***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 |

Reproducible workflow is important because it allows researchers to share their work in a way that others can easily understand, verify, and build upon.

***Stage I: Data Collection***

* Create a new spreadsheet program and enter the given data into it.
* Save the spreadsheet as a CSV file in a new folder named "data".
* Create a metadata file with information about the data and save it in the same "data" folder.
* Add a field in the metadata files to indicate the date of data collection.
* Save both the data and metadata files with appropriate names, e.g., "raw\_data.csv" and "metadata.txt".
* Create a subdirectory named "src" for storing all source code files.
* The Work we can do here is to save Raw data along with information the Information should tell us about the data for example what frailty in the above table and what does the values N (NO) and Y(YES) indicates and we can add fields like date about when the data is collected and from where it is collected etc.,

***Stage II: Data Processing***

* Create a new script file named "clean\_handgrip\_data.ipynb" in the "src" directory.
* Write a script to read the raw data, remove rows with NA yields, blank spaces and save the cleaned data as a new CSV file in a subdirectory named "clean\_handgrip\_data" under the "clean data" directory.
* Add comments in the script file to describe the data cleaning steps.
* Run the "clean\_handgrip\_data.ipynb" script to create the cleaned data file.
* The Work we do here is basically removing the noise from the data to get the accurate analysis in the above table all the data is clean and good to analyze.

***Stage III: Data Analysis***

* Create a new script file named "analysis.ipynb" in the "src" directory.
* Write a script to perform an unpaired two-sample t-test to compare the mean Grip per person between the Frailty and grip\_strength using the clean data.
* Add comments in the script file to describe the analysis steps.
* Save the test results as a text file named "test\_results.txt" in a new subdirectory named "results" under the "data" directory.
* Run the "analysis.ipynb" script to generate the test results.

Overall Folder Structure:

* Handgrip\_project

1. Data\_raw

* raw\_handgrip\_data.csv
* metadata.txt

1. Clean Data/

* clean\_handgrip\_data.csv

1. src/

* analysis.ipynb
* clean\_handgrip\_data.ipynb

1. results/

* output.pdf

This is the Reproducible workflow of a data.

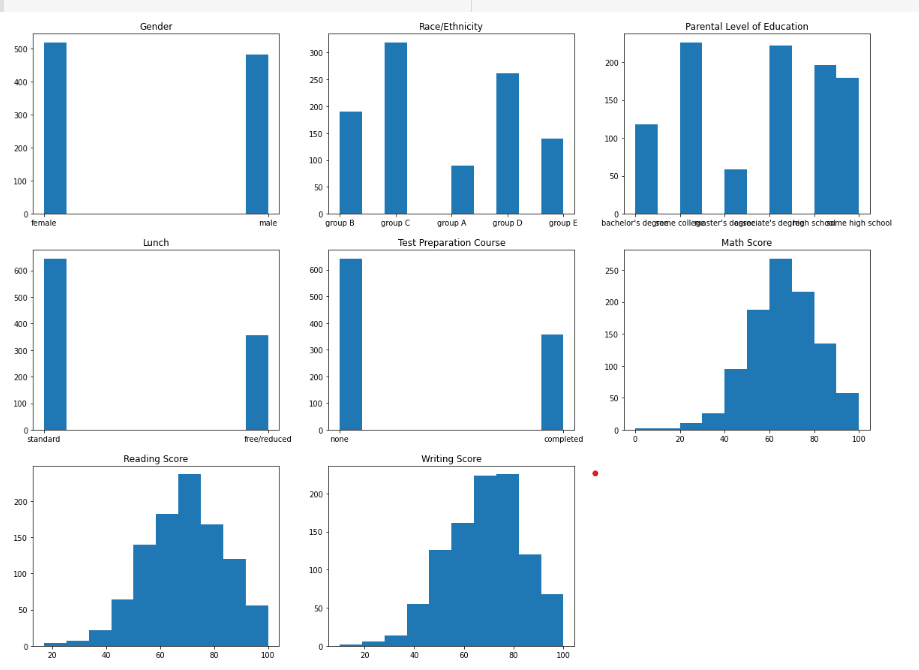
2) Perform 5 data visualization tasks on the student performance dataset given in the link below (create 5 different visualizations). Explain what kind analysis has become easier with each of the visualizations. Create the folder structure for this question similar to question 1. (15 points).

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

***VISUALIZATIONS***

**Histogram**

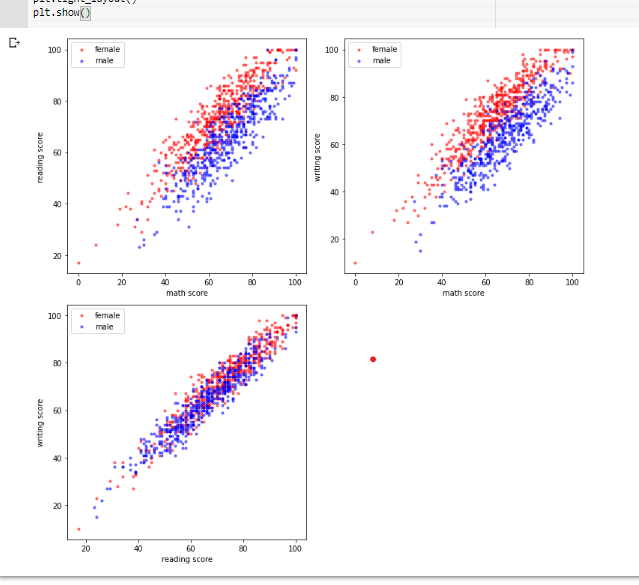
The histogram graph below represents all the fields in the dataset and what are the fields and the different levels of the fields.



The graph represents

* There are more females then the males from gender.
* In Race/Ethnicity group C are higher than the other groups whereas group A is the least.
* From Parental level education some college people is higher and master’s degree is lower
* Lunch is higher for standard and lower for free/reduced.
* There are only 300-400 people completed the test preparation course.
* Math scores are 60 for 250 + people and 100 for 50 people.
* There is 200+ people with the 60-80 reading score and 400+ people is having 60-80 writing score.

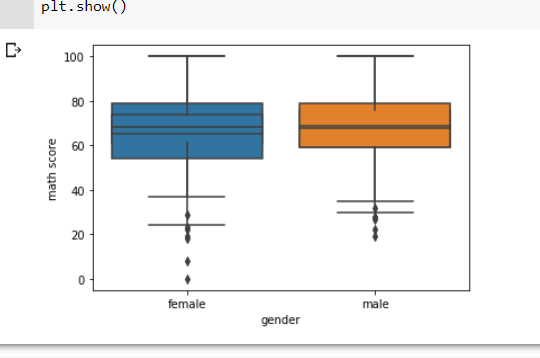
**Scatter Plot**



Here we are seeing is there any trends based on gender for scores like Math, reading and writing.

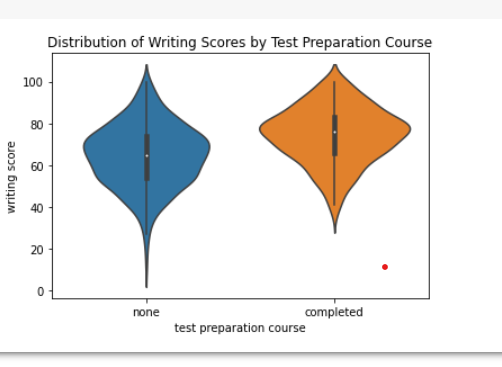
* From scatter plot we can see there is a positive correlation for all the plots so we can say that reading score and math score, math score and writing score and reading score and writing score is having almost perfect positive correlation.
* When we see it based on gender wise then also it is having a positive correlation between the scores.
* This tells us both female and male are scoring in positive correlation for all the scores which means who scores better in writing does better in reading , who scores better in math does better in reading and who scores better in math does better in writing and vice versa.

**Box plot**

****

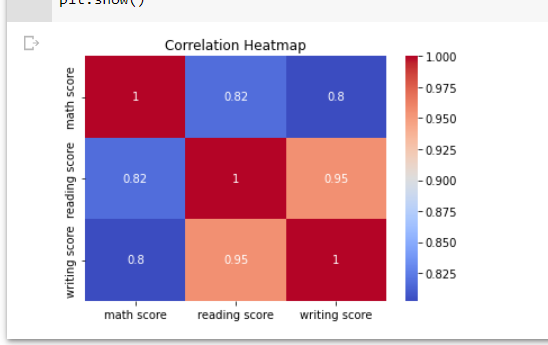
* The box plot is drawn for the gender v/s math score.
* The box plot shows that there are some outliers for both male and female.
* The median for math score is almost same for both the males and females.
* The range of female is little higher than the male.
* Some females are having very least score like 0 .
* The box plot shows the interquartile range, which represents the average score (IQR).

**Violin plot**

****

* A violin plot is similar to a box plot, but also shows the probability density of the data at different values.
* It can be used to compare the distribution of a variable across different categories.
* The test preparation course completed people scored better than the none people and the median is higher for completed people.
* The interquartile range is also better for course completed people.
* The course not completed people also got 100 but also got 0.

Heat Map



* According to the heat map, there is a significant positive link between reading and writing scores as well as between math and reading scores.
* This indicates that students who performed well on the reading exam tended to perform well on the writing exam, and vice versa.
* Students who performed well on the math exam likely to perform well on the reading exam.

https://github.com/SravaniTatimakula/PDS-ASSIGNMENT-1