# IC272-Data Science III


## BY ADARSH SANTORIA (B21176)


You are given the Pima Indians Diabetes Database as a csv file. This data-set is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females with at least 21 years old of Pima Indian heritage. It contains following 9 attributes.

A) pregs: Number of times pregnant
B) plas: Plasma glucose concentration 2 hours in an oral glucose tolerance test
C) pres: Diastolic blood pressure (mm Hg)
D) skin: Triceps skin fold thickness (mm) test: 2-Hour serum insulin (mu U/mL)
E) BMI: Body mass index (weight in kg/(height in m)^2)
F) pedi: Diabetes pedigree function
G) Age: Age (years) class: Class variable (0 or 1)

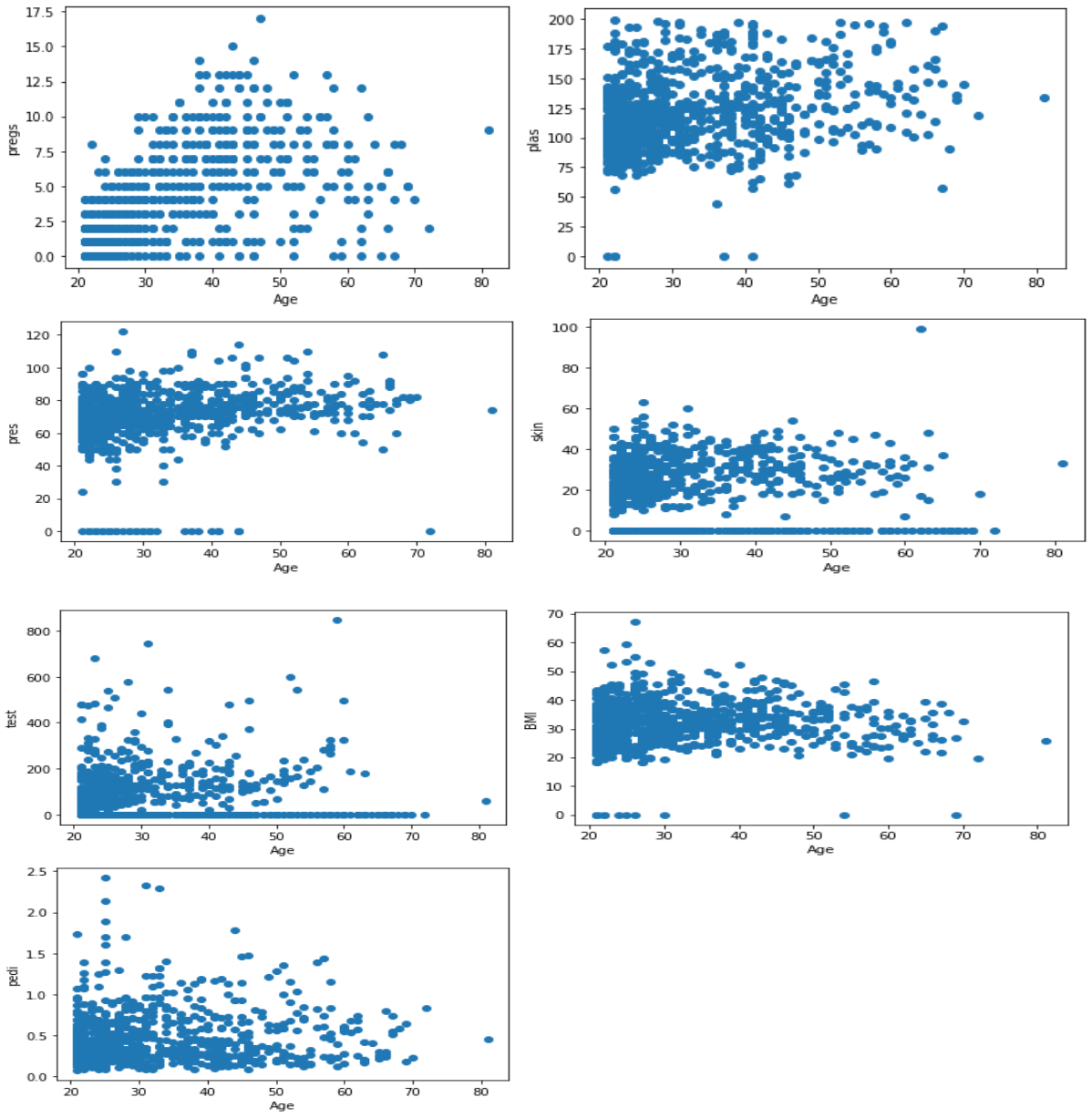
1. Mean, median, mode, minimum, maximum and standard deviation for all the attributes excluding the attribute ‘class’.

```
mean of pregs = 3.8450520833333335
median of pregs = 3.0
mode of pregs = 1
min of pregs = 0
max of pregs = 17
standard deviation of pregs = 3.3673836124089958
mean of plas = 120.89453125
median of plas = 117.0
mode of plas = 100
min of plas = 0
max of plas = 199
standard deviation of plas = 31.95179590820272
mean of pres = 69.10546875
median of pres = 72.0
mode of pres = 70
min of pres = 0
max of pres = 122
standard deviation of pres = 19.343201628981696
mean of skin = 20.536458333333332
median of skin = 23.0
mode of skin = 0
min of skin = 0
max of skin = 99
standard deviation of skin = 15.941828626496939
mean of test = 79.79947916666667
median of test = 30.5
mode of test = 0
min of test = 0
max of test = 846
standard deviation of test = 115.16894926467262
mean of BMI = 31.992578124999998
median of BMI = 32.0
mode of BMI = 32.0
```

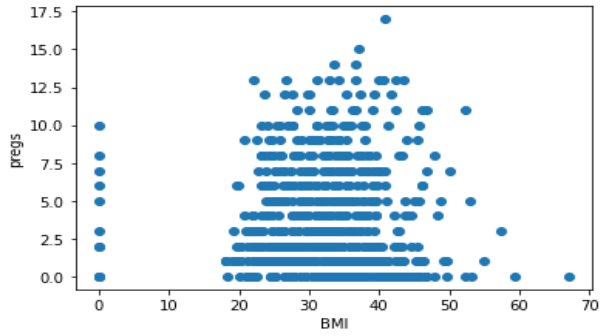
```
min of BMI = 0.0
max of BMI = 67.1
standard deviation of BMI = 7.87902573154013
mean of pedi = 0.47187630208333325
median of pedi = 0.3725
mode of pedi = 0.254
min of pedi = 0.078
max of pedi = 2.42
standard deviation of pedi = 0.3311128160286291
mean of Age = 33.240885416666664
median of Age = 29.0
mode of Age = 22
min of Age = 21
max of Age = 81
standard deviation of Age = 11.752572645994181
```

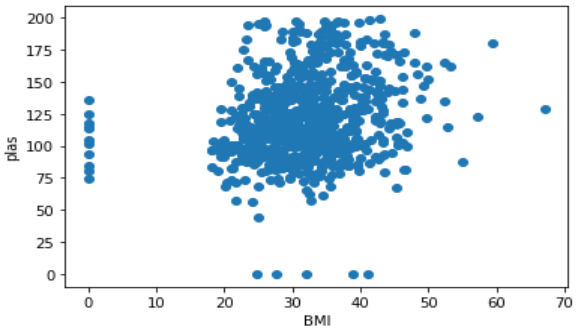
2. Obtain the scatter plot between

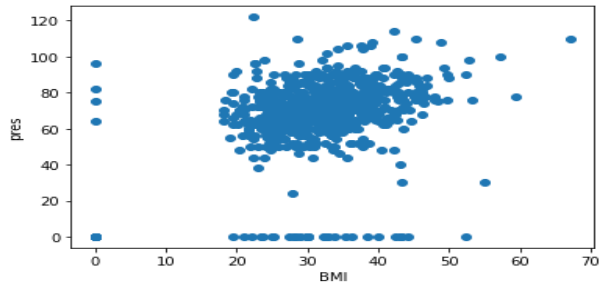
a. ‘Age’ and each of the other attributes, excluding ‘class’

