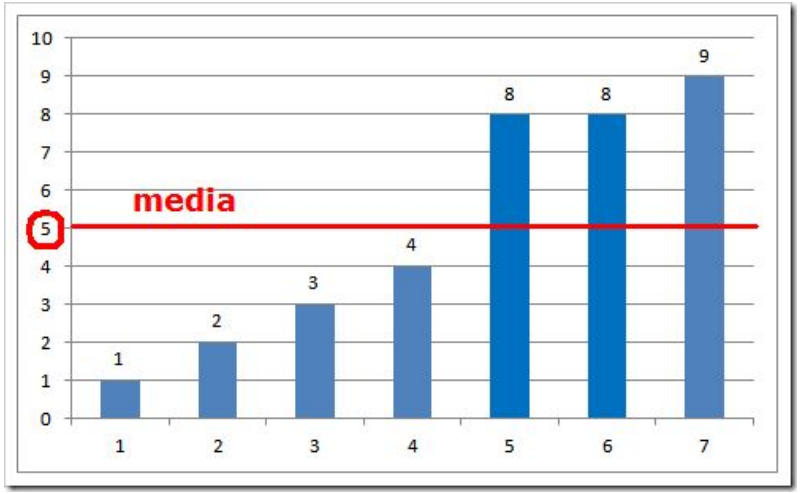
valore al di sotto del quale cade la metà dei dati (valore centrale)

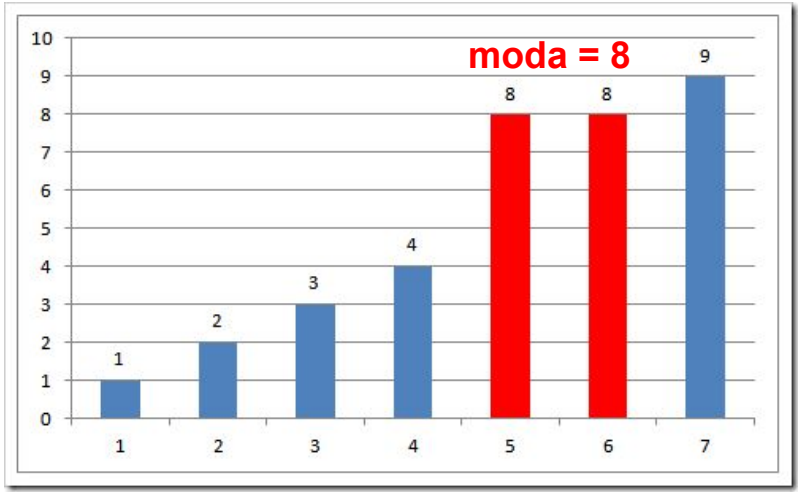
3

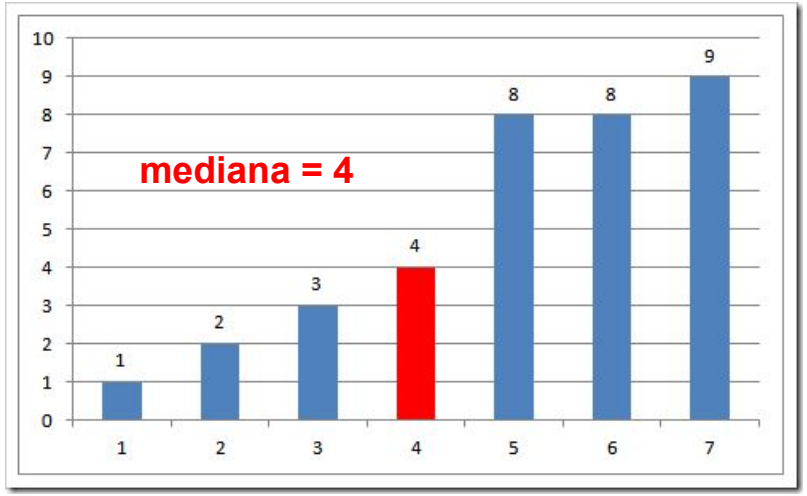
MODA

valore che ricorre con maggiore frequenza


DATI
10
9
8
7
6
5
4
3
2
1
0
8
3
4
8
1
2
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
media
1
2
3
4
8
8
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
moda = 8
1
2
3
4
8
8
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
mediana = 4
1
2
3
4
8
8
9
1
2
3
4
5
6
7

