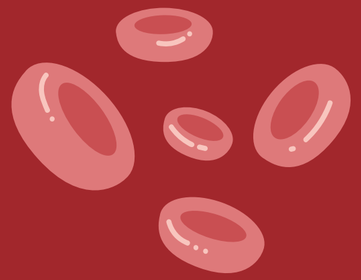
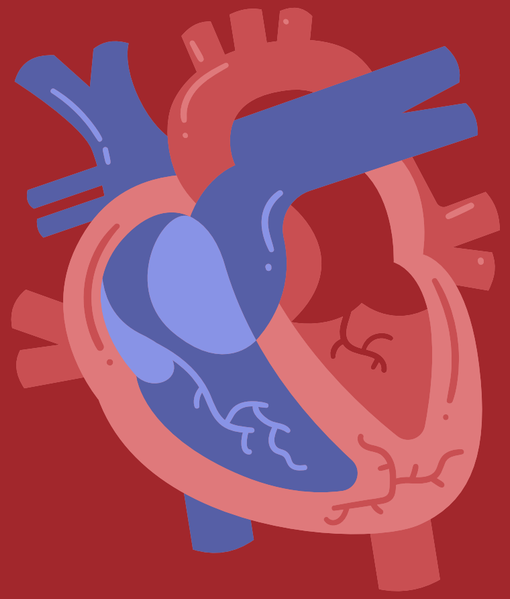
Ann

Final Project





ABOUT HEART DISEASE

Exploring Factors Influencing Health and Lifestyle Choices

BACKGROUND

About **697,000** people in the United States died from heart disease in 2020 - that's **1 in every 5 deaths**.

What are the risk factors for heart disease?

- High blood pressure
- High cholesterol
- Smoking

About half of people in the United States (47%) have at least one of these three risk factors.

DATA CLEANING

```
df.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4238 entries, 0 to 4237
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Gender           4238 non-null   object
 1   age              4238 non-null   int64
 2   education        4133 non-null   object
 3   currentSmoker    4238 non-null   int64
 4   cigsPerDay       4209 non-null   float64
 5   BPMeds           4185 non-null   float64
 6   prevalentStroke  4238 non-null   object
 7   prevalentHyp     4238 non-null   int64
 8   diabetes         4238 non-null   int64
 9   totChol          4188 non-null   float64
 10  sysBP            4238 non-null   float64
 11  diaBP            4238 non-null   float64
 12  BMI              4219 non-null   float64
 13  heartRate        4237 non-null   float64
 14  glucose          3850 non-null   float64
 15  Heart_ stroke    4238 non-null   object
dtypes: float64(8), int64(4), object(4)
memory usage: 529.9+ KB
```

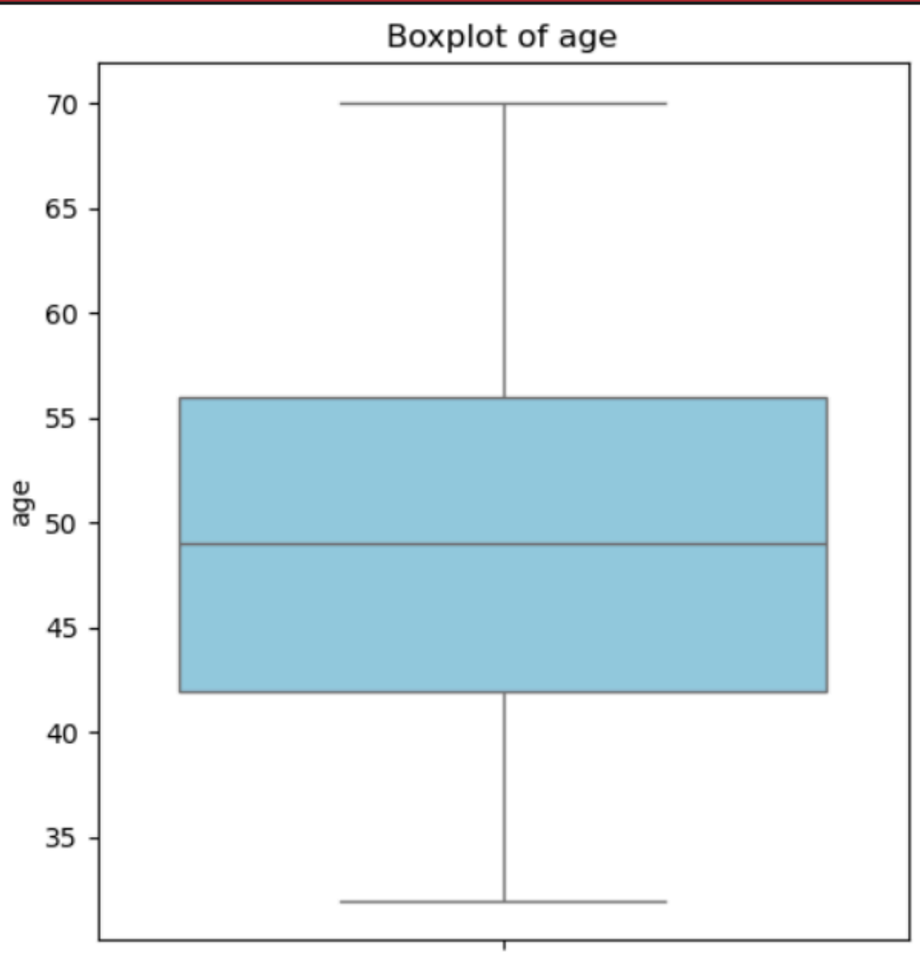
```
df.isnull().sum()
✓ 0.0s
Gender              0
age                 0
education         105
currentSmoker       0
cigsPerDay         29
BPMeds             53
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            50
sysBP               0
diaBP               0
BMI                19
heartRate           1
glucose           388
Heart_ stroke       0
```

