Foundations of Data Science Project - Diabetes Analysis

Context

Diabetes is one of the most frequent diseases worldwide and the number of diabetic patients are growing over the years. The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes.

A few years ago research was done on a tribe in America which is called the Pima tribe (also known as the Pima Indians). In this tribe, it was found that the ladies are prone to diabetes very early. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients were females at least 21 years old of Pima Indian heritage.

Objective

Here, we are analyzing different aspects of Diabetes in the Pima Indians tribe by doing Exploratory Data Analysis.

Data Dictionary

The dataset has the following information:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: A function that scores the likelihood of diabetes based on family history.
- Age: Age in years
- Outcome: Class variable (0: a person is not diabetic or 1: a person is diabetic)

Q 1: Import the necessary libraries and briefly explain the use of each library (3 Marks)

In [3]:
```
import numpy as np  #library used for working with arrays
import pandas as pd #library used for data manipulation and analysis

import seaborn as sns #library used for visualization
import matplotlib.pyplot as plt #library used for visualization
%matplotlib inline
```

Write your Answer here:

Ans 1: write after following coding

Q 2: Read the given dataset (1 Mark)

In [6]:
```
pima = pd.read_csv("diabetes.csv")
```

Q3. Show the last 10 records of the dataset. How many columns are there? (1 Mark)

In [8]:
```
pima.tail(10) # to know the label of columns in the data: pima.colums
```

Out[8]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
758	1	106	76	20	79	37.5	0.197	26	0
759	6	190	92	20	79	35.5	0.278	66	1
760	2	88	58	26	16	28.4	0.766	22	0
761	9	170	74	31	79	44.0	0.403	43	1
762	9	89	62	20	79	22.5	0.142	33	0
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	79	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	20	79	30.1	0.349	47	1
767	1	93	70	31	79	30.4	0.315	23	0

Write your Answer here:

Ans 3: 9

Q4. Show the first 10 records of the dataset (1 Mark)

In [11]:
```
pima.head(10)
```

Out[11]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	79	33.600000	0.627	50	1
1	1	85	66	29	79	26.600000	0.351	31	0
2	8	183	64	20	79	23.300000	0.672	32	1
3	1	89	66	23	94	28.100000	0.167	21	0
4	0	137	40	35	168	43.100000	2.288	33	1
5	5	116	74	20	79	25.600000	0.201	30	0
6	3	78	50	32	88	31.000000	0.248	26	1
7	10	115	69	20	79	35.300000	0.134	29	0
8	2	197	70	45	543	30.500000	0.158	53	1
9	8	125	96	20	79	31.992578	0.232	54	1

Q5. What do you understand by the dimension of the dataset? Find the dimension of the `pima` dataframe. (1 Mark)

In [13]:
```
pima.ndim #return dimension of dataframe/ series
```

Out[13]: 2

Write your Answer here:

Ans 5: 2

Q6. What do you understand by the size of the dataset? Find the size of the `pima` dataframe. (1 Mark)

In [16]:
```
pima.size #to get the total number of elements in DataFrame/ Series
```

Out[16]: 6912

Write your Answer here:

Ans 6: 6912

Q7. What are the data types of all the variables in the data set? (2 Marks)

Hint: Use the info() function to get all the information about the dataset.

In [19]:
```
pima.dtypes.value_counts()
```

Out[19]:
```
int64      7
float64    2
Name: count, dtype: int64
```

Write your Answer here:

Ans 7: 2 float data types and 67 integer data types

Q8. What do we mean by missing values? Are there any missing values in the `pima` dataframe? (2 Marks)

In [22]:
```
pima.isnull().values.any() # To check if data contains null values
```

Out[22]: False

Write your Answer here:

Ans 8: False

Q9. What do the summary statistics of the data represent? Find the summary statistics for all variables except 'Outcome' in the `pima` data. Take one column/variable from the output table and explain all its statistical measures. (3 Marks)

In [25]:
```
pima.iloc[:,0:8].describe()
```

Out[25]:

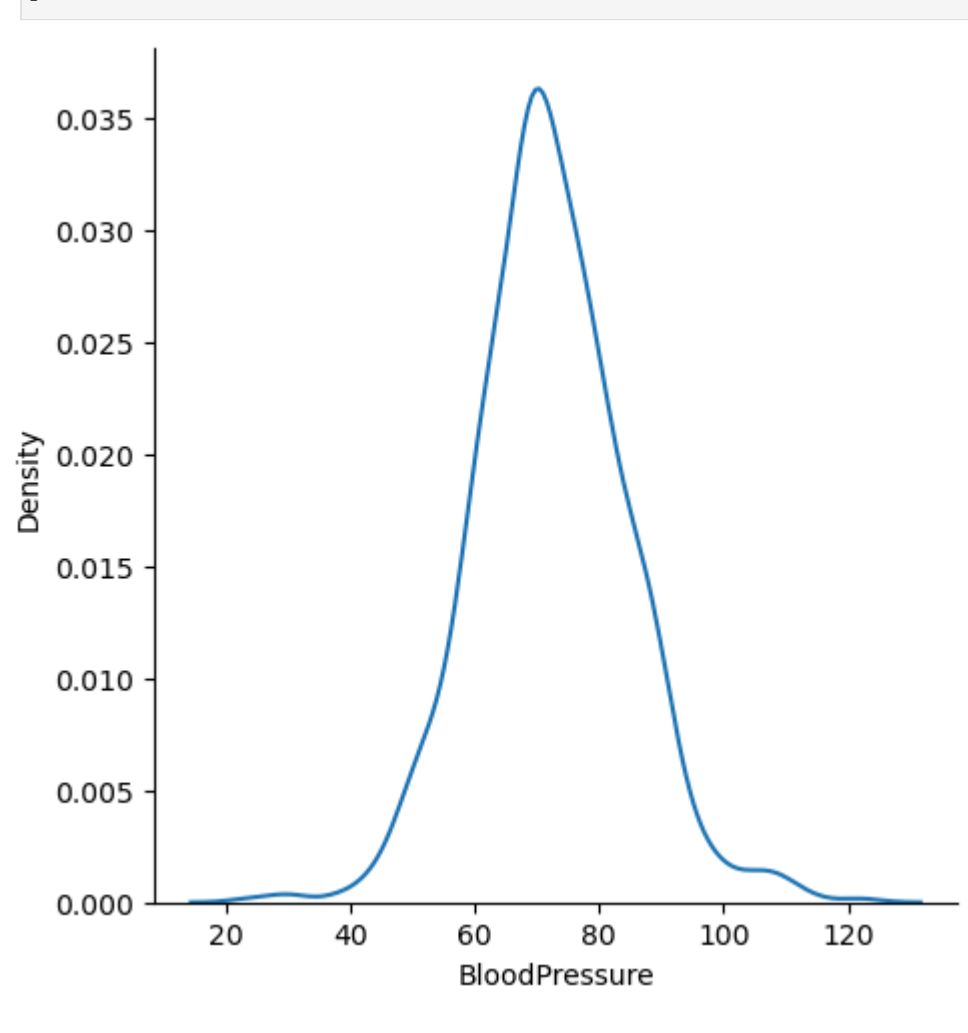
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.675781	72.250000	26.447917	118.270833	32.450805	0.471876	33.240885
std	3.369578	30.436252	12.117203	9.733872	93.243829	6.875374	0.331329	11.760232
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000
25%	1.000000	99.750000	64.000000	20.000000	79.000000	27.500000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	79.000000	32.000000	0.372500	29.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

Write your Answer here:

Ans 9: Highest glucose levels is 199, pregnancies 17 and BMI 67.

Q 10. Plot the distribution plot for the variable 'BloodPressure'. Write detailed observations from the plot. (2 Marks)

In [28]:
```
sns.displot(pima['BloodPressure'], kind='kde')
plt.show()
```



Write your Answer here:

Ans 10: The most people's blood pressure range from 60 to 80.

Q 11. What is the 'BMI' of the person having the highest 'Glucose'? (1 Mark)

In [31]:
```
pima[pima['Glucose']==pima['Glucose'].max()]['BMI']
```

Out[31]:
```
661    42.9
Name: BMI, dtype: float64
```

Write your Answer here:

Ans 11: the person with the highest glucose value(661) has a BMI of 42.9

Q12.

12.1 What is the mean of the variable 'BMI'?

12.2 What is the median of the variable 'BMI'?

12.3 What is the mode of the variable 'BMI'?

12.4 Are the three measures of central tendency equal?

(3 Marks)

In [34]:
```
m1 = pima['BMI'].mean()   # mean
print(m1)
m2 = pima['BMI'].median()   # median
print(m2)
m3 = pima['BMI'].mode()[0]   # mode
print(m3)
```

```
32.45080515543619
32.0
32.0
```

Write your Answer here:

Ans 12: 12: 32.45, 32, 32 seperately Mean, median and mode (central measures of tendency) are equal

Q13. How many women's 'Glucose' levels are above the mean level of 'Glucose'? (1 Mark)

In [37]:
```
pima[pima['Glucose']>pima['Glucose'].mean()].shape[0]
```

Out[37]: 343

Write your Answer here:

Ans 13: 343

Q14. How many women have their 'BloodPressure' equal to the median of 'BloodPressure' and their 'BMI' less than the median of 'BMI'? (2 Marks)

In [40]:
```
pima[(pima['BloodPressure']==pima['BloodPressure'].median()) & (pima['BMI']<pima['BMI'].median())]
```

Out[40]:

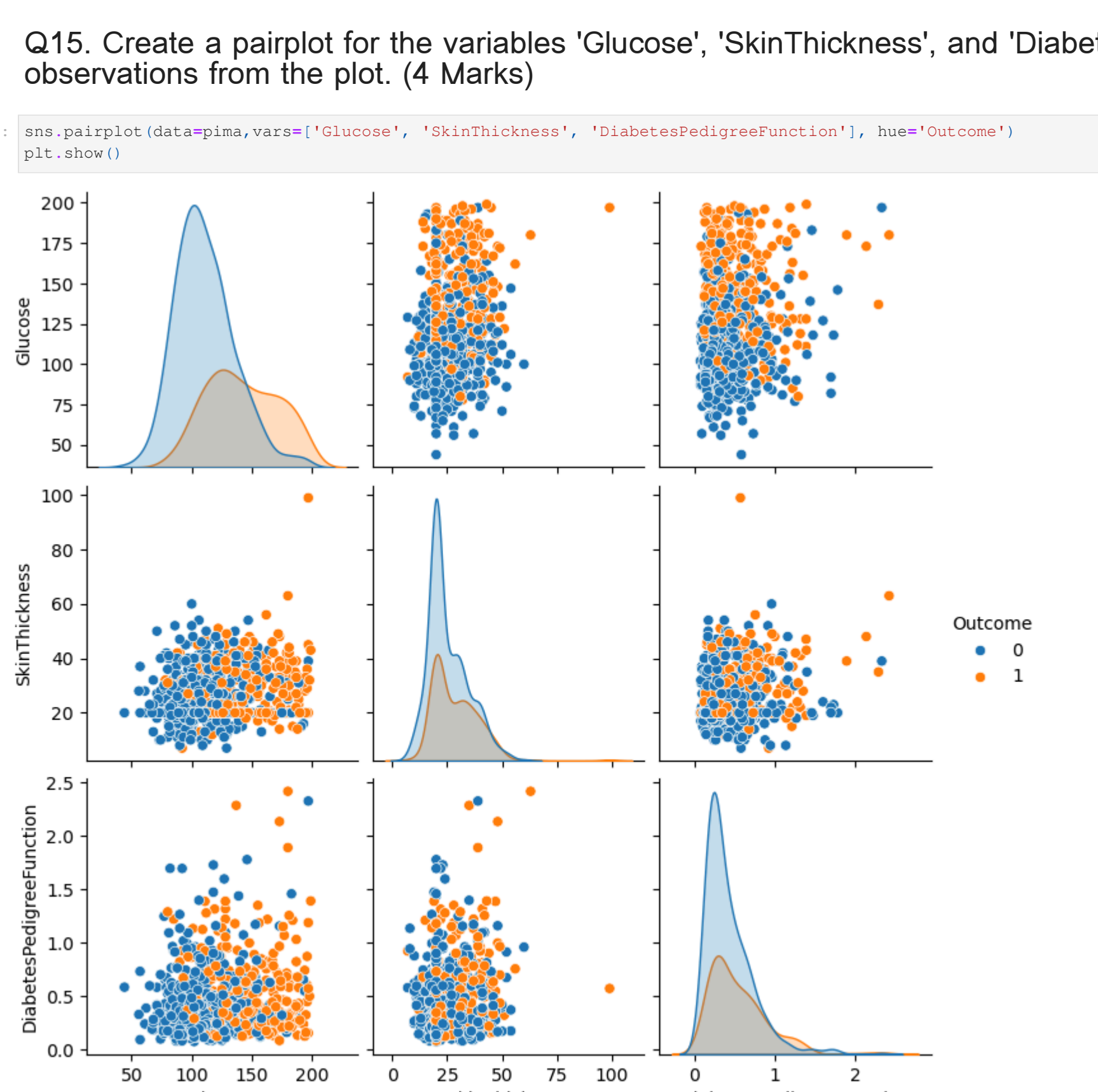
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
14	5	166	72	19	175	25.8	0.587	51	1
93	4	134	72	20	79	23.8	0.277	60	1
103	1	81	72	18	40	26.6	0.283	24	0
205	5	111	72	28	79	23.9	0.407	27	0
299	8	112	72	20	79	23.6	0.840	58	0
325	1	157	72	21	168	25.6	0.123	24	0
330	8	118	72	19	79	23.1	1.476	46	0
366	6	124	72	20	79	27.6	0.368	29	1
380	1	107	72	30	82	30.8	0.821	24	0
393	4	116	72	12	87	22.1	0.463	37	0
406	4	115	72	20	79	28.9	0.376	46	1
446	1	100	72	12	70	25.3	0.658	28	0
460	9	120	72	22	56	20.8	0.733	48	0
488	4	99	72	17	79	25.6	0.294	28	0
497	2	81	72	15	76	30.1	0.547	25	0
510	12	84	72	31	79	29.7	0.297	46	1
568	4	154	72	29	126	31.3	0.338	37	0
615	3	106	72	20	79	25.8	0.207	27	0
635	13	104	72	20	79	31.2	0.465	38	1
644	3	103	72	30	152	27.6	0.730	27	0
717	10	94	72	18	79	23.1	0.595	56	0
765	5	121	72	23	112	26.2	0.245	30	0

