Data Analysis and Visualization

CentraleDigitalLab@Nice

Deborah Dore - ddore@i3s.unice.fr
MARIANNE, I3S, CNRS, INRIA, Université Côte d'Azur

Lesson 1: Basics of Data Analysis

What is a Random Variable?

A random variable (r.v.) X is a function $X: \Omega \rightarrow R$ where Ω is the state space and R is the set of values that the variable can take called Range.

Lesson 1: Basics of Data Analysis

What is a Random Variable?

A random variable (r.v.) X is a function $X: \Omega \rightarrow R$ where Ω is the state space and R is the set of values that the variable can take called Range.

Intuitively, a r.v. is equivalent to a column of your dataset after applying zero or more filters.

Lesson 1: Basics of Data Analysis

What is a Random Variable?

A random variable (r.v.) X is a function $X: \Omega \rightarrow R$ where Ω is the state space and R is the set of values that the variable can take called Range.

Intuitively, a r.v. is equivalent to a column of your dataset after applying zero or more filters.

A random variable can be of different types: numerical or categorical.

Lesson 1: Basics of Data Analysis

What is a Random Variable?

Numerical:

- **Continuous**: Can take on any value within a range, including decimals and fractions. E.g., the height of students in a school (150.2 cm, 165.8 cm …).
- **Discrete**: Can take on specific, separate values and is countable. E.g., the number of cars passing a toll booth in a day (0, 1, 2, 3 …).
    - **Finite Set**: The number of siblings a person has (e.g., 0, 1, 2, 3... up to a reasonable maximum).
    - **Infinite Set**: The number of times you need to roll a dice until you get a six (potentially infinite but countable).

Lesson 1: Basics of Data Analysis

What is a Random Variable?

Categorical: Variables that represent distinct groups or categories.

- **Nominal:** no inherit order. E.g., eye color of individuals (Blue, Brown, Green).
- **Ordinal**: Variables with a meaningful order or ranking. E.g., rating of a restaurant on a scale from 1 to 5 (Poor, Fair, Good, Very Good, Excellent).

Lesson 1: Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Lesson 1: Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Columns

(Random Variables)

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Lesson 1: Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Rows

(Elements of Ω)

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Lesson 1: Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Set of values of a r.v.

(Range R)

Lesson 1: Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Set of values of a r.v.

(Range R)

Lesson 1: Basics of Data Analysis

What is wrong in this dataset?

Age	Height	Degree's level
150	172	Master
26	167	University
22	170	Bachelor
23		Bachelor

Lesson 1: Basics of Data Analysis

What is wrong in this dataset?

Outlier

Age	Height	Degree's level
80	172	Master
26	167	University
22	170	Bachelor
23		Bachelor

Lesson 1: Basics of Data Analysis

What is wrong in this dataset?

Age	Height	Degree's level
80	172	Master
26	167	University
22	170	Bachelor
23		Bachelor

Outlier

Missing value

Lesson 1: Basics of Data Analysis

What is wrong in this dataset?

Age	Height	Degree's level
80	172	Master
26	167	University
22	170	Bachelor
23		Bachelor

Outlier

Wrong value

Missing value

Lesson 1: Basics of Data Analysis

Outlier

An **outlier** is a data point that differs significantly from other observations in a dataset. It can be much higher or lower than the expected range of values, making it stand out.

Lesson 1: Basics of Data Analysis

Outlier

An **outlier** is a data point that differs significantly from other observations in a dataset. It can be much higher or lower than the expected range of values, making it stand out.

It’s important to identify them!

- They can **distort statistical analyses**, such as averages and regression results.
- They might indicate **errors** in data collection or entry.
- They can reveal **insights** about rare or unexpected phenomena.

