DS 203 : Programming for Data Science

Assignment Sheet–5: EDA and Data Visualization

Submission guidelines:

- Prepare an ipython notebook and name it <roll no>.ipynb
- Submit it on Moodle before 11:59pm on Oct 2, 2022.

1. For the data source at https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016, perform the following steps in python using pandas, matplotlib and/or seaborn. Use code cells to perform functions with a comment for each line explaining what it is doing (and using intuitive variable names), and mark-down cells to note down any significant observations after each code cell (e.g., "Variable X appears to be normal distributed"):
    a. Perform basic EDA [2]
    b. Discrete variables: [2]
        i. For each variable, identify the mode value using python commands.
        ii. For each variable, compute the entropy to see if there is diversity in the data. The formula is $-\Sigma_i \, p_i \log p_i$, where i is the index of each category, and p_i is its probability N_i/N.
    c. Continuous variables: [3]
        i. For each variable, print mean, variance, skew, min, max, median, 25th percentile, 75th percentile, and inter-quartile range.
        ii. For each variable, plot box-and-whiskers plots.
        iii. For each variable, plot the histogram three times: with too few bins, too many bins, good number of bins.
    d. Pair-wise interaction: [2]
        i. Pick a two discrete-continuous pairs, and plot box-and-whiskers plot for the continuous variable side-by-side for each value of the discrete variable.
        ii. Plot a heatmap of correlation between all pairs of continuous variables.
2. For the data source at https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016, visualize the following data with the appropriate type of graph, and use the right options to make the graph look readable and professional (such as using legends and axes titles with the legible font size, and exploring color palettes): [6]
    a. Pick the top six countries by average yearly suicides, and display their suicide for each year separately.
    b. For the same six countries compare the mix of age groups. What does the plot tell you about the differences or similarities by country?
    c. Plot an appropriate set of graphs or charts that highlight the consistency of difference between males and females when it comes to suicide rates.
    d. Using an appropriate graph, show the worst year for each generation in the US.
    e. Plot a bihistogram for a few specific countries (for a year, say 2000) for male and female populations by age ranges to highlight some differences in sex ratios between countries. Check

out: https://www.itl.nist.gov/div898/handbook/eda/section3/bihistog.htm and https://stackoverflow.com/questions/62678411/how-to-plot-a-paired-histogram-using-seaborn for ideas

f. Show a bubble plot to show the relation between suicide rates, human development index (HDI), and population. Due to the large spread in population, you might have to use a transform. Is there any interesting observation?