EDA Evaluation

Dataset

1. Replace the null value of all columns with the mode value of those columns.
2. Clean the "WklyStudyHours" column in the dataset, replacing "< 5" with 4, "> 10" with 11, and "5 - 10" with the average of the lower and upper values.
3. Sort the dataset by TestPrep in descending order and then by NrSiblings in ascending order.
4. Create a new column named "ParentEducLevel" that categorizes ParentEduc into two levels: "Higher Education" for those with a bachelor's degree or master’s degree, and "Lower Education" for the rest.
5. Is there a significant difference in the mean weekly study hours between students who completed the test prep and those who did not? Perform a t-test to analyze the data.
6. Use a chi-square test to determine if there is a significant association between the EthnicGroup and ParentEduc variables in the dataset.
7. Use the Shapiro-Wilk test to determine if the distribution of weekly study hours is significantly different from a normal distribution in the dataset.
8. What is the mean weekly study hours for each combination of ParentEduc, LunchType, and PracticeSport?(Do it only into a single Table)
9. What is the median weekly study hours for each combination of Gender, ParentMaritalStatus, and TransportMeans?
10. Create a bar plot to visualize the distribution of students across different EthnicGroups.
11. Create a box plot to compare the distribution of weekly study hours for students who completed the test prep versus those who did not.
12. Create a scatter plot to analyze the relationship between weekly study hours and the number of siblings (NrSiblings).

13. Create a histogram to visualize the distribution of weekly study hours among students.
14. What is the percentage distribution of ParentMaritalStatus categories in the dataset using value_counts?
15. How many unique ParentEduc categories are there in the dataset?
16. Get the rows where the ParentEduc is either "bachelor's degree" or "master's degree" and the LunchType is "standard".
17. Identify and analyze potential outliers in the weekly study hours (WklyStudyHours) column of the dataset. Solve it with Z score and Box plot both methods.
18. Create a new column called "StudyHoursCategory" that categorizes the weekly study hours (WklyStudyHours) into three groups: 'Low' for hours less than 5, 'Medium' for hours between 5 and 10, and 'High' for hours greater than 10. How many students fall into each StudyHoursCategory?