**Assignment 1**

1. 1) Frailty is physical weakness; lack of health or strength. Reduced grip strength in females correlated with higher frailty scores and vice versa. Hand grip strength can be quantified by measuring the amount of static force that the hand can squeeze around a dynamometer. The force has most commonly been measured in kilograms and pounds. The table below represents data from 10 female participants. The Height is measured in inches, Weight in pounds, Age in years, Grip strength in kilograms. Frailty is qualitative attribute indicated the presence or absence of the symptoms. (10 points)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Height | Weight | Age | Grip strength | Frailty |
| 65.8 | 112 | 30 | 30 | N |
| 71.5 | 136 | 19 | 31 | N |
| 69.4 | 153 | 45 | 29 | N |
| 68.2 | 142 | 22 | 28 | Y |
| 67.8 | 144 | 29 | 24 | Y |
| 68.7 | 123 | 50 | 26 | N |
| 69.8 | 141 | 51 | 22 | Y |
| 70.1 | 136 | 23 | 20 | Y |
| 67.9 | 112 | 17 | 19 | N |
| 66.8 | 120 | 39 | 31 | N |

Based on the following table, **you must design AND implement a three‑stage workflow (ingest → process → analyze) with code and organized outputs.** (reference study case in chapter 3). You need to save the raw data in csv file and read it into a pandas data frame and then perform the following:

* 1. Unit standardization
     1. Height\_m = Height\_in \* 0.0254
     2. Weight\_kg = Weight\_lb \* 0.45359237
  2. Feature engineering
     1. BMI = Weight\_kg / (Height\_m \*\* 2) (round to 2 decimals).
     2. AgeGroup (categorical): "<30", "30–45", "46–60", ">60" based on Age\_yr.
  3. Categorical → numeric encoding
     1. Binary encoding: Frailty\_binary (Y→1, N→0, store as int8).
     2. One‑hot encode AgeGroup into columns: AgeGroup\_<30, AgeGroup\_30–45, AgeGroup\_46–60, AgeGroup\_>60
  4. EDA & Reporting

1. Compute summary table: mean/median/std for numeric columns; save to reports/findings.md .
2. Quantify relation of strength ↔ frailty: compute correlation between Grip\_kg and Frailty\_binary, and report it.

2) Perform the 5 data visualization tasks (given below) on the student performance dataset given in the link below. Each figure: 800×600 px, 300 DPI, title, labeled axes/units, legend if applicable, readable ticks. For each, add a 5–8 sentence interpretation to reports. Before the visualizations (analysis step) you need to perform ingestion stage and preprocessing step (missing values etc.) (10 points).

1. V1 — Gender boxplots (math vs reading) *(2 pts)*
   1. Question: Are there gender differences in math vs reading?
   2. Chart: Side‑by‑side boxplots of math score and reading score grouped by gender.
2. V2 — Test prep impact on math *(2 pts)*
   1. Question: Do students who completed test prep score higher in math?
   2. Chart: Any chart of your choice for math score by test preparation course (completed vs none).
3. V3 — Lunch type and average performance *(2 pts)*
   1. Question: Does lunch type (standard vs free/reduced) relate to outcomes?
   2. Chart: Grouped bar chart of mean overall\_avg of all the scores (math, reading, writing) by lunch.
4. V4 — Subject correlations *(2 pts)*
   1. Question: How strongly do the three subjects move together?
   2. Chart: Correlation heatmap for math, reading, writing with annotated coefficients.
5. V5 — Math vs reading with trend lines by test prep *(2 pts)*
   1. Question: How strongly are math and reading scores associated, and do students who completed the test‑preparation course have a *different slope* in the math–reading relationship than those who did not?
   2. Chart: Scatter plot with two straight best‑fit lines (one for each group: completed, none).
      1. X‑axis: reading score
      2. Y‑axis: math score
   3. Color: Points colored by test preparation course (legend must show the two groups and each group’s n).

Data link: <https://app.box.com/s/ji910ez3ycw137rw07xnhielxey7ww41>

Submission:

Create a public GitHub repo and upload the folders for both the questions on the GitHub and submit the link to Canvas.