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. 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 tribe.

Objective

Here, we are analyzing different aspects of Diabetes in the Pima Diabetes Analysis 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 [1]:

```
# remove _____ & write the appropriate library name

import numpy as np
import pandas as pd

import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Write your Answer here:

Ans 1: Numpy is a library for numerical calculations and analysis of arrays Pandas is a library for data manipulation and data processing/analysis Matplotlib.pyplot is a library for data visualization (python plottting) Seaborn is a library for advanced graphical analysis/visualization

Q 2: Read the given dataset (1 Mark)

In [2]:

```
# remove _____ & write the appropriate function name

pima = pd.read_csv("diabetes.csv")
```

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

In [3]:

```
# remove ______ and write the appropriate number in the function

pima.tail(10)
```

Out[3]:

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

Write your Answer here:

Ans 3: There are 9 columns in the dataframe.

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

In [4]:

```
# remove _____ & write the appropriate function name and the number of rows to get in th

pima.head(10)
```

Out[4]:

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

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

In [5]:

```
# remove _____ & write the appropriate function name

pima.shape
```

Out[5]:

```
(768, 9)
```

Write your Answer here:

Ans 5: The dimention of a DataFrame is a tuple of array dimensions that tells the number of rows and columns of a given DataFrame. Our dataset has 768 rows and 9 columns.

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

In [6]:

```
# remove _____ & write the appropriate function name

pima.size
```

Out[6]:

```
6912
```

Write your Answer here:

Ans 6: The size of the dataframe tells us the total numbers of elements (rows x columns), in this case 6912 elements.

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 [7]:

```
# remove _____ & write the appropriate function name

pima.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Write your Answer here:

Ans 7: All the datatype columns are numerical (2 are float, 7 are integers). There are no null values in all the 768 rows. We don't have any object data types.

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

In [8]:

```
# remove _____ & write the appropriate function name

pima.isnull().values.any()
```

Out[8]:

```
False
```

Write your Answer here:

Ans 8: Missing values are all the cells wich are empty/blank, usually marked as NaN (Not a Number). From the above code output (False), we know that there are no missing values in our dataframe.

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 [9]:

```
# remove _____ & write the appropriate function name

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

Out[9]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeF
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	76
mean	3.845052	121.675781	72.250000	26.447917	118.270833	32.450805	(
std	3.369578	30.436252	12.117203	9.733872	93.243829	6.875374	(
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	(
25%	1.000000	99.750000	64.000000	20.000000	79.000000	27.500000	(
50%	3.000000	117.000000	72.000000	23.000000	79.000000	32.000000	(
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	(
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2

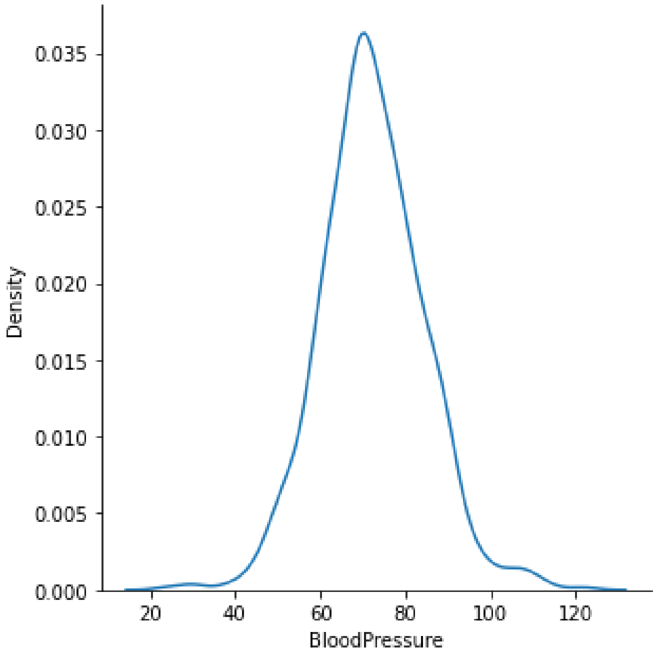
Write your Answer here:

Ans 9: The summary statistics represent the statistical parameters for each column (count, mean, std, min, max, percentiles). For the BloodPressure measurements, the minimum is 24.0, the maximum is 122.0, wich gives a mean of 72.25, with a std of 12.12. Half of the sample has over 72.0 (median). Looking at the mean and median values, and the dispersion (25 quartile and 75 quartile), the distribution appears to approach the standard normal distribution.

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

In [10]:

```
# remove _____ & write the appropriate library name

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



Write your Answer here:

Ans 10: The distribution plot shows the majority of the observations lie between 40 and 100, there is one one peak (unimodal distribution), it looks simmetrical, and appoaches the shape of a normal distribution.

