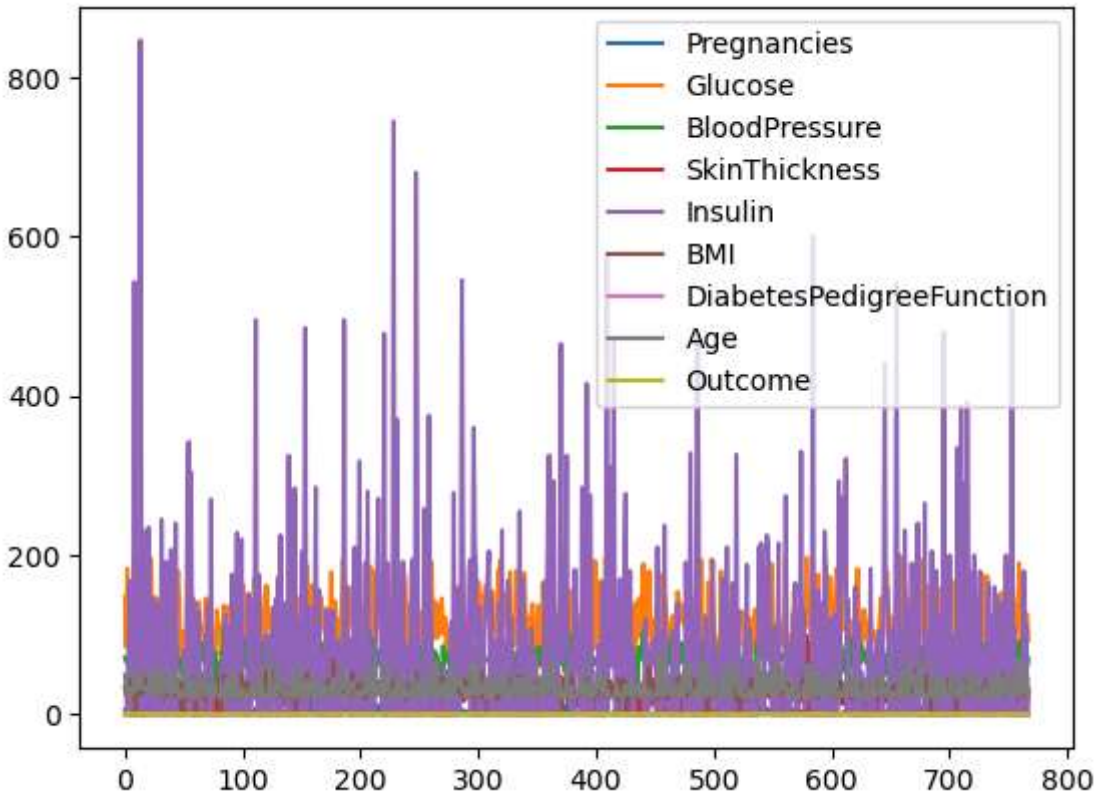
In [1]:
```
 import pandas as pd
data=pd.read_csv(r"C:\Users\DELL\Downloads\archive\diabetes.csv")
print(data)
data.plot()
data.head()
```

```
     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]
```

Out[1]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33



In [5]:
```
print(data.isnull().sum())
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

In [6]:
```
db.fillna(db.mean(),inplce=True)
print(db.isnull().sum())
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[6], line 1
----> 1 db.fillna(db.mean(),inplce=True)
      2 print(db.isnull().sum())

NameError: name 'db' is not defined
```

In [11]:
```
import numpy as np
import matplotlib.pyplot as plt
plt.figure(figsize=(10,6))
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```

In [12]:

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[12], line 1
----> 1 z_scores = stats.zscore(data.select_dtypes(include=['float64', 'int64']))
      2 outliers = (abs(z_scores) > 3).any(axis=1)
      3 print(data[outliers])

NameError: name 'stats' is not defined
```

In [13]:
```
pip install scipy matplotlib seaborn
```

```
Requirement already satisfied: scipy in c:\users\dell\anaconda3\lib\site-packages (1.
10.0)
Requirement already satisfied: matplotlib in c:\users\dell\anaconda3\lib\site-package
s (3.7.0)
Requirement already satisfied: seaborn in c:\users\dell\anaconda3\lib\site-packages
(0.12.2)
Requirement already satisfied: numpy<1.27.0,>=1.19.5 in c:\users\dell\anaconda3\lib\s
ite-packages (from scipy) (1.23.5)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\dell\anaconda3\lib\site-p
ackages (from matplotlib) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\dell\anaconda3\lib\site-packa
ges (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\dell\anaconda3\lib\site-
packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\dell\anaconda3\lib\site-
packages (from matplotlib) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\dell\anaconda3\lib\site-pa
ckages (from matplotlib) (22.0)
Requirement already satisfied: pillow>=6.2.0 in c:\users\dell\anaconda3\lib\site-pack
ages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\dell\anaconda3\lib\site-p
ackages (from matplotlib) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\dell\anaconda3\lib\si
te-packages (from matplotlib) (2.8.2)
Requirement already satisfied: pandas>=0.25 in c:\users\dell\anaconda3\lib\site-packa
ges (from seaborn) (1.5.3)
Requirement already satisfied: pytz>=2020.1 in c:\users\dell\anaconda3\lib\site-packa
ges (from pandas>=0.25->seaborn) (2022.7)
Requirement already satisfied: six>=1.5 in c:\users\dell\anaconda3\lib\site-packages
(from python-dateutil>=2.7->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [14]:
```
z_scores = stats.zscore(data.select_dtypes(include=['float64', 'int64']))
outliers = (abs(z_scores) > 3).any(axis=1)
print(data[outliers])
```

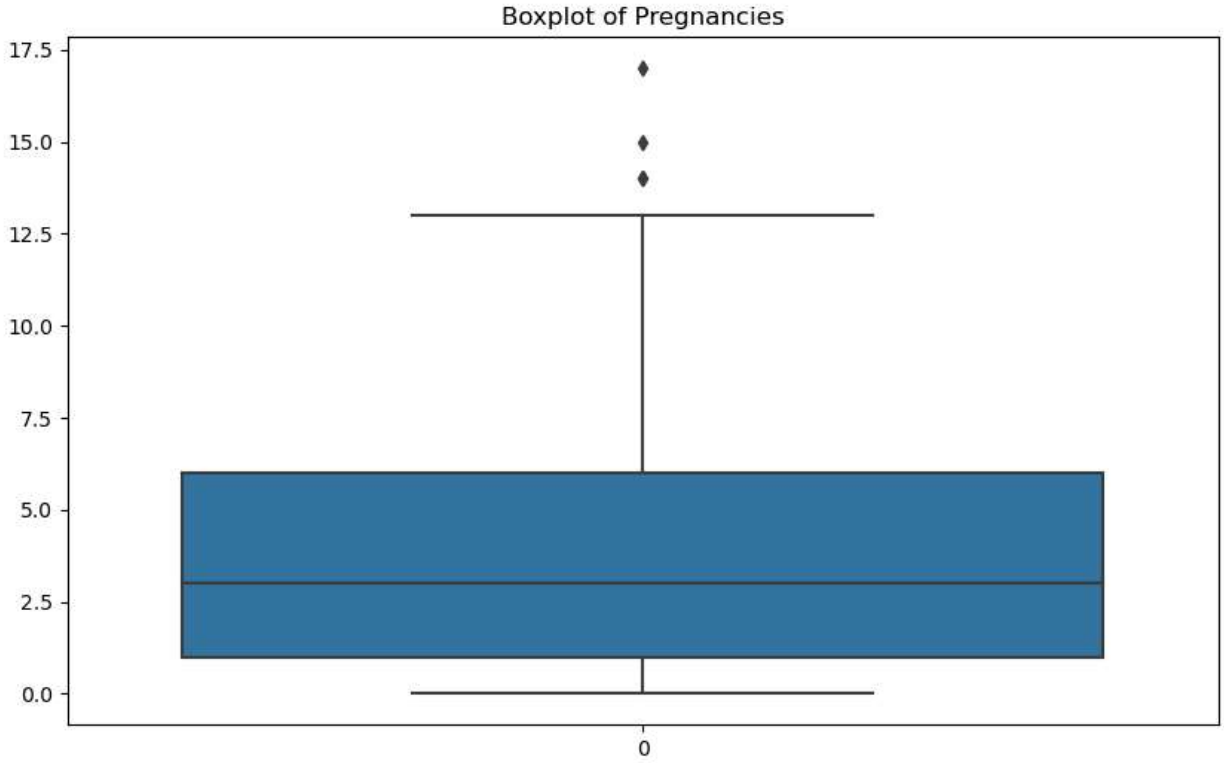
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[14], line 1
----> 1 z_scores = stats.zscore(data.select_dtypes(include=['float64', 'int64']))
      2 outliers = (abs(z_scores) > 3).any(axis=1)
      3 print(data[outliers])

NameError: name 'stats' is not defined
```

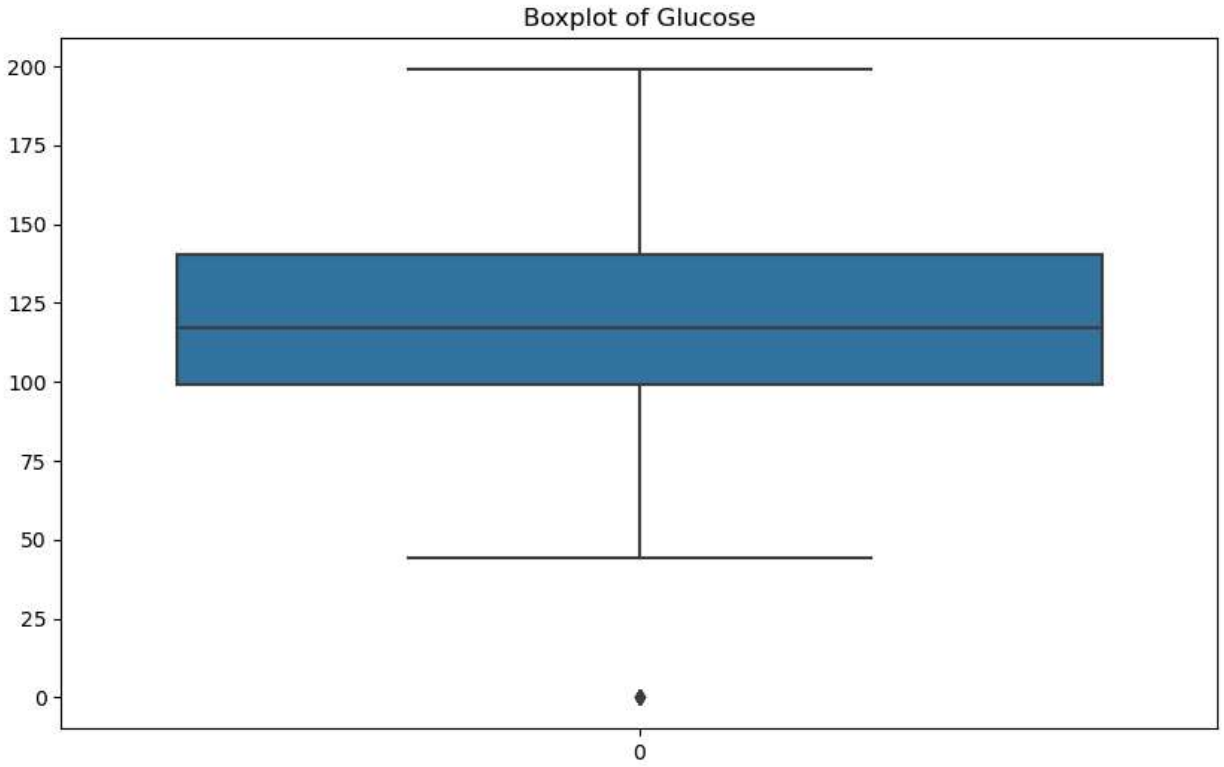
In [15]:
```
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns
```

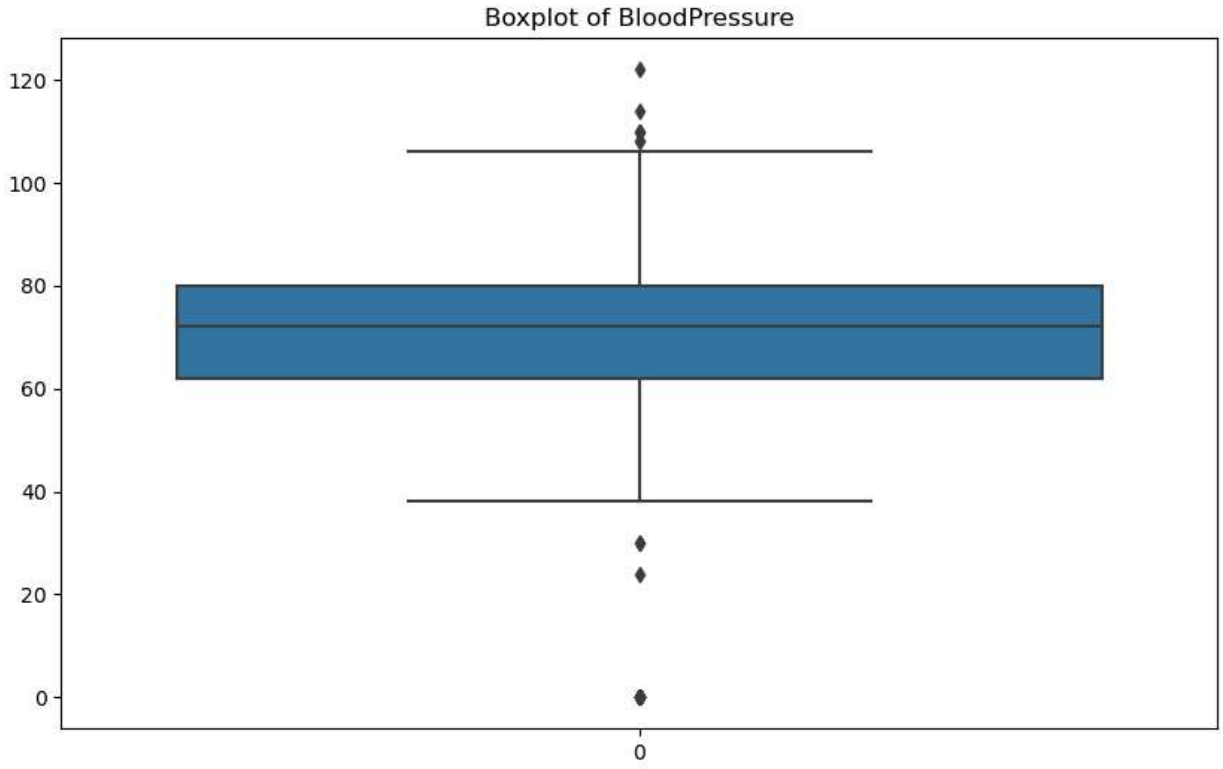
```python
# Load the dataset
data = pd.read_csv(r"C:\Users\DELL\Downloads\archive\diabetes.csv")

# Check for null values
print(data.isnull().sum())

# Compute Z-scores
z_scores = stats.zscore(data.select_dtypes(include=['float64', 'int64']))

# Identify outliers based on Z-scores
outliers = (abs(z_scores) > 3).any(axis=1)
print(data[outliers])

# Calculate IQR and identify outliers based on IQR
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
outlier_condition = ((data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR)))
print(data[outlier_condition.any(axis=1)])

