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 [5]:

```
# 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: Importng numpy provides a numerical library for certain capabilites like using arrays and is also reliable in its efficiency and that it uses up less memory. The usage of pandas in python is vital because it offers in datasets the ability to use tables and columns with the data as well as data structures like series and dataframes. For data visualization, importing seaborn and matplotlib is important because the libraries contain many kinds of plots that can be utilized. Matplotlib includes more simple plots to create while seaborn are more advanced plots in their understanding and building, but both are useful in plotting your data.

Q 2: Read the given dataset (2 Marks)

In [6]:

```
# 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? (2 Marks)

In [28]:

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

pima.tail(10)
```

Out[28]:

	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: There are 9 columns in the dataset excluding the index.

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

In [8]:

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

pima.head(10)
```

Out[8]:

	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. (3 Marks)

In [9]:

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

pima.shape
```

Out[9]:

```
(768, 9)
```

Write your Answer here:

Ans 5: The dimension of the dataset indicates how many rows followed by how many columns there are in the dataframe. The pima dataframe from the output shows that there are 768 rows and 9 columns in the pima dataframe, and by rows it means there are 768 entries in the dataset.

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

In [16]:

```
# 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 6: The dataset has a range index of 768 entries or rows that range from row 0 to row 767, and it has 9 total data columns. From the 9 columns they each have 768 non-null observations which indicates there are not any missing values from any of the observations. Each column has a data type of an integer except for the BMI and DiabetesPedigreeFunction columns that have a float data type, meaning that those columns will have outcomes that do not have whole numbers. The memory used to save the dataframe is at 54.1 kilobytes.

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 [17]:

```
# 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: Columns 0 (Pregnancies), 1 (Glucose), 2 (BloodPressure), 3 (SkinThickness), 4 (Insulin), 7 (Age), and 8 (Outcome) have integer data types and columns 5 (BMI) and 6 (DiabetesPedigreeFunction) have float data types.

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

In [20]:

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

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

Out[20]:

```
False
```

Write your Answer here:

Ans 8: Missing values is whether any of the entries in the dataframe contain any missues values in it. If we pass a key that is not defined then its value will be marked as a NaN in its entry. From the output it looks like there are no missing values in the 768 entries of 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. (5 Marks)

In [50]:

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

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

Out[50]:

	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

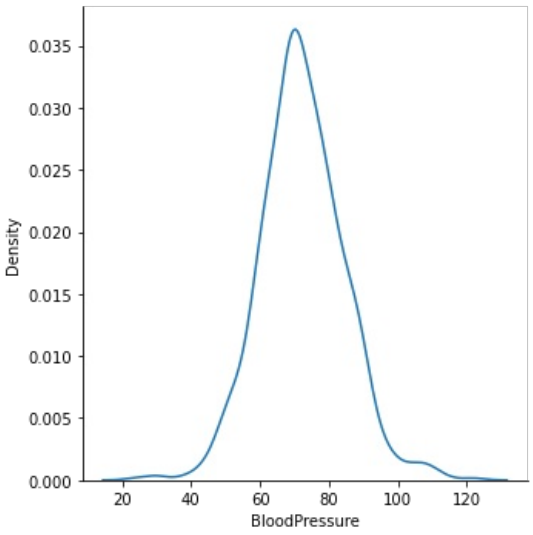
Write your Answer here:

Ans 9: Analyzing the glucose variable from the dataset which totals 768 entries, the average glucose level or mean is 121.675781 among all the women. The calculated std or standard deviation which shows how the dataset spreads in relation to the mean is 30.436252, and that number is one standard deviation away from the mean. The minimum glucose level in the database is 44.000000 and the entry with the maximum glucose level is 199.000000. The 25th percentile or the first quartile represents what percentage is below that quartile in the dataset, which is 99.750000. The 50th percentile, or the second quartile or median, represents what percentage is below that quartile in the dataset, which is 72.000000. The 75th percentile, or the third quartile represents what percentage is below that quartile in the dataset, which is 140.250000. This represents some of the basic statistics of this variable in the dataset.

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

In [22]:

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

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

plt.show()
```



Write your Answer here:

Ans 10: From the distribution plot above it looks as though the range of blood pressure in the dataset is between 60 and 80. The blood pressure variable has a normal distribution and normal bell shaped curve with a mean of around 70. This range observed from the distribution plot for blood pressure could be a factor in the diabetes research done on the Pima Indians tribe in how they relate to one another.

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

In [25]:

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

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

Out[25]:

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

Write your Answer here:

Ans 11: Using the max function, the BMI of the individual with 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 [24]:

```
# 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: The median and mode of the BMI variable are both 32 and the mean is 32.45 so they are all very close to being equal.

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

In [26]:

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

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

Out[26]:

```
343
```

Write your Answer here:

Ans 13: Using the mean function, there are 343 women who have glucose levels over 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 [31]:

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

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

Out[31]:

	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

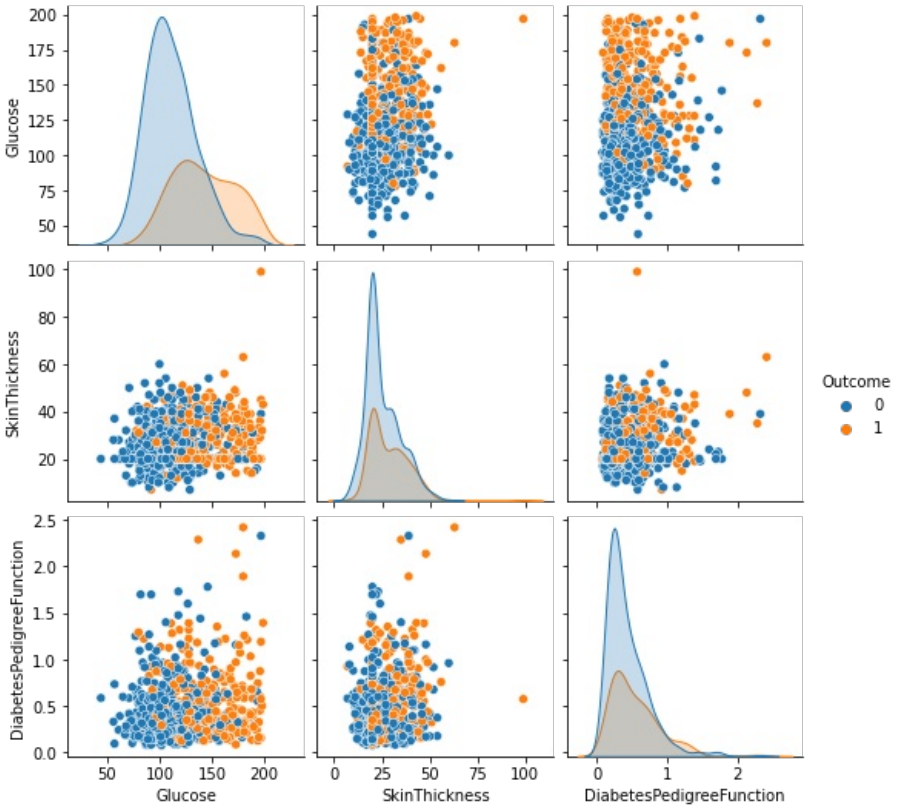
Write your Answer here:

Ans 14: Using the median function, there are 22 women who have the median blood pressure of 72 and also have a BMI median less than 32 which was answered from a previous question.

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

In [32]:

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

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



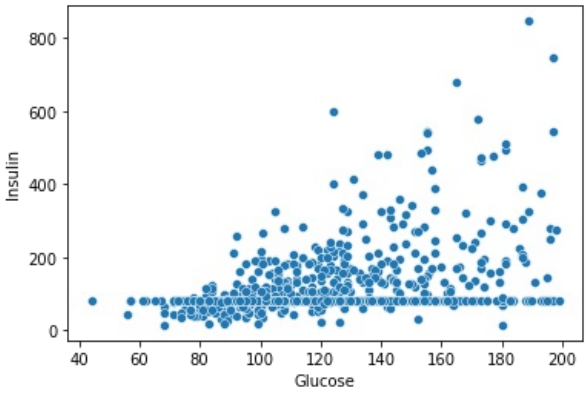
Write your Answer here:

Ans 15: Between the plots of 'Skin Thickness' and 'Glucose', the skin thickness is slightly thicker for people with lower levels of glucose and therefore less people that have diabetes. Overall there shows to be a small positive linear relationship between the two plots with really just one outlier with skin thickness levels. Between the plots of 'DiabetesPedigreeFunction' and 'Glucose', the higher the diabetes pedigree function levels the higher the glocuse levels, showing that there is a little positive linear relationship between the two plots. It does not look like there is any difference in the number of people with diabetes or do not have diabetes in regards to people with higher diabetes pedigree function levels. Between the plots of 'Skin Thickness' and 'DiabetesPedigreeFunction', there is a small indication that the higher the diabetes pedigree function levels are the smaller your skin thickness will be. The plots look as though they lean toward a slightly negative linear relationship. It also does not look like the relationship between skin thickness and diabetes pedigree function impacts whether you have diabetes or not. There is a few outliers in the plot with high skin thickness and a few with low diabetes pedigree function levels as well as a few with high pedigree function levels and lower skin thickness levels.

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

In [34]:

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

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



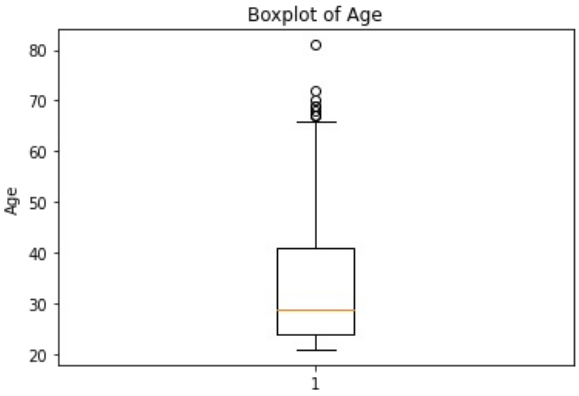
Write your Answer here:

Ans 16: From the scatterplot it looks like that the more the glucose levels increase, the higher the insulin levels goes. This shows between the two variables that there is a positive linear relationship. There does seem to be a particular insulin level that was given to a certain number of individuals who all have different levels of glucose, so it shows in the plot as a flat line going across the x-axis. I do wonder if the two hour insulin serum level was given to a certain group of individuals by design for the research.

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

In [35]:

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

plt.boxplot(pima['Age'])

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



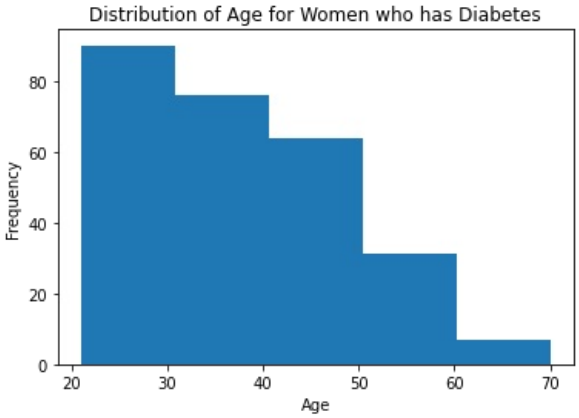
Write your Answer here:

Ans 17: From the plotting of the boxplot there is a few outliers outside the maximum numerical quanities or the upper whisker.

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)

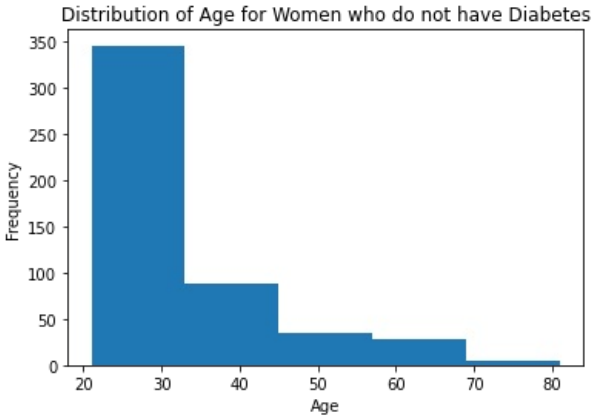
In [36]:

```
# 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 [37]:

```
# 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: My first observation is that both histograms show a negative linear relationship in its distribution, indicating that the frequencies of each plot decline as the age of the women increases. The difference seems to be that for the women that do have diabetes, the frequency only slightly goes down as the women get older, where for the women that do not have diabetes there is a huge drop in frequency at a young age among the number of women before slowly going down after that as the age increases. Considering both these histograms it seems as though the younger population has more issues with diabetes than the older population in the dataset despite the large differences in numbers between who has diabetes and who does not. So overall because the two groups have similar plots it can indicate that this variable only shows what previous research all ready has concluded.

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

In [39]:

```
# 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 Interquartile Range is used to demonstrate where the middle 50 percent of the output is located in the boxplot and how the range is spread out. Because this also includes the median of the dataset, this is where the bulk of the values will be when looking at the variables in the boxplot and when calculating the percentiles. It is also very useful in that it has less of an impact from the outliers that could be in the dataset.

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

In [42]:

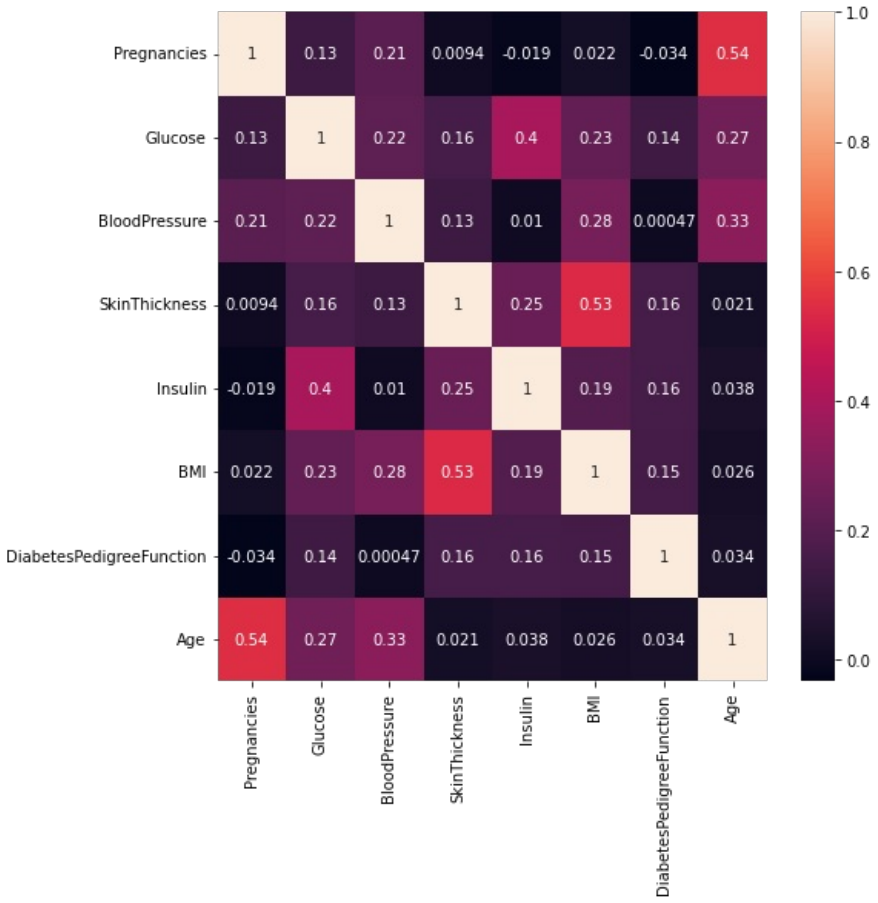
```
# Remove _____ & write the appropriate function name and run the code

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

corr_matrix
```

Out[42]:

	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 [49]:

```
# 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 calculations and plot it looks like most of the relationships and correlations between all of the variables show a positive relationship because they are greater than zero. The exceptions to this are the few negative relationships that are between the variables 'pregancies' and 'insulin' and the variables 'pregnancies' and 'diabetes pedigree function', indicating that they are going in opposite directions. These associations between variables in the diabetes research and help identify patterns in the dataset.