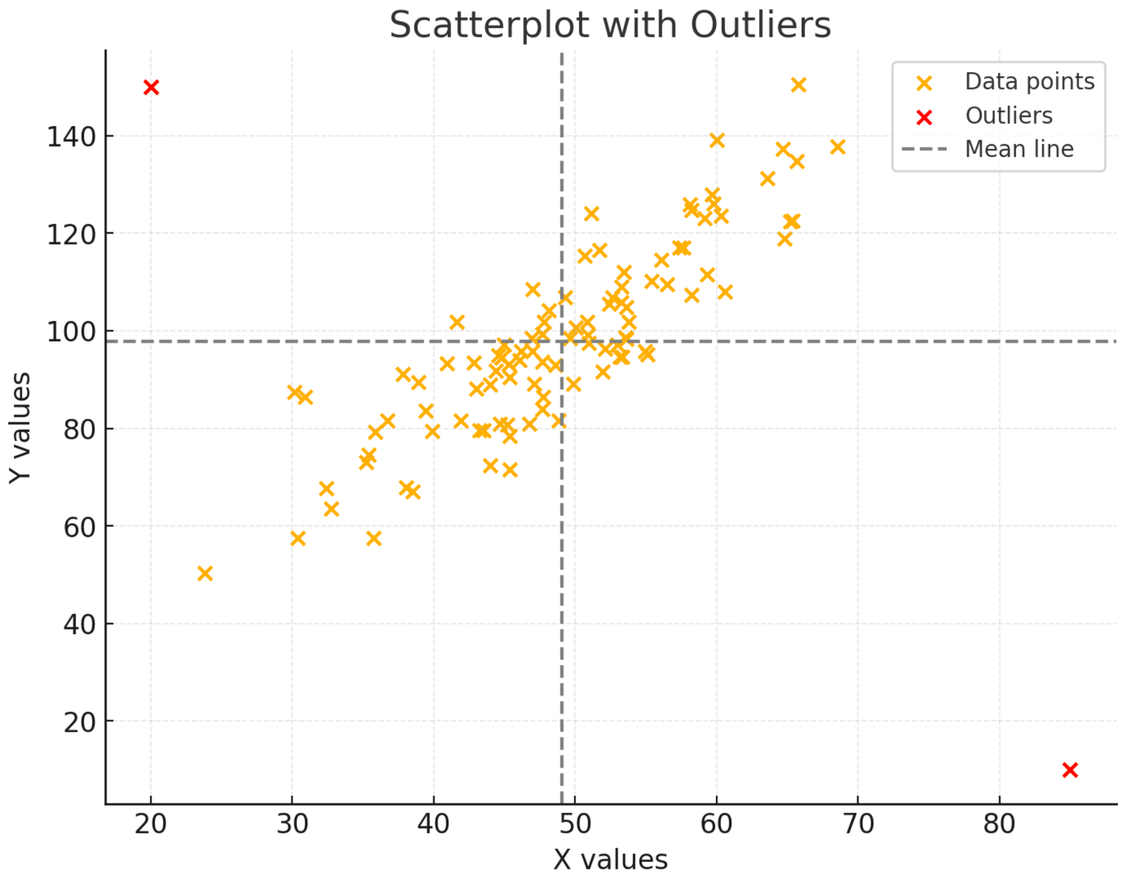
Lesson 1: Basics of Data Analysis

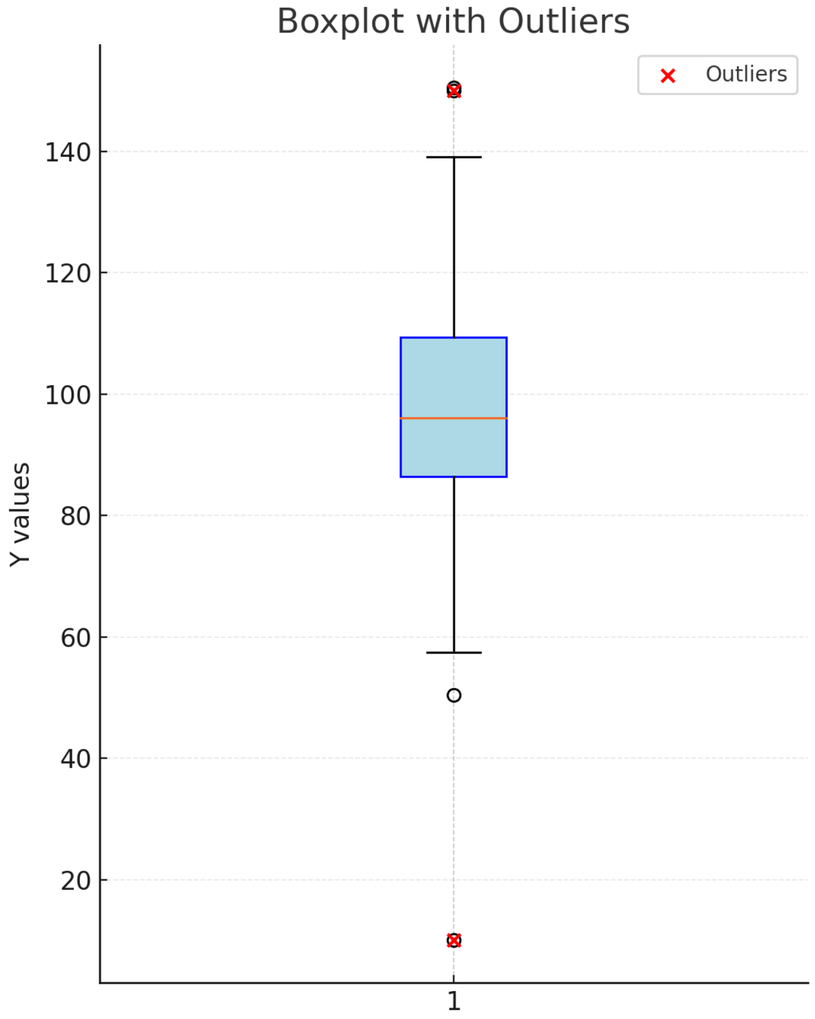
How to identify outliers?

