PRINCIPLES OF DATA SCIENCE ASSIGNMENT 1

**FOLDER STRUCTURE**

Refer below tree for file structure

|--pds\_assn1

| |--pds\_assn1\_1

| | |--data\_clean

| | | |--rawdata\_frality.csv

| | |--data\_raw

| | | |--rawdata\_frality.csv

| | |--src

| | | |--pds\_assn1\_1.ipynb

| |--pds\_assn1\_2

| | |--data\_clean

| | | |--StudentsPerformance.csv

| | |--data\_raw

| | | |--StudentsPerformance.csv

| | |--src

| | | |--pds\_assn1\_2.ipynb

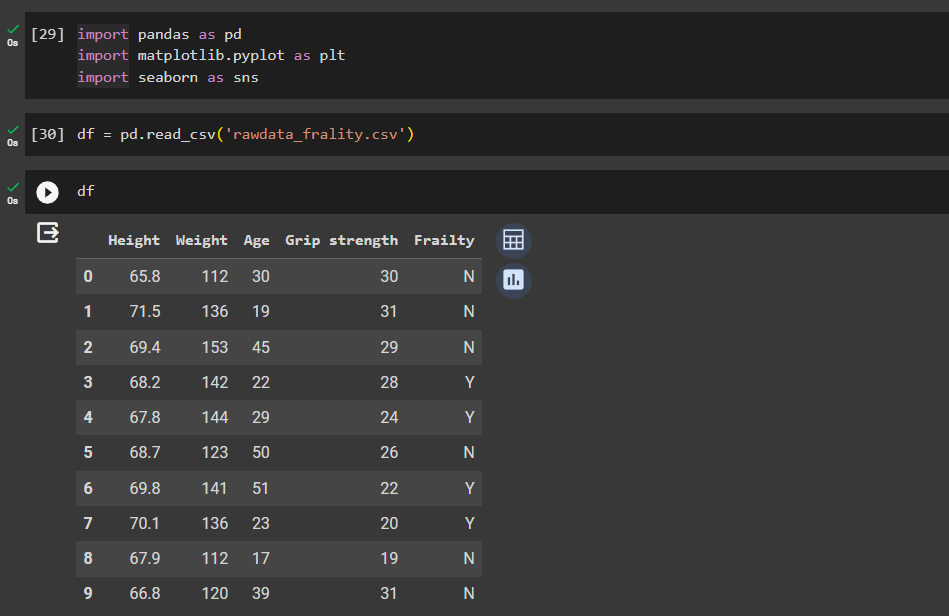


STAGE I: DATA COLLECTION

* Raw data is collected from the question and is converted into a .csv file.
* The file is stored inside a folder named ‘data\_raw.’

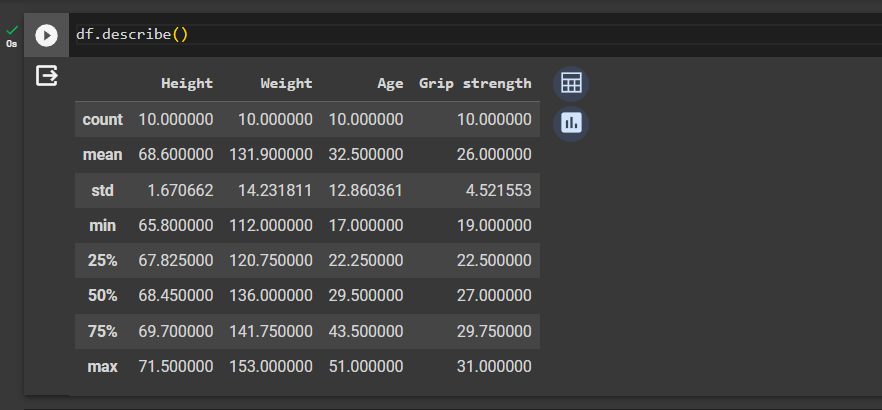
STAGE II: DATA PROCESSING

* File is loaded into google colab for processing and a dataframe is created using the data.
* Data is already cleaned, no pre-processing done.

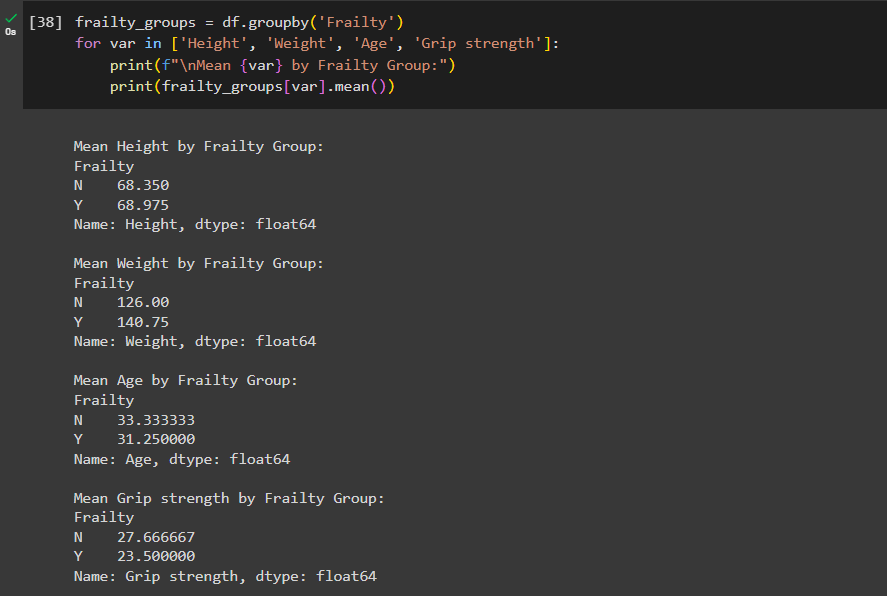


STAGE III: DATA ANALYSIS

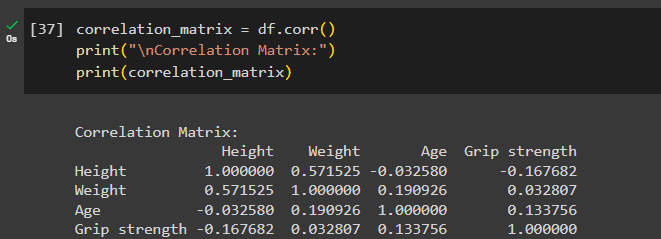
* Analyzed overall distribution of data within all the features.



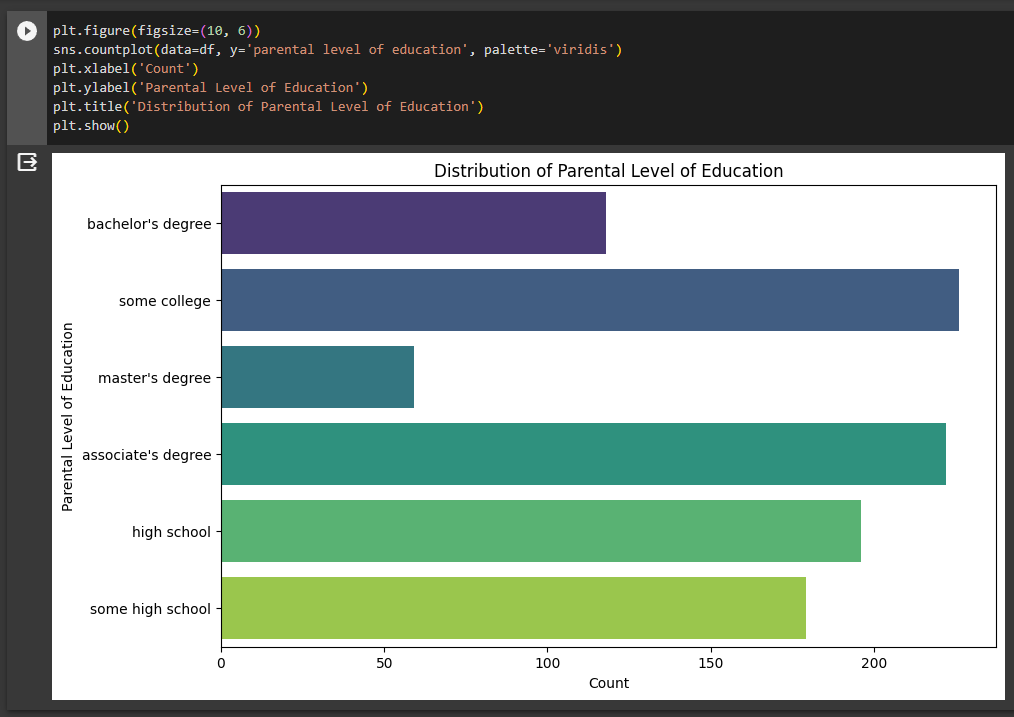
* Performed mean of each feature with women who are frail and who are not to find if there is any trend between them categories.



* Also produced a correlation matrix to see if any features can be ignored to reduce the dimensionality of the data.



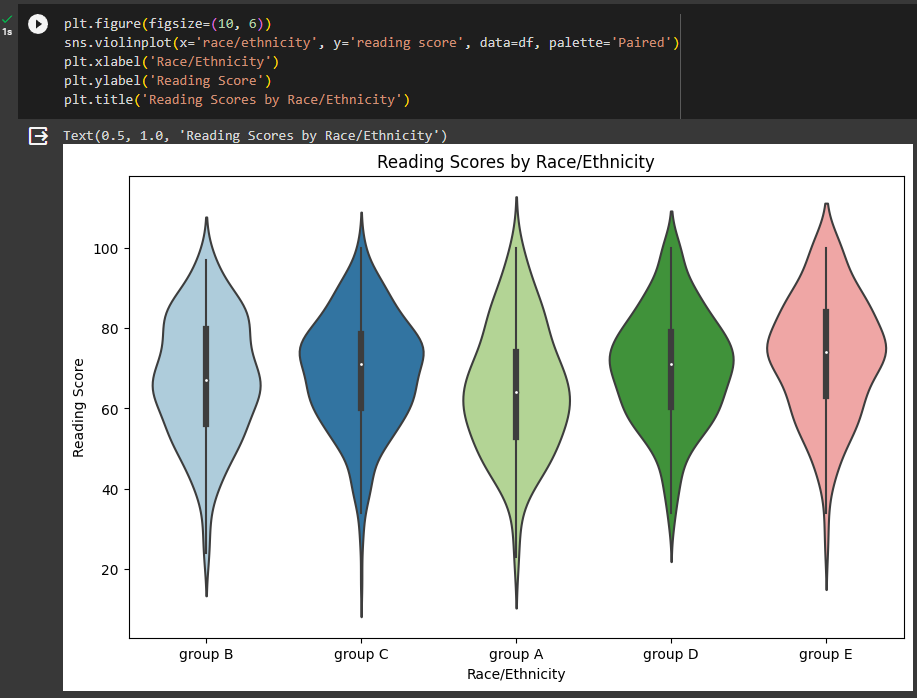
1. **Bar plot for parental level education** – provides insight into the count of students and their parents’ educational background, which might affect the scores.



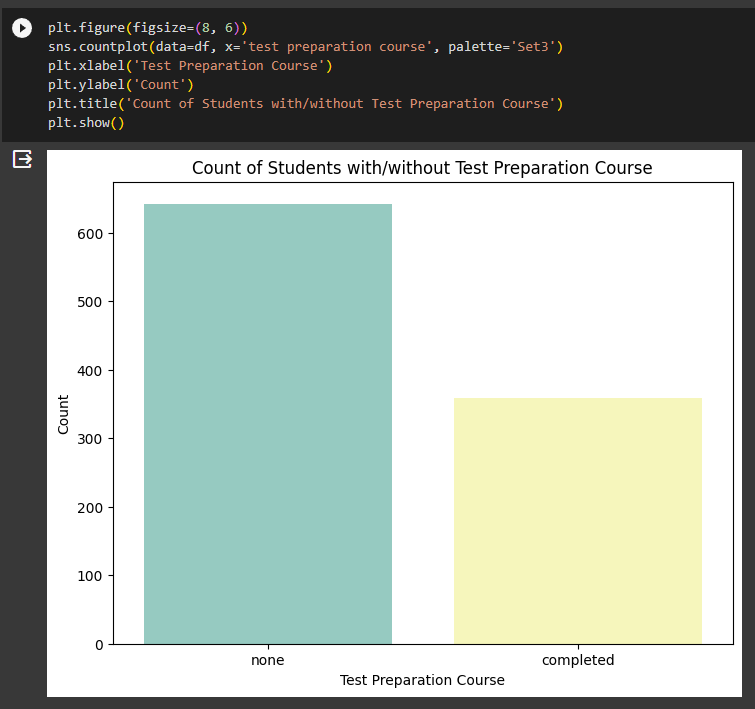
1. **Boxplot of math scores by Gender –** allows to compare math scores and perceive gender-based differences.



1. **Violinplot of reading score by race/ethnicity** - helps to identify similarities/contrasts among data distribution within distinct groups along with the frequency distribution.



1. **Count plot for test preparation completion –** gives you the count of students who completed the course which might affect the final test scores.



1. **Pair plots of all scores** – helps identify any correlation between the subjects.

