DS 203 : Programming for Data Science
Tutorial and Assignment Sheet – 4
Exploratory Data Analysis

Submission guidelines:

- Prepare an ipython notebook and name it <roll no.>.ipynb and submit it on Moodle before 11:59pm on September 16, 2020.

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. Preliminaries:
        i. Read the data file into a data frame.
        ii. Display a portion of the data to get a feel for the dataset.
        iii. Print the number of records.
        iv. Print the number of variables.
        v. Print the datatype of each variable.
        vi. For each variable, print the number of unique values.
        vii. Identify nominal/categorical, ordinal, temporal (time stamps), integer (native but not nominal or ordinal), and continuous variables.
        viii. For each variable, display the number of missing entries.
        ix. Find the number of records with no missing entries.
    b. Discrete variables:
        i. For each variable, plot the frequency of each unique value (histogram).
        ii. For each variable, identify the mode value.
        iii. For each variable, compute the entropy to see if there is diversity in the data.
    c. Continuous variables:
        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.
        iv. For each variable, use QQ-plot to see the extent to which the variable deviates from normal distribution, and how (left-skew, right-skew, or more like uniform distribution)
        v. For each variable, check if the variable deviates is log-normal.
    d. Pair-wise interaction:
        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.
    e. Creative part: Read up on EDA from sources such as https://bolt.mph.ufl.edu/6050-6052/unit-1/role-type-classification/ and https://www.itl.nist.gov/div898/handbook/eda/eda.htm, and perform one more EDA of your choice and share the insight.

2. Repeat the exercise for data at URL https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish.