**PRINCIPLES OF DATA SCIENCE ASSIGNMENT 1**

* **MURALIKRISHNA MAANUKONDA (16330766)**

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 |
| 67.9 | 112 | 17 | 19 | N |
| 66.8 | 120 | 39 | 31 | N |

**Data Reproducibility**

There are three stages of data reproducibility

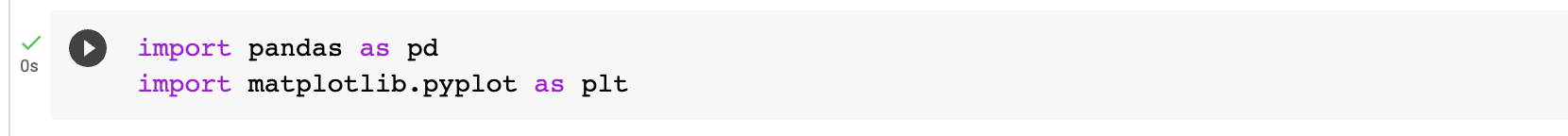
* Data Acquisition
* Data Cleaning/Processing
* Data Analysis

1. **Data Acquisition**

Manual data should be entered to spreadsheet and save it as .csv file here I saved it under the name Book1 2.csv

**Graphical user interface

Description automatically generated with medium confidence**

****

1. **Data Cleaning/Processing**

It is second stage of data reproducibility. Data from the first stage is being cleaned up in this level. Data input, methodical data modification, and data filtering are all required. This data must first be processed before being fed into the model to get improved results. We move on to stage 3 because the data is relatively straightforward and doesn't have any distortion or null numbers in this case.

1. **Data Analysis**

The final step of the workflow involves the analysis of the data using various graphical operations or visualizations to establish the correlation between the variables. We can perform data visualization here because the dataset is straightforward and simple to view, but other researchers might find it difficult to comprehend the details of the analysis. As a consequence, it is important to write comments that properly explain the code so that it can be used for future research projects and understood by as many people as possible.

**Table

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Text, application

Description automatically generated with medium confidence**

**Chart, scatter chart

Description automatically generated**

From the above graph we can observe that people is having the enough grip strength have no fratility. And we can see less grip strength in some people with younger age.

**Table

Description automatically generated**

**Folder structure :**

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

README.txt

**---------Data**

* Book1 2csv

**-----Notebooks**

* height.ipnb

BMI

Age

1. 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>

**5 Data visualization tasks –**

**Scatter plot**

**Histogram**

**Heat map**

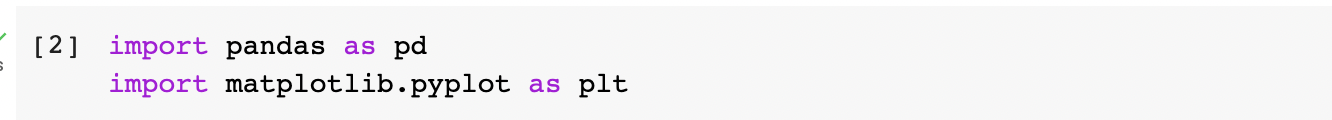
**Pictogram**

**Area plot**

Import libraries and load data and load the .csv file

Graphical user interface, text, application

Description automatically generated



Readthe data from the file StudentsPerformance.csv



**Scatter plot** - The connection between two variables, such as the final score and the number of absences, can be represented visually using a scatter plot. You can spot any correlation between the two factors with the aid of this visualization.

Chart, scatter chart

Description automatically generated

**Histogram -** The position, spread, and skewness of a dataset are all visually represented by a histogram, which also makes it easier to see whether the distribution is symmetric or left- or right-skewed. If it is unimodal, bimodal, or multidimensional as well. Additionally, it can highlight any anomalies or data voids.

Chart, histogram

Description automatically generated

**Heat map** - A heat map makes it simple to see the correlation matrix between a number of different variables. This allows you to see whether there are any strong or weak correlations between the variables and which ones are most closely correlated with one another.

Chart, treemap chart

Description automatically generated

**Pictogram -**  Charts and graphs known as pictograms employ icons and images to represent data. Pictograms, often referred to as "pictographs," "icon charts," "image charts," and "pictorial unit charts," make use of a collection of recurrent icons to represent straightforward facts.

Graphical user interface

Description automatically generated

**Area plot -** An area chart is a line graph's expansion in data visualization. To show how the numerical values of one or more groups vary over a second variable, it combines the line chart and the bar chart. A excellent chart to see how a volume changes over time is an area chart.

Graphical user interface, text, application, email

Description automatically generated

**Folder structure :**

**-------Student performance Visualisation**

README(3).txt

**---------Data**

* Student performance.csv

**-----Notebooks**

* Student performance.ipnb

Scatter plot

Histogram

Heat map

Pictogram

Area plot