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 [7]: # remove _____ & write the appropriate library name

        import numpy as np
        import pandas as pd

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

Write your Answer here:

Ans 1:numpy is used for high function mathmatical problems that have multideminsions. matplotlibis use for static visuabls. pandas is used for satistics and data science. seaborn is also a visual libery that makes attractive visuals for data

Q 2: Read the given dataset (1 Mark)

```
In [8]: # 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 [9]: # remove ______ and write the appropriate number in the function

        pima.tail(10)
```

```
Out[9]:      Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigre
        758            1      106             76             20       79  37.5
        759            6      190             92             20       79  35.5
        760            2       88             58             26       16  28.4
        761            9      170             74             31       79  44.0
        762            9       89             62             20       79  22.5
        763           10      101             76             48      180  32.9
        764            2      122             70             27       79  36.8
        765            5      121             72             23      112  26.2
        766            1      126             60             20       79  30.1
        767            1       93             70             31       79  30.4
```

Write your Answer here:

Ans 3:9

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

```
In [… # remove _____ & write the appropriate function name and the number o

      pima.head(10)
```

Out[10]:

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

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

```
In [11]: # remove _____ & write the appropriate function name

         pima.shape

Out[11]: (768, 9)
```

Write your Answer here:

Ans 5:the dataframe is 768 rows x 9 colums.

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

```
In [12]: # remove _____ & write the appropriate function name

         pima.size
```

```
Out[12]: 6912
```

Write your Answer here:

Ans 6:there is 6912 varibales

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 [13]: # 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:pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function, age outcome.

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

```
In [14]: # remove _____ & write the appropriate function name

         pima.isnull().values.any()

Out[14]: False
```

Write your Answer here:

Ans 8:no mising values

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 [15]: # remove _____ & write the appropriate function name

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

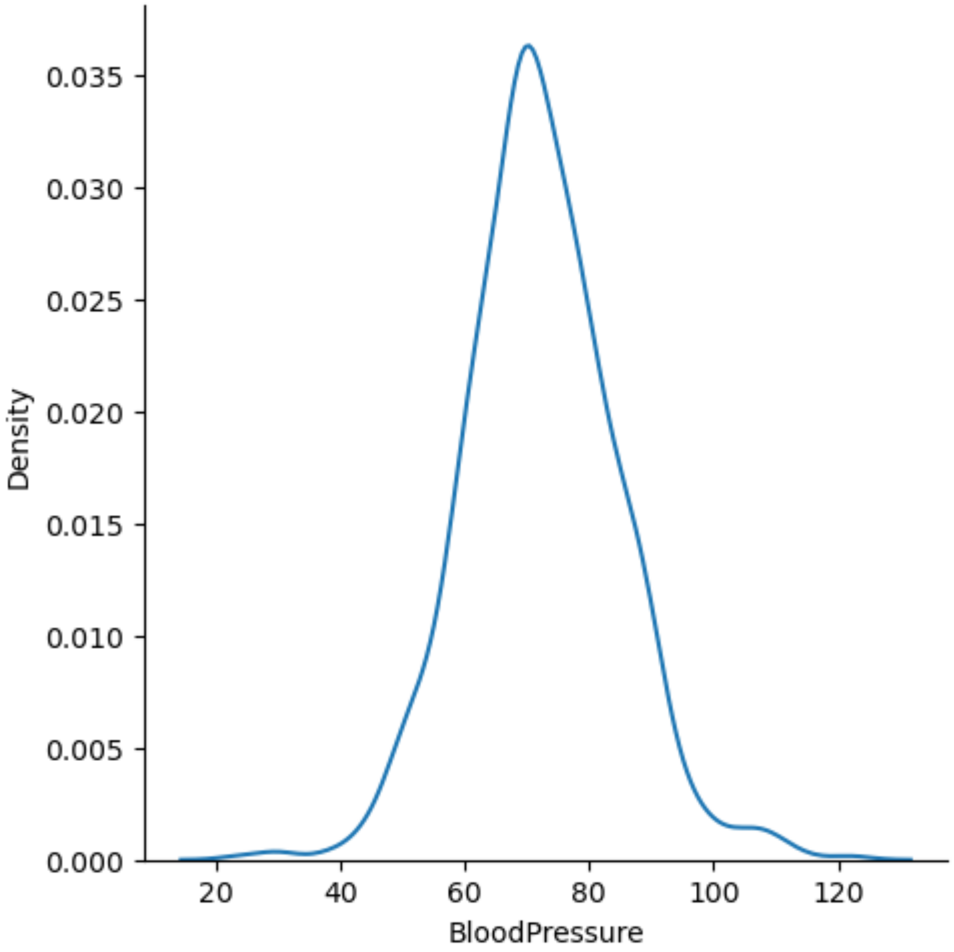
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 8 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   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  
dtypes: float64(2), int64(6)
memory usage: 48.1 KB
```

Write your Answer here:

Ans 9:pregnacies, glucose, blood pressue, skin thickness, age, and insulin have 64 whole number values. bmi, diabetes pedigree function have decimal number. each of which has 64 values.

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

