**Assignment 1**

1) 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 (Inches) | Weight (Pounds) | 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 given dataset, the three stages of reproducible workflow can be designed as follows:

The first step involves Data extraction along with Data acquisition which is continuous process.

| |-- data\_raw

| | |-- data\_raw.csv

| | |-- README.txt

| |-- data\_cleaning

| |-- results

| |-- src

data\_raw <- read.csv(“study\_data.csv”)

The second step we are performing Data cleaning and Data processing which is combined form for processing.

* | |-- data\_raw
* | | |-- data\_raw.csv
* | | |-- README.txt
* | |-- data\_cleaning
* | | |-- data\_cleaning.csv
* | |-- results
* | |-- src
* | | |-- data\_clean.R

data\_raw <- read.csv(“study\_data.csv”)

data\_cleaning <- na.omit(data\_raw[data\_raw$Field != “N”,])

write.csv(data\_cleaning,”data\_cleaning.csv”)

The third step we do the Analysis part.

* | |-- data\_raw
* | | |-- data\_raw.csv
* | | |-- README.txt
* | |-- data\_cleaning
* | | |-- data\_cleaning.csv
* | |-- results
* | | |-- test\_results.txt
* | |-- src
* | | |-- analysis.R
* | | |-- data\_cleaning.R

total\_data <- t.test(data.cleaning$weight, data\_cleaning$Grip, paired = TRUE)

capture.output(total\_data, file = “test\_results.txt”)

I saved the output in the form of txt file.

So these are the stages for Reproducible workflow.

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).

The 5 data visualization tasks on the student performance dataset has been performed and below are the brief analysis of each plot diagram.

1. visualization Gender vs Scores

Here we can compare and analyze the data between Gender and Scores as we can see scores for females and male, the math score for female is less compared to male and reading and writing scores are higher when compared to male scores.

2. visualization Ethnicity vs Scores

Here Ethnicity are divided into different groups from A-E and compared and analyzed them with scores.

3. Race vs male/female:

Here we compared and analyzed the data of females and male with race/ethnicity

4. scatter plot to understand math scores and reading and correlated in any fashion from the plot we can refer that both reading and math scores have a high correlation

5. while training machine learning we generate a correlation matrix of all the variables involved so in this built the correlation matrix between all the score variables

6. box plot or candlestick is generally drawn to understand the disturbing of dataset or its subset in this case we are trying to understand how gender influences the avg test score