

Ques-1:

a. Folder Structure

```
| - - PDS_Assignment_1
|   | - - Ques_1
|   |   | - - data_raw
|   |   |   | - - raw_data.csv
|   |   | - - data_clean
|   |   |   | - - clean_data.csv
|   |   | - - results
|   |   |   | - - height_Distribution.png
|   |   |   | - - height_grip_stren_box_plot.png
|   |   | - - src
|   |   |   | - - analysis.ipynb
|   |   |   | - - clean_data.ipynb
|   |   | - - README.md
```

b. Three Stages of workflow

Stage-1: Data Collection

The data that was given in question-1 was copied and saved into raw_data.csv under the data_raw folder.

	A	B	C	D	E	F	G
1	Height	Weight	Age	Grip strength	Frailty		
2	65.8	112	30	30	N		
3	71.5	136	19	31	N		
4	69.4	153	45	29	N		
5	68.2	142	22	28	Y		
6	67.8	144	29	24	Y		
7	68.7	123	50	26	N		
8	69.8	141	51	22	Y		
9	70.1	136	23	20	Y		
10	67.9	112	17	19	N		
11	66.8	120	39	31	N		
12							

For further processing and analysis, the above file raw_data.csv was loaded into Google Colab using pandas.

```
[5] import os
import pandas as pd

from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive

[3] os.chdir('/content/drive/MyDrive/PDS_Assignment_1/Ques_1')

[6] data = pd.read_csv('data_raw/raw_data.csv')
data
```

	Height	Weight	Age	Grip strength	Frailty
0	65.8	112	30	30	N
1	71.5	136	19	31	N
2	69.4	153	45	29	N
3	68.2	142	22	28	Y
4	67.8	144	29	24	Y
5	68.7	123	50	26	N
6	69.8	141	51	22	Y
7	70.1	136	23	20	Y

Stage-2: Data Processing

The data types of the columns such as Height, Weight, Age, Grip Strength which are numeric in nature were changed from Object type to Float type. For further analysis, this processed data was saved into **clean_data.csv**.

The below script which is used to clean the raw data was saved under the src folder with name **clean_data.ipynb**.

```
+ Code + Text

data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Height          10 non-null    object
1   Weight          10 non-null    object
2   Age             10 non-null    object
3   Grip strength   10 non-null    object
4   Frailty         10 non-null    object
dtypes: object(5)
memory usage: 528.0+ bytes

[9] for i in data.columns[:-1]:
    data[i] = data[i].astype('float')

data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Height          10 non-null    float64
1   Weight          10 non-null    float64
2   Age             10 non-null    float64
3   Grip strength   10 non-null    float64
4   Frailty         10 non-null    object
dtypes: float64(4), object(1)
memory usage: 528.0+ bytes

data.to_csv('/content/drive/MyDrive/PDS_Assignment_1/Ques_1/data_clean/clean_data.csv')
```

Stage-3: Data Analysis

The below picture shows the analysis that was performed on clean_data.csv. The below picture shows the distribution of height and box plot model of Height and Grip Strength features.

The below script which is used to analyse the clean data was saved under the src folder with name **analysis.ipynb**.

