DATS 6101 - Introduction to Data Science

Homework Assignment 3

Create a new Rmarkdown file and complete the below tasks.

Pima Dataset

This exercise uses the Pima.te dataset from the MASS package. The dataset includes measurement on a population of women of Pima Indian heritage living near Phoenix, Arizona. The women were tested for diabetes according to World Health Organization criteria.

The variables in the dataset are:

- npreg: number of pregnancies.
- glu: plasma glucose concentration
- bp: diastolic blood pressure (mm Hg) skin triceps skinfold thickness (mm)
- bmi: body mass index
- ped: diabetes pedigree function
- age: age in years
- type: Yes or No: diabetic by WHO criteria

library(MASS)

Question 1
Use the summary function to get a summary of all the variables in the Pima.te dataset.

Question 2
Print the structure of pima dataset. Something like this:

```
$ npreg   : int  6 1 1 3 2 5 0 1 3 9 ...
$ glu     : int  148 85 89 78 197 166 118 103 126 119 ...
$ bp      : int  72 66 66 50 70 72 84 30 88 80 ...
```

Question 3
Get the variables from the dataset. Example:

```
[1] "npreg"    "glu"      "bp"       "skin"     "bmi"      "ped"
```

Question 4
For bmi and age variables find out the followings:

- The five-number summary and the mean
- Range
- Number of observations

Question 5
How many women are in this dataset? Answer in complete sentences.

Question 6
Select the first 5 observations and first 4 columns/variables from the dataset. Example:

	npreg	glu	bp	skin
1	6	148	72	35
2	1	85	66	29
3	1	89	66	23
4	3	78	50	32
5	2	197	70	45

Question 7
Select the records where bmi is greater than or equal to 50.

Question 8
What percentage of the women have diabetes by WHO criteria. Answer in complete sentences.

Question 9
Obtain a histogram for body mass index.

Question 10
Use this dataset to obtain the T-interval estimate for bmi of the population at 99% and 80% confidence levels. Answer in complete sentences.

CSV File

This dataset contains data on 671 infants with very low (<1600 grams) birth weight from 1981-87 were collected at Duke University Medical Center by Dr. Michael O'Shea, now of Bowman Gray Medical Center. Dataset info can be found here:
http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/Cvlbw.html

Question 11
Import the dataset from "vlbw.csv" file.

Question 12
Obtain a histogram of the length of stay, i.e. the number of day the infants stay in the neonatal intensive care unit (variable hospstay). You might find the histogram way skewed by some extreme outliers. Fix them. (The outlierKD function is one option, for example.)

Question 13
Draw a boxplot for variable lowph.

Question 14
The variable lowph contains the lowest pH in the first 4 days of life. Obtain a histogram of this variable (the variable is called lowph).

Question 15
Obtain the T-interval for lowph at 99.9% confidence level.