Write your Answer here:

Ans 14: 22

Q15. Create a pairplot for the variables 'Glucose', 'SkinThickness', and 'DiabetesPedigreeFunction'. Write your observations from the plot. (4 Marks)

In [43]:
```
sns.pairplot(data=pima,vars=['Glucose', 'SkinThickness', 'DiabetesPedigreeFunction'], hue='Outcome')
plt.show()
```

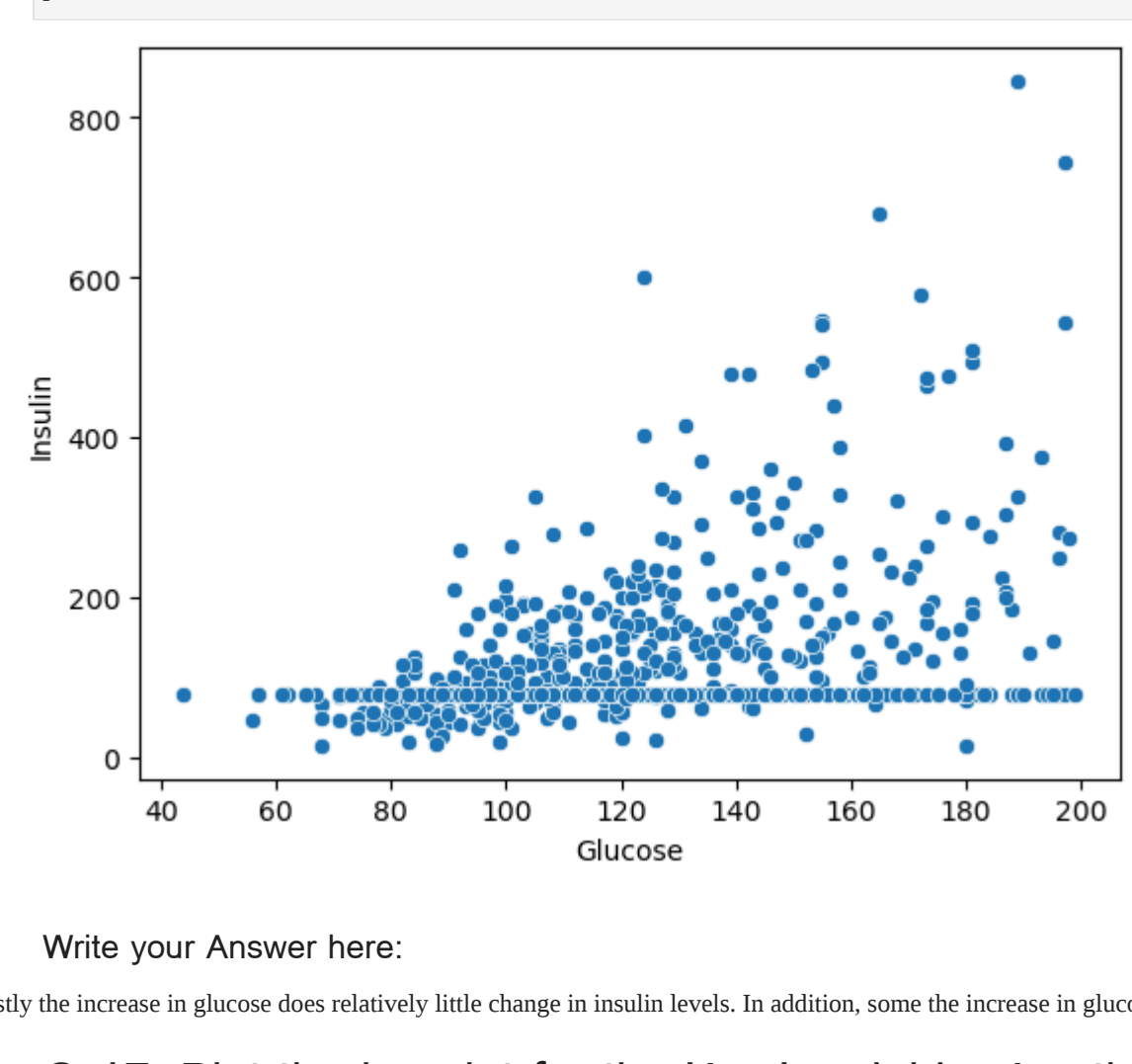


Write your Answer here:

Ans 15: On the diagonal axes, a plot shows the univariate distribution of each variable.

