DS203 -- Assignment 4 : Exploratory Data Analysis

Instructions:

1. Submit a single ipynb file at https://moodle.iitb.ac.in/mod/assign/view.php?id=91244
2. Deadline is Sept 4, 2022
3. Use good coding principles, including using intuitive variable names
4. Make use of pandas, numpy, matplotlib, and seaborn as appropropriate
5. Every line of code should have a comment explaining what it does
6. Every cell of code should have a comment cell before or after noting your observations about whether the results are expected or unexpected, and how it correlates with common sense about the real world
7. All sources of inspiration and code (friends' roll numbers, internet links, teacher's notebooks) should be listed at the end as references.

Steps:

1. Read the data from NDAP_REPORT_7004.csv into a pandas dataframe [1]
2. Print the number of columns and rows [1]
3. Print the number of unique values in each column [2]
4. Print the datatype of each column [1]
5. Print a histogram of number missing values for each column sorted from low to high [2]
6. Plot a histogram of number of missing values for each row sorted from low to high [2]
7. If any column's datatype or values needs to be cleaned, then do so [2]
8. For all discrete variables plot histograms [2]
9. For all continuous variables plot histogram with appropriate number of bins [2]
10. For all pairs of continuous variables, plot the scatter plot and show color-coded correlation matrix [2]
11. Check if the sum of columns J through S matches with column I for all rows [2]
12. Divide columns J through S by column I and store the in new columns as percent (multiply by 100) [2]
13. For the new columns, show box and whiskers plot [1]
14. For the new columns, show box and whiskers plot by rural versus urban [2]
15. For the new columns, plot pair-wise scatter plots and correlation matrix [1]
16. For the new columns, plot Q-Q plot for Gaussian distribution [2]
17. Do some additional EDA of your choice and imagination [3]