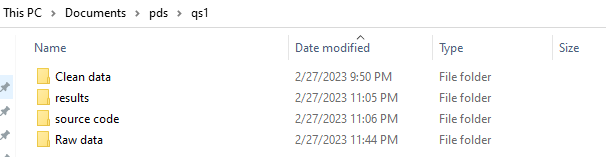
**Assignment 1**

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)

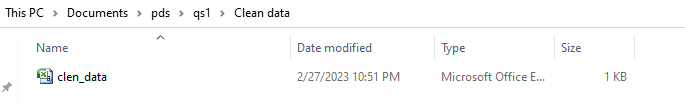
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Height (Inches) | Weight (Pounds) | Age | Grip strength | Frailty |
| 65.8 | 112 | 30 | 30 | N |
| 71.5 | 136 | 19 | 31 | N |
| 69.4 | 153 | 45 | 29 | N |
| 68.2 | 142 | 22 | 28 | Y |
| 67.8 | 144 | 29 | 24 | Y |
| 68.7 | 123 | 50 | 26 | N |
| 69.8 |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

Step 1:**Data collection**

****

Step 2:**Data Processing**

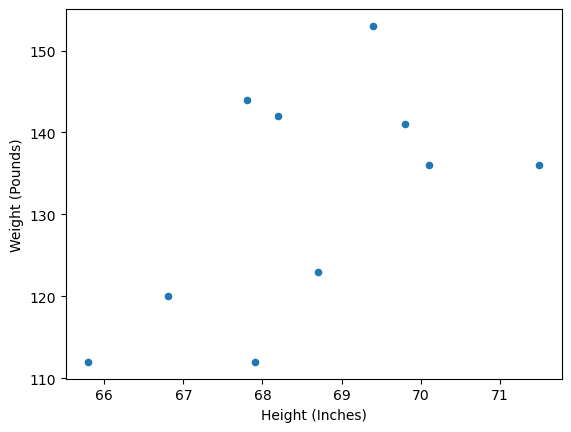
No null values and duplicate values found in the data.so the given data does not require much cleaning. Below file is added to Cleandata folder

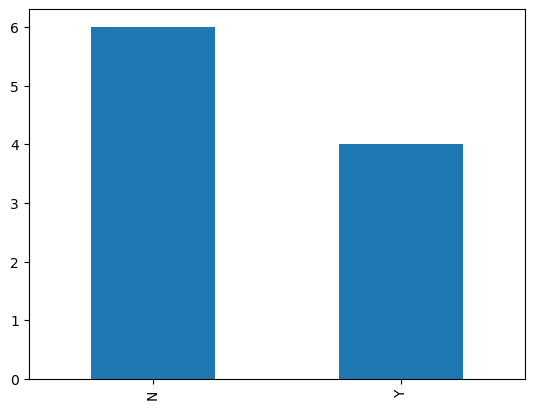


, Below are the scripts and respective code used to create above csv file and clean raw data

Step 3 :**Data Analysis**

performs some basic analysis such as displaying the first 5 rows of the dataframe, calculating summary statistics for the numerical variables creating a cross tabulation of Frailty and Age, creating a scatter plot of Height and Weight, and creating a bar plot of Frailty.





Folder structure:

-qs1

-clean\_data

clen\_data.csv

-raw\_data

Frailty\_data.excel

-results

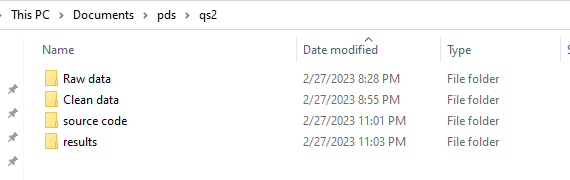
-source code

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>

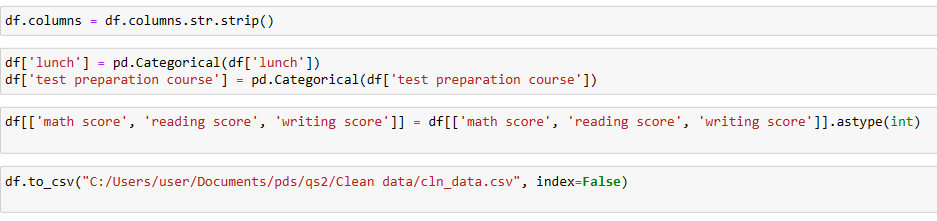
Step 1:

Read the CSV file into a pandas dataframe,

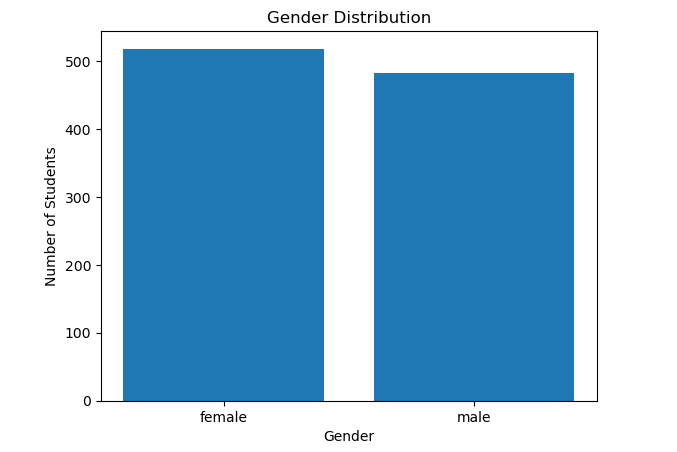


Step 2:Data Processing

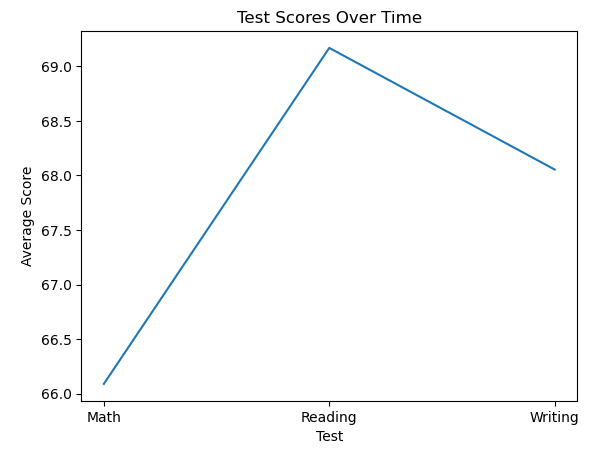
removed any leading or trailing whitespaces from the column names, dropped any missing values from the dataframe, converted the 'lunch' and 'test preparation course' columns to categorical variables and converted the 'math score', 'reading score', and 'writing score' columns to integer data type. Finally, printed the first few rows of the cleaned dataframe using the head() method



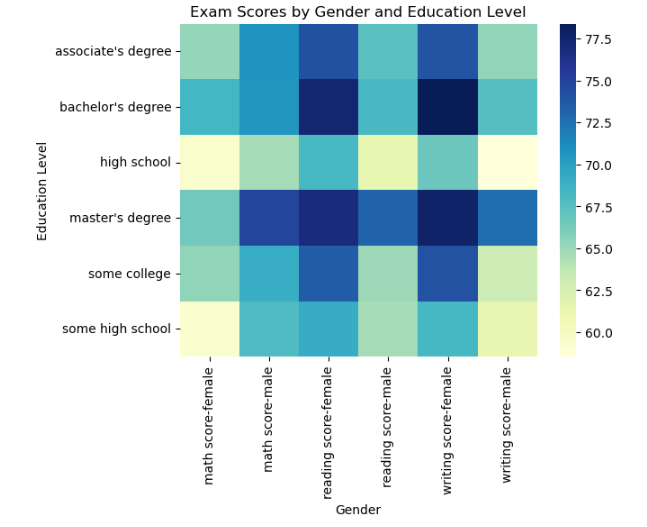
Step 3:Data Analysis



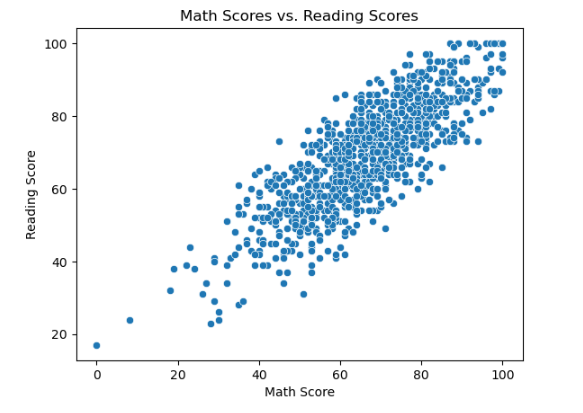
**1.Bar Chart of Gender Distribution:** This visualization can make it easier to compare the number of male and female students in the dataset. This can facilitate analysis related to gender-based trends or biases, such as whether there are any differences in exam scores or other variables based on gender.we can say that there are more number of females than males



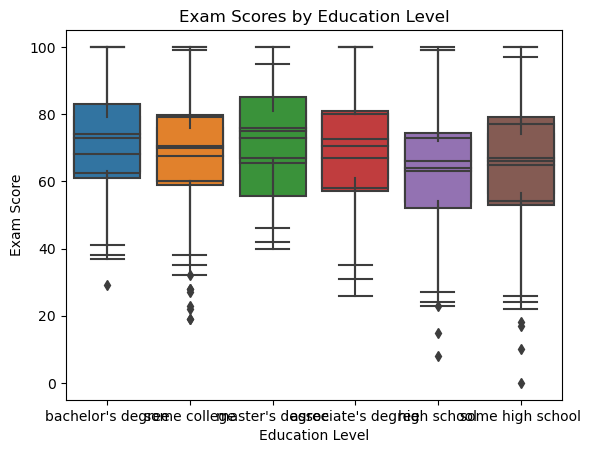
**2.Line Graph of Test Scores Over Time:** This visualization can make it easier to identify any overall trends or patterns in student performance over time. It can facilitate analysis related to the changes or fluctuations in student performance over the years, and to identify any potential factors that may be influencing these changes.We can say average math score is less than reading and writing scores over time.



**3.Heat Map of Exam Scores by Gender and Education Level:** This visualization can make it easier to compare the average scores for each gender and education level, which can facilitate analysis related to the impact of gender and parents' education level on student performance. It can help identify any potential relationships or biases based on both variables and to identify any trends or patterns that may exist.we can also say that female scored more than male when parents completed masters degree

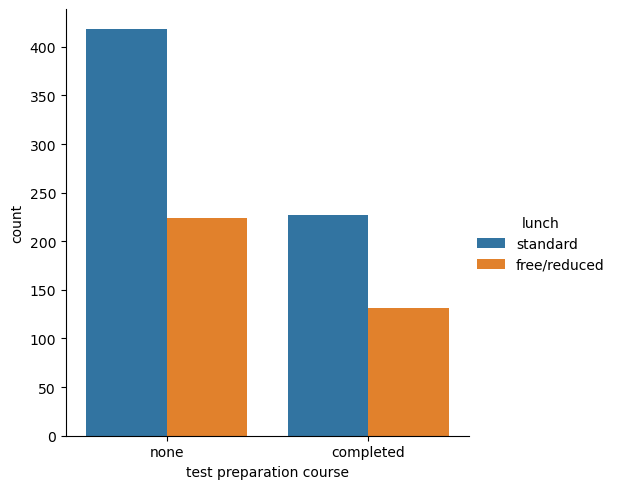


**Scatter Plot of Math Scores vs. Reading Scores:** This visualization can make it easier to identify any potential correlations or patterns between math and reading scores. It can facilitate analysis related to the relationship between the two variables and help identify any trends or patterns that may exist.we can say that scores are mostly correlated



**Box Plot of Exam Scores by Education Level:** This visualization can make it easier to compare the exam scores for each education level, which can facilitate analysis related to the impact of parents' education level on their children's exam scores. It can help identify any differences or similarities between the different groups, and to identify any potential relationships between the two variables.we can say children of parents with masters degree scored maximum of all and with high school scored lowest.

6.



This plot can help us understand if there is any correlation between students' lunch types and their completion of the test preparation course. We can see if there is any trend or pattern in the number of students who completed the course and their lunch type .