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 [1]:
```
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: The Numpy library allows for numerical and scientific computation in python on objects called arrays. The Pandas library supports the creation of data structures and provides the tools to analyze and manipulate these data structures. The Seaborn library allows users to create colorful and visually appealing plots, where as Matplotlib.pyplot draws from the framwework of MATLAB and allows for creation of simpler plots like bars, pies, lines, scatter plots etc.

Q 2: Read the given dataset (2 Marks)

In [2]:
```
##The read_csv() function allows Python to read the mentioned data set into the notebo

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

Out[2]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Ag
0	6	148	72	35	79	33.6	0.627	5
1	1	85	66	29	79	26.6	0.351	3
2	8	183	64	20	79	23.3	0.672	3
3	1	89	66	23	94	28.1	0.167	2
4	0	137	40	35	168	43.1	2.288	3
...	...	...	...	...	...	...	...	
763	10	101	76	48	180	32.9	0.171	6
764	2	122	70	27	79	36.8	0.340	2
765	5	121	72	23	112	26.2	0.245	3
766	1	126	60	20	79	30.1	0.349	4
767	1	93	70	31	79	30.4	0.315	2

768 rows × 9 columns

Q3. Show the last 10 records of the dataset. How many columns are there? (2 Marks)

In [3]:
```
#the tail() function gives the last () rows in the data set

pima.tail(10)
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Ag
758	1	106	76	20	79	37.5	0.197	2
759	6	190	92	20	79	35.5	0.278	6
760	2	88	58	26	16	28.4	0.766	2
761	9	170	74	31	79	44.0	0.403	4
762	9	89	62	20	79	22.5	0.142	3
763	10	101	76	48	180	32.9	0.171	6
764	2	122	70	27	79	36.8	0.340	2
765	5	121	72	23	112	26.2	0.245	3
766	1	126	60	20	79	30.1	0.349	4
767	1	93	70	31	79	30.4	0.315	2

Write your Answer here:

Ans 3: There are nine columns in the diabetes dataset.

Q4. Show the first 10 records of the dataset (2 Marks)

In [4]:
```
#the head() function gives the first () rows in the data set

pima.head(10)
```

Out[4]:

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

Q5. What do you understand by the dimension of the dataset? Find the dimension of the `pima` dataframe. (3 Marks)

In [5]:
```
pima.shape
```

Out[5]:
```
(768, 9)
```

Write your Answer here:

Ans 5: The dimension of the dataset is essentially the number of rows and columns of the dataset.

Q6. What do you understand by the size of the dataset? Find the size of the `pima` dataframe. (3 Marks)

In [6]:
```
size = pima.size
print("The size of the dataset is" , size)
```

```
The size of the dataset is 6912
```

Write your Answer here:

Ans 6: The size of the data set is the product of its rows and columns.

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]:
```
print("Some information about the pima data set: ", end = '\n\n')

pima.info()
```

```
Some information about the pima data set:

<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 variables except BMI and DiabetesPedigreeFunction are integer type. The aforementioned variables are floating type.

Q8. What do we mean by missing values? Are there any missing values in the `pima` dataframe? (4 Marks)

In [8]:
```
#The isnull() function looks for missing values in the dataframe

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

Out[8]:
```
False
```

Write your Answer here:

Ans 8: Missing values in a dataframe are indicated by 'NaN' i.e. NotaNumber. This means that a given variable in the data frame does not have a value associated with it. The code above outputs boolean False when asked about missing values and hence, there are no missing values in this data set.

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. (5 Marks)

In [9]:
```
#The iLoc[] method assigns indices when they are absent/unknown

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

Out[9]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigr
count	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	
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	

