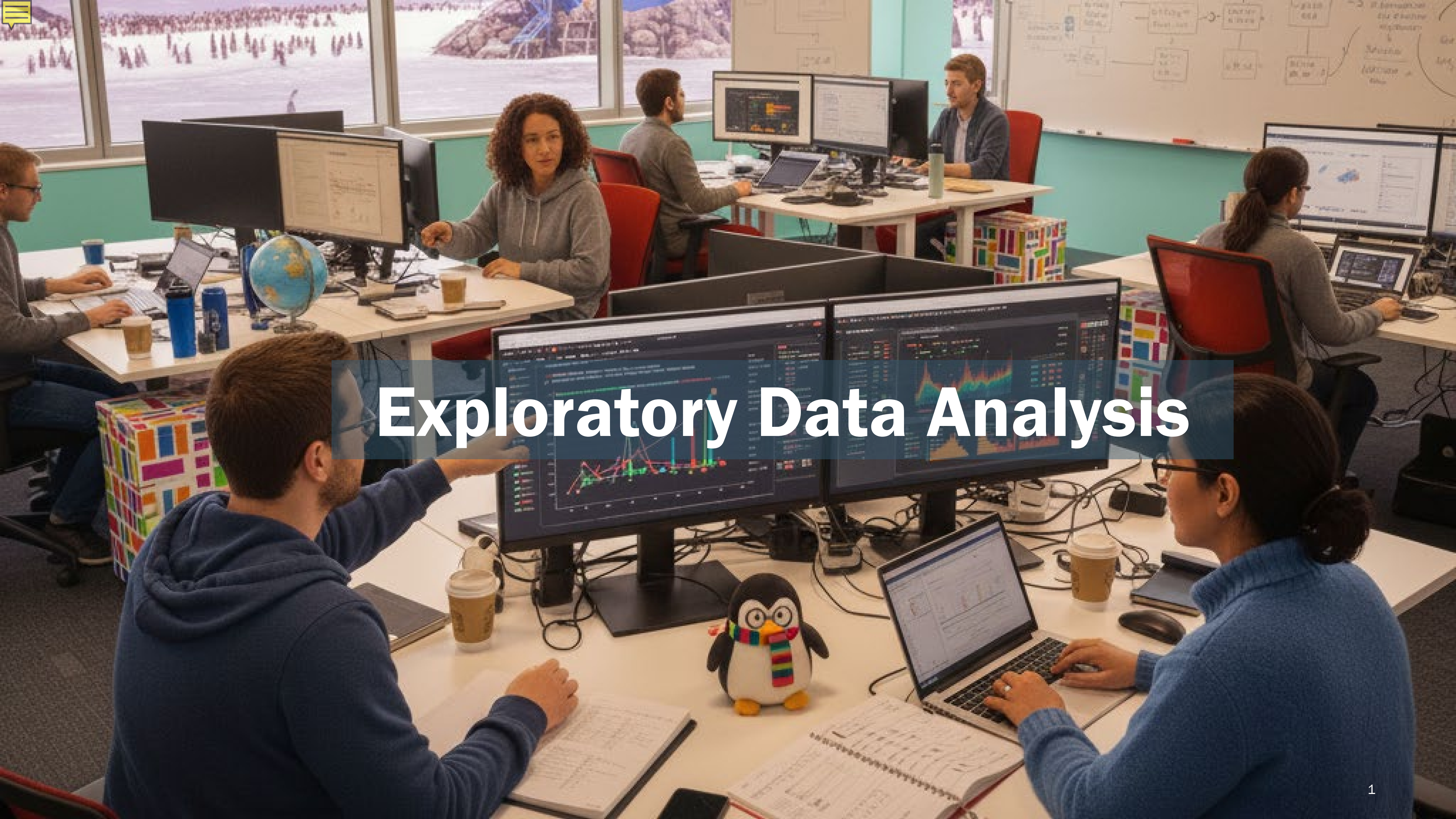
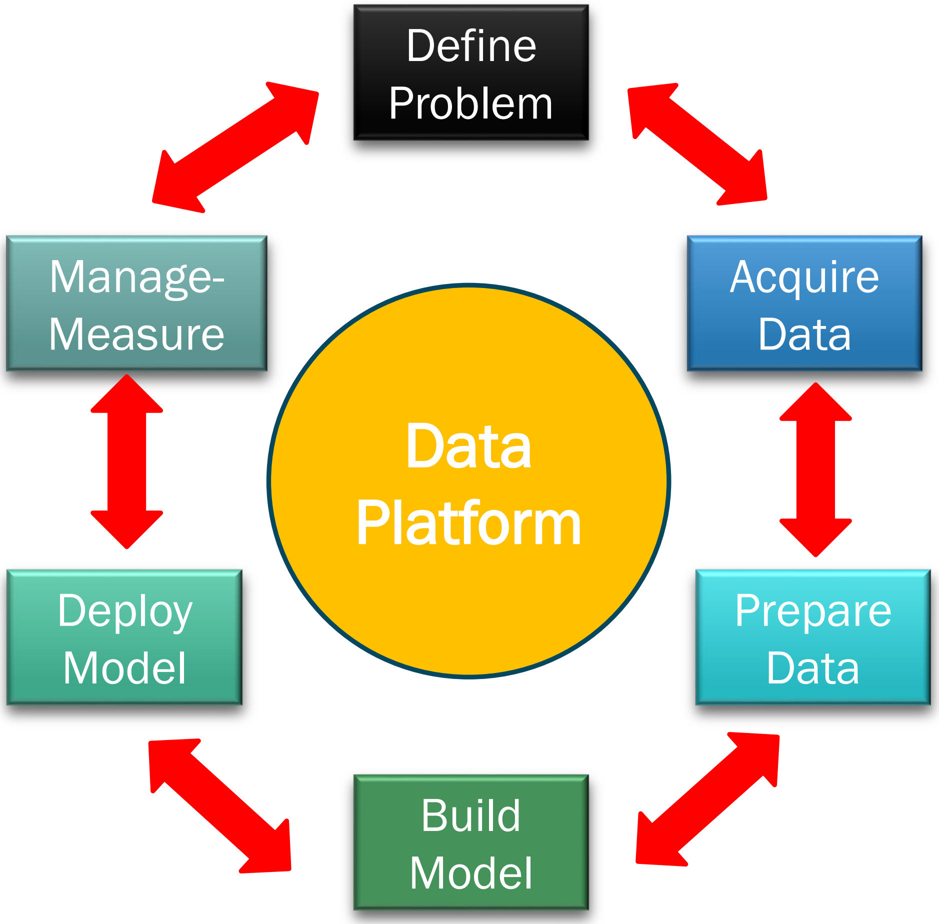


Exploratory Data Analysis

DATA SCIENCE LIFECYCLE



- Identify Data Location
- Manage Metadata
- Extract, Transform, Load
- Join/Merge Tables
- Profile Data
- Data Type and Validation Check
- Regulatory/PII Check

- Clean Data (Missing Values, Outliers, Duplicates, Var names)
- EDA (Univariate, Bivariate, Multivariate analysis, Visualization)
- Feature Engineering
- Prepare Analytic Data Set (ADS) (Standardization/Normalization, Category encoding)

BYU

WHAT IS THE DIFFERENCE BETWEEN DATA PREP AND EDA?

