Module 3. Statistics and the math behind data science

Welcome to module three of the Fundamentals of Data Science course. In this module, we will learn about the math and statistics concepts that underpin data science.

We assume no prior knowledge and will start with the basics.

Math in sports

Unit 3.1 Descriptive statistics

Descriptive statistics are fundamental tools in data science used to summarize and describe the main features of a dataset. They provide a concise way to understand the essential characteristics of data without necessarily analyzing every single data point.

a) **Mean (average):** the sum of all values divided by the number of values.

- Formula: mean $\mu = (\Sigma x) / n$, where x is the individual values and n is the number of values.

Mean or average is one of the most common descriptive statistics. We use it quite often in our daily lives. It is beneficial in summarizing customarily distributed data.

Examples:

- What is the average of 3 numbers, 4, 5, 9?
    - Average/mean = (4+5+9)/3 = 18/3 = 6
- What is the average number of goals scored per match by Manchester United in the Premier League season of 22–23?
    - Manchester United scored 58 goals in 38 matches in the season of 22-23. So their average goals per match is 58/28 = 1.53 goals per match.
- What is the average height of students in a class?
    - Sum all the students' heights in the class and divide it by the number of students: 3011/16 = 188 cm.

b) **Median:** the median is the middle value when the dataset is ordered from the least to the most.

- This metric is helpful to summarize data that is not normally distributed and is skewed.

Examples:

- What is the median of 3 numbers 4, 9, 5?

  - First order the numbers in ascending order: 4, 5, 9

    Median = middle value = 5.

    Notice how the mean (6) and median (5) differ for the same dataset?

    In the case of the dataset having an even number of data points like 4, 9, 8, 5, after ordering them in ascending order, the median is the average of the two middle numbers.

    Ordered dataset: 4, 5, 8, 9

    Average of two middle numbers = (5+8)/2 = 6.5

Here is another example: there are ten students in a class. Here are their heights in centimeters. What are the average and median heights of the class?

155, 175, 185, 153, 182, 192, 156, 177, 183, 155.

Average = (155+175+185+153+182+192+156+177+183+155)/10 = 1711/10=171.1

Ordered data points: 153,155,155,156,**175,177**,182,183,185, 192

Median = 175+177/2 = 176

In this example, the mean is less than the median, and the outliers impact the median less than the mean.

c) **Mode:** the most frequently occurring value in a dataset.

- This is very useful for categorical data.

In the above example of heights in a class

153,**155,155**,156,175,177,182,183,185, 192

Mode = 155 because 155 occurs most frequently.

As you can see, mode is not very good at describing this data type.

Mode is handy for categorical data.

For example, determine the most favorite ice cream flavor in a survey.

Data: vanilla, strawberry, vanilla, chocolate, chocolate, strawberry, vanilla, vanilla.

Mode: vanilla (occurs most frequently)

Unit 3.2 Measures of dispersion

Now, we look at ways to measure how dispersed the data is.

a) **Range**

- Range is the difference between a dataset's maximum and minimum values.
- Formula: Range = Max(x) - Min(x)

Example: what is the range of heights in the class of 10 from the example earlier?

153,155,155,156,175,177,182,183,185, **192**

Max height = 192

Min height = 153

Range = Max – Min = 192 – 153 = 39

Range = 39

b) **Variance**

- Variance is defined as the average of squared deviations from the mean.
- Variance $\sigma^2 = \Sigma(x - \mu)^2 / n$; where μ = Mean, x = individual data point.
- Variance measures the spread of data.

c) **Standard deviation**

- SD = square root of the variance.
- Formula: $\sigma = \sqrt{(\sigma^2)}$
- This is the most widely used measure of dispersion in the same units as original data.

Unit 3.3 Percentiles and quartiles

- **Percentiles**
  - Percentiles are values that divide the dataset into 100 equal parts.

    For example, the median is the 50th percentile. It divides the data into two halves.

    "The median height of the students in a class of N students is 171 cm" – This means exactly 50 % of the students are taller than 171 cm and 50 % shorter than 171 cm.
- **Quartiles**
  - Quartiles are values that divide the dataset into four equal parts.
  - Q1 (25th percentile), Q2 (median), Q3 (75th percentile).
  - Similarly, there are quintiles and so on.
- **Interquartile range (IQR)**
  - IQR is the difference between the 3rd and 1st quartiles.
  - IQR = Q3 - Q1

IQR is a robust measure of spread. It is instrumental in measuring outliers in a dataset.

Here is an example to demonstrate the application of IQR.

A dataset of 14 numbers ordered in ascending order.

