

# AEA. Lab 11: Exploratory Data Analysis

May 8, 2026

**Exploratory data analysis (EDA)** is a 1st step in analyzing the data from an experiment. Types of analysis:

- univariate/multivariate non-graphical
- univariate/multivariate graphical.

**Univariate** methods consider one variable at a time, while **multivariate** methods look at two or more variables to explore relationships.

**Non-graphical** methods: calculation of summary statistics; **graphical** methods summarize the data in a diagrammatic way.

## 1 Univariate non-graphical EDA

To better appreciate the “sample distribution” of a variable; is the population distribution compatible with the sample distribution?; outlier detection.

- **Categorical** data: the values and the frequency of occurrence for each value → **tabulation** of the frequency of each category.
- **Quantitative** (discrete and continuous) data: make preliminary assessments about the population distribution using the data of the observed sample.

**Descriptive Statistics:** to describe the basic features of data; the analysis of the characteristics of the variables in terms of:

- **central tendency**

The indicators:

- The sample (arithmetic) **mean** of  $n$  data values  $x_1, \dots, x_n$  is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- The sample **median** is the middle value after all of the values are put in an ordered list.

For symmetric distributions, the mean and the median coincide.

- The **mode** is the most likely or frequently occurring value.

The most common measure of central tendency is the mean. For skewed distributions, or when there is concern about outliers, the median may be preferred.

- **dispersion**

The indicators:

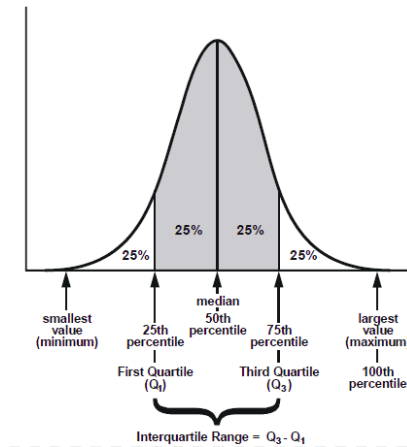
- the **sample variance**, for  $n$  observations  $x_1 \dots x_n$ , is the sum of the squared deviations, divided by  $n - 1$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- the **standard deviation**  $s$ : the square root of the variance  
For Normally distributed data: approximately 95% of the values lie within 2 sd of the mean.

- **interquartile range**

The **quartiles** are values that divide the data into 4 equal parts: 1/4 of the data fall below  $Q_1$ , 1/2 fall below  $Q_2$ , 3/4 fall below  $Q_3$ .



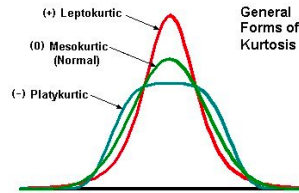
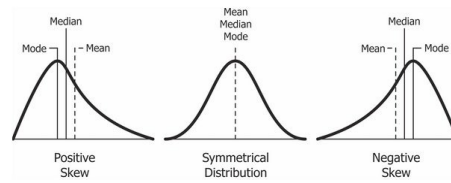
The interquartile range (IQR) is a robust measure of spread.

- **shape of distribution**

- **skewness** is a measure of asymmetry  
For a sample of  $n$  values, the sample skewness:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n \times s^3}$$

Positive skew: long tail on the right.



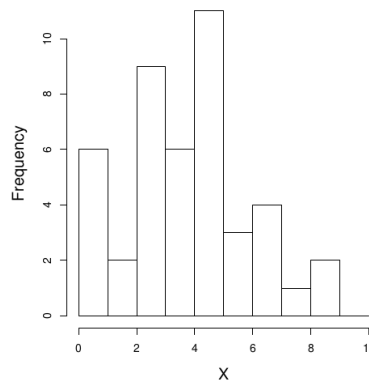
– **kurtosis** is a measure of “peakedness” relative to a Gaussian shape  
 For a sample of  $n$  values, the sample kurtosis:

$$b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n(s^2)^2}$$

Positive kurtosis: values far from the mean are more likely, and the shape of the histogram is peaked in the middle, but with fatter tails.

## 2 Univariate graphical EDA

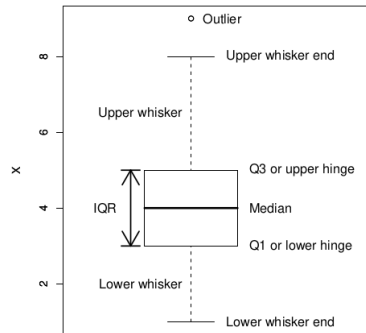
- **Histograms** (for categorical data): a barplot of the tabulation of the data. Each bar represents the frequency or proportion of cases for a range of values.



**Bin:** range of data for each bar.

Used to quickly learn about data: central tendency, spread, modality, shape and outliers.

- **Boxplots**: show extreme values, median and quartiles. You can easily figure out outliers.



Show information about the central tendency, symmetry and skew, outliers; misleading about multimodality.

Description:

- the min value (percentile 0)
- $Q_1$  determine the smallest 25% of values
- median
- $Q_3$  determine the largest 25% of values
- the max value

Analysis:

- "Boxplot outliers": more than 1.5 IQRs above  $Q_3$  or more than 1.5 IQRs below  $Q_1$
- Symmetry: if the median is in the center of the box and the whiskers are the same length  
Positive skewed: the median is closer to  $Q_1$  (mean > median)
- Positive kurtosis: fat tails (a histogram has a lot of values far from the mean, relative to a Gaussian distribution);  
Negative kurtosis: short whiskers.

