**DATA SCIENCE ASSIGNMENT-1**

**Student id-(16352686)**

**Task 1**

**Reproducible Workflow for Frailty Data (5 Points)**

**Introduction**

Frailty is a condition characterized by physical weakness and a lack of strength. Research shows that reduced grip strength in females correlates with higher frailty scores. The given dataset contains 10 female participants' measurements of height, weight, age, grip strength, and frailty status. To ensure a **reproducible workflow**, we implement a structured process consisting of three key stages: **Data Collection & Storage, Data Processing & Cleaning, and Data Analysis & Visualization.**

**1. Data Collection & Storage (Raw Data Stage)**

**Objective:**

* Collect and store raw data in a structured format.
* Ensure the dataset is backed up and easily accessible.

**Actions Taken:**

* The dataset is saved in **CSV format** for easy processing.
* It is stored in the data/raw/ directory within the project structure.

**Folder Structure:**

project/

│── data/

│ ├── raw/ → (store original frailty data file, e.g., frailty\_data.csv)

**2. Data Processing & Cleaning**

**Objective:**

* Ensure the dataset is clean, structured, and ready for analysis.
* Convert categorical variables to numerical form.

**Actions Taken:**

* Load the raw dataset into **Pandas**.
* Convert the Frailty column from categorical (Y/N) to numerical (1/0).
* Check for missing values and inconsistencies.
* Save the cleaned data in the processed/ folder.

**Folder Structure:**

project/

│── data/

│ ├── processed/ → (store cleaned frailty dataset, e.g., frailty\_clean.csv)

│── src/ → (Python scripts for data processing)

**Python Code for Data Cleaning:**

import pandas as pd

# Load raw data

df = pd.read\_csv("project/data/raw/frailty\_data.csv")

# Convert Frailty column to numerical values (Y=1, N=0)

df['Frailty'] = df['Frailty'].map({'Y': 1, 'N': 0})

# Save cleaned dataset

df.to\_csv("project/data/processed/frailty\_clean.csv", index=False)

print("Data cleaning complete!")

**3. Data Analysis & Visualization**

**Objective:**

* Identify patterns and correlations in the data.
* Visualize key insights using plots and summary statistics.

**Actions Taken:**

* Compute **descriptive statistics** such as mean, median, and correlation.
* Create **visualizations** (e.g., grip strength vs. frailty).
* Store results in the results/ folder and images in figures/.
* Document key findings in reports/.

**Folder Structure:**

project/

│── figures/ → (store visualizations, e.g., grip\_strength\_vs\_frailty.png)

│── results/ → (store statistical summaries)

│── reports/ → (final research report)

**Python Code for Analysis & Visualization:**

import matplotlib.pyplot as plt

import seaborn as sns

# Load processed data

df = pd.read\_csv("project/data/processed/frailty\_clean.csv")

# Scatter Plot: Grip Strength vs. Age

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

sns.scatterplot(x=df['Grip strength'], y=df['Age'], hue=df['Frailty'])

plt.title("Grip Strength vs Age (Colored by Frailty)")

plt.savefig("project/figures/grip\_strength\_vs\_age.png")

plt.show()

**Conclusion**

A structured, reproducible workflow is crucial for reliable research. By **storing, cleaning, and analyzing the frailty dataset systematically**, we ensure that results can be validated and reproduced by others. This workflow provides a strong foundation for further studies on frailty and grip strength correlations.

**Task 2**

**Explanation of Data Visualizations (Student Performance Dataset)**

**1. Histogram - Math Score Distribution**

**Purpose:**

* The histogram helps visualize the distribution of math scores among students.
* It shows how frequently different score ranges occur.
* The addition of a Kernel Density Estimate (KDE) curve makes it easier to observe the overall trend.

**Analysis Insight:**

* Helps identify whether the scores are **normally distributed, skewed, or have multiple peaks**.
* Useful for detecting **score concentration** (e.g., most students scoring between 60-80).
* Shows the **spread and variability** of math performance.

**2. Boxplot - Gender vs. Math Score**

**Purpose:**

* A boxplot provides a summary of the distribution of math scores across male and female students.
* It highlights **median scores, interquartile ranges (IQR), and potential outliers**.

**Analysis Insight:**

* Helps **compare male and female performance** in math.
* Shows whether one gender has **higher median scores** than the other.
* Identifies outliers (e.g., students with exceptionally high or low scores).

**3. Scatter Plot - Reading Score vs. Writing Score**

**Purpose:**

* A scatter plot helps visualize the **relationship between reading and writing scores**.
* Each point represents a student, where the x-axis is the reading score and the y-axis is the writing score.

**Analysis Insight:**

* Helps **determine correlation** between reading and writing skills.
* If points form a **clear upward trend**, it indicates that students who perform well in reading also perform well in writing.
* Can reveal **clusters of high- and low-performing students**.

**4. Bar Chart - Average Score per Subject**

**Purpose:**

* This bar chart shows the **average scores** in Math, Reading, and Writing.
* Helps compare subject-wise performance.

**Analysis Insight:**

* Reveals **which subject students perform best in on average**.
* Can highlight potential weaknesses (e.g., if math scores are significantly lower than reading/writing scores, it may suggest the need for intervention).
* Helps educators understand **subject-wise strengths and weaknesses**.

**5. Pie Chart - Test Preparation Course Distribution**

**Purpose:**

* A pie chart illustrates the proportion of students who **completed a test preparation course** versus those who did not.

**Analysis Insight:**

* Helps understand **how many students took additional preparation**.
* Can be used to compare **test preparation effectiveness** (e.g., checking if students who took the course scored higher on average).
* Provides an overview of how many students are actively engaging in extra learning resources.

**Folder Structure for Task 2**

To maintain consistency with Task 1, we structure our project as follows:

project/

│── data/

│ ├── raw/ → (store original student performance dataset, e.g., student\_performance.csv)

│ ├── processed/ → (store cleaned student performance dataset, e.g., student\_performance\_cleaned.csv)

│── figures/ → (store generated visualizations)

│── results/ → (store statistical summaries and analysis outputs)

│── reports/ → (store final analysis reports)

│── src/ → (Python scripts for processing and visualization)

This structure ensures data integrity, reproducibility, and organized project management.

**Conclusion**

These visualizations allow for **better understanding of student performance trends**. Educators can use this analysis to:

* Identify **subject-wise strengths and weaknesses**.
* Determine if **gender disparities exist in performance**.
* Explore the **impact of test preparation courses**.
* Assess the **correlation between different subject scores**.

These insights can guide **curriculum planning, targeted interventions, and overall academic improvements**.