**IBEHS 4C03: STATISTICAL METHODS IN BIOMEDICAL ENGINEERING**

**ASSIGNMENT #1**

**Topics:**

Read and Explore Data

Data Cleaning

Data Visualization

Descriptive Statistics

Probability

**Total marks: 100**

**Submission Instructions:** Electronic submission to the IBEHS 4C03 Avenue to Learn Assignment 1 folder.

**Due:**

**Assignment Submission Format:**

File Names: StudentLastNameFirstNameAssignment#.doc/pdf/py//ipynb etc.

Both your name and student numbers should appear at the top of the document. If separate documents are submitted, then you should submit fully supported answers to the questions in a single document including the plots you made in python, and refer to coding. Regardless, the python file used to generate the plots and any provided answers should also be submitted with any assignment, along with a pdf or word document conversion of the python file.

**Problem 1: [20 points/100 points] Descriptive Statistics**

**Hospital dataset**: The data in the Hospital.csv dataset is collected on people discharged from a hospital as part of a retrospective chart review of antibiotic usage. Use the Hospital dataset to answer the following problems.

1. It is of clinical interest to know if the duration of hospitalization stay (in days of stay) is affected by whether a patient has received antibiotics (Antibiotics=1, No Antibiotics=0). Upload the Hospital dataset. Find the mean, median, 5-number summary information, sample variance, and sample standard deviation of the two subsamples. Discuss and show all the steps to finding these statistics. Write out what these descriptive statistics mean theoretically and specifically for this data. Discuss the distributions of the duration of stay.
2. Next, explore this question descriptively using boxplots for each subsample (use the closest point to 1.5\*IQR for the whisker length). Discuss each of the elements of the boxplot. What does the boxplot tell you about the central tendency and the variability of the data? Discuss the distributions of the duration of stay.
3. Create histograms of the two subsamples and comment on the steps to create histograms, the components, and the distributions.
4. Are there any outliers present in the two subsamples? What impact do they have on the data? Discuss how this affects the answer to question a) if you remove any identified outliers.

**Problem 2: [20 points/100 points] Data Exploration and Data Cleaning**

**Heart Disease dataset: I**mport the heart\_disease\_needs\_data\_cleaning dataset into your statistical software. This dataset is based on the Cleveland Heart Disease dataset, where 303 patients had clinical attributes recorded and the presence or absence of coronary artery (heart) disease. The variables that we will focus on is

* Age (in years old)
* Slope (the slope of the peak exercise ST segment of the ECG: Value 1 = upsloping; Value 2 = flat; Value 3 = down-sloping
* Heart\_Disease (0 is no heart disease and 1 is heart disease). This diagnosis is based on angiographic disease status with Value 0 < 50% diameter narrowing and Value 1 ≥ 50% diameter narrowing.

Review the dataset and inspect the data for data cleaning/data preprocessing.

1. Let’s first look for completeness. For each column of data, how many data values and data rows contain missing data? Briefly explain what you could do with these data values. Now go ahead and delete the rows that have any missing values.
2. Next, look for data-type validity. Since we care about Age, Slope, and Heart\_Disease, let’s check that the data type matches what you would want the data to be used for in an analysis of heart disease. What are the data types for Age, Slope, and Heart\_Disease in the dataset? If the data type does not match what you want it to be, discuss what the data type is that you want and why, and the reasons that you find in the dataset for any datatype mismatch. How are you going to go about fixing these problems? (Hint: There are a lot of way to do this: If you need to remove any values that do not match your desired data type, you can delete the row of data based on the datatype mismatch once you identify, or you can change mismatches to NaNs and then deal with those missing values by deleting by row). In the end, you want the data types to be what you want specified.
3. Now look at data validity. What do you think the limits of plausible data would be for Age, for Slope, and for Heart\_Disease? Go ahead and clean the data based on deleting any nonsense values. Again, delete entire rows to end up with a clean dataset of complete data on Age, Slope, and Heart\_Disease.
4. Create a new variable based on your Age variable that represents being in a category of <=40, 41-50, 51-60, and >60 years old.
5. Choose a graph to represent each of the variables for Age, Age\_Category, Slope, and Heart\_Disease and discuss these variables and why you chose the visualization that you did. Show your graphs and report on how you cleaned these variables for possible future analysis.

**Problem 3:** **[20 points/100 points]: Data Visualization**

Use the following site to find a dataset. Use the dataset to create and discuss one bar graph and one time-series graph. Create the graphs in Python.

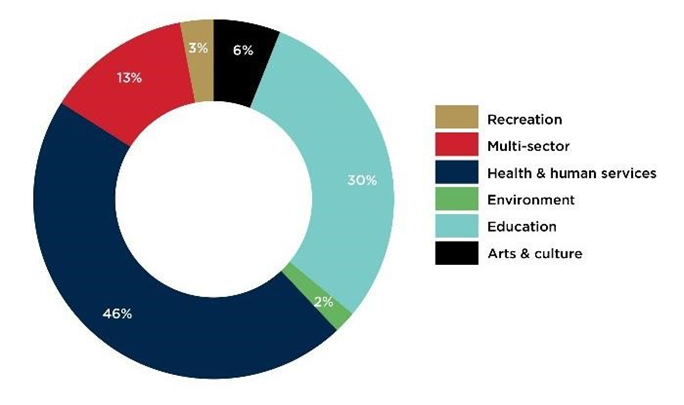
Build interesting graphs and describe what you are showing in the data. Let the graph describe what you are attempting to show, and then describe what you are showing in terms of comparisons or patterns or trends. Aside from using descriptive statistics, also describe what you are visually showing in your graphs in regular words. Submit the graphs and your explanations for your answers. Try to use the guidelines for excellent data visualizations for your graphs.

<https://data.ontario.ca/organization/health>

<https://data.ontario.ca/dataset/confirmed-positive-cases-of-covid-19-in-ontario>

**Problem 4: [20 points/100 points]: Data Visualization**

1. The following figure (from Hamilton’s Community Foundation’s 2019-20 annual report) shows the percentage of grants approved by sector. Using concepts you about making good data visualizations, identify at least two major problems with how this graph was designed. Create a better graph in Python.

****

1. Find a graph from a journal or publication from your discipline and provide a critical appraisal of the visualization of the data. This may be from a journal publication, textbook, news article, lab report, etc. List the features that make it a good or a poor visualization of data. If it is a good plot, discuss why this is, and if it is a bad plot, comment on how you could improve the visualization.

**Problem 5: [20 points/100 points]: Probability**

1. Discuss the General Probability Rules that specify the following:
   1. The range of probabilities
   2. The sum of probabilities
   3. The Complement Rule
   4. The Multiplication Rule for independent events and for dependent (or conditional) events
   5. The Addition Rule for mutually exclusive events
   6. The Addition Rule for non-mutually exclusive events
   7. The Conditional Probability Rule: Bayes Rule
2. A circuit is shown in the following figure. The switches for the circuit operate independently of one another, and the probability of each switch is:

* The probability of A, B being closed is 0.75, 0.9 respectively.
* The probability of C being open is 0.20.

What is the probability that the circuit works? (Note: work these questions out by hand and show your work.)

Shape

Description automatically generated with medium confidence

1. We have four boxes with a composition of defective light bulbs as follows: Box Bi

contains 5%, 40%, 10%, and 25% defective light bulbs for i = 1, 2, 3, and 4, respectively. Pick a box and then pick a light bulb from that box at random. What is the probability that the light bulb is defective?

1. Continuing with the boxes of variably defective light bulbs, suppose the light bulb that you pick is defective. What is the probability that it came from Box 2?