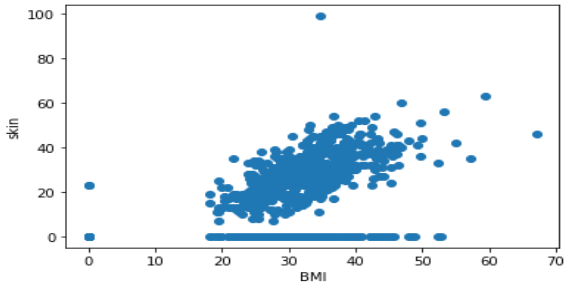


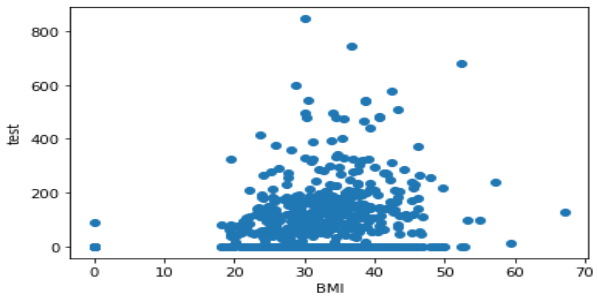
b. ‘BMI’ and each of the other attributes, excluding ‘class’ (You can use matplotlib library).

