CS5530 Assignment 1- Question 1

Question 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)

Answer:
In order to create workflow and make analysis based on given data, I have divided stages to three parts including data collection, data processing and data analysis.

Stage 1 Data collection

|-- GripStrength_project

| |-- data_raw

| | |-- raw_gs_data.csv

| | |-- README.txt

| |--data_clean

| |-- results

| |-- src

- Data should be entered into a spreadsheet program and saved as a CSV file
- Once the file is created, it should be given a name and saved in a useful location (raw_gs_data.csv)
- Creating README.txt file which contains the metadata file. At the same time that data are saved, a metadata file should also be created and saved with it.

Stage 2 Data processing

|-- GripStrength_project

| |-- data_raw

| | |-- raw_gs_data.csv

| | |-- README.txt

| |--data_clean

| | |--clean_gs_data.csv

| | |--README.txt

| |-- results

| |-- src

| | |-- clean_data.R

- New folder **data_clean** will contain the cleaned data file, which is the output of the cleaning process. I named it **clean_gs_data.csv**.

- **clean_data.R** - This script will contain the R code for cleaning the data. It will read the raw data file, perform cleaning tasks like removing missing values and outliers, cleaning unnecessary spaces, converting data types, and output the cleaned data file to the **data_clean** folder.

Stage 3 Data analysis

|-- GripStrength_project

| |-- data_raw

| | |-- raw_gs_data.csv

| | |-- README.txt

| |--data_clean

| | |--clean_gs_data.csv

| |-- results

| | |-- test_results.txt

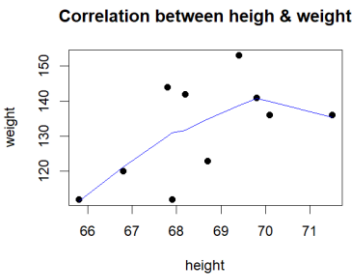
| |-- src

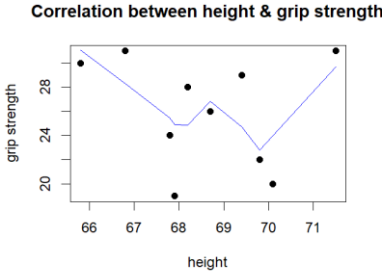
| | |-- analysis.R

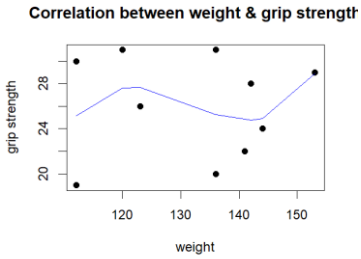
| | |-- clean_data.R

- In this stage, I have made an analysis of the cleaned data. **Results** will contain the data analysis output, such as tables, plots, and any other relevant results.
- The script **analysis.R** will contain the R code for analyzing the data. It will read the cleaned data file, perform analysis tasks like calculating summary statistics and creating plots, and output the results of the analysis to the **results** folder.

In this stage, I created scatter plots to see if there is some correlation analysis between all numeric variables.







Conclusion: As seen in the above graph, grip strength is not correlated directly with weight and height. If people's height is under 70 inches, there is shown a negative correlation between height and grip strength while above 70 inches, these two variables have positive correlation.

Also, I estimated correlations between all numeric variables as follows.

	height	weight	age	grip_strength	frailty
height	1.00000000	0.57152498	-0.03258039	-0.16768188	0.19318608
weight	0.57152498	1.00000000	0.19092588	0.03280675	0.53520014
age	-0.03258039	0.19092588	1.00000000	0.13375644	-0.08365463
grip_strength	-0.16768188	0.03280675	0.13375644	1.00000000	-0.47586687
frailty	0.19318608	0.53520014	-0.08365463	-0.47586687	1.00000000