Q16. Plot the scatterplot between 'Glucose' and 'Insulin'. Write your observations from the plot. (2 Marks)

In [46]:
```
sns.scatterplot(x='Glucose',y='Insulin',data=pima)
plt.show()
```

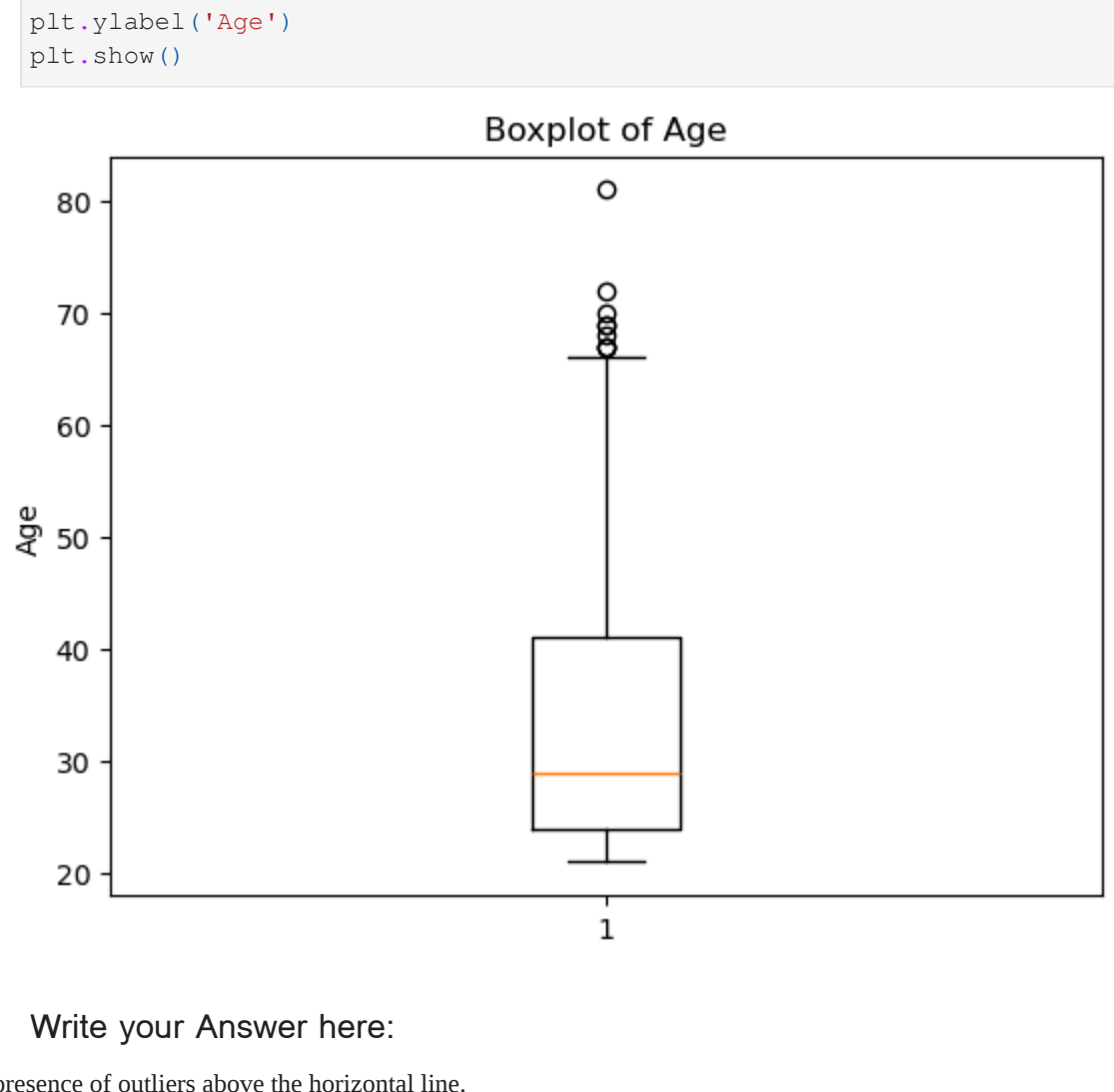


Write your Answer here:

Ans 16: Mostly the increase in glucose does relatively little change in insulin levels. In addition, some the increase in glucose increases in insulin.

Q 17. Plot the boxplot for the 'Age' variable. Are there outliers? (2 Marks)

In [49]:
```
plt.boxplot(pima['Age'])

plt.title('Boxplot of Age')
plt.ylabel('Age')
plt.show()
```

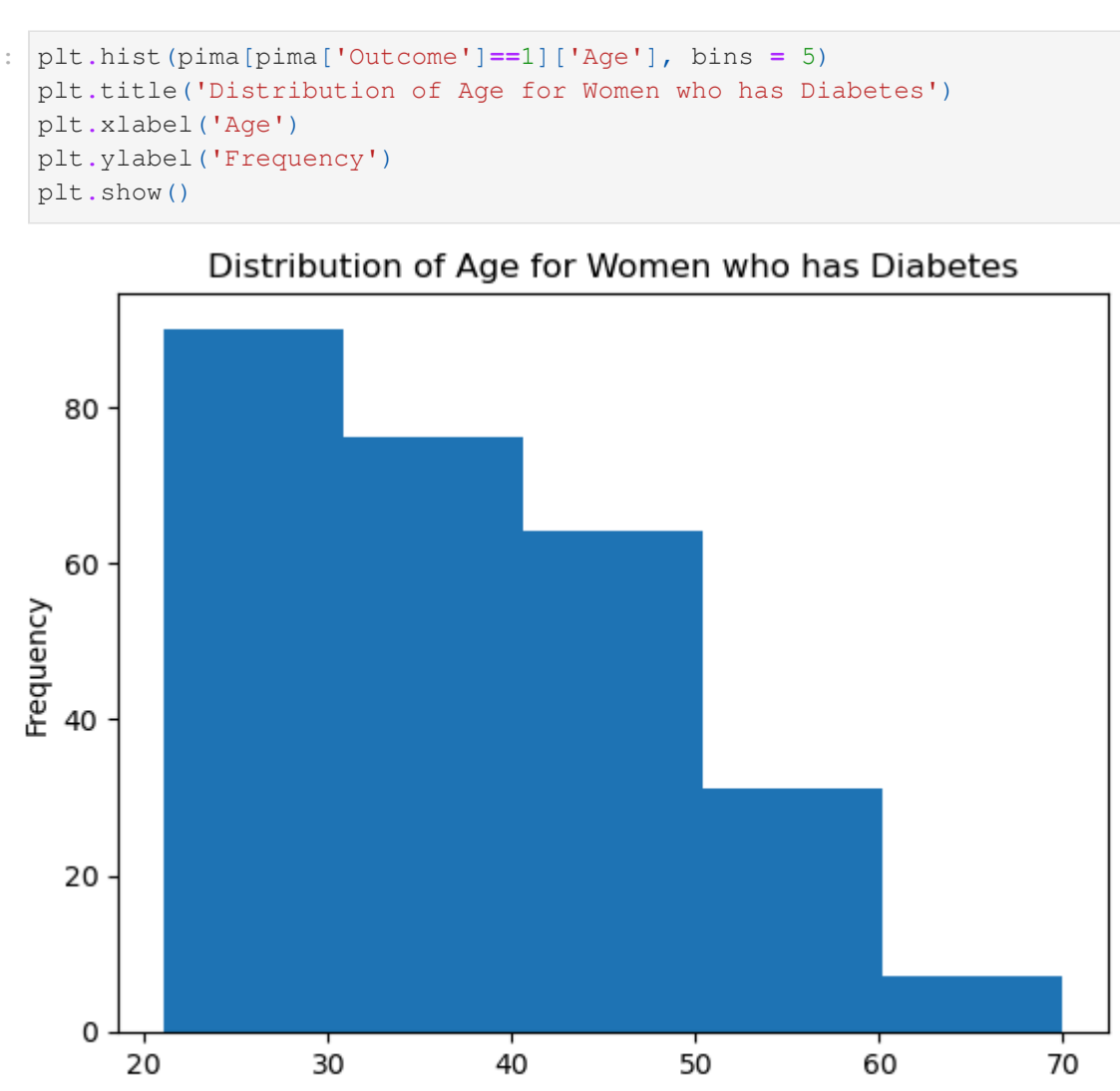


Write your Answer here:

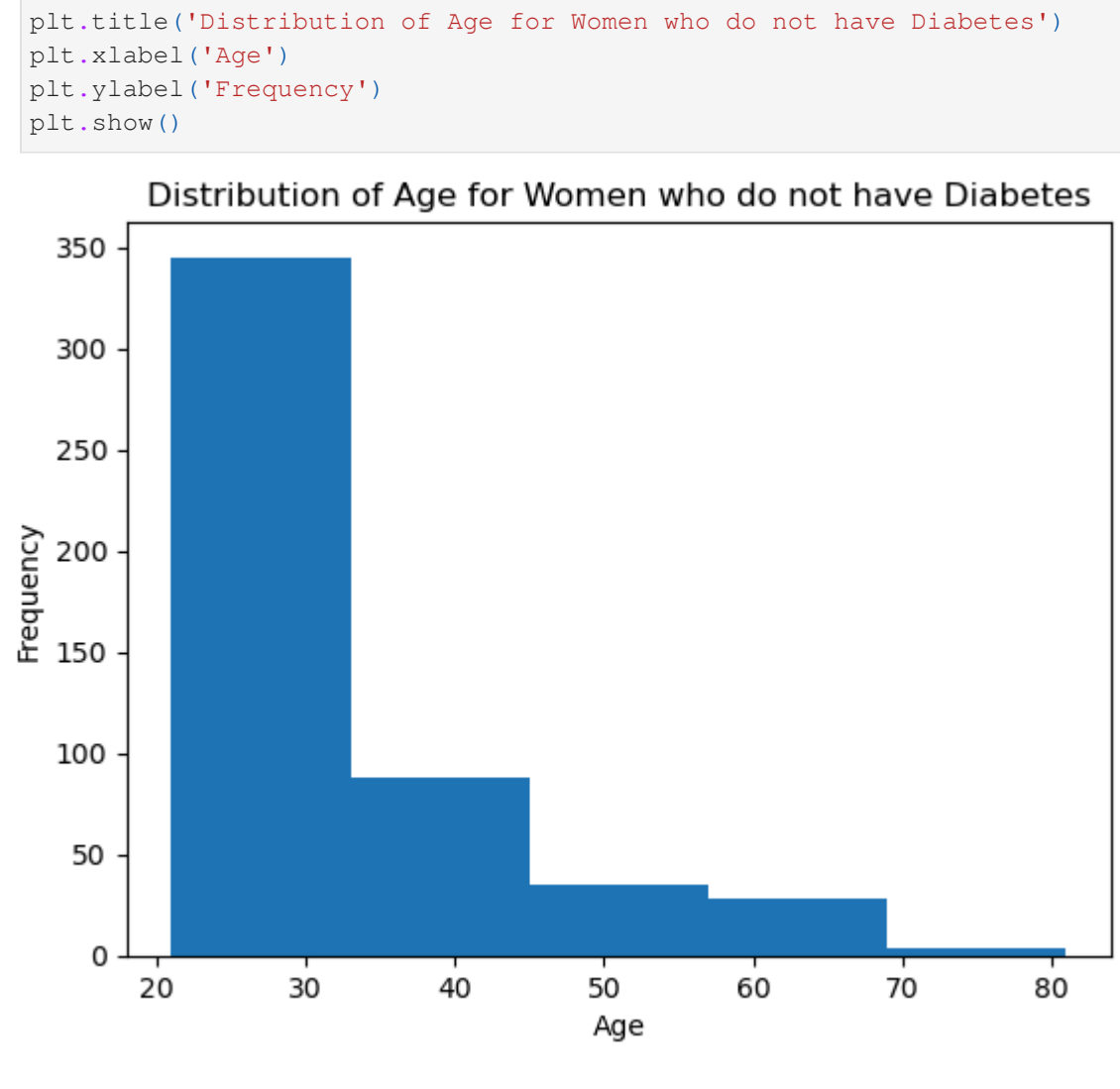
Ans 17: The presence of outliers above the horizontal line.

Q18. Plot histograms for the 'Age' variable to understand the number of women in different age groups given whether they have diabetes or not. Explain both histograms and compare them. (3 Marks)

In [52]:
```
plt.hist(pima[pima['Outcome']==1]['Age'], bins = 5)
plt.title('Distribution of Age for Women who has Diabetes')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



In [53]:
```
plt.hist(pima[pima['Outcome']==0]['Age'], bins = 5)
plt.title('Distribution of Age for Women who do not have Diabetes')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



Write your Answer here:

Ans 18: The woman with diabetes most are from the age between 22 to 30. The younger woman, the easier to get diabetes. The highest number of Women without diabetes range between ages 22 to 33. Woman gets diabetes easier between the age of 22 to 35 (highest risk) and also the highest nember of those without diabetes.

Q 19. What is the Interquartile Range of all the variables? Why is this used? Which plot visualizes the same? (2 Marks)

In [56]:
```
Q1 = pima.quantile(0.25)
Q3 = pima.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
Pregnancies                  5.0000
Glucose                     40.5000
BloodPressure               16.0000
SkinThickness               12.0000
Insulin                     48.2500
BMI                          9.1000
DiabetesPedigreeFunction     0.3825
Age                         17.0000
Outcome                      1.0000
dtype: float64
```

Write your Answer here:

Ans 19: 1. (a)tells us inside what range of the bulk of our data lies (b)it can be calculated by taking the difference between the third quartile and the first quartile within a dataset. 2. It is a methodology that is generally used to filter outliers in a dataset. 3. barplot

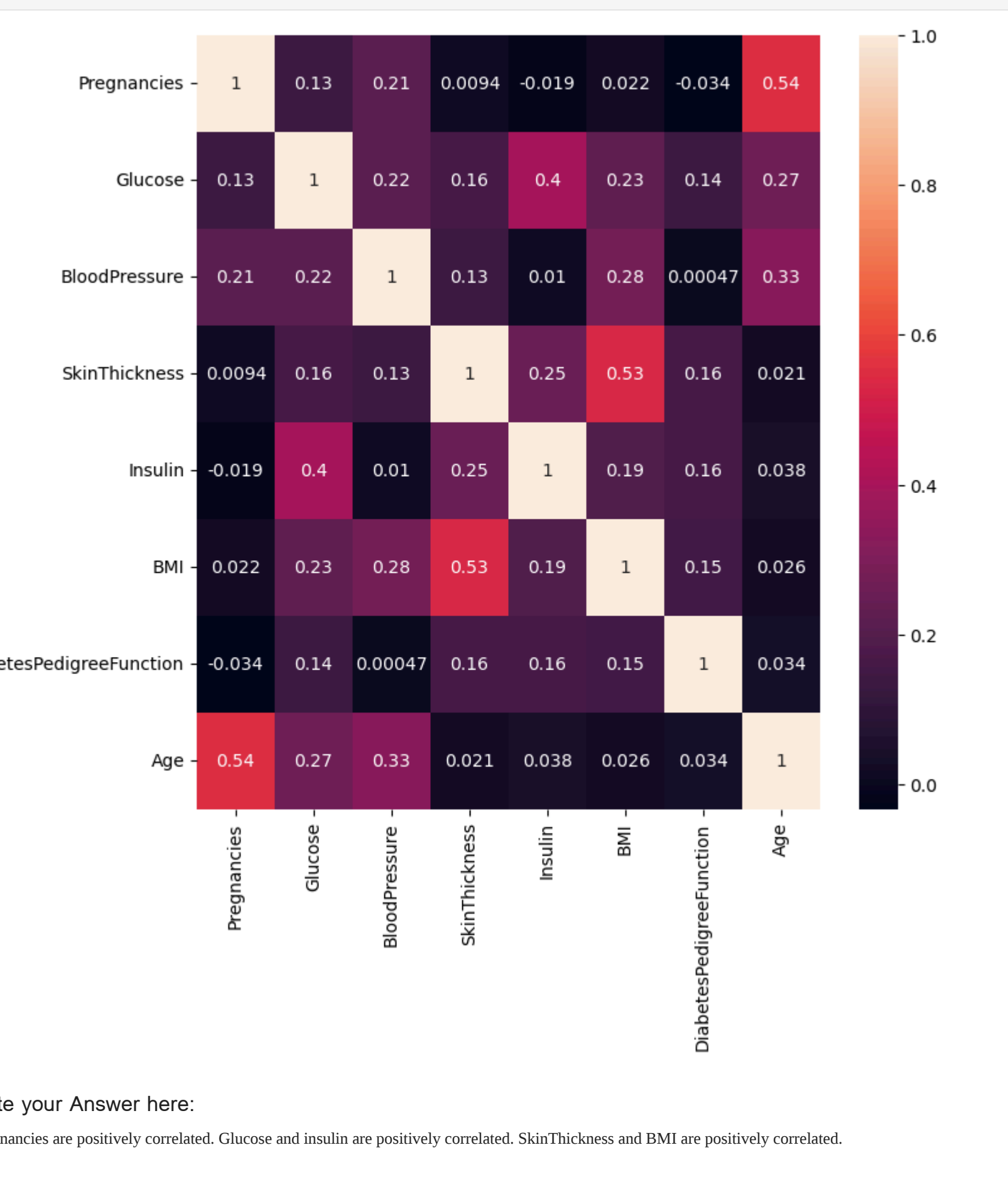
Q 20. Find and visualize the correlation matrix. Write your observations from the plot. (3 Marks)

In [59]:
```
corr_matrix = pima.iloc[:,0:8].corr()

corr_matrix
```

Out[59]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Pregnancies	1.000000	0.128022	0.208987	0.009393	-0.018780	0.021546	-0.033523	0.544341
Glucose	0.128022	1.000000	0.219765	0.158060	0.396137	0.231464	0.137158	0.266673
BloodPressure	0.208987	0.219765	1.000000	0.130403	0.010492	0.281222	0.000471	0.326791
SkinThickness	0.009393	0.158060	0.130403	1.000000	0.245410	0.532552	0.157196	0.020582
Insulin	-0.018780	0.396137	0.010492	0.245410	1.000000	0.189919	0.158243	0.037676
BMI	0.021546	0.231464	0.281222	0.532552	0.189919	1.000000	0.153508	0.025748
DiabetesPedigreeFunction	-0.033523	0.137158	0.000471	0.157196	0.158243	0.153508	1.000000	0.033561
Age	0.544341	0.266673	0.326791	0.020582	0.037676	0.025748	0.033561	1.000000

In [60]:
```
plt.figure(figsize=(8,8))
sns.heatmap(corr_matrix, annot = True)

# display the plot
plt.show()
```



Write your Answer here:

Ans 20: Age and pregnancies are positively correlated. Glucose and insulin are positively correlated. SkinThickness and BMI are positively correlated.