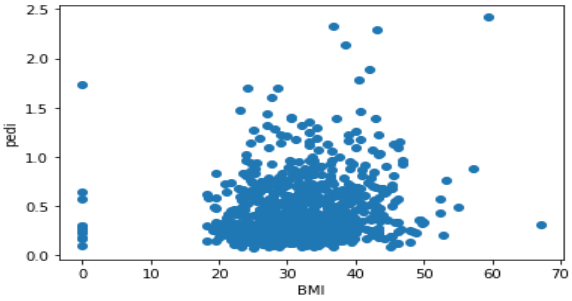


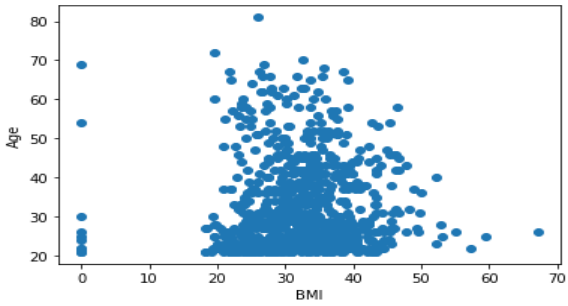



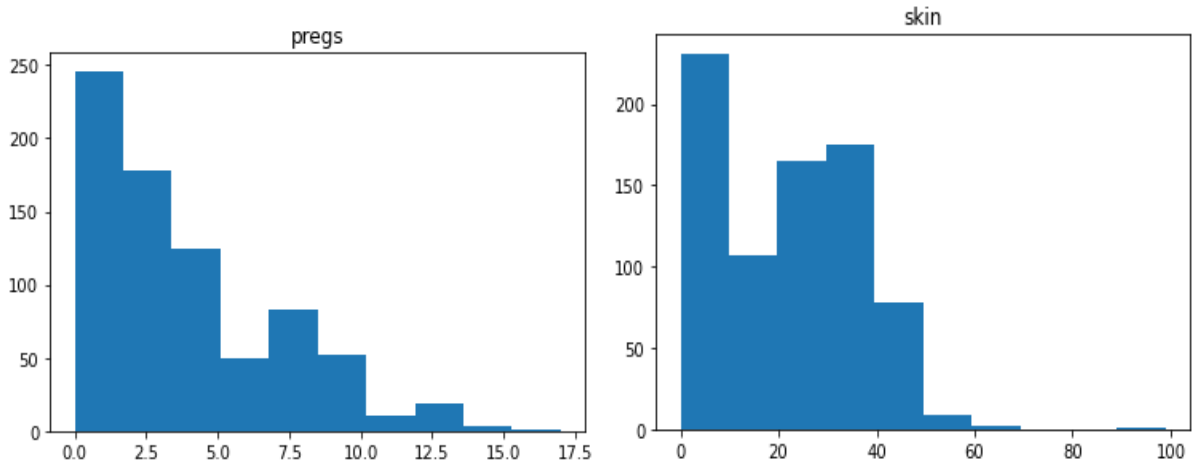



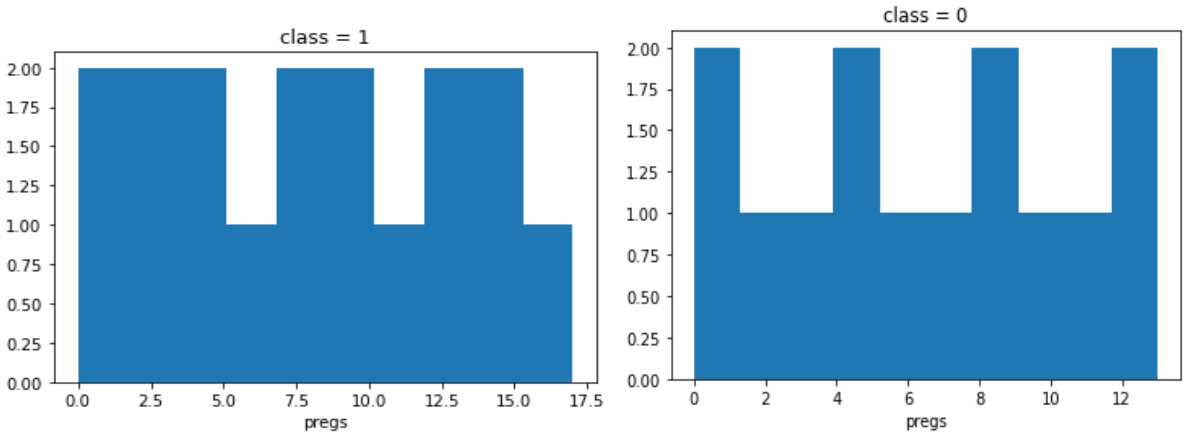

3. Find the value of correlation coefficient in the following cases: a. ‘Age’ with all other attributes (excluding ‘class’). b. ‘BMI’ with all other attributes (excluding ‘class’).

```
corrcoef of Age with pregs = 0.54434122840233992
corrcoef of BMI with pregs = 0.017683090727830580
corrcoef of Age with plas = 0.26351431982433350
corrcoef of BMI with plas = 0.22107106945898280
corrcoef of Age with pres = 0.23952794642136327
corrcoef of BMI with pres = 0.28180528884991063
corrcoef of Age with skin = -0.11397026236774170
corrcoef of BMI with skin = 0.39257320415903880
corrcoef of Age with test = -0.042162954735376820
corrcoef of BMI with test = 0.19785905649310143
corrcoef of Age with BMI = 0.036241870092294105
corrcoef of Age with pedi = 0.033561312434805450
corrcoef of BMI with pedi = 0.14064695254510504
corrcoef of BMI with Age = 0.036241870092294105
```

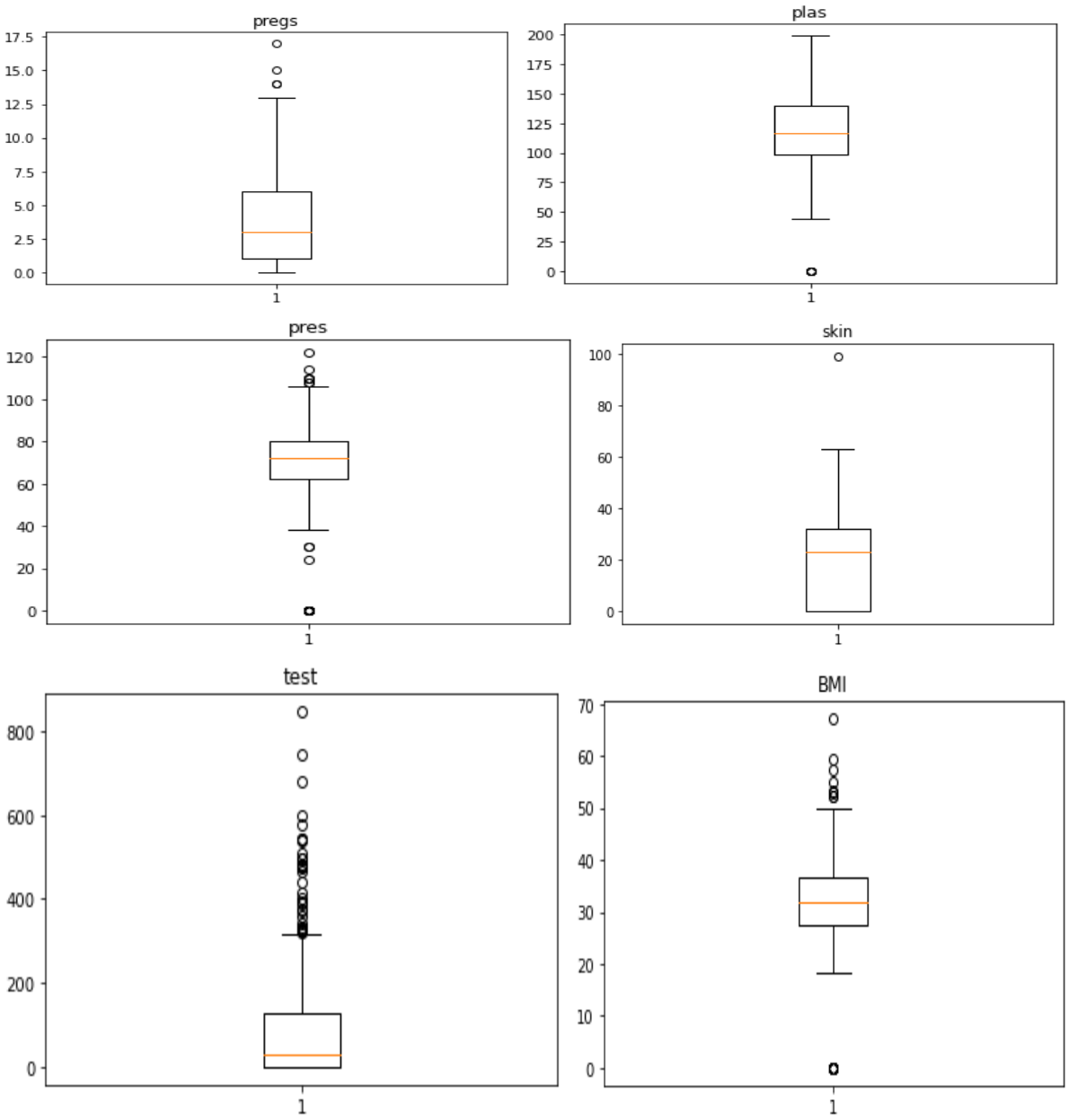
4. Plot the histogram for the attributes ‘preg’ and ‘skin’ (You may use “hist” function from pandas)

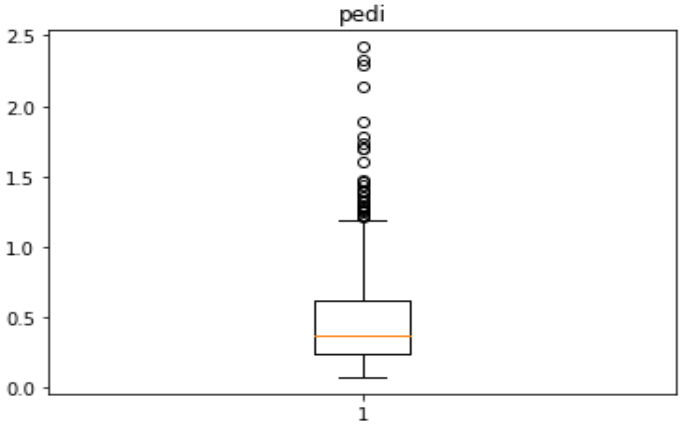


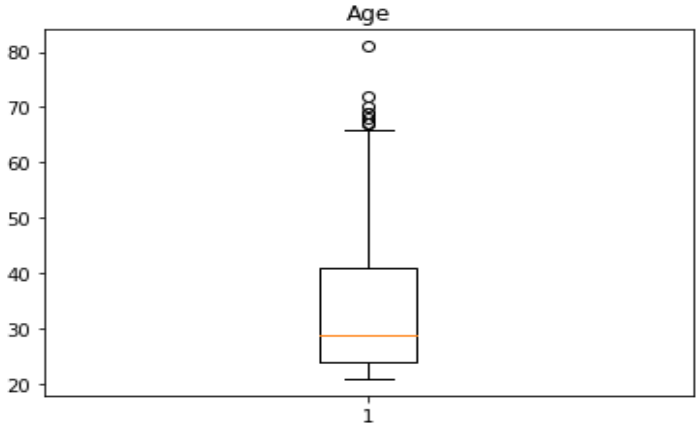

5. Plot the histogram of attribute ‘preg’ for each of the 2 classes individually (Use “groupby” function to group the tuples according to their “class”)




6. Obtain the boxplot for all the attribute excluding ‘class’ (Use “boxplot” function).




pedi
2.5
2.0
1.5
1.0
0.5
0.0
1


Age
80
70
60
50
40
30
20
1