Visualization of Complex Data

DATS 6401

LAB # 5

In this assignment, you will learn how to detect outliers and how to remove them, perform various normality tests and interpret the result of the statistical test. The dataset for this Homework can be found on the course GitHub under 'height-weight.csv.' Display float numbers with 2-digit decimal precisions.

1. Load the height-weight.csv file from the course GitHub and plot line plot [only first one hundred samples] weight and height of females in one graph versus the number of observations [x-axis # of observations]. Add appropriate x-axis label, y-axis label, title [raw data], and legend to your plot. Write down your observation on the sample mean and sample variance from the graph. [5 pts]

2. Plot the histogram of the weight and height of females in on graph. Add appropriate x-axis label, y-axis label, title, and legend to your plot. Write down your observation on the sample mean and sample variance from the graph. [5 pts]

3. Perform a z-transform (without using the libraries and by implementing $z = \frac{y-\bar{y}}{std(y)}$) and plot [only the first one hundred samples] the z-score of weight and height of females versus the number of observations [x-axis # of observations]. Add appropriate x-axis label, y-axis label, title [transformed data], and legend to your plot. Write down your observation on the sample mean and sample variance from the graph. [5 pts]

4. Plot the histogram of the z-transformed weight and height of females in on graph. Add appropriate x-axis label, y-axis label, title, and legend to your plot. Write down your observation on the sample mean and sample variance from the graph. [5 pts]

5. Create a two-by-two subplot that shows the above figures in one subplot. [5 pts]

    a. The 1,1 shows plot of question 1.
    b. The 1,2 shows plot of question 3.
    c. The 2,1 shows the plot of question 2.
    d. The 2,2 shows the plot of question 4.

6. What is the probability that a weight of female be more than 170lb and what is the probability that a lady be 5.5 feet [66 inch] and taller? Use the scipy package in python to answer this question. The program should display the following information on the console: [10 pts]

```
import scipy.stats as st

Sample mean of the lady's weight is ____ lb
Sample mean of the lady's height is ____inches.
Sample std of the lady's weight is _____ lb.
Sample std of the lady's height is _____inches.
The median of the lady's weight is _____lb.
The median of the lady's height is ______ inches.
The probability that a lady weighs more than 170lb is ___%
The probability that a lady be taller than sixty-six inches is ___%
```

7. Normality Test: Plot the QQ plot of the female height. Does the data look Normal? Explain Why? Add appropriate title to the plot. Hint: You need to use the following package: [5 pts]

```
from statsmodels. graphics.gofplots import qqplot
```

8. Normality Test: Plot the QQ plot of the female weight. Does the data look Normal? Explain Why? Add appropriate title to the plot. Hint: You need to use the following package: [5 pts]

```
from statsmodels. graphics.gofplots import qqplot
```

9. Perform a K-S Normality test on the female height and weight data. Display the p-value and statistics of the test for the height and weight [for height and weight a separate test is needed]. Interpret the K-S test [Normal or Not Normal with 99% accuracy] by looking at the p-value. Display the following information on the console: [5 pts]

```
K-S test: statistics= _____ p-value = ______
K-S test:  Weight Female dataset looks _______
```

```
K-S test: statistics= _____ p-value = ______
K-S test:  Height Female dataset looks _______
```

10. Repeat Question 9 with the "Shapiro test". [5 pts]

```
Shapiro test: statistics= _____ p-value = ______
Shapiro test:  Weight Female dataset looks _______
```

```
Shapiro test: statistics= _____ p-value = ______
Shapiro test:  Height Female dataset looks _______
```

11. Repeat Question 9 with the "D'Agostino's K^2 test". [5 pts]

```
da_k_squared test: statistics= _____ p-value = ______
da_k_squared test:  Weight Female dataset looks _______
```

```
da_k_squared test: statistics= _____ p-value = ______
da_k_squared test:  Height Female dataset looks _______
```

12. Using Python calculate Q1, Q3 and IQR for the female height. Then find which range of heights in this dataset is an outlier. Display the following information on the console: [7.5 pts]

```
Q1 and Q3 of the female height is ____ inches & ____ inches.
IQR for the female height is ___ inches.
```

```
Any height < ___ inches and height > ___ inches is an outlier.
```

13. Plot the boxplot of the female height and verify the results in the previous question by looking at outliers on the boxplot graph. [5 pts]

14. Clean the dataset by removing the heights that are outliers in the dataset. Plot the boxplot of the female height after the outliers are removed. Do you still observe outlier(s) in the boxplot of the cleaned data? [7.5 pts]

15. Using Python calculate Q1, Q3 and IQR for the female weight. The find which range of weights in this dataset is an outlier. Display the following information on the console: [7.5 pts]

```
Q1 and Q3 of the female weight is ____ lb & ____ lb.
IQR for the female weight is ___ lb.
Any weight < ___ lb and weight > ___ lb is an outlier.
```

16. Plot the boxplot of the female weight and verify the results in the previous question by looking at outliers on the boxplot graph. [5 pts]

17. Clean the dataset by removing the weights that are outliers in the dataset. Plot the boxplot of the female weight after the outliers are removed. Do you still observe outlier(s) in the boxplot of the cleaned data? [7.5 pts]

Upload a formal **report (as a single pdf**) plus **the .py file** through Canvas by the due date.