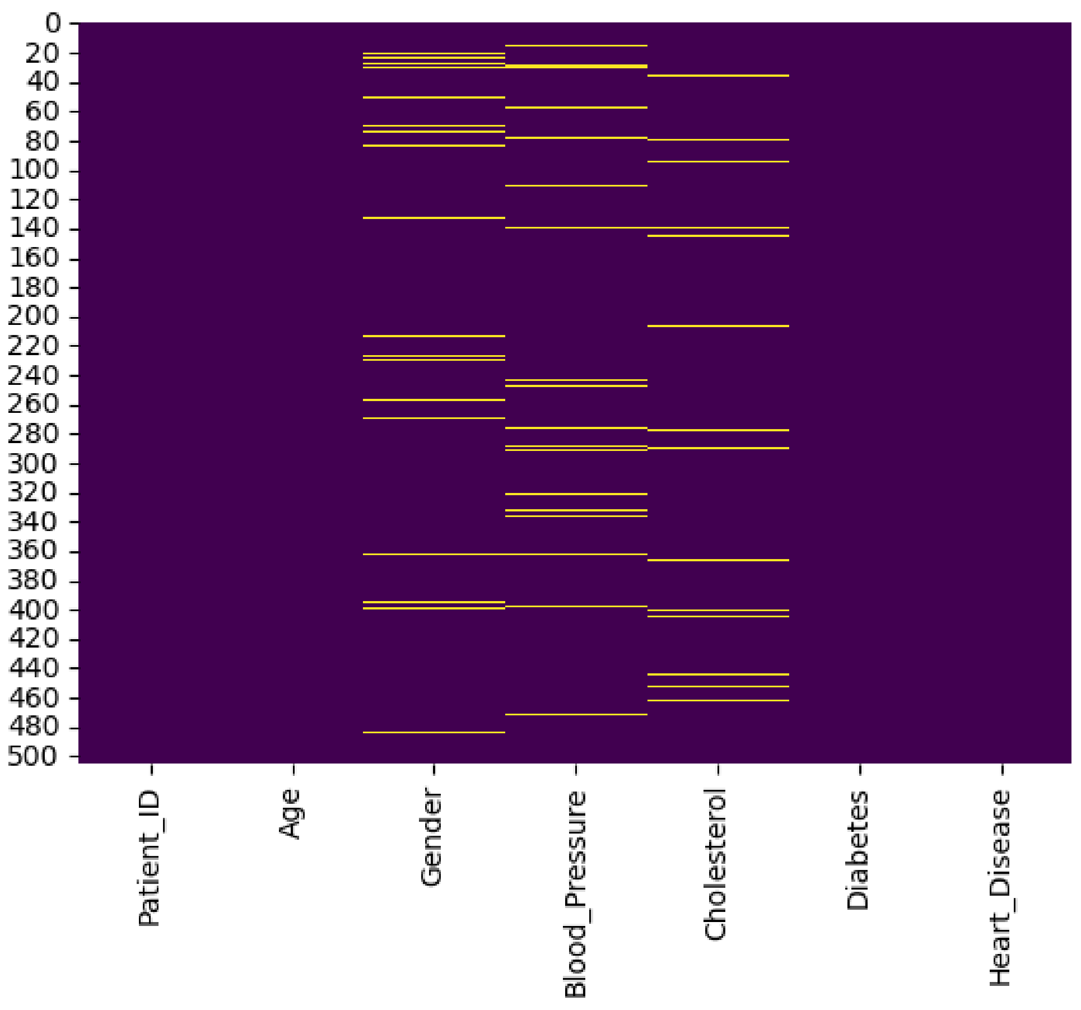
In [7]:

```
import pandas as pd

# Load the dataset (replace with your actual dataset path)
df = pd.read_csv("C:/Users/gundr/Downloads/Day_15_Healthcare_Data.csv")

# Initial data exploration
print(df.head())  # View the first few rows
print(df.info())  # Dataset summary (data types, non-null counts)

# Check for missing values across columns
missing_values = df.isna().sum()
print(missing_values)

# Calculate percentage of missing values for each column
missing_percentage = (df.isna().sum() / len(df)) * 100
print(missing_percentage)
```

```
   Patient_ID  Age  Gender  Blood_Pressure  Cholesterol Diabetes Heart_Disease
0           1   69    Male            95.0        122.0       No            No
1           2   32    Male           129.0        191.0       No            No
2           3   89  Female           101.0        214.0       No            No
3           4   78  Female           142.0        203.0       No            No
4           5   38    Male           160.0        217.0       No            No
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 505 entries, 0 to 504
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Patient_ID      505 non-null    int64  
 1   Age             505 non-null    int64  
 2   Gender          484 non-null    object 
 3   Blood_Pressure  475 non-null    float64
 4   Cholesterol     485 non-null    float64
 5   Diabetes        505 non-null    object 
 6   Heart_Disease   505 non-null    object 
dtypes: float64(2), int64(2), object(3)
memory usage: 27.7+ KB
None
Patient_ID         0
Age                0
Gender            21
Blood_Pressure    30
Cholesterol       20
Diabetes           0
Heart_Disease      0
dtype: int64
Patient_ID        0.000000
Age               0.000000
Gender            4.158416
Blood_Pressure    5.940594
Cholesterol       3.960396
Diabetes          0.000000
Heart_Disease     0.000000
dtype: float64
```

In [10]:

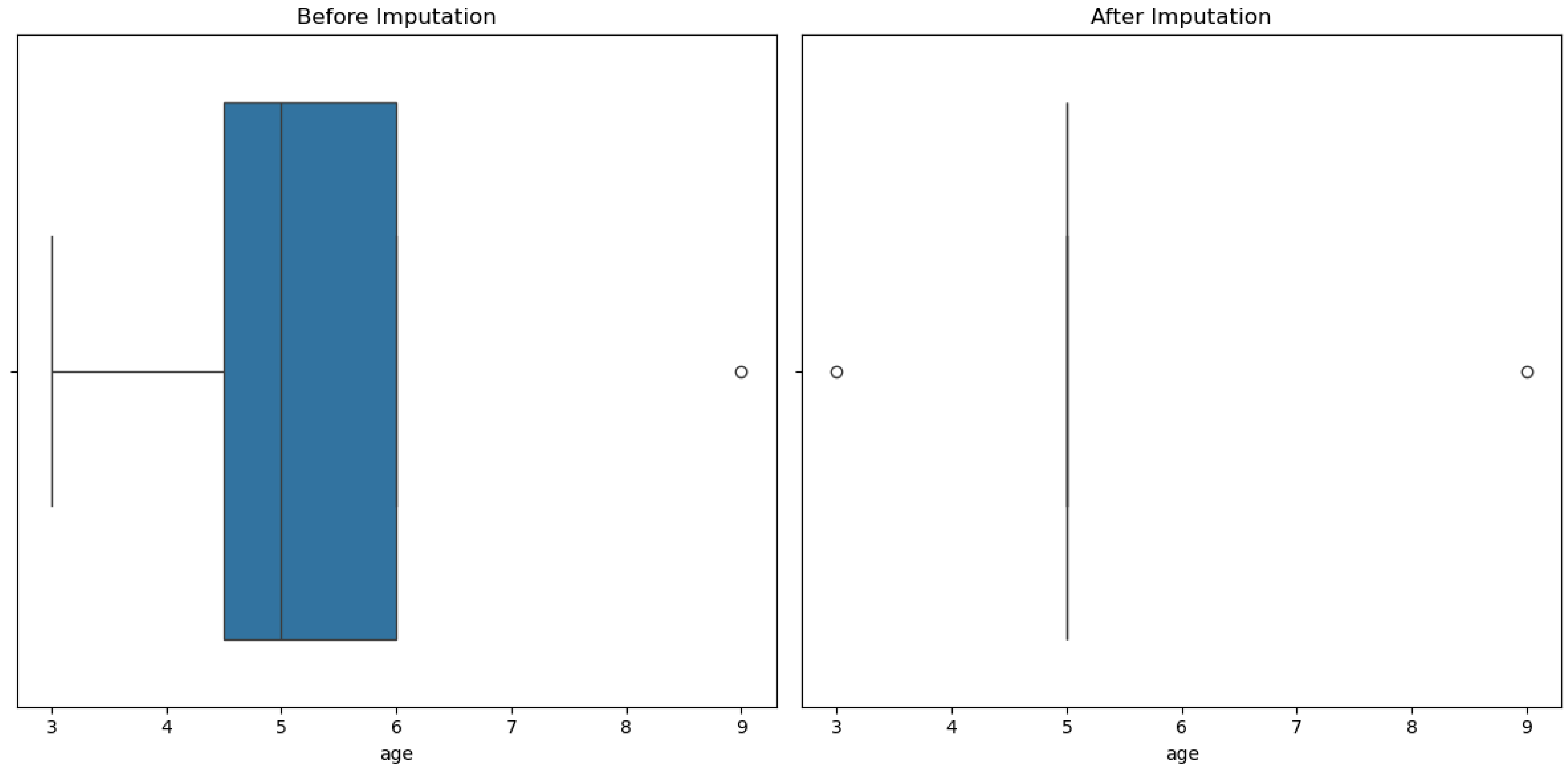
```
import seaborn as sns
import matplotlib.pyplot as plt

# Visualize missing data with a heatmap
sns.heatmap(df.isna(), cbar=False, cmap='viridis')
plt.show()
```



In [39]:

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

# Create the initial DataFrame with some missing values
data = {
    'age': [3, 5, 8, 5, 9]
}
df = pd.DataFrame(data)

# Introduce missing data for testing (e.g., randomly set some values to NaN)
df.loc[2, 'age'] = None  # Set one value as missing

# Before imputation (save a copy)
df_before = df.copy()

# Apply imputation (Median imputation for the 'age' column)
df['age'] = df['age'].fillna(df['age'].median())

# After imputation (save a copy)
df_after = df.copy()

# Compare mean and standard deviation before and after imputation
print(f"Before Imputation Mean: {df_before['age'].mean()}")
print(f"After Imputation Mean: {df_after['age'].mean()}")

print(f"Before Imputation Std Dev: {df_before['age'].std()}")
print(f"After Imputation Std Dev: {df_after['age'].std()}")

# Visualize the impact using boxplots
plt.figure(figsize=(12, 6))

# Before imputation
plt.subplot(1, 2, 1)
sns.boxplot(x=df_before['age'])
plt.title('Before Imputation')

# After imputation
plt.subplot(1, 2, 2)
sns.boxplot(x=df_after['age'])
plt.title('After Imputation')

plt.tight_layout()
plt.show()
```

```
Before Imputation Mean: 5.5
After Imputation Mean: 5.4
Before Imputation Std Dev: 2.516611478423583
After Imputation Std Dev: 2.1908902300206643
```



Before Imputation
3
4
5
6
7
8
9
age
After Imputation
3
4
5
6
7
8
9
age