### 3 Multivariate nongraphical EDA

Show the relationship between two or more variables.

- **Cross-tabulation**, for categorical data  
Column headings - the levels of one variable; row headings the levels of the other variable.

- **Correlation** and **covariance**, for two quantitative variables
  - Sample **covariance**: how much (and in what direction) should we expect one variable to change when the other changes.

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Obs:  $Cov(X, X) = Var(X)$ .

Positive covariance: when one measurement is above the mean, the other will probably also be above the mean, and vv. Covariances near zero: the two variables vary independently of each other.

- Sample **correlation**

$$Cor(X, Y) = \frac{Cov(X, Y)}{s_X s_Y}$$

$s_X$  the standard deviation of  $X$ ,  $s_Y$  the standard deviation of  $Y$ .

Obs:  $Cor(X, X) = 1$ .

Values between -1 and +1 (-1: a “perfect” negative linear correlation, +1: a perfect positive linear correlation, 0: uncorrelation).

Example:  $Cov=-110.6$  as age goes up, strength tends to go down (and vv).  $Cor=-0.91$  (strong negative correlation).

*Spearman’s rank correlation coefficient*: the statistics are calculated from the relative rank of values on each sample.

$$Cor(X, Y) = \frac{Cov(rank(X), rank(Y))}{stdev(rank(X)) * stdev(rank(Y))}$$

The two variables may have a nonlinear relationship.

- **Covariance and correlation matrices** for many quantitative variables: calculate all pairwise covariances and/or correlations.

## 4 Multivariate graphical EDA

- Univariate graphs by category: **Side-by-side boxplots**, for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable.
- **Scatter plots**: for two quantitative variables
  - If one variable is explanatory and the other is outcome, the convention is to put the outcome on the  $y$  axis.
  - Additional categorical variables can be accommodated on the scatterplot by encoding the additional information in the symbol type and/or color.

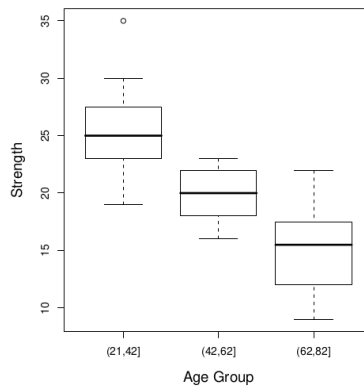


Figure 1: Side-by-side boxplot

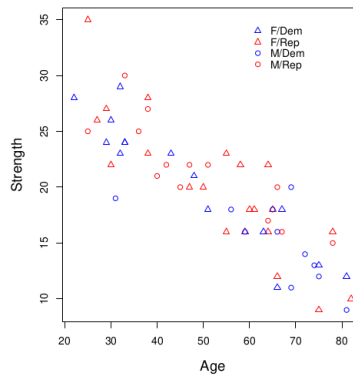


Figure 2: Age vs. strength; different colors and symbols are used to code political party and gender

- **Correlation heatmap** is a graphical representation of the correlation matrix representing the correlation between different variables.

**Homework:** choose a dataset (Chile voting dataset, EDA3) and analyse it.

#### 1. General overview

- Import data

```
import pandas as pd
df = pd.read_csv("data/chile.csv")
```

- Print information about data

```
print(df.head())
print(df.shape)
print(df.info())
```

A quick summary of data: `describe()`

```
print(df.describe())
```

The `describe()` function for numeric data shows:

- the availability of data (count),
- the mean,
- the standard deviation,
- min and max values,
- the percentiles,
- the median

- Count of each category in a categorical attributed series of values:  
`value_counts()`

```
print("\neducation\n", df['education'].count())  
print(df['education'].value_counts())  
print("vote:\n", df['vote'].value_counts())
```

- To filter out the NA/NaN values in a specific column

```
df['income'].dropna()
```

- Statistics per group: `groupby()`

```
df.groupby('region')['income'].mean() #average income by region  
print("\n\ngroupby:", df.groupby(['education', 'vote'])['income'].mean())
```

2. Choose a variable from the dataset and present the data (histogram, box-plot). Compute the central tendency indicators, dispersion indicators and the shape of distribution.

- Compute the mean and the variation of a var

```
print("mean of income", df['income'].mean())  
print("var of income", df['income'].var())
```

- Represent the distribution of data; use

- histograms

```
import matplotlib.pyplot as plt  
#plot income  
plt.hist(df['income'].dropna())  
plt.show()
```

```
#plot population
```

```
num_bins = 10
```

```
plt.hist(df['population'], num_bins, normed=1, facecolor='blue', alpha=0.7)  
plt.show()
```

- boxplots

```
y = list(df.population)
plt.boxplot(y)
plt.show()
```

3. Compute the covariance and correlation for two variables (age and strength, for example). Show the relationship btw them.

- Use `corr()` from `pandas.DataFrame` to create the correlation matrix and `heatmap()` from `seaborn` to create a heat map.
- Use `pearsonr()` SciPy function to compute the Pearson's correlation coefficient between two data samples with the same length.
- Use `scatter()` from `matplotlib.pyplot` for a scatter plot.

### Resources

1. *H. Seltman. Experimental Design and Analysis. Ch.4. Exploratory Data Analysis*
2. python: *Statistics and Machine Learning in Python. 3.2. Pandas: data manipulation (section 3.2.8 Descriptive Statistics) and 3.3. Matplotlib: data visualization).*

Example of exploratory data analysis for Chile dataset  
Deadline: lab 12.