

```
!pip install kagglehub
```

```
Requirement already satisfied: kagglehub in /usr/local/lib/python3.12/dist-  
Requirement already satisfied: kagglesdk<1.0,>=0.1.22 in /usr/local/lib/py  
Requirement already satisfied: packaging in /usr/local/lib/python3.12/dist-  
Requirement already satisfied: pyyaml in /usr/local/lib/python3.12/dist-pa  
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-  
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-pack  
Requirement already satisfied: protobuf in /usr/local/lib/python3.12/dist-  
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/  
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/d  
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python  
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python
```

```
import kagglehub  
path = kagglehub.dataset_download("sullivanova/cardiovascular-disease-d  
print("Path to dataset files:", path)
```

Using Colab cache for faster access to the 'cardiovascular-disease-dataset
Path to dataset files: /kaggle/input/cardiovascular-disease-dataset

```
import numpy as np  
import pandas as pd  
from matplotlib import pyplot as plt  
import seaborn as sns  
import warnings  
warnings.filterwarnings("ignore")  
import os  
print(os.listdir(path))
```

```
['cardio_train.csv']
```

```
df = pd.read_csv(path + "/cardio_train.csv", sep=':')
```

```
df.head()
```

	id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;act
0	0;18393;2;168;62.0;110;
1	1;20228;1;156;85.0;140;
2	2;18857;1;165;64.0;130;
3	3;17623;2;169;82.0;150;1
4	4;17474;1;156;56.0;100;

```
df.dtypes
```

0

```
id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio  object
```

```
dtype: object
```

```
df.columns
```

```
Index(['id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco',  
      dtype='object'])
```

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

0

```
id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio  0
```

```
dtype: int64
```

```
import pandas as pd
```

```
# The root cause of the KeyError is the incorrect separator in pd.read_csv  
# For demonstration and immediate fix within this cell, we will reload the data  
# The ideal fix would be to modify cell 7xJXohVr2_0e directly.  
df = pd.read_csv(path + "/cardio_train.csv", sep=';')
```

```
df['age'] = (df['age']/365).round().astype(int)  
df.drop('id',axis=1,inplace=True)
```

```
df.head()
```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco
0	50	2	168	62.0	110	80	1	1	0	0
1	55	1	156	85.0	140	90	3	1	0	0
2	52	1	165	64.0	130	70	3	1	0	0
3	48	2	169	82.0	150	100	1	1	0	0
4	48	1	156	56.0	100	60	1	1	0	0

```

missing_percentage = df.isnull().sum ()

for col in df.select_dtypes(include=np.number).columns:
    # Corrected 'colums' to 'columns'
    if df[col].isnull().sum()>0:
        df[col].fillna(df[col].mean(), inplace = True )
        # Corrected 'true' to 'True'
        print("\ntotal missing values remaining:",df.isnull().sum().sum())

```

```

df = df[(df['ap_hi'] > 0) & (df['ap_hi'] < 250)]
df = df[(df['ap_lo'] > 0) & (df['ap_lo'] < 200)]
df = df[df['ap_hi'] >= df['ap_lo']]
def remove_outlier_iqr(data,column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    return data[(data[column] >= lower) & (data[column] <= upper)]

```

```

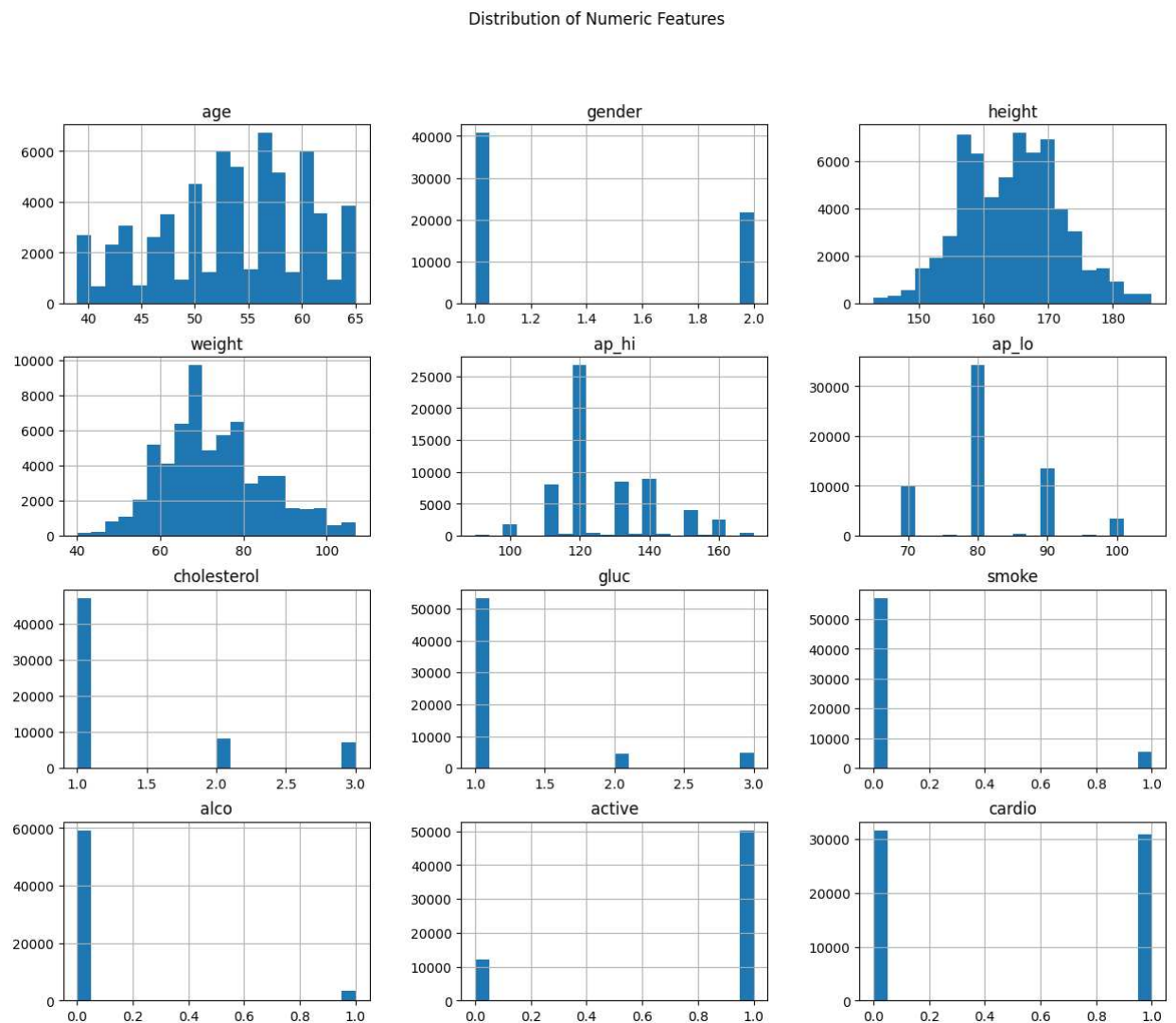
for col in ['age','height','weight','ap_hi','ap_lo']:
    df = remove_outlier_iqr(df,col)
df.head()

```

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco
0	50	2	168	62.0	110	80	1	1	0	0
1	55	1	156	85.0	140	90	3	1	0	0
2	52	1	165	64.0	130	70	3	1	0	0
3	48	2	169	82.0	150	100	1	1	0	0
5	60	1	151	67.0	120	80	2	2	0	0

```
# Fill missing values in columns with mode for column in df
```

```
df.hist(figsize=(15, 12), bins=20)  
plt.suptitle("Distribution of Numeric Features")  
plt.show()
```

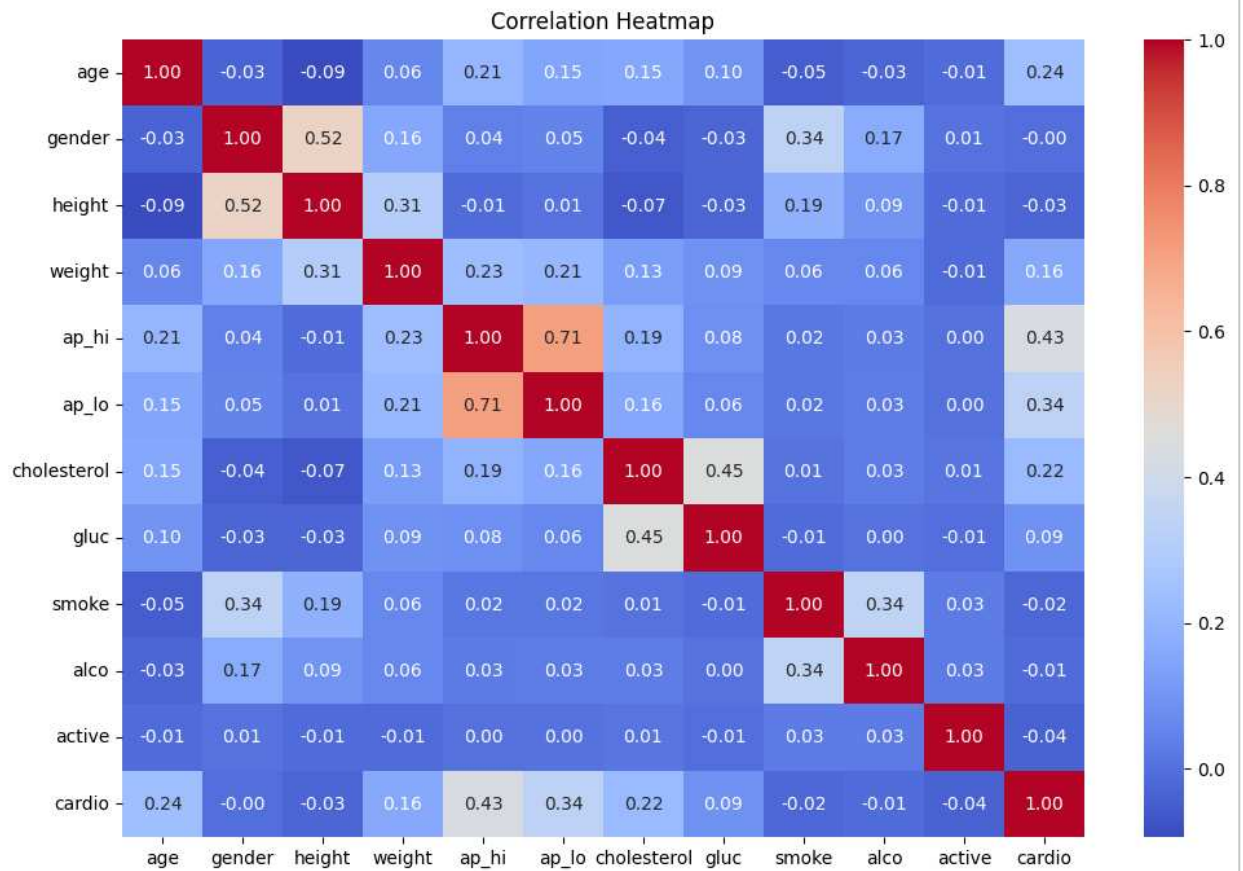


```
df.isnull().sum()
```

	0
age	0
gender	0
height	0
weight	0
ap_hi	0
ap_lo	0
cholesterol	0
gluc	0
smoke	0
alco	0
active	0
cardio	0

dtype: int64

```
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



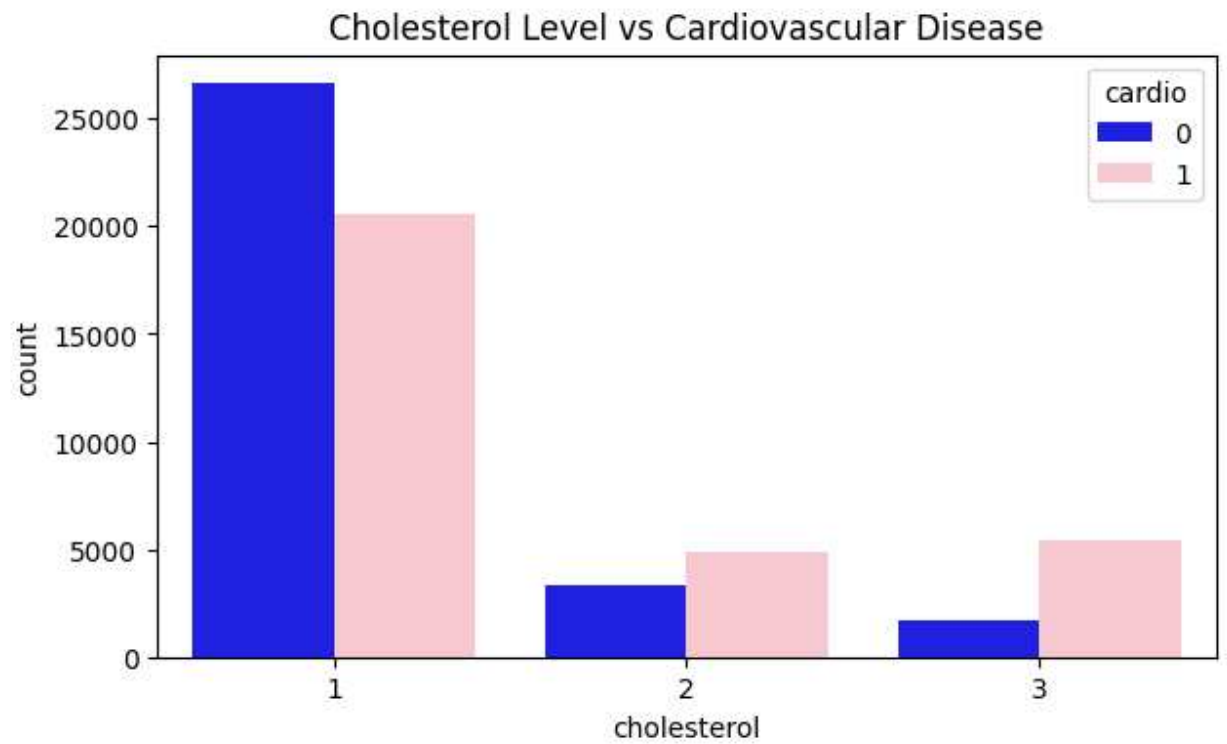
```
df['cardio'].value_counts()
```

