**PDS Assignment – 1**

**Question 2**

Jithendra Pavuluri

16343746

**Loading the Data:**

- The code starts by loading student performance data from a CSV file named clean_data_file.csv using pandas.

**Printing Dataset Info:**

- The info() function is called to display information about the dataset, including the column names, data types, and non-null counts. This helps understand the structure and completeness of the data.

**Generating DataFrame:**

- A DataFrame named Stu_data is created to hold the loaded dataset.

**Understanding the Data:**

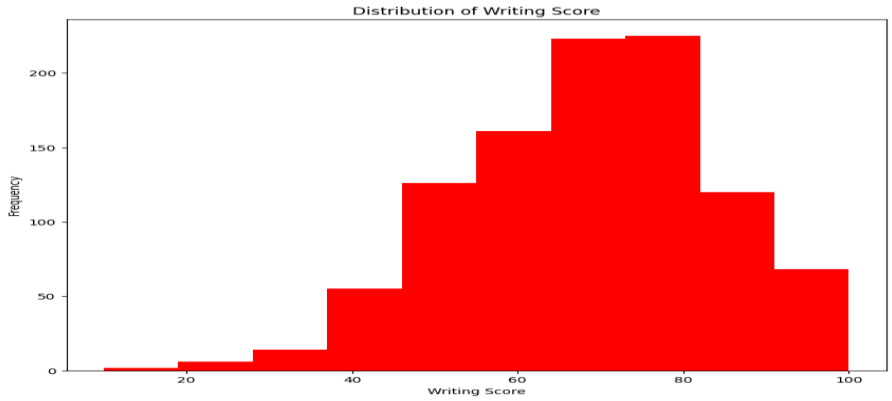
- The code snippet prints:
- Column names using columns.
- Dataset shape (number of rows and columns) using shape.
- First few rows of the dataset using head().
- Summary statistics of the numerical columns using describe().

These steps provide insights into the structure, content, and statistical summary of the dataset.

**Visualizing the Data:**

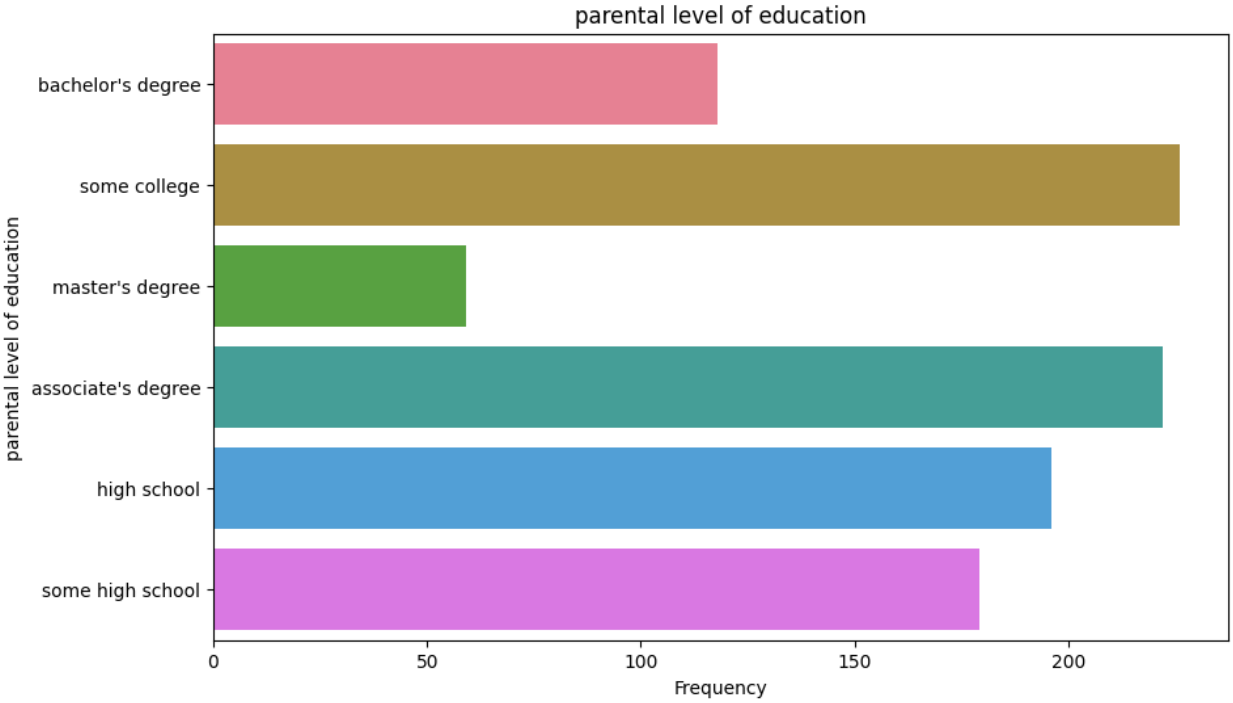
1. **Histogram Plot (Distribution of Writing Score):**

A histogram is plotted to visualize the distribution of writing scores among students.



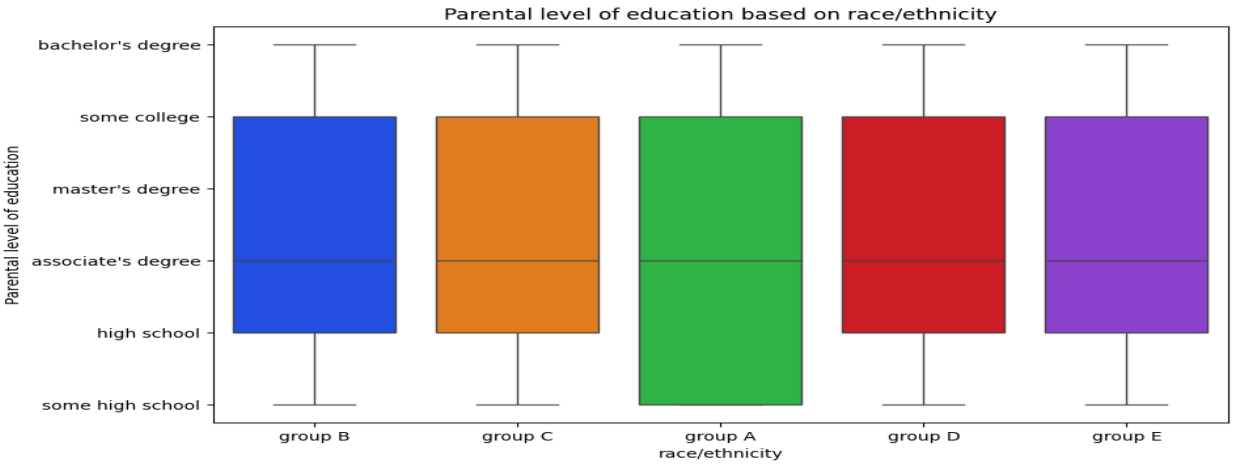
2. **Count Plot (Parental Level of Education):**

A count plot is created to display the frequency of each category in the "parental level of education" feature.



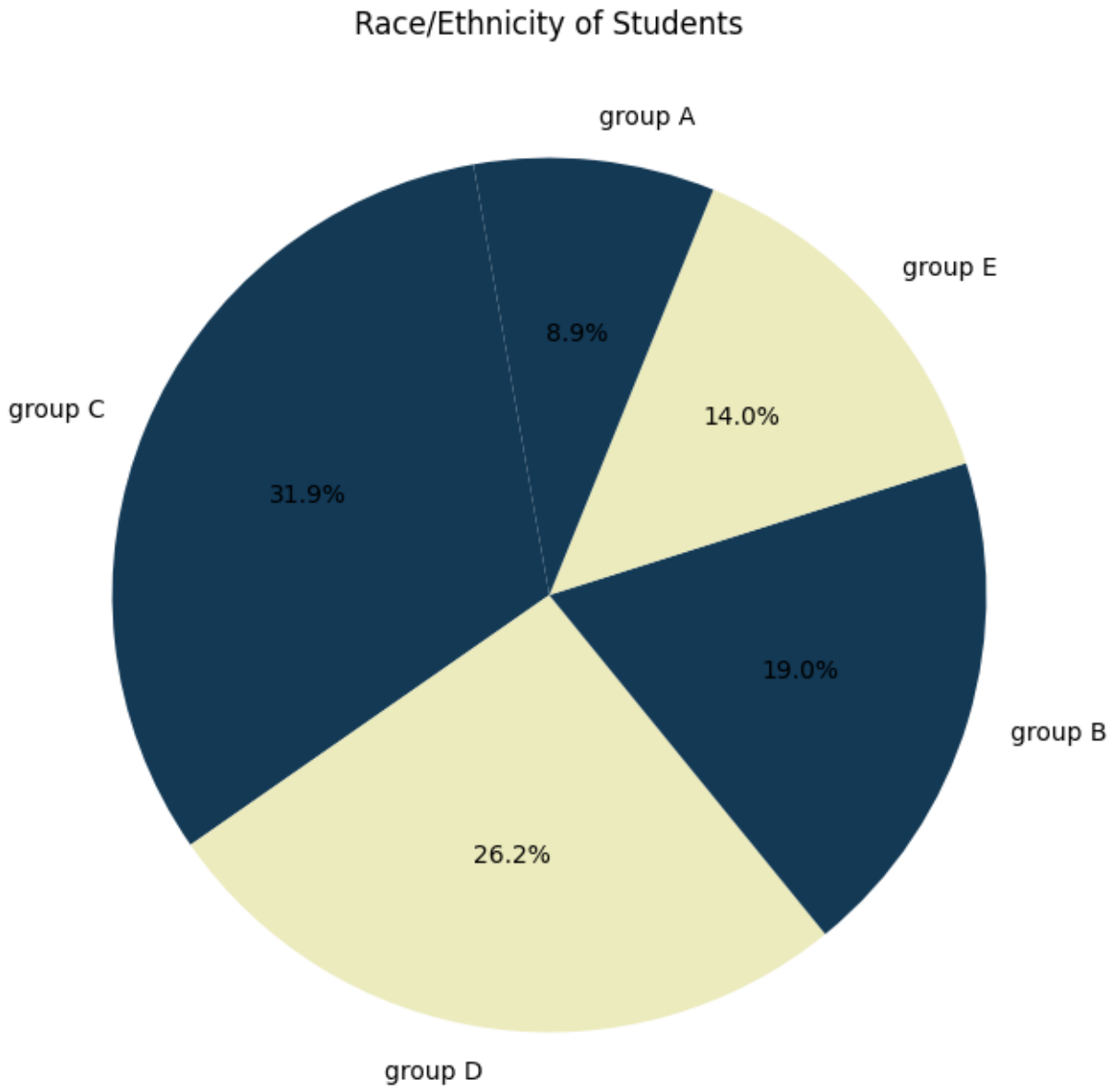

3. **Boxplot (Parental Level of Education vs. Race/Ethnicity):**

A box plot is generated to show the parental level of education based on race/ethnicity, providing insights into educational disparities among different ethnic groups.



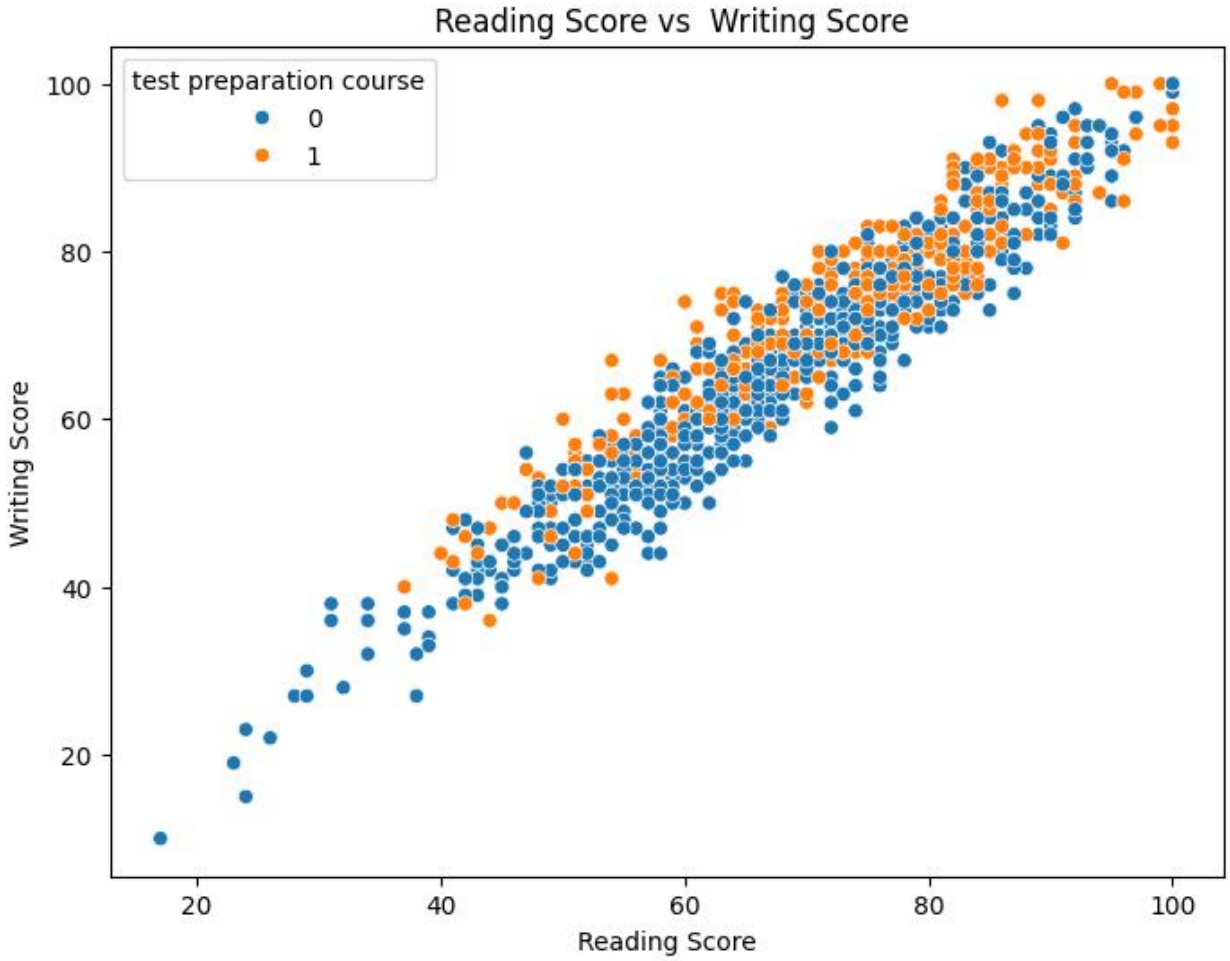
4. **Pie Plot (Race/Ethnicity Distribution):**

A pie chart is drawn to illustrate the distribution of race/ethnicity among students, showing the relative proportions of each group.



5. **Scatter Plot (Reading Score vs. Writing Score):**

A scatter plot is created to explore the relationship between reading scores and writing scores, with the test preparation course indicated by color.




**Results:**

- The visualizations provide insights into various aspects of student performance, including score distributions, demographic representations, and relationships between variables.
- For example, the box plot reveals disparities in parental education levels across different racial/ethnic groups, while the scatter plot indicates the impact of test preparation courses on reading and writing scores.

**Conclusion:**

Overall, the code efficiently loads, explores, and visualizes the student performance dataset, facilitating better understanding and interpretation of the data.