```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

df = pd.read_csv("heart_disease.csv")

df.head()
```

```
    Age  Gender  Blood Pressure  Cholesterol Level Exercise Habits
Smoking  \
0  56.0    Male           153.0              155.0            High
Yes
1  69.0  Female           146.0              286.0            High
No
2  46.0    Male           126.0              216.0             Low
No
3  32.0  Female           122.0              293.0            High
Yes
4  60.0    Male           166.0              242.0             Low
Yes

  Family Heart Disease Diabetes        BMI High Blood Pressure  ...  \
0                  Yes       No  24.991591                 Yes  ...
1                  Yes      Yes  25.221799                  No  ...
2                   No       No  29.855447                  No  ...
3                  Yes       No  24.130477                 Yes  ...
4                  Yes      Yes  20.486289                 Yes  ...

  High LDL Cholesterol Alcohol Consumption Stress Level Sleep Hours  \
0                   No                High       Medium    7.633228
1                   No              Medium         High    8.744034
2                  Yes                 Low          Low    4.440440
3                  Yes                 Low         High    5.249405
4                   No                 Low         High    7.030971

   Sugar Consumption Triglyceride Level  Fasting Blood Sugar  CRP
Level  \
0             Medium              342.0                  NaN
12.969246
1             Medium              133.0                157.0
9.355389
2                Low              393.0                 92.0
12.709873
3               High              293.0                 94.0
12.509046
4               High              263.0                154.0
```

```
10.381259

   Homocysteine Level  Heart Disease Status
0           12.387250                    No
1           19.298875                    No
2           11.230926                    No
3            5.961958                    No
4            8.153887                    No

[5 rows x 21 columns]
```

```
df.tail()
```

```
       Age  Gender  Blood Pressure  Cholesterol Level Exercise Habits
Smoking  \
9995  25.0  Female           136.0              243.0          Medium
Yes
9996  38.0    Male           172.0              154.0          Medium
No
9997  73.0    Male           152.0              201.0            High
Yes
9998  23.0    Male           142.0              299.0             Low
Yes
9999  38.0  Female           128.0              193.0          Medium
Yes

     Family Heart Disease Diabetes        BMI High Blood Pressure  ...
\
9995                   No       No  18.788791                 Yes  ...

9996                   No       No  31.856801                 Yes  ...

9997                   No      Yes  26.899911                  No  ...

9998                   No      Yes  34.964026                 Yes  ...

9999                  Yes      Yes  25.111295                  No  ...


     High LDL Cholesterol Alcohol Consumption Stress Level Sleep Hours
\
9995                  Yes              Medium         High    6.834954

9996                  Yes                None         High    8.247784

9997                  Yes                None          Low    4.436762

9998                  Yes              Medium         High    8.526329

9999                  Yes                High       Medium    5.659394
```

```
      Sugar Consumption Triglyceride Level  Fasting Blood Sugar  CRP
Level  \
9995              Medium               343.0                133.0
3.588814
9996                 Low               377.0                 83.0
2.658267
9997                 Low               248.0                 88.0
4.408867
9998              Medium               113.0                153.0
7.215634
9999                High               121.0                149.0
14.387810

      Homocysteine Level  Heart Disease Status
9995           19.132004                   Yes
9996            9.715709                   Yes
9997            9.492429                   Yes
9998           11.873486                   Yes
9999            6.208531                   Yes

[5 rows x 21 columns]

df.shape

(10000, 21)

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   9971 non-null   float64
 1   Gender                9981 non-null   object
 2   Blood Pressure        9981 non-null   float64
 3   Cholesterol Level     9970 non-null   float64
 4   Exercise Habits       9975 non-null   object
 5   Smoking               9975 non-null   object
 6   Family Heart Disease  9979 non-null   object
 7   Diabetes              9970 non-null   object
 8   BMI                   9978 non-null   float64
 9   High Blood Pressure   9974 non-null   object
 10  Low HDL Cholesterol   9975 non-null   object
 11  High LDL Cholesterol  9974 non-null   object
 12  Alcohol Consumption   9968 non-null   object
 13  Stress Level          9978 non-null   object
 14  Sleep Hours           9975 non-null   float64
 15  Sugar Consumption     9970 non-null   object
```

```
 16  Triglyceride Level    9974 non-null   float64
 17  Fasting Blood Sugar   9978 non-null   float64
 18  CRP Level             9974 non-null   float64
 19  Homocysteine Level    9980 non-null   float64
 20  Heart Disease Status  10000 non-null  object
dtypes: float64(9), object(12)
memory usage: 1.6+ MB
```

```
df.describe()
```

```
                Age  Blood Pressure  Cholesterol Level          BMI  \
count  9971.000000     9981.000000        9970.000000  9978.000000
mean     49.296259      149.757740         225.425577    29.077269
std      18.193970       17.572969          43.575809     6.307098
min      18.000000      120.000000         150.000000    18.002837
25%      34.000000      134.000000         187.000000    23.658075
50%      49.000000      150.000000         226.000000    29.079492
75%      65.000000      165.000000         263.000000    34.520015
max      80.000000      180.000000         300.000000    39.996954

       Sleep Hours  Triglyceride Level  Fasting Blood Sugar    CRP
Level  \
count  9975.000000         9974.000000          9978.000000
9974.000000
mean      6.991329          250.734409           120.142213
7.472201
std       1.753195           87.067226            23.584011
4.340248
min       4.000605          100.000000            80.000000
0.003647
25%       5.449866          176.000000            99.000000
3.674126
50%       7.003252          250.000000           120.000000
7.472164
75%       8.531577          326.000000           141.000000
11.255592
max       9.999952          400.000000           160.000000
14.997087

       Homocysteine Level
count         9980.000000
mean            12.456271
std              4.323426
min              5.000236
25%              8.723334
50%             12.409395
75%             16.140564
max             19.999037
```