```

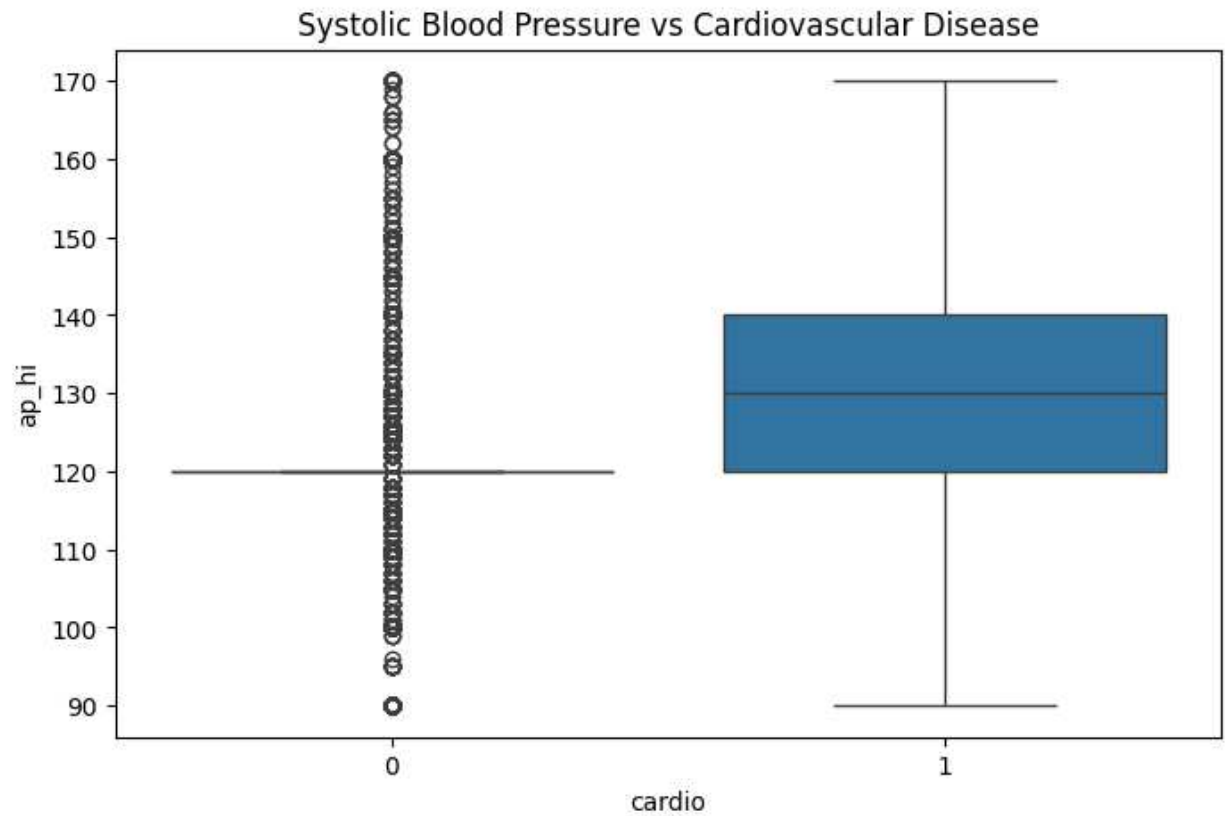
count
cardio
0      31631
1      30869

dtype: int64
```

```
plt.figure(figsize=(7, 4))
sns.countplot(x='cholesterol', hue='cardio', data=df, palette=['blue', 'p
plt.title("Cholesterol Level vs Cardiovascular Disease")
plt.show()
```



```
plt.figure(figsize=(8, 5))  
sns.boxplot(x='cardio', y='ap_hi', data=df)  
plt.title("Systolic Blood Pressure vs Cardiovascular Disease")  
plt.show()
```



```
# distribution of age
plt.figure(figsize=(10, 6))
sns.histplot(df['age'], bins=15, kde=True, color= 'blue')
plt.title('Age Distribution')
plt.xlabel ('Age')
plt.ylabel ('Frequency')
plt.show()
```