Data = {1, 3, 3, 4, 8, 11, 13, 14, 15, 17, 22, 24, 26, 46}

First quartile (Q1) = 5

Second quartile (Q2 or median) = 13.5

Third quartile (Q3) = 20.75

IQR = Q3 – Q1 = 15.75

The limits are calculated as follows:

Lowe limit = Q1 – 1.5*IQR = 5 – 1.5*15.75 = -18.625
Upper limit = Q3 + 1.5*IQR = 20.75 + 1.5*15.75 = 44.375

Any data points above the upper limit or below the lower limit are the outliers.
In this example, data point **46** is an **outlier** as it is greater than the upper limit.

Unit 3.4 Z – Scores

A z-score measures how many standard deviations away a data point is from the mean of a dataset. Here's a brief explanation.

- Z-score = (X - μ) / σ

  Where: X = the raw score;

  μ = the population mean;

  σ = the population standard deviation.

- Z-scores allow you to compare values from different datasets or distributions by standardizing them.

- How to interpret Z-scores?

  Z-scores help you quantify how big of an outlier a data point is from the rest of the population/dataset.

  - A z-score of 0 means the data point is strictly at the mean.
  - Positive z-scores indicate values above the mean.
  - Negative z-scores indicate values below the mean.
  - Roughly 68 % of the data falls within one standard deviation (z-score between -1 and 1).
  - About 95 % falls within two standard deviations (z-score between -2 and 2).
  - About 99.7 % falls within three standard deviations (z-score between -3 and 3).

- Z-scores are used in statistics for standardizing datasets, identifying outliers, and hypothesis testing

  Z-scores example:

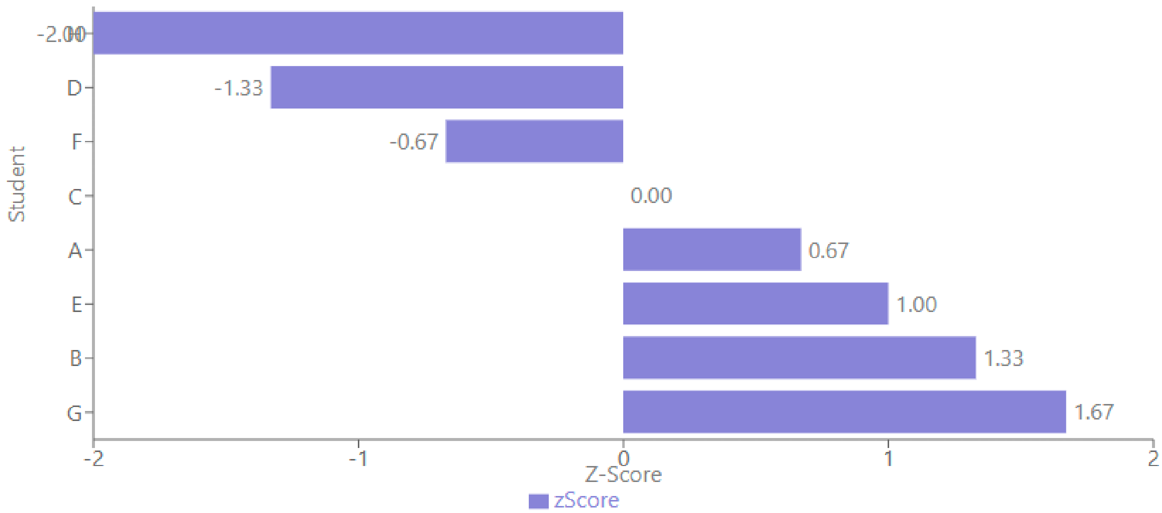
Table 1. Eight students and their scores

Student	Score	Z-score
A	85	0.67
B	92	1.33
C	78	0
D	65	-1.33
E	88	1
F	72	-0.67
G	95	1.67
H	58	-2

Source: own elaboration.

- Mean score = 79.125.
- Standard deviation SD = 12.97.
- Here are their z-scores in column 3.

Figure 1. Z.Score



Source: own elaboration.

How do we interpret this data?

This chart makes it easy to compare the relative performance of the students with the class.

- Student C is right on the average score of the class.

- Students A, B, E, and G are above the mean (positive z-scores).
- Students D, F and H are below the mean (negative z-scores).
- Student G has the highest Z-score – furthest away from the average on the positive side.
- Student H has the lowest Z-score – furthest away from the average on the negative side.

Unit 3.5 Probability

One of the most critical objectives of data science is to make predictions of future outcomes using incomplete information.

Probability is the basis of statistical inference and provides a framework for dealing with uncertainty and making predictions based on incomplete information. This section will explore key probability concepts and their applications in data science.

a) Basic probability concepts

Probability is defined as the extent to which something is possible. It is the likelihood of an occurrence of an event (probability of a dice landing on 2).

Sample space(S): the set of all possible outcomes of an experiment.

Event: a subset of the sample space.

The probability of the occurrence of an event A is P(A).

$P(A) >= 0$ for any event A

$P(S) = 1$

For two events, A and B.

Probability of the occurrence of either A or B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

If A & B are mutually exclusive, then probability of the occurrence of either A or B:

$P(A \cup B) = P(A) + P(B)$

If A & B are independent events, then probability of the occurrence of A and B:

$P(A \cap B) = P(A) * P(B)$

Example: let's use a standard 6-faced die.

Sample space of all outcomes S = {1,2,3,4,5,6}.

The probability that a standard die turns up 1 P(1) = 1/6 = 16.67 %.

The probability that the die turns an even number P(Even) = 3/6 = 50 %.

The probability that the die turns an odd number P(Odd) = 3/6 = 50 %.

P(Even) & P(Odd) are mutually exclusive because either the die will turn up odd (1 or 3 or 5) or even (2 or 4 or 6).

So P(Even or Odd) = P(Even) + P(Odd) = 100 %.

Let us use a different example to demonstrate independent events.

Toss a coin with two faces, head & tails.

The probability of P(Heads) = 50 % and P(Tails) = 50 %.

What is the probability of 2 heads when you toss the coin twice?

Head or tails on the second toss is independent of the outcome of the first toss.

P(H1) = ½ = 50 %

P(H2) = ½ = 50 %

P(H1 AND H2) = ½ * ½ = 25 %

Conditional probability

Conditional probability is the probability of an event occurring, given that another event has already happened. It is denoted as P(A|B), "the probability of A given B".

Mathematically, conditional probability is expressed as $P(A|B) = P(A \cap B) / P(B)$

Where:

- P(A|B) is the conditional probability of A given B.
- $P(A \cap B)$ is the probability of both A and B occurring.
- P(B) is the probability of B occurring.

Based on new information, conditional probability helps us update our beliefs about an event's likelihood. It's used when we want to calculate the probability of an event, taking into account the occurrence of a related event.

Example: flipping a coin twice.

Flip a fair coin twice in a row. Given that we know at least one of the flips was heads, we want to calculate the probability that the first flip was heads.

Let's define our events.

A: the first flip is heads.

B: at least one flip is heads.

We want to calculate P(A|B), the probability that the first flip was heads, given that at least one flip was heads.

Step 1: list all possible outcomes HH, HT, TH, TT (where H is heads and T is tails).

Step 2: calculate P(B), the probability of at least one heads P(B) = 3/4 (HH, HT, TH are acceptable outcomes out of 4 total outcomes).

Step 3: Calculate P(A ∩ B), the probability of the first flip being heads AND at least one heads P(A ∩ B) = 2/4 = 1/2 (HH and HT are favorable outcomes)

Step 4: apply the conditional probability formula P(A|B) = P(A ∩ B) / P(B) = (1/2) / (3/4) = 2/3 ≈ 0.6667 or about 66.67 %.

How do we interpret the result?

Given that, we know at least one of the flips was heads, the probability that the first flip was heads is 2/3, or about 66.67 %.

This is higher than the unconditional probability of getting heads on a single flip (50 %).

Why? Knowing that at least one flip was headed eliminates the TT outcome, and of the remaining outcomes (HH, HT, TH), two out of three start with heads.

Key takeaways:

1. Conditional probability can change our assessment of an event's likelihood based on new information.
2. In this case, knowing that at least one flip was headed increased the probability that the first flip was heads.
3. The conditional probability (66.67 %) is different from both the probability of a single head (50 %) and the probability of at least one head in two flips (75 %).

Conditional probability has a lot of applications in data science.

Bayes theorem, defined as follows, has huge applications in data science.

Bayes' theorem

P(A|B) = P(B|A) · P(A) / P(B)

Where:

- P(A|B) is the posterior probability.
- P(B|A) is the likelihood.
- P(A) is the prior probability.
- P(B) is the marginal likelihood.

Many machine learning algorithms use conditional probability concepts.

- Naive Bayes classifiers: these use Bayes' theorem to predict the probability that a given data point belongs to a particular class.
- Decision trees and random forests: these algorithms often use conditional probability to make splitting decisions.
- Hidden Markov models: these are used in sequential data analysis and rely on conditional probabilities between hidden states and observed data.
- Applications in natural language processing.
    - Language models that predict the probability of a word given the previous words. Examples: Chat GPT, Claude, Google Gemini.
    - Part-of-speech tagging: determining a word's most likely grammatical category given its context.
    - Machine translation: estimating the probability of a translation given the original text.
- Anomaly detection, causal inference, and time series analysis are some other applications of conditional probability and the Bayes' theorem.

Unit 3.6 Inferential statistics

Inferential statistics is a branch that uses sample data to make predictions or inferences about a larger population.

Inferential statistics provides data scientists with powerful tools to make decisions and draw conclusions under uncertainty. Understanding these concepts is crucial for conducting rigorous analyses and building reliable machine-learning models.