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

```
Age                          29
Gender                       19
Blood Pressure               19
Cholesterol Level            30
Exercise Habits              25
Smoking                      25
Family Heart Disease         21
Diabetes                     30
BMI                          22
High Blood Pressure          26
Low HDL Cholesterol          25
High LDL Cholesterol         26
Alcohol Consumption          32
Stress Level                 22
Sleep Hours                  25
Sugar Consumption            30
Triglyceride Level           26
Fasting Blood Sugar          22
CRP Level                    26
Homocysteine Level           20
Heart Disease Status          0
dtype: int64
```

```
df.dropna(inplace=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9500 entries, 1 to 9999
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Age                   9500 non-null   float64
 1   Gender                9500 non-null   object 
 2   Blood Pressure        9500 non-null   float64
 3   Cholesterol Level     9500 non-null   float64
 4   Exercise Habits       9500 non-null   object 
 5   Smoking               9500 non-null   object 
 6   Family Heart Disease  9500 non-null   object 
 7   Diabetes              9500 non-null   object 
 8   BMI                   9500 non-null   float64
 9   High Blood Pressure   9500 non-null   object 
 10  Low HDL Cholesterol   9500 non-null   object 
 11  High LDL Cholesterol  9500 non-null   object 
 12  Alcohol Consumption   9500 non-null   object 
 13  Stress Level          9500 non-null   object 
 14  Sleep Hours           9500 non-null   float64
 15  Sugar Consumption     9500 non-null   object 
 16  Triglyceride Level    9500 non-null   float64
 17  Fasting Blood Sugar   9500 non-null   float64
```

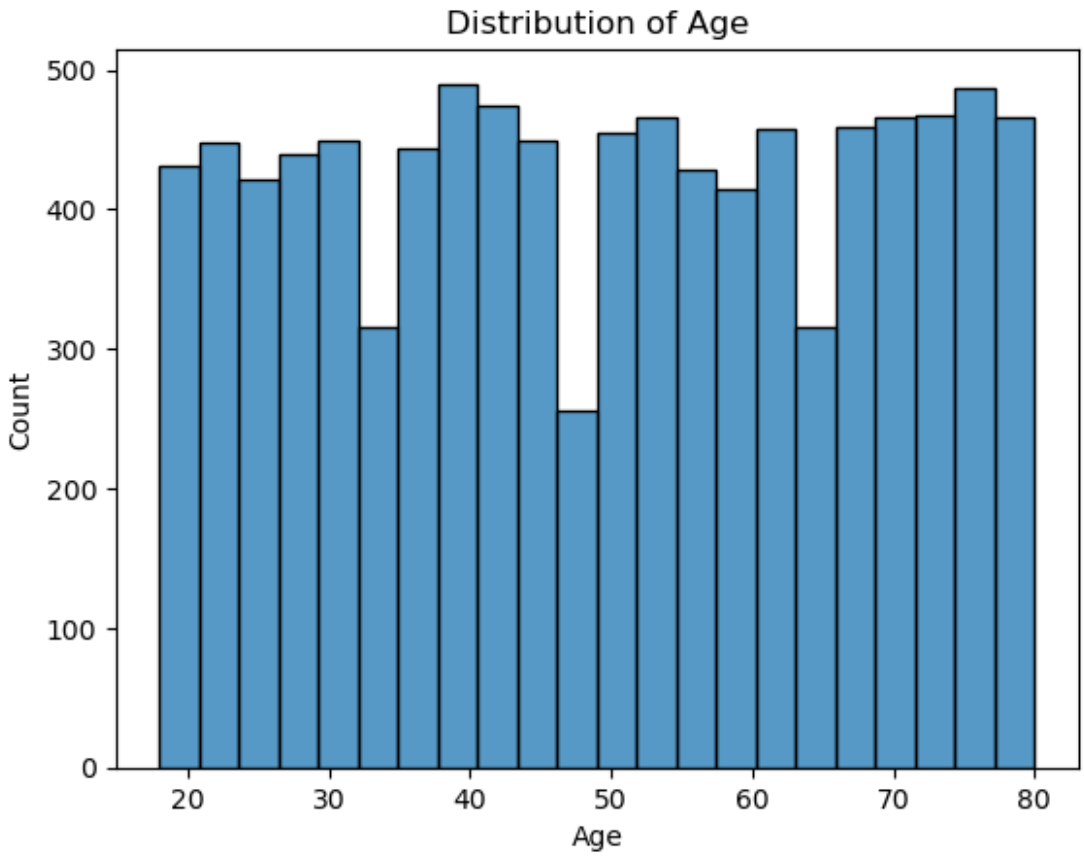
```
 18  CRP Level              9500 non-null   float64
 19  Homocysteine Level     9500 non-null   float64
 20  Heart Disease Status   9500 non-null   object
dtypes: float64(9), object(12)
memory usage: 1.6+ MB
```

```
df[['Age','Cholesterol Level','Sleep Hours']].describe()
```

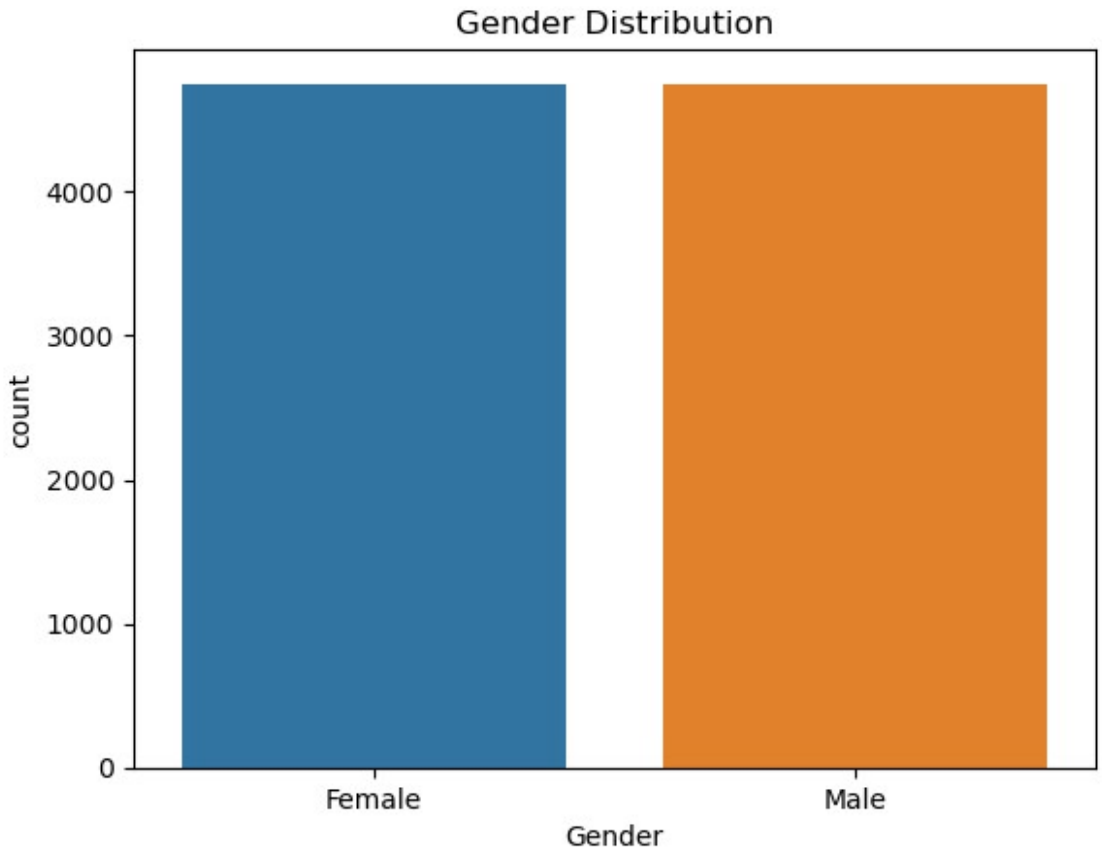
```
                Age  Cholesterol Level  Sleep Hours
count  9500.000000        9500.000000  9500.000000
mean     49.343579         225.295474     6.987953
std      18.213004          43.613240     1.752257
min      18.000000         150.000000     4.000605
25%      34.000000         187.000000     5.448099
50%      49.000000         225.000000     6.998945
75%      65.000000         263.000000     8.531765
max      80.000000         300.000000     9.999952
```

```
data_cleaned = df.dropna()
```

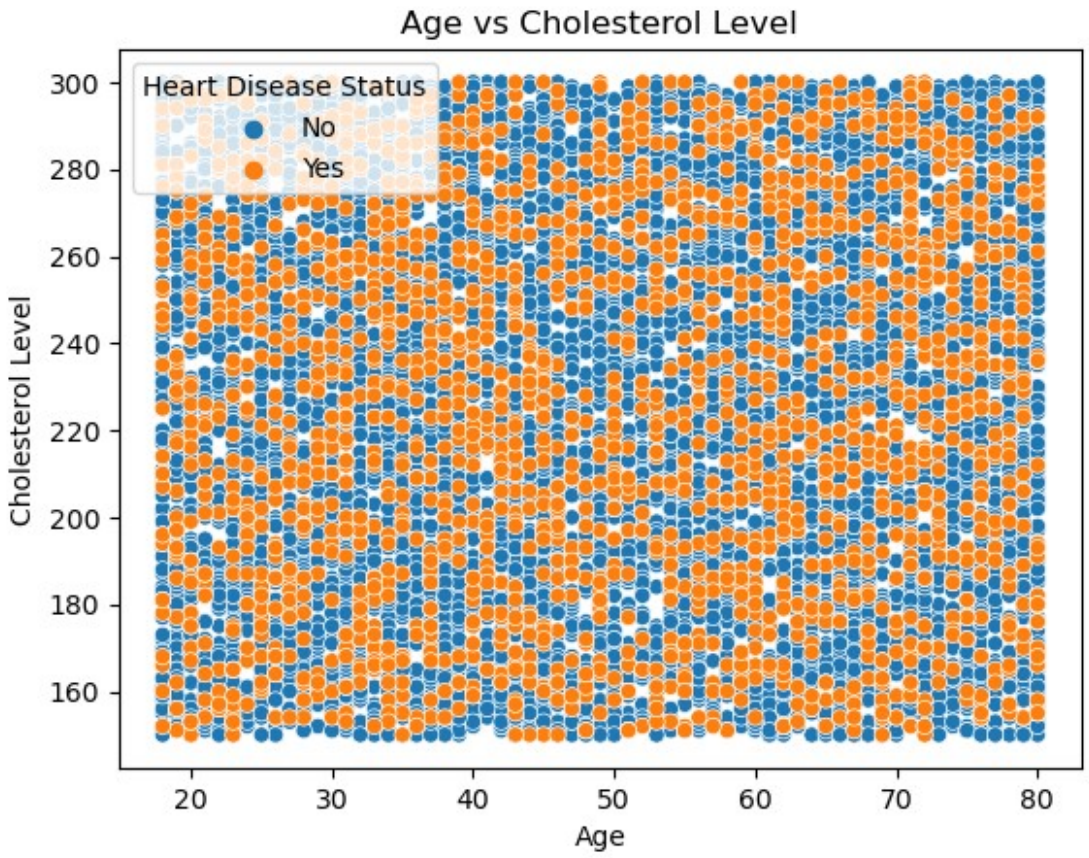
```
import matplotlib.pyplot as plt
import seaborn as sns

# Distribution of Age
sns.histplot(df['Age'])
plt.title('Distribution of Age')
plt.show()
```