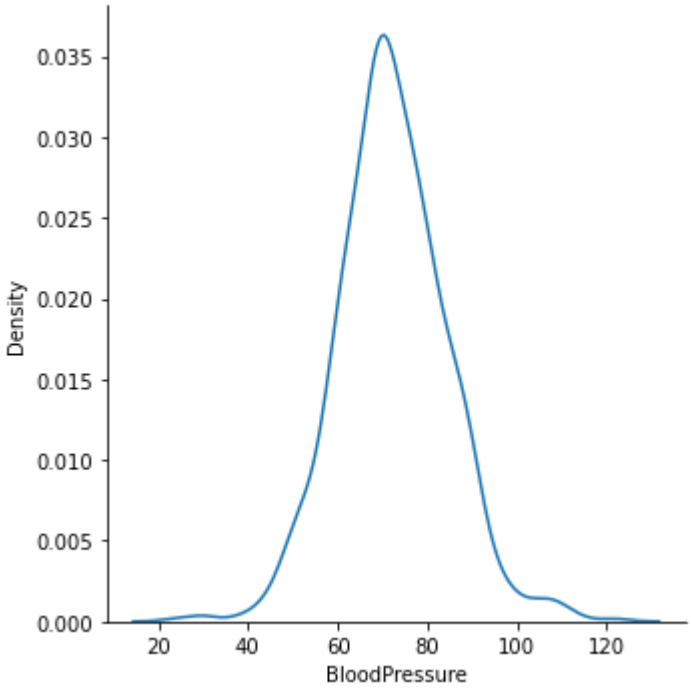
Write your Answer here:

Ans 9: The summary statistics represent the descriptive statistical measures of all the values associated with a particular variable. If we take the Pregnancies variable, we see the count shows that there are 768 samples collected, the mean represents the average number of times the Pima women were pregnant and the std represents the standard deviation or the variability of the values i.e. how far/close they are to each other. The min being 0 tells that there was an instance/instances of no pregnancies. The next three numbers 25%, 50% and 75% represent the quartiles of the pregnancy column. The first quartile tells us that 25% of the women have one or no pregnancies, 50% of the women have more than or less than 3 pregnancies (this is the median) and that 75% of the women have 6 or less pregnancies and therefore only 25% have 6 or more pregnancies. We also know through the max that the maximum times a Pima woman/women were pregnant was 17.

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

In [10]:
```
#displot() for distribution plot
#using seaborn library as sns

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



Write your Answer here:

Ans 10: From the above plot, one can estimate that a large number of Pima women have blood pressure between 60 and 80 mm Hg. Moreover, one can observe that the instances of women having a blood pressure less than 40 or more than 100 are rare.

Q 11. What is the 'BMI' of the person having the highest 'Glucose'? (2 Marks)

In [11]:
```
#max() function being used to find max value in column.

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

Out[11]:
```
661    42.9
Name: BMI, dtype: float64
```

Write your Answer here:

Ans 11: The BMI of the person have the highest glucose is 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?

(4 Marks)

In [12]:
```
m1 = pima['BMI'].mean()  # mean
print("The mean is " , m1)

m2 = pima['BMI'].median()  # median
print("The median is " , m2)

m3 = pima['BMI'].mode()[0]  # mode
print("The mode is " , m3)
```

```
The mean is  32.45080515543617
The median is  32.0
The mode is  32.0
```

Write your Answer here:

Ans 12: The median and mode are equal but the mean is slightly larger.

Q13. How many women's 'Glucose' levels are above the mean level of 'Glucose'? (2 Marks)

In [13]:
```
pima[pima['Glucose'] > pima['Glucose'].mean()].shape[0]

#The .shape[0] function tells us the number of elements in the Glucose column that are
```

Out[13]:
```
343
```

Write your Answer here:

Ans 13: There are 343 women who have a Glucose level above the mean.

