Statistics for Data Science

Datathon

14 Nov 2021

Student_23.csv

Vanshika Goel	**PES1UG20CS484**	H Section	Roll No: 40

In the dataset that was assigned along with all the questions assigned to me, this is my report.

By making use of the pandas, matplotlib and various other related libraries I solved all the questions.
After reading the .csv file, I kept a cell to keep track of all the data types present in the Data Frame as well as the list of sum of all the null values in the table (before carrying out data cleaning).

Making use of the describe() function, I calculated the minimum and maximum numerical values in the math scores, reading scores and writing scores.

After making necessary changes to the dataset to bring it within the correct range; I calculated the percentage for the 3 subjects.

After cleaning the dataset, by filling all the null values in the math score, reading score and writing score columns, with their respective means, I removed all the rows in the table that had null categorical values.

The data visualization of the reading score column was carried out using the histogram chart.
The histogram obtained was of the right-skewed or negative skewed chart.
Formulae:
Range= Max-Min;
Class width=range/sqrt(no of intervals)
Area of histogram is sum of all the products of frequency densities and class widths.

The new column grades was made on the basis of the percentage column, by making use of the grading system as given in the question.

The plot of a grouped bar chart visualizing the distribution of percentage across parental level of education split by gender was made by making use of the seaborn library and the catplot() function in it.

For the Task Questions, I carried out simple and stratified random sampling; as well as calculating the mean math scores for both the samples obtained. The sampling error was calculated and it was observed that the sampling error in stratified random sampling was lower.

The plots for race in all the 2 new samples as well as for the population was obtained.

All the boxplots, against all the races for the 3 scores, was also plotted using matplotlib library of python.

Maximum number of outliers were visually observed in the race vs reading score plot.