Terms and definitions in inferential statistics

- **Population:** the entire dataset.
- **Sample:** a subset of the population selected for the study or experiment.
- **Parameter:** a numerical characteristic of a population (e.g., the population mean).
- **Statistic:** a numerical characteristic of a sample (e.g., the sample mean).

- **Central limit theorem:** distribution of sample means approaches normal as sample size increases.
- **Standard error:** standard deviation of a sampling distribution.
- **Null hypothesis (H0):** a statement of no effect or difference, assumed to be true until evidence suggests otherwise.

  Example: "Home teams do not have any advantage over away teams in terms of winning probability".
- **Alternative hypothesis (Ha):** a statement that contradicts the null hypothesis.

  Example: "Home teams have a distinct advantage over away teams in terms of winning probability".
- **P-value:** the probability of obtaining results at least as extreme as the observed results, assuming the null hypothesis is true.
- **Confidence interval:** a range of values likely to contain the proper population parameter with a certain confidence level.
- **Confidence level:** the probability that the confidence interval contains the proper population parameter.
- **Type I error:** rejecting the null hypothesis when it is true (false positive).
- **Type II error:** failing to reject the null hypothesis when it is false (false negative).

- **Hypothesis testing:** using sample data to test a claim about a population parameter, such as whether a new drug is more effective than a placebo.

  A pharmaceutical company claims their new drug reduces cholesterol levels by 30 % on average. To test this claim we have the following.

  Null hypothesis: the drug reduces cholesterol by 30 %.

  Alternative hypothesis: the drug does not reduce cholesterol by 30 %.

  A sample of patients would take the drug, and their cholesterol levels would be measured before and after treatment.

  Researchers would use statistical tests (like a t-test) to determine whether the sample results significantly differ from the claimed 30 % reduction.

Based on the p-value, they would either reject or fail to reject the null hypothesis.

- **Estimating confidence intervals:** to estimate the average height of adults in a country, researchers might measure the height of 1000 randomly selected adults.

  They could calculate the mean height of this sample, let's say it's 170 cm.

  They would also calculate a margin of error, say ±2 cm.

The 95 % confidence interval would be 168-172 cm.

Researchers are 95 % confident that the population's mean height falls within this range.

- **Analysis of variance (ANOVA):** this method compares means across multiple groups to determine if there are significant differences. It could be used to compare the effectiveness of different teaching methods.

Divide students into three groups, each taught using a different method.

Measure student performance after the teaching period.

Use ANOVA to determine if there are statistically significant differences in mean performance between the groups.

If significant differences are found, post hoc tests can identify which specific groups differ from each other.

- **Chi-square tests:** analyze the relationship between categorical variables and examine whether there's a significant association between gender and voting preferences.

Collect data on the gender and voting choice of a sample of voters.

Create a contingency table showing the count of males/females voting for each party.

Perform a chi-square test of independence. This test compares the observed frequencies to the expected frequencies if there were no association.

A low p-value would suggest a significant association between gender and voting preference.

4. Regression

Regression analysis is a fundamental data science technique that models the relationship between variables. It's essential for prediction, forecasting, and understanding the impact of different factors on an outcome of interest. This section explores the basics of regression techniques, their assumptions, implementations, and applications in data science.

- **Simple linear regression**

Simple linear regression models the relationship between two variables using a straight line.

Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

- y is the dependent variable;
- x is the independent variable;
- β_0 is the y-intercept;
- β_1 is the slope;
- ε is the error term.

Assumptions

1. Linearity: the relationship between x and y is linear.
2. Independence: observations are independent of each other.
3. Homoscedasticity: constant variance of residuals.
4. Normality: residuals are normally distributed.

- **Multiple linear regression**

Extends simple linear regression to include multiple independent variables.

Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

Assumptions

a) Same as simple linear regression, plus.
b) No multicollinearity: independent variables are not highly correlated with each other.

- **Polynomial regression**

Model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

- Adding polynomial terms to model nonlinear relationships.
- Overfitting concerns.

- **Logistic regression**
  - For binary outcomes. It is widely used for binary classification problems.
  - Logit function: $P(y=1|x) = 1 / (1 + e^{(-z)})$ where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

Unit 3.7 Linear regression analysis example

The following is a simple linear regression analysis example, predicting future sales based on advertising expenditure.

Collect data on past sales and corresponding advertising budgets.

Plot these data points and fit a line (or curve) that best represents the relationship.

Use the equation of this line to predict future sales for a given advertising budget.

For example, the equation might be: sales = 100,000 + 5 * advertising budget.

