DAI-101 Assignment 1 (EDA)

Conclusion

Data Cleaning

1. My chosen cdataset includes 6607 rows and 20 columns
2. Out of those 20 columns seven were numerical columns and 13 were numerical columns
3. Dataset also included NaN Values after removing them total rows became 6378

Univariate Analysis

1. For univariate analysis I made bar plots of all the categorical data columns
2. For all numerical data columns I made histogram and boxplots
3. After analyzing we got that
    - Majority of students are studying twenty hours per week
    - Majority of students have attendance around 80%
    - Majority of students get seven hours of sleep
    - Most of students take three tutorial sessions per month
    - Marks distribution for current score follows normal distribution without any skewness

Multivariate Analysis

1. Firstly We prepared the heatmap for all the numerical data columns
2. After Analyzing we got that
    - Attendance has highest positive correlation with Exam Score
    - Hours Studied Per Week also have a significant correlation with Exam Score
    - Sleep hours have a close to zero but negative correlation with Exam Score
3. Finally We encoded the categorical data into numerical data and again prepared the heatmap for all the columns
4. After analysing this heatmap we got
    - Tutoring Session and Previous scores also have slightly positive correlation with Exam Scores
    - Distance from home, Teacher quality, access to resources and parental involvement also have negative correlation with exam score.