Instructions

- This homework assignment is worth 69 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: September 29, 2023 by 11:59 pm.**

Exercise 1

(4 points) What is the difference between histograms and bar charts? Explain.

Exercise 2

(4 points) Data on 1,500 students' height were collected at a larger university in the East Coast. Which of the following is the best chart for presenting the information?

(a) A pie chart

(b) A Pareto chart

(c) A side-by-side bar chart

(d) A histogram

Exercise 3

(4 points) Which of the following statements about the median is **not true**?

(a) It is more affected by extreme values than the arithmetic mean.

(b) It is a measure of central tendency.

(c) It is equal to Q_2.

(d) It is equal to the mode in bell-shaped "normal" distributions.

Exercise 4

(4 points) Which of the following statistics **can not** be determined from a box-plot?

(a) The median

(b) The mean

(c) The first quartile

(d) The third quartile

(e) The interquartile range

Exercise 5

Following are the summary statistics of for one data set A.

Min	0.066
Q_1	1.42
Q_2	2.60
Q_3	6.02
Max	10.08

(a) (3 points) Compute the range and the interquartile range for the data set A

(b) (3 points) Do the summary statistics for the dataset A provide enough information to construct a box-plot? Explain why.

Exercise 6

(4 points) If the largest value of a data set is doubled, which of the following is **FALSE**?

(a) The mean increases.

(b) The standard deviation increases.

(c) The interquartile range increases.

(d) The range increases.

(e) The median remains unchanged.

Exercise 7

(5 points) The five-number summary for scores on a DATA-101 exam are:

$$\begin{aligned} \text{Min} &= 35 \\ Q_1 &= 68 \\ Q_2 &= 77 \\ Q_3 &= 83 \\ \text{Max} &= 97 \end{aligned}$$

In all 196 students took the test. About how many had scores between 77 and 83?

Exercise 8

Consider the automobile data set posted on blackboard. The Automobile dataset has a different characteristic of an auto such as body-style, wheel-base, engine-type, price, mileage, horsepower and many more. **In Python**, answer the following:

(a) (3 points) Using the `pandas` library, read the csv datafile and create a data-frame called `autos`.

(b) (3 points) Using the appropriate commands, report the number of observations and variables.

(c) (4 points) Create a bar chart of `body_style`. Comment on the chart.

(d) (4 points) Create a boxplot of price. Comment on the plot.

(e) (5 points) Create a scatter-plot between `horsepower` and `average_mileage`. Comment on the plot.

Exercise 9

Consider the Wisconsin breast cancer diagnostic data set from the UCI Machine Learning Repository (`for more info click here`). The Wisconsin breast cancer data set contains information on 569 biopsies, each with 32 features. One feature is an id number, another is the cancer diagnosis, "`M`" to indicate malignant or "`B`" to indicate benign. The other 30 numeric measurements comprise the mean, standard error, and worst (that is, largest) value for 10 different characteristics of the digitized cell nuclei. Notice that all these features are related to the shape and size of the cell nuclei. **In Python**, answer the following:

(a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `breast_cancer`.

(b) (3 points) Using the proper commands, report the number of observations and variables.

(c) (4 points) Create a pie chart of diagnosis. Comment on the chart.

(d) (4 points) Create a side-by-side boxplot of `radius_mean` for two types of diagnosis. Comment on the plot.

(e) (5 points) Create a scatter plot of `radius_mean` and `area_mean`. Comment on the plot.