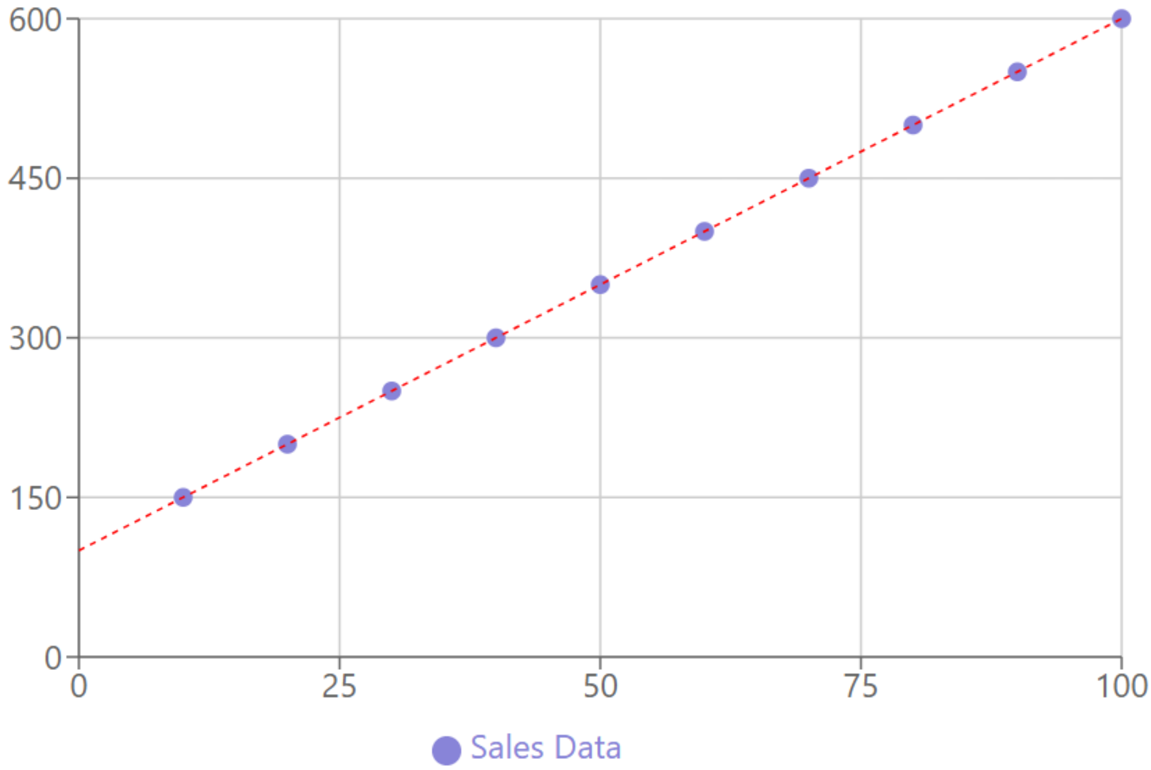
This suggests that for every dollar spent on advertising, sales increase by $5.

Table 2. Advertising and sales data

Advertising	10	20	30	40	50	60	70	80	90	100
Sales	150	200	250	300	350	400	450	500	550	600

Source: own elaboration.

Figure 2. Advertising budget vs. sales graph with the regression line



Source: own elaboration.

- X-axis: represents the advertising budget in thousands of dollars.
- Y-axis: represents the sales in thousands of dollars.
- Blue dots: each dot represents a data point, showing the sales for a given advertising budget.
- Red dashed line: the regression line represents the linear relationship between advertising and sales.

The regression line follows the equation we mentioned earlier: sales = 100,000 + 5 * advertising budget. In the graph, this translates to:

- Y-intercept at 100 (when advertising is 0, sales are $100,000).
- For every increase of 1 unit on the x-axis (1,000 in advertising), the y-axis increases by five units (5,000 in sales).
- This visualization helps to see the positive correlation between advertising budget and sales. As the advertising budget increases, sales tend to grow.

- **Metrics to evaluate a regression model**

  Regression metrics are crucial for evaluating the performance of regression models. They help us understand how well our model fits the data and make comparisons between different models. Here's a detailed look at the most common regression metrics:

1. **Mean squared error (MSE)**
   Formula
   $MSE = (1/n) * \Sigma(yi - \hat{y}i)^2$
   where:
   - n is the number of observations
   - yi is the actual value
   - ŷi is the predicted value

Interpretation
MSE measures the average squared difference between the predicted and actual values. It penalizes more significant errors more heavily than smaller ones. Lower MSE indicates better model performance.

2. **Root mean squared error (RMSE)**

   Formula
   $RMSE = \sqrt{MSE} = \sqrt{((1/n) * \Sigma(yi - \hat{y}i)^2)}$

   Interpretation
   RMSE is the square root of MSE. It has the same units as the dependent variable, making it easier to interpret. Like MSE, lower values indicate a better fit.