Indici di Variabilità

Descrivere la variabilità dei dati

1

MAXIMUM valore massimo
MINIMUM valore minimo
RANGE Differenza tra il valore massimo e il valore minimo

2

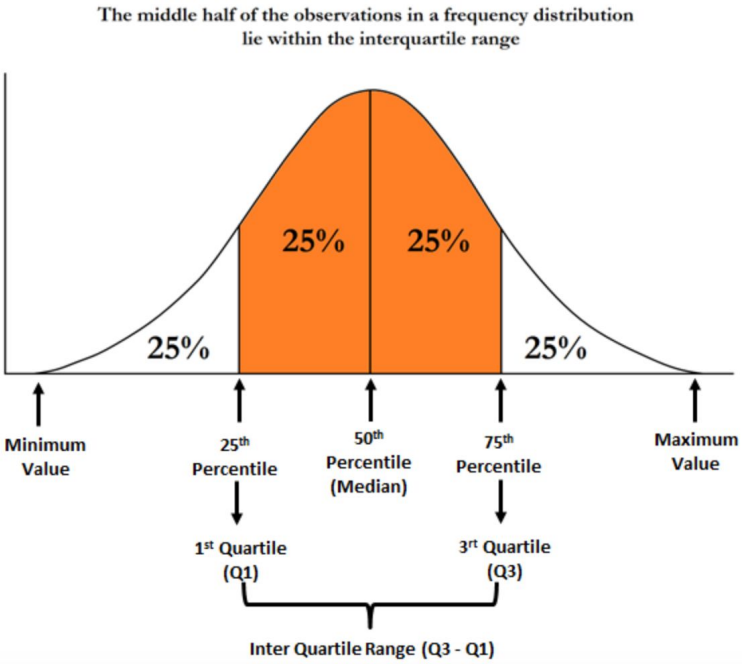
QUARTILE

3

VARIANZA
dispersione dei valori del dataset attorno al valor medio. La deviazione standard è la radice quadrata della varianza

Quartile

I quartili dividono un set di dati in 4 parti uguali e si riferiscono ai valori del punto tra i quarti. Il Quartile inferiore (Q1) è il punto tra il 25% più basso di valori e il 75% più alto di valori. È anche chiamato il 25 ° percentile. Il secondo quartile (Q2) è il centro del set di dati. È anche chiamato 50 ° percentile, o mediana. Il quartile superiore (Q3) è il punto tra il 75% più basso e il 25% più alto di valori. È anche chiamato il 75 ° percentile.



Non-Visual EDA in Python

pandas-profiling

```
pip install pandas-profiling
```

```
from pandas_profiling import ProfileReport

profile = ProfileReport(df, title="my_title")

profile.to_file("my_report.html")
```

Exercise

Source Dataset

2022 Ukraine Russia War

Equipment losses & Death Toll & Military Wounded & Prisoner of War of russians

Fields

Personnel
Prisoner of War
Armored Personnel Carrier
Multiple Rocket Launcher
Aircraft
Anti-aircraft warfare
Drone
Field Artillery
Fuel Tank
Helicopter
Military Auto
Naval Ship
Tank

POW - Prisoner of War,
MRL - Multiple Rocket Launcher,
APC - Armored Personnel Carrier,
SRBM - Short-range ballistic missile,
drones:
UAV - Unmanned Aerial Vehicle,
RPA - Remotely Piloted Vehicle.

Tasks

- Download the datasets
- Load them in two Pandas Dataframe
- Merge them as follows: `df_tot = df_p.join(df_e, on='date')`
- Produce a Pandas Profiling report
- Identify at least two relevant questions your data can answer
- Try to answer to the questions

Possible Questions

- Which is the average number of daily dead soldiers?
- Is there a correlation between the number of dead soldiers and the use of drones?