**Assignment 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. Based on the following table, design the three stages of reproducible workflow, includes the work you can do and the folder structure in each stage (reference study case in chapter 3).  (5 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 |

The three stages of reproducible workflow are **data collection, data cleaning, and data analysis**.

**The data collection** process for this dataset would either include surveying/measuring the features given (height,weight,age,grip strength, frailty) or retrieving from the source that provides this data if it’s a 3rd party that does the collection. This is all assuming the process was repeated for more than 10 female participants in the same structure as above.

**For data cleaning**, the first step should be to inspect the data. Check first to make sure the shape of the data matches. Are the number of columns and rows the same? Next check the validity of the values which are entered into the dataset. Are there any missing/null/invalid values? Are all the measurements in the same format and standardized? The best way to do this in a reproducible workflow would be to create an automated script to do these things to replicate the process.

Ideally the data would come in a csv file. Pandas is a great library for this in Python. Pandas can convert the csv data into a dataframe. From there Pandas has imported functions for checking not only the shape of the data but also for null values. One could also use it to check for values outside of a certain range for a particular feature or column. It would also be good to quantize the frailty values from ‘N’ & ‘Y’ to 0 & 1. This will make any analysis related to these categories much easier to do later on.

Once all of this is complete, it is important to have a location to either dump or save the edited dataset instance (without overwriting the original dataset).

**For data analysis**, first one must have an idea of what they want to be analyzed. Is there a target variable or some output being measured? What’s the significance or importance of the data? In this case it appears to be frailty. So, it might be good to do a linear regression of different values on frailty or visualize with a correlation matrix.

2) Perform 5 data visualization tasks on the student performance dataset given in the link below (create 5 different visualizations). Explain what kind analysis has become easier with each of the visualizations. Create the folder structure for this question similar to question 1. (15 points).

I want to make an important note that I chose to make the visualizations sequential to make it easy for the viewer to visualize and understand which categories performed best and worst. This is not to draw any conclusions about demographic data without further analysis.A graph of a bar chart

Description automatically generated with medium confidenceA graph of different colored bars

Description automatically generatedA graph of different colored vertical lines

Description automatically generated

A graph of bar graph

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

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.