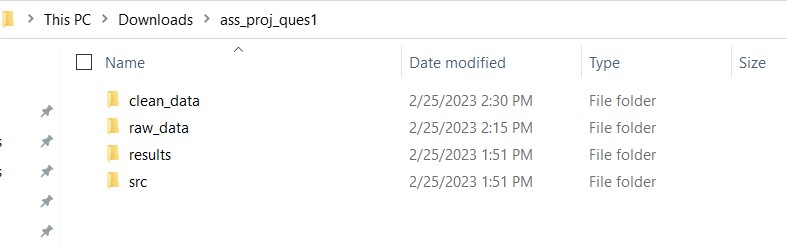
1) Based on the following table, design the three stages of reproducible workflow, includes the work you can do and the folder structure in each stage (reference study case in chapter 3).  (5 points)

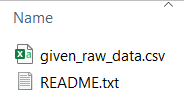
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Height (Inches) | Weight (Pounds) | Age | Grip strength | Frailty |
| 65.8 | 112 | 30 | 30 | N |
| 71.5 | 136 | 19 | 31 | N |
| 69.4 | 153 | 45 | 29 | N |
| 68.2 | 142 | 22 | 28 | Y |
| 67.8 | 144 | 29 | 24 | Y |
| 68.7 | 123 | 50 | 26 | N |
| 69.8 |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

# Stage 1: Data Collection

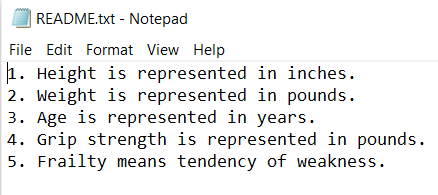
Below is the initial folder structure.



Below are the only files stored in ass\_proj\_ques1 and in raw\_data.

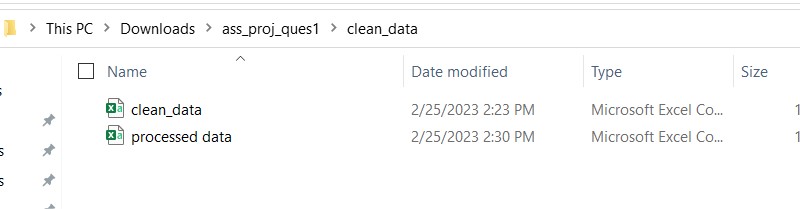


Below is the data in README.txt



# Stage 2: Data Processing

Below files are added to clean\_data folder



Below are the scripts and respective code used to create above csv files





Above .py files are stored in src folder

**Folder structure:**

-ass\_proj\_ques1

-clean\_data

clean\_data.csv processed\_data.csv

-raw\_data

given\_raw\_data.csv README.txt

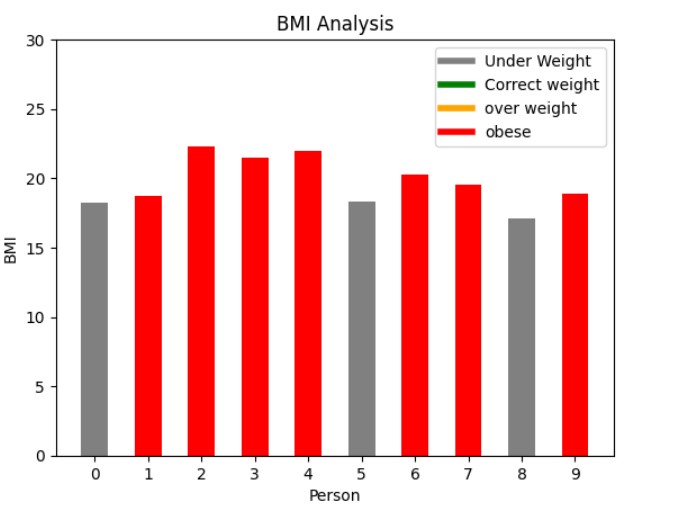
-results

-src

data\_cleaning.py pre\_process.py

# Stage 3: Data Analysis

Below are the two-analysis drawn from the pre proceed data.



Chart, box and whisker chart

Description automatically generated

Below are the respective scripts to generate above visuals. These scripts are stored in src folder.





