Data Science Workshop-1 (CSE 2195)

ASSIGNMENT-8: DATA CLEANING

Name	CGPA	Percentage of Attendance	Country
Ram	8.8	92	India
Rehman	7.9	91	India
Mary	8.3	90	Pakistan
Joseph	6.2	89	Pakistan
James	5.7	88	UAE
Abdullah	8.6	87	Israel
Abdullah	8.7	87	Turkey
Gita	7.6	86	Ukraine
Krishna	8.1	85	Pakistan
Daniel	8.6	84	India
Rahim	Nan	82	Turkey
Juliet	Nan	83	Nan

1. Create the above table using pandas dataframe.
2. a. Write a python program to print the first 6 rows of the above table.
   b. Write a python program to print the first 6 columns of the above table.
3. Count the number of missing values in each column. Which column has maximum number of missing values.
4. Replace the Nan values in CGPA column, with the mean value of that column.
5. Drop the rows containing Nan values from the table, update the original table, don't create a new one.
6. Compute the minimum, 25th percentile, median, 75th, and maximum of CGPA and Attendance column.
7. Consider the following rule for cgpa conversion to create a grade column, for the above students and add it to the original dataframe.

9-10	A+
8-8.9	A
6.5-7.9	B
0-6.4	F

.

8. Count the number of students passed with different grades.
9. Select students only from India and Pakistan.
10. Create a college name column, where for Ram, Rehman, Joseph, add the college name to ITER, for rests keep it nan.
11. Delete the college name column as it contains so many nan values.

12. Drop the Percentage of Attendance, Country columns and rows below index 5(rows 6 on wards).
13. We want to give fellowships to those, whose attendance is more than 85% and CGPA is more than 8.5. Write a code to find out those student Names only(Not any other information about those students).
14. Write the above program using pandas loc method.
15. In the ”Name” column, convert each individual name to upper case.
16. Update the country column by taking only first 3characters of the Country. For example India will be updated to Ind.
17. Drop the duplicate rows, by considering duplicate names in the Name column.
18. Find the student with highest CGPA. Your output should show only the name and country of the student.
19. Lets see the variation in the CGPA and Percentage of Attendance. Difference between the performance of the best and worst student. (Find the difference between the highest and lowest of CGPA and Percentage of Attendance column.)
20. Sort the students according to the CGPA, and Percentage of Attendance , i.e. if two students CGPA score is same then student with highest Percentage of Attendance will appear first.
21. Create a rank column according to the cgpa of the students.
22. Plot the CGPA vs Percentage of Attendance column. Can you conclude any relation between CGPA vs Percentage of Attendance just by looking at the graph.
23. Find the covariance and correlation between CGPA vs Percentage of Attendance. Can you validate your established relation in previous ques by the covariance and correlation.
24. Use one hot encoding(get_dummies method) on Country column, replace the Country column in the dataframe with the dummies encoded columns.
25. Convert the column names to lower case.
26. Find the top 25% students with respect to the CGPA.(Find the 3rd quartile for the CGPA column, compare all student’s marks with that value, if cgpa is greater than that value, then show the student’s name.)
27. Find the outliers with respect to CGPA column. If any present show their name.
28. Do a random sampling by choosing 5students randomly from the above data set. Find their average performance.