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

In [11]:

```
# remove _____ & write the appropriate function name

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

Out[11]:

```
661    42.9
Name: BMI, dtype: float64
```

Write your Answer here:

Ans 11: The person with the highest glucose value 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 [12]:

```
# remove _____ & write the appropriate function name

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

```
32.45080515543617
32.0
32.0
```

Write your Answer here:

Ans 12: With a mean of 32.45, a median of 32.0, and a mode of 32.0, we have the suggestion that the BMI distribution is quite normal, with no skewness. (In a Normal Distribution the mean=median=mode).

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

In [13]:

```
# remove _____ & write the appropriate function name

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

Out[13]:

```
343
```

Write your Answer here:

Ans 13: There are 343 women with Glucose levels above the mean level.

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 [14]:

```
# remove _____ & write the appropriate column name

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

Out[14]:

```
22
```

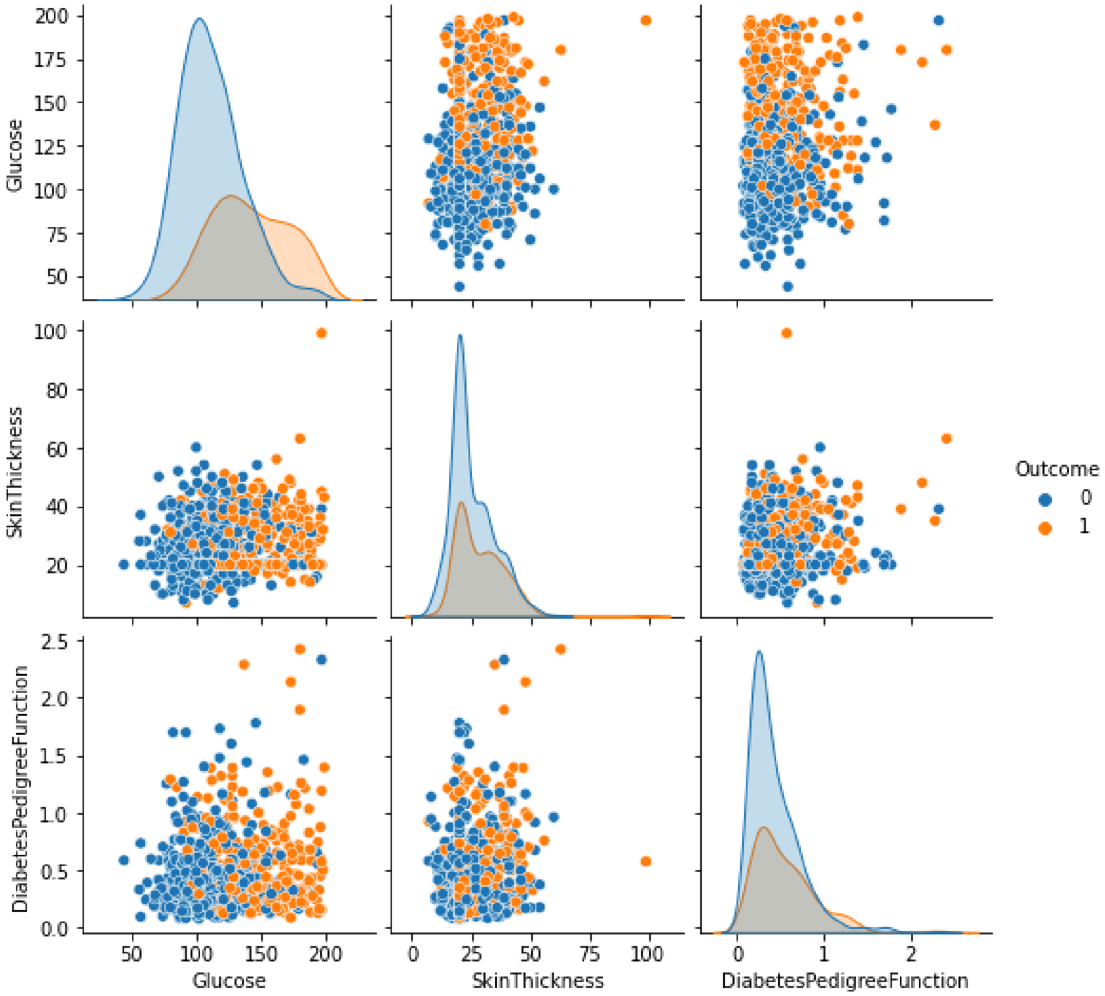
Write your Answer here:

Ans 14: There are 22 women that meet these 2 criteria.

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

In [15]:

```
# remove _____ & write the appropriate function name

sns.pairplot(data=pima,vars=['Glucose', 'SkinThickness', 'DiabetesPedigreeFunction'], hu
plt.show()
```



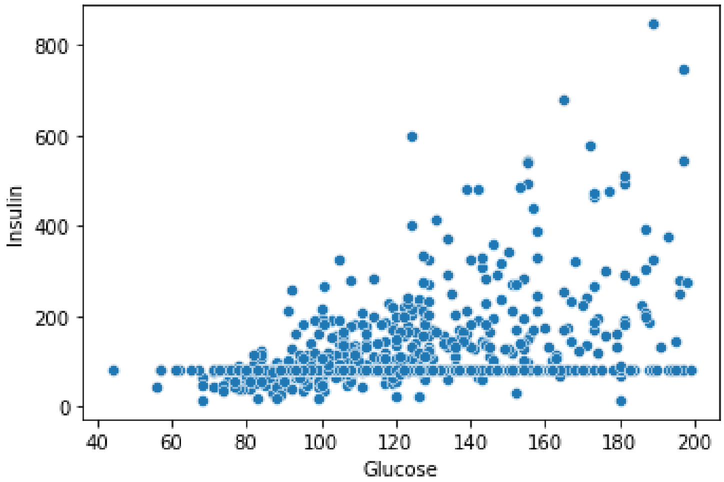
Write your Answer here:

Ans 15: From the above pairplot, we can see that DiabetesPedigreeFunction has a positive impact on Glucose variable. For the other pairs, they seem to be uncorrelated. In general, the women with Outcome=1 (diabetics), have higher values of Glucose. Also, all three distributions as right skewed, but diabetic women (Outcome=1) show lower frequencies.

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

In [16]:

```
# remove _____ & write the appropriate function name

sns.scatterplot(x='Glucose',y='Insulin',data=pima)
plt.show()
```



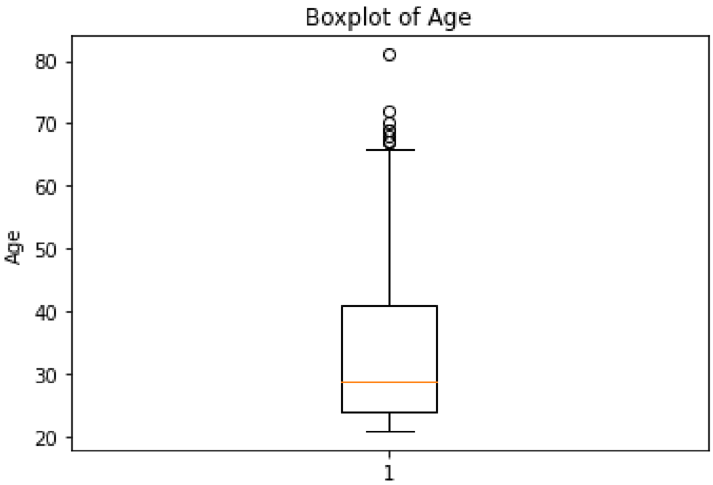
Write your Answer here:

Ans 16:There is a positive correlation between Glucose and Insulin. As Glucose increases, we have higher Insulin values.

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

In [17]:

```
# remove _____ & write the appropriate function and column name

plt.boxplot(pima['Age'])

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



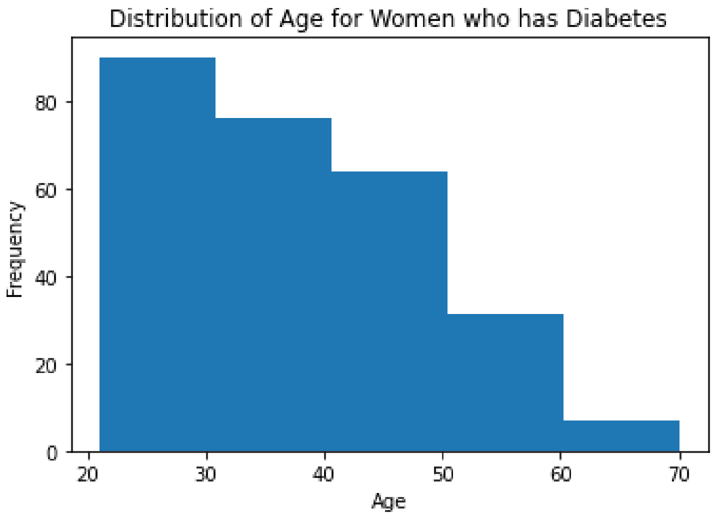
Write your Answer here:

Ans 17: Yes there are various outliers in the upper range of the distributions. These are extreme values that are more than 1.5*IQR away from Q3 (top of the box). The distribution is very posivively skewed.

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)

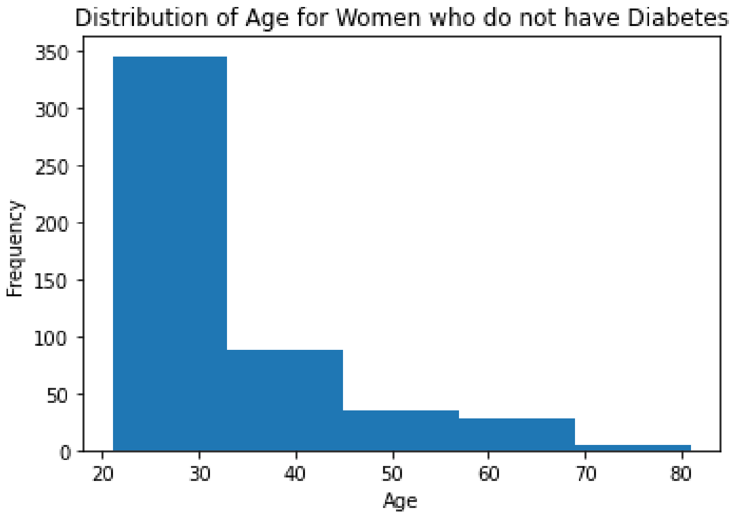
In [18]:

```
# remove _____ & write the appropriate function and column name

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 [19]:

```
# remove _____ & write the appropriate function and column name

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: Women without Diabetes are younger (mostly in their 20's), and women with Diabetes are spread over a wider age range (20 to 50). Also, in both distributions, the frequency decreases with advancing age.

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

In [20]:

```
# remove _____ & write the appropriate variable name

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: The IQR is used get a sense on the variability of the majority of the data (50%). We can visualize this with a Boxplot.

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

In [21]:

```
# remove _____ & write the appropriate function name and run the code

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

corr_matrix
```

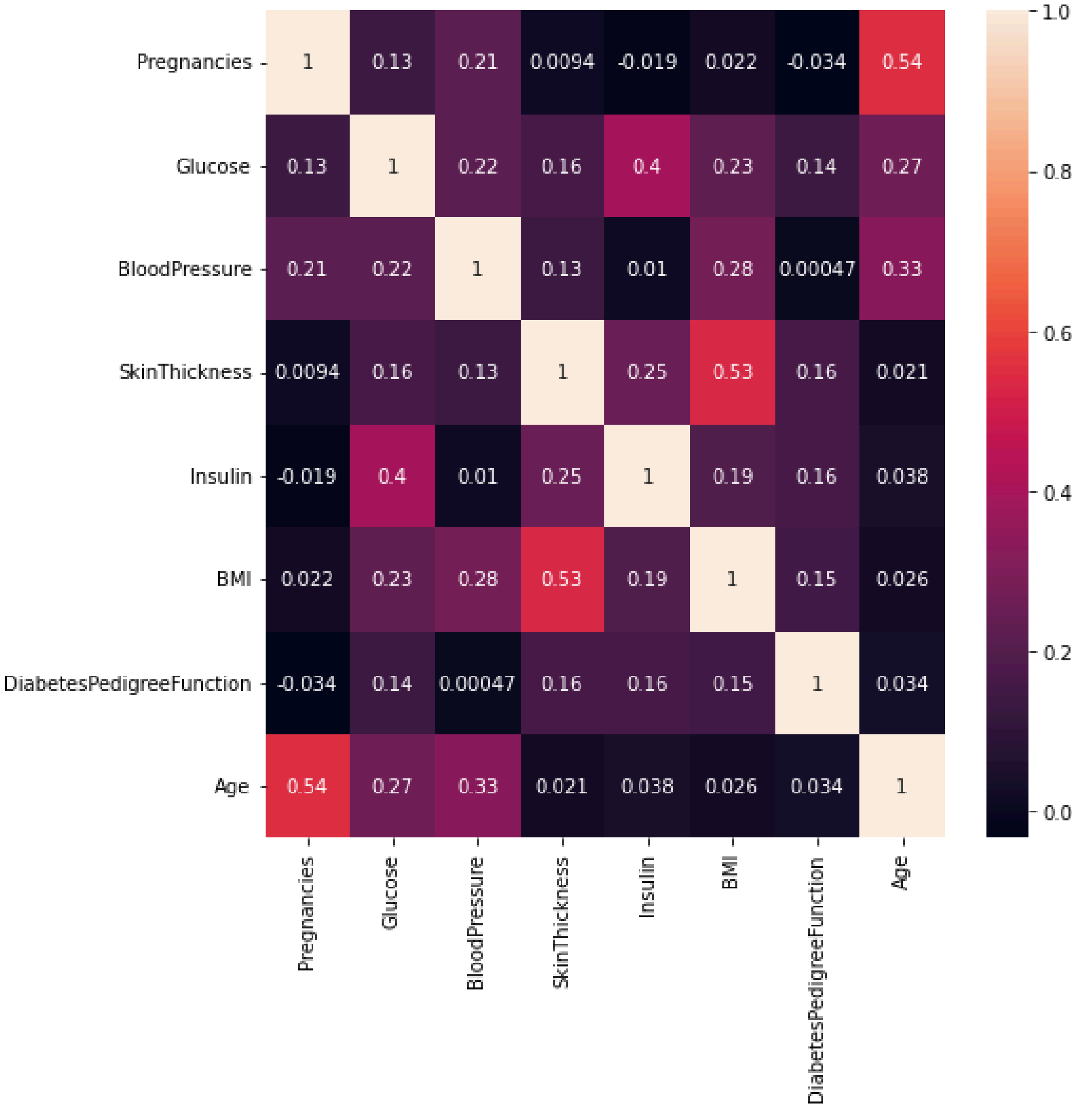
Out[21]:

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

In [22]:

```
# remove _____ & write the appropriate function name

plt.figure(figsize=(8,8))
sns.heatmap(corr_matrix, annot = True)
```

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



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Pregnancies	1	0.13	0.21	0.0094	-0.019	0.022	-0.034	0.54
Glucose	0.13	1	0.22	0.16	0.4	0.23	0.14	0.27
BloodPressure	0.21	0.22	1	0.13	0.01	0.28	0.00047	0.33
SkinThickness	0.0094	0.16	0.13	1	0.25	0.53	0.16	0.021
Insulin	-0.019	0.4	0.01	0.25	1	0.19	0.16	0.038
BMI	0.022	0.23	0.28	0.53	0.19	1	0.15	0.026
DiabetesPedigreeFunction	-0.034	0.14	0.00047	0.16	0.16	0.15	1	0.034
Age	0.54	0.27	0.33	0.021	0.038	0.026	0.034	1

Write your Answer here:

Ans 20: There are no very strong correlations between the variables. But still, there are moderate correlations between Age & Pregnancies (0.54), SkinThickness & BMI (0.53), Glucose & Insulin (0.40), all of these are positive correlations. There are no significant negative correlations.