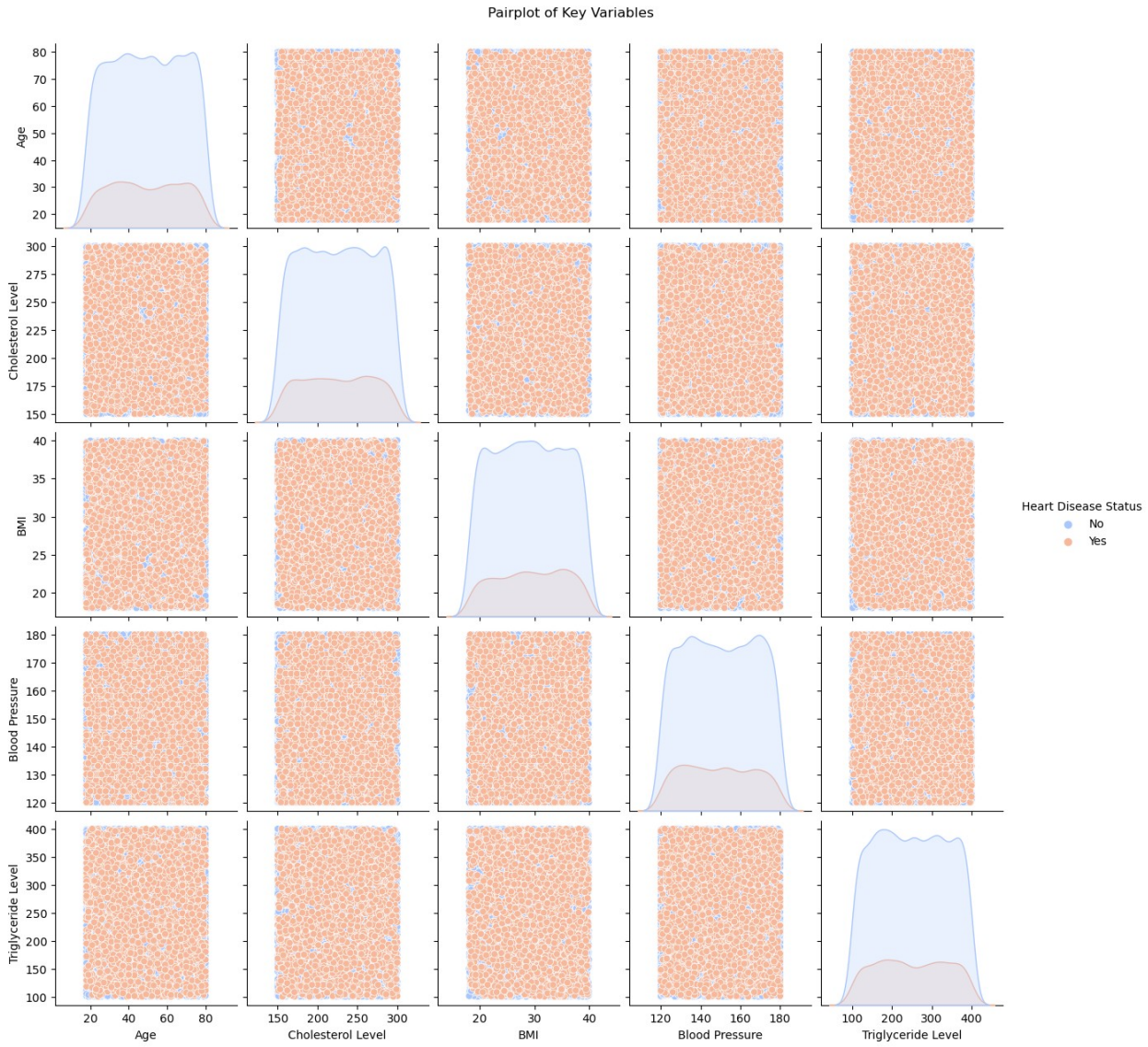

```
sns.countplot(x='Gender', data=df)
plt.title('Gender Distribution')
plt.show()
```



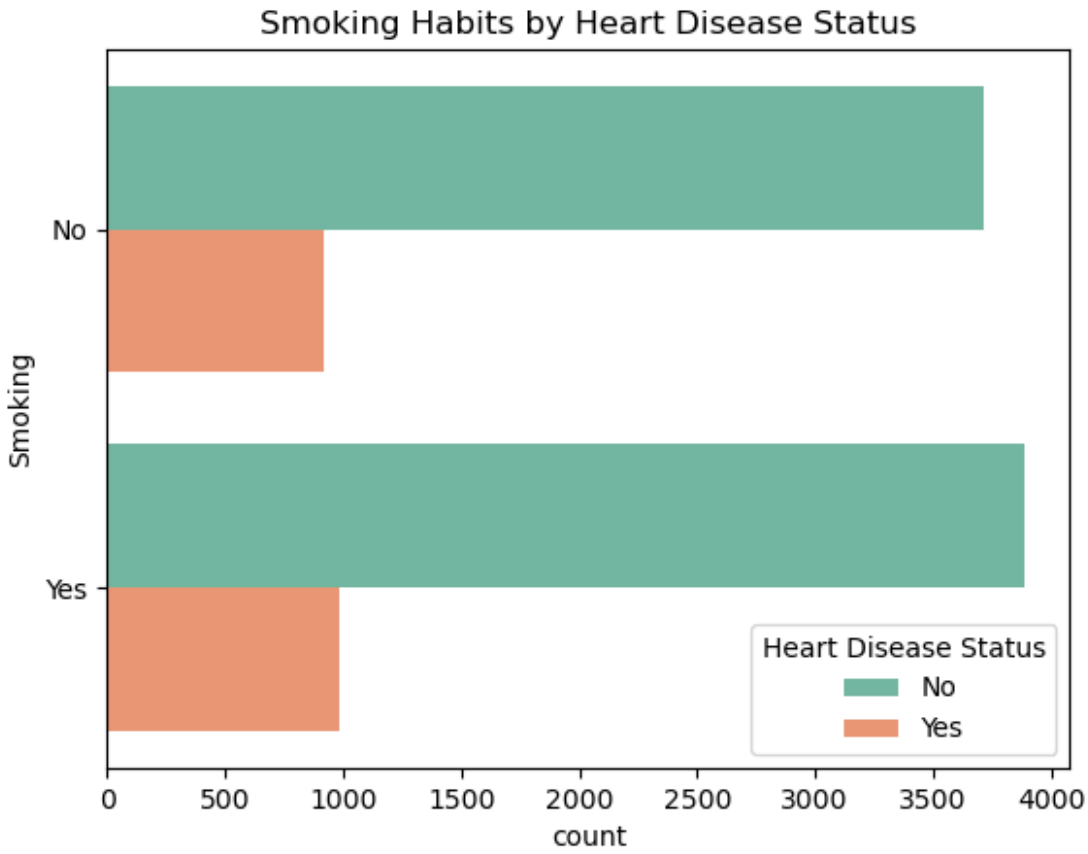

```
sns.scatterplot(x='Age', y='Cholesterol Level', hue='Heart Disease
Status', data=df)
plt.title('Age vs Cholesterol Level')
plt.show()
```



