In [1]:

```
import pandas as pd
df = pd.read_csv(r'C:\Users\mishu\OneDrive\Documents\Most-Recent-Cohorts-Institution.c
print(df)
```

```
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0              6      148             72             35        0  33.6   
1              1       85             66             29        0  26.6   
2              8      183             64              0        0  23.3   
3              1       89             66             23       94  28.1   
4              0      137             40             35      168  43.1   
..           ...      ...            ...            ...      ...   ...   
763           10      101             76             48      180  32.9   
764            2      122             70             27        0  36.8   
765            5      121             72             23      112  26.2   
766            1      126             60              0        0  30.1   
767            1       93             70             31        0  30.4   

     DiabetesPedigreeFunction  Age  Outcome  
0                       0.627   50        1  
1                       0.351   31        0  
2                       0.672   32        1  
3                       0.167   21        0  
4                       2.288   33        1  
..                        ...  ...      ...  
763                     0.171   63        0  
764                     0.340   27        0  
765                     0.245   30        0  
766                     0.349   47        1  
767                     0.315   23        0  

[768 rows x 9 columns]
```

In [2]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [3]:

```
df.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
```

In [4]:

```
df.isnull
```

Out[4]:

```
<bound method DataFrame.isnull of      Pregnancies  Glucose  BloodPressure  SkinThick
ness  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age  Outcome
0                       0.627   50        1
1                       0.351   31        0
2                       0.672   32        1
3                       0.167   21        0
4                       2.288   33        1
..                        ...  ...      ...
763                     0.171   63        0
764                     0.340   27        0
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0

[768 rows x 9 columns]>
```

In [5]:

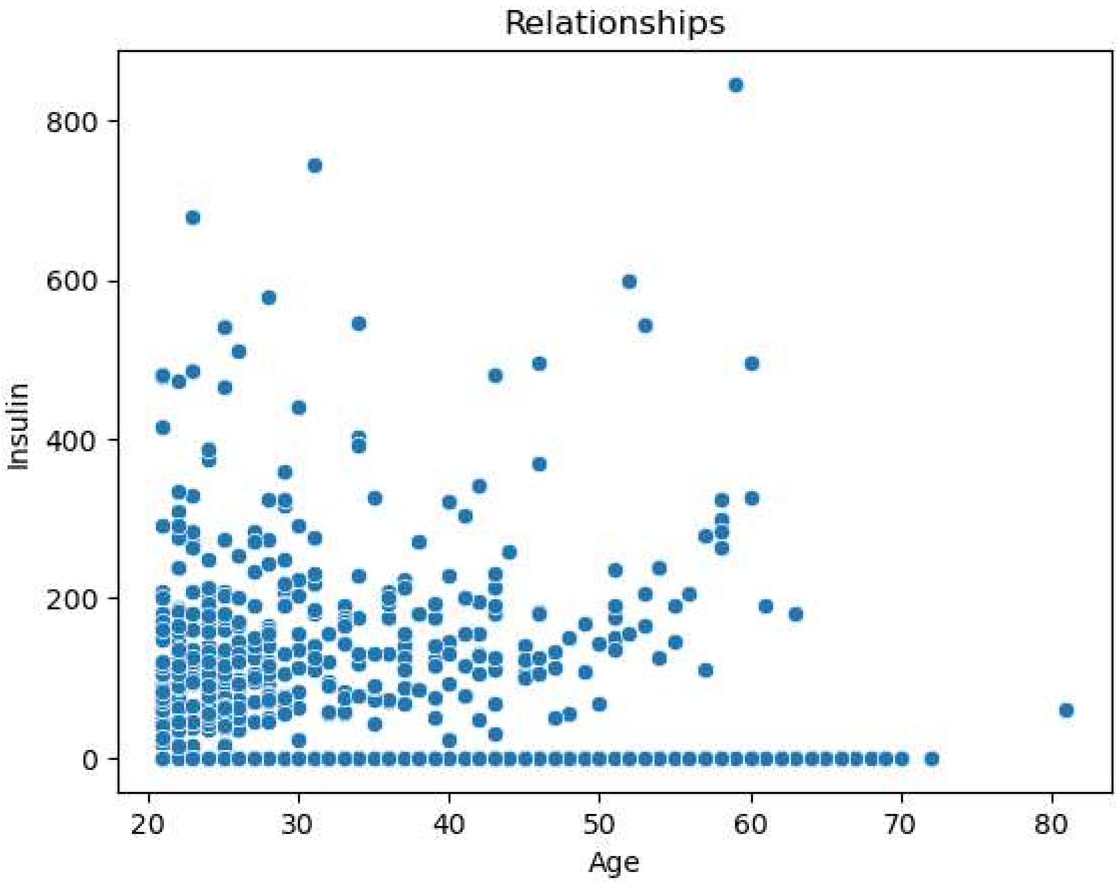
```
df.shape
```

Out[5]:

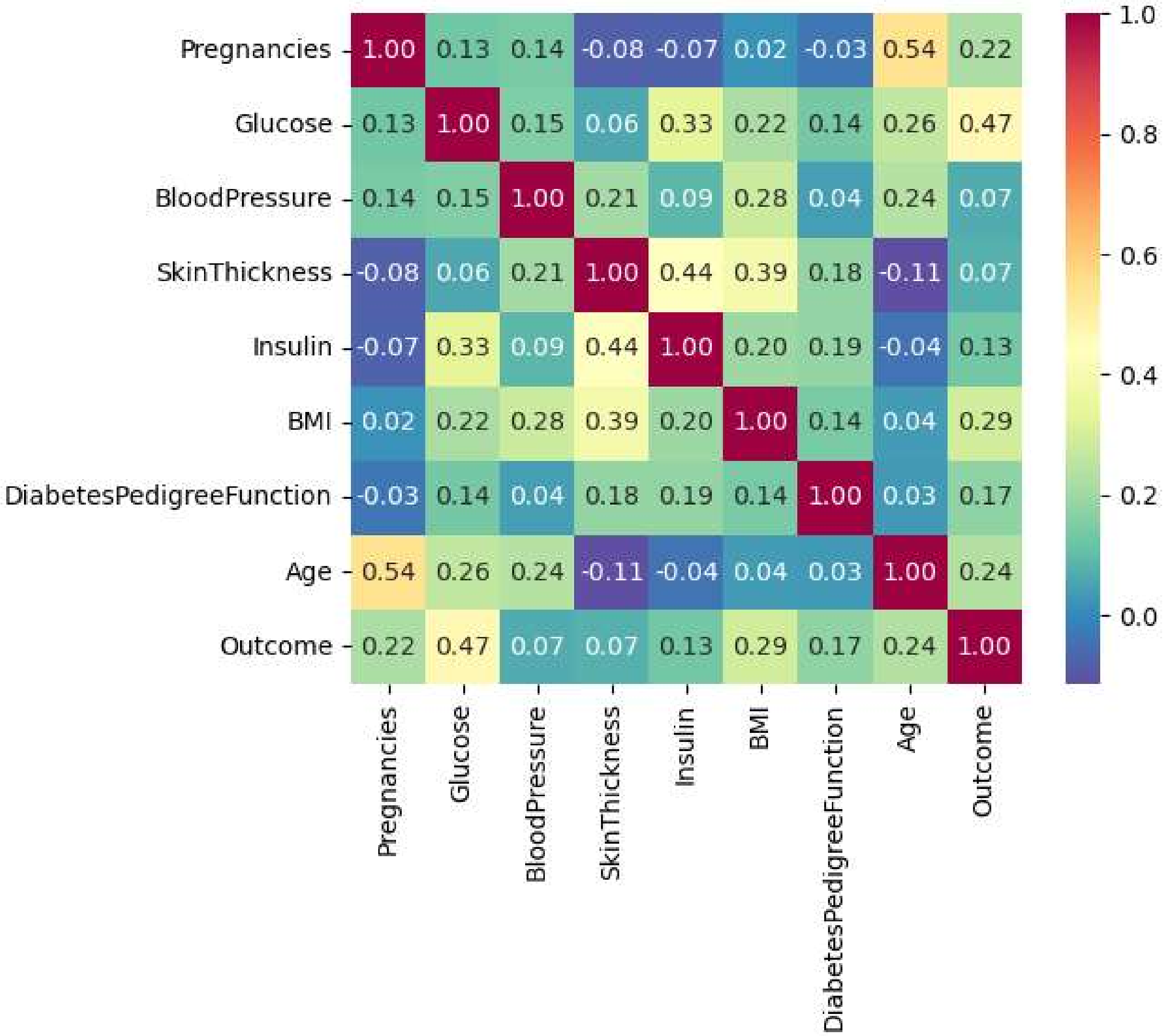
```
(768, 9)
```

In [6]:

```
sns.scatterplot(x = df['Age'] , y = df['Insulin'], palette = "Dark2")
plt.title("Relationships")
plt.show()
```



```
In [7]: corrmat = df.corr()
        hm = sns.heatmap(corrmat,
                         cbar=True,
                         annot=True,
                         square=True,
                         fmt='.2f',
                         annot_kws={'size': 10},
                         yticklabels=df.columns,
                         xticklabels=df.columns,
                         cmap="Spectral_r")
        plt.show()
```

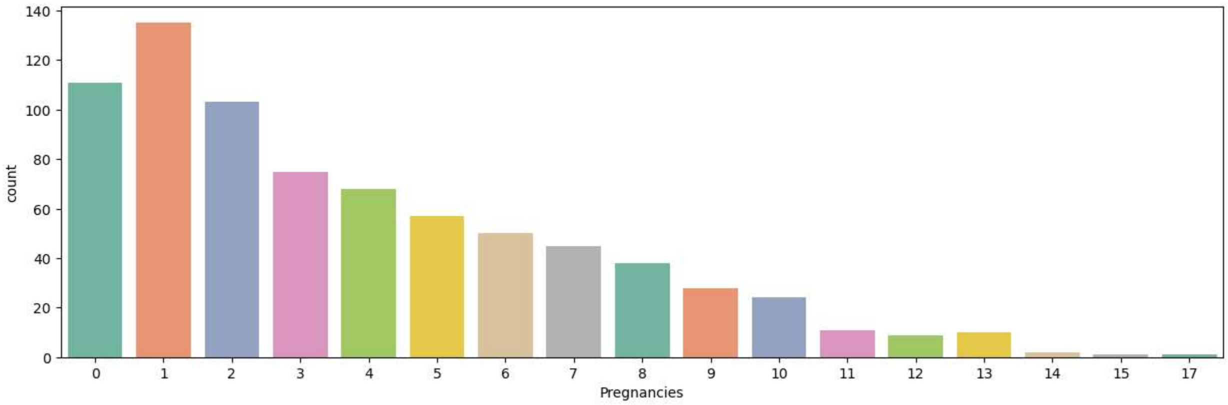


In [8]:

```
plt.figure(figsize = (15,10))

plt.subplot(2,1,1)
sns.countplot(x = 'Pregnancies', palette = 'Set2', data = df)
```

Out[8]:

```
<AxesSubplot:xlabel='Pregnancies', ylabel='count'>
```



In [12]:

```
df.describe()
```

Out[12]:

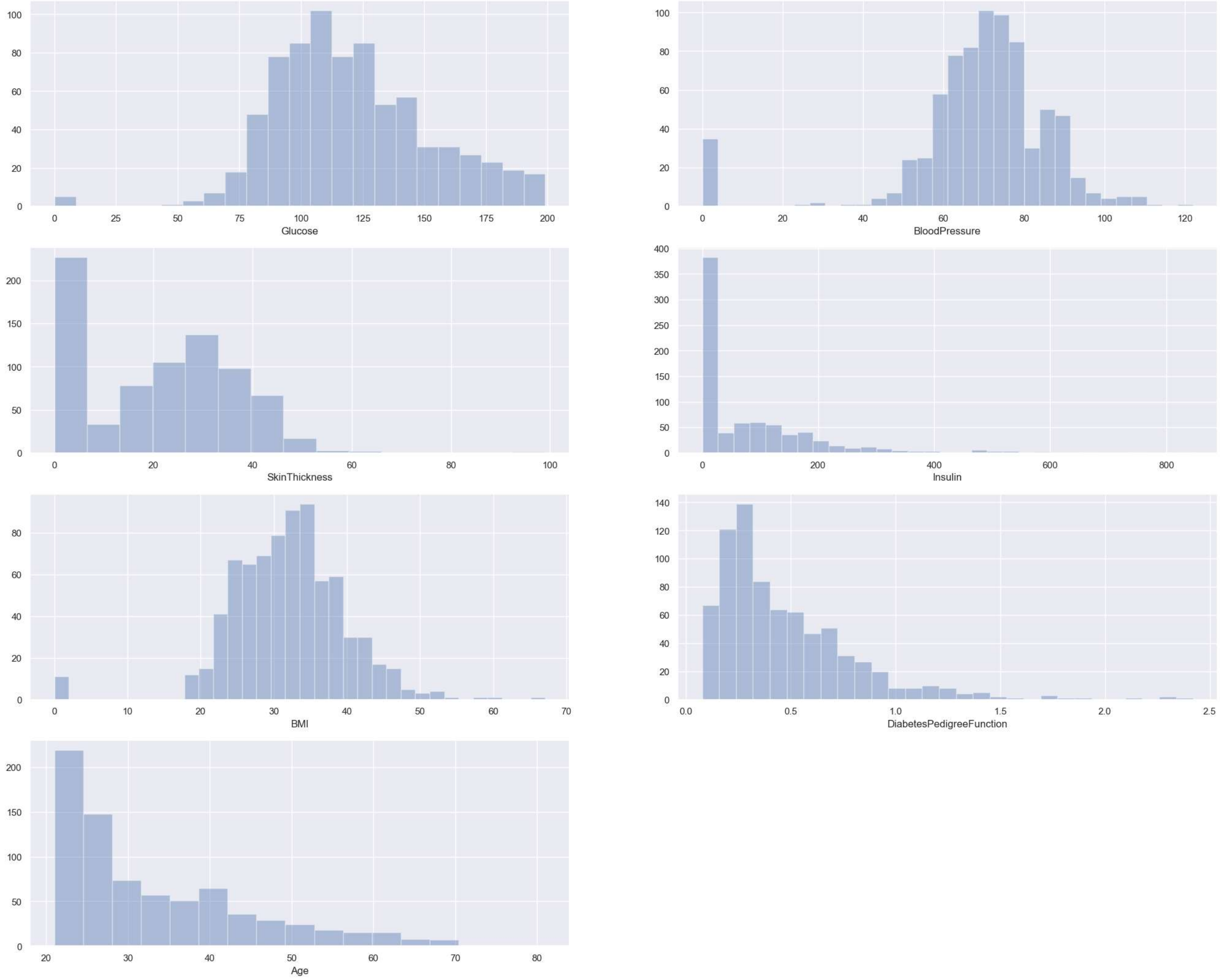
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigr
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	
50%	3.000000	117.000000	72.000000	23.000000	30.500000	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	

In [14]:

```
plt.figure(figsize = (25,20))
sns.set(color_codes = True)

plt.subplot(4,2,1)
sns.distplot(df['Glucose'], kde = False)

plt.subplot(4,2,2)
sns.distplot(df['BloodPressure'], kde = False)

plt.subplot(4,2,3)
sns.distplot(df['SkinThickness'], kde = False)

plt.subplot(4,2,4)
sns.distplot(df['Insulin'], kde = False)

plt.subplot(4,2,5)
sns.distplot(df['BMI'], kde = False)

plt.subplot(4,2,6)
sns.distplot(df['DiabetesPedigreeFunction'], kde = False)

plt.subplot(4,2,7)
sns.distplot(df['Age'], kde = False)
```

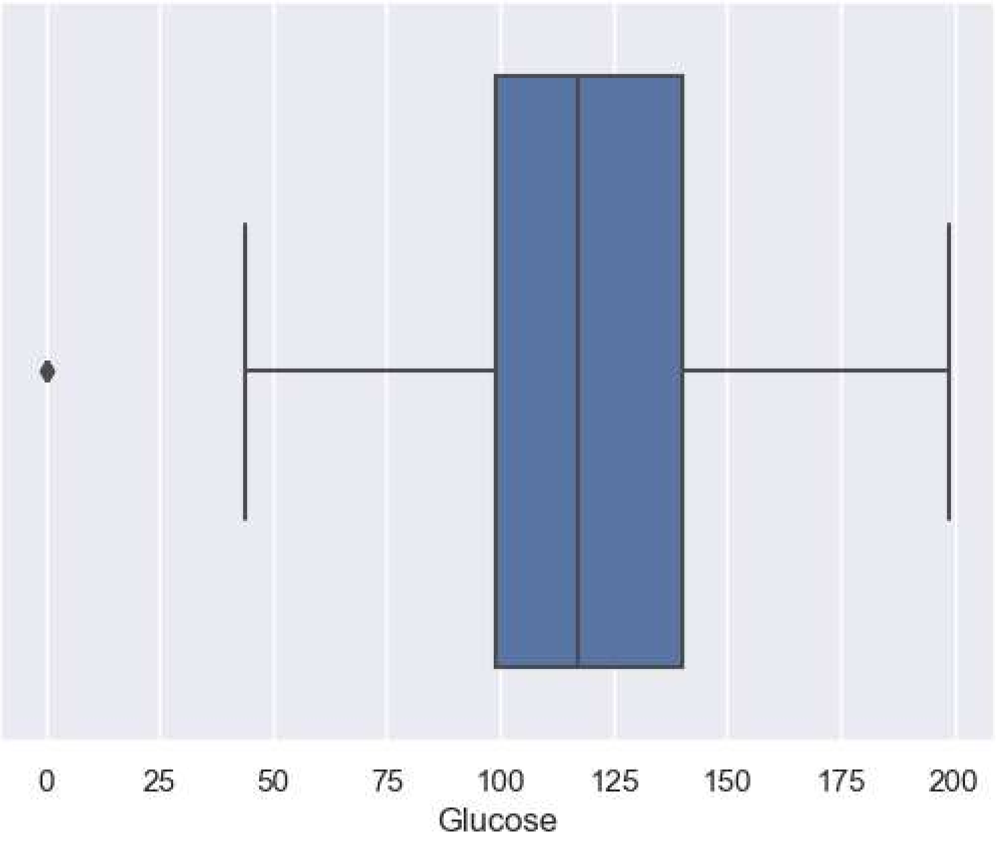
Out[14]:

```
<AxesSubplot:xlabel='Age'>
```



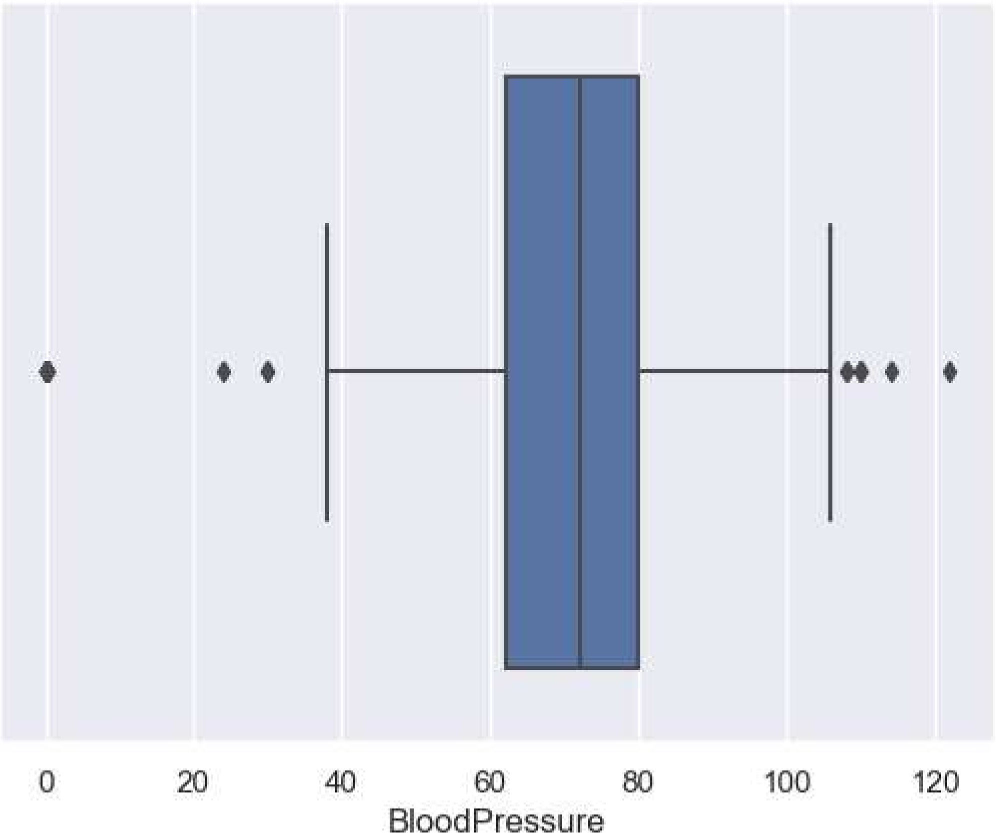
In [15]:
```
sns.boxplot(x=df["Glucose"])
```

Out[15]:
```
<AxesSubplot:xlabel='Glucose'>
```



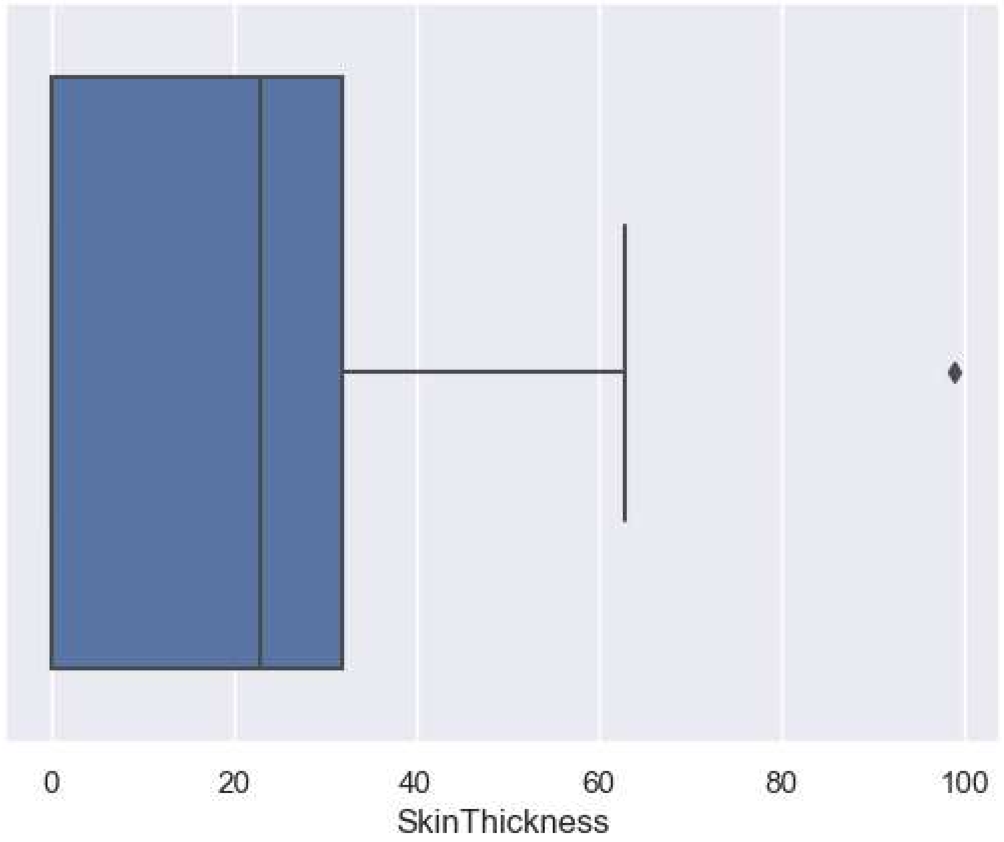
In [16]:
```
sns.boxplot(x=df["BloodPressure"])
```

Out[16]:
```
<AxesSubplot:xlabel='BloodPressure'>
```



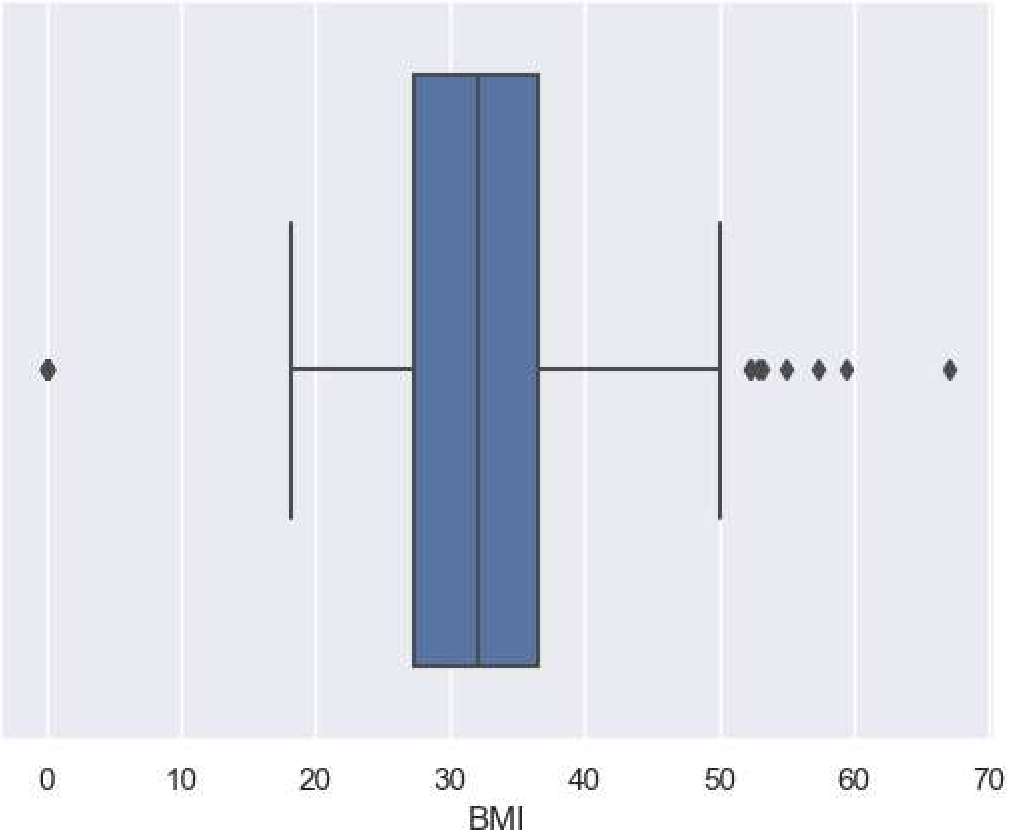
In [17]:
```
sns.boxplot(x=df["SkinThickness"])
```

Out[17]:
```
<AxesSubplot:xlabel='SkinThickness'>
```



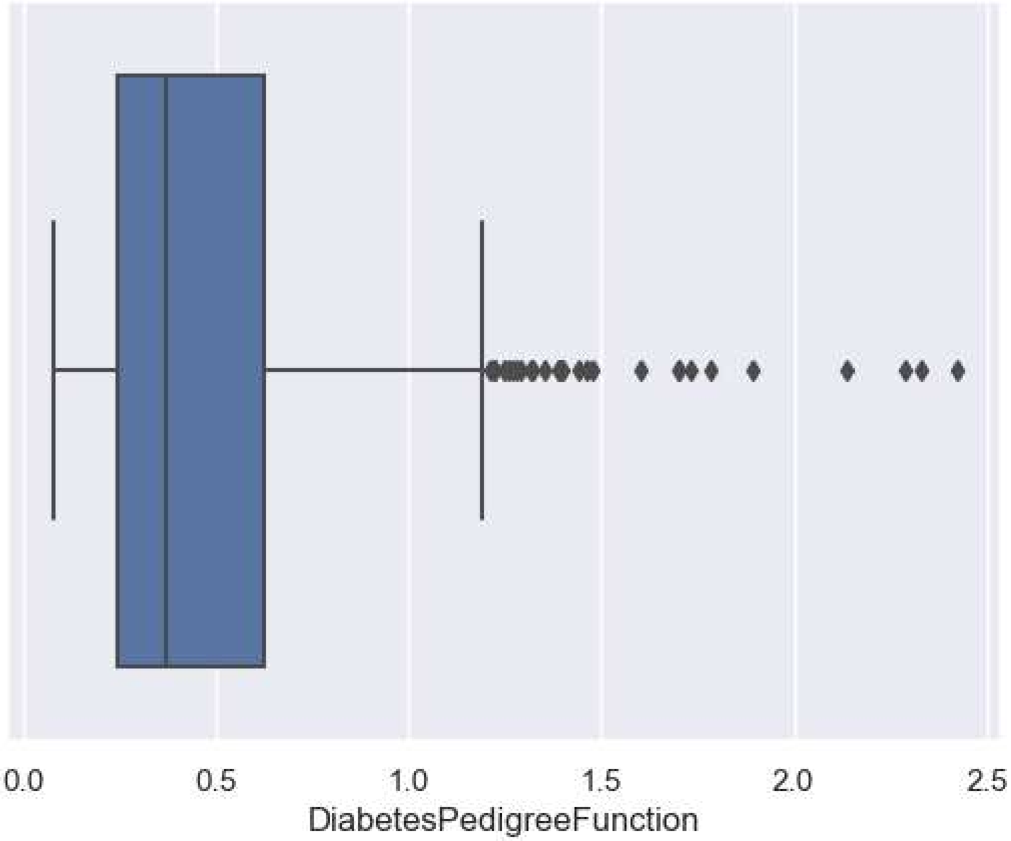
In [18]:
```
sns.boxplot(x=df["BMI"])
```

Out[18]:
```
<AxesSubplot:xlabel='BMI'>
```



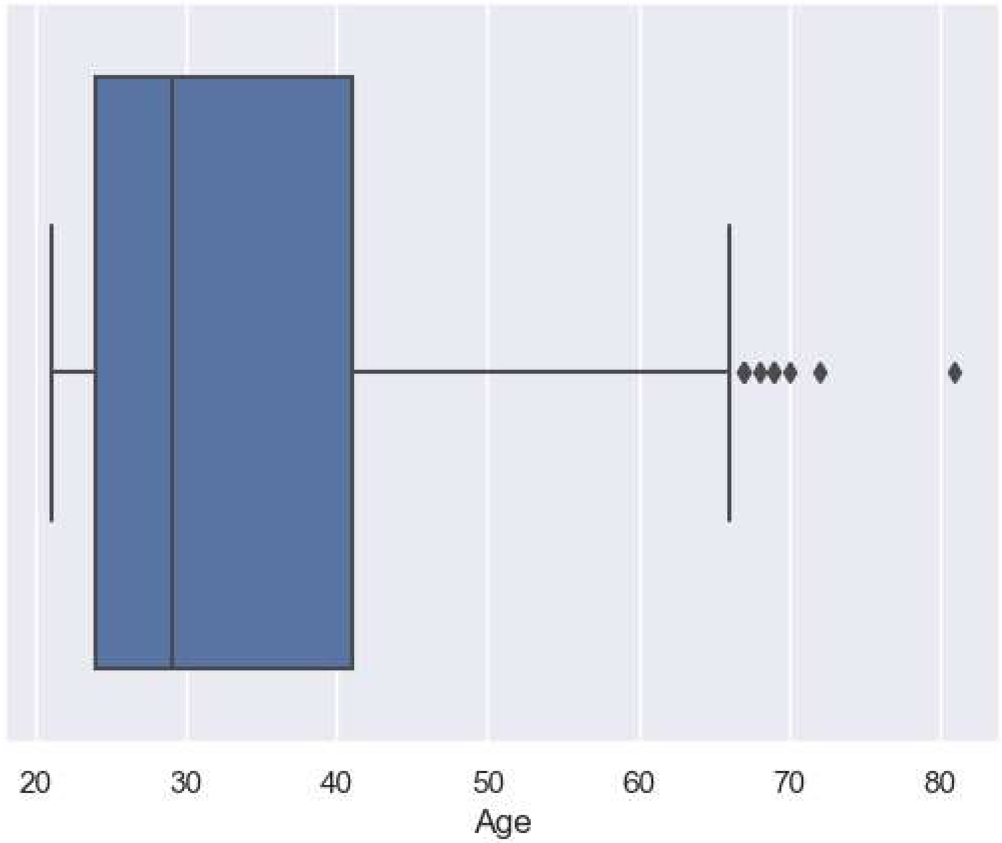
In [19]:
```
sns.boxplot(x=df["DiabetesPedigreeFunction"])
```

Out[19]:
```
<AxesSubplot:xlabel='DiabetesPedigreeFunction'>
```



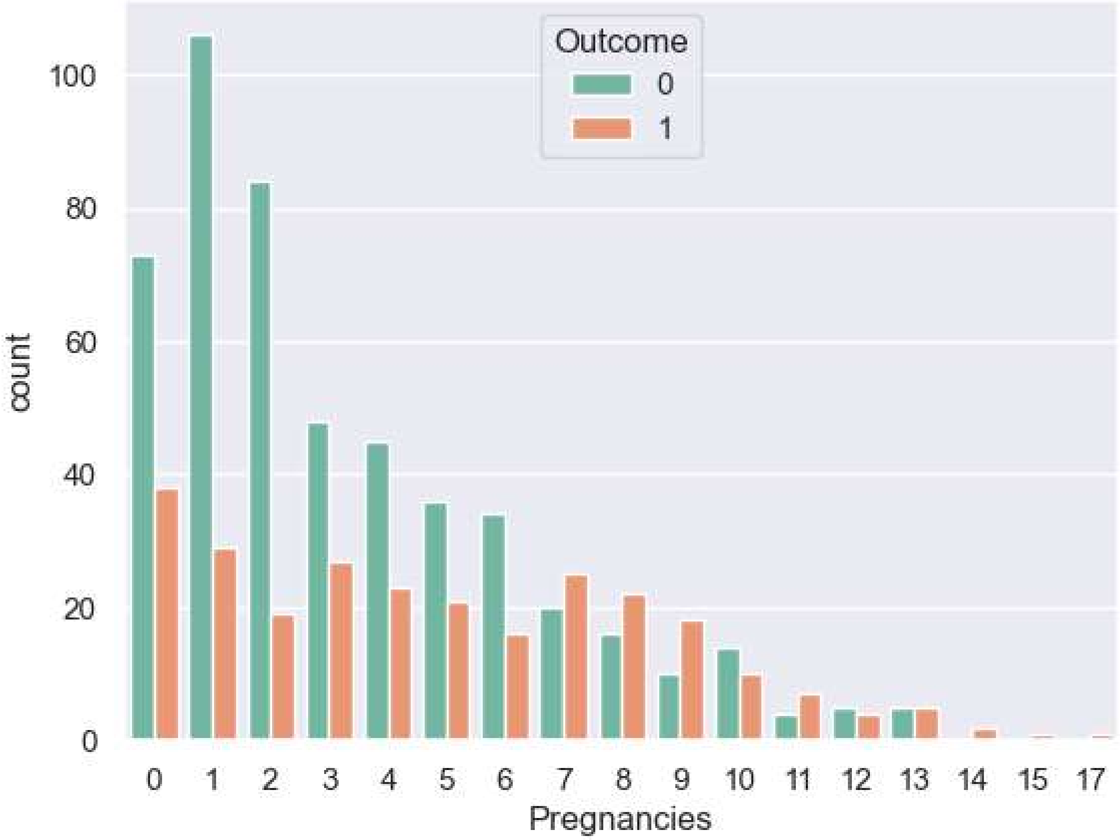
In [20]:
```
sns.boxplot(x=df["Age"])
```

Out[20]:
```
<AxesSubplot:xlabel='Age'>
```



In [21]:
```
sns.countplot(x = 'Pregnancies', hue = 'Outcome', palette = 'Set2', data = df)
```

Out[21]:
```
<AxesSubplot:xlabel='Pregnancies', ylabel='count'>
```

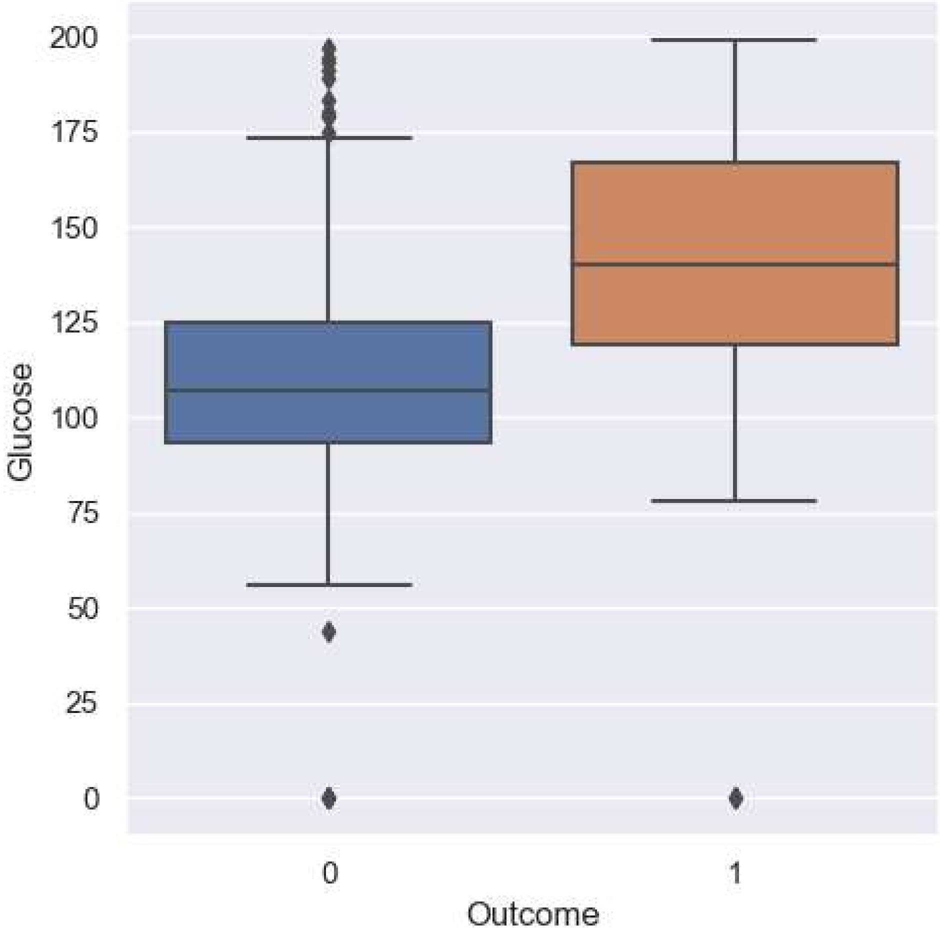


In [22]:

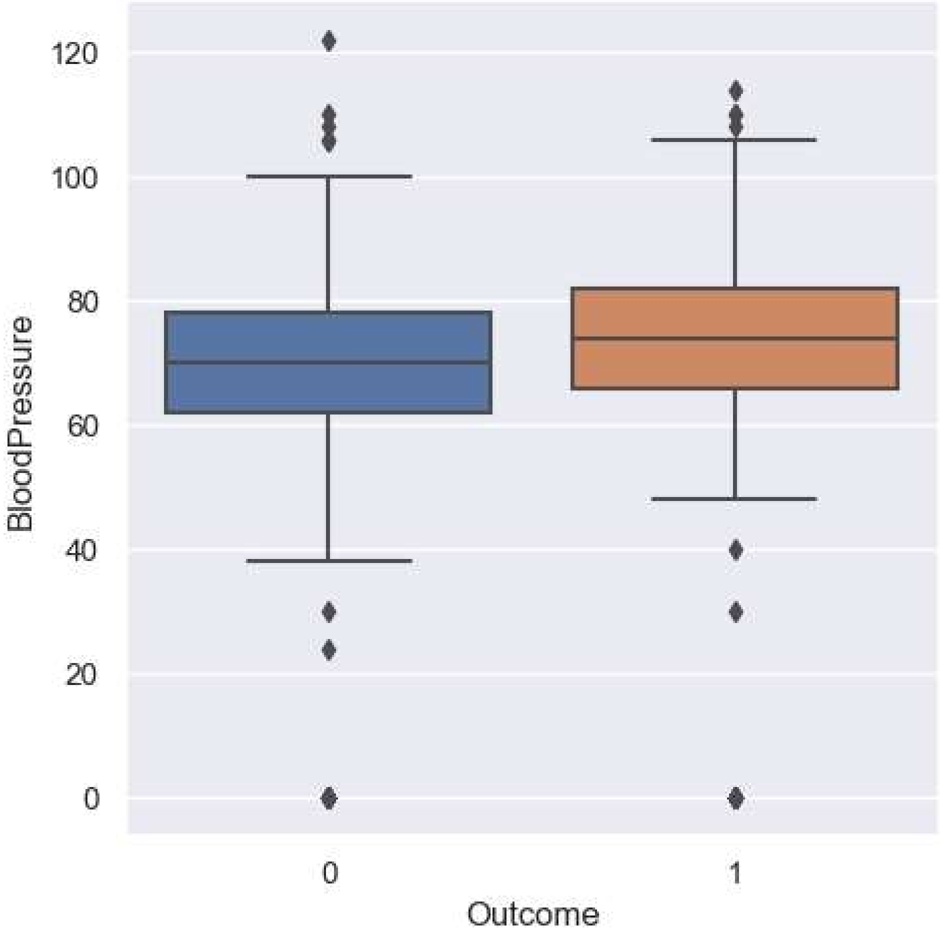
```
sns.catplot(x = 'Outcome', y="Glucose", kind="box", data = df)
```

Out[22]:

```
<seaborn.axisgrid.FacetGrid at 0x2a425ff3100>
```

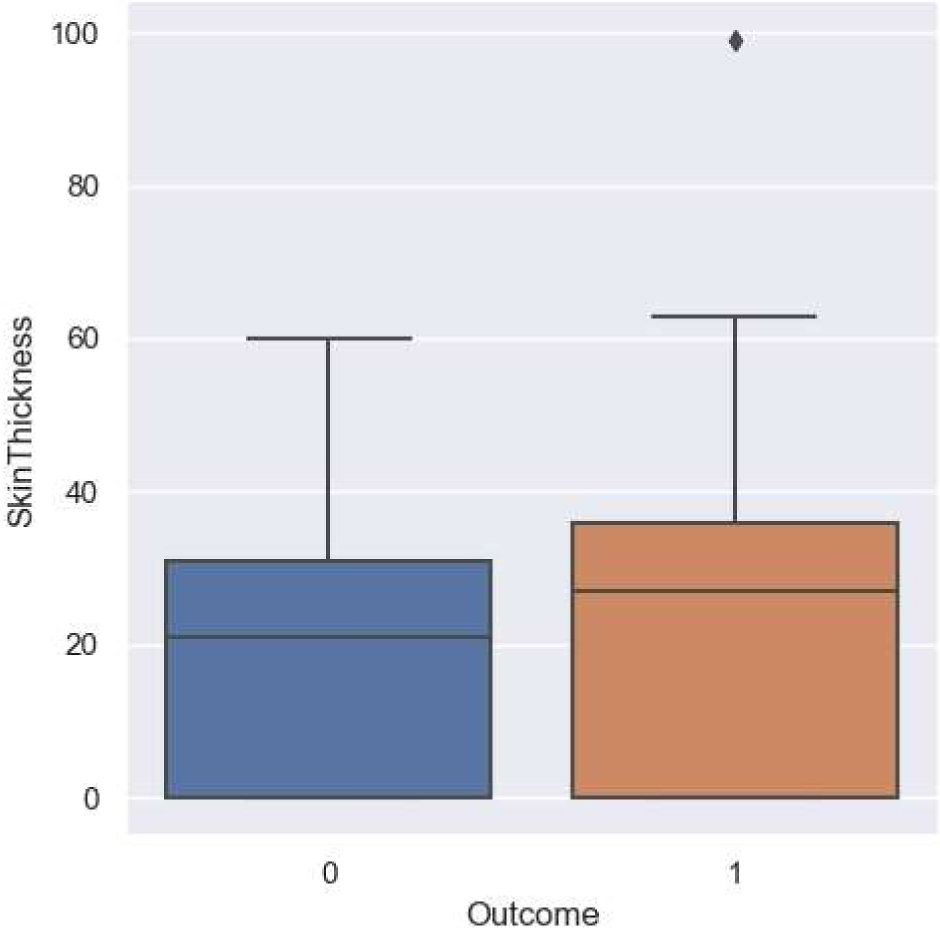


```
In [23]: sns.catplot(x = 'Outcome', y="BloodPressure", kind="box", data = df)

Out[23]: <seaborn.axisgrid.FacetGrid at 0x2a427d9b790>
```