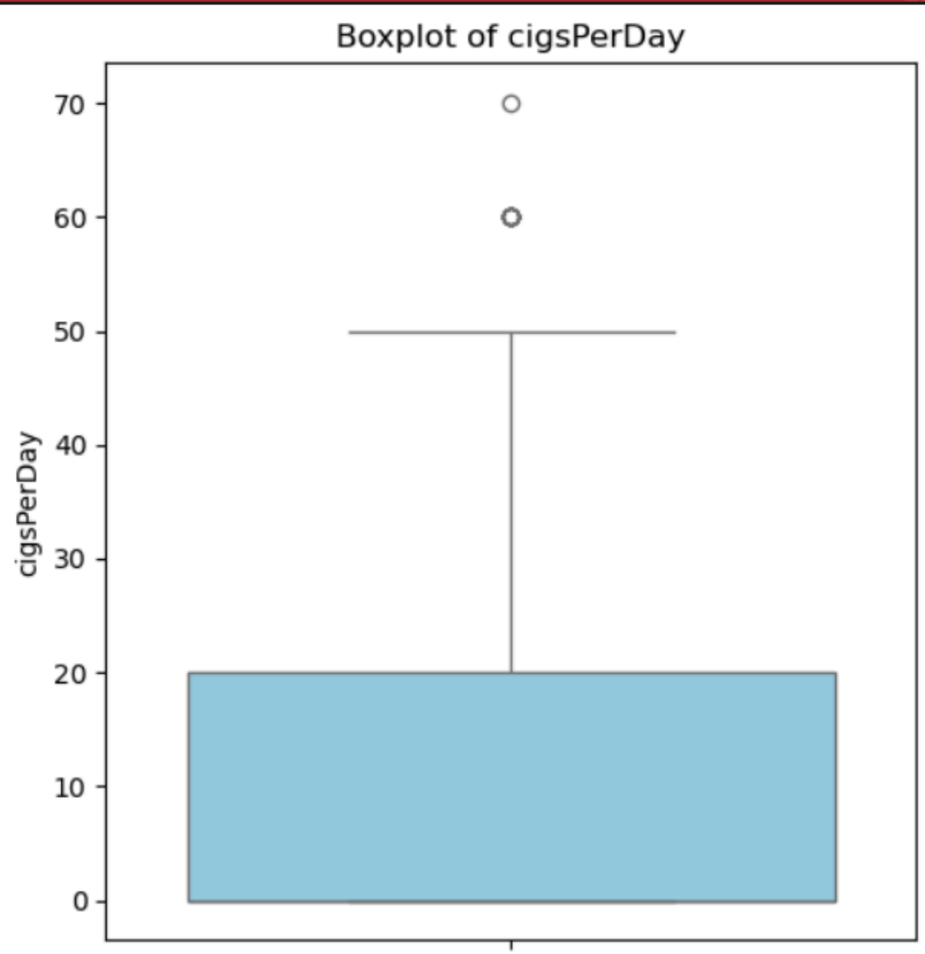
Delete **582** rows, around **13%** of original data

```
df_cleaned = df.dropna()

print(f"Original number of rows：{le
print(f"Number of rows after cleanin
✓ 0.0s
Original number of rows：4238
Number of rows after cleaning：3656
```

EDA (EXPLORATORY DATA ANALYSIS)

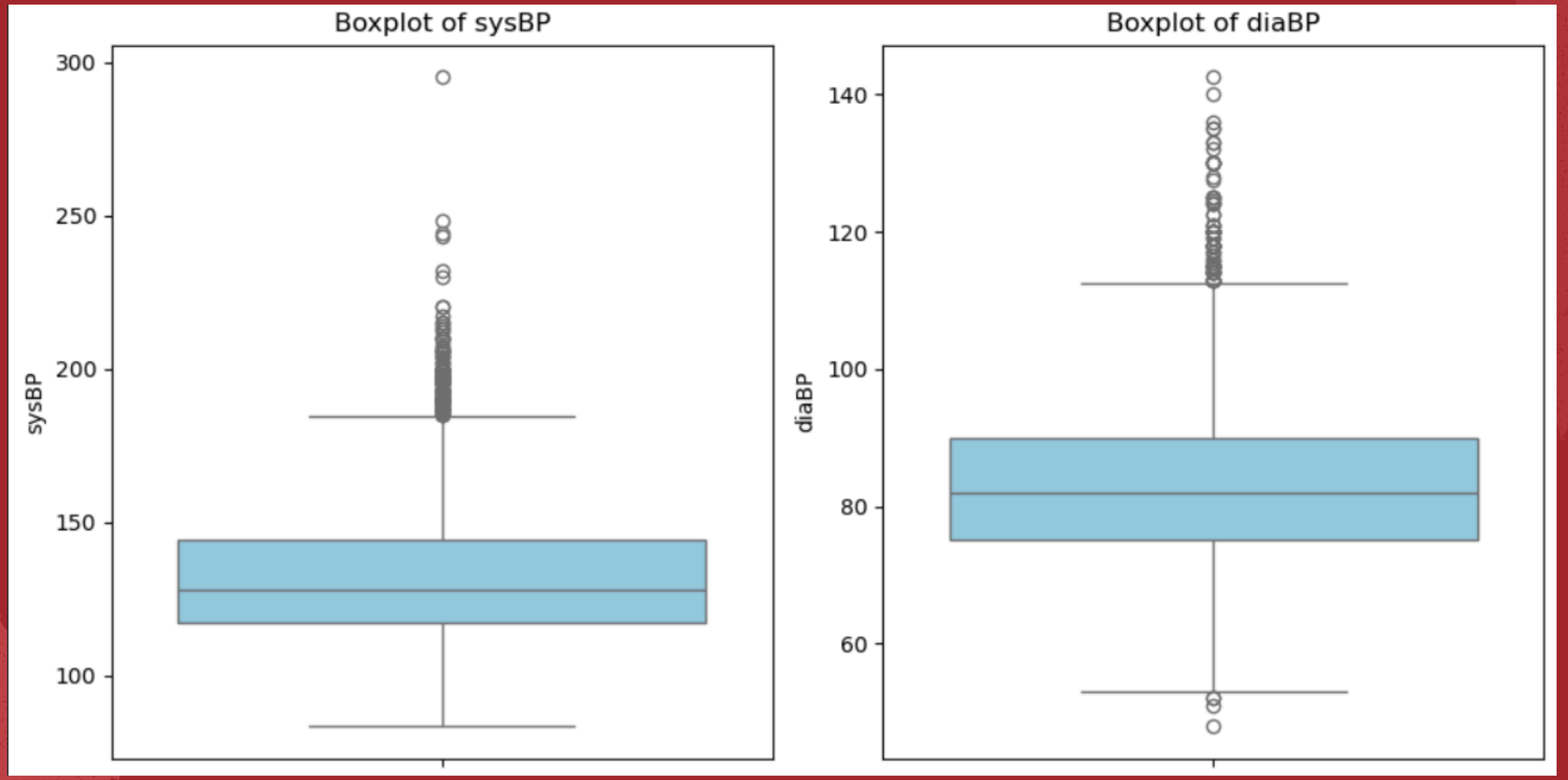




Most people between **40 and 55** years old.

Most people smoking **less than 20** cigarettes per day.

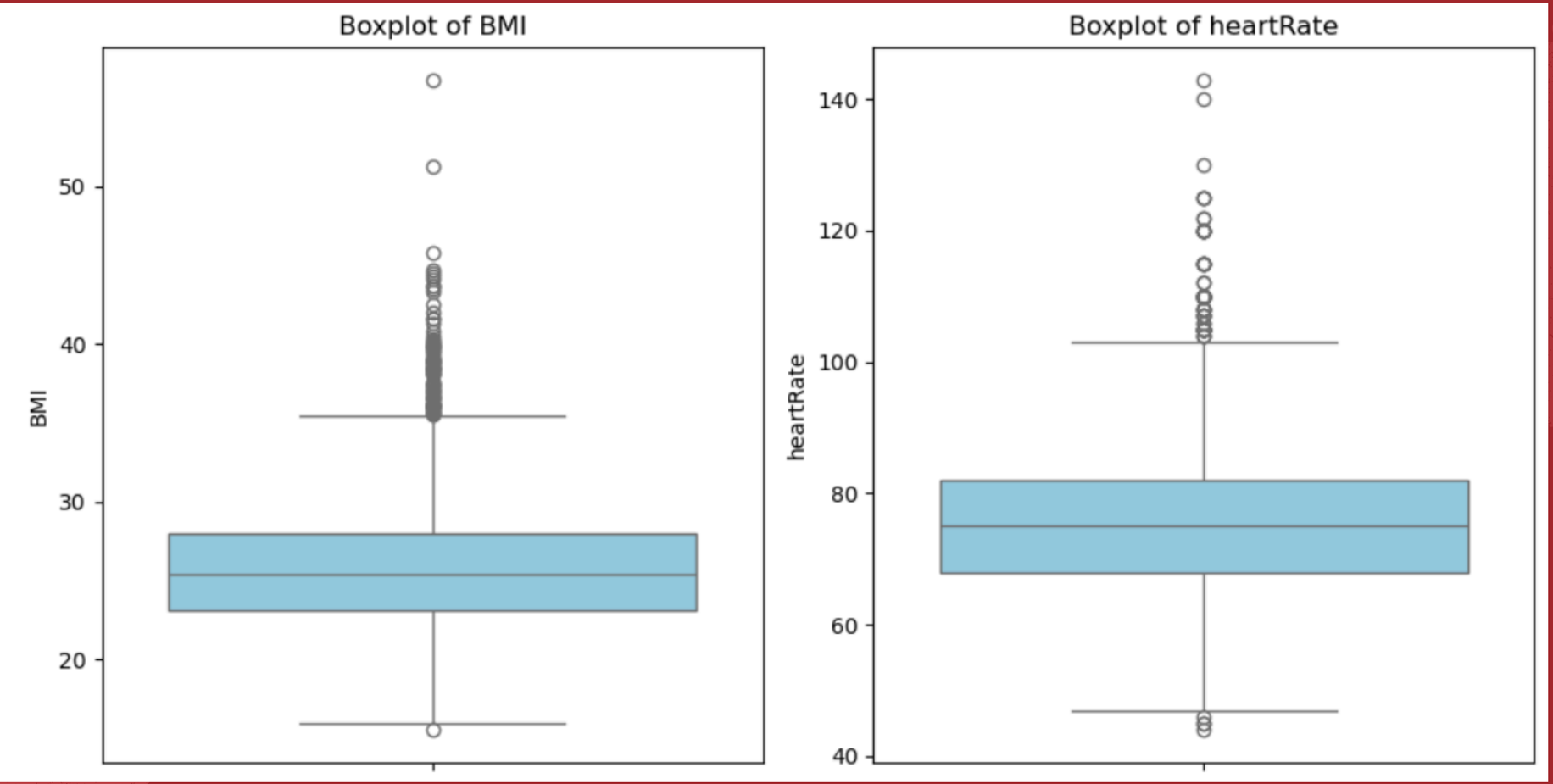
EDA (EXPLORATORY DATA ANALYSIS)



The majority ranging **between 110 and 150** mmHg.

Most values lie **between 75 and 90** mmHg.

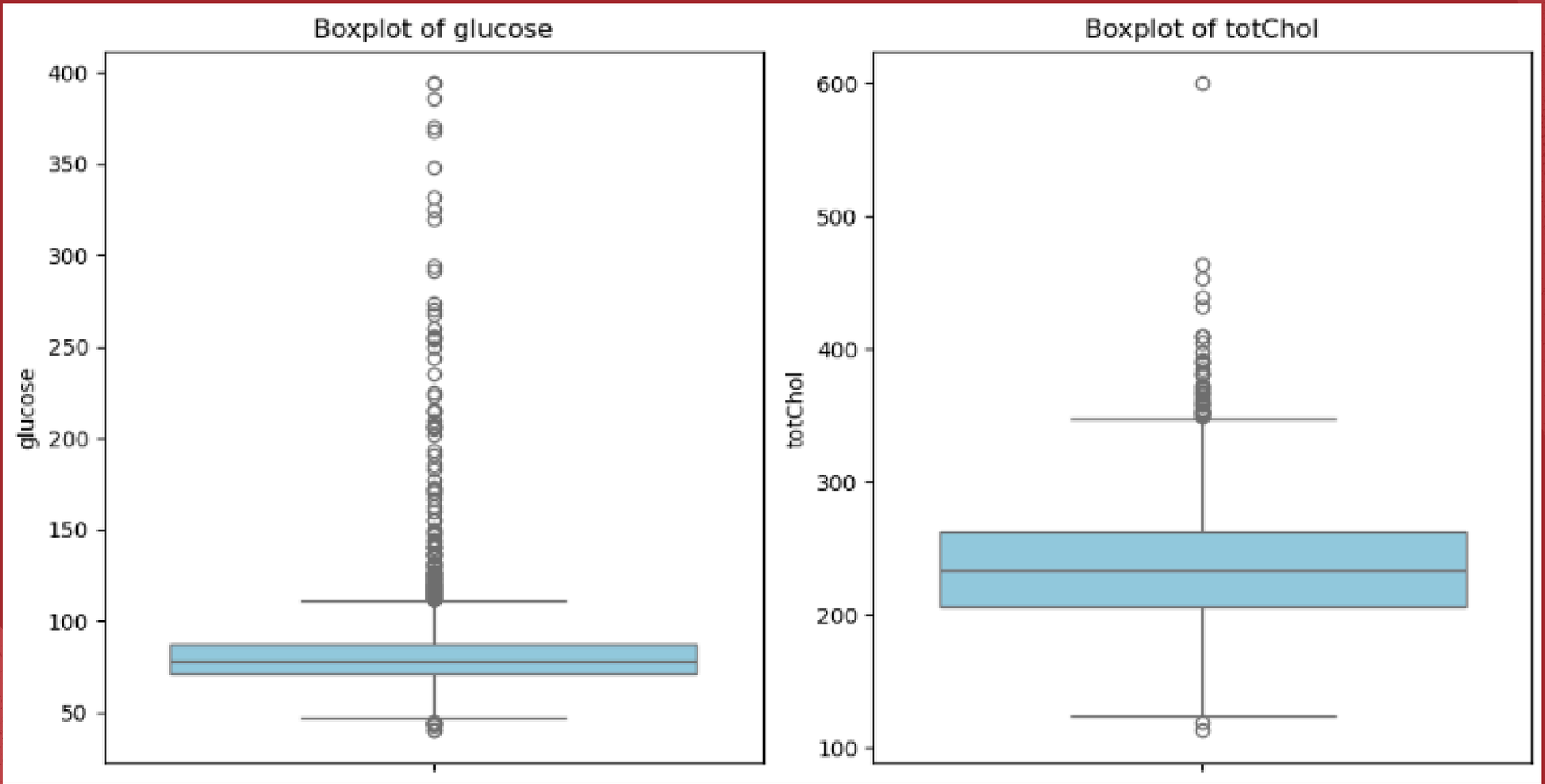
EDA (EXPLORATORY DATA ANALYSIS)



Most people have a BMI **between 23 and 30**

Moderately symmetrical with a few outliers

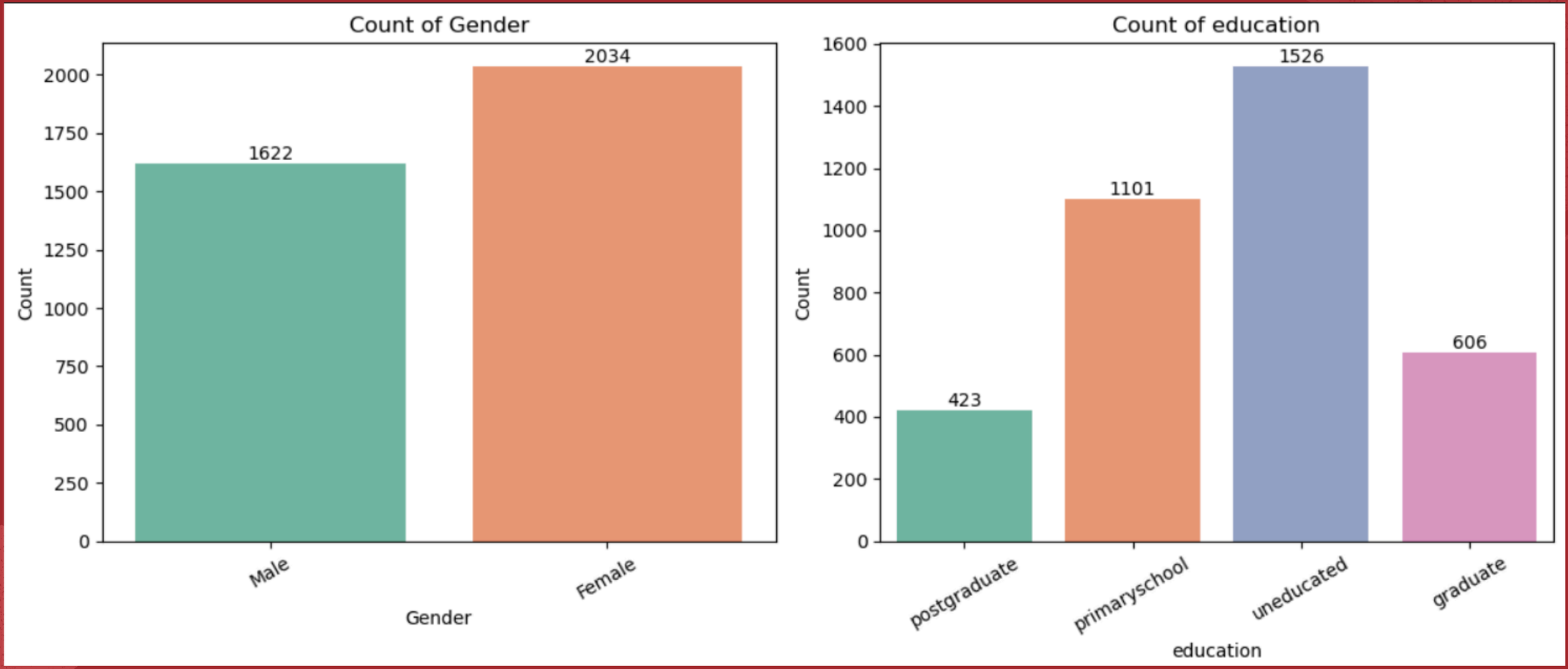
EDA (EXPLORATORY DATA ANALYSIS)



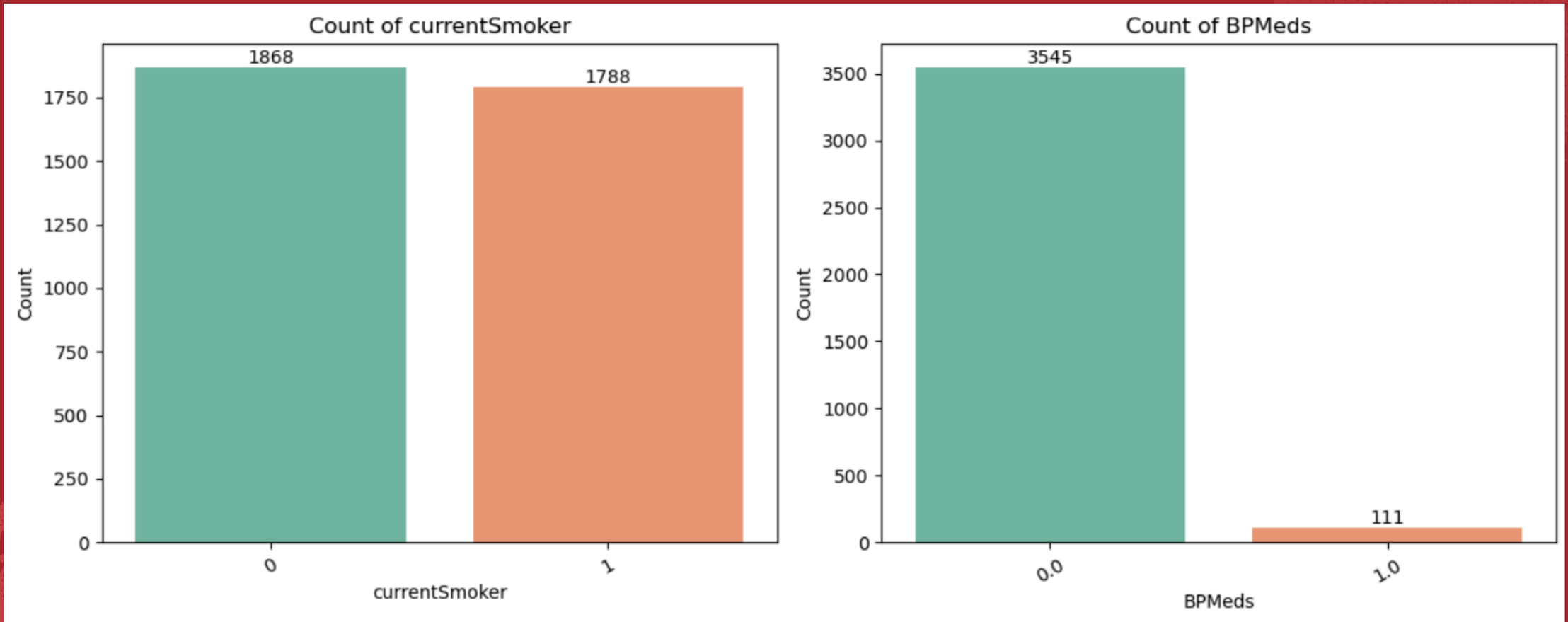
A large number of high outliers, possibly indicating diabetes risk.