3. **Mean absolute error (MAE)**

   Formula
   $MAE = (1/n) * \Sigma|yi - \hat{y}i|$

   Interpretation
   MAE measures the average absolute difference between predicted and actual values. It's less sensitive to outliers than MSE/RMSE. Lower values indicate a better fit.

4. **R-squared (coefficient of determination)**

   Formula

   $R^2 = 1 - (SSres / SStot)$

where:

- SSres = $\Sigma(yi - \hat{y}i)^2$ (sum of squares of residuals)
- SStot = $\Sigma(yi - \bar{y})^2$ (total sum of squares)
- $\bar{y}$ is the mean of observed data

Interpretation

R^2 represents the proportion of variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1, with 1 indicating perfect fit. However, it can be misleading for non-linear relationships and doesn't account for model complexity.

5. **Adjusted R-squared**

Formula

$$\text{Adjusted } R^2 = 1 - [(1 - R^2)(n - 1) / (n - k - 1)]$$

where:

- n is the number of observations
- k is the number of predictors

Interpretation

Adjusted R^2 modifies R^2 to account for the number of predictors in the model. It penalizes the addition of unnecessary variables. It can be damaging and is always lower than R^2. It's useful for comparing models with different numbers of predictors.

6. **Mean absolute percentage error (MAPE)**

Formula

$$\text{MAPE} = (100/n) * \Sigma|(yi - \hat{y}i) / yi|$$

Interpretation

MAPE expresses accuracy as a percentage of the error. It's scale-independent, which makes it easy to compare across different datasets. However, it can be inflated by small actual values and can't be used when valid values are zero.

- **Choosing the right metric to evaluate your model**

The choice of metric depends on your specific problem and data:

- Use RMSE when significant errors are particularly undesirable.

- Use MAE when you want to treat all errors equally.
- Use R^2 or Adjusted R^2 when you want to know how well your model explains the variance in your data.
- Use MAPE when you want a percentage error that is easily interpreted across different scales.

Considering multiple metrics to get a comprehensive view of your model's performance is often beneficial.

Unit 3.7 Data plots and visualizations

As a data scientist, getting into the habit of plotting the data is very handy. Plotting data helps in data analysis, communication, and decision-making.

- Data plots simplify complex datasets into easily understandable visual formats.
  - This helps identify trends and patterns in the data.
  - Spot outliers and anomalies.
  - Understand the relationship between variables.

- Interpretation
  - Helps in quickly interpreting and speeds up data comprehension.

- Communication
  - Plots and graphs can convey information more clearly and persuasively to audiences with diverse backgrounds.

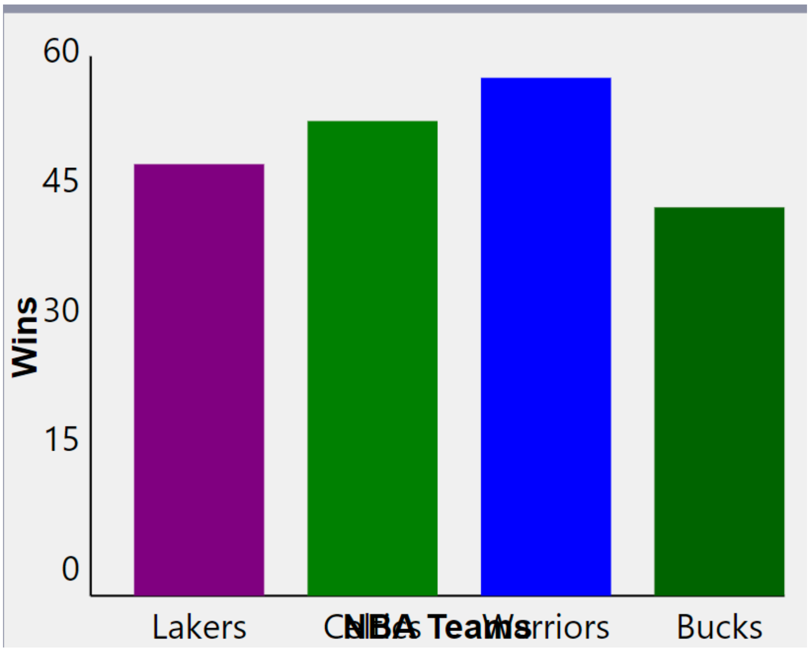
- Pattern recognition
  - Data visualizations can reveal hidden patterns that are not obvious when looking at a table of numbers.
  - Cyclical trends in time-series data.
  - Clusters in scatter plots.

- Data quality
  - Help assess data quality by revealing missing data and mal-formatted data.
  - New questions and ideas
  - Visualizing data in the appropriate format can help raise new questions and hypotheses for further investigation.

1. **Bar charts**
    - Show comparisons between categories.
    - Good for displaying discrete, categorical data.

Example: NBA # of wins by team.

Figure 3. Bar chart of NBA # of wins by team

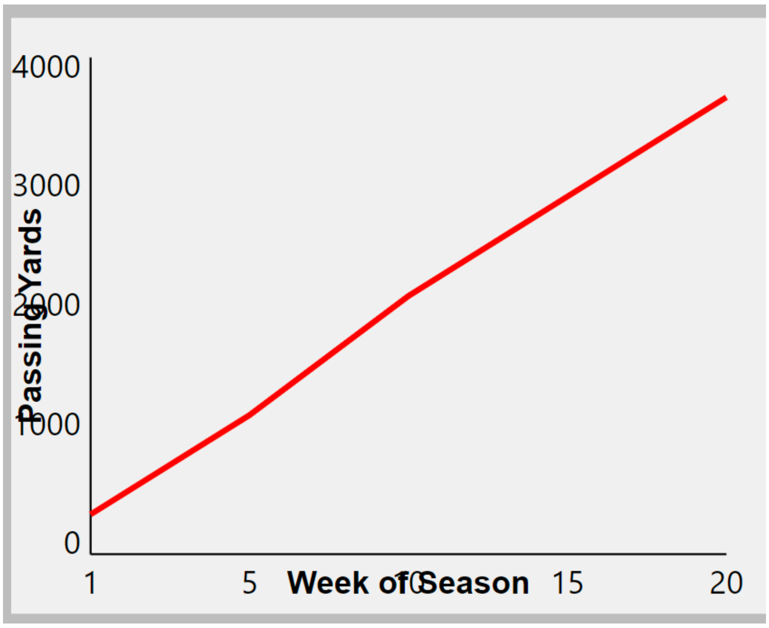


Source: own elaboration.

2. **Line graphs**
    - Display trends over time or continuous data
    - Ideal for showing changes and patterns

Example: NFL Passing yards by week of the season.

Figure 4. Line graph of NFL Passing yards by week of the season



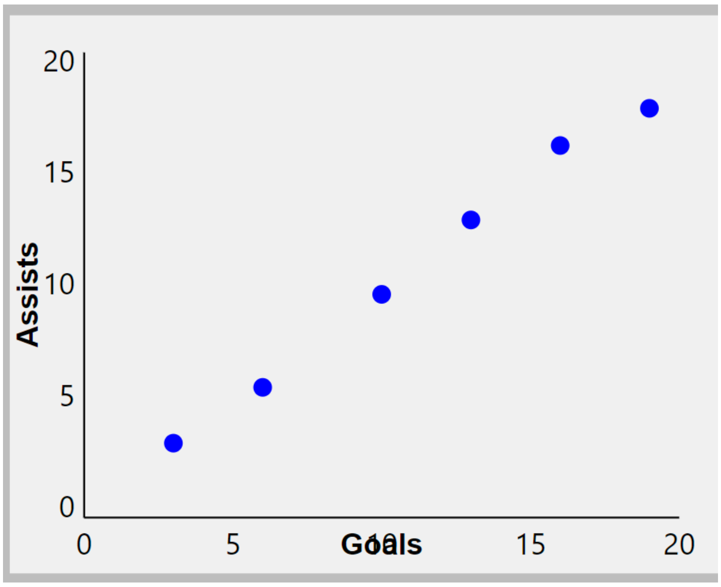
Source: own elaboration.

3. Scatter plots

- Show relationships between two variables.
- Useful for identifying correlations or clusters.
- Applications: height vs. weight, advertising spend vs. sales.

Example: goals and assists in soccer.

Figure 5. Scatter plot of goals and assists in soccer

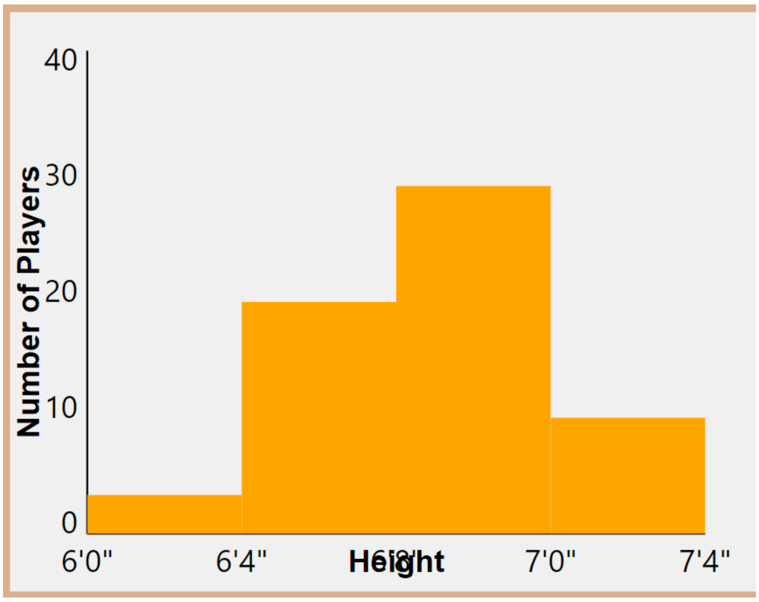


Source: own elaboration.

4. **Histograms**
    - Show relationships between two variables.
    - Useful for identifying correlations or clusters.
    - Applications: height vs. weight, advertising spend vs. sales.

Example: histogram of the heights of NBA players.

Figure 6. Histogram of the heights of NBA players

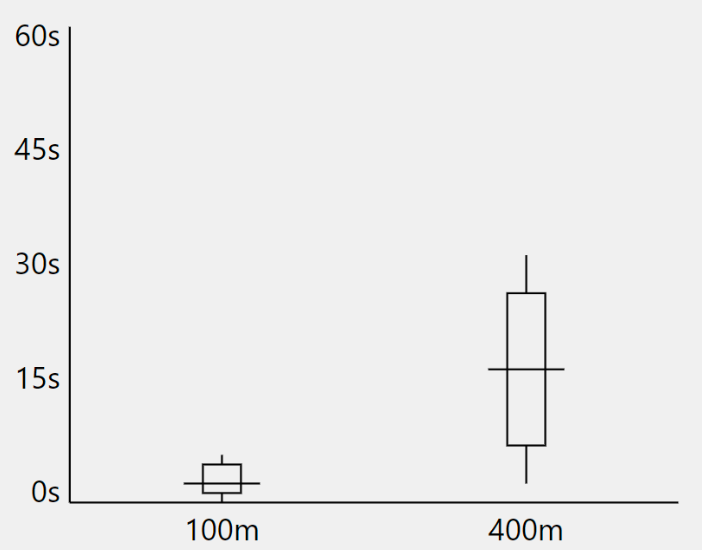


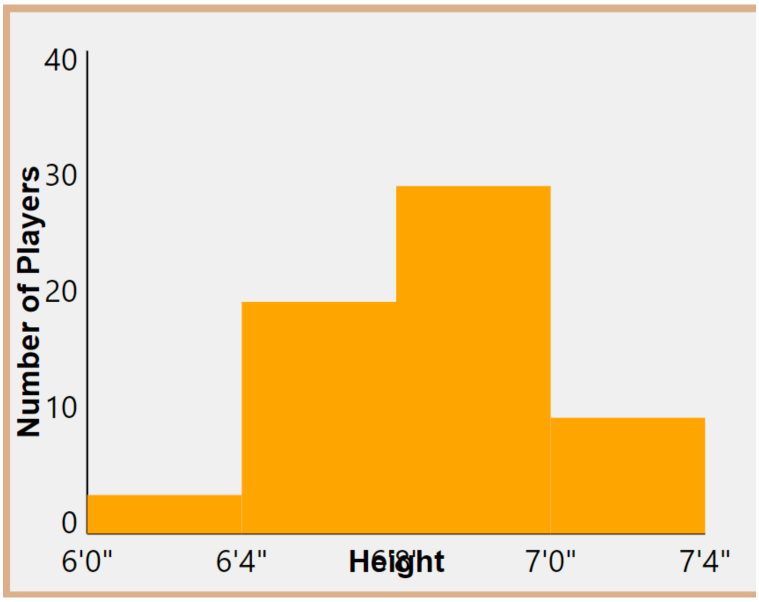
Source: own elaboration.

5. **Box plots**
    - Summarize data distribution.
    - Show median, quartiles, and potential outliers.
    - Applications: comparing distributions, identifying data spread.

Example: Running times of 100m & 400m races.

Figure 7. Box plot of running times of 100m & 400m races





Source: own elaboration.

Suggested bibliography

Casella, G., & Berger, R. W. (2024). *Statistical inference*. CRC Press.

Descriptive statistics definition of psychology. https://pdfprof.com/PDF_DOC/PDF_Documents/331542/4/101/descriptive+statistics+definition+psychology

Gelman, A. (2013). *Bayesian Data Analysis: Texts in Statistical Science*. CRC Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of Statistical Learning: Data Mining, Inference, and prediction*. Springer-Verlag New York.

Larson, M. G. (2006). Descriptive Statistics and Graphical Displays. Circulation. https://doi.org/10.1161/circulationaha.105.584474

Sachkov, I. N., Turygina, V. F., & Ford, V. (2019). Development of FEM programs for assessing the risk of electrical breakdown of devices operating in high humidity conditions. https://doi.org/10.1063/1.5133571

Sathianandan, T. V., Safeena, P. K., & Rahman, M. R. (2017). Basic Statistics. https://core.ac.uk/download/79425841.pdf