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['
```

Out[14]:
```
22
```

Write your Answer here:

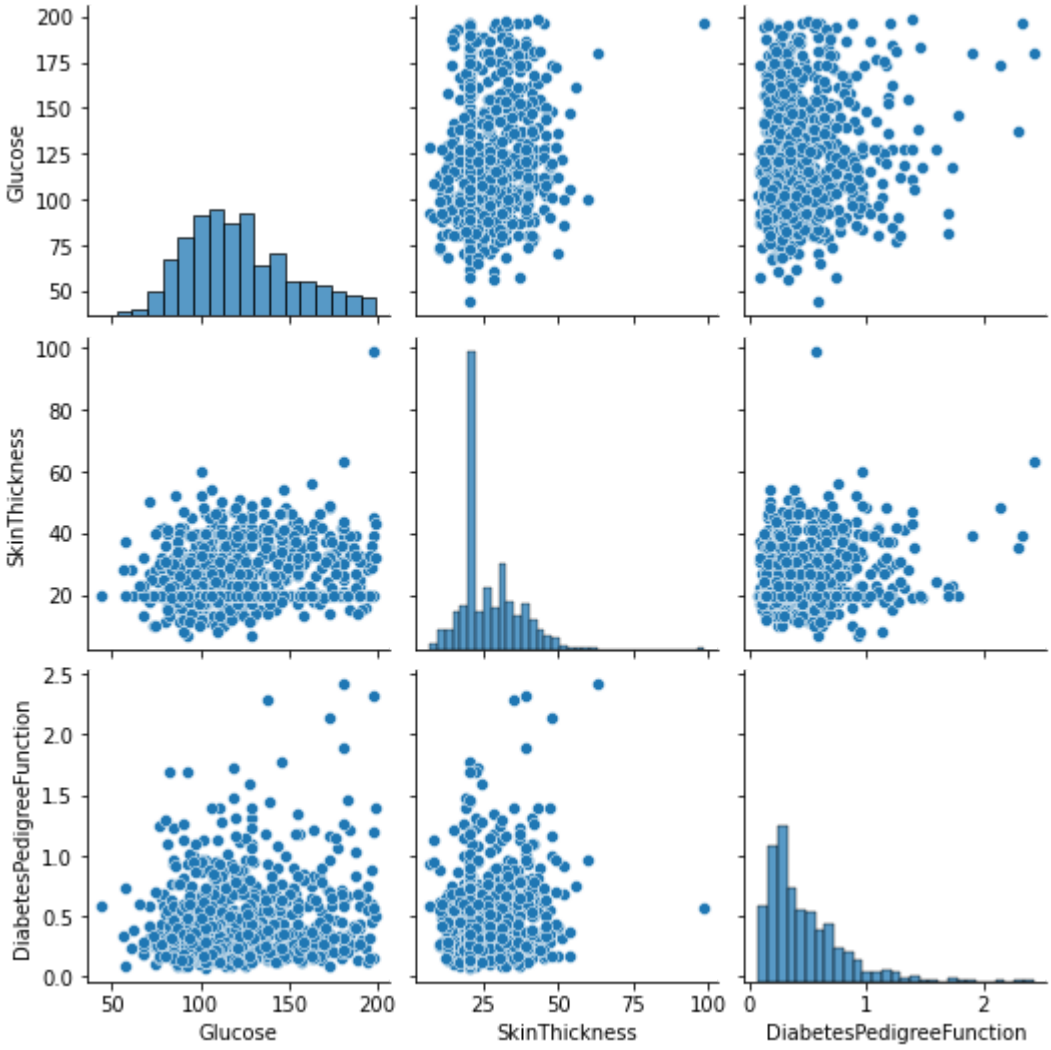
Ans 14: Twenty two women have their blood pressure equal to and less than the median

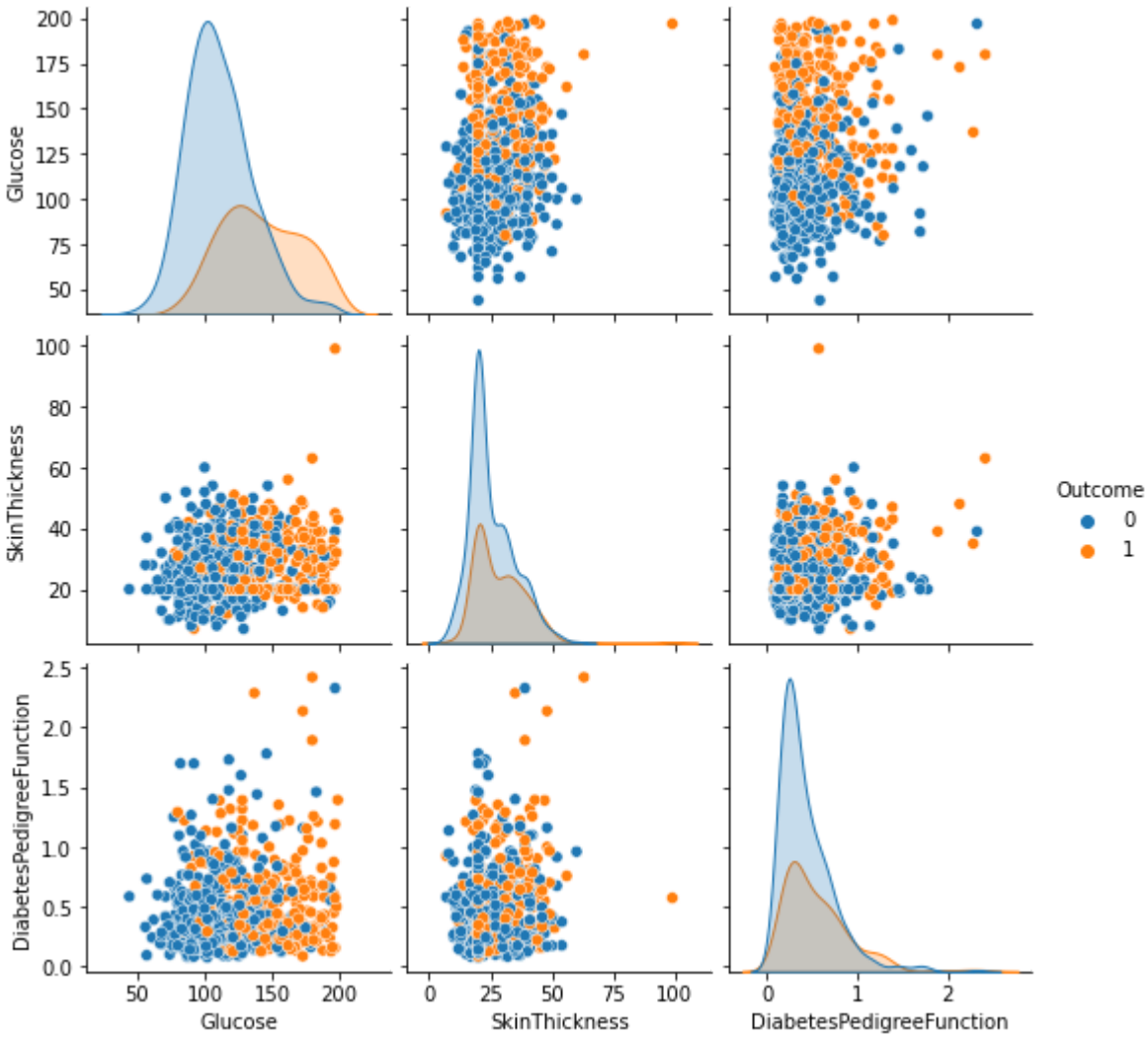
Q15. Create a pairplot for the variables 'Glucose', 'SkinThickness', and 'DiabetesPedigreeFunction'. Write your observations from the plot. (3 Marks)

In [15]:
```
# Remove _____ & write the appropriate function name

#pairplot plotted with and without Outcome variable
```

```
sns.pairplot(data = pima, vars = ['Glucose', 'SkinThickness', 'DiabetesPedigreeFunctio
print('\n')
sns.pairplot(data = pima, vars = ['Glucose', 'SkinThickness', 'DiabetesPedigreeFunctio
plt.show()
```





Write your Answer here:

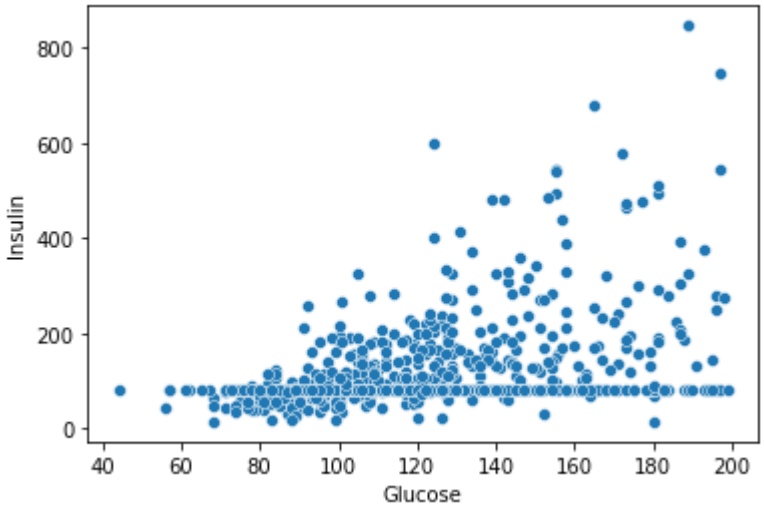
Ans 15: Both the pairplots show interesting correlations between different aspects of the data. For instance, in the one above it appears that skin thickness and diabetes pedigree function (dpf) are correlated in that lower dpf is associated with lower skin thickness in many cases. The similar relation in the 2nd plot shows that higher skin thickness and higher dpf results in more cases of diabetes. The 2nd plot shows that glucose above 125 and dpf between 0 and 1 results in many cases of diabetes

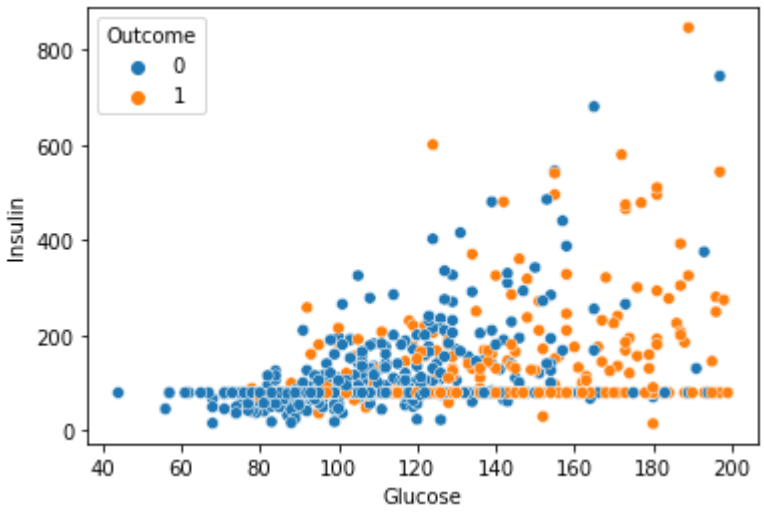
Q16. Plot the scatterplot between 'Glucose' and 'Insulin'. Write your observations from the plot. (4 Marks)

In [16]:
```
# Remove _____ & write the appropriate function name

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

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





Write your Answer here:

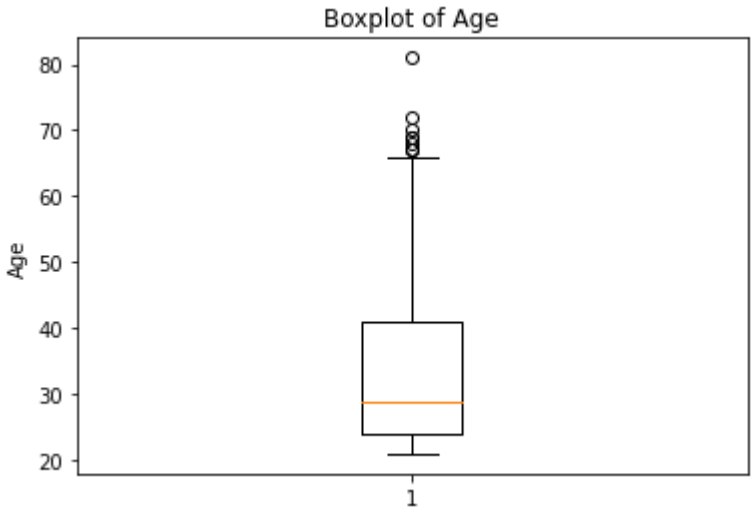
Ans 16: From the first plot one can say that there is a significant number of women who have glucose levels between 60 and 120 and take less than 200 mu of insulin. It also seems that higher glucose levels are resulting in higher insulin dosage. However, the 2nd scatterplot shows that a significant number of women who have glucose above 120 and insulin dosage between 100 and 300 are diabetic. Comparing the two plots, one can also confirm that low glucose levels are associated with low doses of insulin.

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()
```

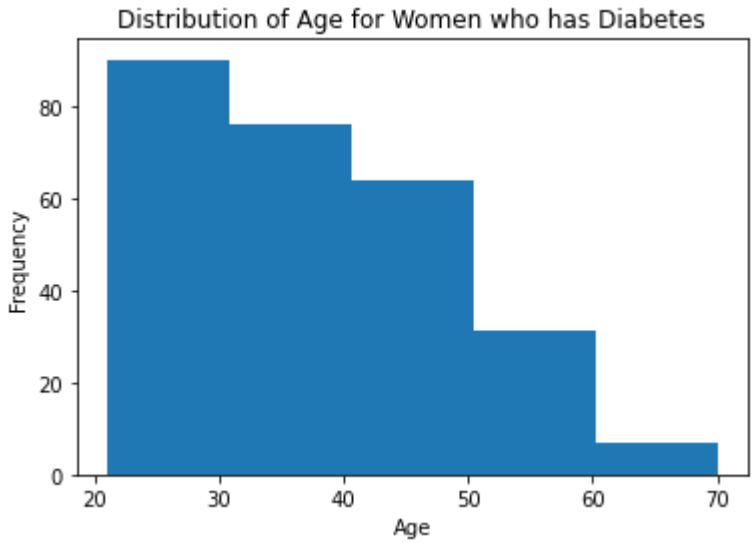


Write your Answer here:

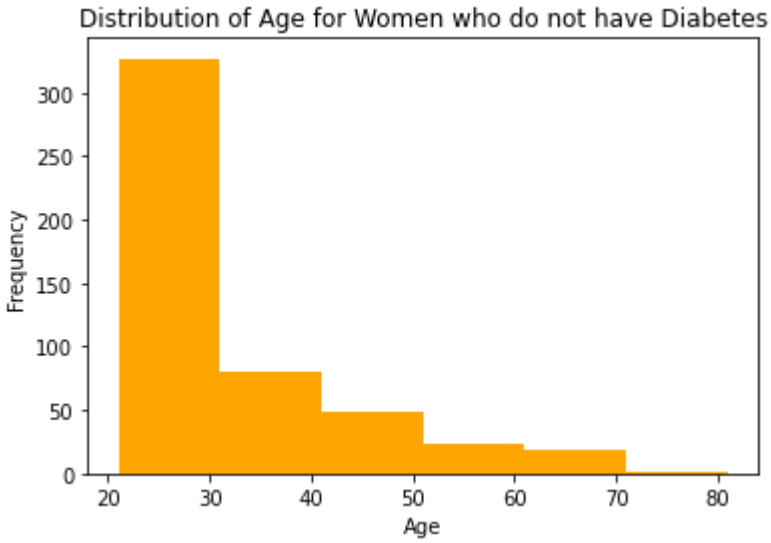
Ans 17:There is one outlier in the sample of women who is around 80 years old.

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. (5 Marks)

In [18]:
```
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]:
```
plt.hist(pima[pima['Outcome'] == 0]['Age'], bins = 6, color = "Orange")
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: Both histograms show that most women between 20 and 30 years of age have and do not have diabetes. However, that can be explained by the larger number of non-diabetic women in the tribe. In the first histogram, women between 20 and 30 years of age are the most diabetic, even when compared with women of other age groups. A similiar pattern is observed for the non-diabetic women. However, after comparing the frequencies on the histograms there seems to be more diabetic than non-diabetic women between 40 and 50 years.

Q 19. What is the Interquartile Range of all the variables? Why is this used? Which plot visualizes the same? (5 Marks)

In [20]:
```
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 interquartile range gives the range of the middle 50% of the data for a particular variable. The boxplot visualizes it.

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

In [21]:
```
corr_matrix = pima.iloc[ : ,0 : 8].corr()

corr_matrix
```

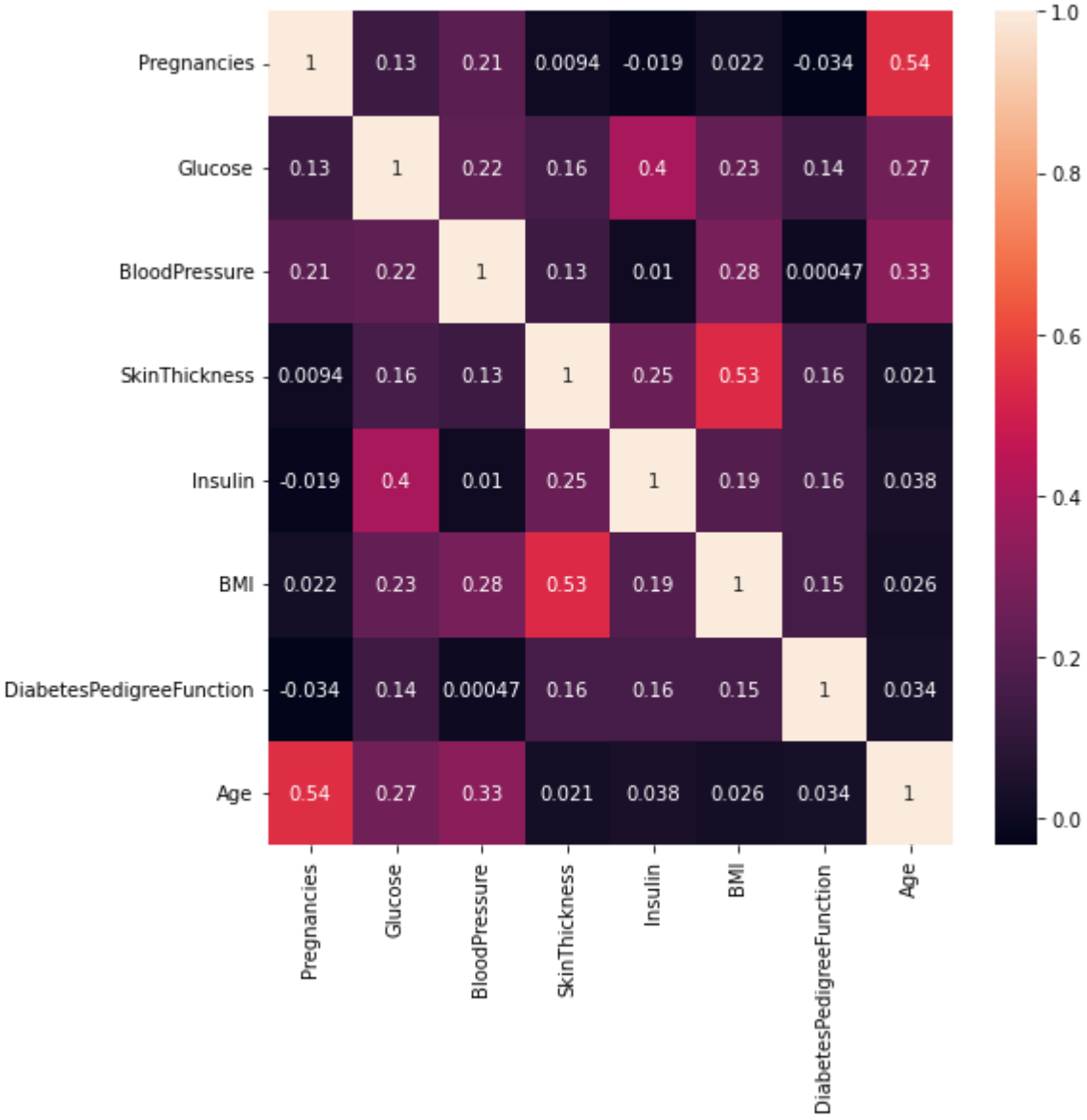
Out[21]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	D
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()
```



Write your Answer here:

Ans 20: From the plot, we can tell that the the darker the color of a box, the weaker is the correlation between the two variables, but the lighter the color the stronger the correlation. The plot is essentially a symmetric matrix, with correlation between the similiar variables on the diagonals and the other ones around it. Glucose and insulin have a high correlation of 0.4.

Thank you.