

```
In [2]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('Diabetes Missing Data.csv', delimiter=';')
print(df.head())
```

	Pregnant	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	\
0	6	148.0	72.0	35.0	NaN	33.6	
1	1	85.0	66.0	29.0	NaN	26.6	
2	8	183.0	64.0	NaN	NaN	23.3	
3	1	89.0	66.0	23.0	94.0	28.1	
4	0	137.0	40.0	35.0	168.0	43.1	

	Diabetes_Pedigree	Age	Class
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

```
In [3]: print(df.isnull().sum())
```

```
Pregnant      0
Glucose        5
Diastolic_BP  35
Skin_Fold     227
Serum_Insulin 374
BMI           11
Diabetes_Pedigree  0
Age           0
Class         0
dtype: int64
```

```
In [8]: for col in ['Glucose', 'Diastolic_BP', 'Skin_Fold', 'Serum_Insulin', 'BMI']:
        median_value = df[col].median()
        df[col] = df[col].fillna(median_value)

print(df.isnull().sum())
```

```
Pregnant      0
Glucose        0
Diastolic_BP  0
Skin_Fold     0
Serum_Insulin 0
BMI           0
Diabetes_Pedigree  0
Age           0
Class         0
dtype: int64
```

```
In [9]: import numpy as np

numeric_cols = ['Glucose', 'Diastolic_BP', 'Skin_Fold', 'Serum_Insulin', 'BMI'],

from scipy.stats import zscore

df_clean = df.copy()
```

```

z_scores = np.abs(zscore(df_clean[numeric_cols]))

rows_to_remove = (z_scores > 3).any(axis=1)

df_clean = df_clean[~rows_to_remove]

print("ROWS AFTER DELETE:", df_clean.shape[0])

```

ROWS AFTER DELETE: 730

```

In [10]: print("ROWS BEFORE DELETE:\n", df[numeric_cols].describe())
         print("\nROWS AFTER DELETE:\n", df_clean[numeric_cols].describe())

```

ROWS BEFORE DELETE:

	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI \
count	768.000000	768.000000	768.000000	768.000000	768.000000
mean	121.656250	72.386719	29.108073	140.671875	32.455208
std	30.438286	12.096642	8.791221	86.383060	6.875177
min	44.000000	24.000000	7.000000	14.000000	18.200000
25%	99.750000	64.000000	25.000000	121.500000	27.500000
50%	117.000000	72.000000	29.000000	125.000000	32.300000
75%	140.250000	80.000000	32.000000	127.250000	36.600000
max	199.000000	122.000000	99.000000	846.000000	67.100000

	Age
count	768.000000
mean	33.240885
std	11.760232
min	21.000000
25%	24.000000
50%	29.000000
75%	41.000000
max	81.000000

ROWS AFTER DELETE:

	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI \
count	730.000000	730.000000	730.000000	730.000000	730.000000
mean	120.193151	72.171233	28.708219	130.464384	32.165616
std	29.704714	11.363045	8.139798	55.457216	6.487230
min	44.000000	38.000000	7.000000	15.000000	18.200000
25%	99.000000	64.000000	25.000000	120.000000	27.425000
50%	116.000000	72.000000	29.000000	125.000000	32.250000
75%	138.000000	80.000000	32.000000	125.000000	36.275000
max	199.000000	108.000000	54.000000	392.000000	52.900000