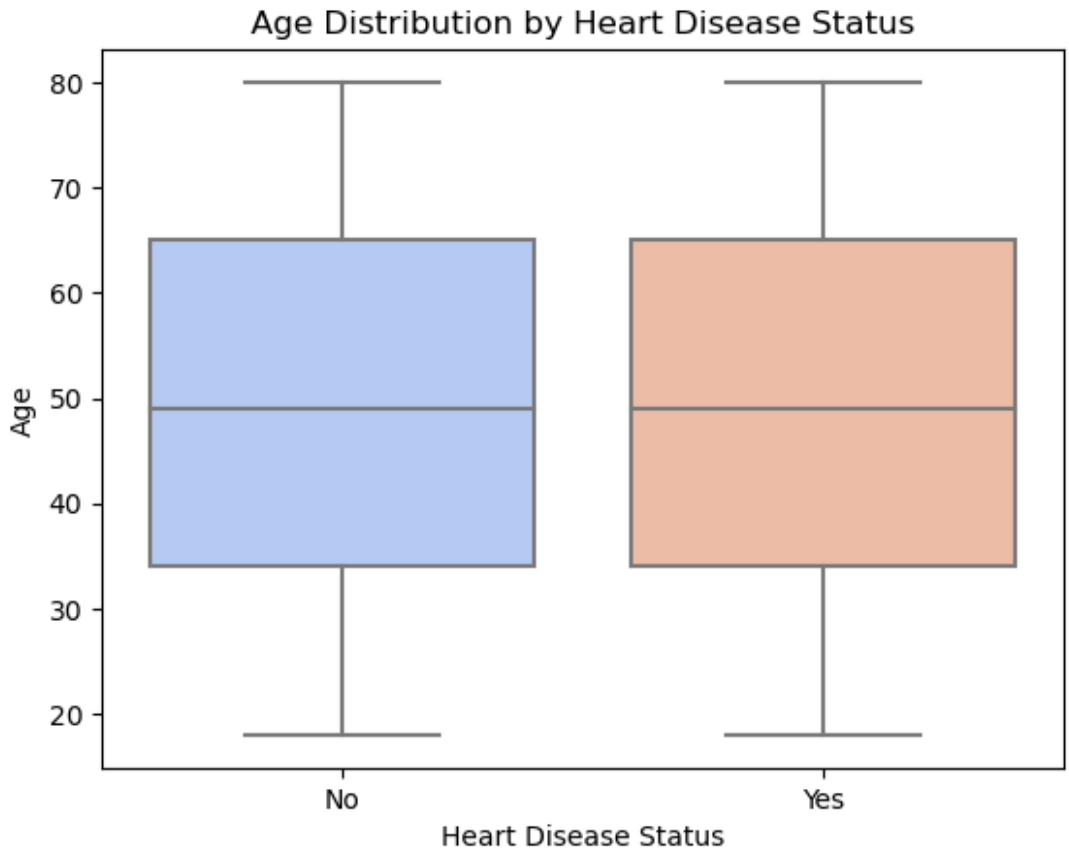

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.pairplot(df,
             vars=['Age', 'Cholesterol Level', 'BMI', 'Blood
Pressure', 'Triglyceride Level'],
             hue='Heart Disease Status',
             diag_kind='kde',
             palette='coolwarm')
plt.suptitle('Pairplot of Key Variables', y=1.02)
plt.show()
```



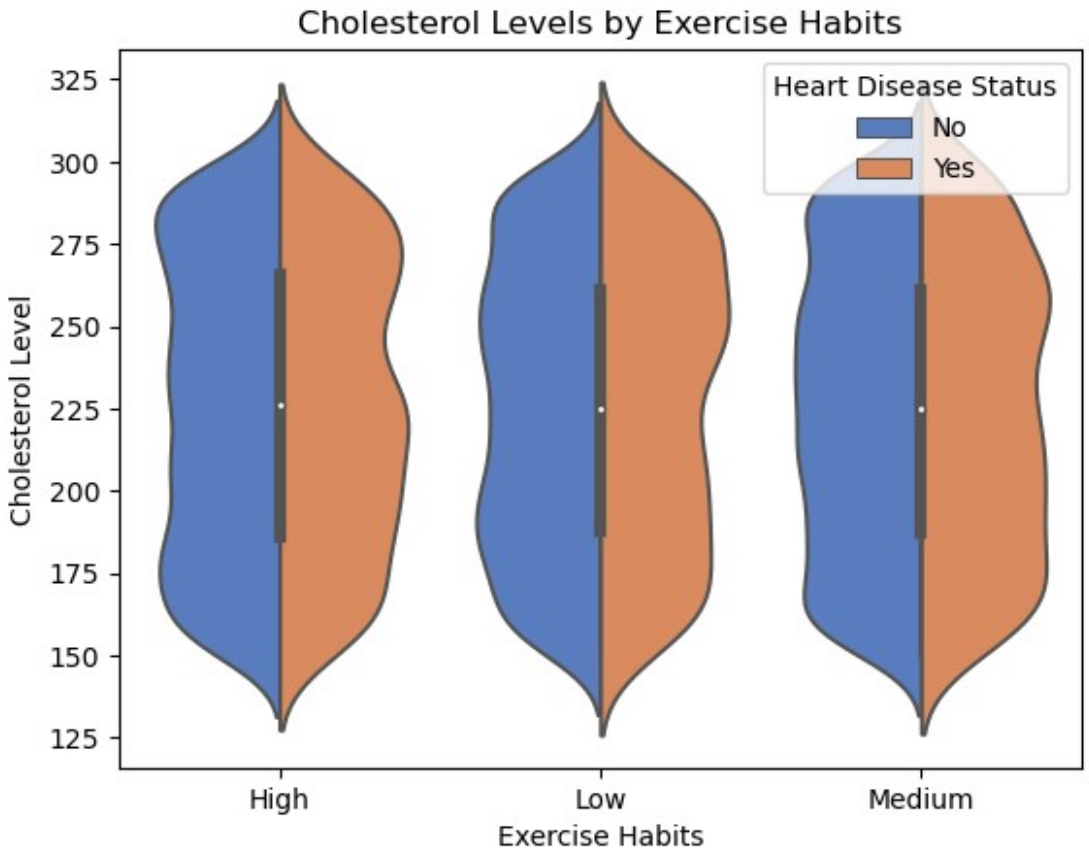

```
sns.countplot(y='Smoking', data=df, hue='Heart Disease Status',
palette='Set2')
plt.title('Smoking Habits by Heart Disease Status')
plt.show()
```



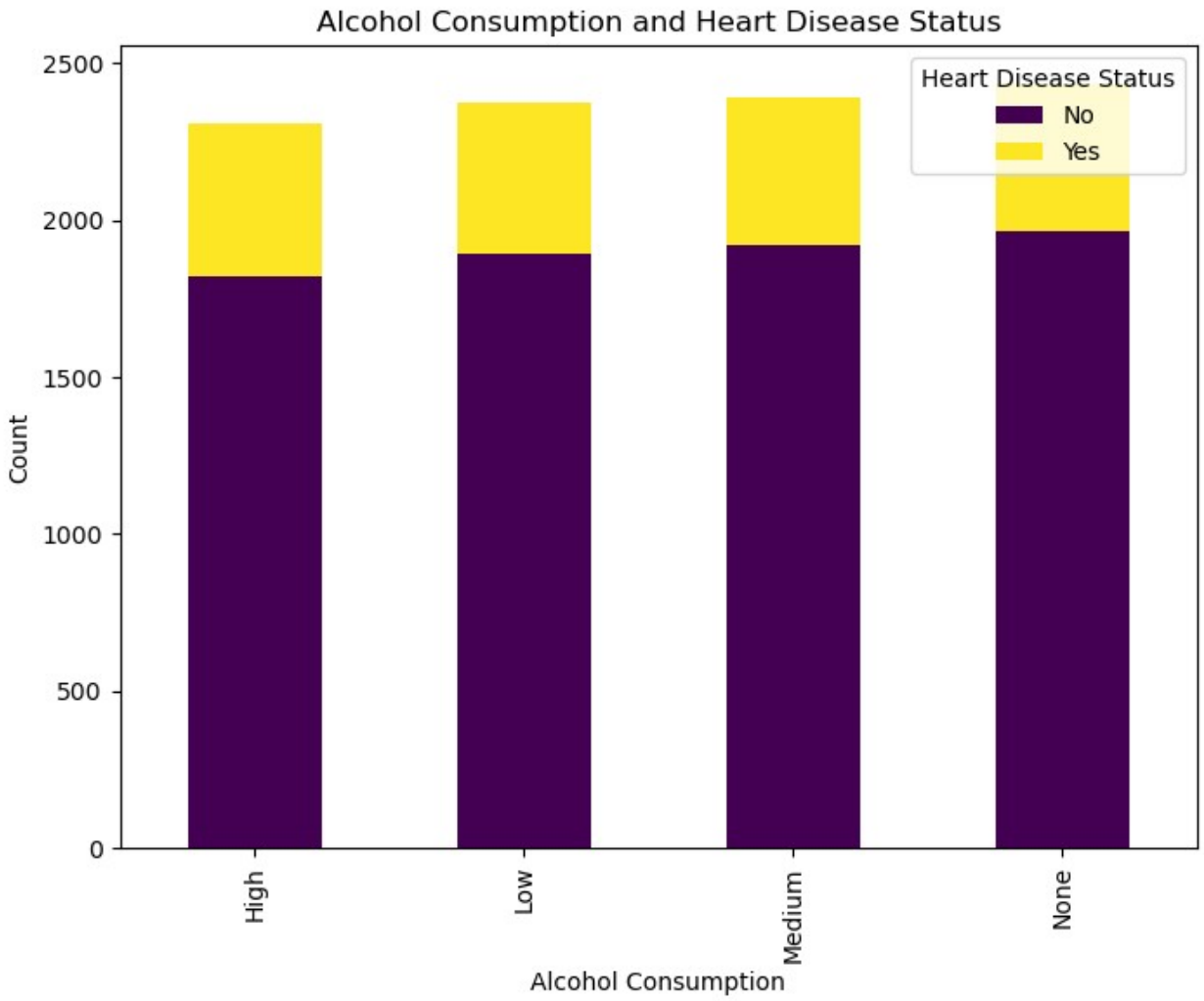

```
sns.boxplot(x='Heart Disease Status', y='Age', data=data_cleaned,
palette='coolwarm')
plt.title('Age Distribution by Heart Disease Status')
plt.show()
```



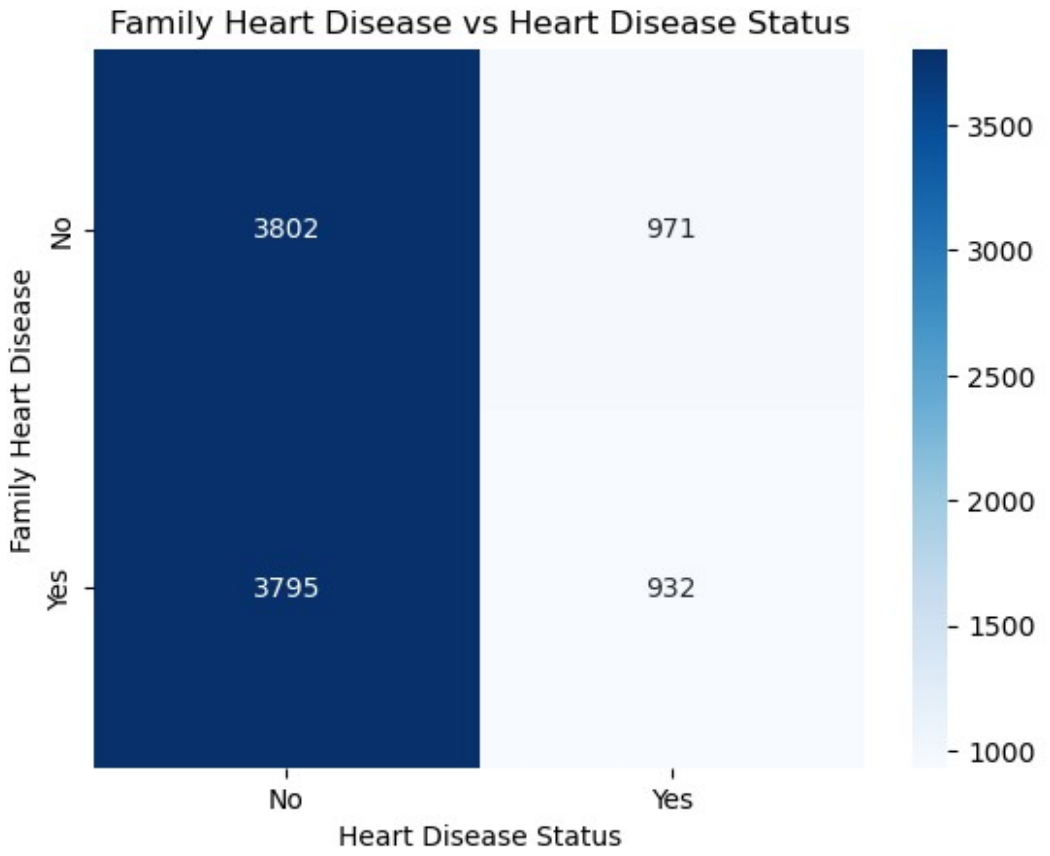

```
sns.violinplot(x='Exercise Habits', y='Cholesterol Level', hue='Heart
Disease Status',
               data=df, split=True, palette='muted')
plt.title('Cholesterol Levels by Exercise Habits')
plt.show()
```



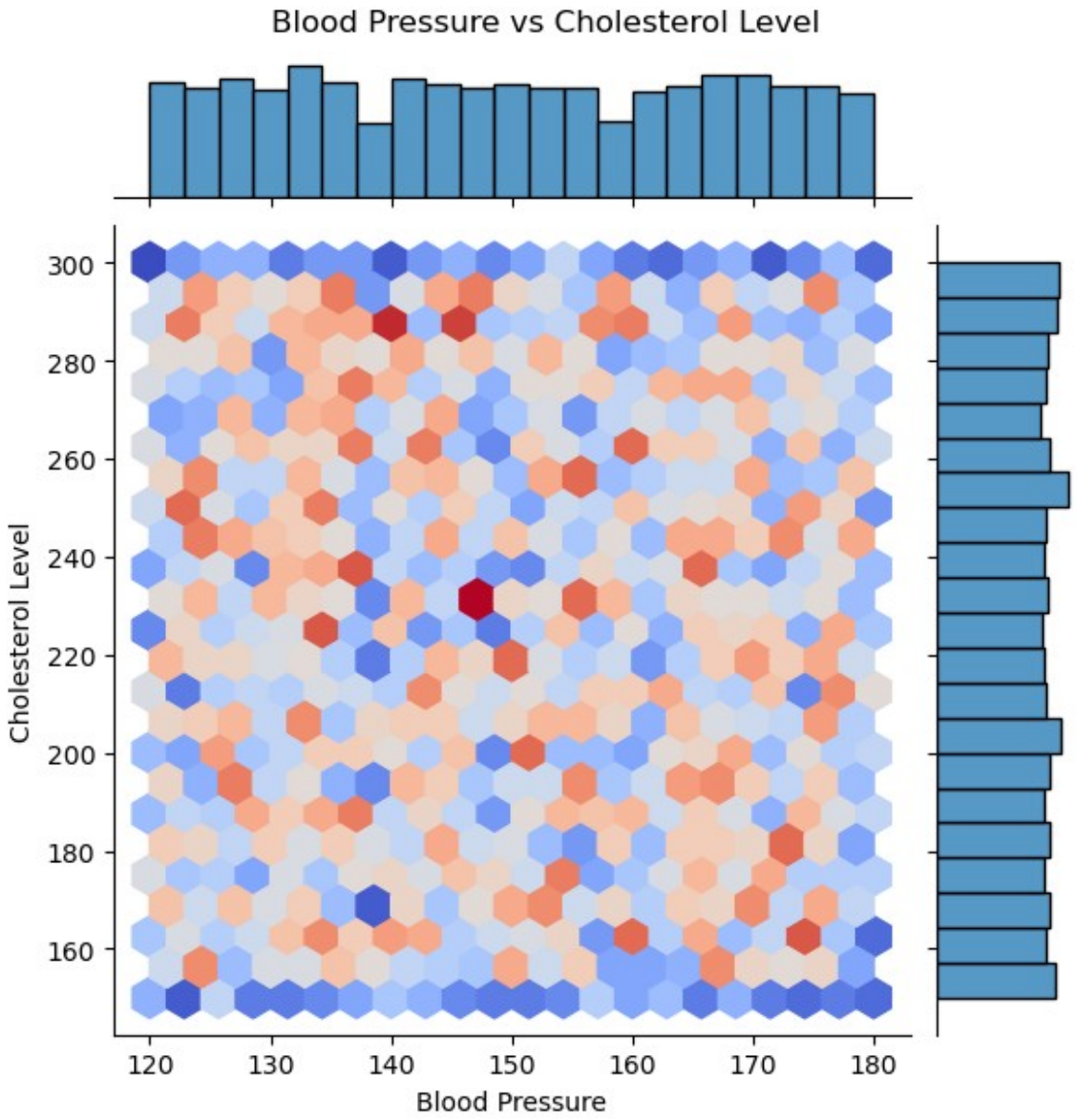

```
alcohol_vs_disease = df.groupby(['Alcohol Consumption', 'Heart Disease
Status']).size().unstack()
alcohol_vs_disease.plot(kind='bar', stacked=True, figsize=(8, 6),
colormap='viridis')
plt.title('Alcohol Consumption and Heart Disease Status')
plt.ylabel('Count')
plt.show()
```



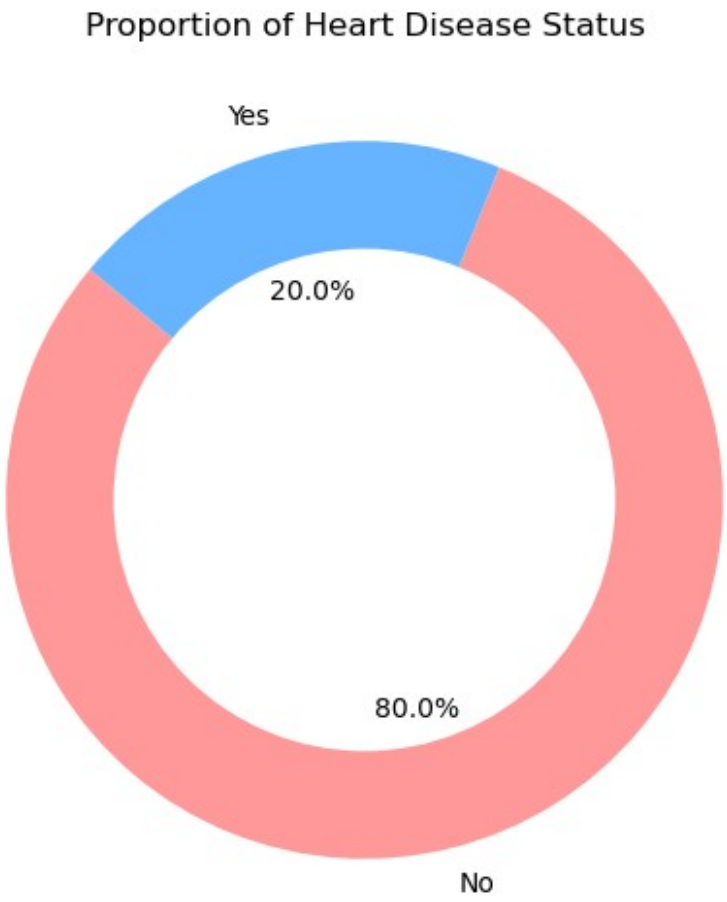

```
cat_relationship = pd.crosstab(df['Family Heart Disease'],
data_cleaned['Heart Disease Status'])
sns.heatmap(cat_relationship, annot=True, fmt='d', cmap='Blues')
plt.title('Family Heart Disease vs Heart Disease Status')
plt.show()
```



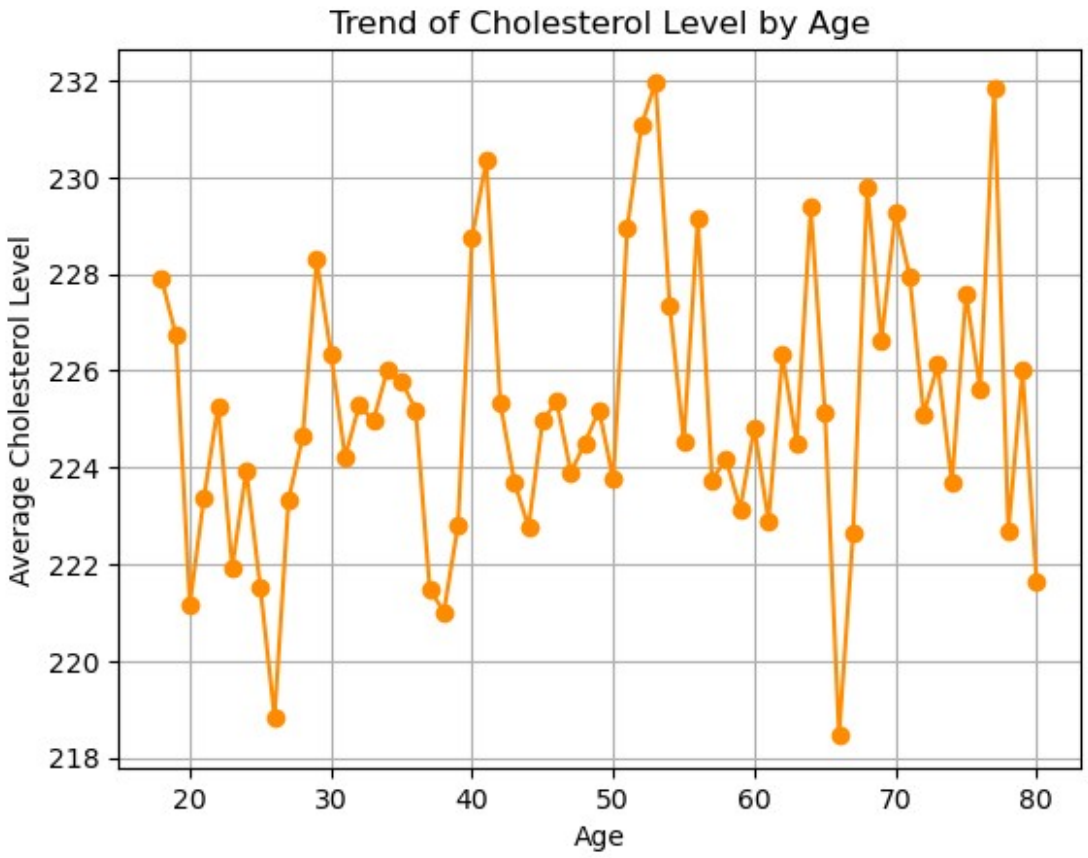

```
sns.jointplot(x='Blood Pressure', y='Cholesterol Level', data=df,
kind='hex', cmap='coolwarm')
plt.suptitle('Blood Pressure vs Cholesterol Level', y=1.02)
plt.show()
```




```
heart_disease_proportions = df['Heart Disease Status'].value_counts()

plt.figure(figsize=(6, 6))
plt.pie(heart_disease_proportions,
labels=heart_disease_proportions.index, autopct='%1.1f%%',
startangle=140,
        colors=['#ff9999','#66b3ff'])

centre_circle = plt.Circle((0, 0), 0.70, fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.title('Proportion of Heart Disease Status')
plt.show()
```




```
age_cholesterol_trend = df.groupby('Age')['Cholesterol Level'].mean()

plt.plot(age_cholesterol_trend.index, age_cholesterol_trend.values,
marker='o', linestyle='-', color='darkorange')
plt.title('Trend of Cholesterol Level by Age')
plt.xlabel('Age')
plt.ylabel('Average Cholesterol Level')
plt.grid(True)
plt.show()
```


Trend of Cholesterol Level by Age
Average Cholesterol Level
232
230
228
226
224
222
220
218
20
30
40
50
60
70
80
Age