Median in totChol Around 240

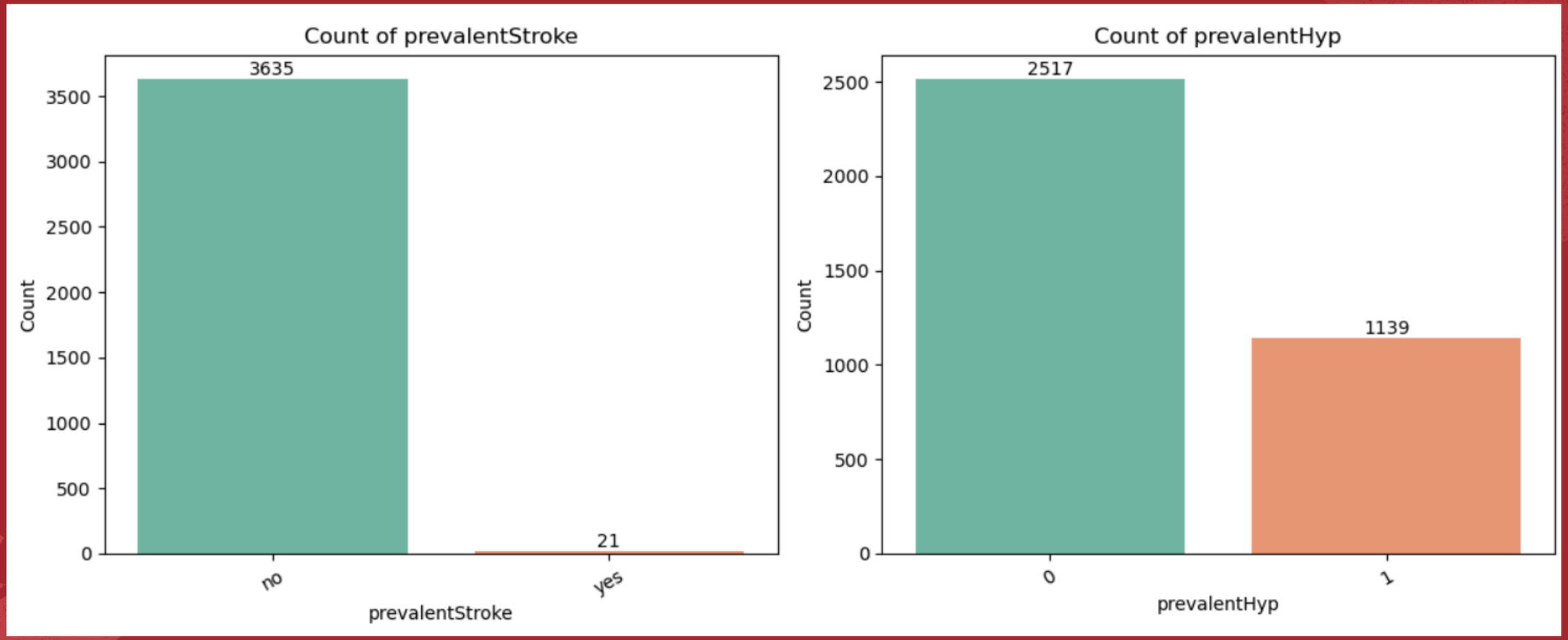
EDA (EXPLORATORY DATA ANALYSIS)



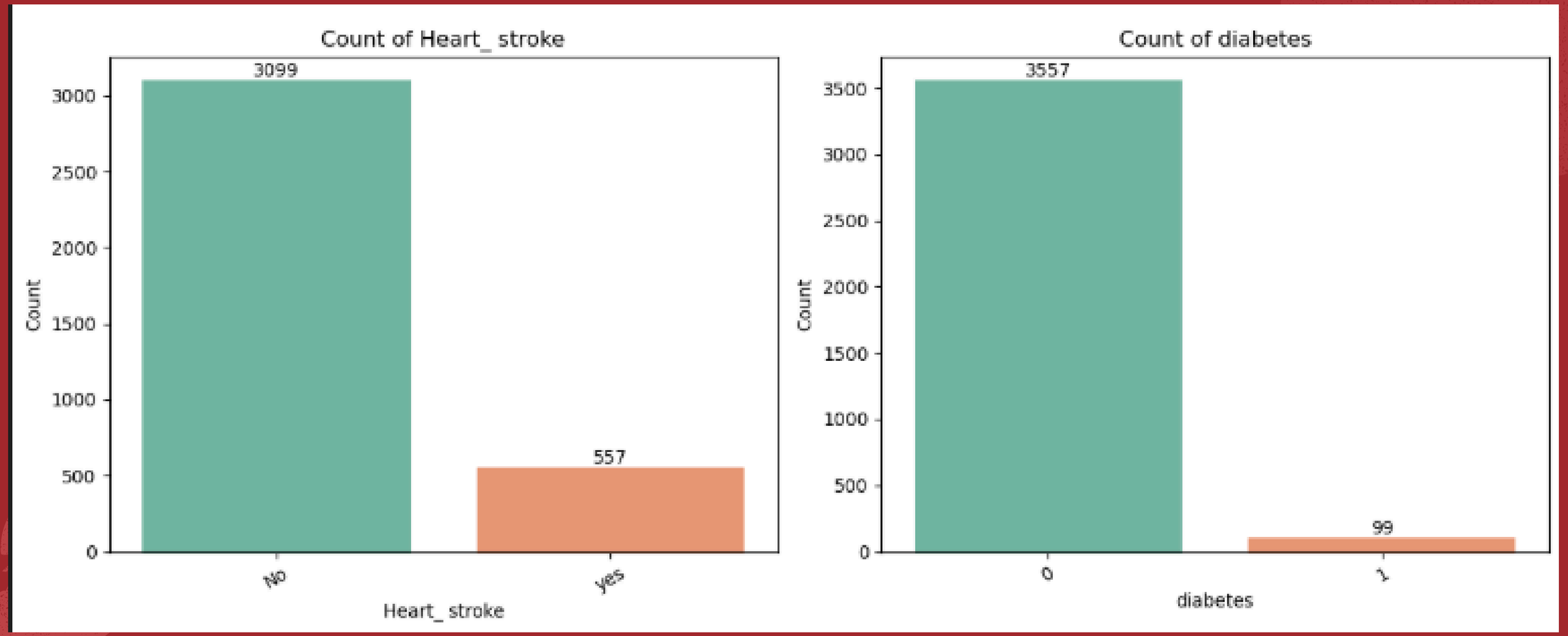
EDA (EXPLORATORY DATA ANALYSIS)



EDA (EXPLORATORY DATA ANALYSIS)



EDA (EXPLORATORY DATA ANALYSIS)



HYPOTHESIS TESTING

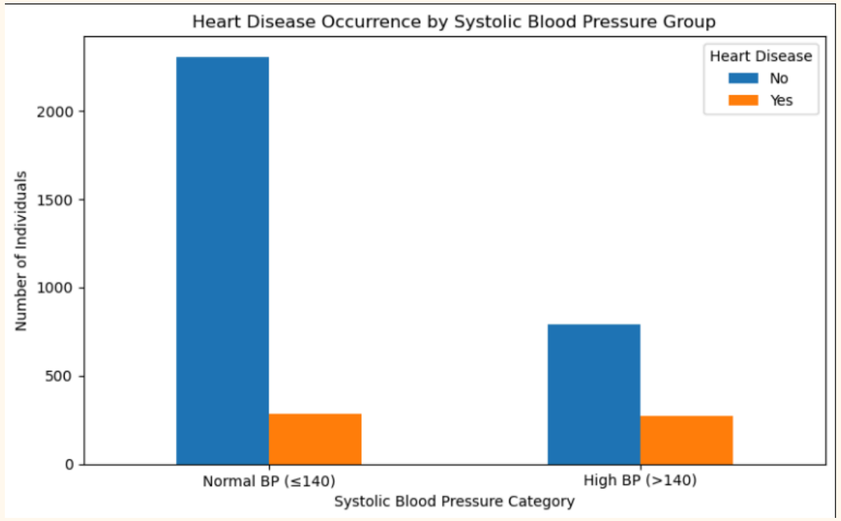
1. According to clinical standards (systolic BP ≥140), are individuals with hypertension more likely to have heart disease? (Chi-Square Test)

1

```
Contingency Table:
Heart_ stroke    No  yes
high_sysBP
0              2307  285
1               792  272

Chi-square statistic: 122.8506
P-value: 0.0000
Degrees of freedom: 1
```

2



Individuals with hypertension **are more likely** to have heart disease and should be considered a high-risk group.

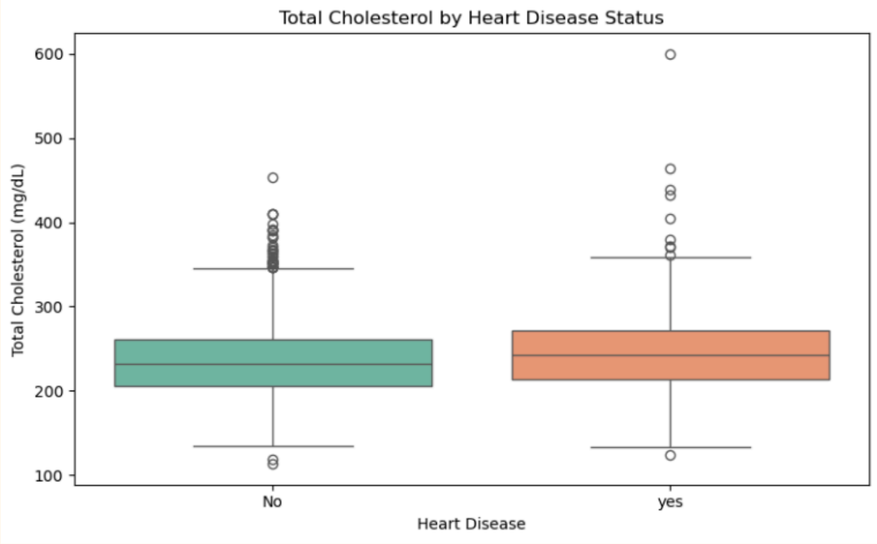
HYPOTHESIS TESTING

2. Is the average total cholesterol level (totChol) higher among individuals with heart disease than those without? (T-test)

1

```
T-statistic: 5.1066
P-value: 0.0000
```

2



People with heart disease tend to have higher cholesterol levels on average, suggesting a possible positive relationship between cholesterol and heart disease risk.

HYPOTHESIS TESTING

3. Do individuals with heart disease smoke more cigarettes per day (on average) compared to those without? (T-test)

1

```
T-statistic: 2.9522
P-value: 0.0033
```

2

Individuals with heart disease tend to smoke more, reinforcing smoking as a potential risk factor for heart disease.

High-risk individuals requiring follow-up

	Gender	age	education	currentSmoker	cigsPerDay	BPMeds	\
327	Male	56	uneducated	1	60.0	0.0	
721	Male	59	uneducated	1	60.0	0.0	
1054	Male	58	primaryschool	1	60.0	0.0	
1452	Male	39	uneducated	1	60.0	0.0	
1468	Male	50	uneducated	1	60.0	0.0	
1488	Male	37	postgraduate	1	60.0	0.0	
1849	Male	48	uneducated	1	60.0	0.0	
2709	Male	46	uneducated	1	60.0	0.0	
3008	Male	40	graduate	1	70.0	0.0	
3928	Male	67	primaryschool	1	60.0	0.0	

	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	\
327	no	0	0	246.0	125.0	79.0	29.64	
721	no	1	0	298.0	153.5	105.0	25.05	
1054	no	1	0	250.0	150.0	97.0	32.00	
1452	no	0	0	215.0	112.0	65.0	23.60	
1468	no	1	0	340.0	134.0	95.0	30.46	
1488	no	0	0	254.0	122.5	82.5	23.87	
1849	no	0	0	252.0	104.0	73.5	23.03	
2709	no	0	0	285.0	121.0	82.0	27.62	
3008	no	1	0	210.0	132.0	86.0	31.57	
3928	no	1	0	261.0	170.0	100.0	22.71	

	heartRate	glucose	Heart_ stroke	high_sysBP	Heart_stroke_binary
327	70.0	85.0	No	0	0
721	70.0	84.0	No	1	0
1054	75.0	65.0	No	1	0
1452	59.0	78.0	No	0	0
1468	85.0	86.0	No	0	0
1488	88.0	83.0	No	0	0
1849	70.0	77.0	No	0	0
2709	70.0	79.0	No	0	0
3008	98.0	80.0	No	0	0
3928	72.0	79.0	yes	1	1

HYPOTHESIS TESTING

4. Is the average blood glucose level (glucose) significantly higher among individuals with heart disease? (T-test)

1

```
T-statistic: 4.6041
P-value: 0.0000
```

2

Individuals with heart disease tend to have higher blood glucose levels, suggesting that high glucose or diabetes may be an important risk factor associated with heart disease.

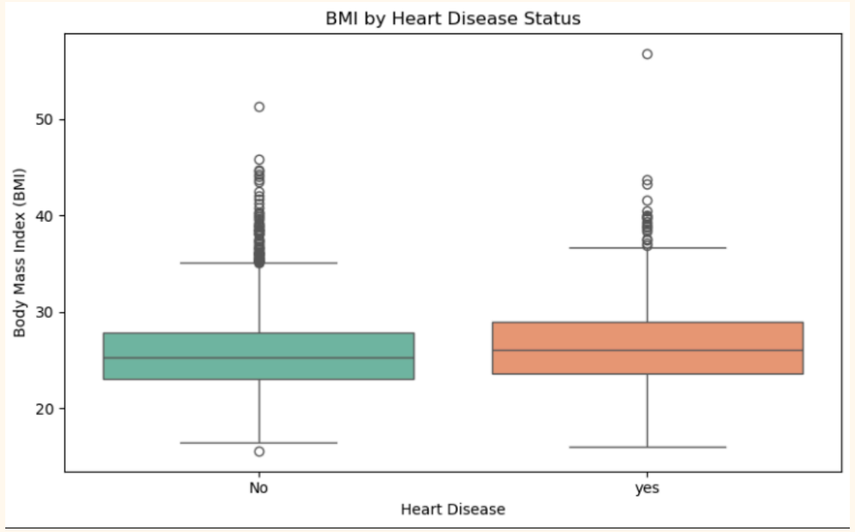
HYPOTHESIS TESTING

5. Is the average body mass index (BMI) higher among individuals with heart disease than those without? (T-test)

1

```
T-statistic: 4.5453
P-value: 0.0000
```

2



There is a **statistically significant association between higher BMI and heart disease**, reinforcing the role of excess weight as a key risk factor.

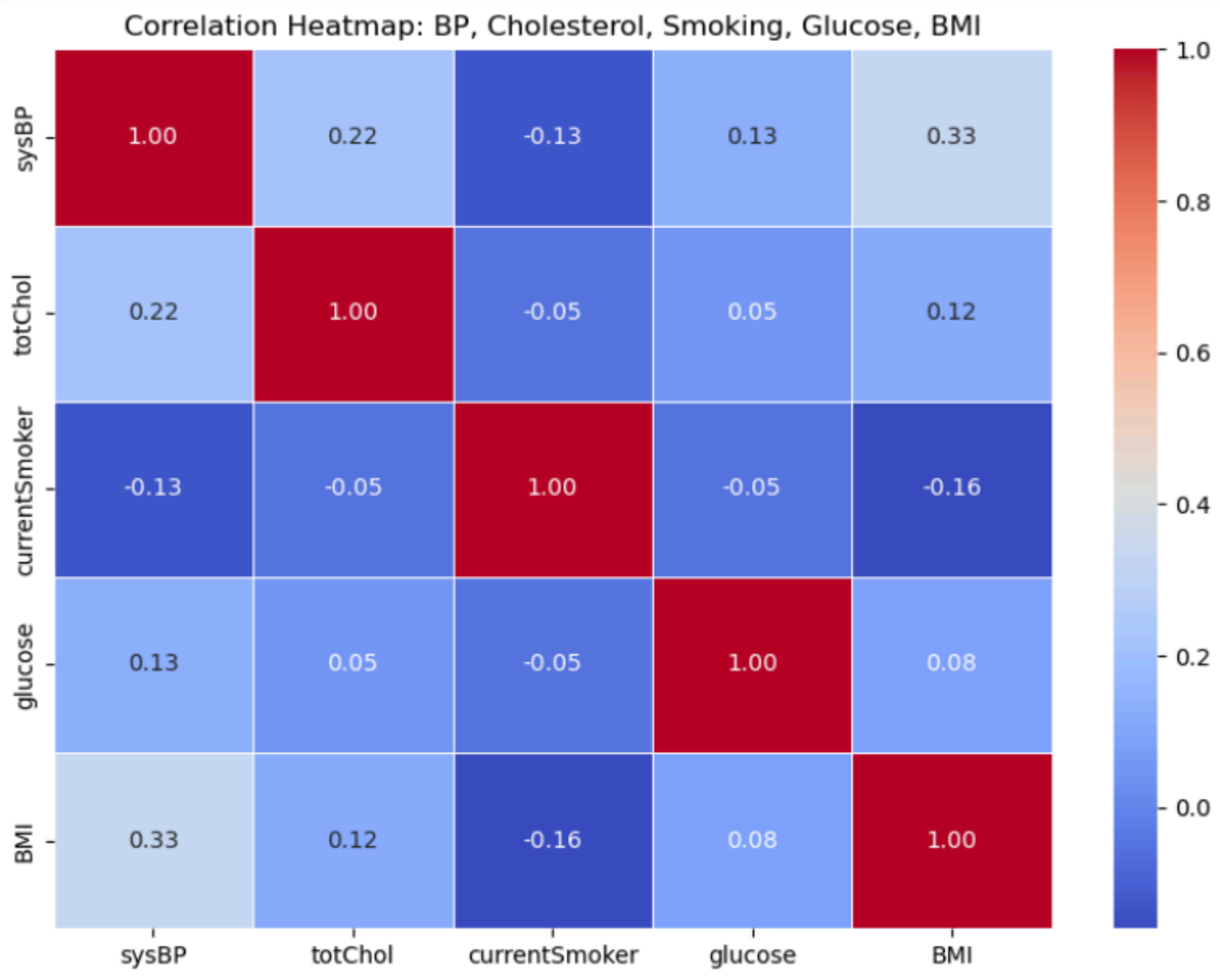
SUMMARY CONCLUSION

- **B** Blood pressure
- **S** Smoking
- **B** Blood glucose
- **B** BMI
- **C** Cholesterol

High blood pressure, cholesterol, smoking, blood glucose, and BMI **are all significantly linked to heart disease**. Targeted prevention strategies and lifestyle interventions focusing on these risk factors are highly recommended to reduce the risk of heart disease.

CORRELATION TESTING

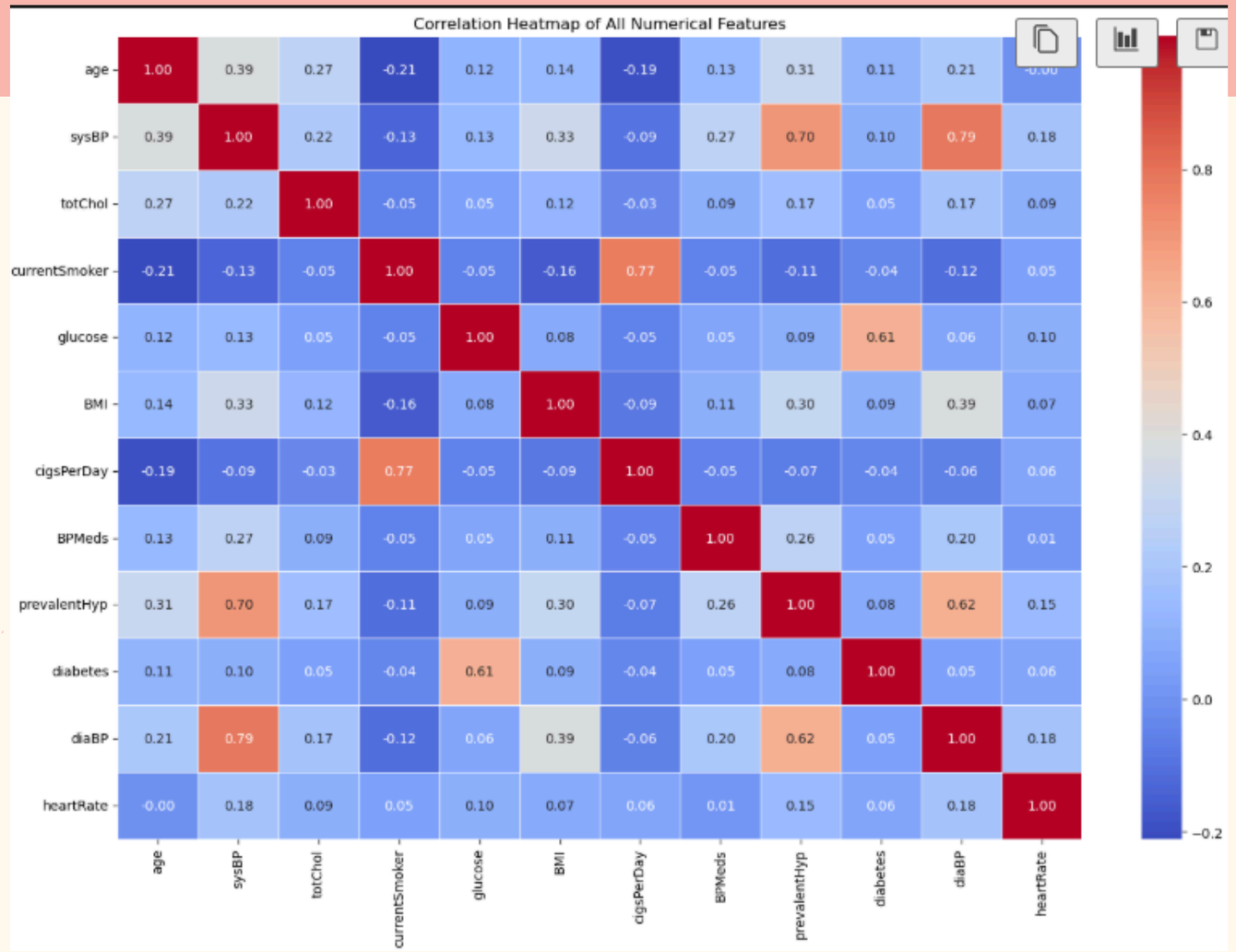
1. Correlation Heatmap: BP, Cholesterol, Smoking, Glucose, BMI



Most of these five health risk factors exhibit weak to moderate linear correlations, suggesting that while they may collectively contribute to health outcomes, **they are not strongly linearly related to each other**.

CORRELATION TESTING

2. Correlation Heatmap of All Numerical Features :



The correlation of 0.61 between glucose and total cholesterol indicates a moderately strong positive linear relationship.

- **Metabolic Syndrome Link : Both glucose and cholesterol are part of metabolic health markers.**

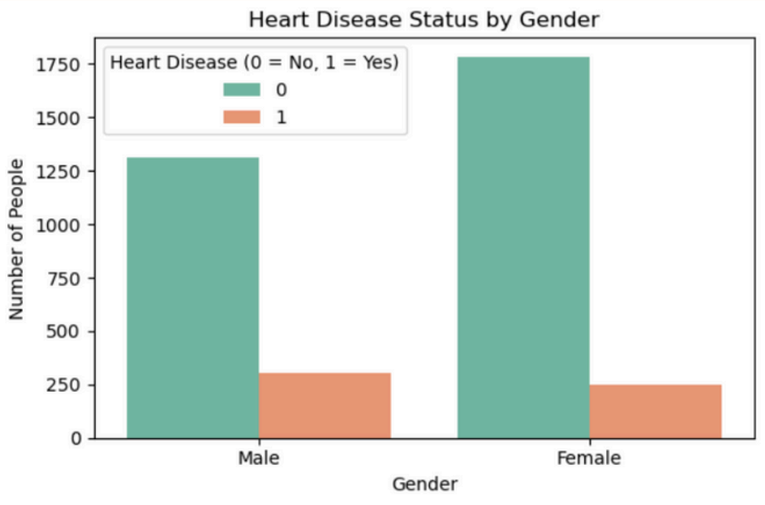
CORRELATION TESTING

3. Is there a statistically significant association between gender and the presence of heart disease? (Phi Coefficient)

1

```
Contingency Table:
Heart_ stroke     No  yes
Gender_binary
0.0             1784  250
1.0             1315  307
P-value: 0.0000
Phi Coefficient (ϕ): 0.0910
```

2



Males have a slightly higher proportion of heart disease compared to females

However, the strength of the association is very weak (φ = 0.0910), indicating that while gender may be related to heart disease status, the effect size is minimal.

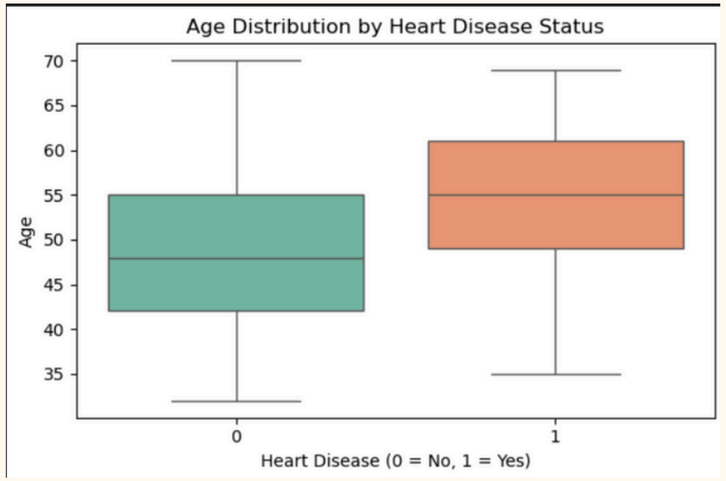
CORRELATION TESTING

4. Is there a statistically significant association between age and the presence of heart disease? (Point-Biserial Correlation)

1

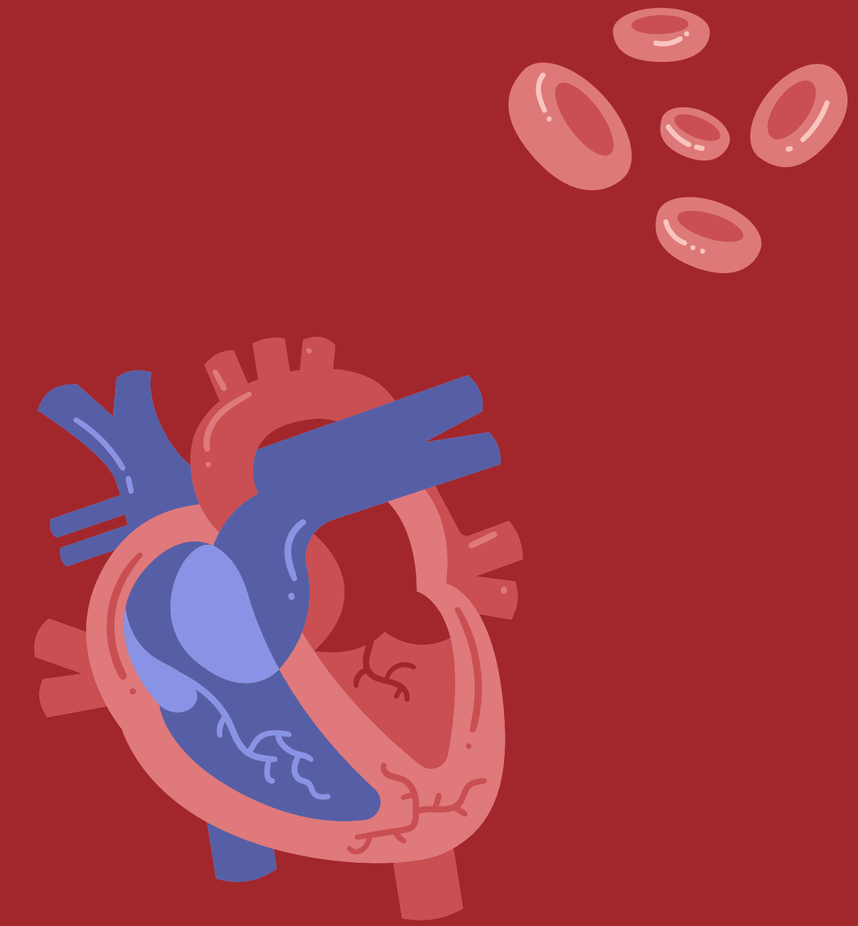
```
int-Biserial correlation coefficient: 0.2338
value: 0.0000
```

2



Statistically **significant positive association between age and the presence of heart disease** (r = 0.2338)

- While the correlation strength is moderate-weak, the result suggests that older individuals are more likely to have heart disease
- The boxplot further supports this finding, showing a noticeably higher median and overall age distribution for individuals with heart disease



These findings reinforce clinical understanding that hypertension, high cholesterol, smoking, high glucose, and obesity are major risk factors for heart disease. **Early screening, lifestyle interventions, and long-term monitoring of at-risk individuals are essential to reducing the burden of heart disease**.