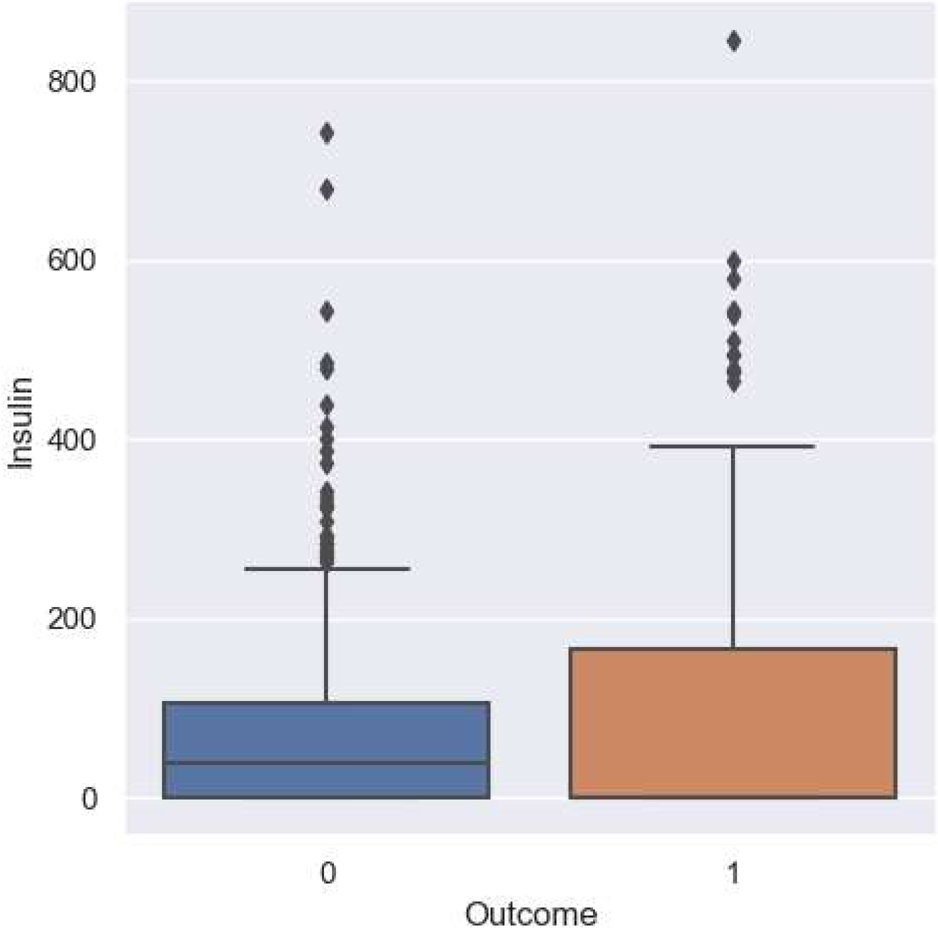
In [24]:
```
sns.catplot(x = 'Outcome', y="SkinThickness", kind="box", data = df)
```

Out[24]:
```
<seaborn.axisgrid.FacetGrid at 0x2a426083700>
```

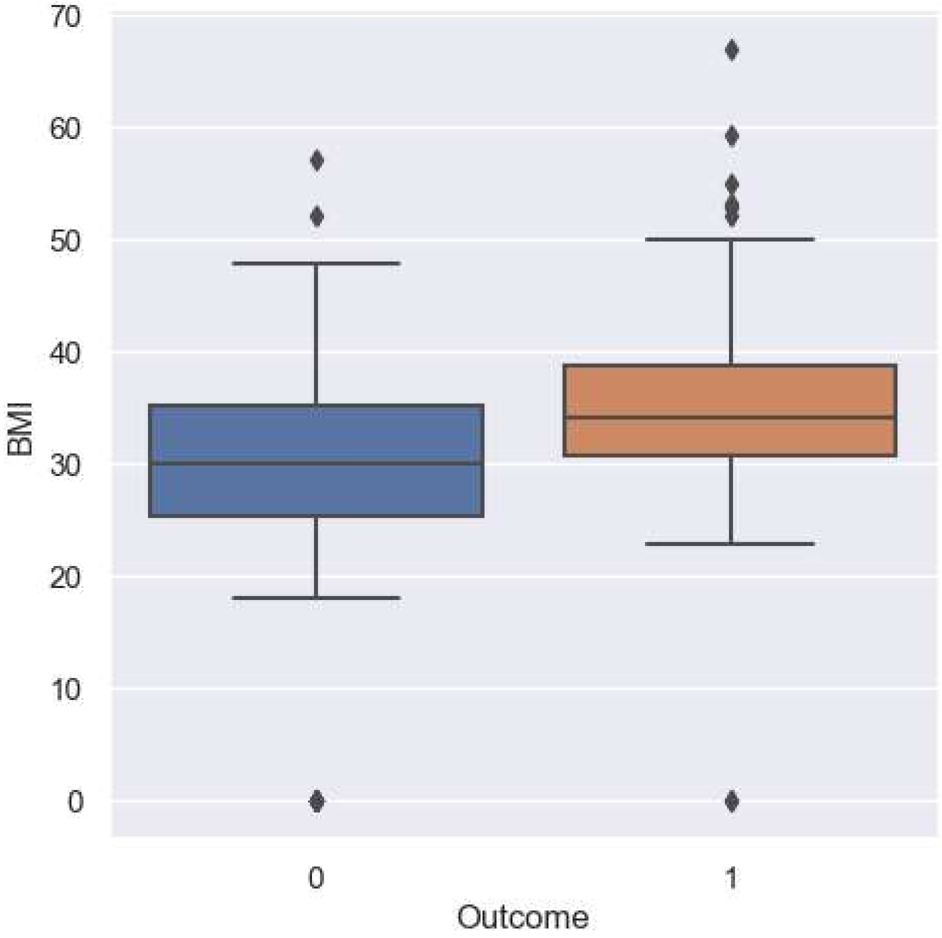


In [25]:
```
sns.catplot(x = 'Outcome', y="Insulin", kind="box", data = df)
```

Out[25]:
```
<seaborn.axisgrid.FacetGrid at 0x2a426157a00>
```

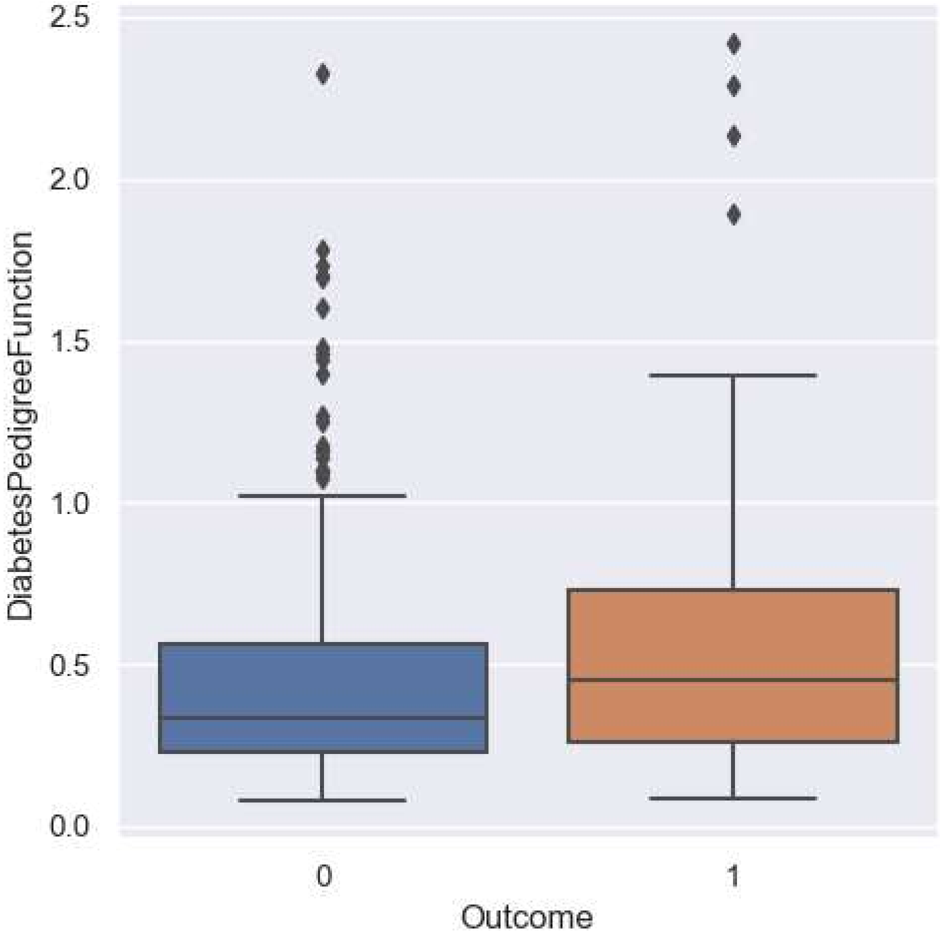


```
In [26]: sns.catplot(x = 'Outcome', y="BMI", kind="box", data = df)

Out[26]: <seaborn.axisgrid.FacetGrid at 0x2a428e0ceb0>
```



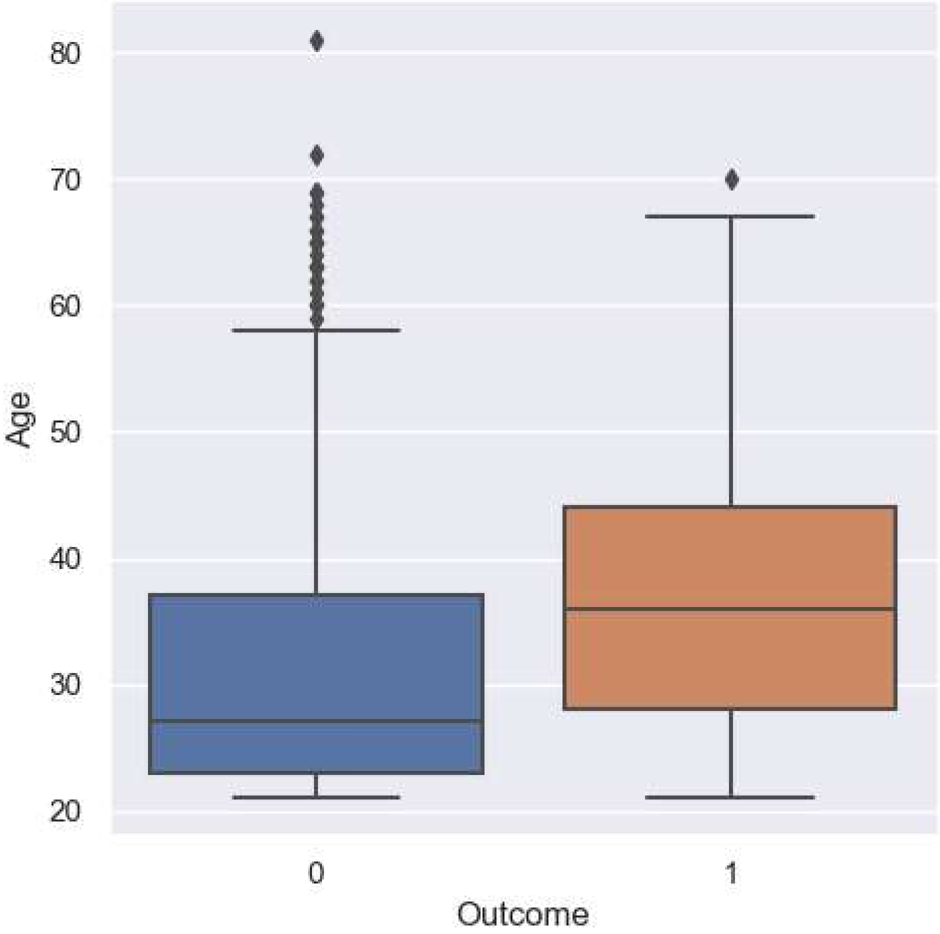
In [27]:
```
sns.catplot(x = 'Outcome', y="DiabetesPedigreeFunction", kind="box", data = df)
```

Out[27]:
```
<seaborn.axisgrid.FacetGrid at 0x2a425ff3880>
```



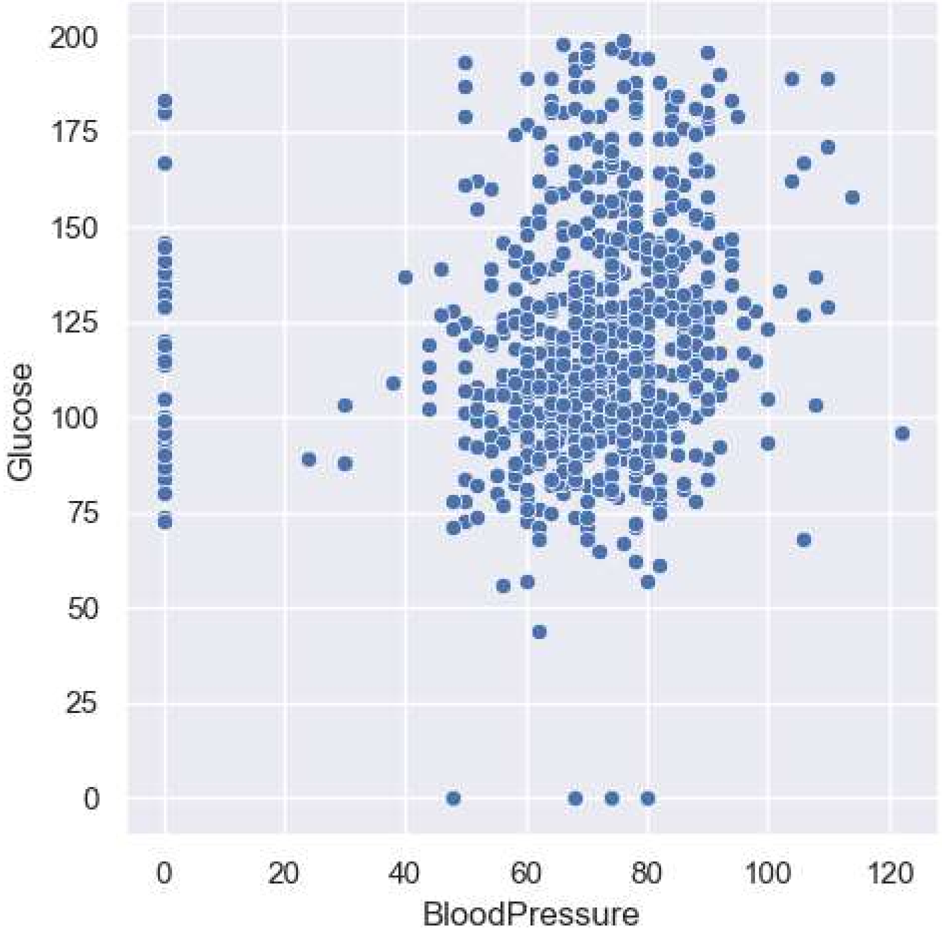
In [28]:
```
sns.catplot(x = 'Outcome', y="Age", kind="box", data = df)
```

Out[28]:
```
<seaborn.axisgrid.FacetGrid at 0x2a428ef0dc0>
```



```
In [29]: sns.relplot(x='BloodPressure', y = 'Glucose' , data = df)
```

```
Out[29]: <seaborn.axisgrid.FacetGrid at 0x2a428df4970>
```



```
In [48]: X = df.drop('Outcome', axis = 1)
```

```
In [49]: X = X.values
```

```
In [50]: y = df['Outcome']
```

```
In [51]: columns = df.drop('Outcome', axis = 1).columns
```

```
In [52]: from sklearn.feature_selection import SelectKBest
         from sklearn.feature_selection import chi2

         features = X
         target = y

         best_features = SelectKBest(score_func = chi2,k = 'all')
         fit = best_features.fit(features,target)

         featureScores = pd.DataFrame(data = fit.scores_,index = list(columns),columns = ['Chi
```

```
In [53]: featureScores.sort_values(by = 'Chi Squared Score', ascending = False)
```

Out[53]:

	Chi Squared Score
Insulin	2175.565273
Glucose	1411.887041
Age	181.303689
BMI	127.669343
Pregnancies	111.519691
SkinThickness	53.108040
BloodPressure	17.605373
DiabetesPedigreeFunction	5.392682

In [ ]: