

---

---

# Exploratory Data Analysis

Unit 3

---

---

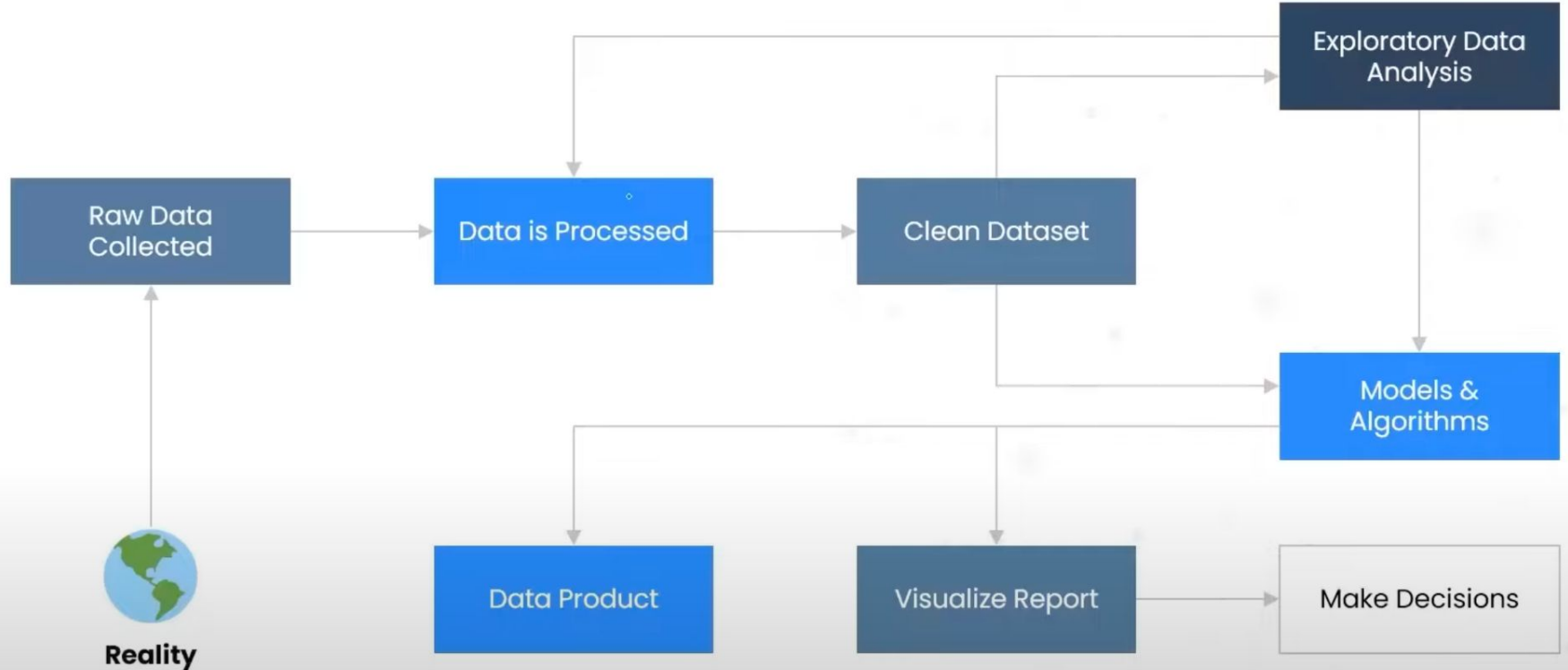
# Introduction to EDA

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, focusing on analyzing and visualizing datasets to uncover patterns, relationships, and insights without preconceived hypotheses

EDA involves using graphical representations and summary statistics to explore data sets.

The primary goal is to understand the underlying structure of the data, identify anomalies, and generate hypotheses for further analysis.

# Data Analytics/Science Process



# Steps in EDA

Understand the Data: Familiarize yourself with the dataset, including its structure and the types of variables involved.

Data Collection: Gather data from various sources such as databases, APIs, or web scraping.

Data Cleaning: Address issues like missing values, duplicates, and incorrect data types to ensure data quality.

Data Transformation: Normalize or standardize data as needed and create new features through feature engineering.

Data Integration: Combine data from different sources to form a comprehensive dataset for analysis.

# Descriptive Statistics

Descriptive statistics is a fundamental aspect of statistical analysis that focuses on summarizing and presenting the characteristics of a data set.

It serves as a crucial tool in both quantitative and qualitative research by providing clear insights into the data without making inferences or predictions about a larger population.

# Descriptive statistics

```
graph TD; A[Descriptive statistics] -.-> B[Distribution]; A -.-> C[Measures of central tendency]; A -.-> D[Measures of variability]; C -.-> E[Mean]; C -.-> F[Median]; C -.-> G[Mode]; D -.-> H[Range]; D -.-> I[Standard deviation]; D -.-> J[Variance]; D -.-> K[Interquartile range];
```

Distribution

Measures of central  
tendency

Measures of variability

Mean

Median

Mode

Range

Standard deviation

Variance

Interquartile range

# Mean

The mean, or average, is calculated by summing all values in a dataset and dividing by the number of values.

$$\text{Mean} = \frac{\sum_{i=1}^n x_i}{n}$$

In finance, the mean is used to analyze investment returns.

For example, if an investor tracks the annual returns of a stock over five years as 10%, 12%, -5%, 8%, and 15%, the mean return can be calculated as:

$$\text{Mean Return} = (10 + 12 - 5 + 8 + 15) / 5 = 8\%$$

This average helps investors gauge the stock's performance over time and make informed decisions about future investment

# Median

The median is the middle value in a dataset when arranged in ascending order. If there is an even number of observations, the median is the average of the two middle numbers.

In real estate, agents often use the median home price to provide a better understanding of market conditions.

For instance, if home prices in a neighborhood are \$200,000, \$250,000, \$300,000, \$350,000, and \$1,000,000, the median price is \$300,000.

This measure is less affected by extreme values (like the \$1 million home) than the mean would be



# Mode

The mode is the value that appears most frequently in a dataset.

In market research, companies analyze customer preferences to determine which product features are most popular.

For example, if a smartphone company finds that most customers prefer blue phones (with sales data showing blue sold 150 units compared to other colors selling fewer), blue becomes the mode of their color preference data.

This insight helps guide product development and marketing strategies

# Standard Deviation

Standard deviation measures the amount of variation or dispersion in a set of values.

A low standard deviation indicates that values tend to be close to the mean, while a high standard deviation indicates that values are spread out over a wider range.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$$

# Standard Deviation

In finance, standard deviation is used to assess investment risk.

For example, if two stocks have the same mean return but one has a higher standard deviation, it indicates that its returns are more volatile and thus riskier.

Investors may prefer stocks with lower standard deviations for stability

# Variance

Variance quantifies how much the numbers in a dataset differ from the mean. It is essentially the square of the standard deviation.

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

# Variance

In quality control processes in manufacturing, variance can help determine product consistency.

If a factory produces widgets with weights averaging 10 grams but with high variance, it indicates significant weight discrepancies among products.

Reducing variance can lead to improved quality assurance

---

---

# Hypothesis Testing

---

---

# Introduction

Hypothesis testing is a statistical method used to make inferences or draw conclusions about a population based on sample data.

It involves formulating a hypothesis, collecting data, and then using statistical techniques to determine whether the evidence supports the hypothesis.

# Significance Level ( $\alpha$ )

The significance level is the threshold for deciding whether to reject the null hypothesis.

Common values are 0.05 (5%) and 0.01 .

It represents the probability of making a Type I error, which occurs when the null hypothesis is incorrectly rejected.



# Null Hypothesis ( $H_0$ )

The null hypothesis is a statement that there is no effect, no difference, or no relationship between variables in the population. It serves as the default assumption that any observed differences in sample data are due to random chance rather than a true effect.

Examples of Null Hypothesis:

- In a clinical trial, the null hypothesis might state that a new drug has no effect on patients compared to a placebo.
- In a study comparing two teaching methods, the null hypothesis could assert that there is no difference in student performance between the two methods.

The null hypothesis is typically tested using statistical methods, and the outcome of the test will either lead to its rejection or failure to reject.

# Alternate Hypothesis (H1)

The alternative hypothesis is a statement that contradicts the null hypothesis. It proposes that there is a significant effect, difference, or relationship between the variables being studied. If the null hypothesis is rejected, the alternative hypothesis is considered to be supported.