🔑 Takeaway:

- They highly overlap
- You need both
- In practice, you combine the approaches
- Data Preparation is about fixing (cleaning & structuring).
- EDA is about exploring (diagnosis, discovery, hypotheses).

Aspect	Data Preparation	Exploratory Data Analysis (EDA)
Primary Goal	Clean and transform data to make it ready for modeling	Understand the data, discover patterns, generate insights
Focus	Cleaning, formatting, and transformation	Analysis, visualization, hypothesis generation
Typical Questions	*Are there missing values? Are formats consistent? Are features encoded properly?*	*What does the distribution look like? Are variables correlated? Are there anomalies?*
Key Activities	- Handling missing data - Encoding categorical variables - Normalization/standardization - Removing duplicates & outliers - Data type conversions - Train/test splitting	- Summary statistics (mean, median, variance) - Visualizations (histograms, scatterplots, heatmaps) - Detecting trends, clusters, anomalies - Correlation analysis
Output	Clean, well-structured dataset ready for modeling	Insights, hypotheses, understanding of data structure and relationships
When Used	After or alongside EDA, before modeling	Early stage of analysis, often iteratively with preparation
Tools/Methods	Data wrangling libraries (Pandas, NumPy, SQL), preprocessing functions (scikit-learn, PySpark, AWS Glue)	Visualization libraries (Matplotlib, Seaborn, Plotly), statistical summaries, hypothesis tests

WHAT IS EXPLORATORY DATA ANALYSIS?

- **Definition:** Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. It's about understanding the data and finding insights.
- **Goal:** To uncover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.
- **Analogy:** Think of it like being a detective. Before you solve the case (build a model), you need to investigate the scene (your data) thoroughly.

WHY IS EDA IMPORTANT?

- Before Modeling: EDA is a critical step before building predictive models. A good model on bad data is still a bad model!
- **Pattern Discovery:** Reveals underlying structures, patterns, and relationships within the data.
- **Hypothesis Generation:** Helps formulate hypotheses about the data that can be tested later.
- **Feature Engineering:** Informs decisions about how to create new features.
- **Communication:** Provides a clearer understanding to communicate insights to stakeholders.
- Many times, EDA is the objective of the project.

THE DATA PREP / EDA WORKFLOW

Process 4.5.1: Exploratory data analysis.

Step	Description
Step 1: Understand the data	This is done as part of the ”Profile Data” step before analysis. Objective is to understanding the data types, numbers, ranges, overall cleanliness
Step 2: Detect and address Outliers and missing data	Data Cleaning stage. Data needs to be cleaned before analysis; otherwise, analysis could be skewed by dirty data.
Step 3: Describe the shape of each feature of the data.	Use descriptive univariate statistics and visualization to characterize data distributions for each feature.
Step 4: Identify and address correlations between features	(Bivariate and Multivariate Analysis) Assess whether features with high (+/-) correlations can be dropped.
Step 5: Hypothesize	Formulate questions and potential answers based on observations.

BYU

MISSING DATA

"**Missing Completely at Random" (MCAR)** refers to where the missingness occurs purely by chance and has no relationship with any of the variables in the dataset, either the missing ones or the observed ones.

Imagine a survey where respondents are asked to provide their age and income, and some survey forms are accidentally lost due to a clerical error.

"**Missing at Random" (MAR)** refers to where the missingness can be explained by other known (observed) variables in the dataset, but not by the value of the missing variable.

Survey respondents are asked to report their income. Some respondents might not report their income, and the likelihood of not reporting income could depend on their education level (observed) and not on the actual income itself (missing).

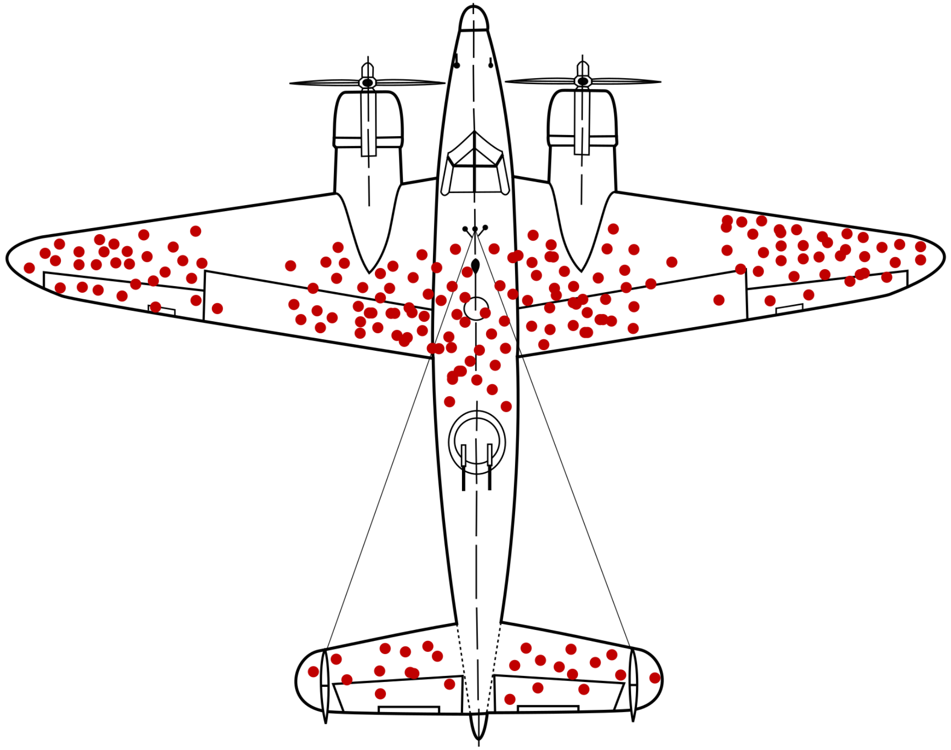
Missing Not at Random" (MNAR) refers to when the missingness is not random and cannot be explained by the observed data. Instead, it is systematically related to the unobserved, missing values.

In contrast, if higher-income individuals are systematically less likely to report their income regardless of other variables, the data would be "Missing Not at Random" (MNAR).

MISSING DATA MORE EXAMPLES

“Missing Completely at Random” (MCAR) refers to where the missingness occurs purely b relationship with any of the vari the missing ones or the observe

A lab instrument randomly fails to record , regardless of patient age, lesterol.



“Missing at Random” (MAR) re missingness can be explained b variables in the dataset, but not variable.

<35 years) are less likely to get red, but given their age, related to the actual cholesterol

Missing Not at Random” (MNA missingness is not random and cannot be explained by the observed data. Instead, it is systematically related to the unobserved, missing values.

high cholesterol hide their results, so missingness is directly tied to high cholesterol levels.

MISSING DATA--WHAT YOU DO DEPENDS ON THE TYPE OF MISSING

- MCAR → scattered randomly. Safe to drop missing rows (unbiased).
- MAR → missingness related to observed variable (e.g., Age). Use imputation or models that adjust for related variables.
- MNAR → missingness related to the unobserved (missing) variable itself (e.g., high cholesterol). Requires explicit modeling of the missingness mechanism.
  - [This is straightforward with the WWII airplane example, but can be very complex otherwise.]

MISSING DATA QUIZ 1

1. Which of the following best defines “Missing Completely at Random” (MCAR)?

a) The missingness is related to both observed and missing data.
b) The missingness is related to observed data but not missing data.
c) The missingness is unrelated to both observed and missing data.
d) The missingness is related to the missing data only.

2. In which of the following scenarios is data considered “Missing Not at Random” (MNAR)?

a) Data is missing due to a random system failure.
b) Patients with higher incomes are less likely to report their income.
c) Missingness is related to other observed variables, like age or gender.
d) Missing data occurs randomly across the dataset.

3. Which type of missing data is easiest to handle without introducing bias in analyses?

a) Missing Not at Random (MNAR)
b) Missing Completely at Random (MCAR)
c) Missing at Random (MAR)
d) None of the above

4. Which of the following is true about “Missing at Random” (MAR)?

a) Missingness is explained by variables that are not observed.
b) Missingness is random and unrelated to any variables in the dataset.
c) Missingness is related to observed variables but not the missing data itself.
d) Missingness is dependent on the value of the missing data.

MISSING DATA QUIZ 2

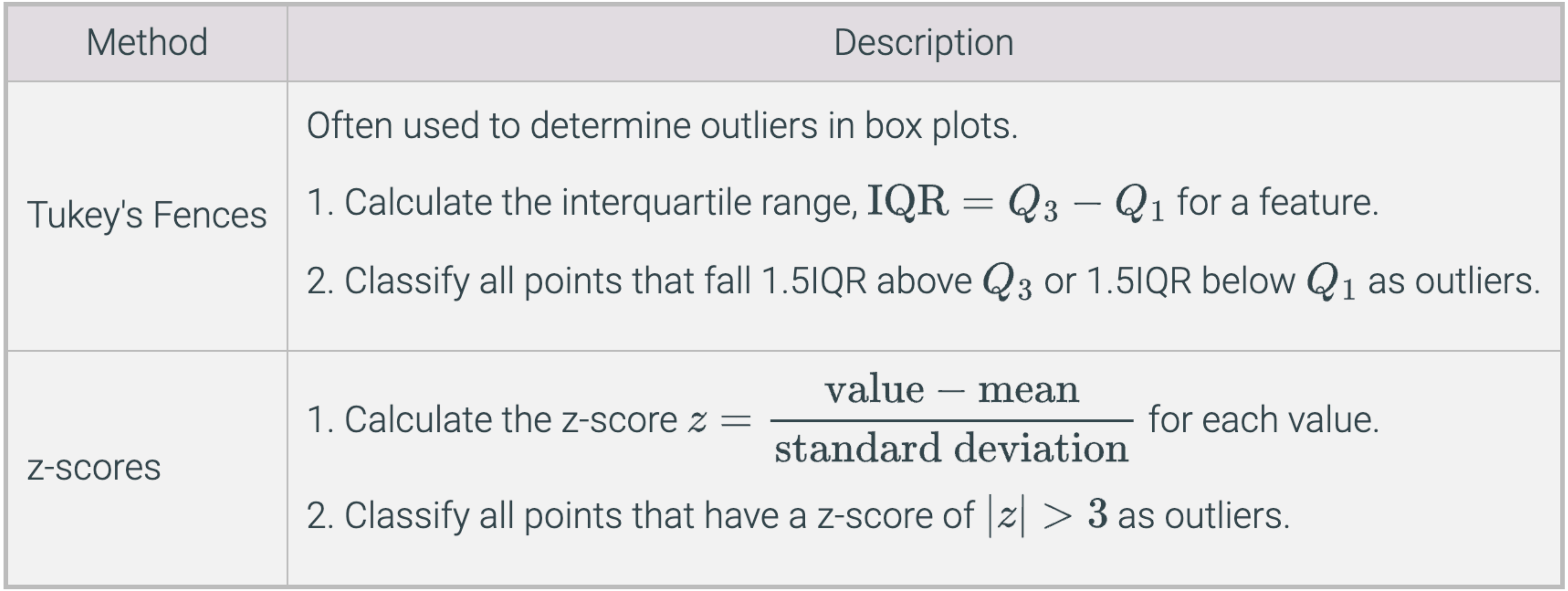
5. Which of the following techniques can help handle data that is “Missing at Random” (MAR)?
a) Removing all missing data rows.
b) Imputing missing values based on observed variables.
c) Ignoring the missing data and proceeding with the analysis.
d) Using the mean of the missing variable to fill in the gaps.

6. If data is Missing Completely at Random (MCAR), removing missing data points will not introduce bias.
(True / False)

7. Data that is Missing Not at Random (MNAR) can be safely ignored without any effect on the analysis.
(True / False)

8. Missing at Random (MAR) means the missingness is related to the unobserved, missing data itself.
(True / False)

OUTLIER DETECTION



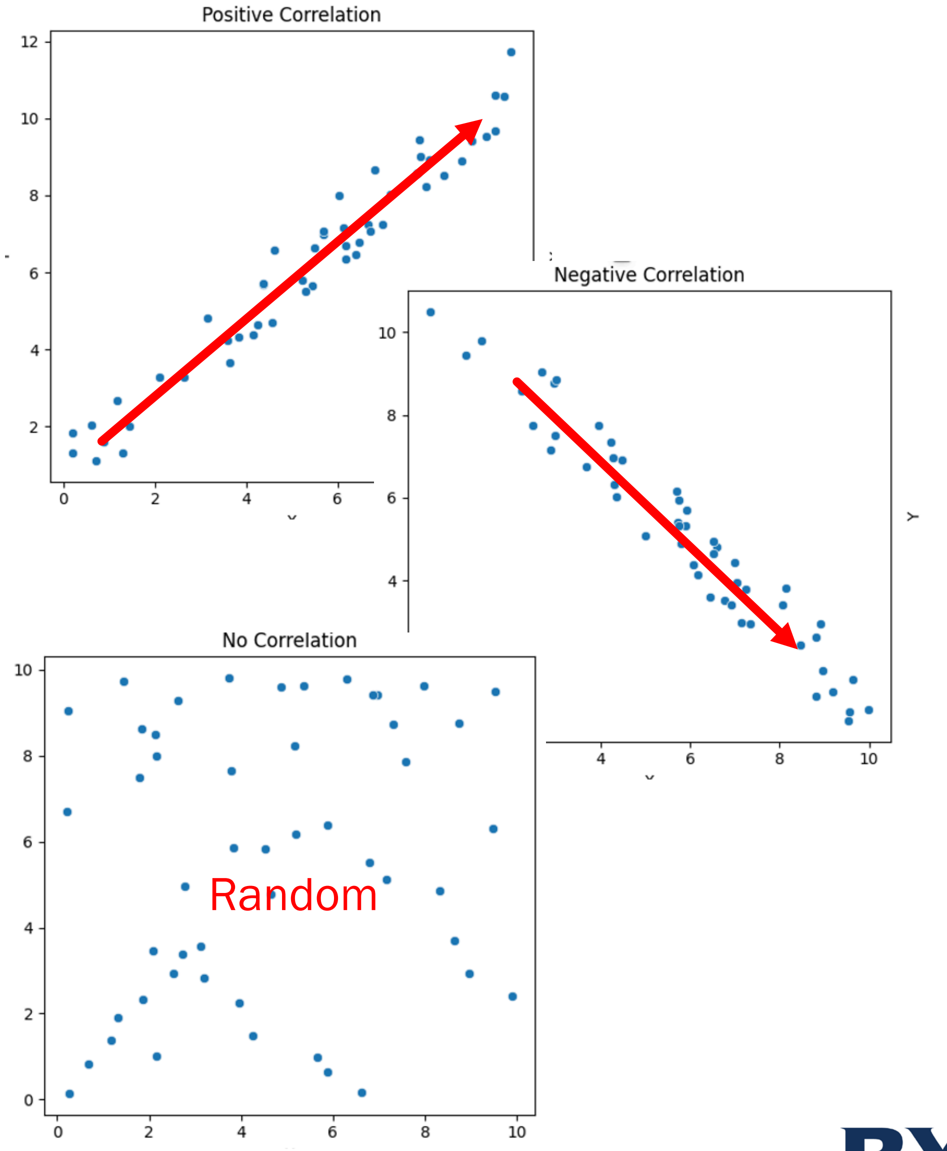
Method	Description
Tukey's Fences	Often used to determine outliers in box plots. 1. Calculate the interquartile range, $\mathrm{IQR} = Q_3 - Q_1$ for a feature. 2. Classify all points that fall 1.5IQR above Q_3 or 1.5IQR below Q_1 as outliers.
z-scores	1. Calculate the z-score $z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$ for each value. 2. Classify all points that have a z-score of $\|z\| > 3$ as outliers.

DEALING WITH OUTLIERS

Cause of outlier	Description of cause	Method of cleaning
Data entry error	Human error entering a value into the dataset	If the intended value can be inferred, correct the value.
Measurement error	Error in an instrument recording the value	If the error is consistently biased from the actual value and can be identified, remove the bias.
Data processing error	Accident in data manipulation	If feasible, redo the manipulation to remove the error.
Sampling error	Error in combining data from multiple populations or groups	Explore dataset for outliers that may be from different populations or groups.
Natural outlier	Not an error. This value, though extreme, belongs in the data.	Models should be checked for stability of results with and without the outlier.

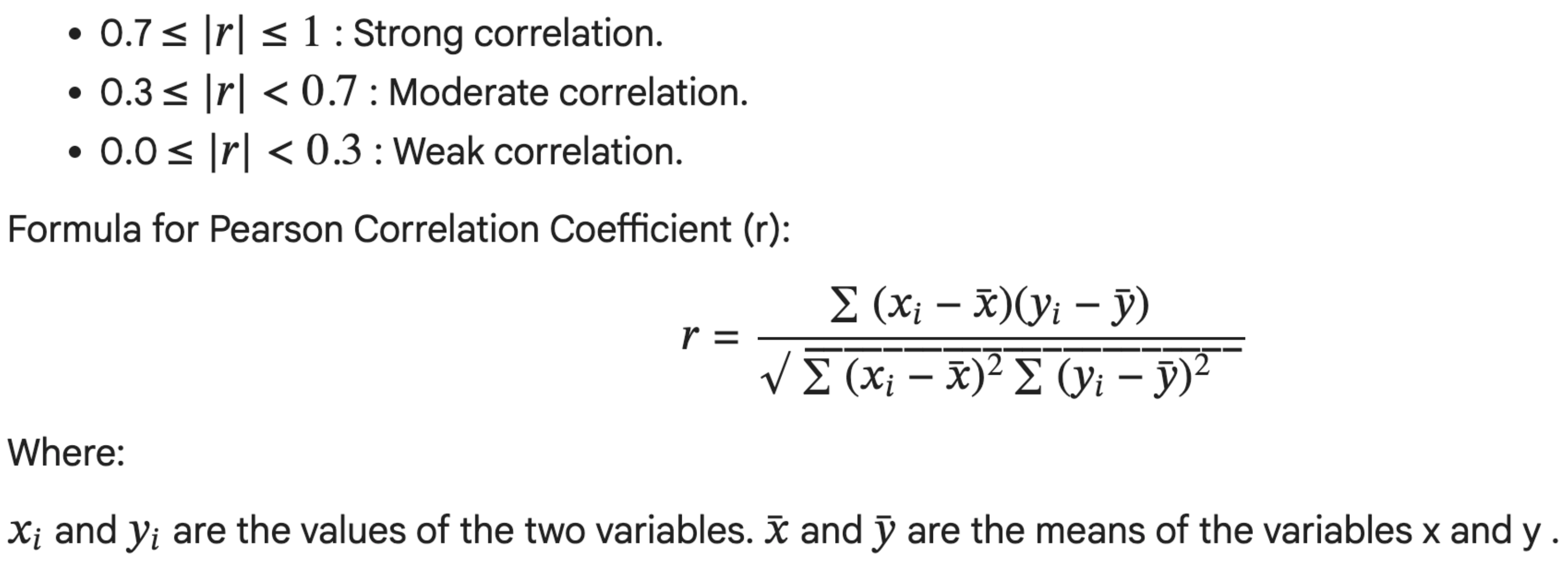
CORRELATION

- Correlation is a **statistical measure** that describes the degree to which two variables move in relation to each other.
- If **one variable increases while the other tends to increase**, they are **positively correlated.**
- If **one increases while the other decreases,** they are **negatively correlated.**
- The correlation is close to zero if they don’t move together systematically.



