Transform

Prof. MSc. Fernando Prass

fprass@gmail.com

github.com/fernandoprass/etl

Who am I?

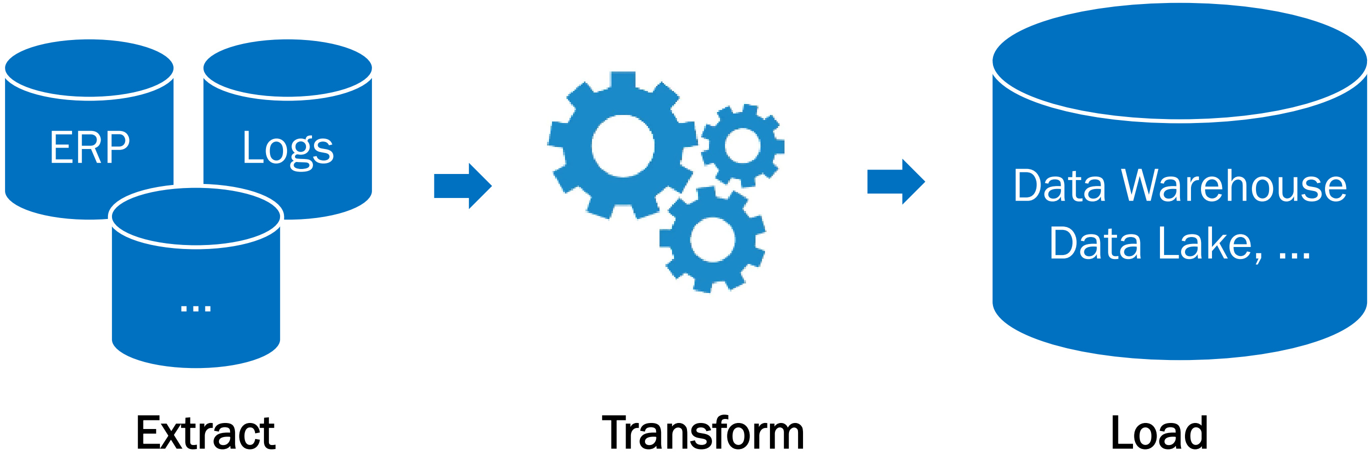
- Fernando Sarturi Prass
- Bachelor in Information Systems
- Master in Computer Science
- Senior Software Engineer (tech leader) at Workday

- fprass@gmail.com
- github.com/fernandoprass
- twitter.com/oFernandoPrass

- linkedin.com/in/fernandoprass
- lattes.cnpq.br/2919187023046130

Extract, TRANSFORM, Load (ETL)

- Extract, **TRANSFORM**, Load (ETL) *“is a data integration process that combines, cleans and organizes data from multiple sources into a single, consistent data set for storage in a data warehouse, data lake or other target system”.* (IBM)



Transform

- In the staging area the data is transformed for its intended use.

  Transformed can mean:

  - Filtering, cleaning, aggregating, de-duplicating, and validating;
  - Performing calculations, translations, summarizations or conversions (units of measurement, currency, ...);
  - Conducting audits to ensure data quality and compliance, and computing metrics.
  - Removing, encrypting or protecting private data (GDPR);
  - Formatting the data into tables or joined tables to match the target schema, change/update titles and description.

Agenda

- Review
  - Types of Variables
  - Mean vs Moda vs Median
- Missing Values and Outliers
- Encoding Categorical Variables
- Distance Measurements

Review

- Types of Variables
- Mean vs Moda vs Median

Types of Variables

A variable is defined as an attribute of an object of study, can be classified
as categorical or quantitative.

You ***MUST*** *know what types of variables you are working with to choose the appropriate form of transformation, the appropriate ML algorithm, and correctly interpret the results of your study.*

Types of Variables

Remember: *You MUST know what types of variables you are working on.*

- **Quantitative Variables** have numerical values with consistent intervals (height, weight, distance, …).
- **Categorical Variables** are those that provide groupings that may have no logical order, or a logical order with inconsistent differences between groups (e.g., the difference between 1st place and 2 second place in a race is not equivalent to the difference between 3rd place and 4th place).

