IARE
INSTITUTE OF AERONAUTICAL ENGINEERING
(An Autonomous Institute affiliated to JNTUH, Hyderabad)
Dundigal, Hyderabad - 500 043

LABORATORY WORK BOOK

Name of the Student : HIMAKAR C

Class : CSE-B Semester : VI

Course Code : ACIC08 Course Name : DMKD Laboratory

Roll Number									
2	1	9	5	1	A	0	5	6	5

Name of the Course Faculty : Dr. D DURGA BHAVANI Faculty ID : IARE10921

Exercise Number : Week Number : 04 Date : 23/4/24

S. No.	Exercise Number	EXERCISE NAME	MARKS AWARDED					
			Aim/ Preparation	Algorithm / Procedure / Performance in the Lab	Source Code Calculations and Graphs	Program Execution Results and Error Analysis	Viva - Voce	Total
			4	4	4	4	4	20
1	4.1	Load data, describe data	4	4	4	4	4	20
2		and identify missing, outlier						
3		data items						
4	4.2	Find correlation among all attributes						
5	4.3	visualize correlation matrix						
6								
7								
8								
9								
10								
11								
12								

Signature of the Student

Signature of the Fac

START WRITING FROM HERE

Load data, describe given data and identify missing, outlier data items.

The program will load Pima Indians Diabetes Dataset, display first few rows and summary statistics, identify missing data and visualize outliers using boxplots for each column.

```
import pandas as pd
import matplotlib.pyplot as plt
d = pd.read_csv('pima-diabetes.csv')
print('First few rows of dataset is:')
print(d.head())
print('summary statistics of dataset is:')
print(d.describe())
print('Missing data is:')
print(d.isnull().sum())
plt.figure(figsize = (10, 6))
bp = d.boxplot(column = ['Pregnancies', 'Glucose',
    'BloodPressure', 'SkinThickness', 'Insulin', 'BMI',
    'DiabetesPedigreeFunction', 'Age', 'Outcome'])
plt.xticks(rotation = 45)
plt.title('BoxPlot showing outliers:')
plt.show()
```

ROLL NUMBER :

INPUT/OUTPUT:

First few rows of dataset is:

Pregnancies	Glucose	BloodPressure	...	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	...	0.627	50
1	1	85	66	...	0.351	31
0	2	8	183	64	...	0.672
32	1	3	1	89	66	...
0.167	21	0	4	0	137	40
...	2.288	33	1			

[5 rows x 9 columns]

Summary statistics of dataset is:

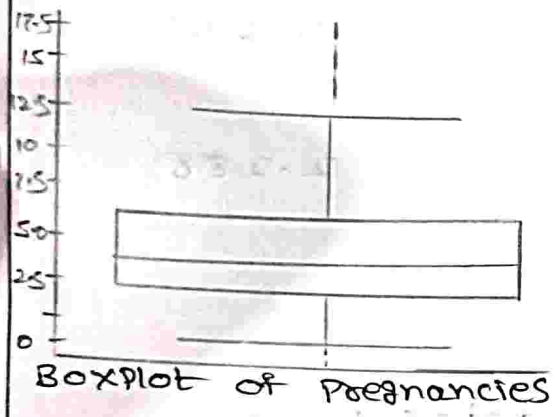
	Pregnancies	Glucose	...	Age	Outcome
count	768.00000	768.00000	...	768.00000	768.0000
mean	3.845052	120.894531	...	33.240885	0.348958
std	3.369578	31.972618	...	11.760232	0.476951
min	0.0000	0.0000	...	20.0000	0.0000
25%	1.00000	99.00000	...	24.00000	0.0000
50%	3.00000	117.00000	...	29.0000	0.0000
75%	6.00000	140.25000	...	41.00000	1.0000
max	17.00000	199.00000	...	81.00000	1.0000

ROLL NUMBER :

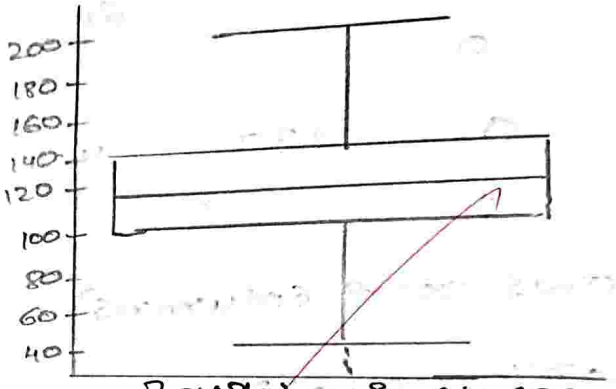
Missing data is:

Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigree Function	0
Age	0
Outcome	0

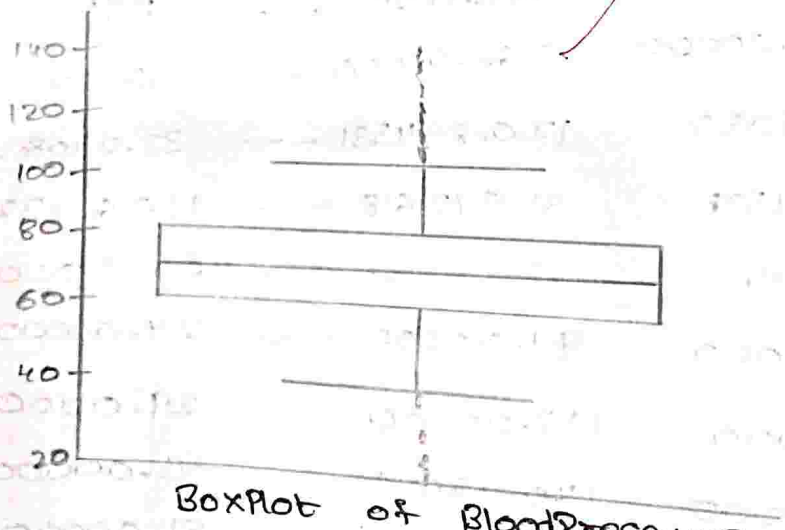
dtype: int64



Boxplot of Pregnancies

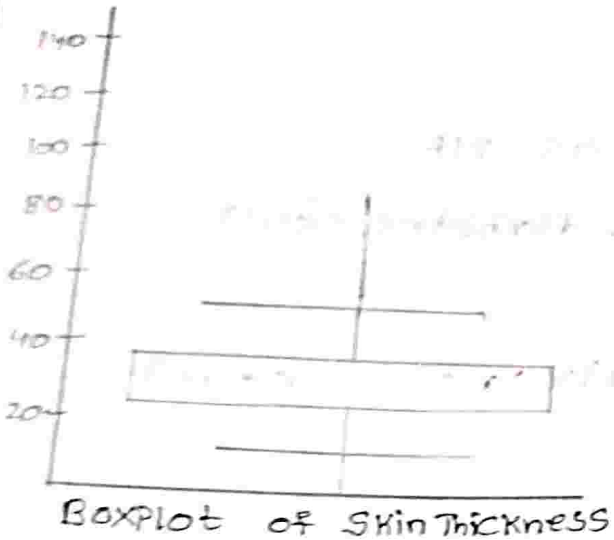


Boxplot of Glucose

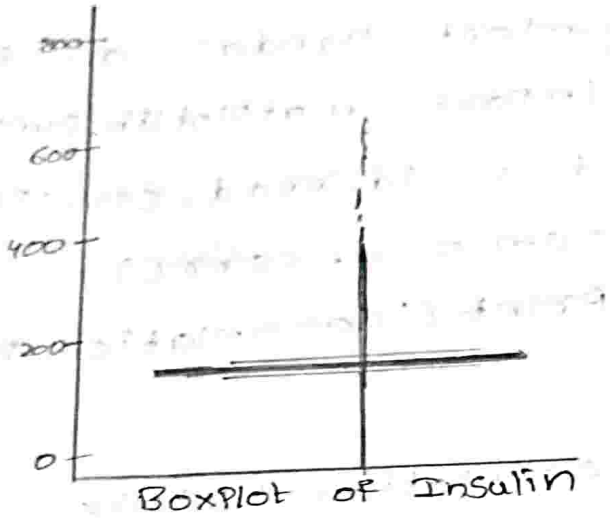


Boxplot of BloodPressure

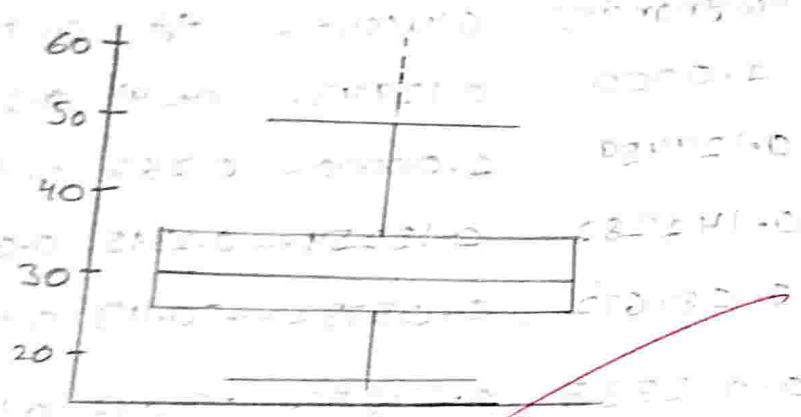
ROLL NUMBER :