Folder structure:

-ass\_proj\_ques1

-clean\_data

cleaned\_data.csv processed\_data.csv

-raw\_data

given\_raw\_data.csv README.txt

-results

-src

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).

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

Submission:

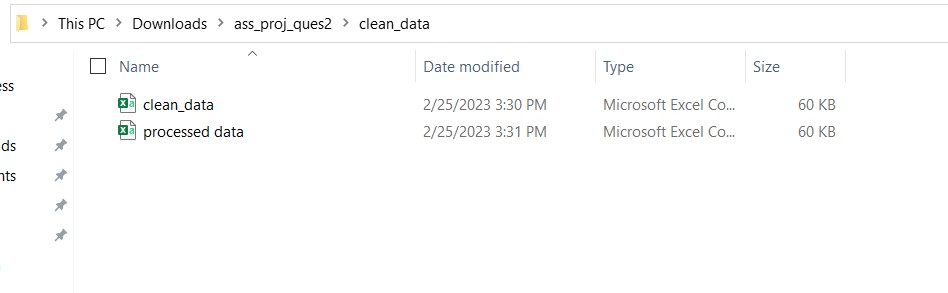
Create a public GitHub repo and upload the folders for both the questions on the GitHub and submit the link to Canvas.

# Stage 1: Data Collection

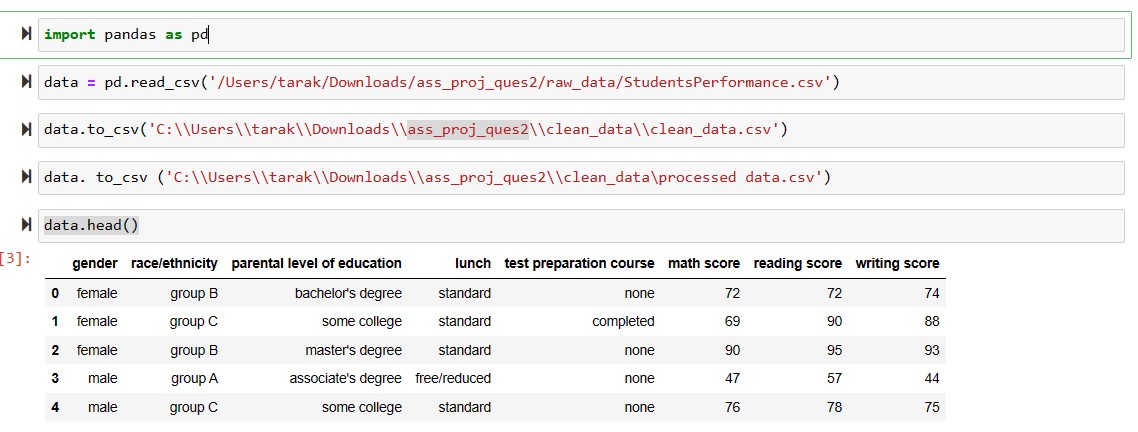
# 

# Stage 2: Data Processing

Below files are added to clean\_data folder



Below are the scripts and respective code used to create above csv files.



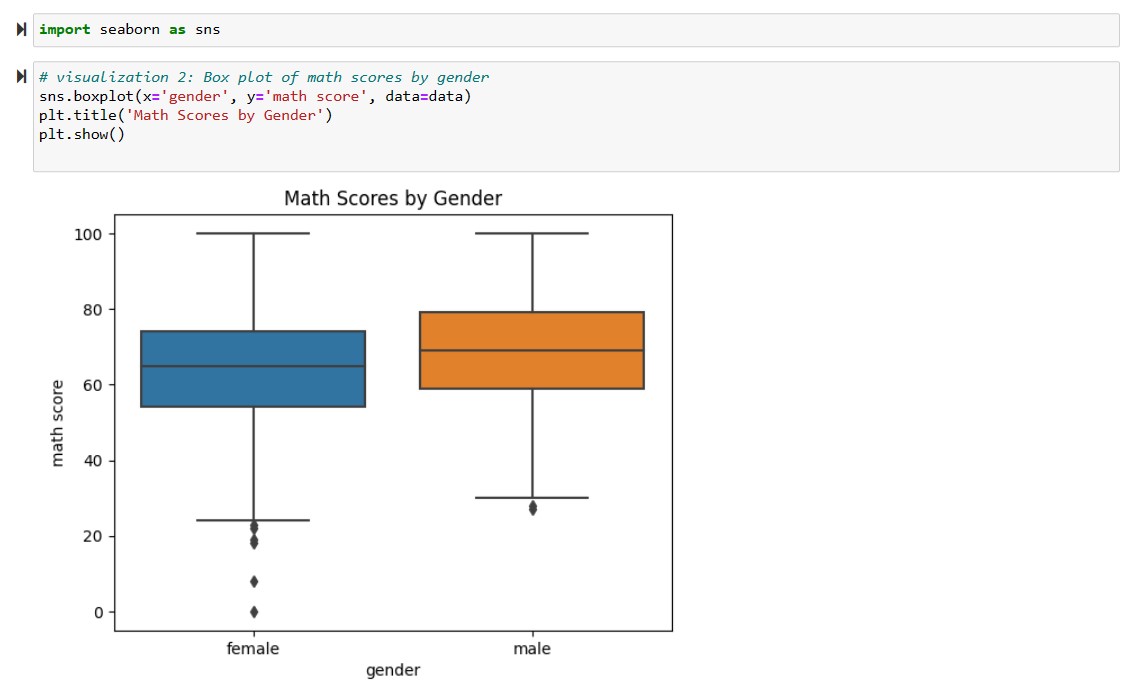
# Stage 3: Data Analysis

1. data visualization tasks are:

1) Bar plot for gender distribution.



2) Box plot of math scores by gender.



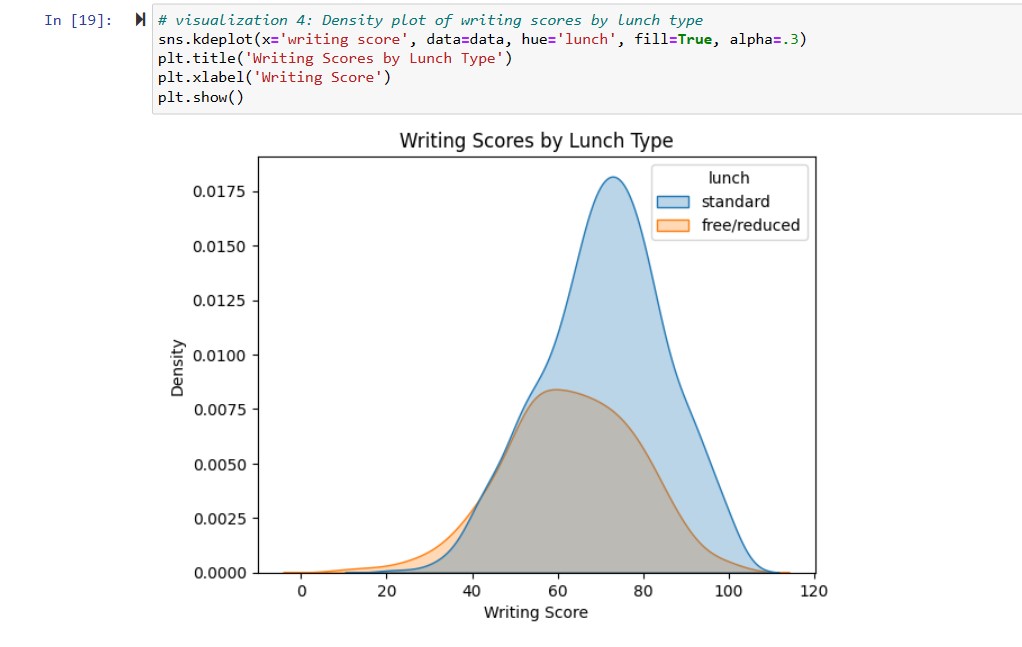
3) Scatter plot of math scores vs

reading scores.

Chart, scatter chart

Description automatically generated

4) Density plot of writing scores by lunch type.



5) Heatmap of correlation matrix.