PEARSON CORRELATION

- The most common type of correlation measure is the **Pearson correlation coefficient (r)**, which ranges from **-1 to +1**.
- Interpretation of Pearson Correlation Coefficient (r):
  - r = 1 : Perfect positive linear relationship. As one variable increases, the other variable increases proportionally.
  - r = -1 : Perfect negative linear relationship. As one variable increases, the other decreases proportionally.
  - r = 0 : No linear relationship between the variables.



- 0.7 ≤ $|r| \leq 1$: Strong correlation.
- 0.3 ≤ $|r| < 0.7$: Moderate correlation.
- 0.0 ≤ $|r| < 0.3$: Weak correlation.

Formula for Pearson Correlation Coefficient (r):

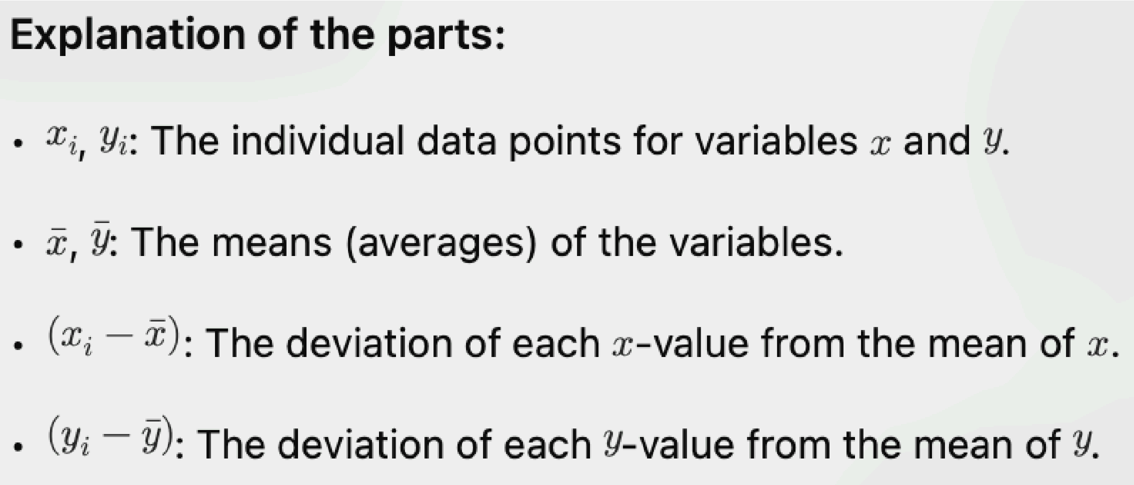
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

x_i and y_i are the values of the two variables. $\bar{x}$ and $\bar{y}$ are the means of the variables x and y .

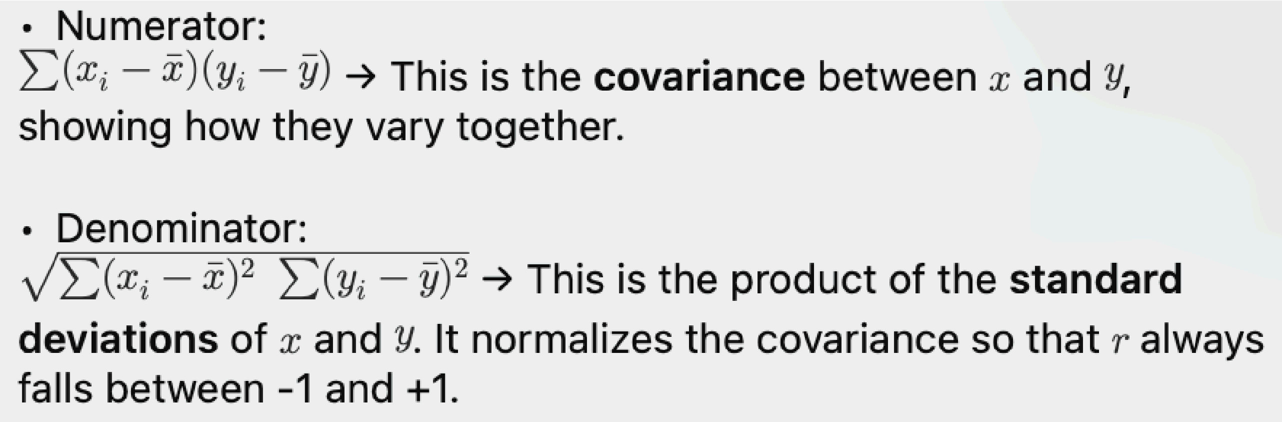
PEARSON CORRELATION INTERPRETATION

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



Explanation of the parts:

- x_i, y_i: The individual data points for variables x and y.
- $\bar{x}$, $\bar{y}$: The means (averages) of the variables.
- $(x_i - \bar{x})$: The deviation of each x-value from the mean of x.
- $(y_i - \bar{y})$: The deviation of each y-value from the mean of y.



- Numerator:

$\sum(x_i - \bar{x})(y_i - \bar{y})$ → This is the **covariance** between x and y, showing how they vary together.

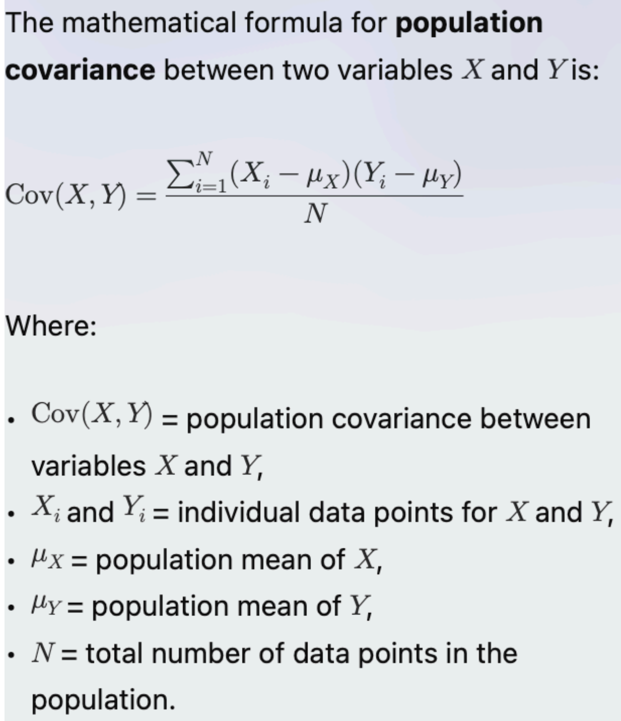
- Denominator:

$\sqrt{\sum(x_i - \bar{x})^2 \ \sum(y_i - \bar{y})^2}$ → This is the product of the **standard deviations** of x and y. It normalizes the covariance so that r always falls between -1 and +1.

BYU

COVARIANCE VS CORRELATION

Covariance
Covariance measures how the deviation of one variable from its mean is related to the deviation of another variable from its mean
(-∞, +∞)



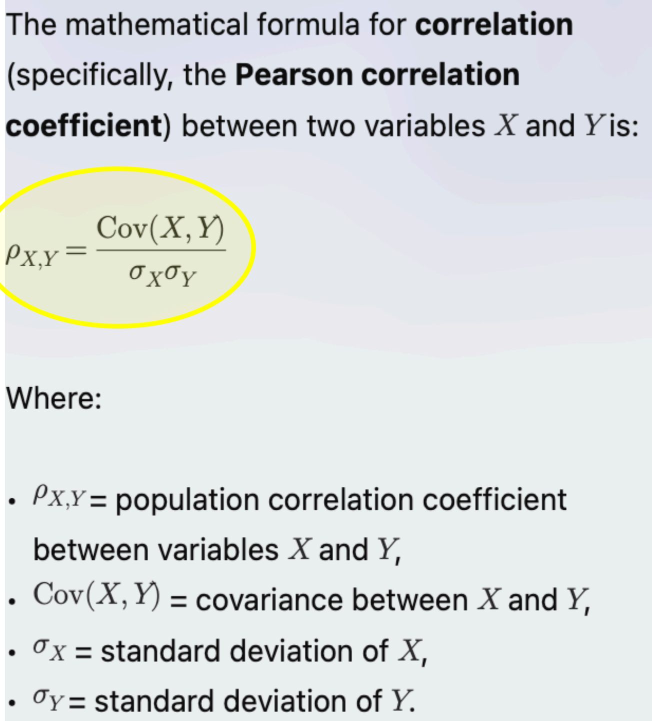
The mathematical formula for **population covariance** between two variables X and Y is:

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Where:

- $\mathrm{Cov}(X,Y)$ = population covariance between variables X and Y,
- X_i and Y_i = individual data points for X and Y,
- μ_X = population mean of X,
- μ_Y = population mean of Y,
- N = total number of data points in the population.

Correlation
Correlation measures how strongly the two variables are related to each other
[-1, 1]



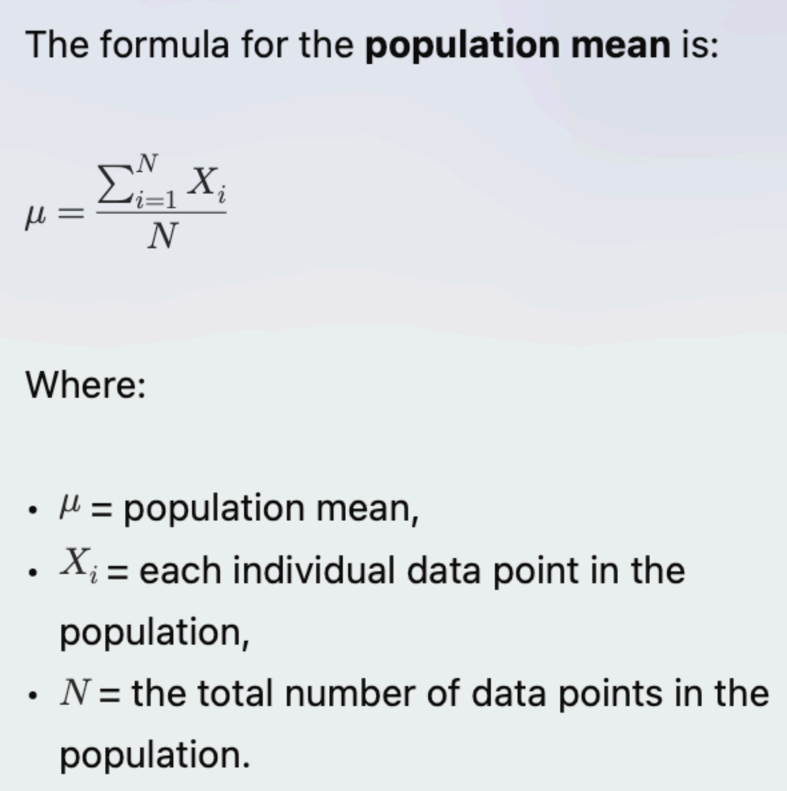
The mathematical formula for **correlation** (specifically, the **Pearson correlation coefficient**) between two variables X and Y is:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where:

- $\rho_{X,Y}$ = population correlation coefficient between variables X and Y,
- $\mathrm{Cov}(X,Y)$ = covariance between X and Y,
- σ_X = standard deviation of X,
- σ_Y = standard deviation of Y.

MEAN, VARIANCE, STANDARD DEVIATION

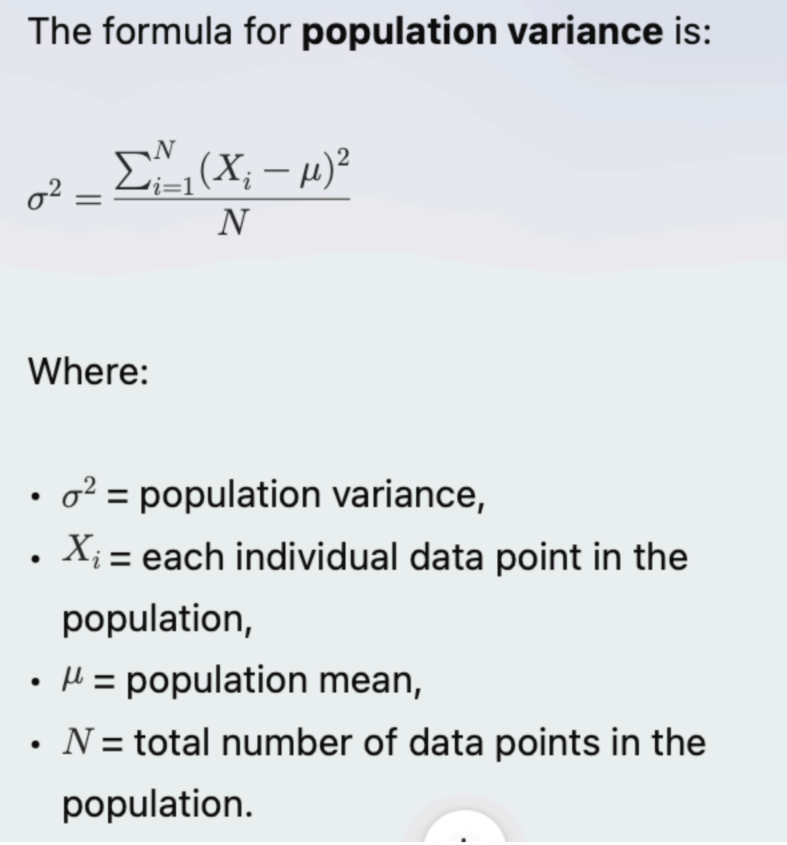


The formula for the **population mean** is:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Where:

- μ = population mean,
- X_i = each individual data point in the population,
- N = the total number of data points in the population.

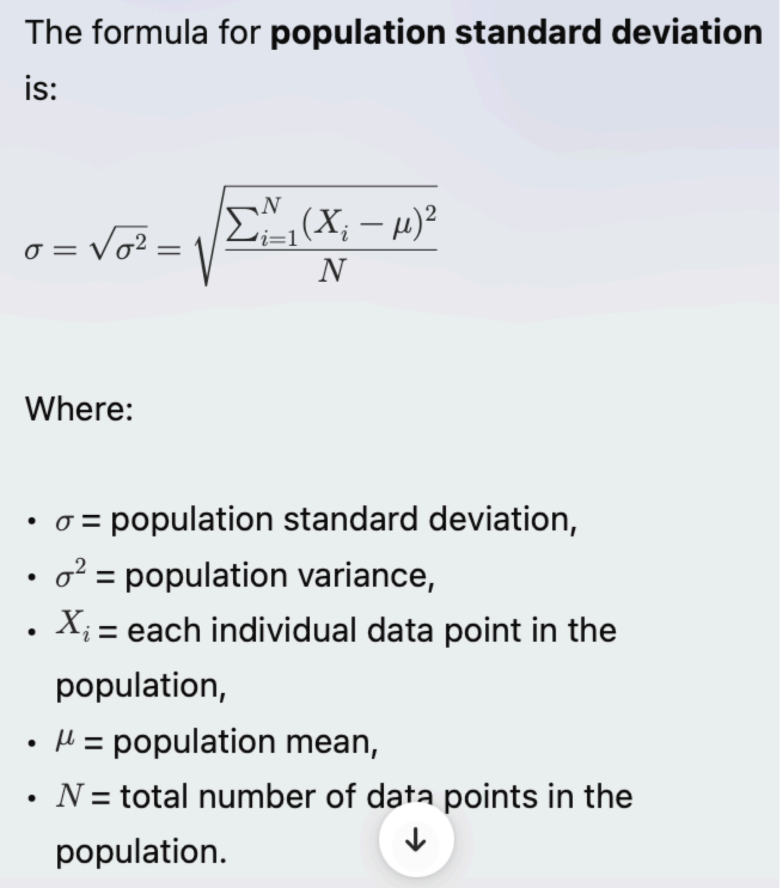


The formula for **population variance** is:

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

Where:

- σ^2 = population variance,
- X_i = each individual data point in the population,
- μ = population mean,
- N = total number of data points in the population.



The formula for **population standard deviation** is:

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

Where:

- σ = population standard deviation,
- σ^2 = population variance,
- X_i = each individual data point in the population,
- μ = population mean,
- N = total number of data points in the population.

WHY IS CORRELATION IMPORTANT IN EDA?

During EDA, correlation helps you understand **relationships between variables**, which is crucial for:

- Feature Selection
  - Strongly correlated features may provide **redundant information.**
  - Detecting this helps avoid multicollinearity (important in regression models).
- Understanding Patterns
  - Correlations reveal how variables interact.
  - For example, in sales data, *advertising spend* may strongly correlate with *revenue*.
- Detecting Potential Causation Candidates
  - Correlation doesn’t prove causation, but it highlights variables worth deeper investigation.
- Dimensionality Reduction
  - Highly correlated variables can sometimes be combined or reduced with techniques like **PCA (Principal Component Analysis).**
- Identifying Data Issues
  - Unexpected correlations (or lack of correlation) may point to **data errors, outliers, or biases**.

BYU

HEATMAP OF DATA CORRELATIONS

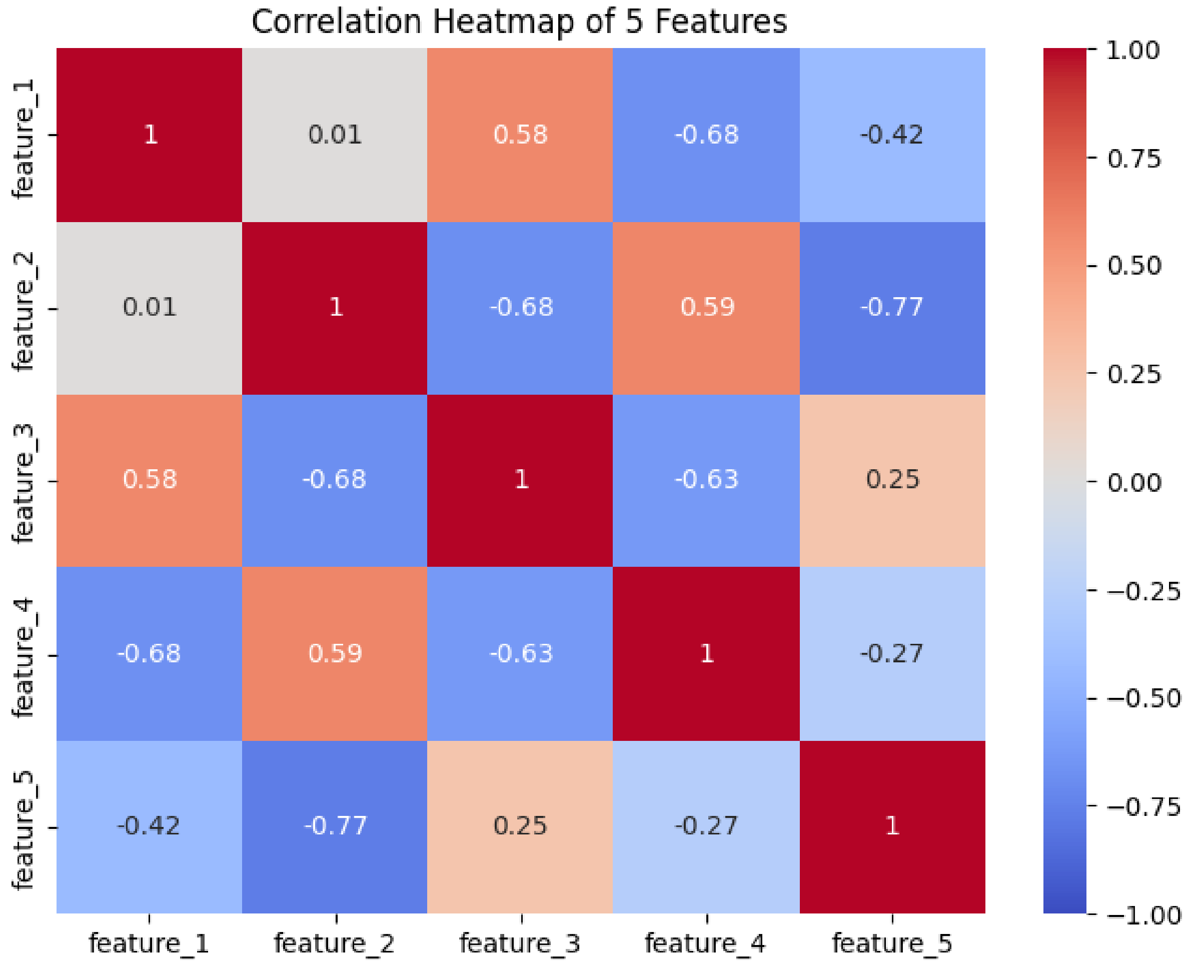
- A heatmap is an excellent way to visualize the correlation between multiple features.
- Using Seaborn, a Python visualization library built on top of Matplotlib, you can create a heatmap to display the correlation matrix between many features.