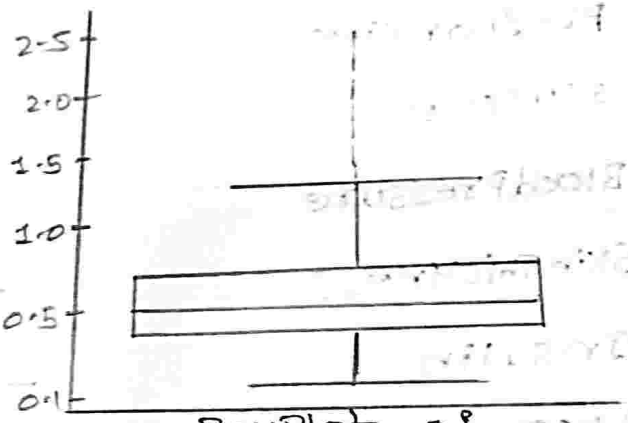
BoxPlot of SkinThickness



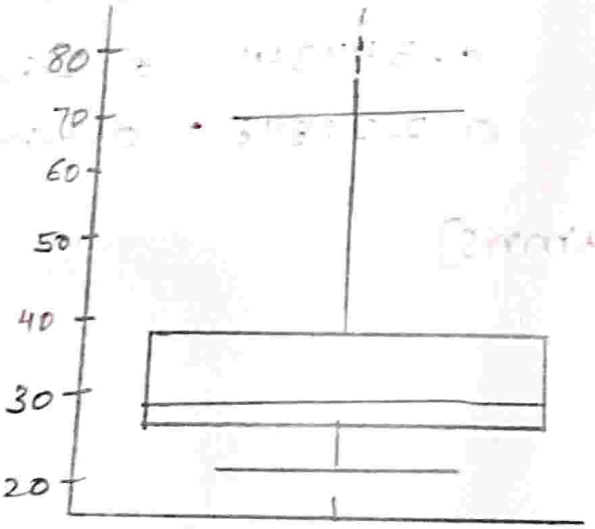
BoxPlot of Insulin



BoxPlot of BMI



BoxPlot of
Diabetes Pedigreefunctio



BoxPlot of Age

ROLL NUMBER :

4.2 Find correlation among all attributes

```
import pandas as pd
import matplotlib.pyplot as plt
d = pd.read_csv('Pima-diabetes.csv')
com = d.corr()
Print ('Correlation Matrix \n: ', com)
```

INPUT/OUTPUT:

correlation Matrix

	Pregnancies	Glucose	---	Age	Outcome
Pregnancies	1.0000	0.12945	---	0.544	0.221
Glucose	0.129459	1.00000	---	0.263	0.4665
BloodPressure	0.141282	0.15259	---	0.2395	0.065
SkinThickness	-0.081672	0.05732	---	-0.113	0.0747
Insulin	-0.073535	0.331357	---	-0.042	0.1305
BMI	0.017683	0.221071	---	0.0362	0.2926
DiabetesPedigreeFunction	-0.033523	0.137337	---	0.1000	0.238
Age	0.544341	0.263514	---	1.0000	0.1738
Outcome	0.221898	0.466581	---	0.2383	1.000

[9 rows x 9 columns]

ROLL NUMBER :

Visualize correlation Matrix

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
d = pd.read_csv('Pima-diabetes.csv')
cm = d.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(cm, annot=True, cmap='hsv', fmt=".2f")
plt.title('Correlation Matrix of PID Dataset is:')
plt.xticks(rotation = 45)
plt.yticks(rotation = 0)
plt.show()
```

ROLL NUMBER :

Correlation Matrix of Pima Indians Diabetes Dataset

	Pregnancies	Glucose	BP	Skin	Insulin	BMI	DPF	Age	Outcome
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BP	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
Skin	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.03
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
DPF	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00

Pregnancies Glucose BP Skin Thickness Insulin BMI DPF Age Outcome