	Age
count	730.000000
mean	32.956164
std	11.293600
min	21.000000
25%	24.000000
50%	29.000000
75%	40.000000
max	68.000000

```

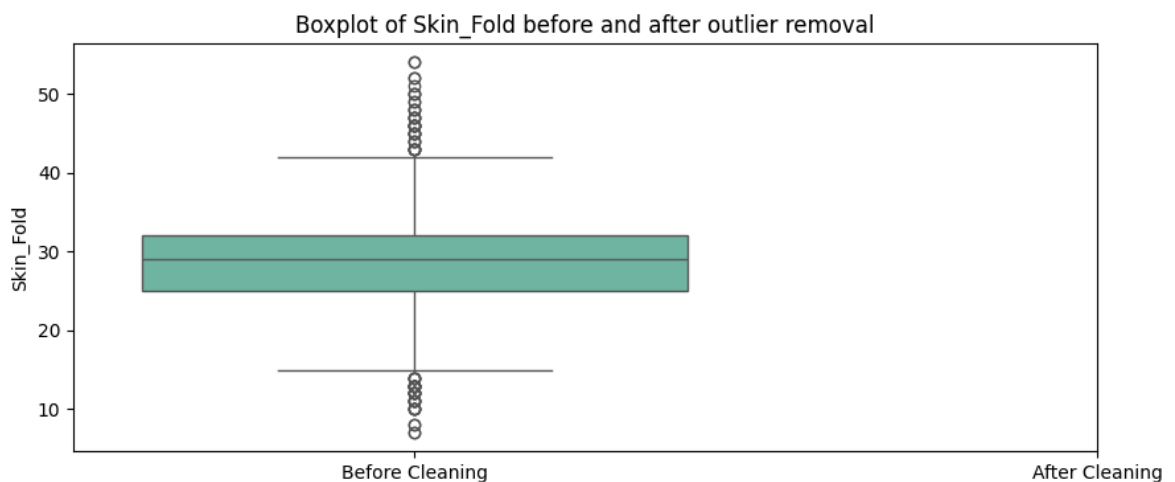
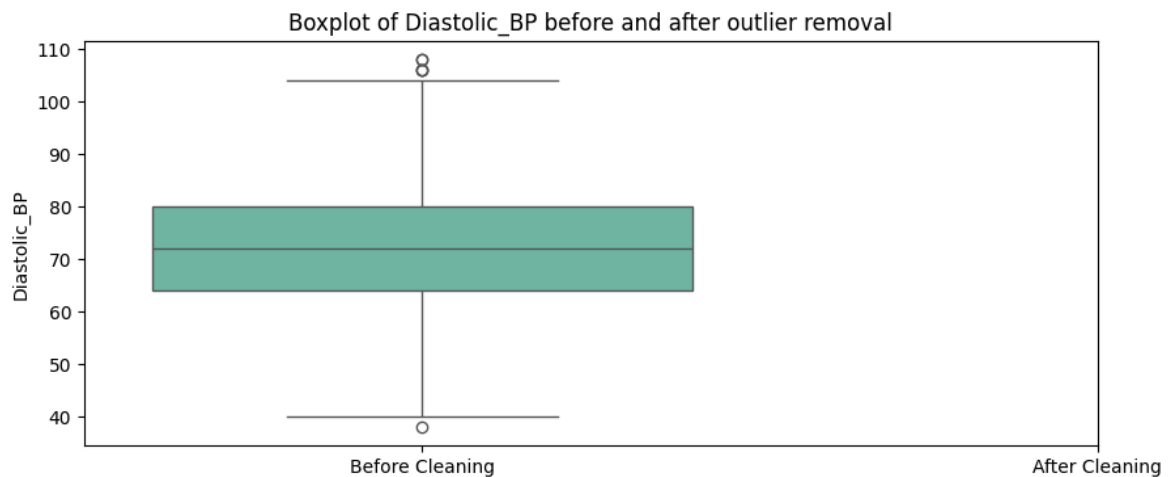
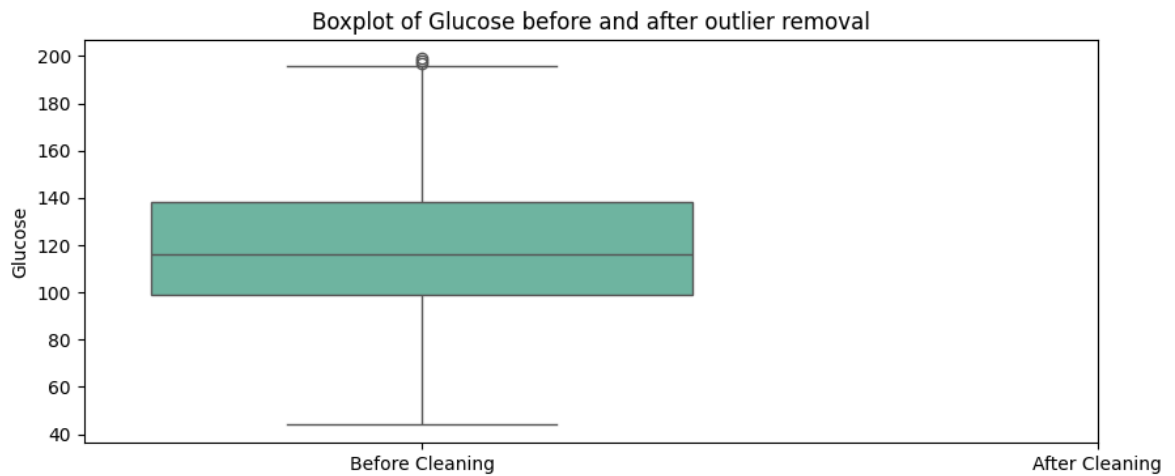
In [17]: import matplotlib.pyplot as plt
         import seaborn as sns

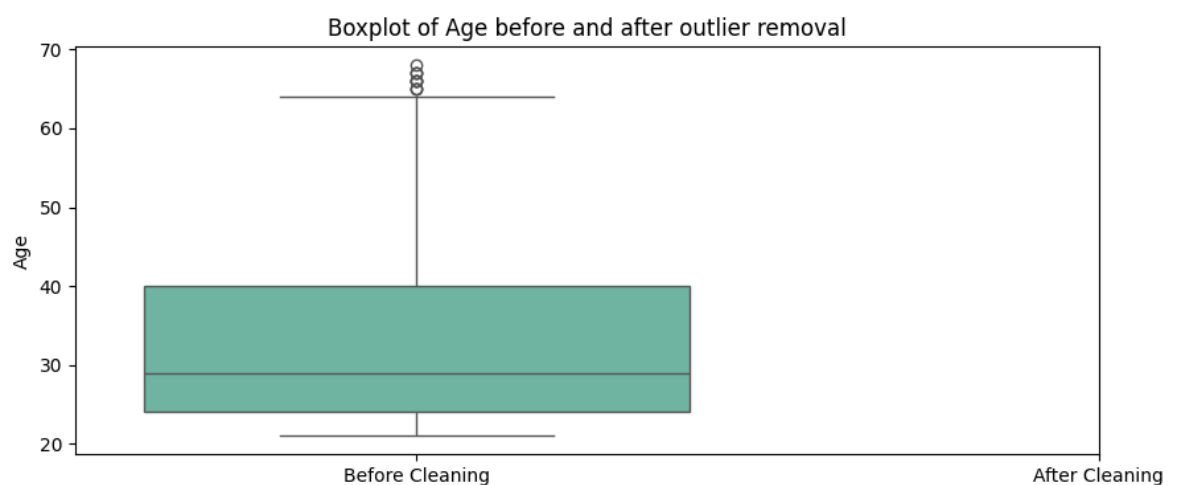
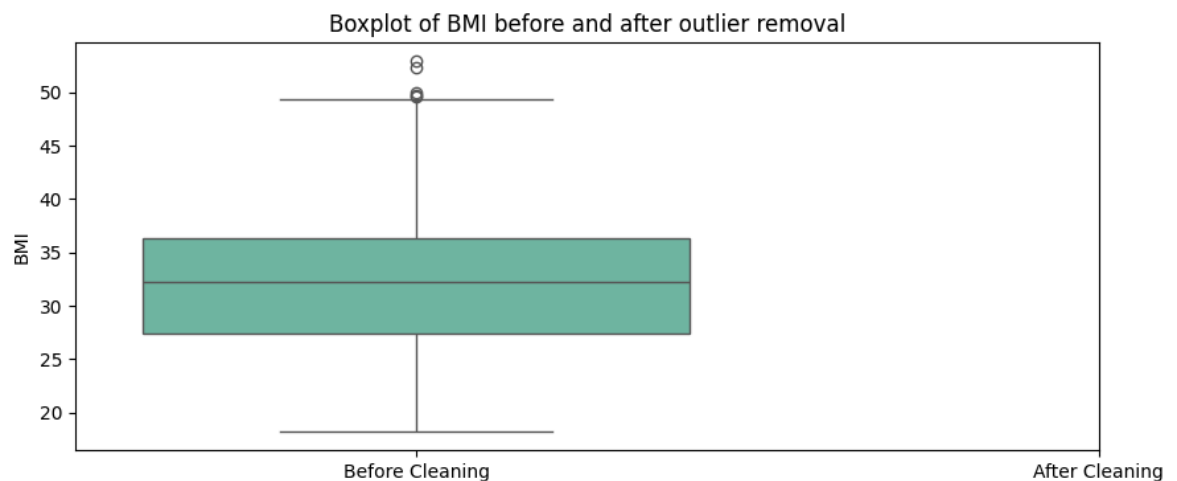
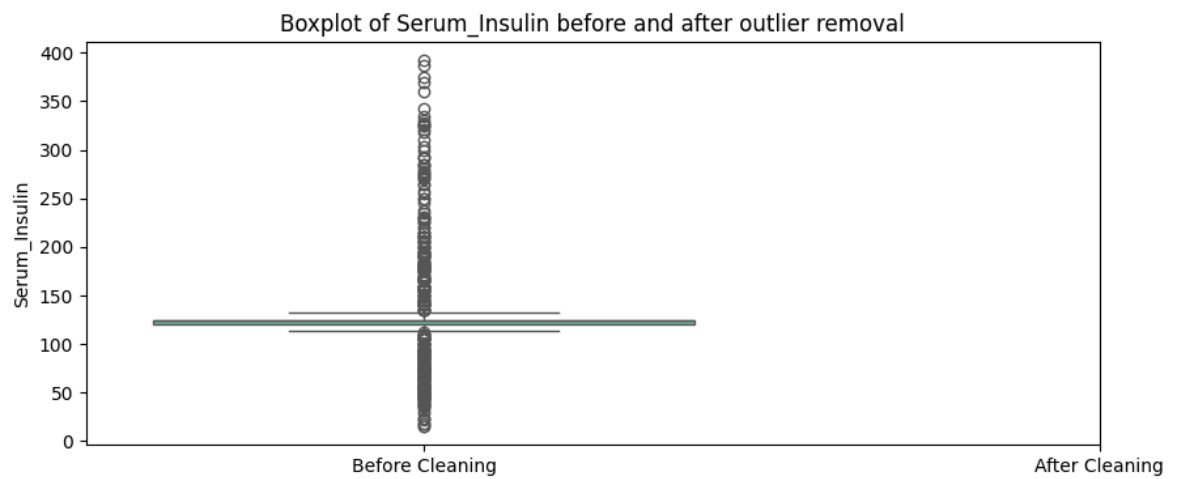
         numeric_cols = ['Glucose', 'Diastolic_BP', 'Skin_Fold', 'Serum_Insulin', 'BMI',
                        'Age']

         for col in numeric_cols:

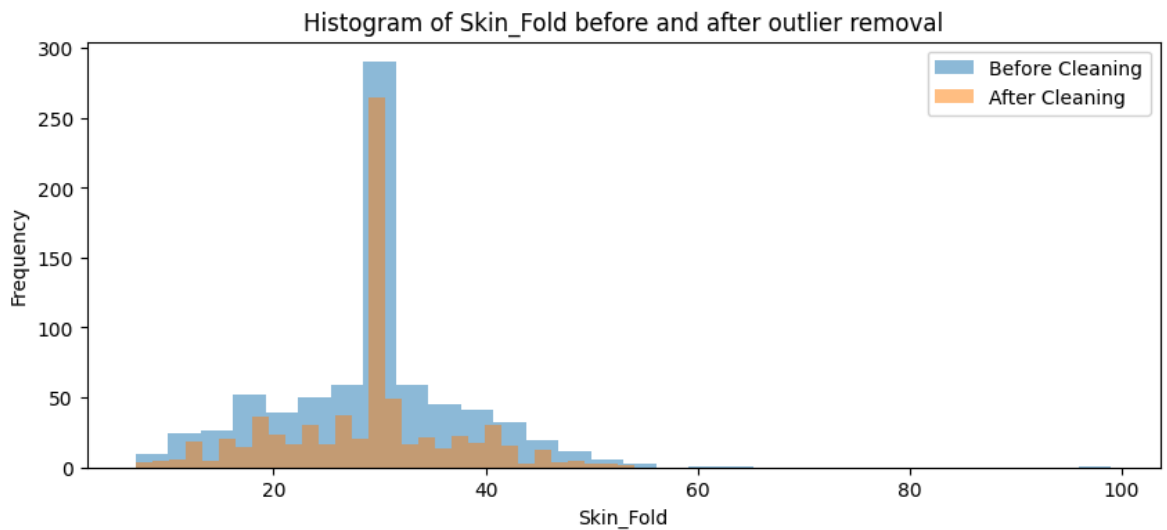
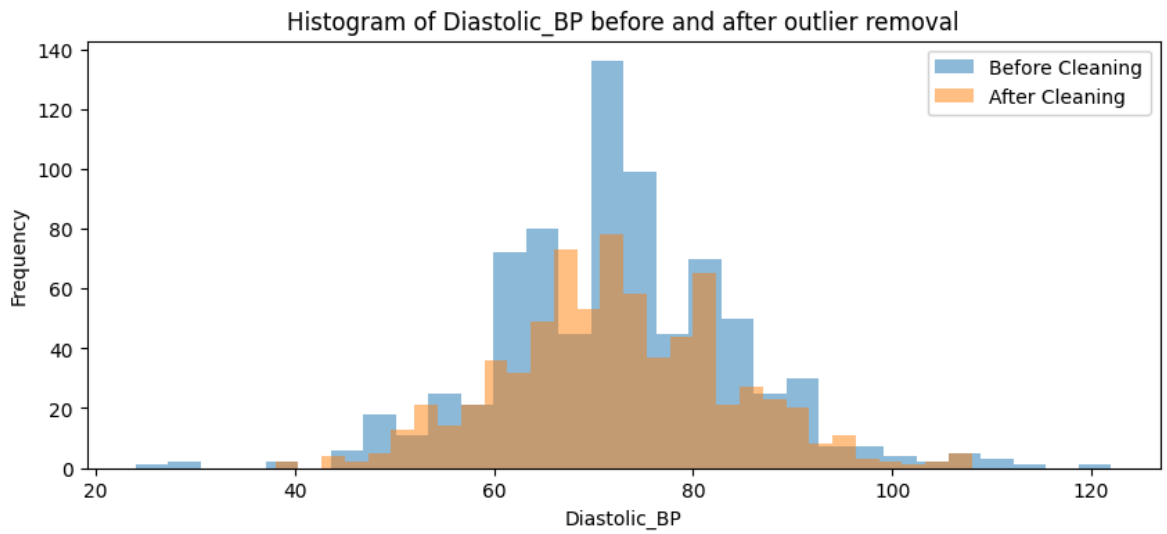
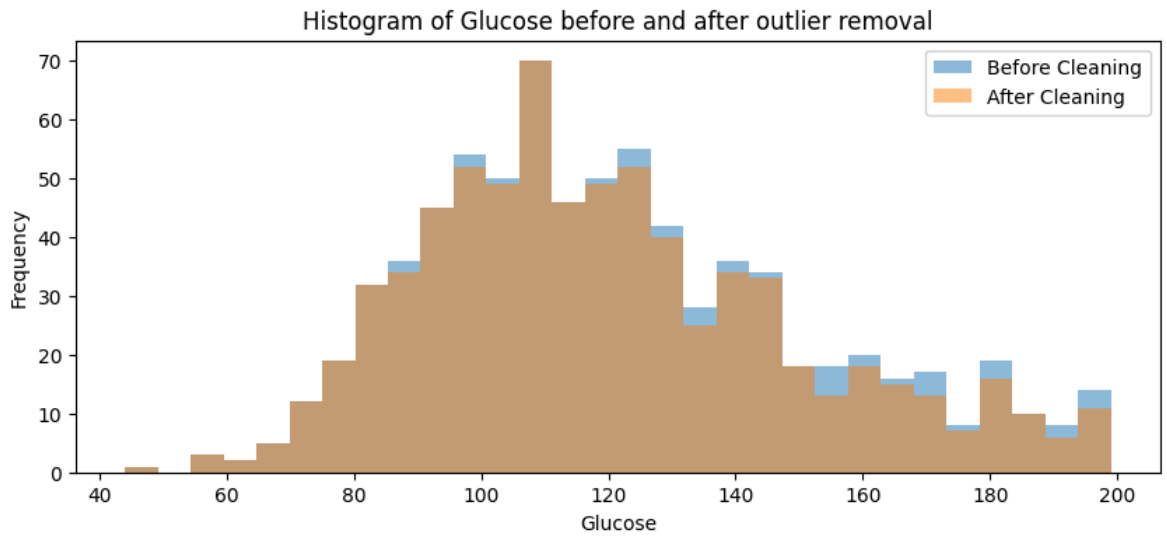
```

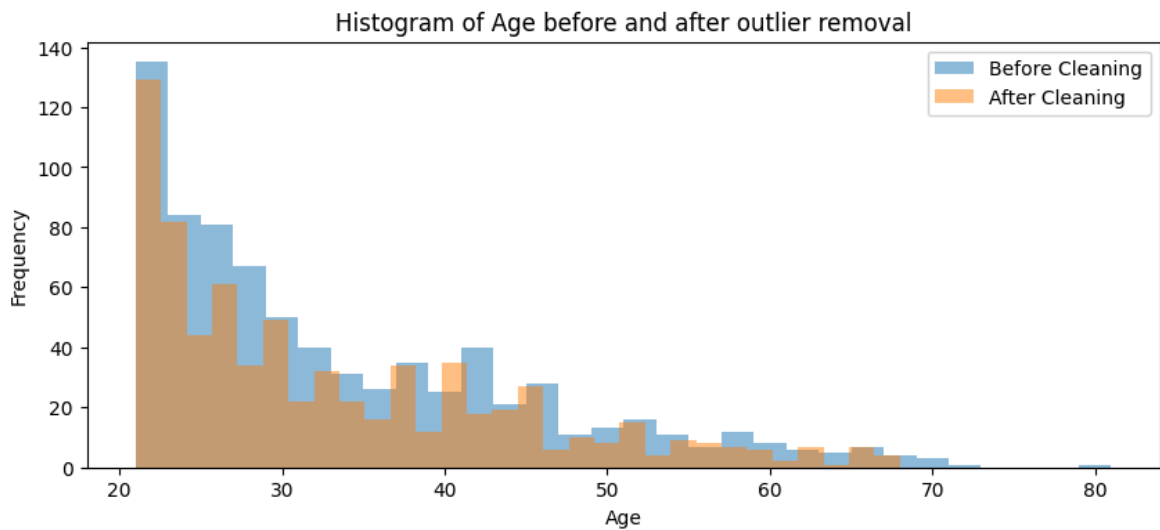
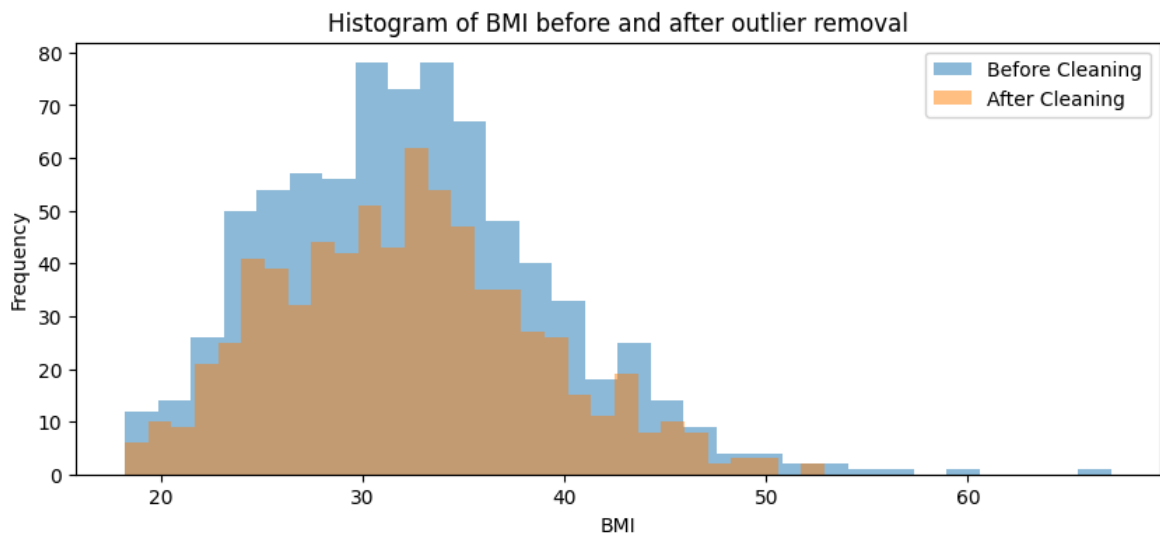
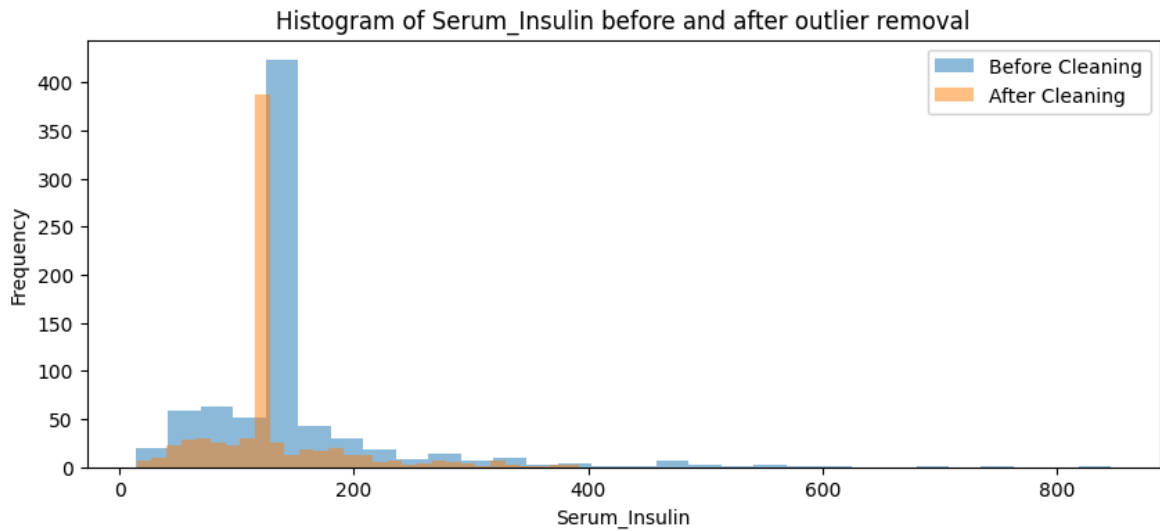
```
plt.figure(figsize=(10, 4))
sns.boxplot(data=[df[col], df_clean[col]], palette="Set2")
plt.xticks([0, 1], ['Before Cleaning', 'After Cleaning'])
plt.title(f'Boxplot of {col} before and after outlier removal')
plt.ylabel(col)
plt.show()
```



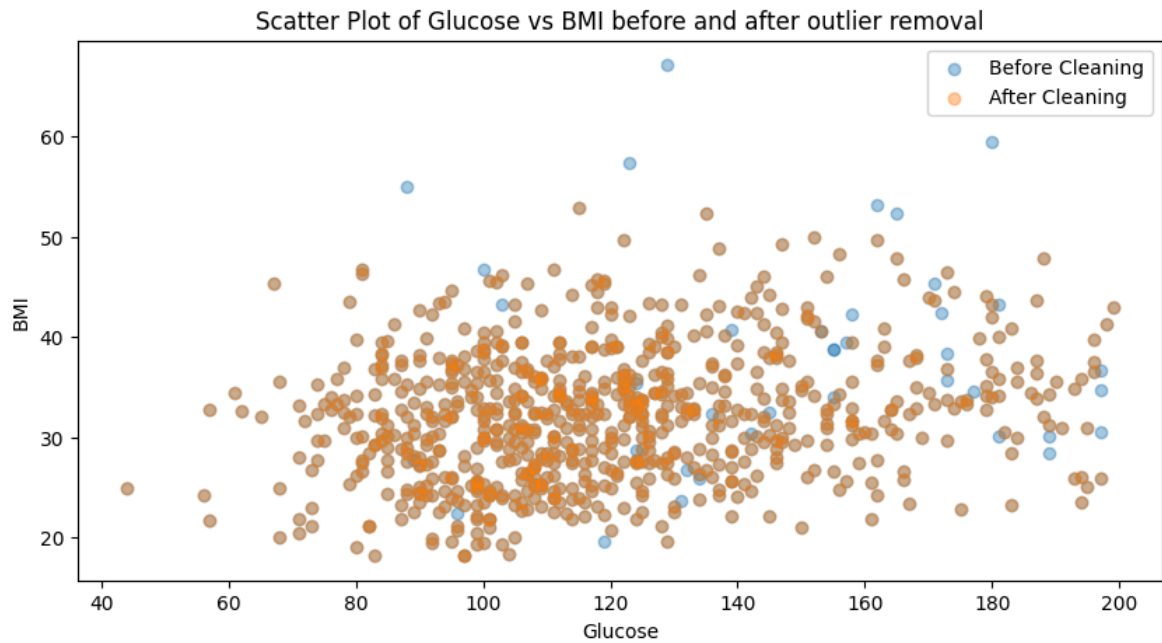


```
In [18]: for col in numeric_cols:
plt.figure(figsize=(10, 4))
plt.hist(df[col], bins=30, alpha=0.5, label='Before Cleaning')
plt.hist(df_clean[col], bins=30, alpha=0.5, label='After Cleaning')
plt.title(f'Histogram of {col} before and after outlier removal')
plt.xlabel(col)
plt.ylabel('Frequency')
plt.legend()
plt.show()
```





```
In [21]: plt.figure(figsize=(10, 5))
plt.scatter(df['Glucose'], df['BMI'], alpha=0.4, label='Before Cleaning')
plt.scatter(df_clean['Glucose'], df_clean['BMI'], alpha=0.4, label='After Cleaning')
plt.xlabel('Glucose')
plt.ylabel('BMI')
plt.title('Scatter Plot of Glucose vs BMI before and after outlier removal')
plt.legend()
plt.show()
```

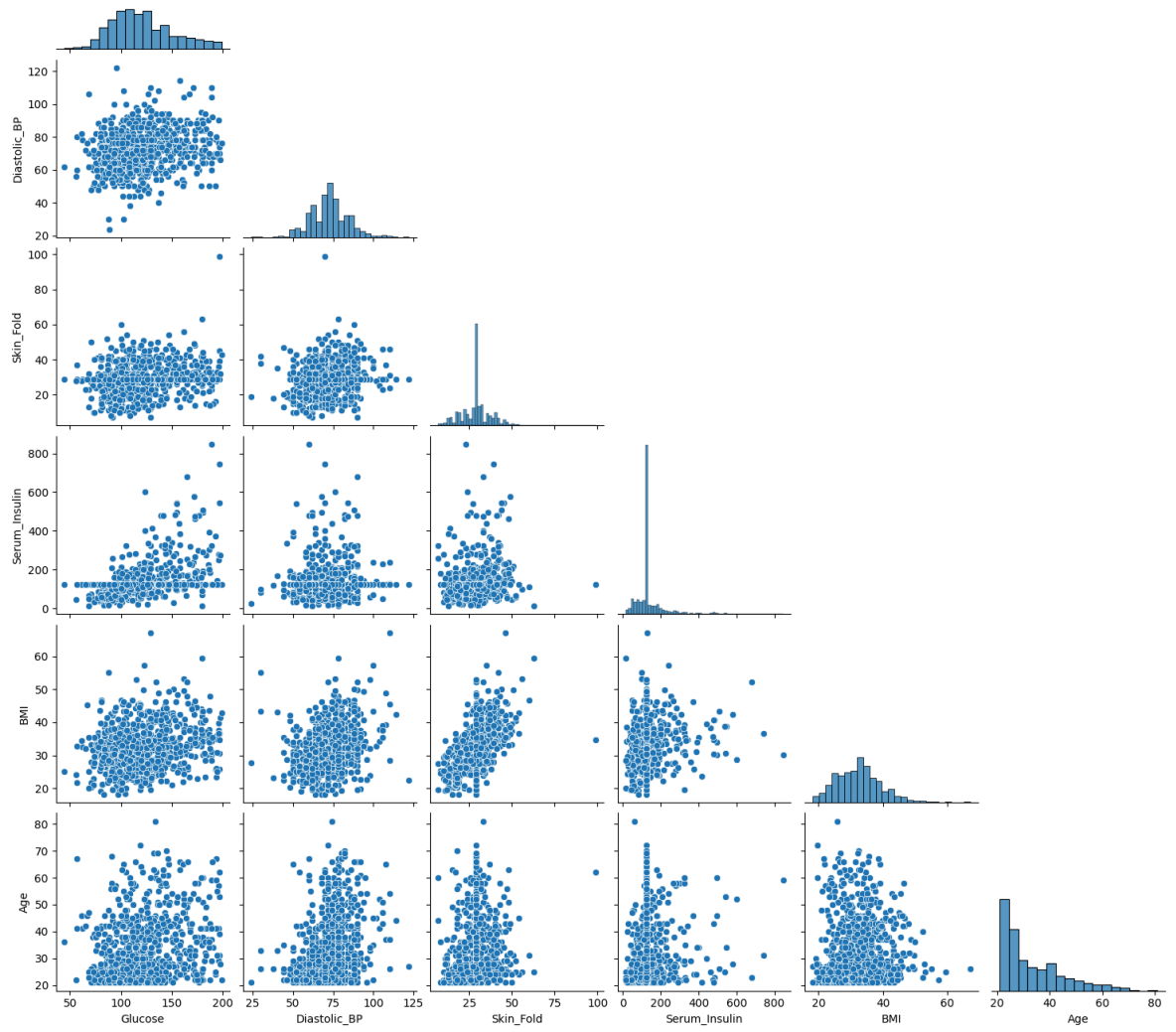


```
In [20]: import seaborn as sns

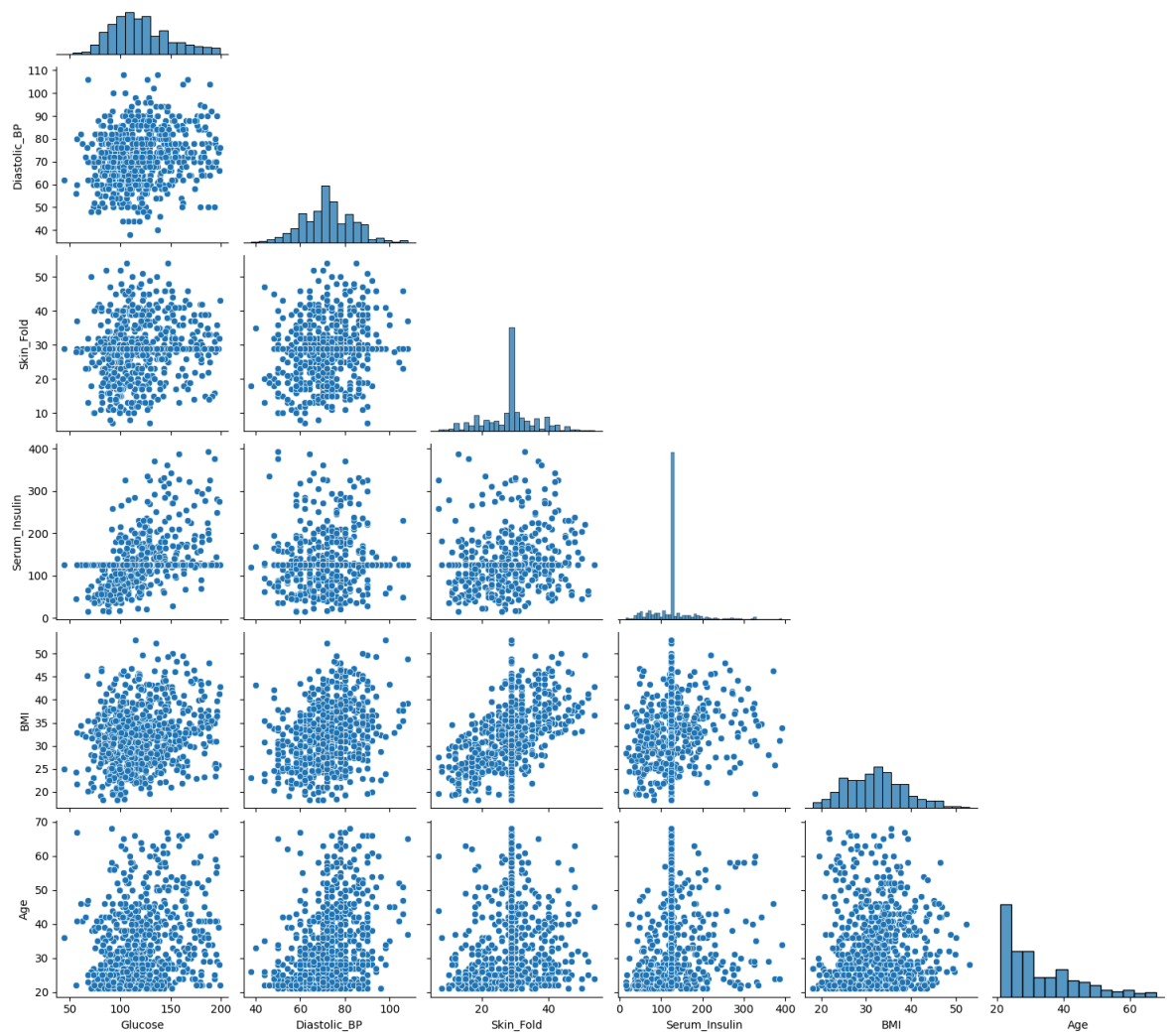
# Pairplot before cleaning
sns.pairplot(df[numeric_cols], corner=True)
plt.suptitle('Pairplot before outlier removal', y=1.02)
plt.show()

# Pairplot after cleaning
sns.pairplot(df_clean[numeric_cols], corner=True)
plt.suptitle('Pairplot after outlier removal', y=1.02)
plt.show()
```

Pairplot before outlier removal



Pairplot after outlier removal



In []: