DAV Report – Week 10 & 11
Dataset 3 – Heart Monitoring

I ADARSH, Section PG *(Roll no. 12)*, Reg. no. 240905294

1. Objective of the analysis

1.1 Introduction to the dataset

- The given dataset gives us details about various health indicators and risk factors associated with heart disease.
- These indicators include age, gender, blood pressure, cholesterol levels, smoking habits and exercise patterns.
- The objective of this analysis is to analyze the risk of heart disease based on various factors like smoking, blood pressure etc.

1.2 Variables in the dataset

```
<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
```

2. Data exploration

2.1 Data exploration of the attributes: Blood Pressure, Cholesterol Level, Triglyceride Level, Fasting Blood Sugar, CRP Level, Homocysteine Level.

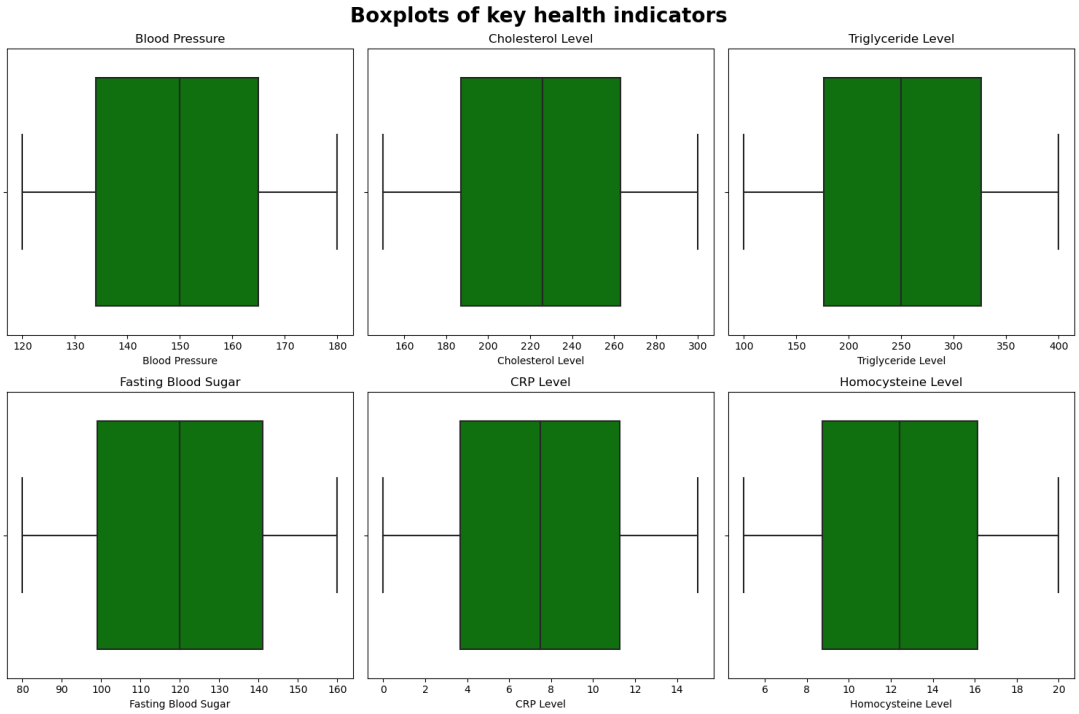
Five number summary of the attributes

	Blood Pressure	Cholesterol Level	Triglyceride Level	Fasting Blood Sugar	CRP Level	Homocysteine Level
min	120.0	150.0	100.0	80.0	0.003647	5.000236
25%	134.0	187.0	176.0	99.0	3.674126	8.723334
50%	150.0	226.0	250.0	120.0	7.472164	12.409395
75%	165.0	263.0	326.0	141.0	11.255592	16.140564
max	180.0	300.0	400.0	160.0	14.997087	19.999037

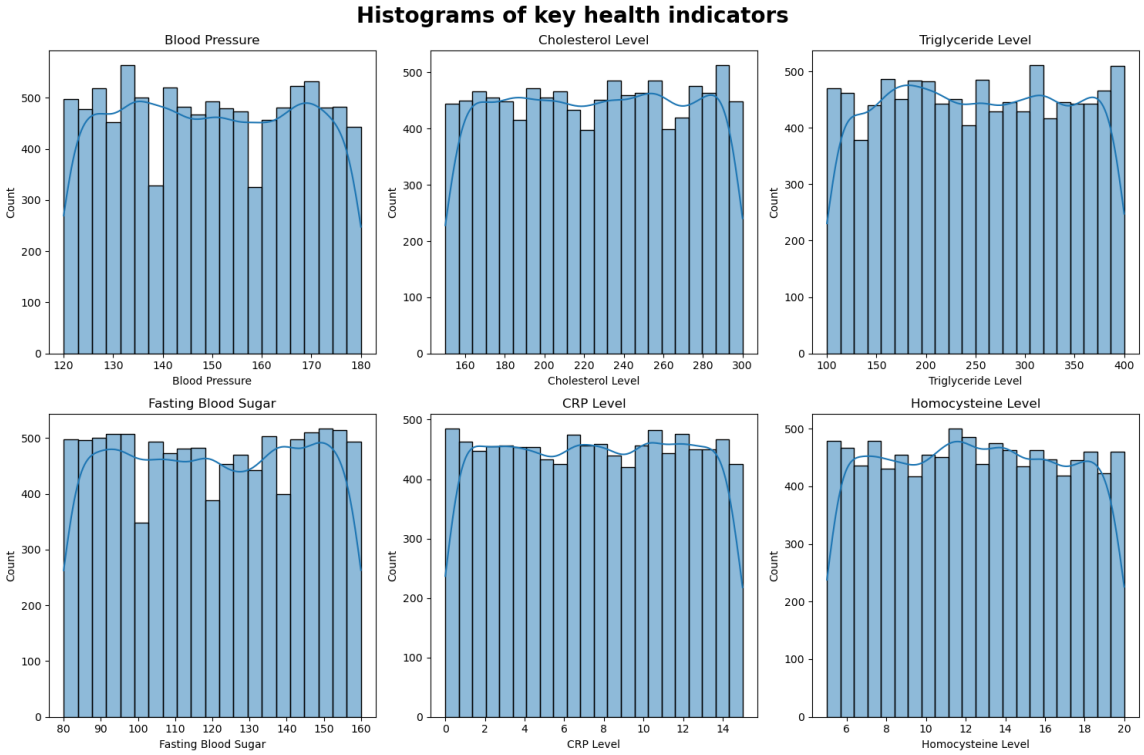
Null values in the attributes

```
Blood Pressure          19
Cholesterol Level       30
Triglyceride Level      26
Fasting Blood Sugar     22
CRP Level               26
Homocysteine Level      20
dtype: int64
```

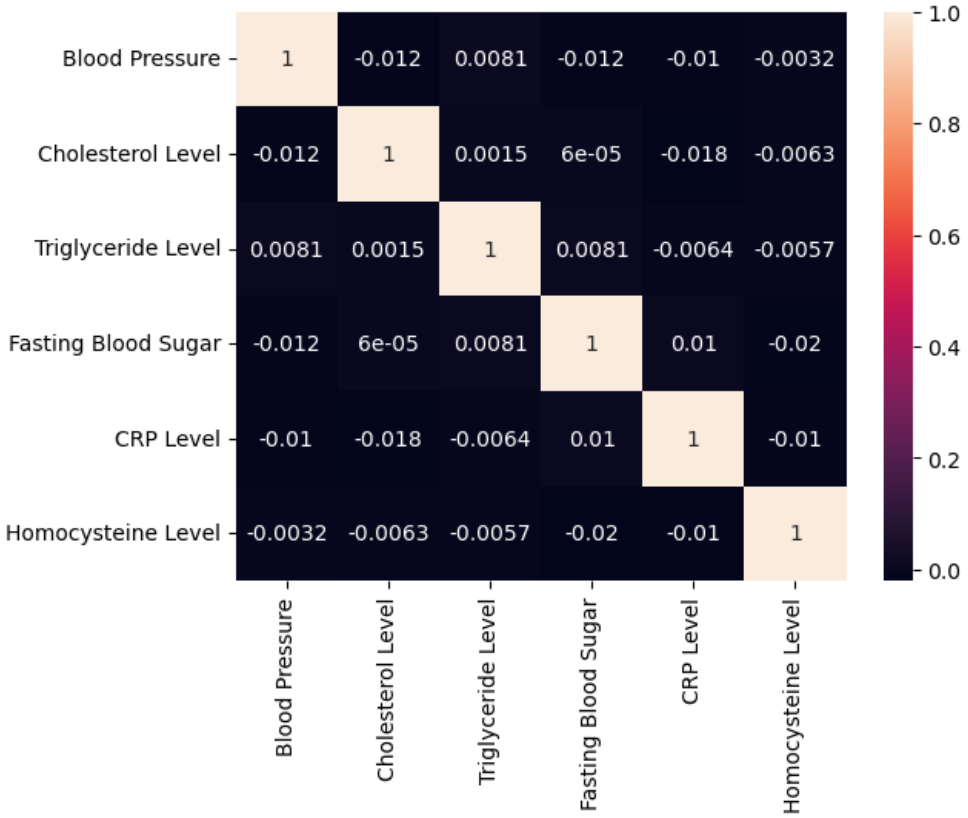
Plotting boxplots to analyze the distribution and outliers for each attribute



Plotting histograms for each attribute to analyze frequency distribution

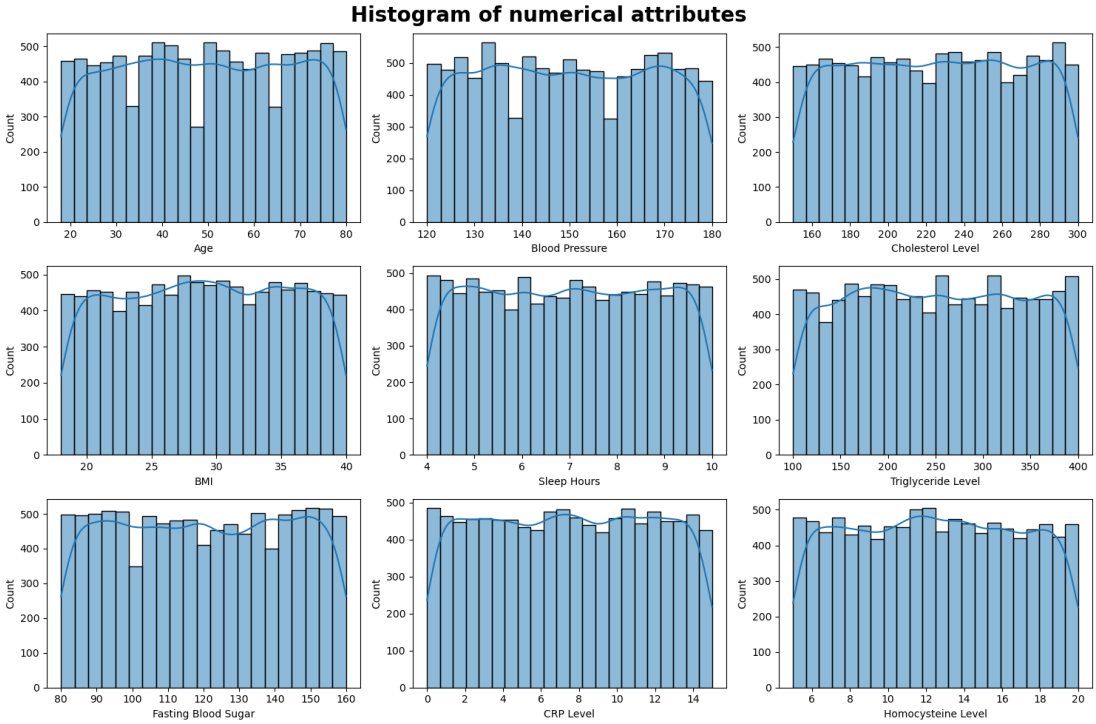


Plotting heatmap to analyze the correlation between each attribute



2.2 Data exploration of all numerical attributes

Histograms of all numerical attributes to analyze frequency distribution

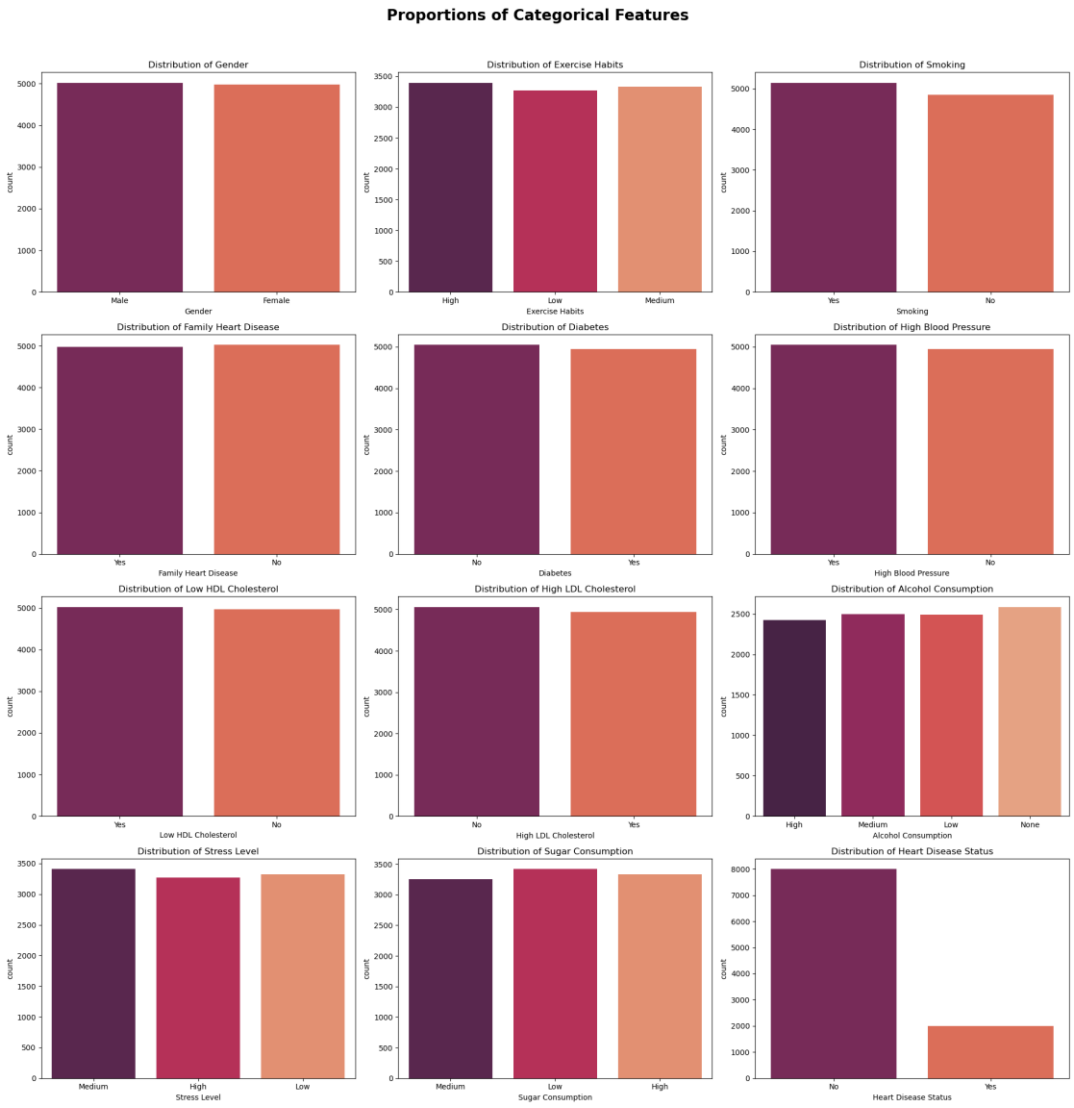


2.3 Data exploration of categorical attributes

Unique values in categorical attributes

```
Male      5022
Female    4978
Name: Gender, dtype: int64

High      3397
Medium    3332
Low       3271
Name: Exercise Habits, dtype: int64

Yes    5148
No     4852
Name: Smoking, dtype: int64

No     5025
Yes    4975
Name: Family Heart Disease, dtype: int64

No     5048
Yes    4952
Name: Diabetes, dtype: int64

Yes    5048
No     4952
Name: High Blood Pressure, dtype: int64

Yes    5025
No     4975
Name: Low HDL Cholesterol, dtype: int64

No     5062
Yes    4938
Name: High LDL Cholesterol, dtype: int64

None      2586
Medium    2500
Low       2488
High      2426
Name: Alcohol Consumption, dtype: int64

Medium    3409
Low       3320
High      3271
Name: Stress Level, dtype: int64

Low       3420
High      3330
Medium    3250
Name: Sugar Consumption, dtype: int64

No     8000
Yes    2000
Name: Heart Disease Status, dtype: int64
```

Count plots of categorical attributes



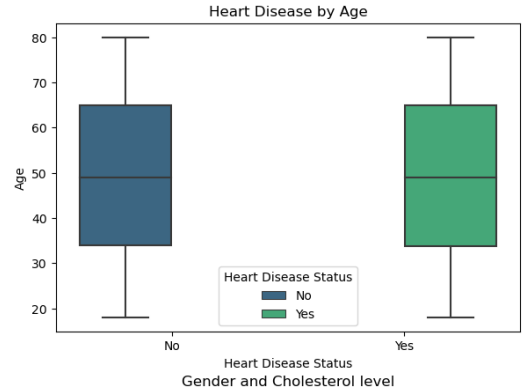
2.3 Data exploration of example attributes

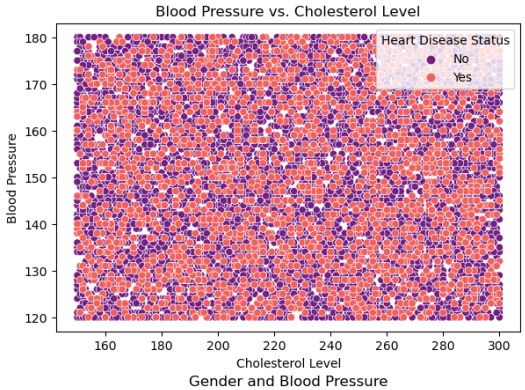
Examples:

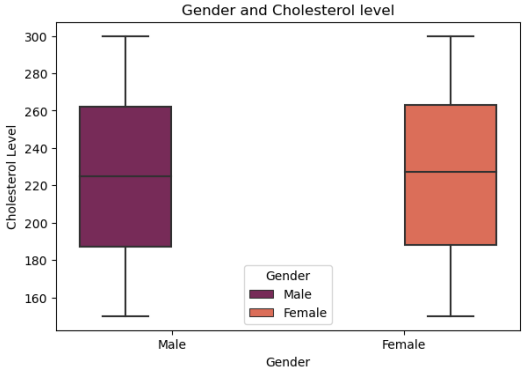
How does age correlate with heart disease risk?

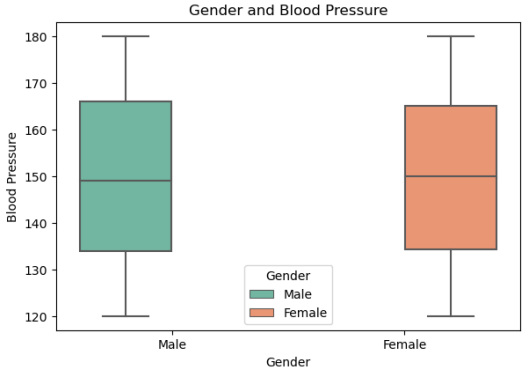
Is there a relation between cholesterol levels and blood pressure?

Does gender affect cholesterol levels or blood pressure?

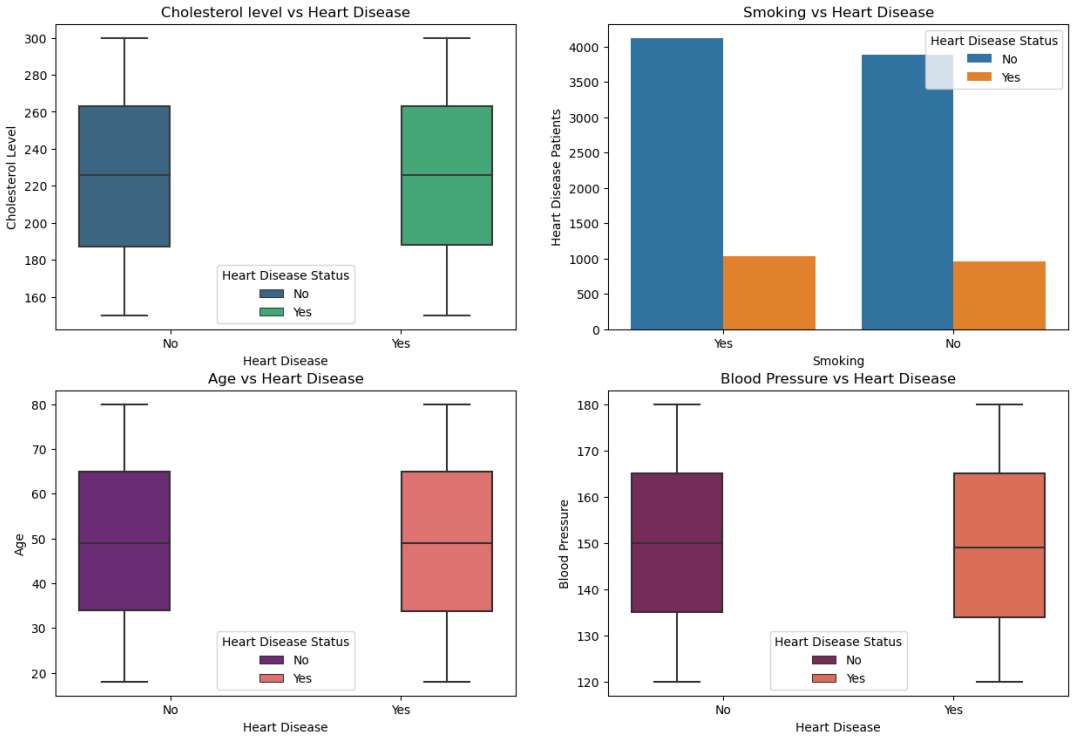






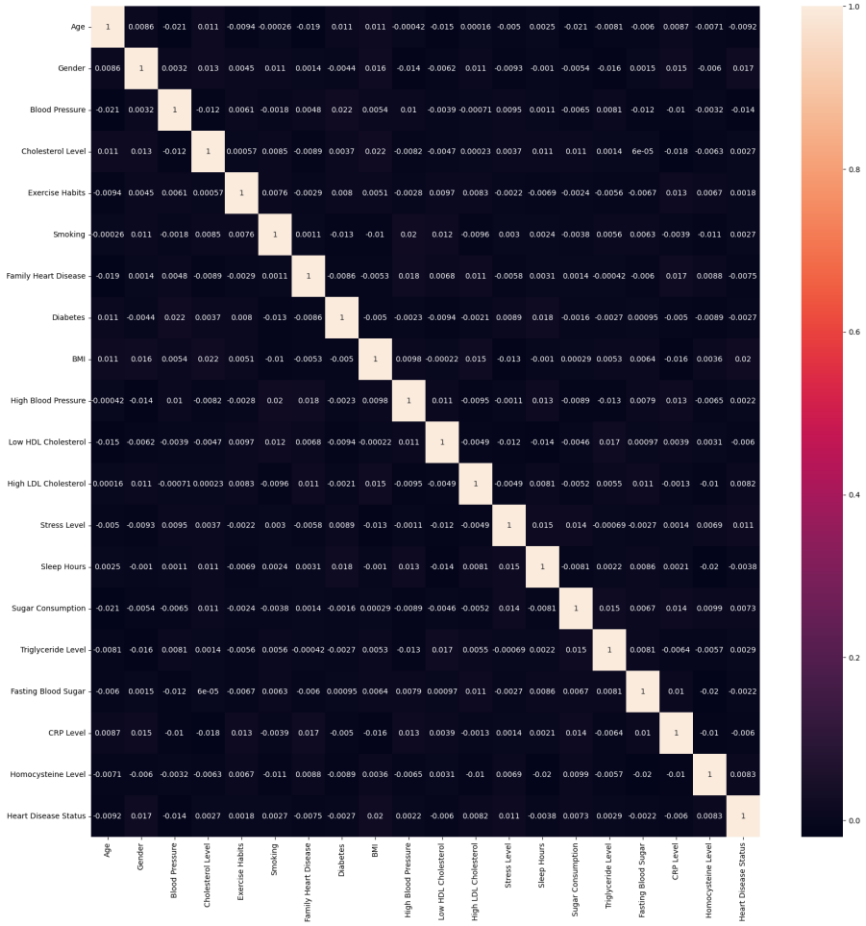


Do higher cholesterol levels indicate a higher risk of heart disease?
How does smoking status affect heart disease occurrence?
Do older individuals with high blood pressure have a significantly higher risk?



2.4 Data exploration of all attributes

Heatmap of all attributes *(categorical data was converted into numerical data)*



3. Data cleaning

3.1 Summary of data before cleaning

Five number summary before cleaning

	Age	Blood Pressure	Cholesterol Level	BMI	Sleep Hours	Triglyceride Level	Fasting Blood Sugar	CRP Level	Homocysteine Level
min	18.0	120.0	150.0	18.002837	4.000605	100.0	80.0	0.003647	5.000236
25%	34.0	134.0	187.0	23.658075	5.449866	176.0	99.0	3.674126	8.723334
50%	49.0	150.0	226.0	29.079492	7.003252	250.0	120.0	7.472164	12.409395
75%	65.0	165.0	263.0	34.520015	8.531577	326.0	141.0	11.255592	16.140564
max	80.0	180.0	300.0	39.996954	9.999952	400.0	160.0	14.997087	19.999037

Null values before cleaning

```
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
```

3.2 Summary of data after cleaning (filling numerical data with median, categorical data with mode)

Null values after cleaning

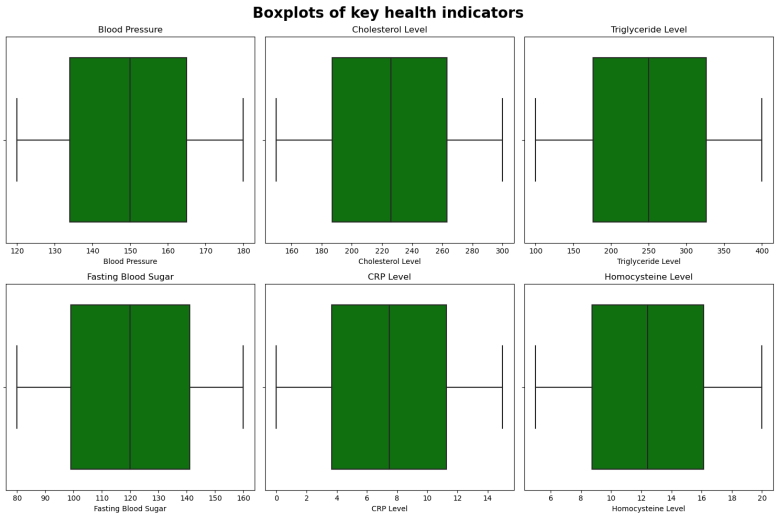
```
Age                    0
Gender                 0
Blood Pressure         0
Cholesterol Level      0
Exercise Habits        0
Smoking                0
Family Heart Disease   0
Diabetes               0
BMI                    0
High Blood Pressure    0
Low HDL Cholesterol    0
High LDL Cholesterol   0
Alcohol Consumption    0
Stress Level           0
Sleep Hours            0
Sugar Consumption      0
Triglyceride Level     0
Fasting Blood Sugar    0
CRP Level              0
Homocysteine Level     0
Heart Disease Status   0
dtype: int64
```

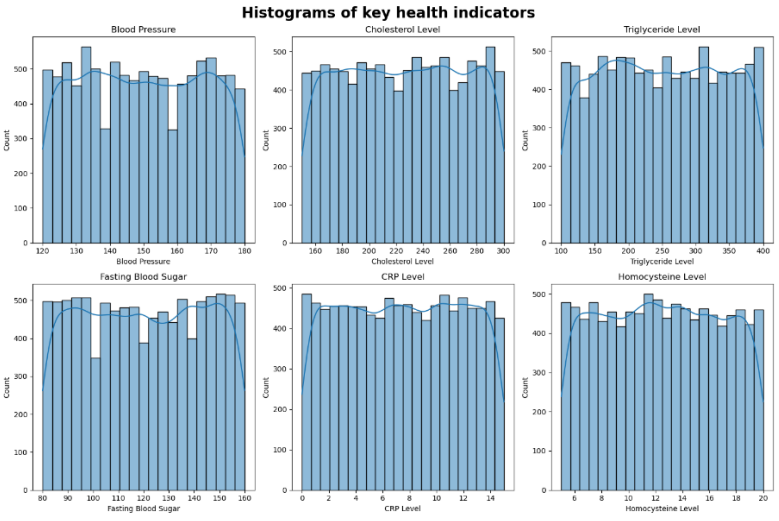
Five number summary after cleaning

	Age	Blood Pressure	Cholesterol Level	BMI	Sleep Hours	Triglyceride Level	Fasting Blood Sugar	CRP Level	Homocysteine Level
min	18.0	120.0	150.0	18.002837	4.000605	100.0	80.0	0.003647	5.000236
25%	34.0	134.0	187.0	23.668887	5.455288	176.0	99.0	3.681800	8.729771
50%	49.0	150.0	226.0	29.079492	7.003252	250.0	120.0	7.472164	12.409395
75%	65.0	165.0	263.0	34.509009	8.527938	326.0	141.0	11.244879	16.130968
max	80.0	180.0	300.0	39.996954	9.999952	400.0	160.0	14.997087	19.999037

4. Analysis and insights

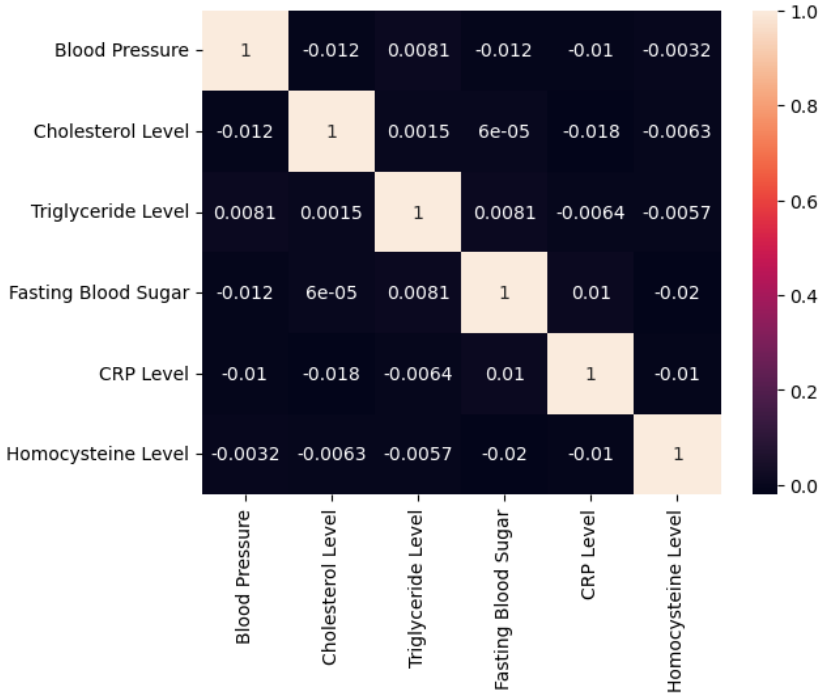
4.1 Distribution of data across given features





Data is roughly evenly distributed across the given features of Blood Pressure, Cholesterol Level, Triglyceride Level, Fasting Blood Sugar, CRP Level, Homocysteine Level.

4.2 Relationship of data between given features



No significant relationship between the given features of Blood Pressure, Cholesterol Level, Triglyceride Level, Fasting Blood Sugar, CRP Level, Homocysteine Level.

4.3 Outliers

Age	Gender	Blood Pressure	Cholesterol Level	Exercise Habits	Smoking	Family Heart Disease	Diabetes	BMI	High Blood Pressure	...	High LDL Cholesterol	Alcohol Consumption	Stress Level	Sleep Hours	Sugar Consumption	Trigl

0 rows × 21 columns

Using IQR to check for outliers, there are no outliers present in the data.

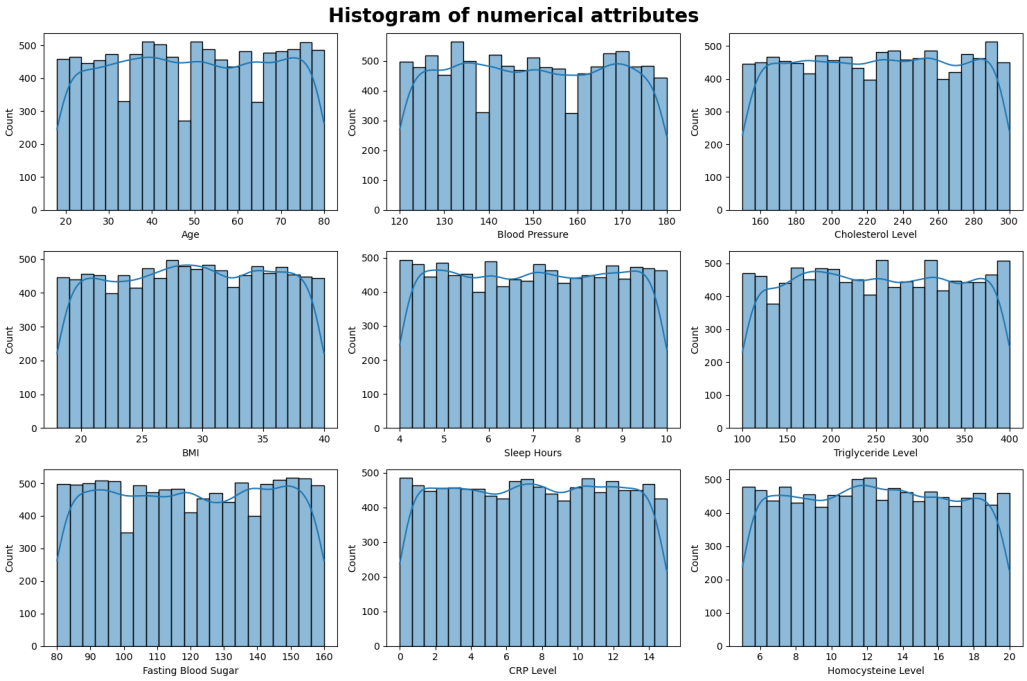
4.4 Difference in data before and after cleaning

	Age	Blood Pressure	Cholesterol Level	BMI	Sleep Hours	Triglyceride Level	Fasting Blood Sugar	CRP Level	Homocysteine Level
min	18.0	120.0	150.0	18.002837	4.000605	100.0	80.0	0.003647	5.000236
25%	34.0	134.0	187.0	23.658075	5.449866	176.0	99.0	3.674126	8.723334
50%	49.0	150.0	226.0	29.079492	7.003252	250.0	120.0	7.472164	12.409395
75%	65.0	165.0	263.0	34.520015	8.531577	326.0	141.0	11.255592	16.140564
max	80.0	180.0	300.0	39.996954	9.999952	400.0	160.0	14.997087	19.999037

	Age	Blood Pressure	Cholesterol Level	BMI	Sleep Hours	Triglyceride Level	Fasting Blood Sugar	CRP Level	Homocysteine Level
min	18.0	120.0	150.0	18.002837	4.000605	100.0	80.0	0.003647	5.000236
25%	34.0	134.0	187.0	23.668887	5.455288	176.0	99.0	3.681800	8.729771
50%	49.0	150.0	226.0	29.079492	7.003252	250.0	120.0	7.472164	12.409395
75%	65.0	165.0	263.0	34.509009	8.527938	326.0	141.0	11.244879	16.130968
max	80.0	180.0	300.0	39.996954	9.999952	400.0	160.0	14.997087	19.999037

Values have changed, but not by much due to very few numeric null data.

4.5 Distribution of numerical data



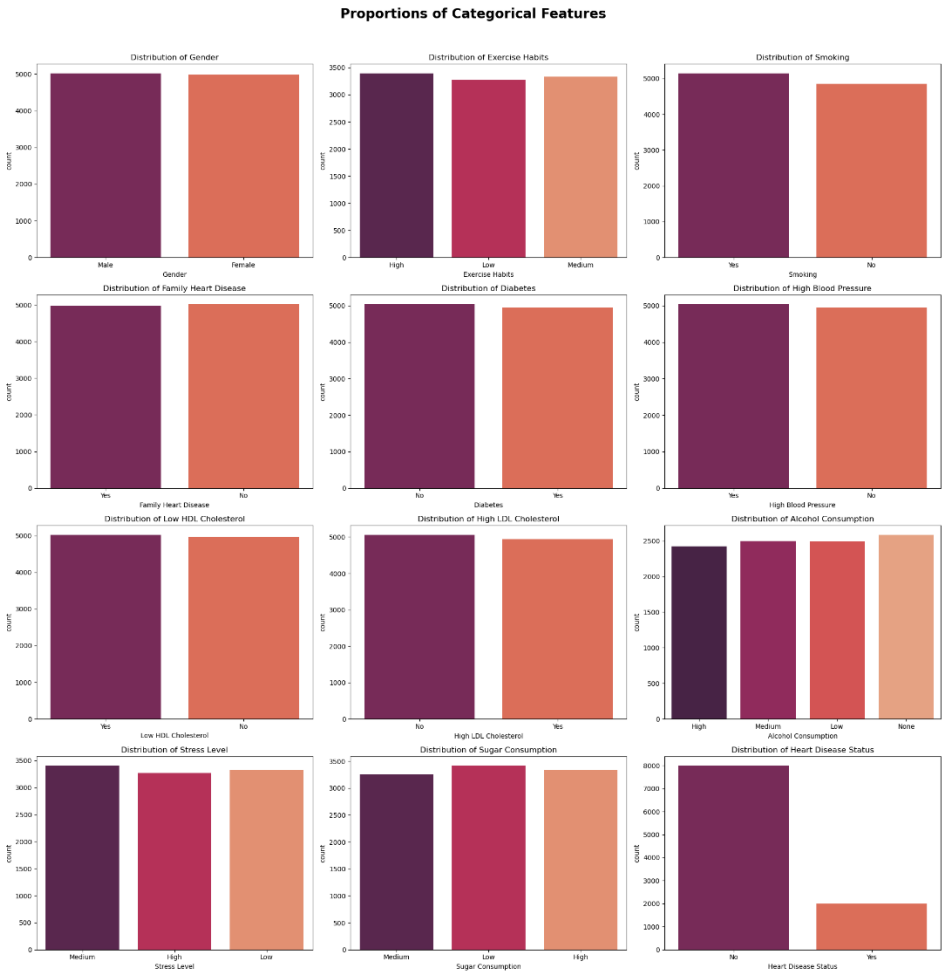
Data is roughly evenly distributed across the above features.

4.6 Skewness in numerical data

```
Age                   -0.006657
Blood Pressure         0.013842
Cholesterol Level     -0.007250
BMI                   -0.021368
Sleep Hours            0.000121
Triglyceride Level     0.006216
Fasting Blood Sugar   -0.008885
CRP Level             -0.004074
Homocysteine Level     0.007959
dtype: float64
```

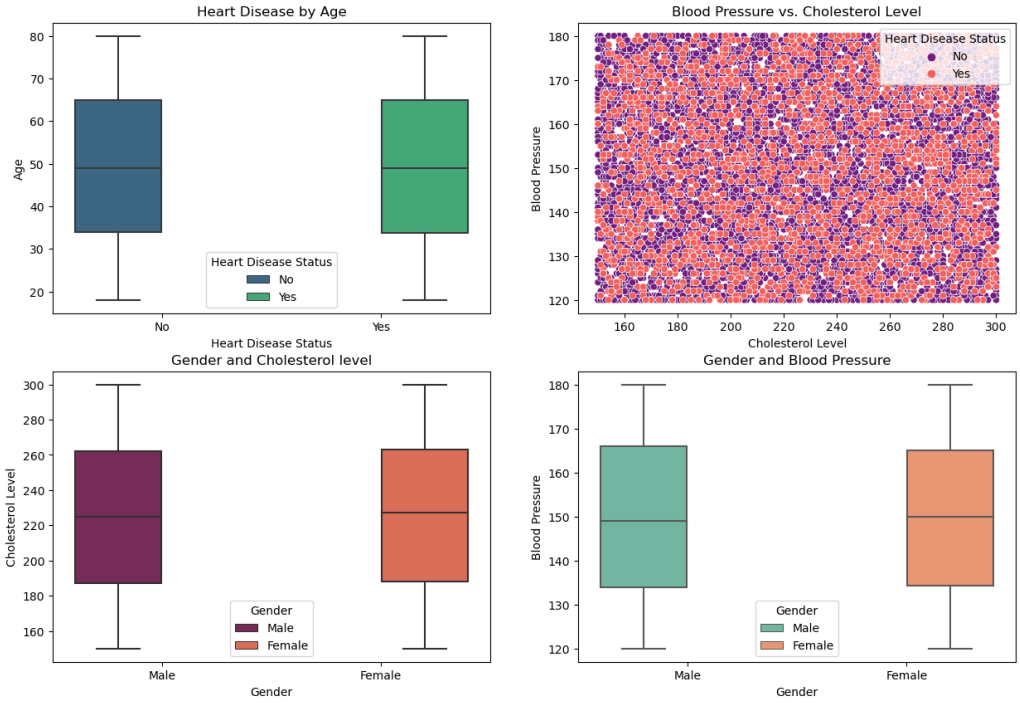
There is no significant skewness in numerical data.

4.7 Distribution of categorical data



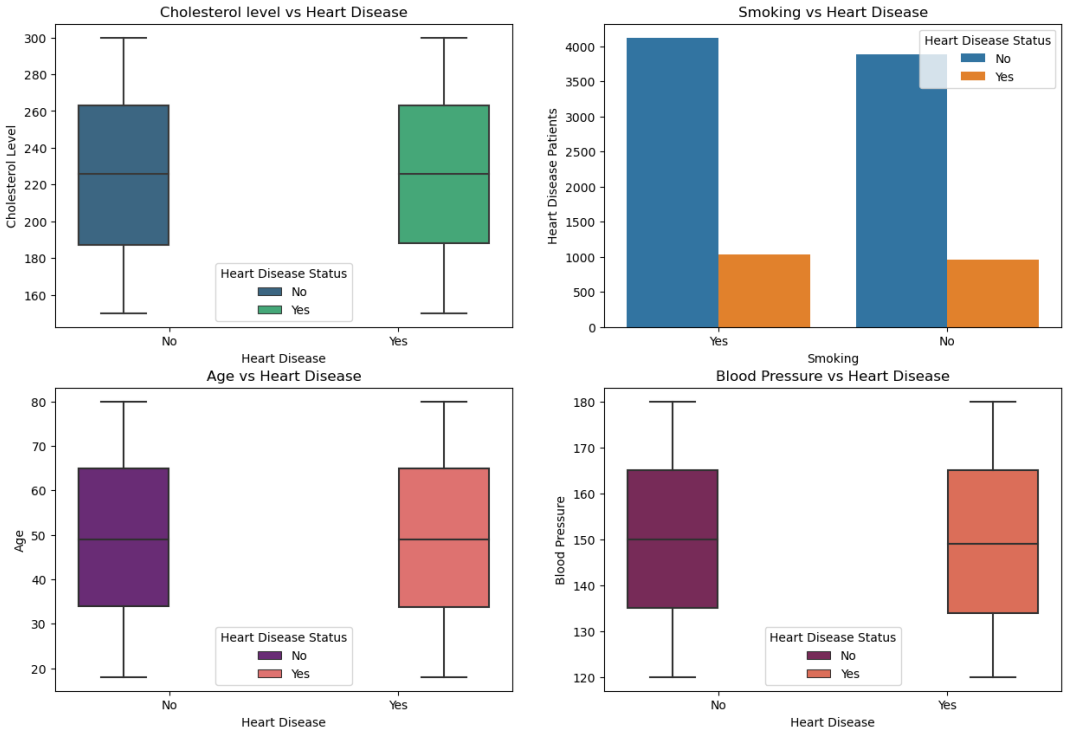
Almost all values are highly evenly distributed except Heart Disease Status and Alcohol Consumption due to filling of a large number of missing values with the mode.

4.8 Correlation between example attributes



- **Heart disease by Age:** Median and variance is almost identical for both patients and non-patients of heart disease. Thus, age alone does not determine heart disease significantly.
- **Blood Pressure vs Cholesterol Level:** Highly even scatter plot, data is not correlated.
- **Gender and Cholesterol Level:** Cholesterol level is roughly the same in both genders, except for a slightly higher median in females. However, the difference is too small to make any difference in the data.
- **Gender and Blood Pressure:** Blood pressure is roughly the same in both genders, except for a slightly higher median in females. However, the difference is too small to make any difference in the data.

4.9 Analyzing example feature interactions



- **Cholesterol Level vs Heart Disease:** The median cholesterol level appears to be slightly higher for individuals with heart disease compared to those without. There are potential outliers in both groups, representing individuals with exceptionally high cholesterol levels.
- **Smoking vs Heart Disease:**
  - For both smokers and non-smokers, the number of individuals without heart disease is significantly higher than the number of individuals with heart disease.
  - Among those with heart disease, there are more patients who are smokers compared to those who are non-smokers.
  - Conversely, among those without heart disease, there are more non-smokers than smokers. This suggests a possible association between smoking and heart disease.
- **Age vs Heart Disease:** The median age seems to be higher for individuals with heart disease compared to those without. There are potential outliers in both groups, indicating individuals who are significantly younger or older than the majority in their respective groups.
- **Blood Pressure vs Heart Disease:** The median blood pressure appears to be slightly higher for individuals with heart disease compared to those without. There are potential outliers in both groups, representing individuals with unusually high or low blood pressure levels.

5. Conclusion

Using this dataset and the analysis on various attributes, we can predict and analyze the risk of heart disease among different groups of people.