Practice (HW) 1

Due: February 07, 2019

- Asking questions to TAs and collaborating with classmates are encouraged, but copying, sharing, or distributing any material is *strictly prohibited*. Homework should be students' original work.
- Please submit
    1) SAS code (.SAS) with detailed comments
    2) PDF document with relevant output and interpretations
- Late homework will not be accepted.

1. Pima Indians

This dataset is from a study conducted by the National Institute of Diabetes and Digestive Kidney Diseases on 768 adult female Pima Indians living near Phoenix. The dataset contains the following variables:

Variable	Description
pregnant	Number of pregnancies
glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
blood	Diastolic blood pressure (mmHg)
triceps	Triceps skin fold thickness (mm)
insulin	2-hour serum insulin (mu U/ml)
bmi	Body mass index (weight in kg/(height in m^2))
pedigree	Diabetes pedigree function
age	Age (years)
test	Whether the patient showed signs of diabetes (0=negative, 1=positive)

The data may be obtained from UCI Repository of machine learning databases at http://www.ics.uci.edu/~mlearn/MLRepository.html.

a) Create a library called 'practice' and import the dataset named 'pima' in the library. Try all 3 different types.

1) Pima.txt 2) Pima.xlsx 3) Pima.csv

b) Label the variables as follows:

Variable	Label
pregnant	Number of times pregnant
glucose	Plasma glucose concentration
blood	Diastolic blood pressure (mmHg)
triceps	Triceps skin fold thickness (mm)
insulin	2-hour serum insulin (mu U/ml)
bmi	Body mass index
pedigree	Diabetes pedigree function
age	Age (years)
test	Sign of diabetes

Print the first 5 observations of dataset 'pima' with the labels above.

c) Create the following formats:

Format name	Values
posneg	1 = Positive, 0 = Negative
Bmilevel	BMI < 18.5 = Underweight
	18.5 ≤ BMI <25 = Healthy
	25 ≤ BMI < 30 = Overweight
	BMI ≥ 30 = Obese

Print the first 5 observations of dataset 'pima'. Apply the formats 'posneg' and 'bmilevel' for variables **test** and **bmi**, respectively.

d) Create a new dataset 'young': Include subjects younger than 27 years old with no sign of diabetes. Use the formats created in **c)** in DATA step for permanent formatting. Keep the following variables only: glucose, triceps, bmi, age, and test.

Print the first 5 observations of dataset 'young'.

e) Create two new datasets 'bplow' and 'bphigh': Use one DATA step to create two datasets 'bplow' and 'bphigh' from the dataset 'pima'. A subject will be included in the dataset 'bplow' if diastolic BP < 60. If one's diastolic BP > 80, then that individual will be included in the dataset 'bphigh'. Keep only 4 variables (BMI, glucose, blood, and insulin).

Print the first 5 observations of two datasets 'bplow' and 'bphigh'.

2. Athletic Shoes

A distributor of athletic shoes has an end of summer sale. He wants to combine information from two data files to create a new one containing only the article style and the final price (after adjustment).

Data 'Regular'

Style	ExerciseType	RegularPrice
MaxFlight	Running	142.99
LightStep	Walking	73.99
ZoomAirborne	Running	112.99
ZipSneak	C-Train	92.99

Data 'Discount'

ExcerciseType	DiscountRate
Running	0.30
Walking	0.20
C-Train	0.25

a) Merge two datasets by ExerciseType and name the merged dataset 'Shoes'.

b) In the dataset 'Shoes', create a new variable 'NewPrice', showing the price after adjustment. Print only the shoe styles and the new prices.