```
In [16]: # remove _____ & write the appropriate library name
         import seaborn as sns
         import matplotlib.pyplot as plt
         sns.displot(pima['BloodPressure'], kind='kde')
         plt.show()
```



Write your Answer here:

Ans 10:the majority have a blood pruessre in the 70. very little people have it in the 40s and under and 100 and above

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

```
In [17]: # remove _____ & write the appropriate function name

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

```
Out[17]: 661    42.9
         Name: BMI, dtype: float64
```

Write your Answer here:

Ans 11:the person with the hightest glucose level has a marker 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 [18]: # 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 3 measurement are not all the same.the mean is off by 0.4.

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

```
In [19]: # remove _____ & write the appropriate function name

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

Out[19]: 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 [… # remove _____ & write the appropriate column name

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

Out[20]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigr
14	5	166	72	19	175	25.8	
93	4	134	72	20	79	23.8	
103	1	81	72	18	40	26.6	
205	5	111	72	28	79	23.9	
299	8	112	72	20	79	23.6	
325	1	157	72	21	168	25.6	
330	8	118	72	19	79	23.1	
366	6	124	72	20	79	27.6	
380	1	107	72	30	82	30.8	
393	4	116	72	12	87	22.1	
406	4	115	72	20	79	28.9	
446	1	100	72	12	70	25.3	
460	9	120	72	22	56	20.8	
488	4	99	72	17	79	25.6	
497	2	81	72	15	76	30.1	
510	12	84	72	31	79	29.7	
568	4	154	72	29	126	31.3	
615	3	106	72	20	79	25.8	
635	13	104	72	20	79	31.2	
644	3	103	72	30	152	27.6	
717	10	94	72	18	79	23.1	
765	5	121	72	23	112	26.2	

Write your Answer here:

Ans 14 there are 22 women

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

In […

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

sns.pairpolt(data=pima,vars=['Glucose', 'SkinThickness', 'DiabetesPed
plt.show()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_21692\3835635000.py in <module>
      1 # remove _____ & write the appropriate function name
      2 
----> 3 sns.pairpolt(data=pima,vars=['Glucose', 'SkinThickness', 'Diabetes
PedigreeFunction'], hue='Outcome')
      4 plt.show()

AttributeError: module 'seaborn' has no attribute 'pairpolt'
```

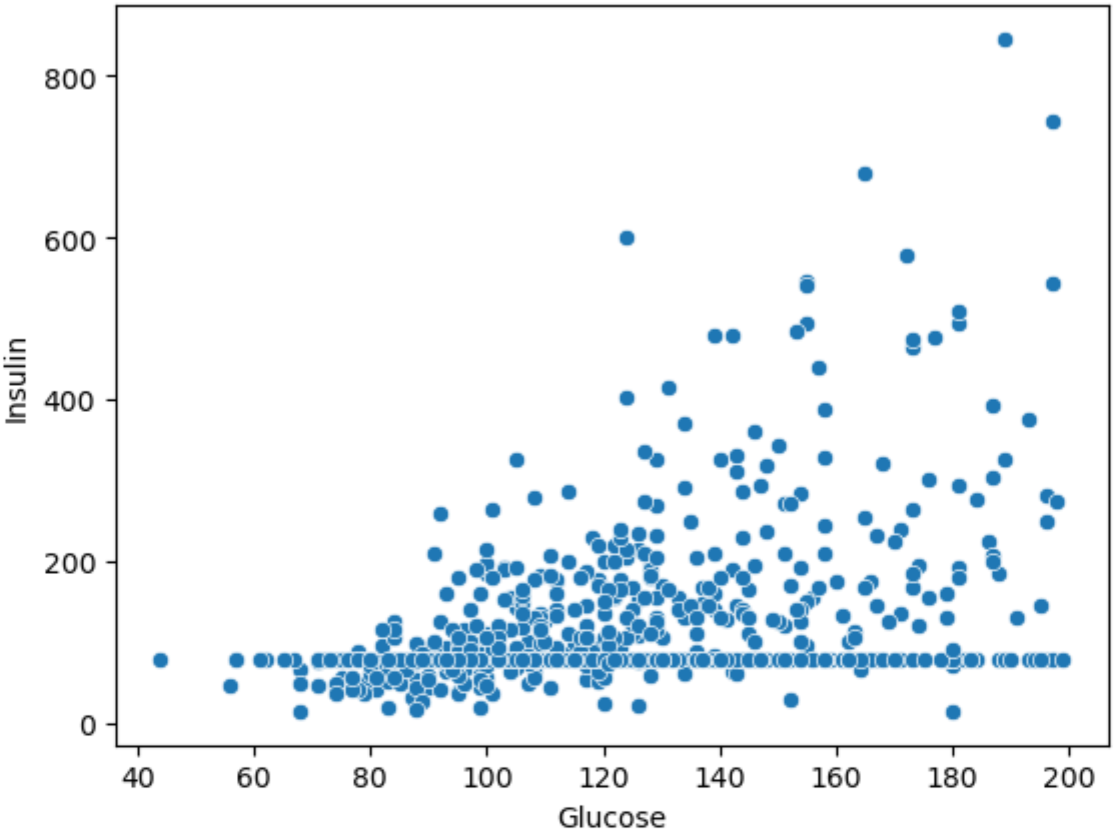
Write your Answer here:

Ans 15:there is not negative or positve correlation, however there is a range where central tendacyy is and a cut from both tails.

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

```
In [40]: # remove _____ & write the appropriate function name

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



Write your Answer here:

Ans 16: there is a positive correlation with glucose and insulin. the higher glucose gose up the higher insulin goses up.

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

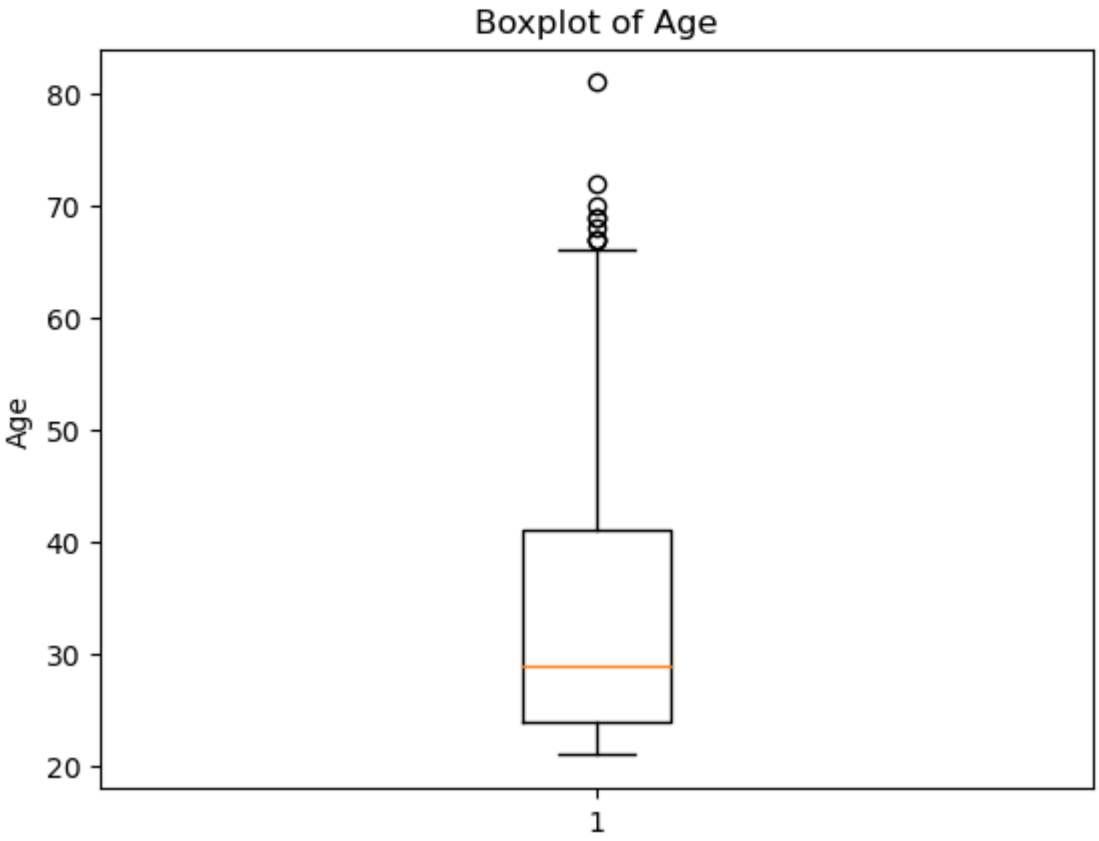
```
In [24]: import numpy as np
         import pandas as pd

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

```
In [37]: # remove _____ & write the appropriate function and column name

         plt.pyplot.boxplot(pima['Age'])

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



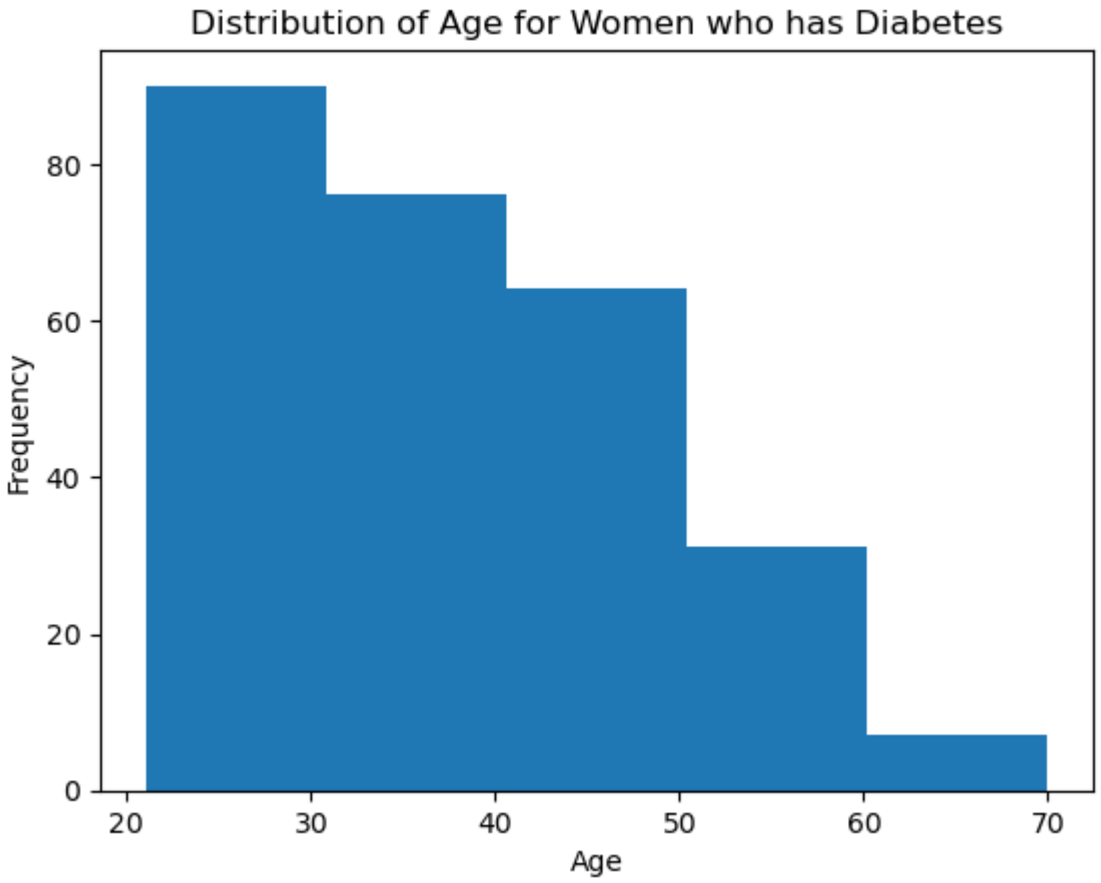
Write your Answer here:

Ans 17:there are 6 outliers.

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)

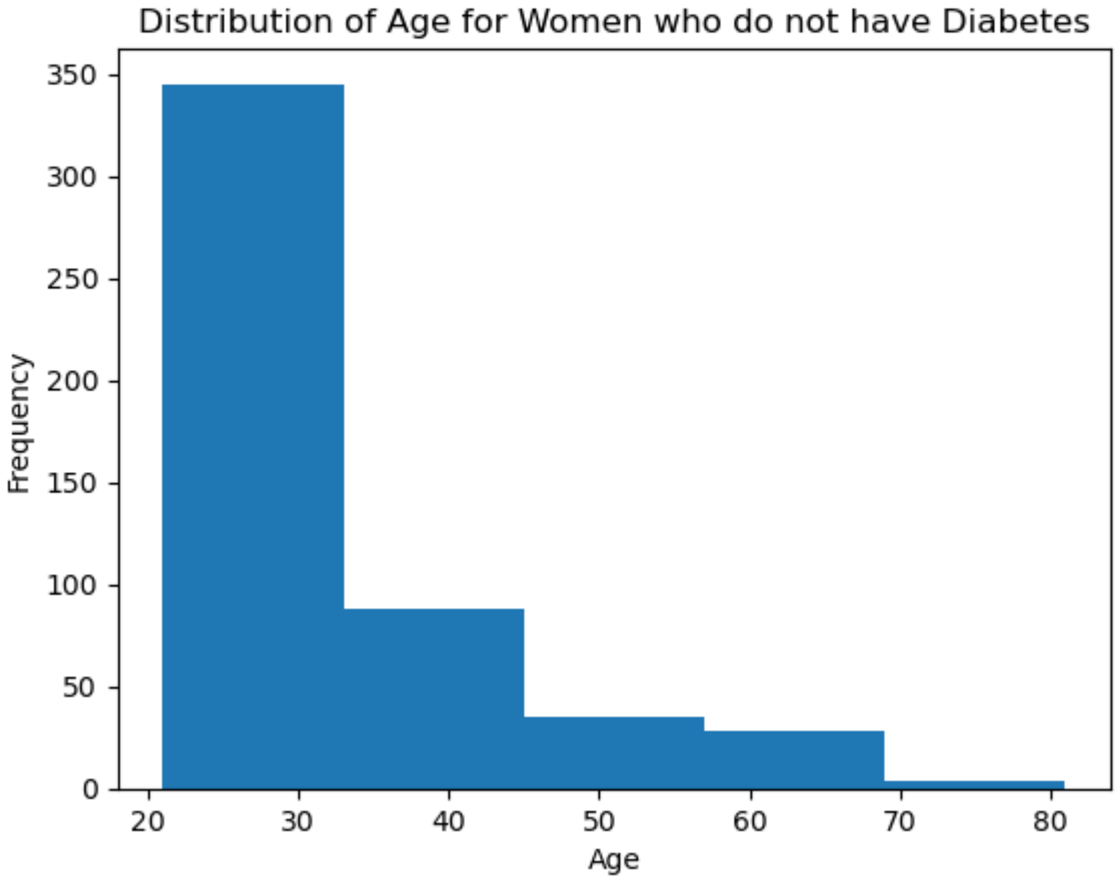
```
In [36…  # remove _____ & write the appropriate function and column name

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



```
In [41…  # remove _____ & write the appropriate function and column name

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



Write your Answer here:

Ans 18:women how have diabetes has more frequency in ages20 -60 compared to women who do not have diabetes have a drop off thats significant after age 30.

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

```
In [42]: # 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: its used to see the first Q1 and see the differene between the Q1 and Q3.the box plot visualizes the same information as the interquartile range. It shows the minimum, first quartile, median, third quartile, and maximum of a dataset, which are the same values as the interquartile range.

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

```
In [43…  # remove _____ & write the appropriate function name and run the co

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

         corr_matrix
```

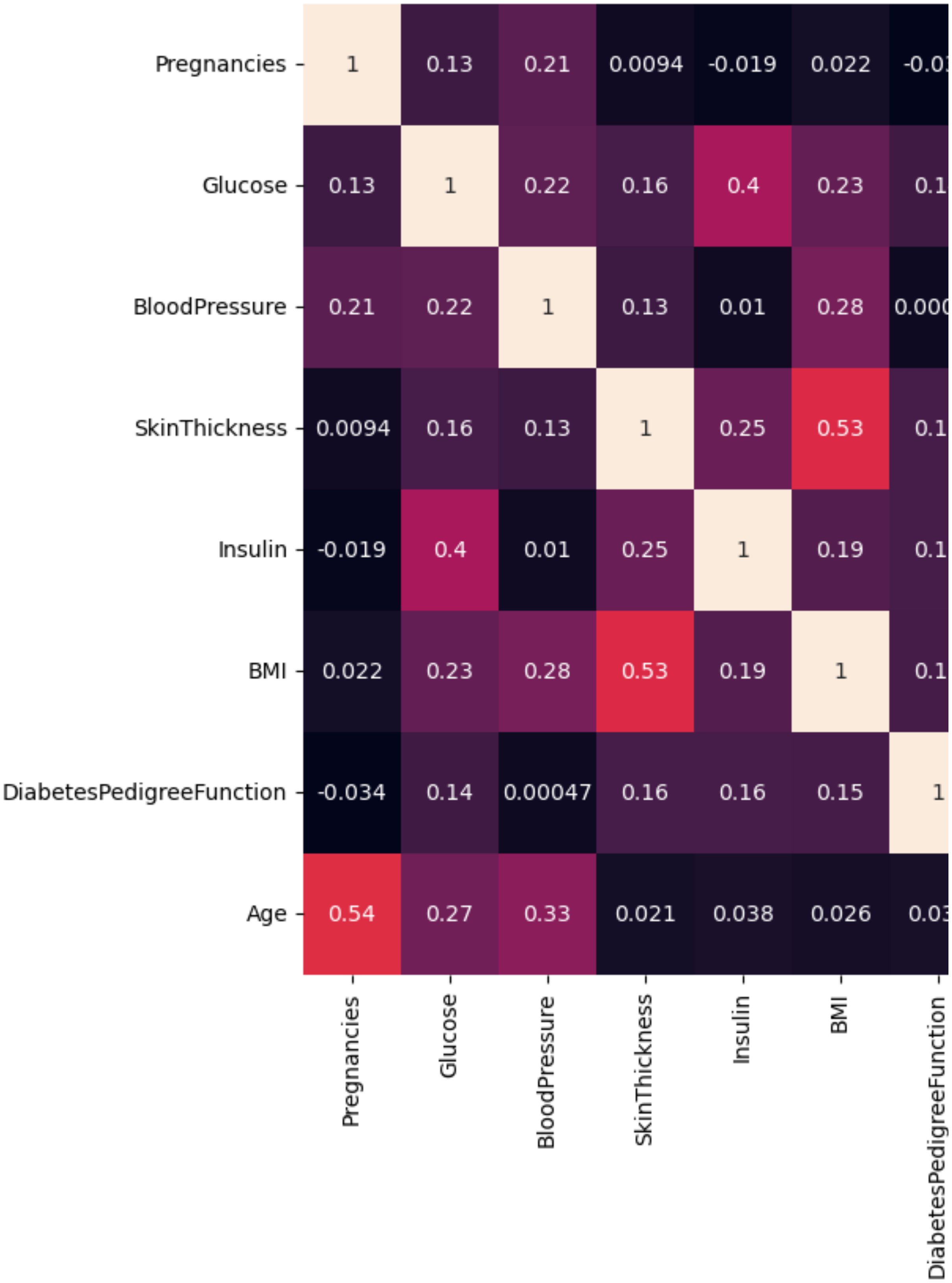
Out[43]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insu
Pregnancies	1.000000	0.128022	0.208987	0.009393	-0.0187
Glucose	0.128022	1.000000	0.219765	0.158060	0.3961
BloodPressure	0.208987	0.219765	1.000000	0.130403	0.0104
SkinThickness	0.009393	0.158060	0.130403	1.000000	0.2454
Insulin	-0.018780	0.396137	0.010492	0.245410	1.0000
BMI	0.021546	0.231464	0.281222	0.532552	0.1899
DiabetesPedigreeFunction	-0.033523	0.137158	0.000471	0.157196	0.1582
Age	0.544341	0.266673	0.326791	0.020582	0.0376

```
In [46]: # remove _____ & write the appropriate function name

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

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



Write your Answer here:

Ans 20: age has a strong correlation with pregancy. diabetes pedegree overall has a week correlation. bmi has a strong correlation with skin thickness. insulin has a moderate correlation with glucose.

```
In [ ]:
```