**ASSIGNMENT-1**

**Name-** Sai Priyatham Gummadi

**St.ID -**16331093

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Height(Inches)** | **Weight(Pounds)** | **Age** | **Grip strength** | **Fraility** |
| **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 |
|  | 120 | 39 | 31 | N |

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)

**Data Reproducibility**

There are three stages of data reproducibility.

1. Data Acquisition

2. Data Cleaning/Processing

3. Data Analysis

**Stage 1-Data Acquisition:**

We must manually enter the data into an Excel sheet and then save it as a.csv file.

Graphical user interface

Description automatically generated

**Stage 2-Data Cleaning/Processing**

The second stage, known as data cleaning or processing, entails cleaning up the data that was generated in the previous stage. We go on to step 3 since the provided data are clear, uncomplicated, and free of noise or null values. Cleaning is not necessary for these data.

Graphical user interface

Description automatically generated with low confidence

**Stage 3-Data Analysis.**

It is the last step in the reproduction of data. In this part of the study, we look at the data using different mathematical techniques or visuals to identify correlations between the variables.

#graph represents the scatter and visualition plots which represnts age on x-axis and Grip strength on y-axis

#here in the graph blue plot represent fertaility class-'N' and Green color represents class fertaility-'y'

Graphical user interface

Description automatically generated

We see that individuals older than 20 who do not exhibit fertility have strong grips. Moreover, some of the individuals with BMI weakness are younger than 20.

Table

Description automatically generated

We get to the conclusion that 80% of people who are older than 30 and have a high body mass index are strong. And only 40% of those under 30 are physically fit. Hence, those above the age of 30 are stronger than those under 30.

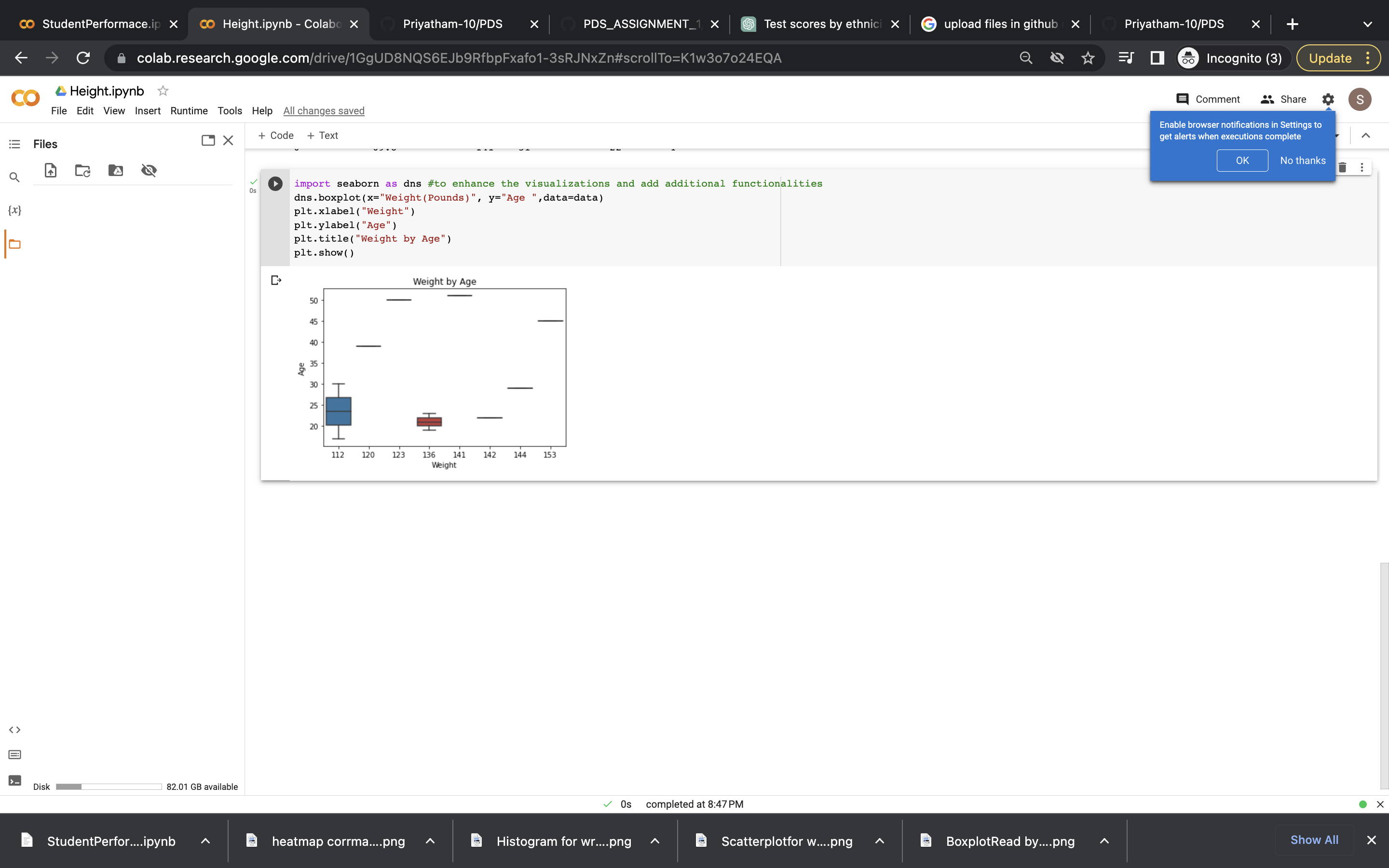
**SOURCE CODE**

Graphical user interface, application

Description automatically generated

Graphical user interface, text

Description automatically generated



Folder Structure

**-------content**

🡪 **README(1).txt**

----**Data**

🡪 **Book\_PDS.csv**

**----Notebooks**

🡪 **Height.ipynb**

----**Output**

**🡪Age.png**

**🡪BMI.png**

**🡪Weight.png**

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.

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

Here we are given with Student Performance data which consists of 8 attributes such as Gender, Race/Ethnicity, Parental Level of Education, Lunch, Test Preparation Course, Math Score, Reading Score, Writing Score. In the given data with the help of all these attributes a wide range of student’s performance is analysed. The Basic explanation for all these attributes are:

• Gender: The gender of each student.

• Race/Ethnicity: The group for which each student belongs to.

• Parental Level of Education: The education qualification of each student’s parents.

• Lunch: The lunch category of each student

• Test Preparation Course: Data about test preparation course completed by student.

• Math Score: The math score of each student.

• Reading Score: The reading score of each student.

• Writing Score: The writing score of each student.

The given data consists of 1000rows and as well as 8 columns. Out of the given 8 columns Math Score, Reading Score and Writing Score are the columns with Numerical values. Performing some basic statistical operations on this numerical data has given details about minimum values, maximum values, mean and standard deviation. Remaining comes under categorical values as they contain two or more categories in their data.

There are different kinds of data visualization schemes. Let us consider 5 of them and perform them on our given data which is as shown below.

**Task-1:Bar Plot for Gender Distribution**

With this visualization task we can see the parental level of education distribution of students in the dataset. We can analyse how much data contains and how they compare to each other.

Graphical user interface, application

Description automatically generated

**Type-2: Box Plot of reading score by gender**

This visualization task compares the reading scores for male and female students(gender).It can help to identify any gender based differences in reading scores, which can be useful in identifying areas where one gender might need additional resources.

Chart

Description automatically generated

**Type-3: The Scatter plot of writing score versus reading scores.**

The Scatter plot of writing score versus reading scores gives us clear understanding between 2 variables. From this scatter plot we can see a positive correlation between reading and writing score which means students perform well in both reading and writing. This type of visualization helps us identifying patterns in data.

Graphical user interface

Description automatically generated

**Type-4 Histogram plot for distribution of the writing score**

This Visualization makes it easy to see the distribution of the writing scores. We can see that the scores are normally distributed and that most students scored between 65 and 82.

Bin = 5

Graphical user interface, application

Description automatically generated

Bin = 10

Graphical user interface

Description automatically generated

**Type -5 Heatmap task**

The Visualization task results in showing pairwise correlation between math and reading scores. The heatmap shows that there is a moderate positive correlation between math and reading scores(0.82).This visualization makes it easier to identify the relationships between the different test scores and understand how they are correlated with each other.

Graphical user interface, application

Description automatically generated

Top of Form

**Type-6: Pictogram Plot**

By Performing these pictogram visualizations, it becomes easier to compare different aspects of the student performance dataset, providing a visual representation of the data and making it easier to identify patterns. In this we create the pictogram for math test results that how many students failed and how many are passed.

Graphical user interface, text

Description automatically generated

**Folder Structure:**

----Student Performance Visualization.

🡪 README.txt

----Data

🡪 StudentPerfomance.csv

----Notebooks

🡪 StudentPerformance.ipynb

----Output

🡪Barplot for ParentLE.png

🡪BoxplotRead by Gender.png

🡪heatmap corrmatrix.png

🡪Histogram for writingscores.png

🡪scatterplotfor read vs write.png