1. **Visual Methods**: Scatterplots, boxplots, or histograms.
2. **Statistical Methods**:
    1. Z-scores (values more than 3 standard deviations away from the mean).
    2. Interquartile Range (IQR): Points outside $1.5 \times \text{IQR}$ beyond the first and third quartiles.
3. **Model-Based Approaches**: Identifying outliers based on residuals in regression or clustering methods.

Lesson 1: Basics of Data Analysis

Example





Lesson 1: Basics of Data Analysis

What is wrong in this dataset?

Age	Height	Degree's level
80	172	Master
26	167	University
22	170	Bachelor
23		Bachelor

Outlier

Wrong value

Missing value

Lesson 1: Basics of Data Analysis

How to correct outliers, wrong values and missing values?

1. **Outliers:**
    1. Leave them as they are (if they don’t affect too much the statistics)
    2. Apply transformations to reduce their impact (es. Log)
    3. Replace with an estimate (mean/median/mode/predict with ML models)
    4. Remove whole row (only if the dataset is sufficiently big)
2. **Wrong values:**
    1. Check the source of the data for corrections
    2. Replace with an estimate (mean/median/mode/predict with ML models)
    3. Remove whole row (only if the dataset is sufficiently big)
3. **Missing values:**
    1. Replace with an estimate (mean/median/mode/predict with ML models)
    2. Remove whole row (only if the dataset is sufficiently big)

Lesson 1: Basics of Data Analysis

How to correct outliers, wrong values and missing values?

	Leave them as they are	Apply transformations	Replace with estimate	Remove row	Check the source
Outliers	✅	✅	✅	✅	✅
Wrong Values			✅	✅	✅
Missing Values			✅	✅	

Lesson 1: Basics of Data Analysis

Replace with an estimate

1. **Mean & Median (only numeric variables):**
    1. **Mean:** $\frac{\sum_{i=1}^{n} x_i}{n}$ where x_i is each element in the column and n in the total number of elements
    2. **Median:** the middle value when the column is sorted in ascending order.
2. **Mode (both):**
    1. The most frequently occurring value in the column.
3. **Machine Learning model prediction (both):**
    1. **KNN Estimation**: Predict missing values based on the closest similar data points (neighbors).
    2. **Regression Imputation**: Use a regression model to predict missing values based on relationships in the data.

Lesson 1: Basics of Data Analysis

Categorical Variable Encoding

Categorical: Variables that represent distinct groups or categories.

- **Nominal:** no inherit order. E.g., eye color of individuals (Blue, Brown, Green).
- **Ordinal**: Variables with a meaningful order or ranking. E.g., rating of a restaurant on a scale from 1 to 5 (Poor, Fair, Good, Very Good, Excellent).

Lesson 1: Basics of Data Analysis

Categorical Variable Encoding

Categorical: Variables that represent distinct groups or categories.

- **Nominal:** no inherit order. E.g., eye color of individuals (Blue, Brown, Green).
- **Ordinal**: Variables with a meaningful order or ranking. E.g., rating of a restaurant on a scale from 1 to 5 (Poor, Fair, Good, Very Good, Excellent).

Why Encode Them?

- Many machine learning models require numerical input.
- Proper encoding improves model performance and interpretability.

Lesson 1: Basics of Data Analysis

Encoding Techniques

One Hot Encoding: converts categories into binary columns.

- [“Red," "Green," “Blue"] → [100, 010, 001].
- **Pros:** No ordinal assumptions; widely supported.
- **Cons:** Can lead to high-dimensional data.

Red	Green	Blue
1	0	0
0	1	0
0	0	1

Lesson 1: Basics of Data Analysis

Encoding Techniques

One Hot Encoding: converts categories into binary columns.

- [“Red," "Green," “Blue"] → [100, 010, 001].
- **Pros:** No ordinal assumptions; widely supported.
- **Cons:** Can lead to high-dimensional data.

Red	Green	Blue
1	0	0
0	1	0
0	0	1

Label Encoding: Assigns unique integer values to each category.

- ["Red," "Green," "Blue"] → [0, 1, 2].
- **Pros**: Simple and memory-efficient.
- **Cons:** Assumes ordinal relationship, which might mislead models.

Red	0
Green	1
Blue	2

Demo with notebook

Lesson 1: Basics of Data Analysis

Summary

- A random variable (r.v.) X is a function X: Ω → R where Ω is the state space and R is the set of values that the variable can take called Range.
  - A r.v. can be numerical (continuous/discrete) or categorical (nominal/ordinal)
- Outliers and missing values are important to detect: different ways to detect them!
- Computers only understand numbers: convert categorical variables into numerical!