Examples of Alternative Hypothesis:

- In the clinical trial mentioned earlier, the alternative hypothesis would state that the new drug does have an effect on patients.
- In the teaching methods study, the alternative hypothesis could claim that one teaching method leads to better student performance than the other.

# Hypothesis testing process

1. State the Hypotheses: Clearly define the null and alternative hypotheses.
2. Choose a Significance Level (
3.  $\alpha$
4.  $\alpha$ ): Common choices are 0.05 or 0.01, which represent the probability of rejecting the null hypothesis when it is actually true (Type I error).
5. Collect Data: Gather sample data relevant to the hypotheses.
6. Perform Statistical Test: Use an appropriate statistical test (e.g., t-test, chi-square test) to analyze the data.
7. Make a Decision: Based on the p-value obtained from the test

# T-test

A t-test is a statistical hypothesis test used to determine whether there is a significant difference between the means of one or two groups. It is particularly useful when the sample sizes are small and the population standard deviation is unknown. The t-test can be categorized into three types:

**One-Sample T-Test:** Compares the mean of a single group against a known value (e.g., a population mean).

**Independent Two-Sample T-Test:** Compares the means of two independent groups to see if they are significantly different from each other.

## Assumptions of the T-Test:

- The data are continuous and normally distributed.
- The samples are independent (for independent t-tests).
- The variances of the two groups are equal (for independent t-tests).
- The sample data are randomly sampled from the population.

Usage: T-tests are commonly used in various fields, including psychology, medicine, and social sciences, to analyze experimental data and draw conclusions about populations based on sample data.

# Chi-square test

The chi-square test is a statistical test used to determine if there is a significant association between categorical variables. It compares the observed frequencies in each category to the frequencies expected if there were no association between the variables.

There are two main types of chi-square tests:

1. Chi-Square Test of Independence: Tests whether two categorical variables are independent of each other. For example, it can be used to determine if gender is independent of voting preference.
2. Chi-Square Goodness of Fit Test: Tests whether the observed distribution of a single categorical variable matches an expected distribution.

## Assumptions of the Chi-Square Test:

- The data are categorical.
- The observations are independent.
- The sample size should be sufficiently large, typically at least 5 expected frequencies per category.

Usage: Chi-square tests are widely used in market research, genetics, and social sciences to analyze categorical data and understand relationships between variables.

# Importing the libraries

```
import numpy as np
```

```
from scipy import stats
```



## Create some sample data

```
# Sample data for traditional method (mean = 75)
```

```
traditional_scores = np.array([70, 75, 80, 65, 78, 72, 74, 76, 77, 79])
```

```
# Sample data for new method (mean > 75)
```

```
new_method_scores = np.array([80, 85, 82, 90, 88, 84, 86, 91, 87, 89])
```

```
# Perform a one-sample t-test
```

```
t_statistic, p_value = stats.ttest_1samp(new_method_scores, 75)
```

```
print(f'Test Statistic: {t_statistic}')
```

```
print(f'P-value: {p_value}')
```

# Level of significance

$\alpha = 0.05$

# Hypothesis Testing

if  $p\_value < \alpha$ :

    print("Reject the null hypothesis: The new teaching method is effective.")

else:

    print("Fail to reject the null hypothesis: No significant effect of the new teaching method.")

## Another data set

```
import numpy as np
```

```
import pandas as pd
```

```
# Set a seed for reproducibility
```

```
np.random.seed(42)
```

```
# Generate daily sales data for each store over 31 days
```

```
dates = pd.date_range(start='1/1/2023', end='1/31/2023')
```

```
# Generate sales data for Store A (mean = 5000, std = 500)
```

```
store_a_sales = np.random.normal(loc=5000, scale=500, size=len(dates))
```

```
# Generate sales data for Store B (mean = 4500, std = 400)
```

```
store_b_sales = np.random.normal(loc=4500, scale=400, size=len(dates))
```

```
# Create the dataset
```

```
df = pd.DataFrame({  
    'date': np.tile(dates, 2), # Repeat the dates for both stores  
    'store_id': ['A'] * len(dates) + ['B'] * len(dates), # Store identifiers  
    'sales': np.concatenate((store_a_sales, store_b_sales)) # Combine sales  
data  
})
```

```
# Sort the dataset by date
```

```
df = df.sort_values('date').reset_index(drop=True)
```

```
print(df.head(10)) # Display the first 10 rows of the dataset
```

```
from scipy.stats import ttest_ind
```

```
# Filter data for each store
```

```
store_a_sales = df.loc[df['store_id'] == 'A', 'sales']
```

```
store_b_sales = df.loc[df['store_id'] == 'B', 'sales']
```

```
# Perform two-sample t-test
```

```
t_stat, p_value = ttest_ind(store_a_sales, store_b_sales)
```

```
# Set significance level
```

```
alpha = 0.05
```

```
# Evaluate results
```

```
if p_value < alpha:
```

```
    print("Reject the null hypothesis. There is a significant difference in  
    average daily sales between the two stores.")
```

```
else:
```

```
    print("Fail to reject the null hypothesis. There is no significant difference  
    in average daily sales between the two stores.")
```



```
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency
```

```
# Sample data
```

```
data = {
    'Names': ['Alice', 'Bob', 'Charlie', 'David', 'Eva', 'Frank', 'Grace', 'Hannah', 'Ian',
              'Jack'],
    'Degree': ['Bachelor', 'Master', 'Bachelor', 'PhD', 'Master', 'Bachelor', 'PhD',
              'Master', 'Bachelor', 'PhD'],
    'Years_of_Experience': [2, 5, 3, 10, 7, 1, 8, 4, 6, 9],
    'Salary': [50000, 70000, 55000, 90000, 75000, 45000, 95000, 80000, 60000,
              100000]
}
```

```
# Create DataFrame
```

```
df = pd.DataFrame(data)
```

```
# Step 1: Categorize Salary
```

```
def categorize_salary(salary):
```

```
    if salary < 60000:
```

```
        return 'Low'
```

```
    elif 60000 <= salary < 80000:
```

```
        return 'Medium'
```

```
    else:
```

```
        return 'High'
```

```
df['Salary_Category'] = df['Salary'].apply(categorize_salary)
```

```
# Step 2: Create Contingency Table
```

```
contingency_table = pd.crosstab(df['Degree'],  
df['Salary_Category'])
```

```
print("Contingency Table:")
```

```
print(contingency_table)
```

```
# Step 3: Perform Chi-Square Test
```

```
chi2, p, dof, expected = chi2_contingency(contingency_table)
```

```
# Output results
```

```
print(f"\nChi-Square Statistic: {chi2}")
```

```
print(f"P-value: {p}")
```

```
print(f"Degrees of Freedom: {dof}")
```

```
print(f"Expected Frequencies:\n{expected}")
```

```
# Hypothesis Testing
```

```
alpha = 0.05
```

```
if p < alpha:
```

```
    print("Reject the null hypothesis: There is a significant relationship between Degree and  
Salary Category.")
```

```
else:
```

```
    print("Fail to reject the null hypothesis: There is no significant relationship between Degree  
and Salary Category.")
```

```
# Visualizations
```

```
# 1. Bar plot of Degree vs Salary Category
```

```
plt.figure(figsize=(8, 5))
```

```
sns.countplot(data=df, x='Degree', hue='Salary_Category', palette='Set2')
```

```
plt.title('Count of Degrees by Salary Category')
```

```
plt.xlabel('Degree')
```

```
plt.ylabel('Count')
```

```
plt.legend(title='Salary Category')
```

```
plt.show()
```

# 2. Box plot of Salary by Degree

```
plt.figure(figsize=(8, 5))
```

```
sns.boxplot(data=df, x='Degree', y='Salary', palette='Set3')
```

```
plt.title('Salary Distribution by Degree')
```

```
plt.xlabel('Degree')
```

```
plt.ylabel('Salary')
```

```
plt.show()
```

### # 3. Histograms of Years of Experience and Salary

```
plt.figure(figsize=(12, 5))
```

#### # Histogram for Years of Experience

```
plt.subplot(1, 2, 1)
```

```
plt.hist(df['Years_of_Experience'], bins=5, color='skyblue',  
edgecolor='black')
```

```
plt.title('Distribution of Years of Experience')
```

```
plt.xlabel('Years of Experience')
```

```
plt.ylabel('Frequency')
```

```
# Histogram for Salary
```

```
plt.subplot(1, 2, 2)
```

```
plt.hist(df['Salary'], bins=5, color='salmon', edgecolor='black')
```

```
plt.title('Distribution of Salary')
```

```
plt.xlabel('Salary')
```

```
plt.ylabel('Frequency')
```

```
plt.tight_layout()
```

```
plt.show()
```



# ANOVA Test

ANOVA, or Analysis of Variance, is a statistical method used to determine whether there are significant differences between the means of three or more groups.

It helps researchers understand if the variations in data are due to actual differences between groups or if they can be attributed to random chance.

- One-Way ANOVA: Compares the means of three or more independent groups based on one independent variable. For example, testing the effectiveness of three different diets on weight loss.
- Two-Way ANOVA: Assesses the impact of two independent variables on a dependent variable and can also evaluate interaction effects between the two independent variables. For example, examining the effects of diet type and exercise level on weight loss.

# ANOVA

ANOVA enables us to test for significance of differences among more than two sample means

This can also be defined as the extension of t-test

Test Statistics for ANOVA is F-test

# Assumption for ANOVA

Sample follows Normal Distribution

Samples have been selected randomly and independently

Each group should have common variance

Data are independent

# Basics of ANOVA

Null Hypothesis: The means for all the groups are same.

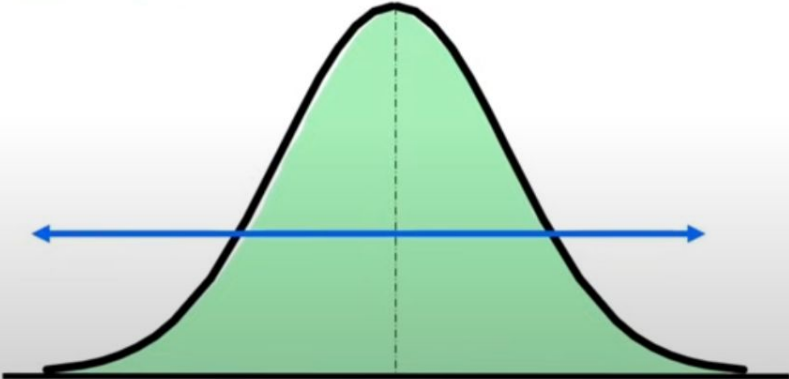
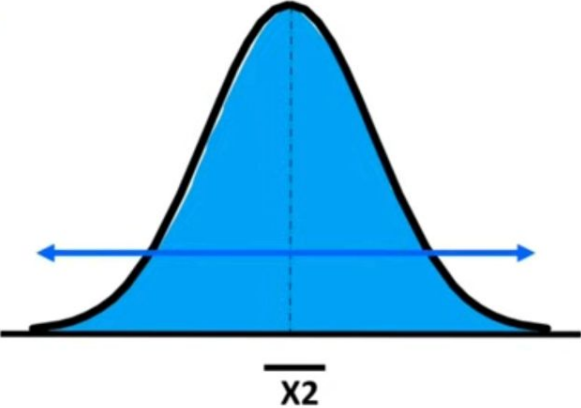
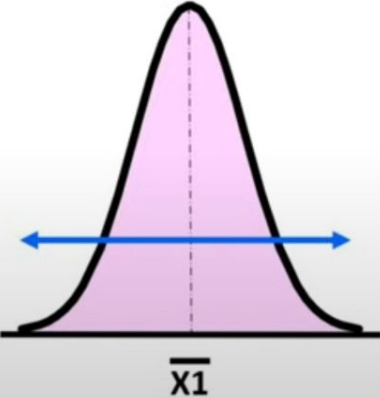
$$\mu_1 = \mu_2 = \mu_3 = \dots = \mu_n$$

Alternative Hypothesis: The means are different for atleast one pair of groups

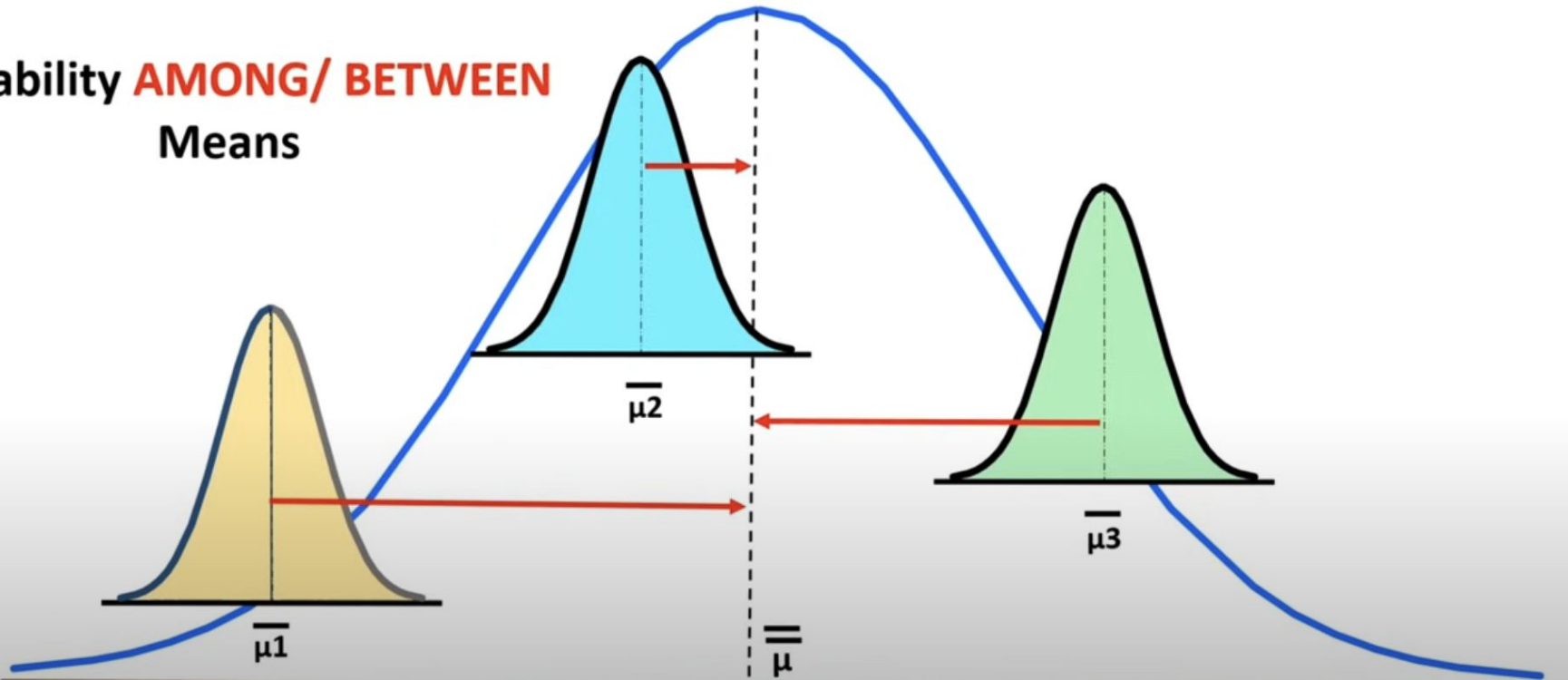
$$\mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \neq \mu_n$$

ANOVA = Variance between / Variance Within

Variability **AROUND/ WITHIN**  
distribution



Variability **AMONG/ BETWEEN**  
Means

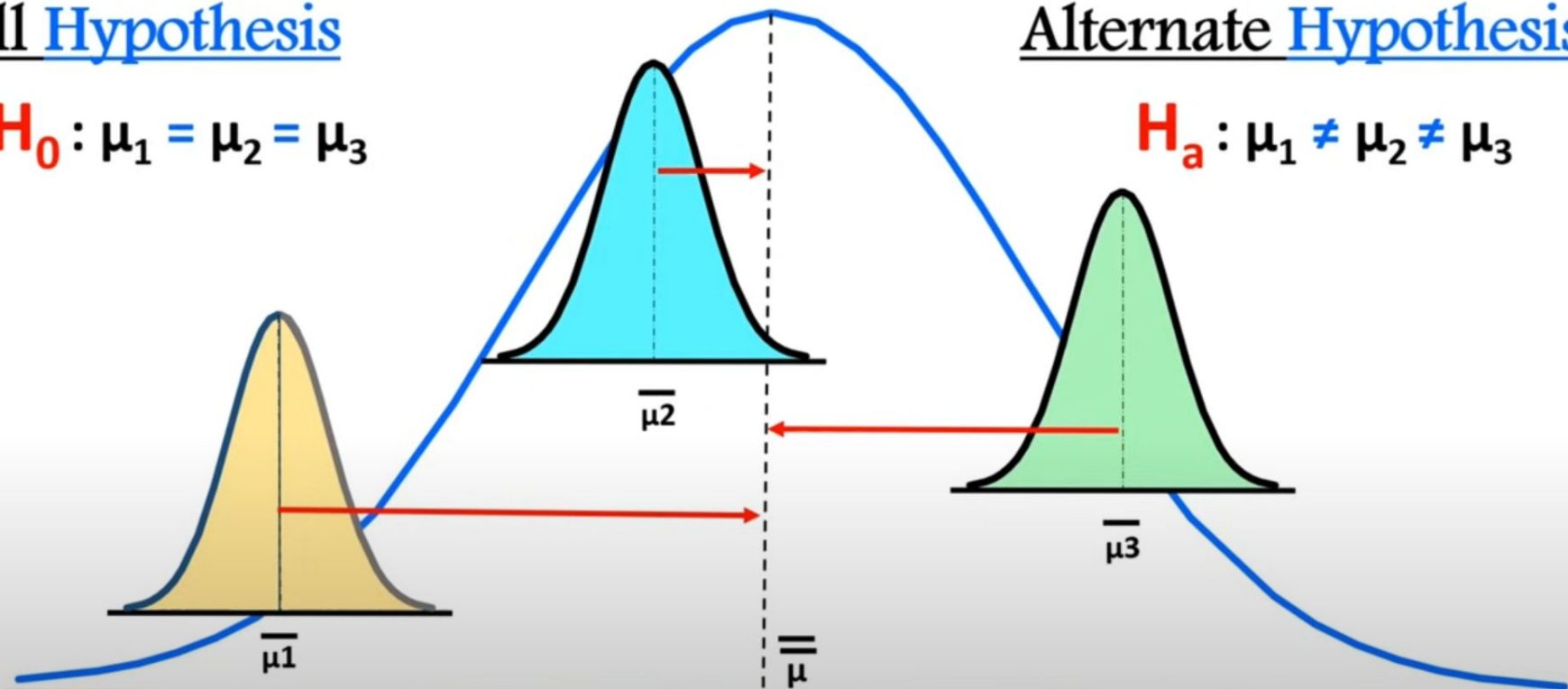


## Null Hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3$$

## Alternate Hypothesis

$$H_a: \mu_1 \neq \mu_2 \neq \mu_3$$





$$\underline{\text{ANOVA}} = \frac{\text{Variance Between}}{\text{Variance Within}}$$

$$\underline{\text{Total Variance}} = \text{Variance Between} + \text{Variance Within}$$

Variance Between

Variance Within

$> 1$

Reject  $H_0$

Variance Between

Variance Within

$< 1$

Fail to Reject  $H_0$

Variance Between

Variance Within

$= 1$

Fail to Reject  $H_0$

Sno	Method A	Method B	Method C
1.	10	8	9
2.	9	9	8
3.	8	10	7
4.	7.5	8	10
5.	8.5	8.5	9
6.	9	7	8
7.	10	9.5	7
8.	8	9	10
9.	8	7	9
10.	9	10	8
Group Mean	8.7	8.6	8.5

Overall Mean **8.6**

**Between** Group Variation =  $10 \times (8.7 - 8.6)^2 + 10 \times (8.6 - 8.6)^2 + 10 \times (8.5 - 8.6)^2$

**Between** Group Variation = **0.2**

**Within** Group Variation:  $\sum (X_{ij} - X_j)^2$

**Where:**

$\Sigma$ : a symbol that means "sum"  
 $X_{ij}$ : the  $i^{\text{th}}$  observation in group  $j$   
 $X_j$ : the mean of group  $j$

**Method A:**  $(10 - 8.7)^2 + (9 - 8.7)^2 + (8 - 8.7)^2 + (7.5 - 8.7)^2 + (8.5 - 8.7)^2 + (9 - 8.7)^2 + (10 - 8.7)^2 + (8 - 8.7)^2 + (8 - 8.7)^2 + (9 - 8.7)^2 = \mathbf{6.6}$

**Method B:**  $(8 - 8.6)^2 + (9 - 8.6)^2 + (10 - 8.6)^2 + (8 - 8.6)^2 + (8.5 - 8.6)^2 + (7 - 8.6)^2 + (9.5 - 8.6)^2 + (9 - 8.6)^2 + (7 - 8.6)^2 + (10 - 8.6)^2 = \mathbf{10.9}$

**Method C:**  $(9 - 8.5)^2 + (8 - 8.5)^2 + (7 - 8.5)^2 + (10 - 8.5)^2 + (9.5 - 8.5)^2 + (8 - 8.5)^2 + (7 - 8.5)^2 + (10 - 8.5)^2 + (9 - 8.5)^2 + (8 - 8.5)^2 = \mathbf{10.5}$

**Within** Group Variation:  $6.6 + 10.9 + 10.5 = \mathbf{28}$

$$\frac{\text{Variance Between}}{\text{Variance Within}} = \frac{0.2}{28} = 0.0071 < 1$$

Fail to Reject  $H_0$

“Means are very close to overall mean and distribution overlap is hard to distinguish”.

```
import pandas as pd
import numpy as np
from scipy import stats
import statsmodels.api as sm
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Create the dataset
```

```
data = {
```

```
    'Age': [25, 32, 41, 28, 35, 45, 30, 38, 43],
```

```
    'Income': [50000, 65000, 75000, 55000, 70000, 80000, 60000, 72000,  
78000],
```

```
    'Education': ['Bachelor', 'Master', 'PhD', 'Bachelor', 'Master', 'PhD',  
'Bachelor', 'Master', 'PhD']
```

```
}
```

```
df = pd.DataFrame(data)
```

# 1. T-test between Bachelor's and Master's

```
bachelor_income = df[df['Education'] == 'Bachelor']['Income']
```

```
master_income = df[df['Education'] == 'Master']['Income']
```

```
t_stat, p_value_ttest = stats.ttest_ind(bachelor_income, master_income)
```

```
print(f"T-test results: t-statistic = {t_stat}, p-value = {p_value_ttest}")
```

# 2. One-way ANOVA

```
anova_result = ols('Income ~ Education', data=df).fit()
```

```
anova_table = sm.stats.anova_lm(anova_result, typ=2)
```

```
print("\nOne-way ANOVA results:")
```

```
print(anova_table)
```

# 3. Regression analysis

```
regression_model = ols('Income ~ Age + Education', data=df).fit()
```

```
print("\nRegression analysis results:")
```

```
print(regression_model.summary())
```

```
# Data Visualization
```

```
# Set the style for seaborn
```

```
sns.set(style="whitegrid")
```

```
# 1. Box Plot of Income by Education
```

```
plt.figure(figsize=(8, 5))
```

```
sns.boxplot(x='Education', y='Income', data=df, palette='Set2')
```

```
plt.title('Income Distribution by Education Level')
```

```
plt.xlabel('Education Level')
```

```
plt.ylabel('Income')
```

```
plt.show()
```



```
# 2. Scatter Plot of Age vs Income
```

```
plt.figure(figsize=(8, 5))
```

```
sns.scatterplot(x='Age', y='Income', hue='Education', data=df, palette='Set1',  
s=100)
```

```
plt.title('Scatter Plot of Age vs Income')
```

```
plt.xlabel('Age')
```

```
plt.ylabel('Income')
```

```
plt.legend(title='Education Level')
```

```
plt.show()
```

### # 3. Histograms for Age and Income

```
fig, axes = plt.subplots(1, 2, figsize=(12, 5))
```

#### # Histogram for Age

```
sns.histplot(df['Age'], bins=5, kde=True, ax=axes[0], color='skyblue')
```

```
axes[0].set_title('Distribution of Age')
```

```
axes[0].set_xlabel('Age')
```

```
axes[0].set_ylabel('Frequency')
```

```
# Histogram for Income
```

```
sns.histplot(df['Income'], bins=5, kde=True, ax=axes[1], color='salmon')
```

```
axes[1].set_title('Distribution of Income')
```

```
axes[1].set_xlabel('Income')
```

```
axes[1].set_ylabel('Frequency')
```

```
plt.tight_layout()
```

```
plt.show()
```

# Trend Identification

A trend refers to the general direction in which something is changing or developing over time.

It can be observed in various contexts, such as social behaviors, economic indicators, fashion, and market movements.

Trends can be identified through various analytical methods, including statistical analysis and visual tools like trendlines in graphs.

Trend identification is crucial in various fields, including environmental science, finance, and social sciences, business, trading

---

---

# Correlations

---

---

# Introduction

Correlation is a statistical measure that describes the degree to which two or more variables move in relation to each other.

Understanding correlation is essential in various fields, including finance, healthcare, and social sciences, as it helps identify relationships between variables, which can inform decision-making and predictive modeling.

Correlation quantifies the strength and direction of a relationship between two variables. It is typically represented by a correlation coefficient, denoted as  $r$ , which ranges from -1 to +1

- Positive Correlation ( $r > 0$ ): Indicates that as one variable increases, the other variable also tends to increase. For example, there may be a positive correlation between hours studied and exam scores.
- Negative Correlation ( $r < 0$ ): Suggests that as one variable increases, the other variable tends to decrease. An example could be the relationship between temperature and heating costs; as temperatures rise, heating costs typically fall.
- No Correlation ( $r = 0$ ): Implies that there is no discernible relationship between the variables.

---

---

# Analyzing trends

---

---



# Mann-Kendall Test

The Mann-Kendall test (MK test) is a widely used statistical method for assessing trends in time series data.

This non-parametric test is particularly valuable in various fields, including environmental science, hydrology, and climate studies, where it helps to determine whether a variable exhibits a consistent upward or downward trend over time.

# Introduction

The primary purpose of the Mann-Kendall test is to evaluate monotonic trends in data over time.

A monotonic trend means that the variable of interest consistently increases or decreases, regardless of whether this trend is linear.

The test does not require the data to follow a normal distribution, making it suitable for a wide range of datasets,

# Hypothesis

The Mann-Kendall test operates under two hypotheses:

Null Hypothesis ( $H_0$ ): There is no monotonic trend in the data.

Alternative Hypothesis ( $H_a$ ): A monotonic trend exists (this can be either positive or negative)

# Steps

The test involves comparing each data point with all preceding points to determine the direction of change:

For each pair of observations, if the later observation is greater than the earlier one, it contributes positively to the trend.

Conversely, if it is smaller, it contributes negatively.

The total number of positive and negative contributions is calculated to derive a statistic that indicates the presence and direction of a trend

## Checking the direction of sign changes

$$\text{sgn}(x_j - x_i) = \begin{cases} +1 & \text{if } x_j - x_i > 0 \\ 0 & \text{if } x_j - x_i = 0 \\ -1 & \text{if } x_j - x_i < 0 \end{cases}$$

## Finding the total sum

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_j - x_i)$$

# Variance and z-score

$$\text{var}(S) = \frac{1}{18} [n(n-1)(2n+5) + \sum_{i=1}^m t_i(t_i-1)(2t_i)+5]$$

$$Z_s = \begin{cases} \frac{S-1}{\sqrt{\text{Var}(S)}} & : \text{if } S > 0 \\ 0 & : \text{if } S = 0 \\ \frac{S+1}{\sqrt{\text{Var}(S)}} & : \text{if } S < 0 \end{cases}$$

Confidence level	Z-value
90%	1.65
91%	1.7
92%	1.75
93%	1.81
94%	1.88
95%	1.96
96%	2.05
97%	2.17
98%	2.33
99%	2.58

# Interpreting the result

A positive Z-score indicates that there is a tendency for the data to increase over time, suggesting an upward trend. This means that more recent observations are generally higher than earlier ones.

Conversely, a negative Z-score suggests a downward trend, indicating that more recent observations tend to be lower than those observed earlier.

# Statistical Significance

- To determine whether the observed trend is statistically significant, compare the absolute value of the Z-score to a critical value from the standard normal distribution (for example, at a significance level of  $\alpha=0.05$ , the critical value is approximately  $\pm 1.96$ ).
- If  $|Z_{\text{MK}}| > |Z_{1-\alpha/2}|$ , where  $|Z_{1-\alpha/2}|$  corresponds to the critical value, you reject the null hypothesis (which states that there is no trend) and conclude that a significant trend exists in the time series data.



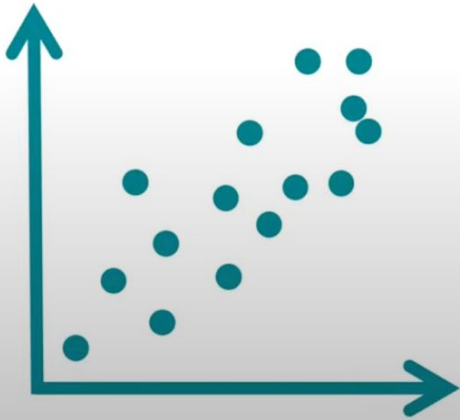
# Spearman's Rank

Spearman's rank correlation is a statistical method used to assess the strength and direction of the relationship between two ranked variables

Spearman's rank correlation coefficient, denoted as  $\rho$  (rho) or sometimes  $r_s$ , measures how well the relationship between two variables can be described using a monotonic function.

It is particularly useful when the data does not meet the assumptions required for Pearson's correlation

**Spearman's rank correlation** examines the **relationship** between two variables.

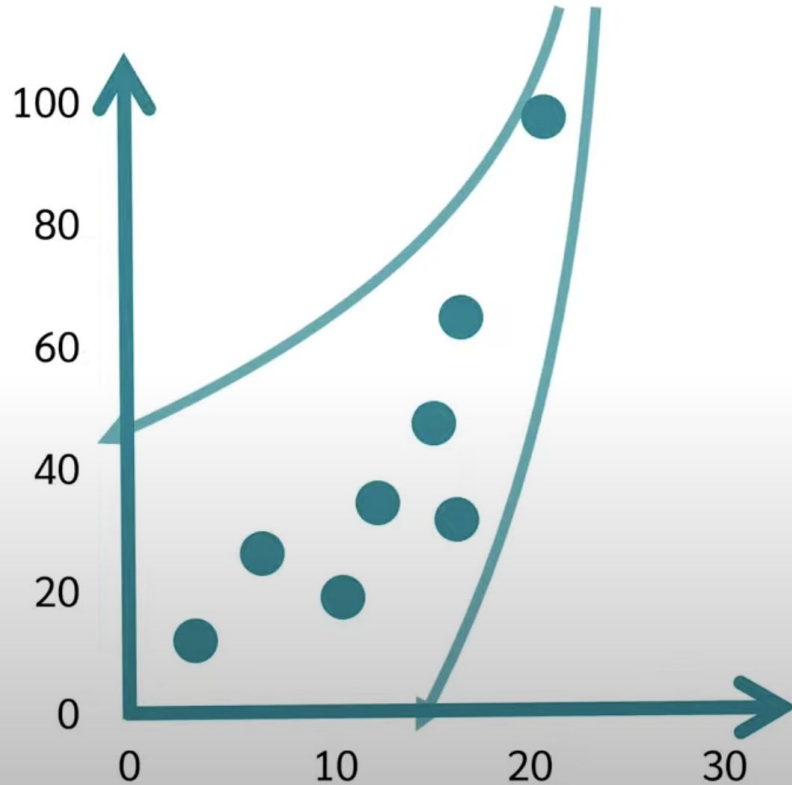


Isn't that exactly what the **Pearson correlation** does?

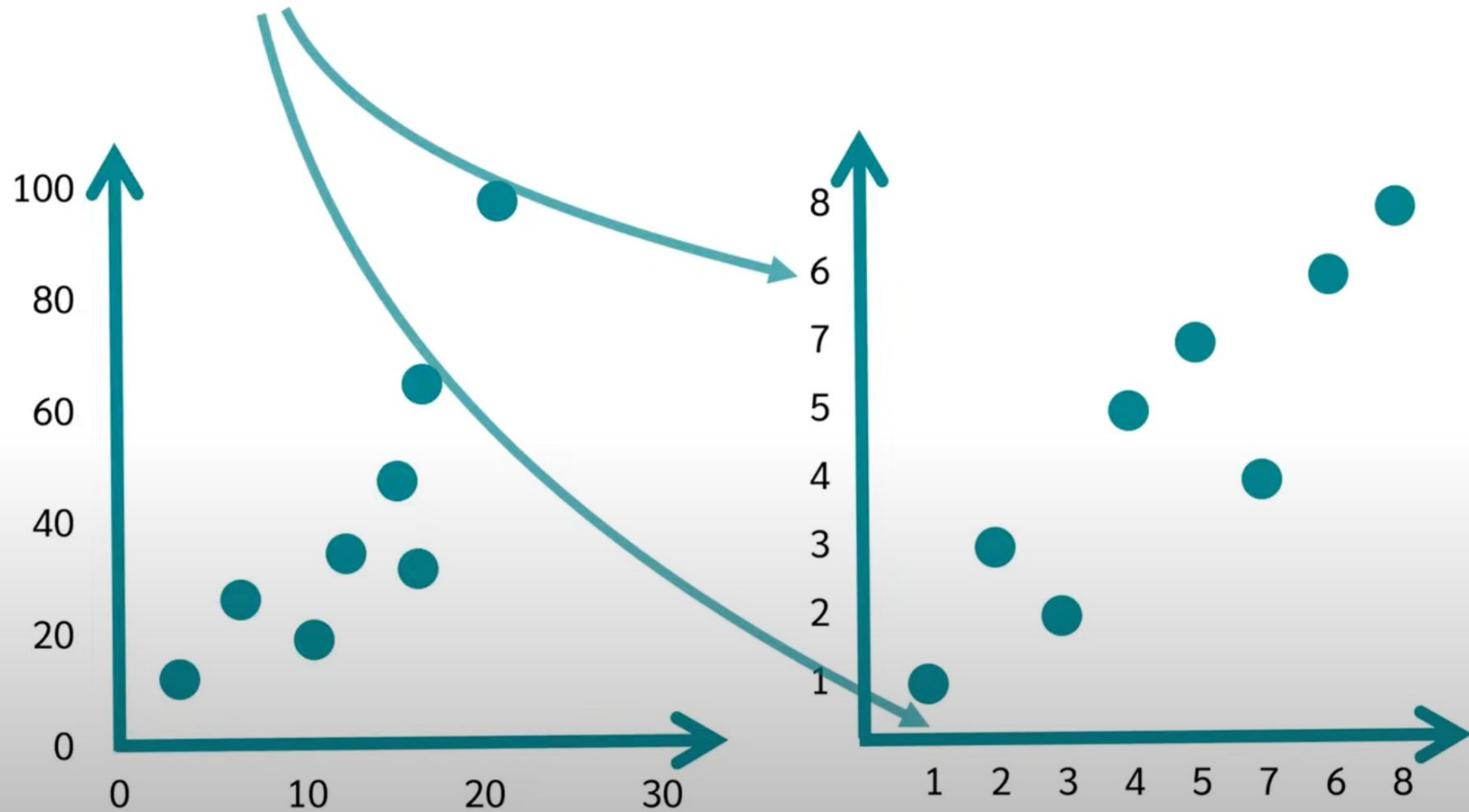


# Spearman correlation

does not use the **raw data**,



but the **ranks** of the data.





12  
15  
17  
18  
20  
21  
22  
26

We measured the **reaction time** of **8 computer players** and asked their **age**.


When we calculate a **Pearson correlation**, we simply take the two variables **reaction time** and **age**



12  
15  
17  
18  
20  
21  
22  
26

14  
25  
20  
35  
45  
30  
60  
95

When we calculate a **Pearson correlation**, we simply take the two variables **reaction time** and **age** and calculate the **Pearson correlation coefficient**.


$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$













12  
15  
17  
18  
20  
21  
22  
26



14  
25  
20  
35  
45  
30  
60  
95

However, we now want to calculate the **Spearman rank correlation**,  
so first we assign a **rank** to each  
person for **reaction time** and **age**.

		
	12	14
	15	25
	17	20
	18	35
	20	45
	21	30
	22	60
	26	95



The **reaction time** is  
already sorted by size.



**12** is the smallest value,  
so gets **rank 1**



12	1	14
15	2	25
17	3	20
18	4	35
20	5	45
21	6	30
22	7	60
26	8	95

We are now doing the same with **age**.



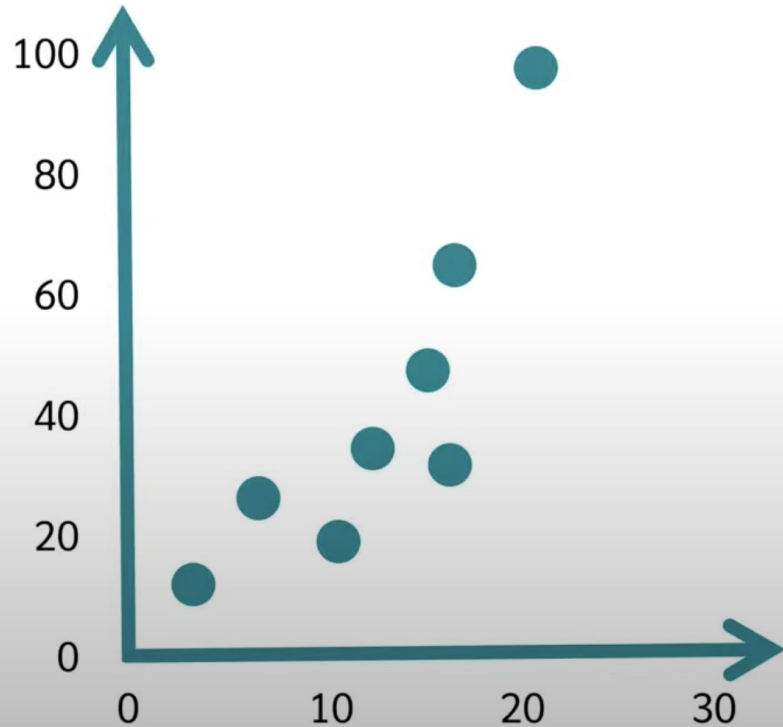
Here we have the **smallest** value,

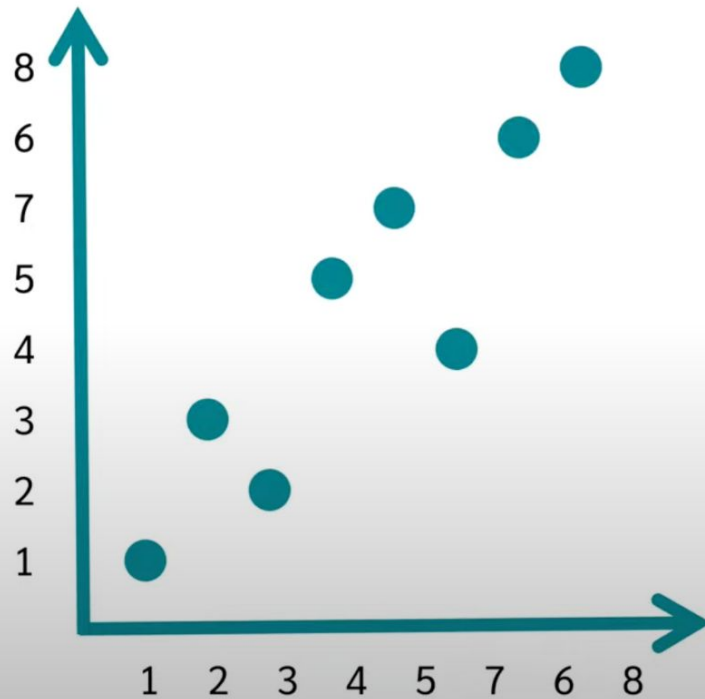
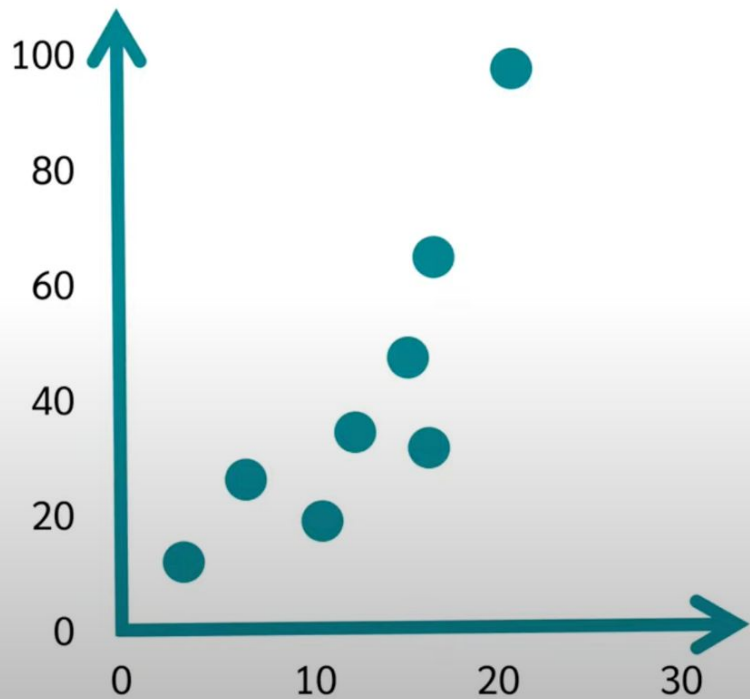


12	1
15	2
17	3
18	4
20	5
21	6
22	7
26	8

14	1
25	3
20	2
35	5
45	6
30	4
60	7
95	8

Let's take a look at this in a **scatter plot**.








If there are **no rank ties**, we can also use this **equation** to calculate the **Spearman correlation**.

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$$

$n$  is the **number of cases**


and  $d$  is the **difference in ranks** between the two variables.

				d	d <sup>2</sup>	
	12	1	14	1	1-1 = 0	0
	15	2	25	3	2-3 = -1	1
	17	3	20	2	3-2 = 1	1
	18	4	35	5	4-5 = -1	1
	20	5	45	6	5-6 = -1	1
	21	6	30	4	7-4 = 2	4
	22	7	60	7	7-7 = 0	0
	26	8	95	8	8-8 = 0	0
					<hr/>	Σ 8

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$$

$n$ , which is the number of people, is 8.

If we put everything in, we get a  
**correlation coefficient** of 0.9.


$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{48}{504} = 0.90$$

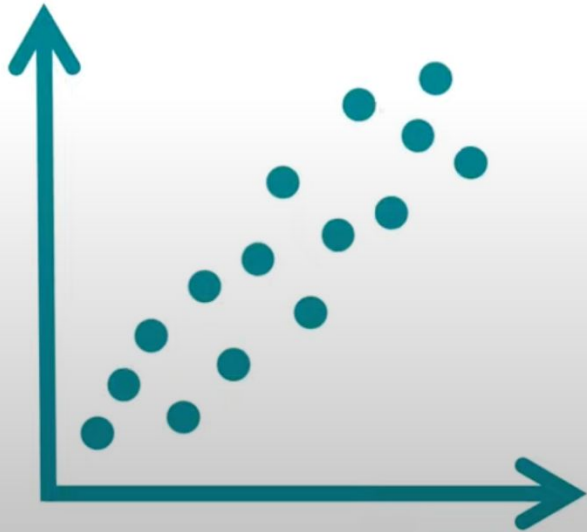
Just like the **Pearson correlation coefficient  $r$**   
the **Spearman correlation coefficient  $r_s$**  also varies between **-1** and **1**.



With the help of the **coefficient**, we  
can now **determine two things**.



- 1 How **strong** the **correlation** is
- 2 and in **which direction** the **correlation** goes.

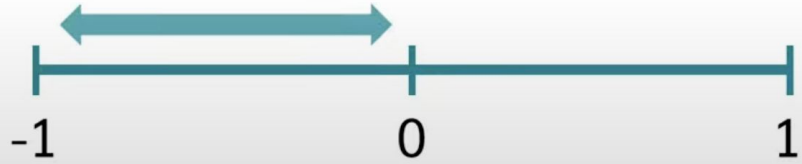


The **strength of the correlation**, can be read in a **table**.

Amount of r	Strength of the correlation
$0.0 < 0.1$	no correlation
$0.1 < 0.3$	low correlation
$0.3 < 0.5$	medium correlation
$0.5 < 0.7$	high correlation
$0.7 < 1$	very high correlation

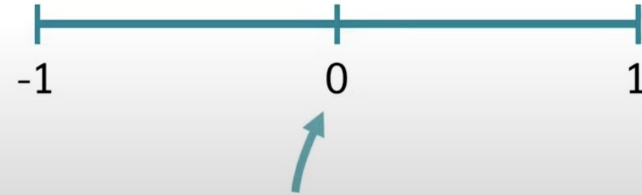
From Kuckartz et al.: Statistik, Eine verständliche Einführung, 2013, p. 213

If we have a coefficient between **-1** and **less than 0**, there is a **negative correlation**,



thus a **negative relationship** between the variables.

If we have a coefficient between **greater than 0** and **1**, there is a **positive correlation**, that is, a **positive relationship** between the two variables.



If the result is **0**, we have **no correlation**.

Often, however, starting from a **sample**,  
we want to **test a hypothesis**  
about the **population**.



Population



Sample



We calculated the **correlation coefficient** from the **sample data**.  
Now we can **test** if the **correlation coefficient** is **significantly different** from **0**.



Thus, the **null hypothesis** is:

The correlation coefficient  **$r = 0$**   There is no relationship.

And the **alternative hypothesis** is:

The correlation coefficient  **$r \neq 0$**   There is a relationship.

Whether the correlation coefficient is significantly different from zero based on the **sample** collected



can be checked using a **t-test**.

Where  $r$  is the **correlation coefficient**

and  $n$  is the **sample size**.

$$t = \frac{r \cdot \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

A **p-value** can then be calculated from the test **statistic t**.

$$t = \frac{r \cdot \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

If the **p-value** is less than the specified **significance level**, which is **usually 5%**, then the **null hypothesis** is rejected, otherwise it is not.



# Sen's slope

Sen's slope, or the Theil-Sen estimator, is a robust statistical method used to determine the slope of a trend in univariate time series data.

It is particularly effective in the presence of outliers and provides a non-parametric alternative to traditional linear regression methods.

This method estimates the slope of the trend line by calculating the median of all possible slopes between pairs of data points.

Sen's slope is commonly applied in fields such as environmental science, hydrology, and climate studies to analyze trends in data such as temperature, rainfall, and other time series measurements.

# Steps in calculation

1. Arrange your data points in order of time
2. For each pair of points  $(x_i, y_i)$  and  $(x_j, y_j)$  where  $j > i$ , calculate the slope using the formul

$$Q_{ij} = \frac{y_j - y_i}{x_j - x_i}$$

3. After calculating all possible slopes, find the median of these slopes.

This median value represents Sen's slope estimator.

# Interpretation

**Positive Slope:** A positive value indicates an upward trend in the data over time (i.e., as time increases, the variable also tends to increase).

**Negative Slope:** A negative value indicates a downward trend (i.e., as time increases, the variable tends to decrease).

**Magnitude:** The unit of Sen's slope reflects the change in the response variable per unit change in time (e.g., units per year).

Sen's slope is widely used for:

Trend Analysis: In hydrological studies to assess changes in precipitation or temperature over time.

Environmental Monitoring: To evaluate long-term trends in ecological data.

Climate Change Studies: To analyze shifts in climate variables over extended periods.

# Data Visualization Techniques

## Histogram

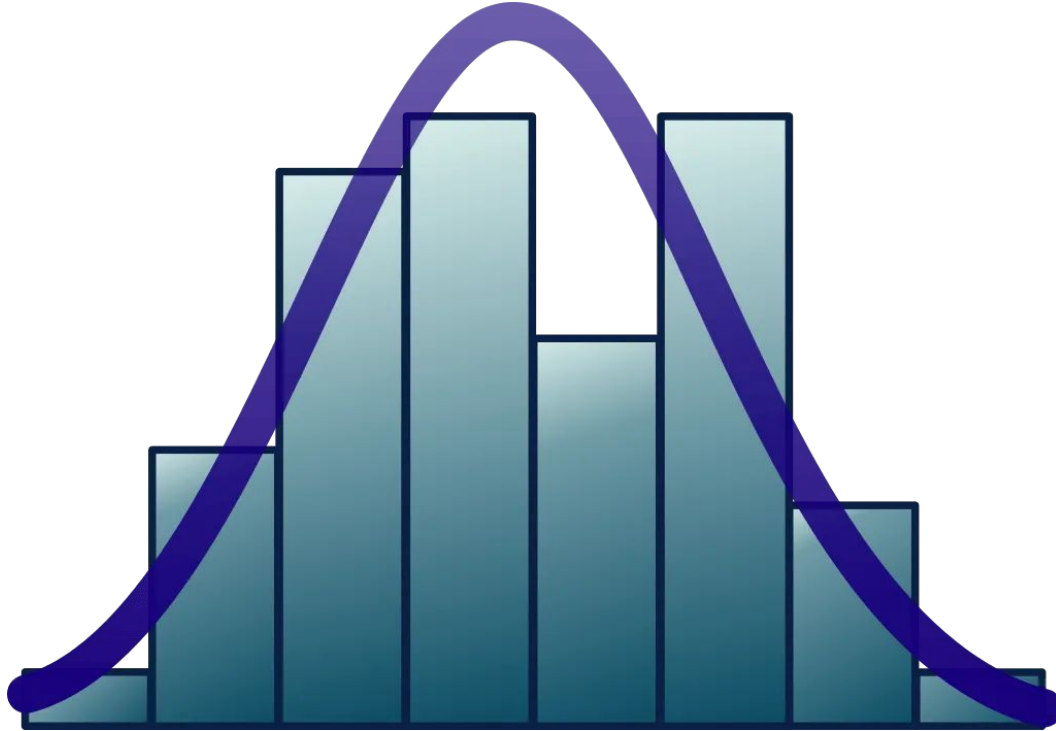
A histogram is a graphical representation that displays the distribution of numerical data. It uses bars to show the frequency of data points within specified intervals.

Histograms are useful for understanding the distribution of continuous data, such as age or income.

They help identify skewness, outliers, and the shape of the distribution.

# Data Visualization Techniques

## Histogram



# Data Visualization Techniques

## Box Plots

A box plot, also known as a box-and-whisker plot, displays the distribution of data based on quartiles. It shows the median, quartiles, and outliers.

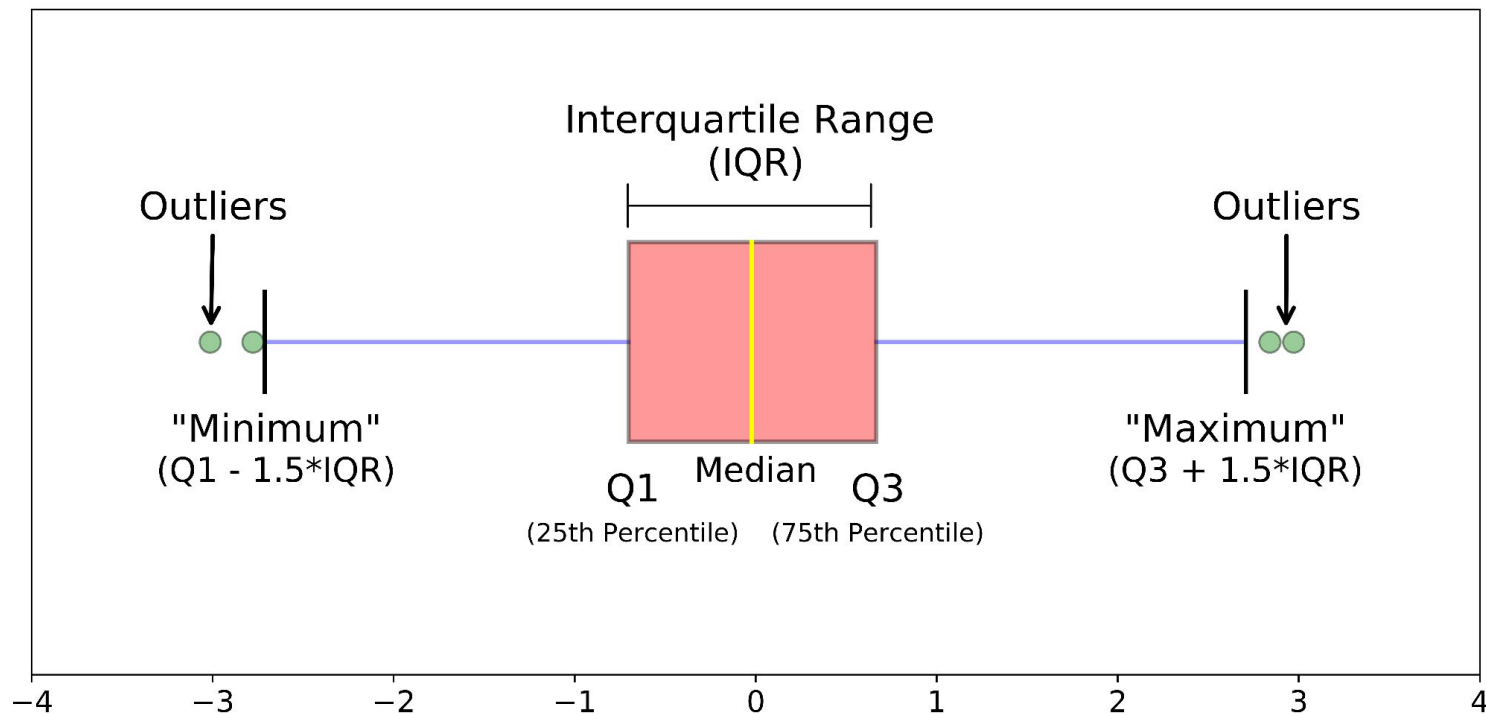
Box plots are effective for comparing the distribution of data across different groups or categories.

They highlight the median, interquartile range (IQR), and outliers, providing insights into data variability and skewness



# Data Visualization Techniques

## Box Plots



# Data Visualization Techniques

## Scatter Plot

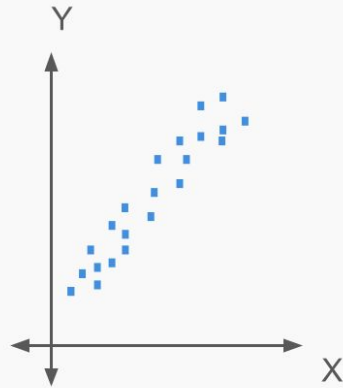
A scatter plot is a graph that displays the relationship between two continuous variables. Each point on the plot represents a pair of values.

Scatter plots are used to visualize correlations or relationships between variables.

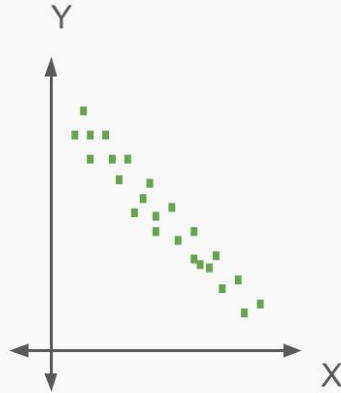
They help identify patterns, such as linear or non-linear relationships, and outliers.

# Data Visualization Techniques

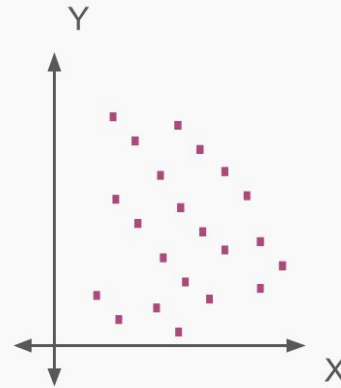
## Scatter Plot Correlation Examples



Positive  
Correlation



Negative  
Correlation



No  
Correlation

# Data Visualization Techniques

## HeatMaps

A heat map is a graphical representation of data where values are depicted by color. It is often used to show relationships between two variables or to display density.

Heat maps are useful for visualizing complex data, such as correlations between variables or geographical distributions.

They provide a clear visual representation of data density or intensity, making it easier to identify patterns or hotspots.

# Data Visualization Techniques

## HeatMaps

### Project Risks Analysis & Actions