Steps:

1. Compute the correlation matrix between the features using pandas.

2. Plot the heatmap using Seaborn's heat map() function()

3. Colors indicate relative correlation

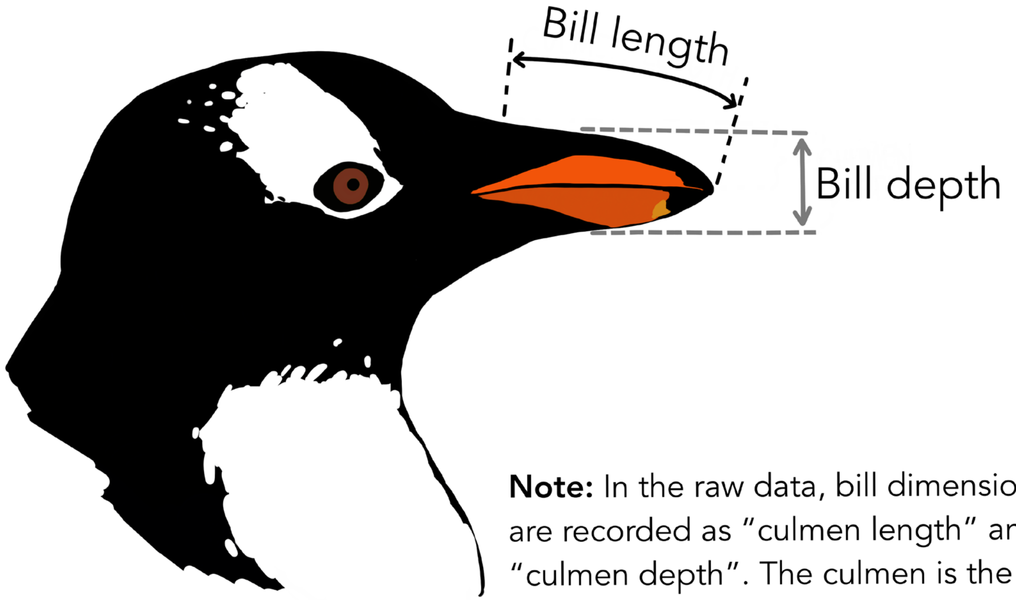


EDA CASE STUDY: PALMER PENGUINS

- **Scenario:** Imagine we are data scientists studying penguin populations. We have a dataset containing various measurements.
- **Goal of EDA:** To understand the characteristics of different penguin species, how they vary across islands, and potential factors influencing their size and dimensions.
- This understanding will inform ecological studies, conservation efforts, or even predictive models for species identification.
- **Data collectors:** Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program.
- **Time frame:** Measurements were taken from 2007–2009.
- **Location:** Palmer Archipelago, Antarctica.

PALMER PENGUINS DATASET VARIABLES

- **species**: Adélie, Chinstrap, and Gentoo.
- **island**: Biscoe, Dream, and Torgersen.
- **bill_length_mm**: Length of the penguin's culmen (bill).
- **bill_depth_mm**: Depth of the penguin's culmen.
- **flipper_length_mm**: Length of the penguin's flipper.
- **body_mass_g**: The penguin's body mass.
- **sex**: Male or female.



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

BYU

HANDS-ON ACTIVITY: EXPLORE YOUR PENGUINS!

Instructions:

- **Load the data:** Ensure you have the penguins_cleaned DataFrame from our earlier steps.
- **Task 1: Summarize 'island'**
  - Calculate the frequency of penguins on each island.
  - Create a bar plot of island counts.
- **Task 2: Visualize 'flipper_length_mm' vs. 'body_mass_g'**
  - Create a scatter plot of flipper_length_mm vs. body_mass_g.
  - Color the points by sex.
- **Task 3: Formulate a Hypothesis**
  - Based on your plots from Task 1 and 2, write down one observation or a simple hypothesis (e.g., "Penguins on Island X tend to be larger").

WHAT WE CAN LEARN

- Key takeaways from the dataset

- **Summary:**
  - The three penguin species in the study show distinct patterns in their physical characteristics (size and bill shape).
  - The data can be effectively used for exploratory analysis, visualization, and training simple machine learning models.
  - The project highlights the value of accessible and approachable datasets for educational purposes.

- **Multiple Choice:** Which of these Pandas functions is best for getting a quick summary of descriptive statistics (mean, std, min, max) for numerical columns?
  - a) df.info()
  - b) df.describe()
  - c) df.head()
  - d) df.isnull().sum()
- **True/False:** A high correlation coefficient (e.g., 0.9) between two variables always implies that one causes the other.
- **Short Answer:** Why might you choose to use df.fillna(df['column'].median()) instead of df.fillna(df['column'].mean()) for missing numerical data?
- **Multiple Choice:** If you want to visualize the distribution of a single continuous numerical variable, which plot is most appropriate?
  - a) Scatter Plot
  - b) Bar Chart
  - c) Histogram
  - d) Heatmap