**PDS ASSIGNMENT**

**NAMBURI SWETHA**

**16330344**

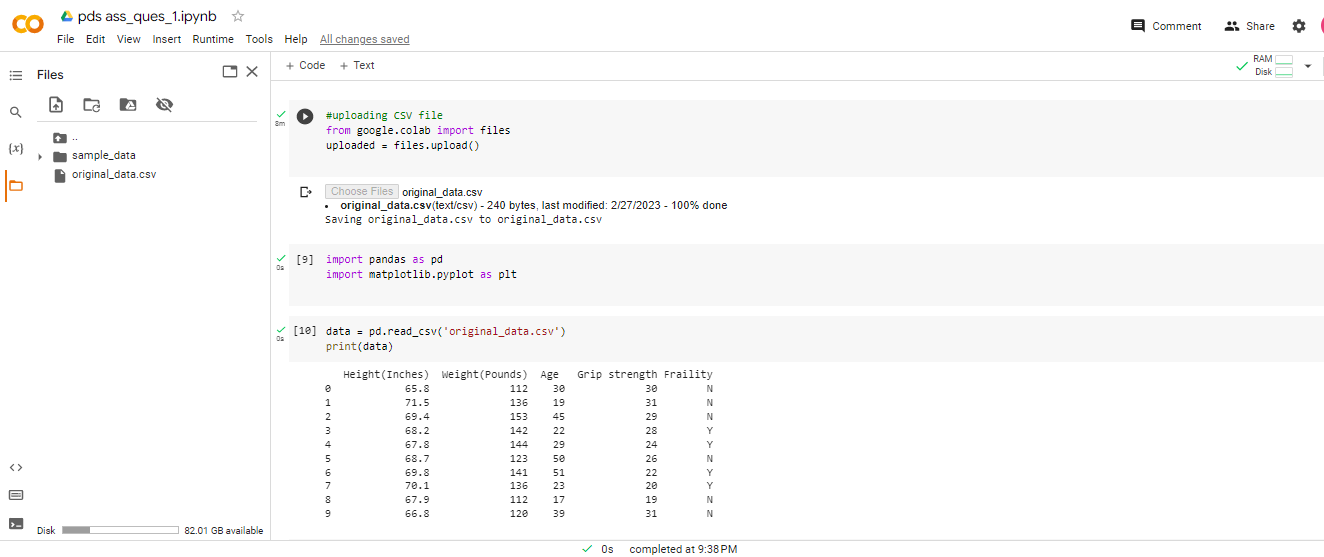
1) Based on the following table, design the three stages of reproducible workflow, includes the work you can do and the folder structure in each stage (reference study case in chapter 3).  (5 points)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

There are three stages for reproducible workflow which are

DATA COLLECTION:

We need to collect the data from source and convert it into csv file.Here the original\_data is the raw data file and we uploaded it.



DATA CLEANING AND PROCESSING:

In this stage, the raw data will be cleaned and pre-processed at this step. The following are the duties that must be completed at this stage:

* We need to Open the input.csv file(raw\_data.csv) located in the raw data folder to load the raw data.
* Remove any duplicates, missing numbers, or outliers from the data to clean it up.
* Save the cleaned data to a new file called output.csv(clean\_data.csv) in a new folder you create called data/cleaned data.
* Since the supplied data are unambiguous, straightforward, and devoid of noise or null values, we move on to step 3. For these data, cleaning is not essential.

DATA ANALYSIS:

At this phase, we will analyze and examine the cleaned data to learn more and spot any patterns or connections. At this stage we need to do the following tasks:

* We need to store the cleaned and processed data in the data/cleaned data folder by loading it from the output.csv file.
* Use various visualization tools, such as scatter plots, histograms, and box plots, to analyze the data.
* Under the exploratory analysis folder, create a new folder named figures and save the created charts there.



We are using scatterplot to visualize the relationship between age and grip strength in the data. From this we can conclude that younger generations are having less grip strength.

Folder structure:

-ass\_proj\_ques1

-clean data

clean\_data.csv processed\_data.csv

-raw data

raw\_data.csv README.txt

-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>

Submission:

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

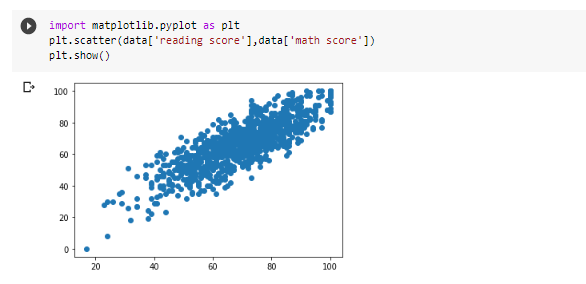
**1.VISUALIZATION OF AVERAGE SCORES BY TEST PREPARATION COURSE**

Here the mean of three scores such as math score, reading score and writing score is calculated and it easier to compare the performance of students who completed the test preparation course to those who didn’t. This can help in analyzing whether the course has any significant impact on the scores.



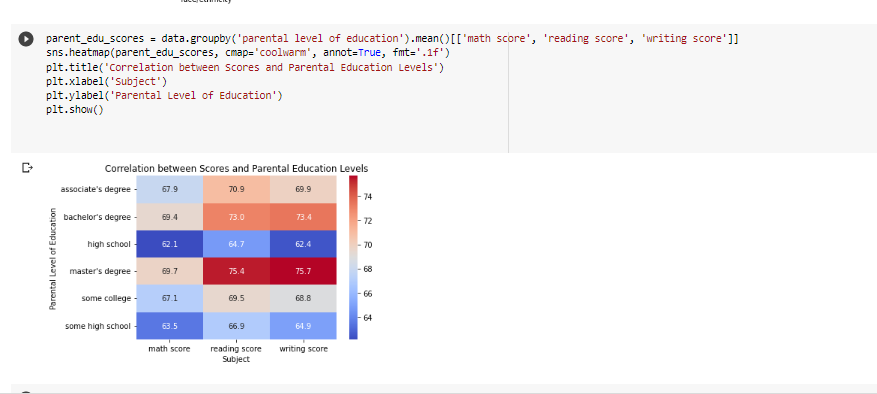
**2.VISUALIZATION OF MATH SCORE VS READING SCORE**

This visualization helps us to find the correlation between the two scores using the scatter plot of the math and reading results. As a result, it will be easier to determine whether there is any association between the results and whether students who perform well in maths also frequently perform well in reading.



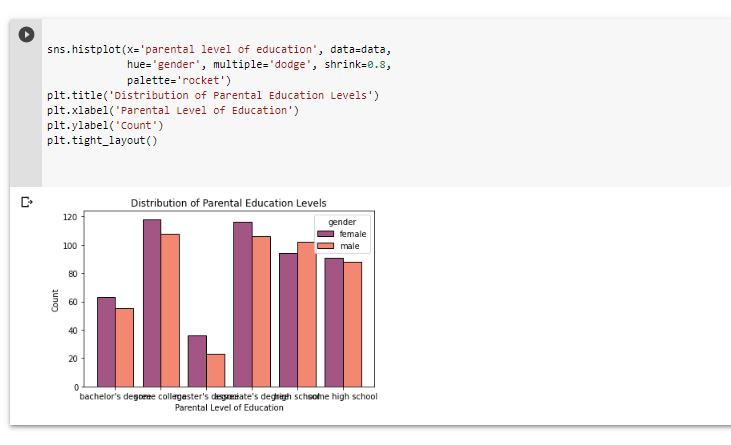
**3.VISUALIZATION OF PARENTAL LEVEL OF EDUCATION WITH SCORES**

The code group the data by the column labeled "parental level of education," compute the means of the columns labeled "math score," "reading score," and "writing score," and then produce a heatmap using the seaborn library to show the relationships between the scores and the parental level of education.With the 'fmt' argument, the heatmap displays the mean scores as annotations with an accuracy of one decimal place using the 'coolwarm' colormap.



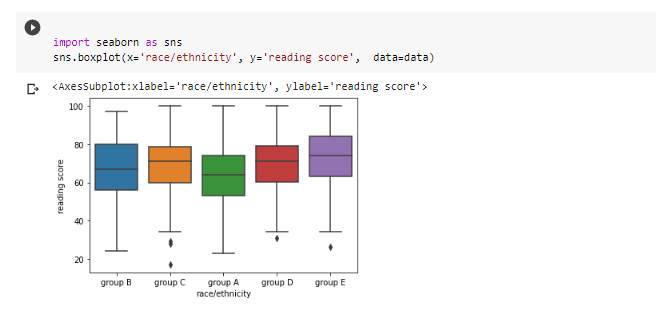
**4.VISUALIZATION OF DISTRIBUTION OF PARENTAL EDUCATION LEVELS:**

In this Visualization, using the count of people in each category on the y-axis and the various categories of parental education levels on the x-axis, the code provided creates a histogram plot of the distribution of parental education levels. A histogram plot showing the distribution of parental education levels for each gender category in the dataset is produced by the code as a whole by the information provided.



5.**VISUALIZATION OF READING SCORE IN RESPECT TO RATE/ETHNICITY:**

The code produces a boxplot that displays the distribution of reading scores across the dataset's various race/ethnicity categories.The boxplot is made using the sns.boxplot() function, with the x and y parameters set to "race/ethnicity" and "reading score," respectively, to indicate the variables on the x- and y-axes, and the data parameter set to the DataFrame data that contains the necessary data.



Folder Structure:

----Student Performance Visualization.

🡪 README.txt

----Data

🡪 StudentPerfomance.csv

----Notebooks

🡪 PDS\_ass\_ques2.ipynb

----Output

🡪RESULT1.png

🡪RESULT2.png

🡪RESULT3.png

🡪RESULT4.png

🡪RESULT5 png