Quantitative Variables

- Quantitative Variable consists of numerical insights that represent quantities or measurements, they are numerical values.
- Examples include:
  - Population size of a city
  - Number of different tree species in a forest
  - Number of students in a class
  - Height of an individual
  - Number of rooms in a house

Categorical Variables

- Categorical variable, also known as *qualitative* or *nominal* data, comprises information that can be grouped into distinct categories or labels. They are variables that take on names or labels. Examples include:
  - Marital status (single, married, divorced)
  - Smoking status (smoker, non-smoker)
  - Eye color (blue, black, white)
  - Level of education (high school, bachelor's degree, doctoral)

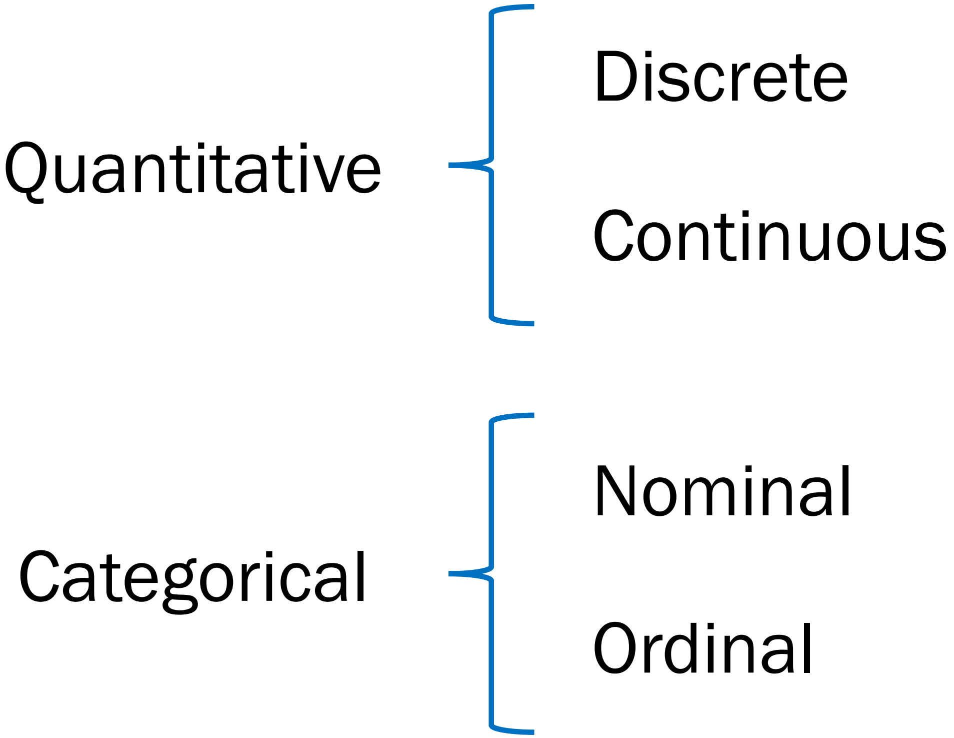
Categorical Variables

- It conveys descriptive details about the qualitative attributes or characteristics of an object, individual, or event.
- Categorical data is non-numeric and often expressed using words or labels.
- Since it does not provide numerical numbers, it is not amenable to mathematical operations or calculations. They are analysed using frequencies, proportions, or percentages.

Quantitative vs Categorical Variables

	Quantitative Data	Categorical Data
Type of Values	Numerical values	Labels or categories
Nature	Quantitative	Qualitative
Mathematical Operations	Extensive (mathematical calculations, statistical analysis)	Limited (frequencies, proportions)
Level of Measurement	Interval or ratio	Nominal or ordinal
Examples	Height, weight, …	Genders, Countries, …

Variables - Classification



Quantitative

- Discrete
- Continuous

Categorical

- Nominal
- Ordinal

Quantitative Discrete

- These variables can only assume specific, distinct values that you cannot subdivide. Usually, you count them, and the results are integers.
- Examples:
  - The number of birds in a cage (e.g., 10 birds).
  - The number of books on the shelf.
  - The population of a country (whole numbers).

Quantitative Continuous

- These variables can take on any numeric value within a defined range, can be meaningfully split into smaller parts, including fractional and decimal values.
- Examples:
  - Distance (measured kilometer or miles).
  - Time (measured in minutes or seconds).
  - Temperature (measured in degrees Celsius or Fahrenheit).

Categorical Nominal

- These variables represent categories without any intrinsic ranking, that is, in no defined order.
- Examples:
  - Colors: Red, Blue, Green
  - Types of Cuisine: Italian, Chinese, Mexican
  - Religions: Christianity, Islam, Buddhism

Categorical Ordinal

- These variables represent categories with a clear order or ranking.
- Examples:
  - Education Level: High School < Bachelor’s < Master’s < PhD
  - Customer Satisfaction: Unsatisfied < Neutral < Satisfied < Very Satisfied
  - Military Ranks: Soldier < Sergeant < Lieutenant < Captain

Binary Variables

- A binary variable can only take one of two values. It is usually represented as Boolean (True or False), String (Yes or No) or Integer (0 or 1, where 0 typically indicates the absence of the attribute, and 1 indicates its presence).
- Examples:
  - Gender: Male or Female
  - Lamp status: on or off
  - Test result: Pass or Fail

Question

- Is binary variable a different type or can it be classified as a nominal or ordinal categorical variable?

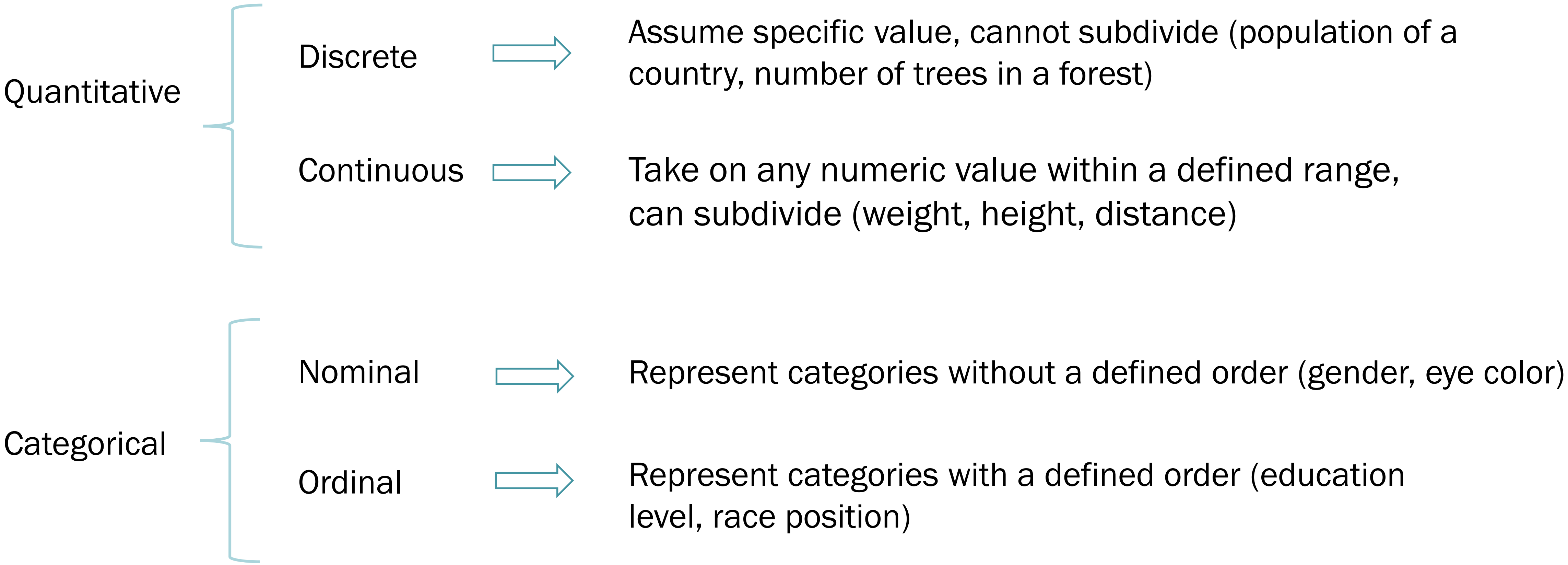
Symmetric Binary vs Asymmetric Binary

- A **symmetric binary** variable has equivalent states, that is, both have the same weight. There is not difference between the characteristic being present (1) or absent (0).
- An **asymmetric binary** variable does not have equivalent weights for the two states. There is a difference between the characteristic being present or absent.
  - This type of variable is commonly used to represent the presence or absence of a certain disease. By convention, the value 1 is used for the most important result, in this case, that the patient has the

So ...

Binary variables can be considered nominal or ordinal, depending on the context

Variables Classification - Resume



Quantitative

- Discrete → Assume specific value, cannot subdivide (population of a country, number of trees in a forest)
- Continuous → Take on any numeric value within a defined range, can subdivide (weight, height, distance)

Categorical

- Nominal → Represent categories without a defined order (gender, eye color)
- Ordinal → Represent categories with a defined order (education level, race position)

Graphic Representation

- **Discrete** variable are often use bar charts to represent discrete data, while **Continuous** use **h**istograms and scatterplots are commonly used for continuous data.
- **Nominal variables** are often visualized using pie charts or bar graphs, while **Ordinal** variables can be represented with bar graphs that reflect the inherent order of the categories.

Review

- Mean
- Moda
- Median

Mean

- The **mean**, or average, is calculated by adding up all the numbers in a dataset and then dividing by the count of numbers.
- Dataset (n =10)
  - [6, 2, 7, 3, 4, 5, 1, 2, 8, 9]
- Sum the elements
  - 6 + 2 + 7 + 3 + 4 + 5 + 1 + 2 + 8 + 9 = 47
- Divide the sum by the count of numbers: 47 / 10 = 4.7

Moda

- The **mode** is the value that appears most frequently in a data set. A set of data may have one mode, more than one mode, or no mode at all.
- Dataset (n = 10): [6, 2, 7, 3, 4, 5, 1, 2, 8, 9]
- In this dataset, the number **2** appears twice, more frequently than any other number, so the mode of this dataset is 2.

Median

- The **median** is the value separating the higher half from the lower half of a data sample. For a dataset with an odd number of observations, it is the middle number when all observations are arranged in ascending order. For an even number of observations, it is the average of the two middle numbers.

Median – Even number of elements

- Dataset (n = 10):
  - [6, 2, 7, 3, 4, 5, 1, 2, 8, 9]
- When we arrange them in ascending order:
  - [1, 2, 2, 3, **4**, **5**, 6, 7, 8, 9]
- The median would be the average of the two middle numbers:
  - (4 + 5) / 2 = 4.5
- So, the median of this dataset is **4.5**.

Median – Odd number of elements

- Dataset (n = 11):
  - [2, 8, 4, 9, 1, 5, 3, 8, 6, 0, 2]
- When we arrange them in ascending order:
  - [0, 1, 2, 2, 3, **4**, 5, 6, 8, 8, 9]
- So, the median of this dataset is **4**.

Mean vs Moda vs Median

- Dataset (n = 10): [6, 2, 7, 3, 4, 5, 1, 2, 8, 9]
- Mean (Average):
  - 6 + 2 + 7 + 3 + 4 + 5 + 1 + 2 + 8 + 9 = 47 => 47 / 10 = 4.7
- Mode (Most frequent number):
  - [6, **2**, 7, 3, 4, 5, 1, 2, 8, 9] => 2 (it appears twice)
- Median (Middle value when data is ordered):
  - [1, 2, 2, 3, **4**, **5**, 6, 7, 8, 9] => 4 and 5 =>= (4+5)/2 = 4.5

Mean vs Moda vs Median in Excel

- English version:
  - Mean: =AVERAGE(range)
  - Mode: =MODE.SNGL(range) or =MODE.MULT(range)
  - Median: =MEDIAN(range)
- Portuguese Version:
  - Mean: =MÉDIA(intervalo)
  - Mode: =MODA.ÚNICO(intervalo) or =MODA.MULT(intervalo)
  - Median: =MED(intervalo)

Mean vs Moda vs Median in Python

```
from statistics import mean, median, mode
data = [6, 2, 7, 3, 4, 5, 1, 2, 8, 9]
mean_value = mean(data)
median_value = median(data)
mode_value = mode(data)
# Print the results
print("Mean:", mean_value)
print("Median:", median_value)
print("Mode:", mode_value)
```

Note: if there are multiple modes in the dataset and you’re using Python’s statistics module’s *mode* function, it will raise an error. In such cases, you might want to use *multimode* function from the same module to get all modes.

Missing Values

Missing values occur when no data value is stored for a variable in an observation. They can happen for various reasons, such as data corruption or failure to record the data.

How to minimize the impact

- **Deletion:** Remove records with missing values if they are not significant.
- **Imputation:** Fill in missing values with estimates, such as the mean, median, or mode of the column.
- **Prediction Models:** Use algorithms like k-nearest neighbors or regression models to predict and fill missing values.

How to minimize the impact

- **Multiple Imputation:** Use statistical models to estimate multiple values for each missing entry.
- **Maximize Data Collection:** Improve the data collection process to prevent missing data.

Outliers

Outliers are data points that differ significantly from other observations in a dataset. They can arise due to variability in the measurement or may indicate experimental errors.

Outliers - How to minimize the impact

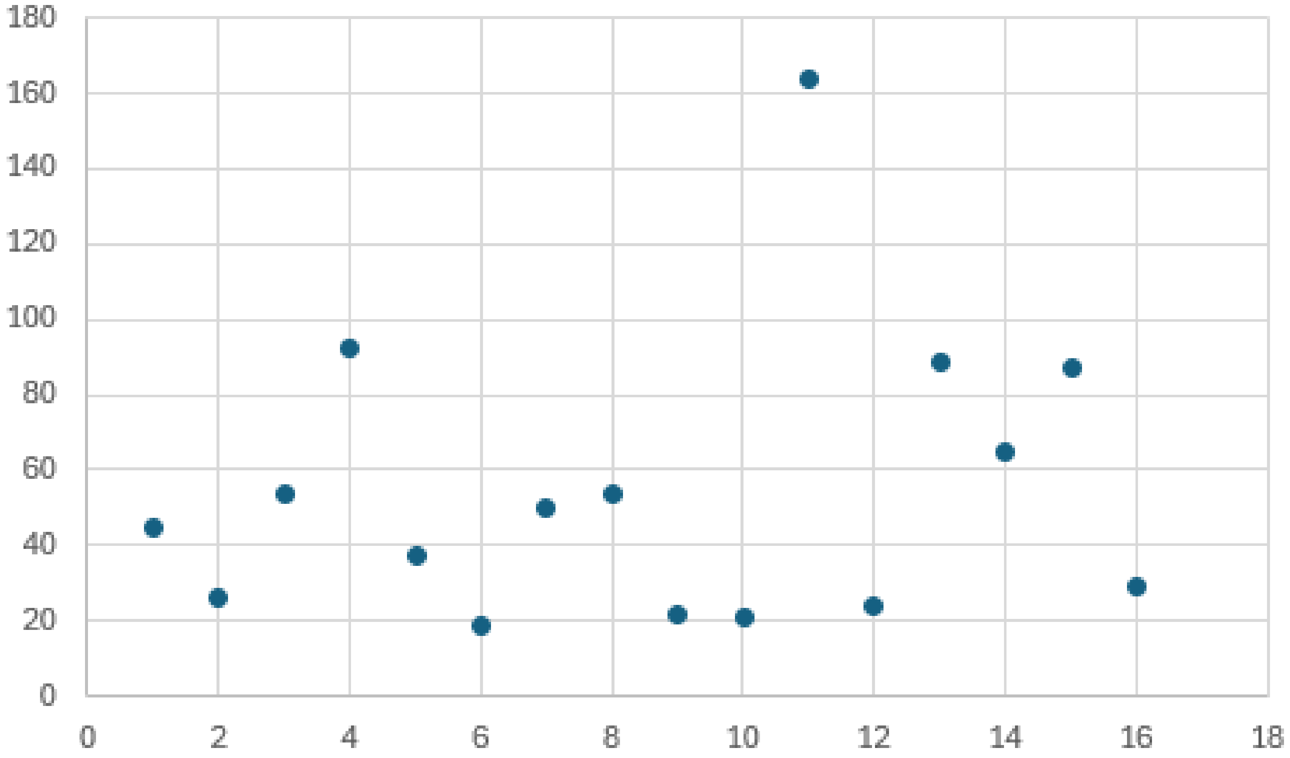
- **Data Cleaning:** Check for data entry errors or misclassifications.
- **Robust Statistical Measures:** Use median and interquartile range instead of mean and standard deviation.
- **Data Transformation:** Apply transformations like log or square root to reduce the effect of extreme values.
- **Outlier Detection Methods:** Use statistical tests like Z-score or Grubbs' test to detect outliers.
- **Filtering:** Set thresholds based on domain knowledge to filter out improbable data points.

How to know if you have an Outliers?

- Graphical Analysis
- Mathematic Analysis
  - The Interquartile Range (IQR)
  - The z-score

Graphical Analysis

- Dataset :[45, 26, 54, 92, 37, 19, 50, 54, 22, 21, 164, 24, 89, 65, 87, 29]



The Interquartile Range (IQR)

- The interquartile range (IQR) is the difference between the 75th percentile (Q3) and the 25th percentile (Q1) in a dataset. It measures the spread of the middle 50% of values.
- An element is an outlier if it is 1.5 times the interquartile range greater than the third quartile (Q3) or 1.5 times the interquartile range less than the first quartile (Q1).

IQR Explained

- Calculate the Q3 and Q1
- IQR = Q3 – Q1
- If Element < Q1 – (1,5 * IQR) or Element > Q3 + (1,5 * IQR) then the element is an Outlier

IQR in Python

```
import numpy as np

data = np.array([45,26,54,92,37,19,50,54,22,21,164,24,89,65,87,29])

q3, q1 = np.percentile(data, [75 ,25])

iqr = q3 - q1

print(iqr)
```

Z-score

- A z-score tells you how many standard deviations a given value is from the mean. The calculation formula is **z** = (X – μ) / σ. Where:
  - X is a single raw data value
  - μ is the mean of the dataset
  - σ is the standard deviation of the dataset
- Any element that has a z-score less than -3 or greater than 3 is an Outlier (sometimes a z-score of 2.5 is used instead of 3)

Z-score in Python

```
import numpy as np

data = np.array([45,26,54,92,37,19,50,54,22,21,164,24,89,65,87,29])

z_scores = (data - np.mean(data)) / np.std(data)

print(z_scores)
```

Encoding Categorical Variables

- In the realm of machine learning, encoding qualitative variables is crucial for several reasons:
  - Numerical Representation: ML algorithms typically work with numerical data. Encoding categorical variables translate them into a format that models can understand.
  - Preserving Meaning: Encoding transform categories into numbers while preserving their meaning. E.g., converting “red,” “green,” and “blue” into numeric codes (0, 1, 2) maintains the color context.
  - Enhancing Model Performance: Proper encoding ensures that models learn effectively from categorical features

Label Encoding

- Label encoding, also know as ordinal encoding, assigns each unique category value a different integer. For example:
  - Red -> 0
  - Green -> 1
  - Blue -> 2

Label Encoding – Pros vs Cons

- Pros
  - Simple to implement
  - Results in low cardinality
- Cons
  - Assumes ordinal relationship between categories (order matters)
  - Can distort linear models

Label Encoding – In Python

```
colors = ['Red', 'Green', 'Blue']

# Define a dictionary to map colors to numeric labels
color_mapping = {'Red': 0, 'Green': 1, 'Blue': 2}

# Encode the colors
encoded_colors = [color_mapping[color] for color in colors]

for color, encoded_label in zip(colors, encoded_colors):
    print(f"{color} → {encoded_label}")
```

One-Hot Encoding

- One-hot encoding creates new binary columns indicating the presence/absence of each category value.

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

In general, **one-hot encoding is preferable for categorical data** in most cases since it avoids assuming ordinal relationships.

One-Hot Encoding – Pros vs Cons

- Pros:
  - Accounts for no natural ordering between categories
  - Works for both linear and nonlinear models
- Cons:
  - Generates more sparse data
  - Increases number of features

One-Hot Encoding in Python

```python
colors = ['Red', 'Green', 'Blue']

# Create a dictionary to map each color to an index
color_index = {color: idx for idx, color in enumerate(colors)}
one_hot_encoded = []
for color in colors:
    # Create a binary list for each color
    encoded = [0] * len(colors)
    # Set the index of the color to 1
    encoded[color_index[color]] = 1
    one_hot_encoded.append(encoded)

# Display the one-hot encoded colors
for color, encoded in zip(colors, one_hot_encoded):
    print(f"{color}: {encoded}")
```

Multicollinearity

- Certainly! Multicollinearity occurs when independent variables in a regression model are correlated.

- This correlation poses a problem because independent variables should ideally be independent of each other. When the degree of correlation between variables is high, it can cause issues during model fitting and result interpretation.

Why Is Multicollinearity a Problem?

- In regression analysis, we aim to isolate the relationship between each independent variable and the dependent variable.

- The interpretation of a regression coefficient assumes that you can change one independent variable while holding others constant.

- When variables are highly correlated, changes in one variable tend to be associated with shifts in another. This makes it challenging to estimate the relationship between each independent variable and the dependent variable independently.

Distance Measurements

- Euclidean Distance
- Jaccard Coefficient

Jaccard Coefficient

- The **Jaccard Coefficient**, also known as the **Jaccard Similarity Index**, measures the similarity between two sets of data. It quantifies the proportion of shared elements between the sets. The index ranges from 0 to 1, where 1 indicates complete similarity (both sets are identical) and 0 indicates no common elements.
- Formula: $J(A, B) = \frac{\text{Number elements in the Intersection}}{\text{Number elements in the Union}}$

Who are the most similar?

Name	Height
John	1,70 m
Paul	1,78 m
Peter	1,90 m

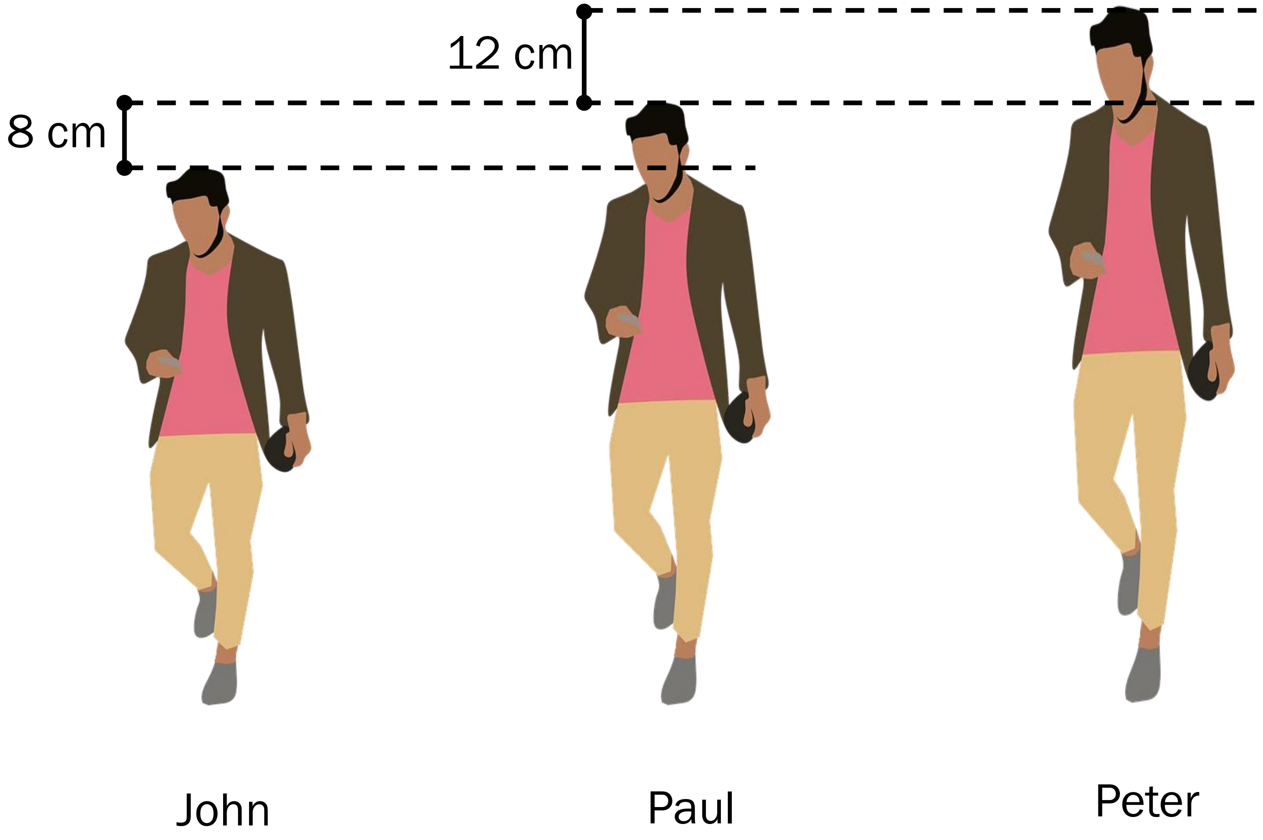
John

Paul

Peter

Who are the most similar?

