CDS 1020-1: Introduction to Computational Data Science
Homework 2

Coupling exploratory analysis and statistical inference methods, one can extract meaning from data. Using Python exclusively we will employ various statistical techniques to derive conclusions. For implementations completed solely from scratch (i.e., using NumPy), an additional two points will be given for the respective question (eligible questions labeled with *).

Assigned: Monday, February 20th, 2023, 11:59pm
Deadline: Friday, March 3rd, 2023, 11:59pm

20 points

For the second and third part of the homework, you may perform descriptive and inferential statistics on a National Health and Nutrition Examination Survey (NHANES) dataset available through the Centers for Disease Control and Prevention. The data is stored in the file “NHANES.csv.” For the fourth and fifth part, you may use statistical techniques on the Sleep Study dataset contained in the file “SleepStudy.csv.”

Part I: Distributions

1. Generate random data with a Gaussian distribution. Set the sample size to 100,000. Confirm the values for the mean and standard deviation with print statements. Plot the distribution using a histogram. Is your distribution a standard Gaussian distribution? Explain your reasoning. Share your code and output.

Part II: Conduct Exploratory Analysis on the NHANES Dataset

2. State the number of participants in the dataset. Of the features in the dataset, how many are categorical variables? How many are numerical variables? For any categorical variables, specify at least one nominal variable and one ordinal variable. For numerical variables, specify at least one discrete variable and one continuous variable. Explain the reasoning for your classification. If any of the aforementioned types of variables are absent, please specify.
    a. Examine the data types of each variable. What are the different data types across the dataset? Share your code and output.
3. Calculate the summary statistics of mean, lower quartile, median, upper quartile, IQR, range, minimum, maximum, and standard deviation for the *SleepHrsNight* variable. Share your code and output.

4. Describe the center, shape and spread for the *SleepHrsNight* variable using a plot and text. Do the values for the variable show an approximately normal distribution? Explain your reasoning. Share your code and output.

Part III: Perform Inferential Methods on the NHANES Dataset

5. Based on this sample, do people in the general population get on average more than 8 hours of sleep per night (*SleepHrsNight* > 8)? Conduct a hypothesis test, specifying the null and alternative hypotheses, using the notation H_0 and H_A. Is this a one-tailed or two-tailed hypothesis test? Use a t-statistic as the test statistic in your formulation and an alpha value of 0.05. Explain your reasoning for rejecting or failing to reject the null hypothesis. Share your code and output. *
6. Calculate the standardized amount of sleep or z-value for a participant in the study that gets 10 hours of sleep (*SleepHrsNight* = 10). Based on its sign (+ or -), what does this z-value indicate relative to the population mean? Hint: this will require calculating the mean and standard deviation of the sampling distribution for the feature in this sample. Explain your reasoning. Share your code and output. *

Part IV: Conduct Exploratory Analysis on the Sleep Study Dataset

7. Calculate the summary statistics of mean, lower quartile, median, upper quartile, IQR, range, minimum, maximum, and standard deviation for the *AverageSleep* variable. Share your code and output.
8. To prepare for regression analysis (performed in Unit 2), call the Pandas function get_dummies( ) to convert a chosen categorical feature into a numerical feature. State the chosen feature. Explain in detail why this method subsequently requires dropping a column. Share your code and output.

Part V: Perform Inferential Methods on the Sleep Study Dataset

9. Based on the sample, estimate the proportion of those in the general population who have had an all-nighter (*AllNighter* = 1), using hypothesis testing. Specify the null and alternative hypotheses, using the notation H_0 and H_A. Is this a one-tailed or two-tailed hypothesis test? Use a z-statistic as the test statistic in your formulation and an alpha value of 0.05. Explain your reasoning for rejecting or failing to reject the null hypothesis. Share your code and output. *
10. By constructing a confidence interval for the sample data, determine if there is a true difference between the average *DepressionScore* and the average *AverageSleep* in the population. Use a confidence level of 0.95. Explain your reasoning. Share your code and output. *

Submission:
Submit a standard Python file (.py) with your code using the name “[last name]_HW2” to Canvas. For visualizations and text answers, please use a document (.pdf) with appropriate numbering. The submission to Canvas will consist of these two files.