Assignment 1

Question 1:

Answer:

To perform a reproducible analysis on the dataset provided, we can follow a three-stage workflow: Data Preparation, Data Analysis, and Reporting. The folder structure and each stage is described briefly below:

Stage 1: Data Preparation

- Data Collection:
  - Gather and verify the integrity of the dataset.
- Data Cleaning:
  - Handle missing values, outliers, and data errors.
- Data Transformation:
  - Convert units if necessary and encode qualitative attributes.

Folder Structure:

/data_preparation

/data

- raw_dataset.csv

- cleaned_dataset.csv

/scripts

- data_cleaning.py

Stage 2: Data Analysis

- Exploratory Data Analysis (EDA):
  - Visualize the data and calculate relevant statistical measures.
- Model Building:
  - If necessary, construct a model to predict frailty based on other attributes.
- Hypothesis Testing:
  - Test the hypothesis about the relationships between grip strength and frailty.

Folder Structure:

/data_analysis

/visualizations

- eda_plots.png

/scripts

- data_analysis.py

Stage 3: Reporting

- Interpret Results:
  - Summarize and interpret the results of the analysis.
- Documentation:
  - Document the analysis, results, and methods used.
  -

Folder Structure:

/reporting

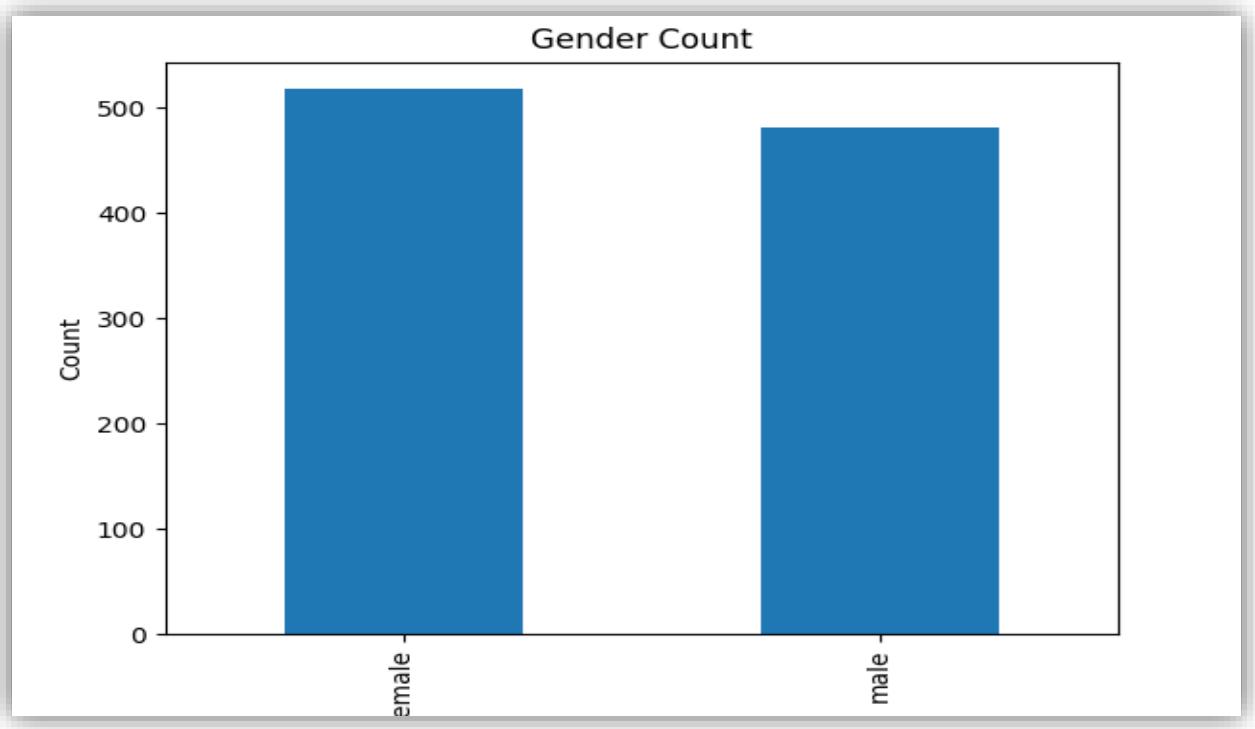
/documents

- final_report.pdf

- methodology.docx

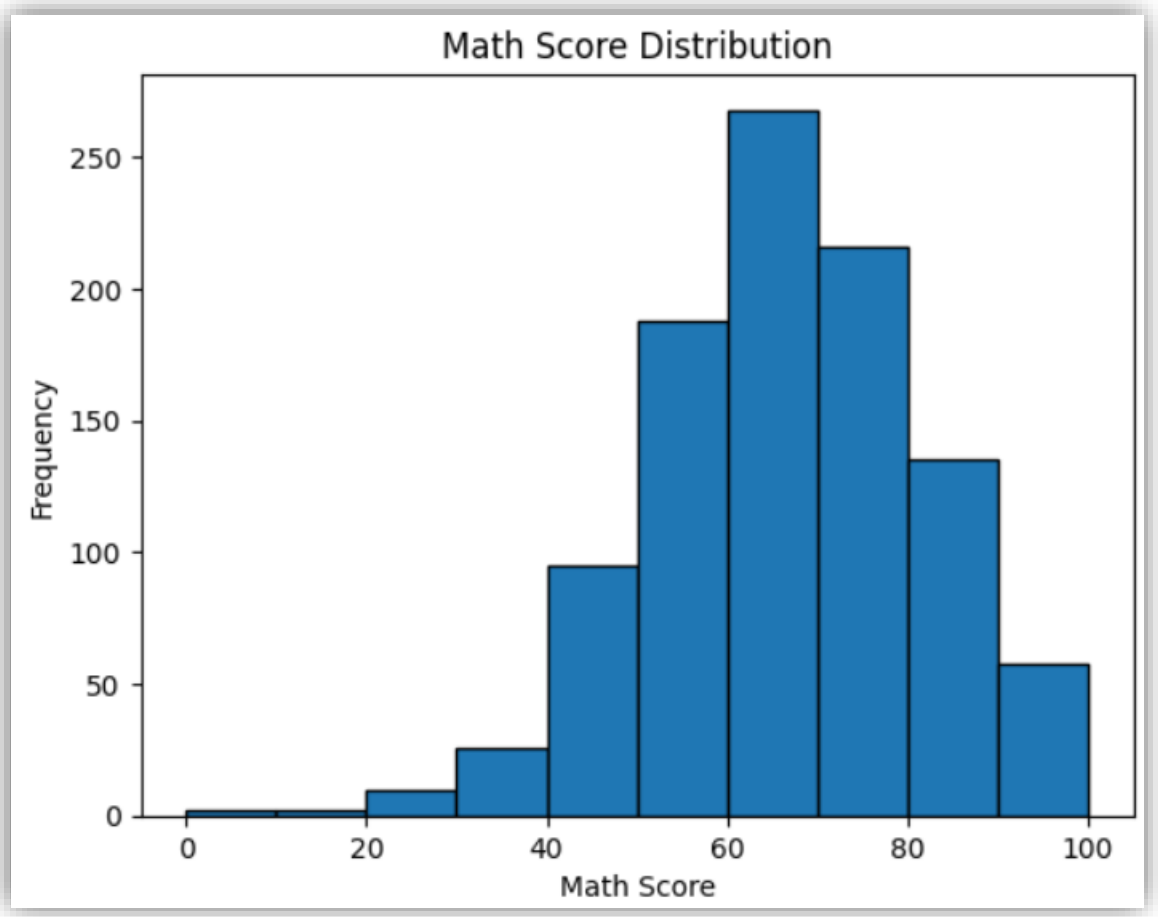
Question 2:

1. Count plot for gender



This count plot gives us the analysis for the count of male and female in the dataset. This plot roughly shows almost 520 female and 490 males. We get to know the total number of male and female student here.

2. Histogram for math scores

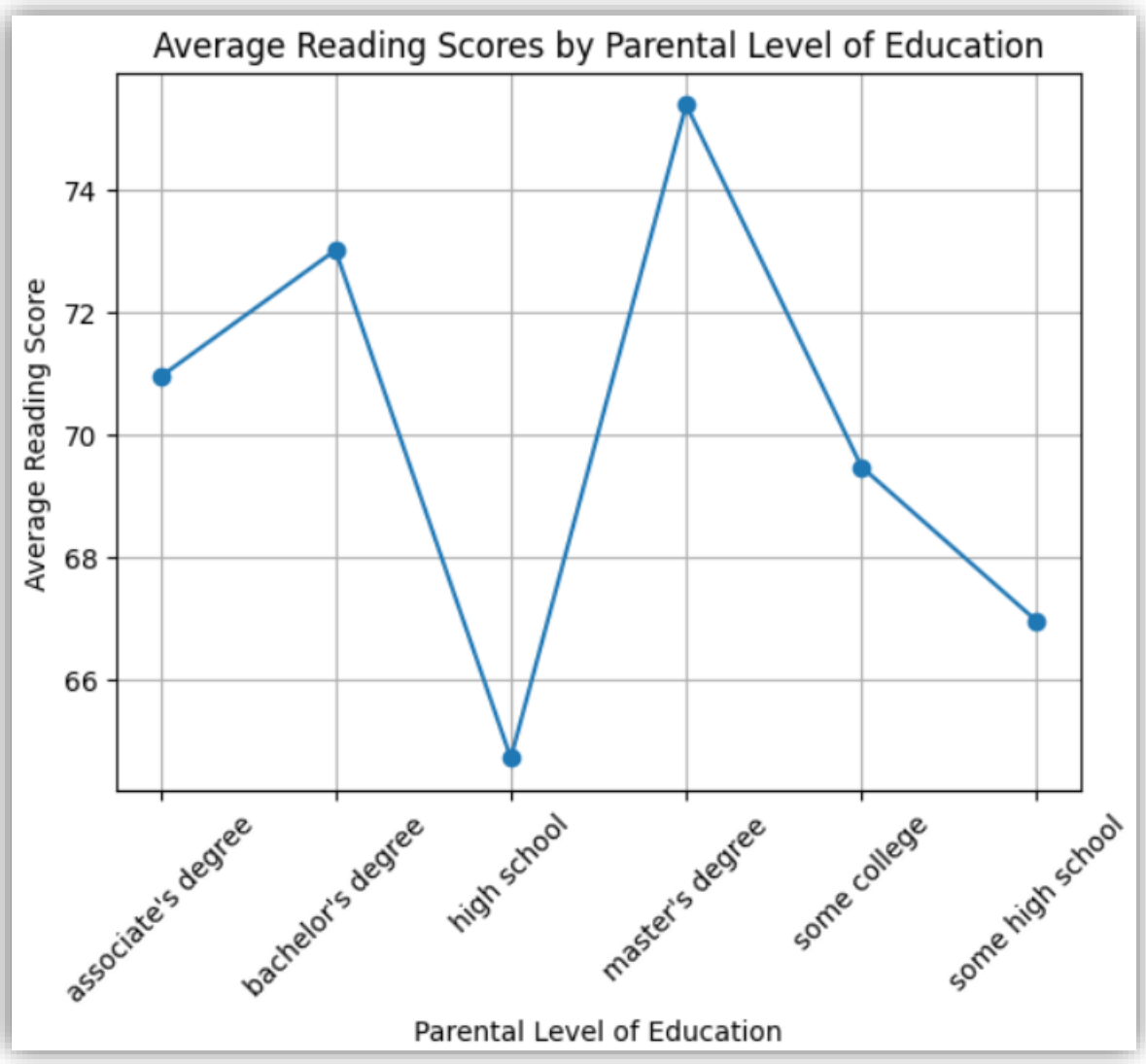


The histogram of math scores reveals most students score between 60 and 80, peaking at a frequency of 250. This concentration suggests a majority average performance in math. The decrease in frequency on either side of the 60 to 80 score range indicates fewer students achieving very high or very low scores.

3. Average Reading Scores by Parental Level of Education

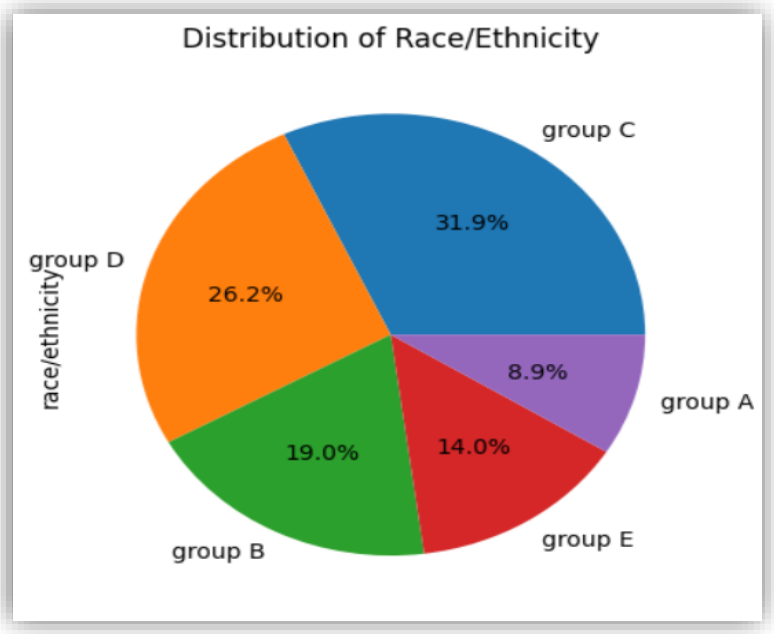
The line plot of average reading scores by parental level of education highlights discernible trends in reading performance based on parental education levels. It's clear that students with parents holding a master's degree tend to have the highest average reading scores, above 74, indicating a possible positive influence of parental education on reading abilities. On the other end, students whose parents only

completed high school score below 66, suggesting a potential area of focus for educational intervention and support. The visualization facilitates easy comparison and analysis of these trends, helping identify areas where additional resources or attention might be beneficial for enhancing students' reading performance.



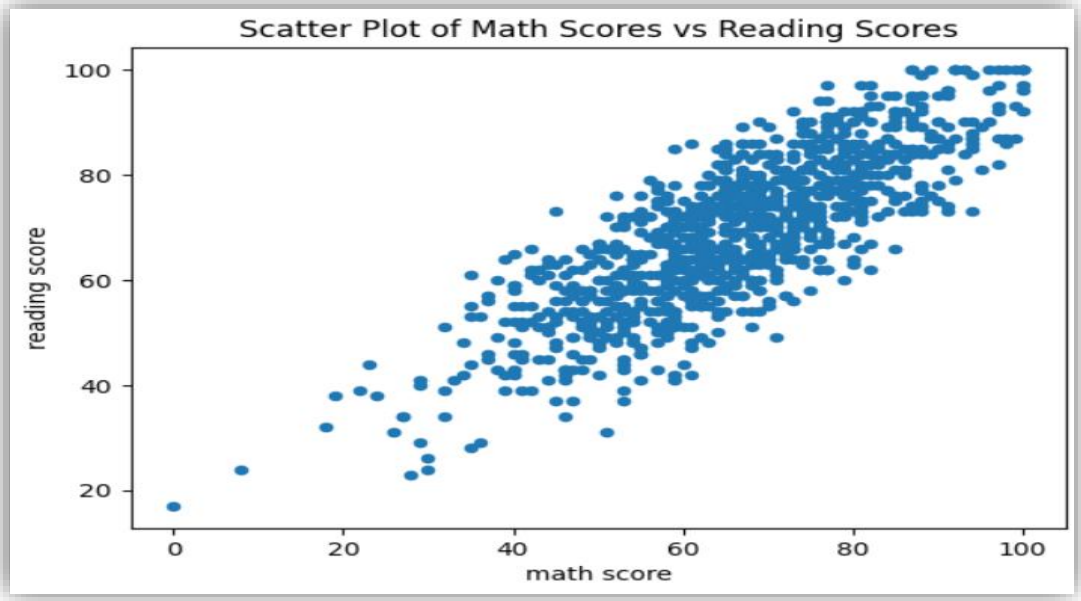
4. Distribution of Race/Ethnicity

This visualization has made it easier to understand the percentage of race/ethnicity of students. It is quite easy to understand by just looking at the pie chart. It gives us the student's race percentage wise,



5. Scatter plot for math vs reading scores.

The scatter plot displaying math versus reading scores reveals a notable correlation between the two subjects, particularly for scores ranging from 50 to 80. This pattern suggests students generally perform similarly in both subjects. This visualization aids in swiftly pinpointing students who excel in one subject but not the other, allowing for targeted educational support and resources to enhance overall academic achievement.


Scatter Plot of Math Scores vs Reading Scores
reading score
math score
100
80
60
40
20
0
20
40
60
80
100