Name	Height
John	1,70 m
Paul	1,78 m
Peter	1,90 m



Euclidean Distance

- The **Euclidean Distance** measures the straight-line distance between two points in Euclidean space. It reflects the length of the shortest path connecting those points, as if you were stretching a string tightly between them. Imagine a bird flying directly from one spot to another without any turns or bends—that's the Euclidean distance.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

p, q **=** two points in Euclidean n-space

q_i*, *p_i = Euclidean vectors, starting from the origin of the space (initial point)

n = n-space

Who are the most similar?

Name	Height	Weight	Eye Color	Country	Marital Status	Has Sons	Has Car
John	1,70 m	78 kg	Blue	England	Married	Yes	Yes
Paul	1,78 m	85 kg	Brown	Ireland	Married	No	No
Peter	1,90 m	98 kg	Blue	England	Single	Yes	Yes

Jaccard Coefficient - Example

- Datasets: A: ({0, 1, 2, 5, 6, 8, 9}) and B: ({0, 2, 3, 4, 5, 7, 9})
- Find the total number of elements present in both datasets (intersection):
  - A ∩ B = {0, 2, 5, 9} => 4
- Find the total number of observations in either set (union):
  - A ∪ B = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} => 10
- Calculate the Jaccard Similarity:
  - J(A, B) = 4 / 10 => 0.4

Create your own Distance Measurement

- You can create your own distance measurement, as long as you follow four basic premises:

1. There is no negative distance: $d(i,j) \geq 0$
2. The distance from an object to itself is always zero: $d(i,i) = 0$
3. Distance is a symmetric function: $d(i,j) = d(j,i)$
4. The straight-line distance between *i* and *j* is less than any distance that passes through a third point *h*, as long as *h* is not a point that belongs to the straight line that joins *i* and *j* (triangular inequality principle): $d\,(i, j) \leq d\,(i, h) + d(h, j)$

Credit Score file columns

- Name: uses common surnames and given names.
- Date of Birth: format YYYY-MM-DD or DD/MM/YYYY
- Salary: the annual value.
- Gender: male, female.
- Marital Status: single, married, divorced, widowed
- Country: Brazil, USA, Italy, Ireland, England
- Number of Sons: a whole number
- Degree of Education: none, associate, bachelor, master and doctoral
- Has Car and Own Home are binary variables
- Rent Value: $0,00 if own home is true; otherwise usually 10% to 30% of salary.
- Credit Score: whole number between 0 and 5.

Credit Score file columns - tips

- Check if the values in categorical variables are as expected
- *None* can have different meanings: it can be a category, zero/nothing, or an unspecified/unknown value. Evaluate case by case
- Look for discrepancies, for example: low salary and high rent (and vice versa); has a house, but pays rent; low age and very high salary, ...

Final Report – What is Expected

- For each column, describe:
  - Any assumption you made (e.g., if a binary variable is empty, it was considered FALSE)
  - Translations (e.g., Brasil was translated to Brazil)
  - Technique used to reduce impact of outliers and missing values
  - Delete cases: explain why you decide to delete this case
  - Data Origin: if you get data for an external source, mention it

References

- Discrete vs. Continuous Data: What’s the Difference?
  - www.g2.com/articles/discrete-vs-continuous-data
- Encoding Categorical Data: One-Hot vs Label Encoding
  - dataheadhunters.com/academy/encoding-categorical-data-one-hot-vs-label-encoding
- Transform Variables in Regression
  - www.studysmarter.co.uk/explanations/engineering/engineering-mathematics/transform-variables-in-regression
- Variable Transformation
  - bookdown.org/mike/data_analysis/variable-transformation.html
- HAN, Jiawei; KAMBER, Micheline. Data Mining: Concepts and Techniques. San Diego: Academic Press, 2001.
- PRASS, Fernando S. Estudo comparativo entre algoritmos de análise de agrupamentos em data mining. Florianópolis, 2004.

Thank you

Fernando Prass

fprass@gmail.com