Task 1: Data Cleaning

The preprocessing phase focused on handling missing data and standardizing the dataset's structure. Any gaps in the dataset were identified and filled using conventional techniques like replacing them with mean or mode values. For the Height column, which contained measurements in both feet and inches, a provided formula was applied to convert all entries to centimeters. Numerical and categorical data were properly assigned to their respective data types, and inconsistencies in column names were resolved to align with standard conventions. Certain columns, such as Hometown and FavoriteApp, were removed because they couldn't be meaningfully populated using these methods. The final cleaned dataset was organized and saved as "Cleaned_datafile.csv," ensuring it matched the provided structure and values.

Task 2: Data Exploration

Exploratory Data Analysis (EDA) techniques in Python were employed to investigate the cleaned dataset. Descriptive statistics, including measures like mean, median, and mode, were computed for numerical columns to gain insights into the data's distribution. The dataset was categorized and analyzed to identify trends, such as the average credit amount by age group. Pearson's correlation coefficient was calculated to evaluate relationships between variables. Outliers were identified, and data visualizations were created to better understand the dataset's distribution, structure, and variability. These efforts provided a clearer picture of the dataset's overall characteristics and patterns.

Task 3: Data Visualization

Using the statistical analysis , several visualizations were created to better interpret the dataset. A scatter plot was generated to EXPLAIN the correlation between age and credit amount, highlighting their relationship . Moreover, a bar chart was prepared to display the top 10 female students with the highest weights, ranking them in descending order for clear comparison. These visualizations provided meaningful insights and of the data trends.

self questions:-

1.What is the distribution of weights across different genders?

A box plot was created to compare weight distributions across genders, highlighting key statistics like median, variability, and outliers. This visualization gives clear insights into patterns and differences in weight between genders, enhancing the dataset's interpretability.

2.What is the correlation between age and weight for all students?

A scatter plot was created to display the relationship between age and weight for all students. By using color coding for genders, the plot highlights differences and patterns in the distribution. This visualization helps to identify trends, such as whether weight changes with age or if the relationship changes between genders, providing deeper insights into the dataset's characteristics.

3.What is the distribution of students based on their gender ?

A pie chart was created to visualize the gender distribution within the dataset. The chart displays the proportion of two genders as percentages, with labels and different colors for . This visualization provides an easy-to-understand overview of the dataset's gender composition, highlighting any imbalance in representation.