# Plot box plots for each numeric column
for column in data.select_dtypes(include=['float64', 'int64']).columns:
    plt.figure(figsize=(10, 6))
    sns.boxplot(data[column])
    plt.title(f'Boxplot of {column}')
    plt.show()
```

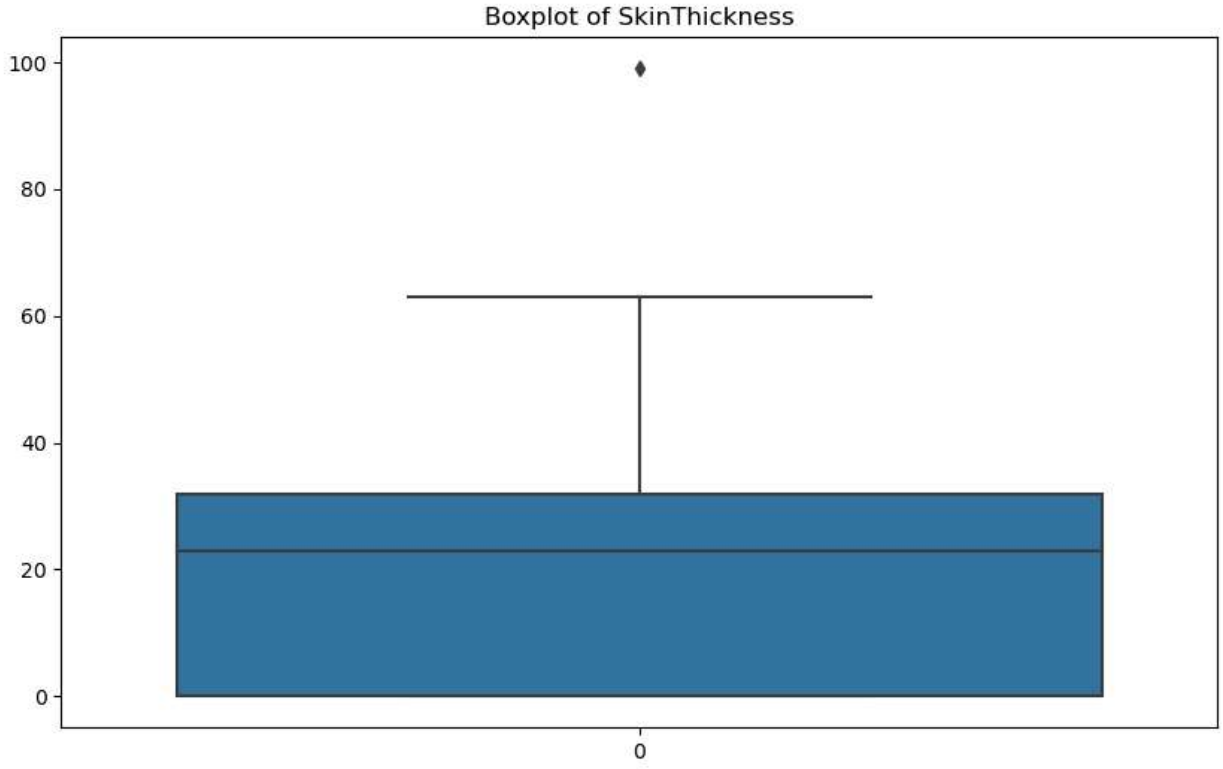
```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
4              0      137             40             35      168  43.1
7             10      115              0              0        0  35.3
8              2      197             70             45      543  30.5
9              8      125             96              0        0   0.0
13             1      189             60             23      846  30.1
..           ...      ...            ...            ...      ...   ...
695            7      142             90             24      480  30.4
697            0       99              0              0        0  25.0
703            2      129              0              0        0  38.5
706           10      115              0              0        0   0.0
753            0      181             88             44      510  43.3

     DiabetesPedigreeFunction  Age  Outcome
4                       2.288   33        1
7                       0.134   29        0
8                       0.158   53        1
9                       0.232   54        1
13                      0.398   59        1
..                        ...  ...      ...
695                     0.128   43        1
697                     0.253   22        0
703                     0.304   41        0
706                     0.261   30        1
753                     0.222   26        1

[80 rows x 9 columns]
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
4              0      137             40             35      168  43.1
7             10      115              0              0        0  35.3
8              2      197             70             45      543  30.5
9              8      125             96              0        0   0.0
12            10      139             80              0        0  27.1
..           ...      ...            ...            ...      ...   ...
706           10      115              0              0        0   0.0
707            2      127             46             21      335  34.4
710            3      158             64             13      387  31.2
715            7      187             50             33      392  33.9
753            0      181             88             44      510  43.3

     DiabetesPedigreeFunction  Age  Outcome
4                       2.288   33        1
7                       0.134   29        0
8                       0.158   53        1
9                       0.232   54        1
12                      1.441   57        0
..                        ...  ...      ...
706                     0.261   30        1
707                     0.176   22        0
710                     0.295   24        0
```

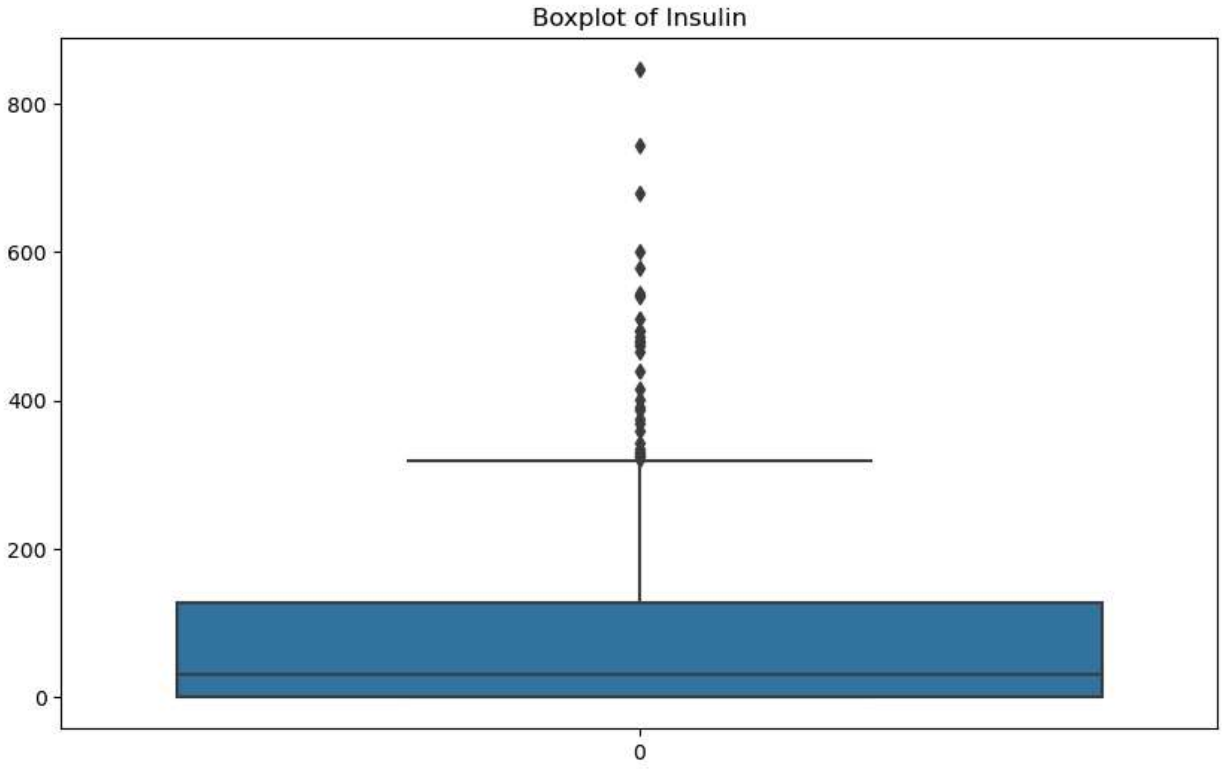
```
715                   0.826   34        1
753                   0.222   26        1

[129 rows x 9 columns]
```

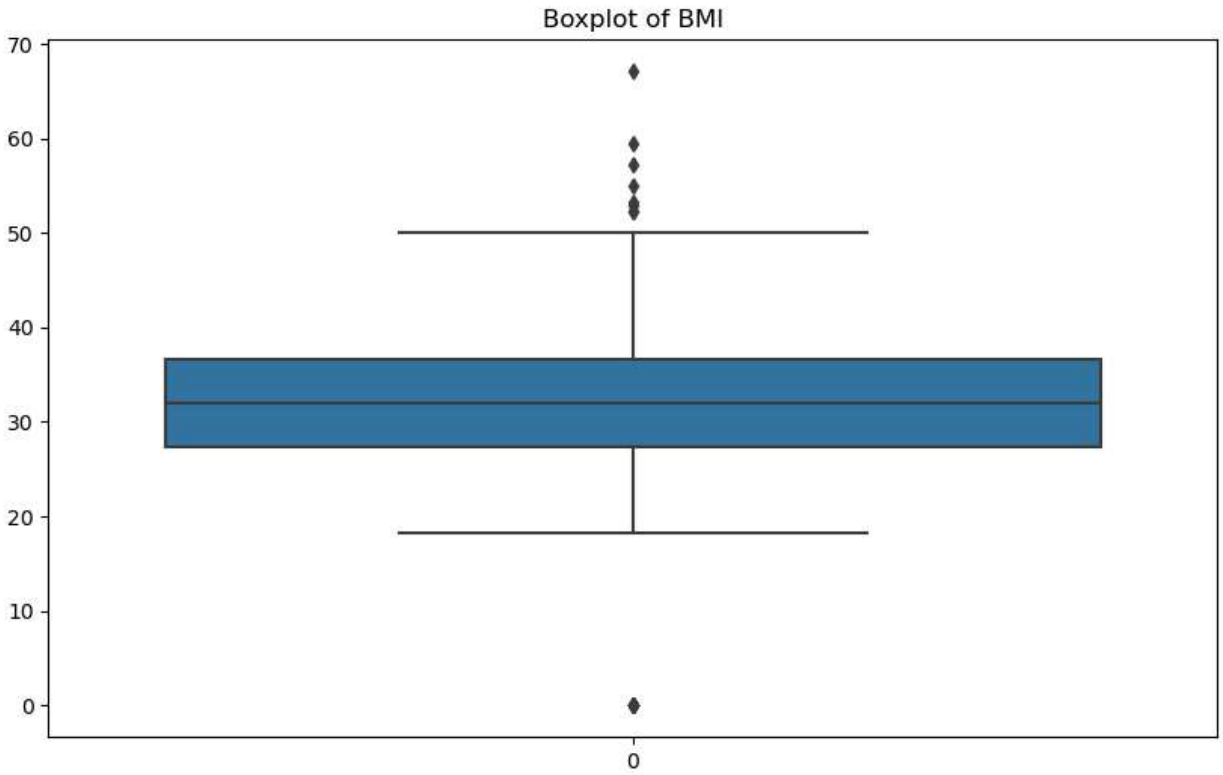


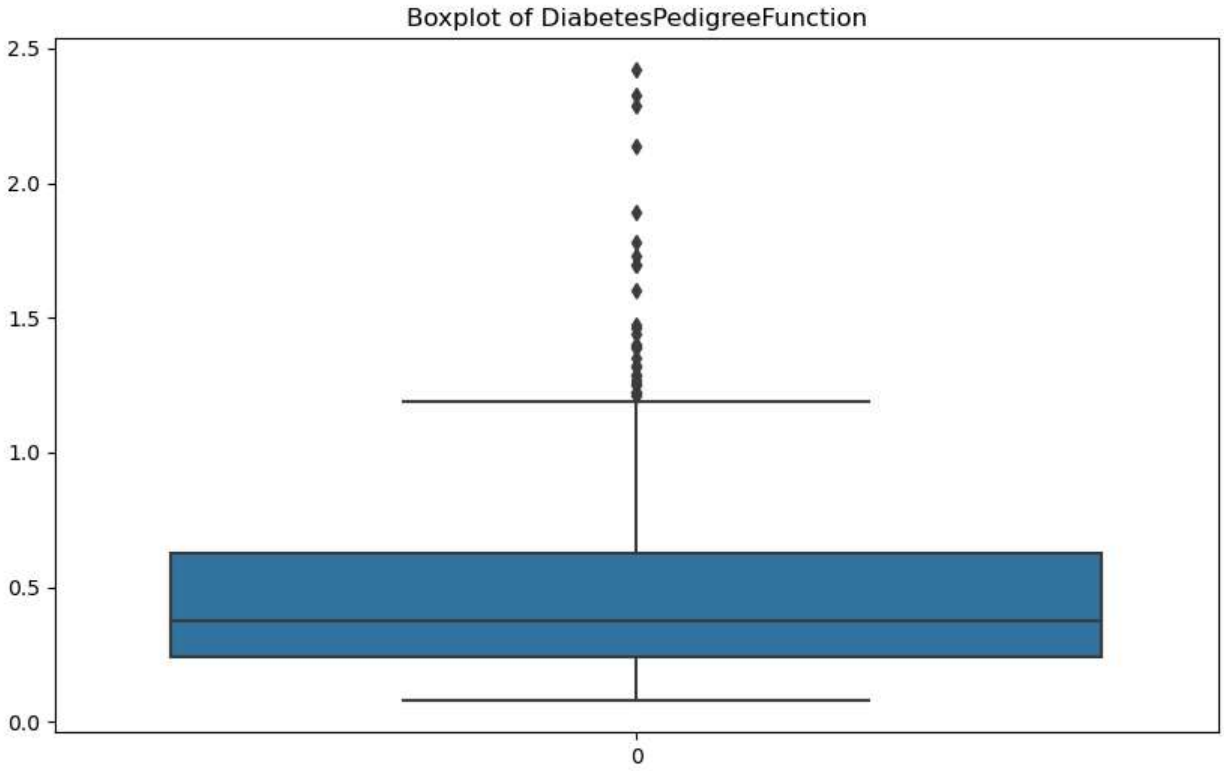


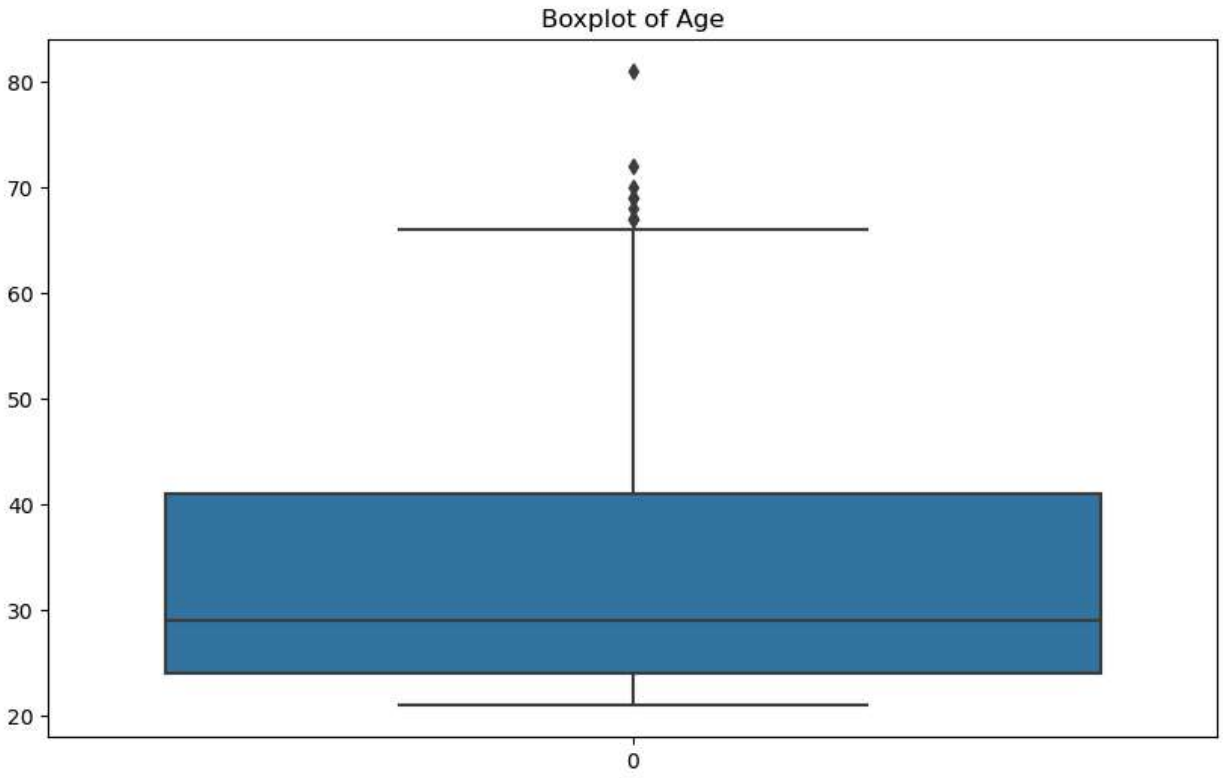

Boxplot of BloodPressure
120
100
80
60
40
20
0
0

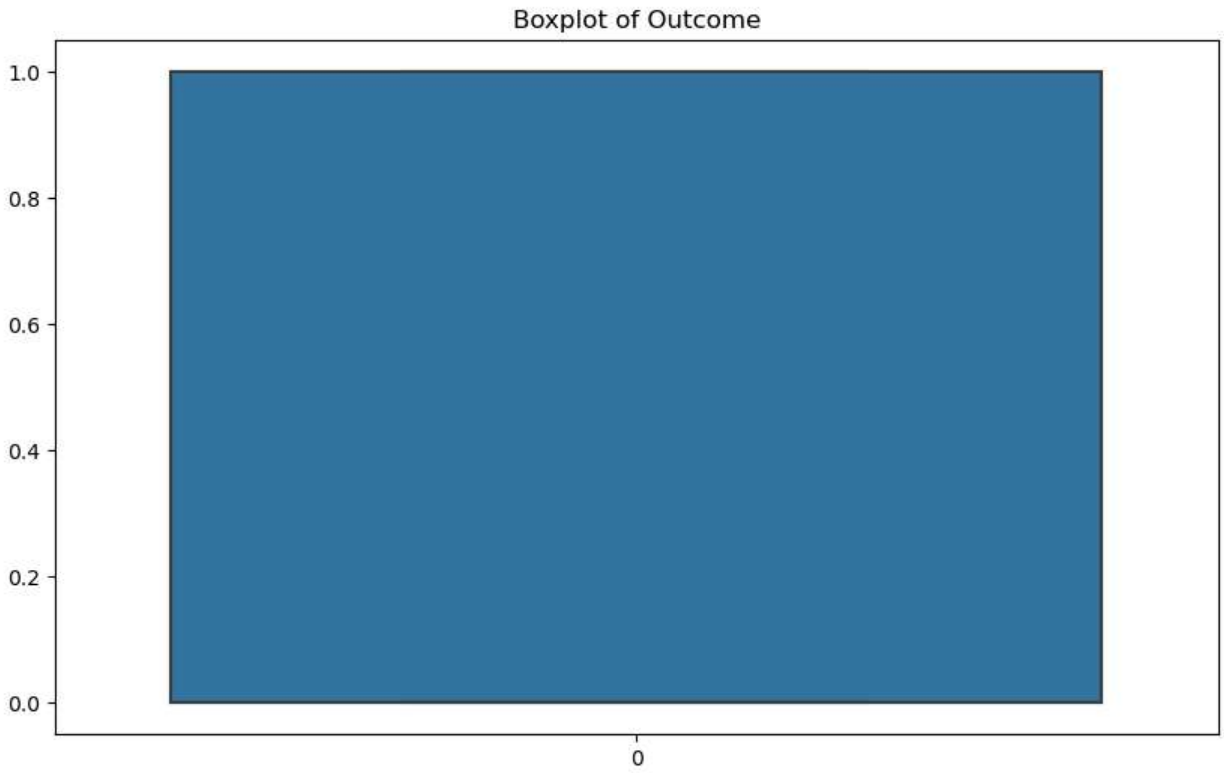

Boxplot of SkinThickness
100
80
60
40
20
0
0


Boxplot of Insulin
800
600
400
200
0
0


Boxplot of BMI
70
60
50
40
30
20
10
0
0


Boxplot of DiabetesPedigreeFunction
2.5
2.0
1.5
1.0
0.5
0.0
0

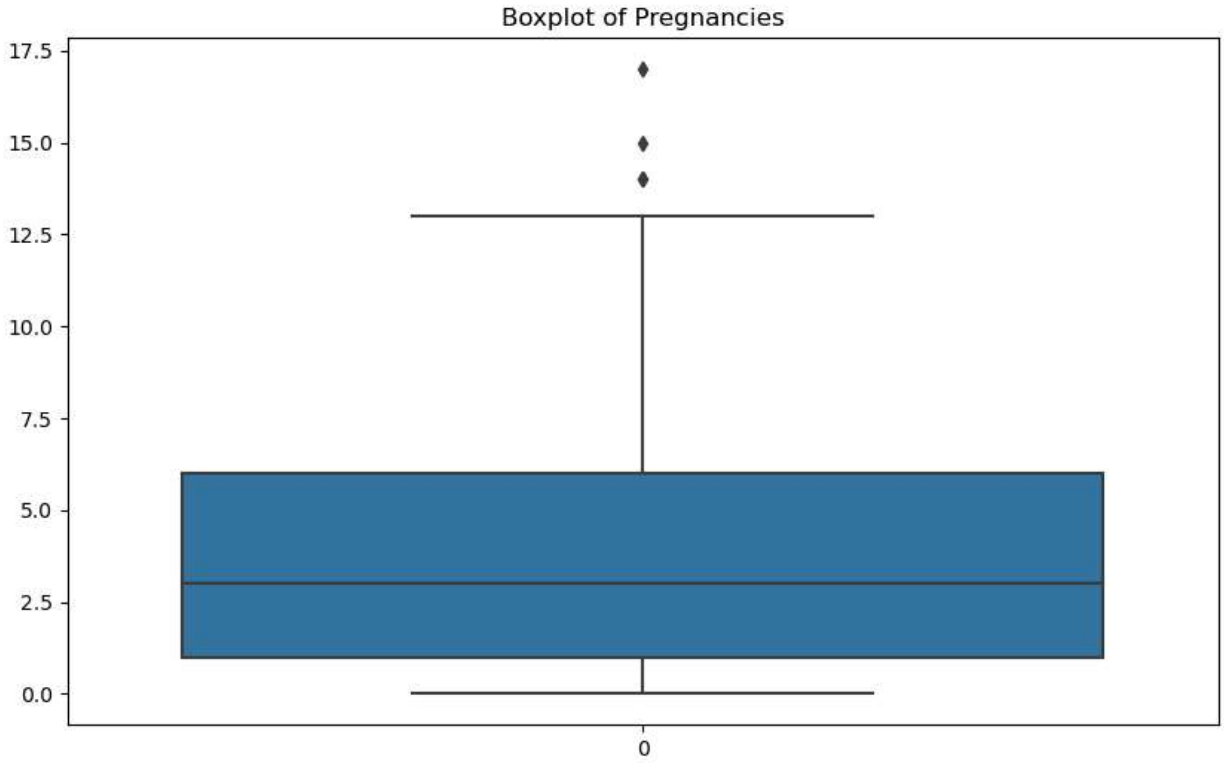

Boxplot of Age
80
70
60
50
40
30
20
0

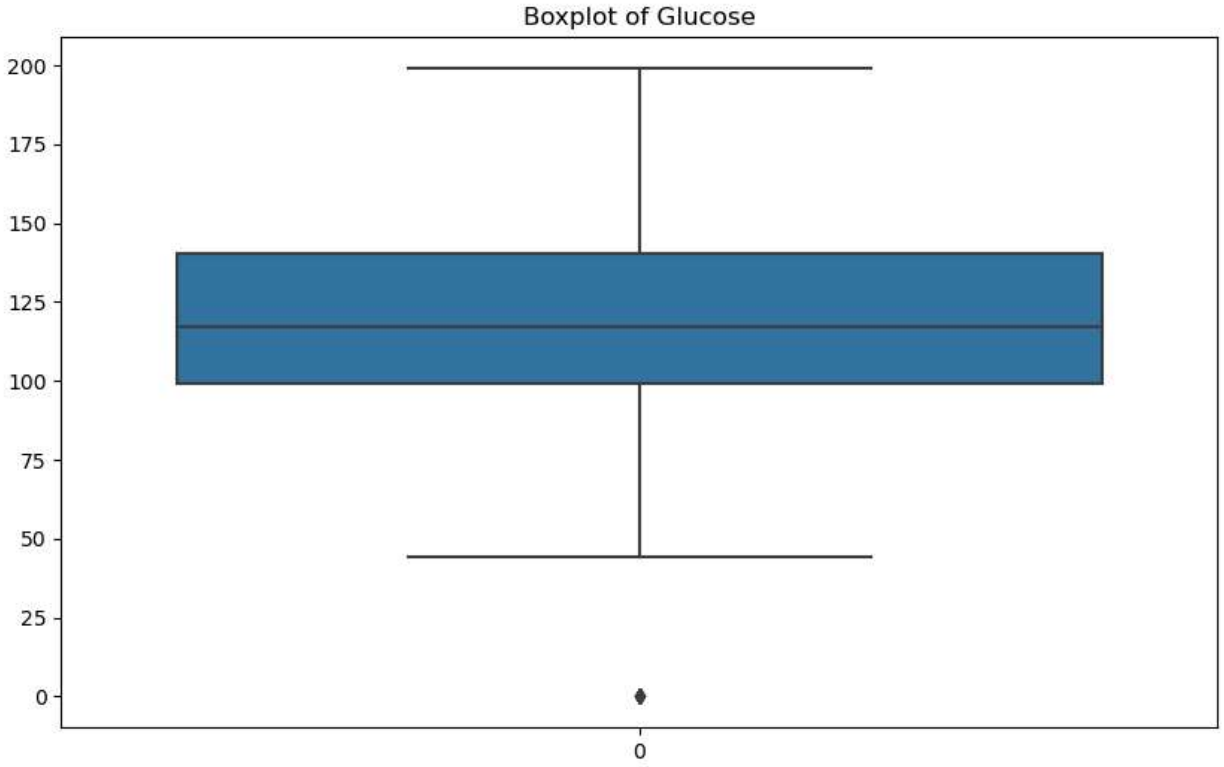


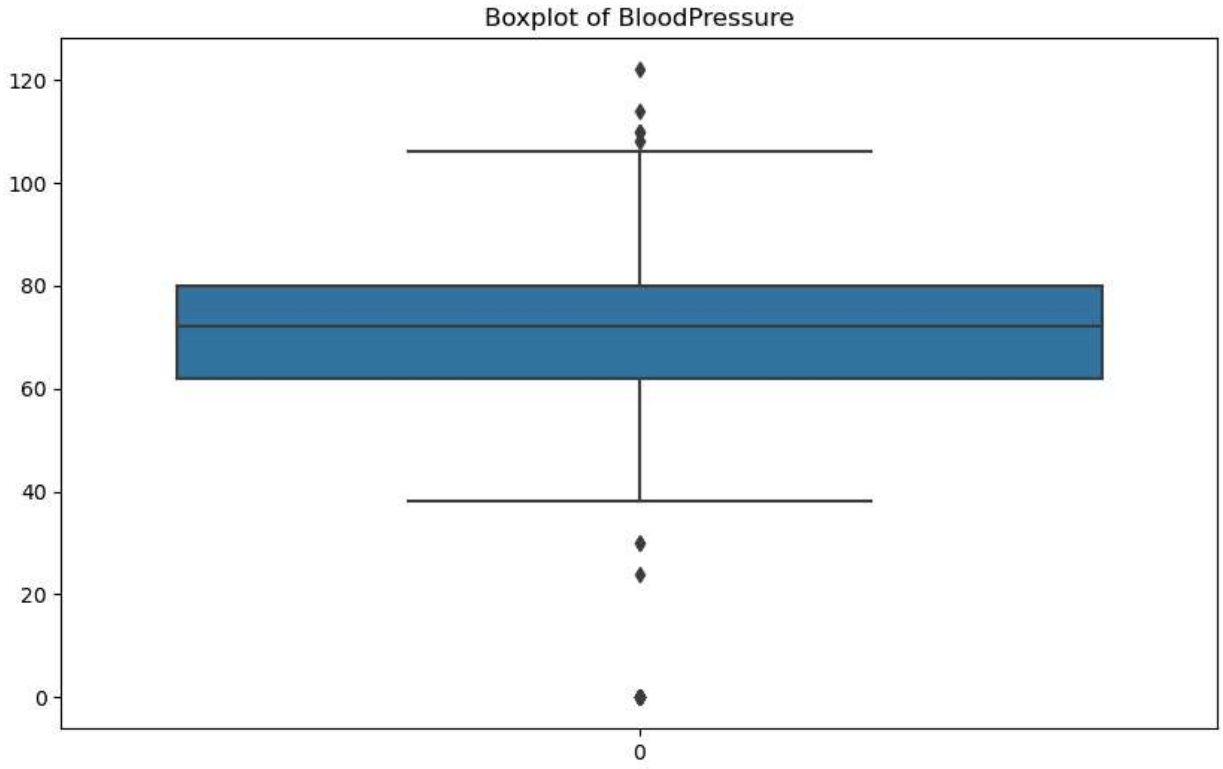
In [16]:
```
print(data.isnull().sum())
```

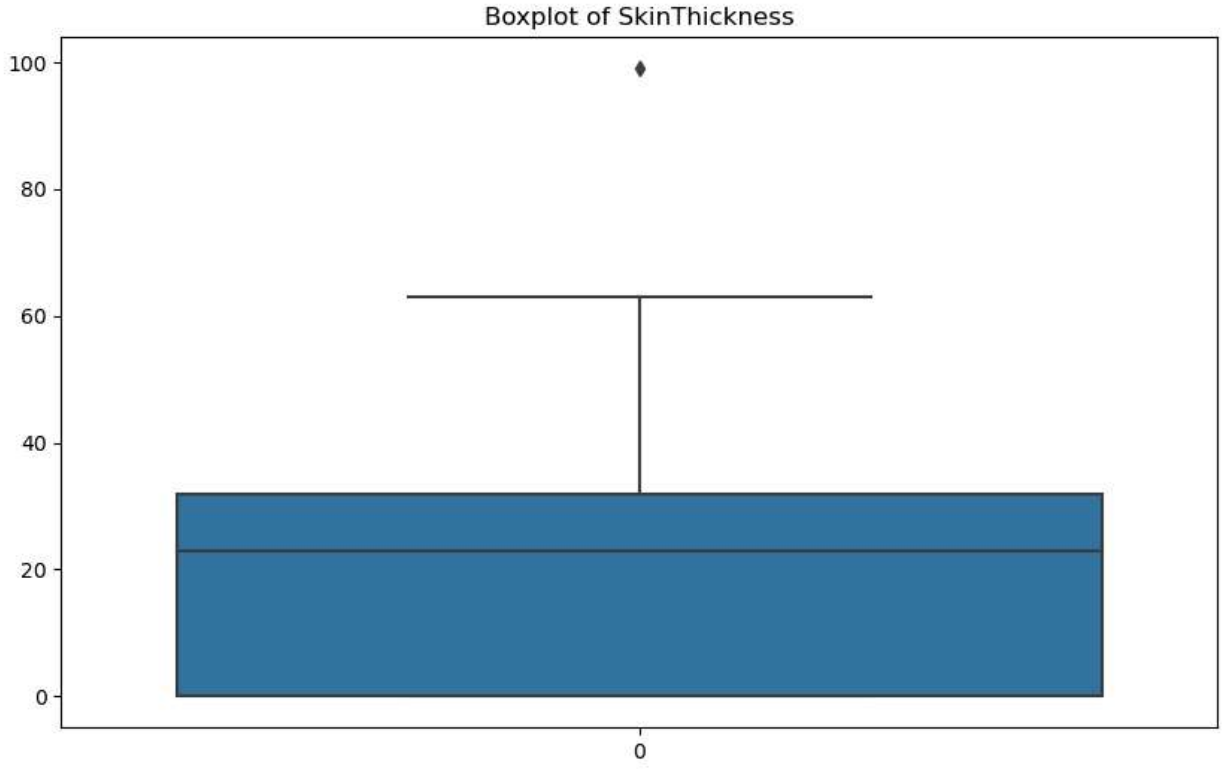
```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

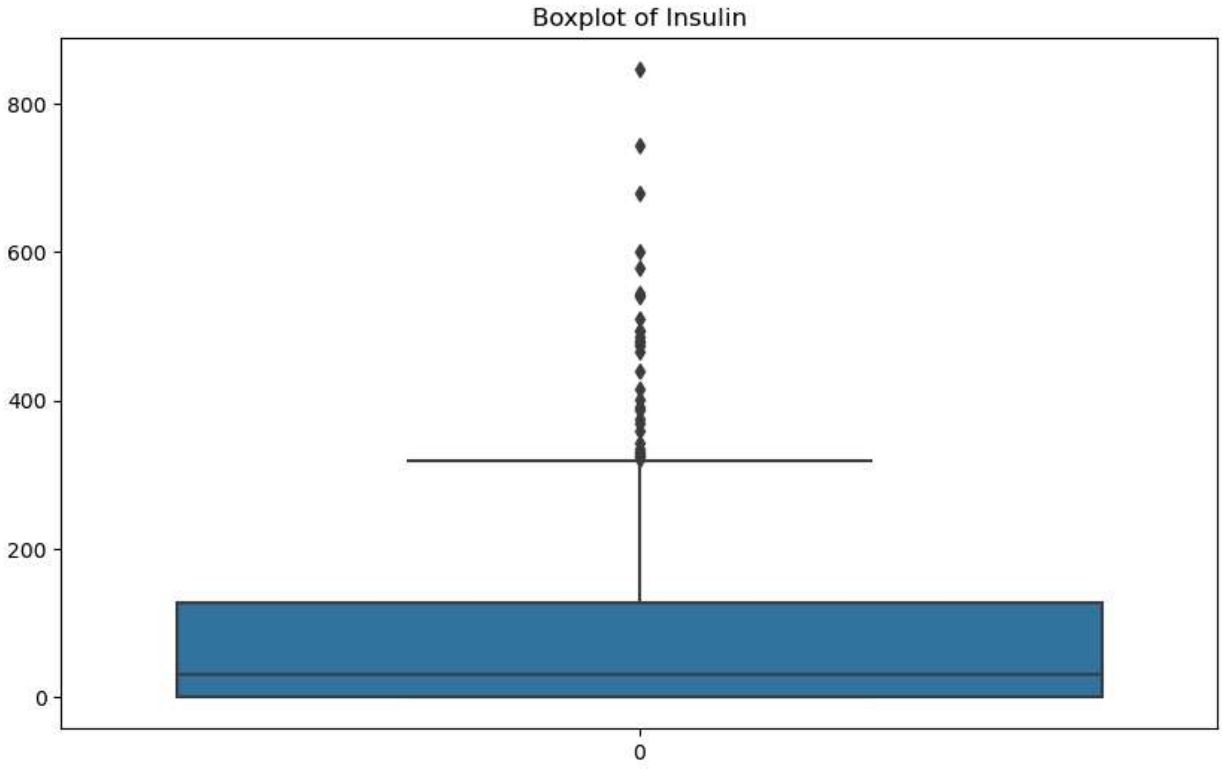
In [17]:
```
for column in data.select_dtypes(include=['float64', 'int64']).columns:
    plt.figure(figsize=(10, 6))
    sns.boxplot(data[column])
    plt.title(f'Boxplot of {column}')
    plt.show()
```

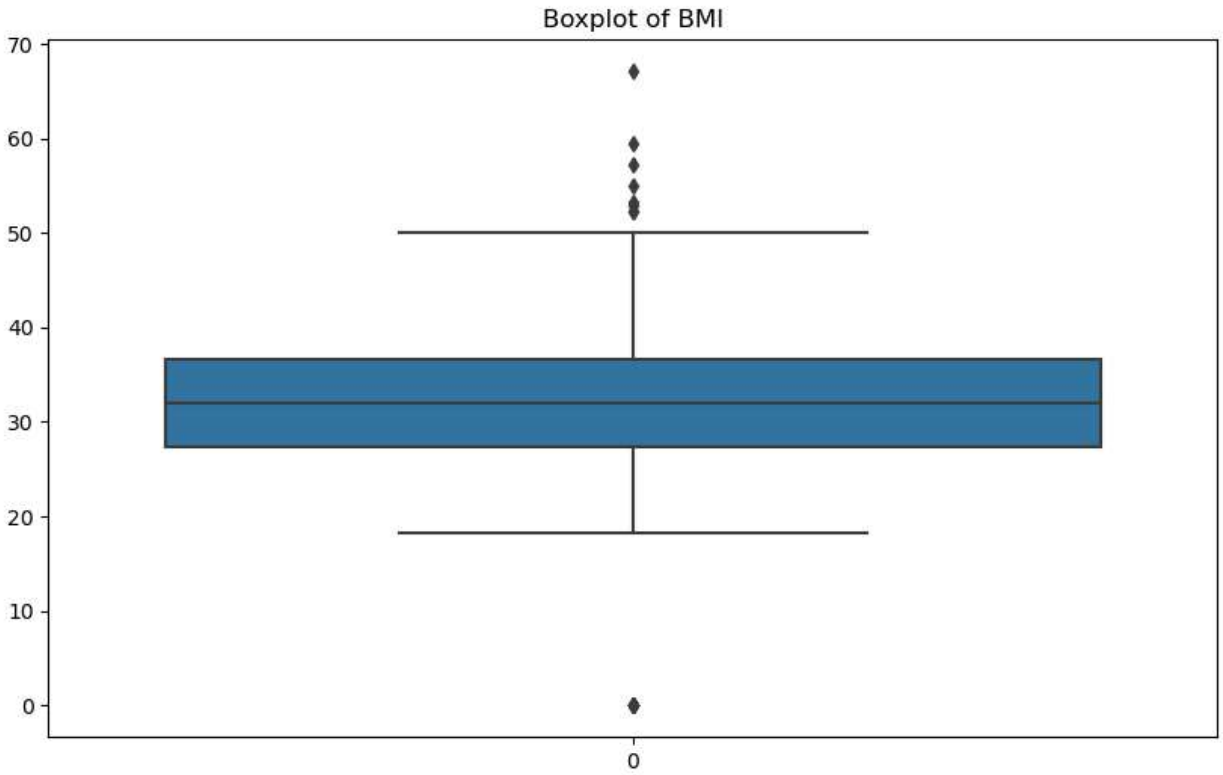

Boxplot of Pregnancies
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0

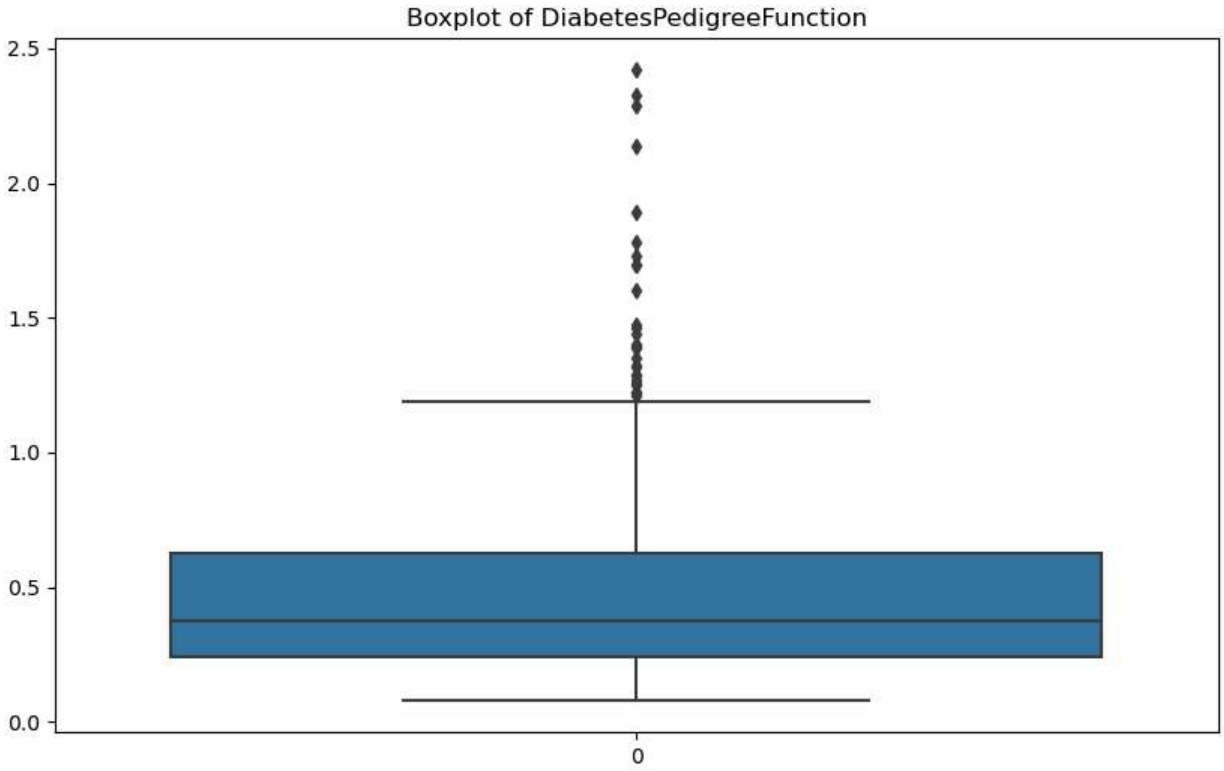

Boxplot of Glucose
200
175
150
125
100
75
50
25
0
0

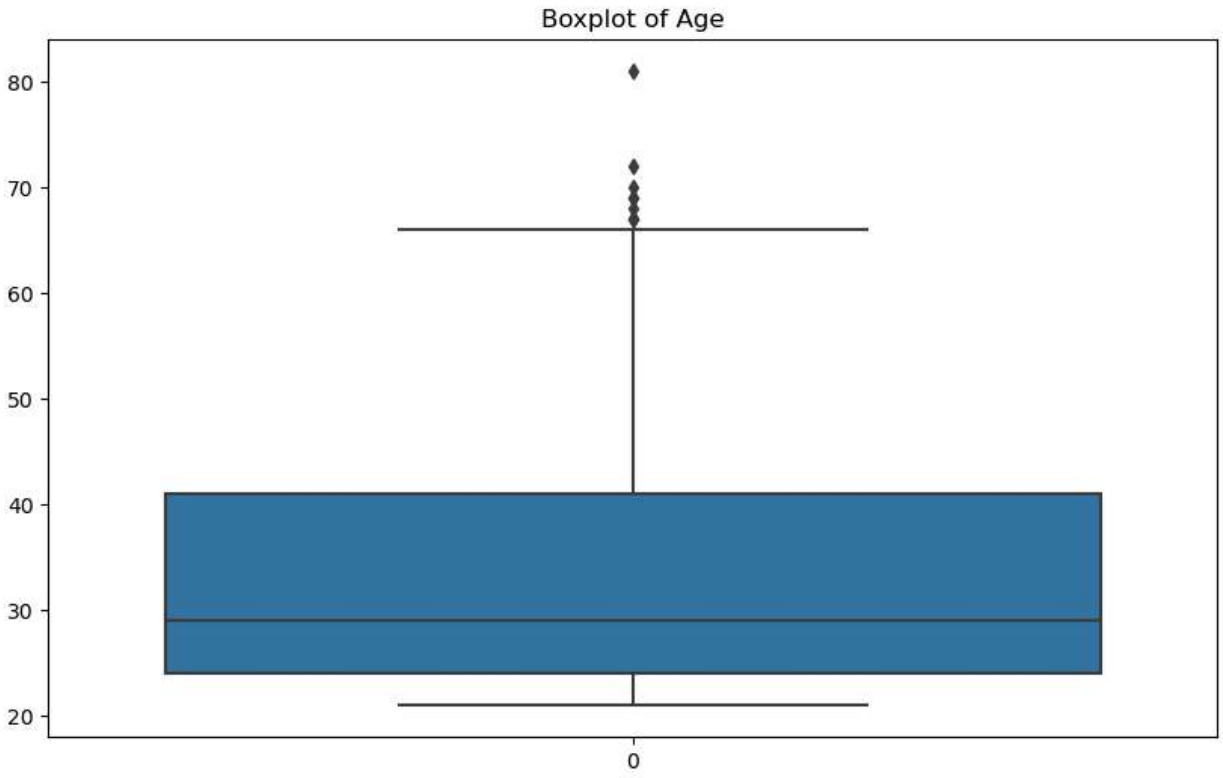

Boxplot of BloodPressure
120
100
80
60
40
20
0
0

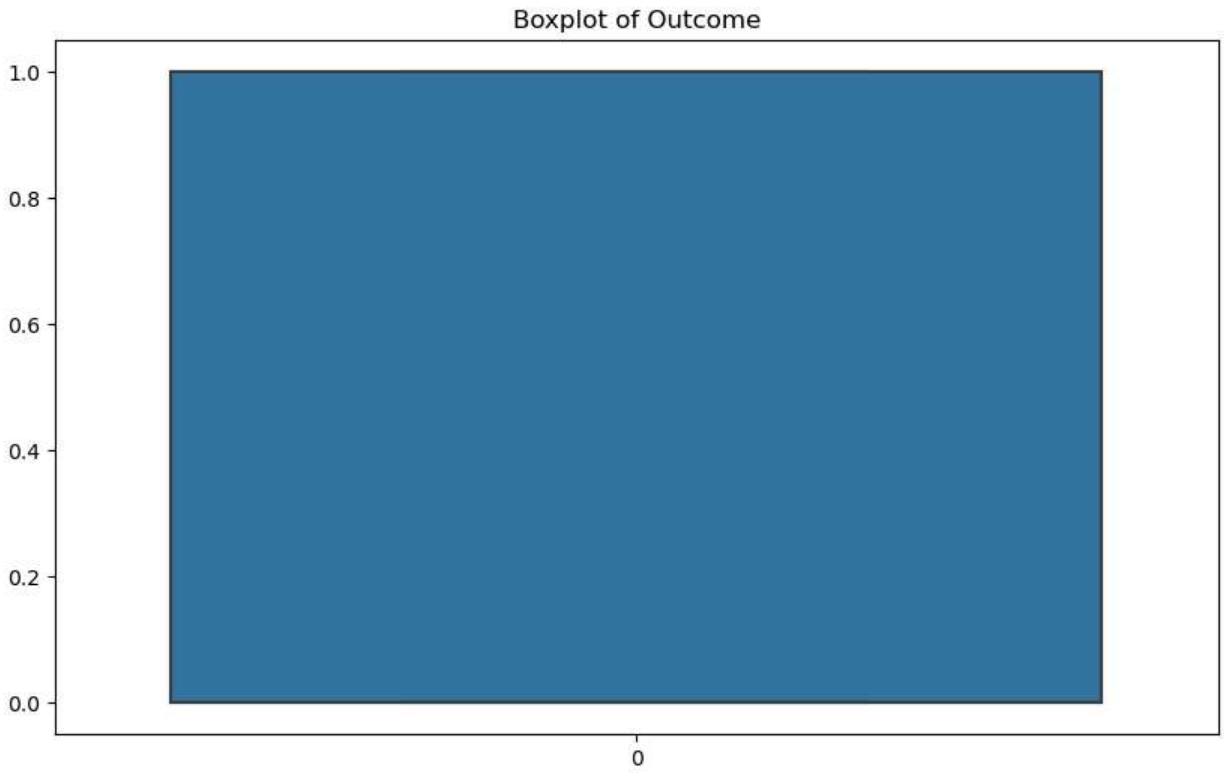

Boxplot of SkinThickness
100
80
60
40
20
0
0


Boxplot of Insulin
800
600
400
200
0
0


Boxplot of BMI
70
60
50
40
30
20
10
0
0


Boxplot of DiabetesPedigreeFunction
2.5
2.0
1.5
1.0
0.5
0.0
0


Boxplot of Age
80
70
60
50
40
30
20
0



In [18]:
```
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
outlier_condition = ((data < (Q1 - 1.5 * IQR)) | (data > (Q3 + 1.5 * IQR)))
print(data[outlier_condition.any(axis=1)])
```

```
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
4              0      137             40             35      168  43.1
7             10      115              0              0        0  35.3
8              2      197             70             45      543  30.5
9              8      125             96              0        0   0.0
12            10      139             80              0        0  27.1
..           ...      ...            ...            ...      ...   ...
706           10      115              0              0        0   0.0
707            2      127             46             21      335  34.4
710            3      158             64             13      387  31.2
715            7      187             50             33      392  33.9
753            0      181             88             44      510  43.3

     DiabetesPedigreeFunction  Age  Outcome
4                       2.288   33        1
7                       0.134   29        0
8                       0.158   53        1
9                       0.232   54        1
12                      1.441   57        0
..                        ...  ...      ...
706                     0.261   30        1
707                     0.176   22        0
710                     0.295   24        0
715                     0.826   34        1
753                     0.222   26        1

[129 rows x 9 columns]
```

In [ ]: