Cardivascular Disease Predication

Exploratory Data Analysis (EDA)

For Exploratory Data Analysis (EDA), we first imported the necessary libraries.

The main library used for handling data is `pandas`, which allows us to create and manipulate dataframes.

For data visualization, we use `matplotlib` and `seaborn` to generate informative plots.

We also use `scipy` and `numpy` for statistical analysis and numerical computations.

Importing Required Libraries

•Initial Data Inspection:

Out[2]:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	ac
0	0	18393	2	168	62.0	110	80	1	1	0	0	
1	1	20228	1	156	85.0	140	90	3	1	0	0	
2	2	18857	1	165	64.0	130	70	3	1	0	0	
3	3	17623	2	169	82.0	150	100	1	1	0	0	
4	4	17474	1	156	56.0	100	60	1	1	0	0	
5	8	21914	1	151	67.0	120	80	2	2	0	0	
6	9	22113	1	157	93.0	130	80	3	1	0	0	
7	12	22584	2	178	95.0	130	90	3	3	0	0	
8	13	17668	1	158	71.0	110	70	1	1	0	0	
9	14	19834	1	164	68.0	110	60	1	1	0	0	

Check data types

```
id                 int64
age                int64
gender             int64
height             int64
weight           float64
ap_hi              int64
ap_lo              int64
cholesterol        int64
gluc               int64
smoke              int64
alco               int64
active             int64
cardio             int64
dtype: object
```

Null Values

```
id             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
```

Check Duplicates

```
0
```

Dimensionality

```
(70000, 13)
```

- ### Descriptive Statistics:

Out[8]:

	id	age	gender	height	weight	ap
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000
mean	49972.419900	19468.865814	1.349571	164.359229	74.205690	128.817
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	154.011
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	16020.000

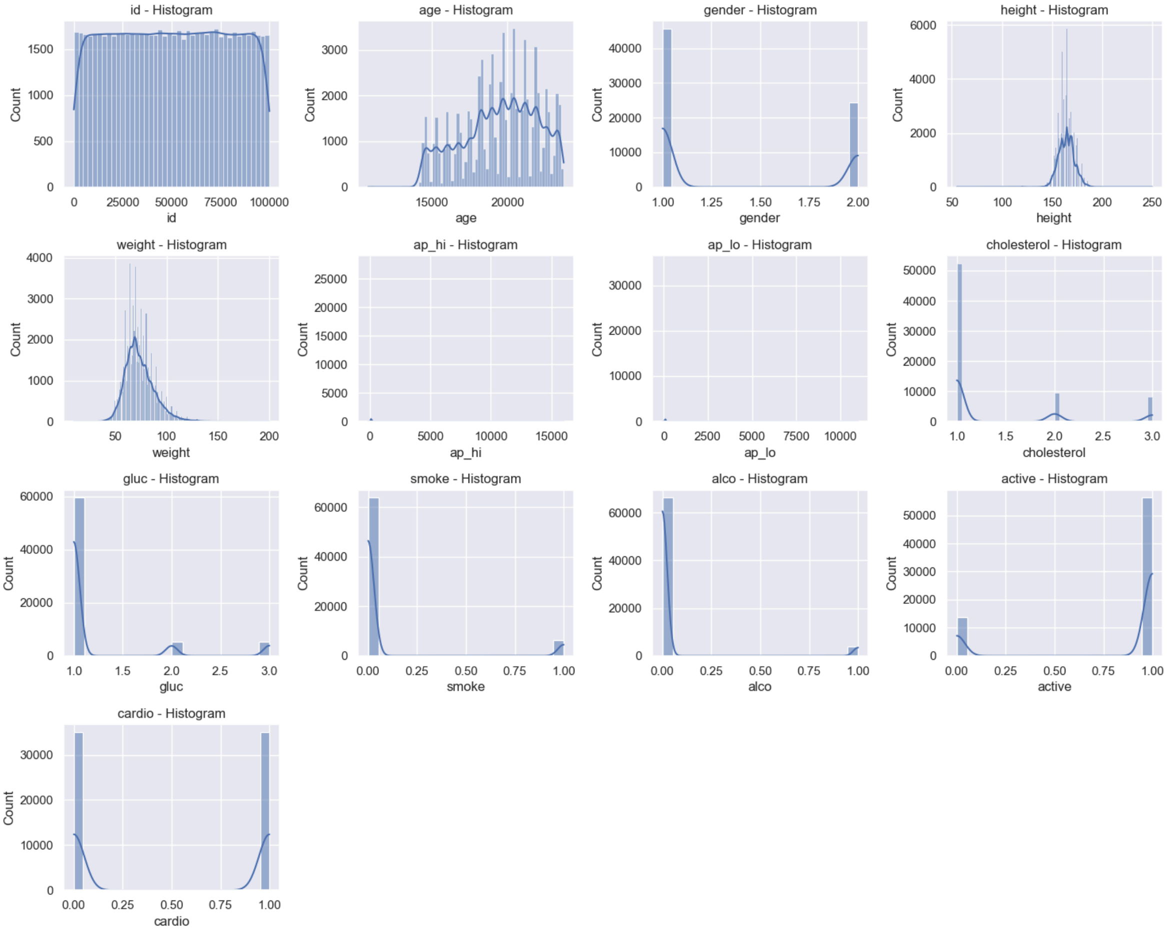
The issues should be addressed to clean the data before proceeding with machine learning models

Age Column: The mean age value is unusually high (19468.87), indicating that the age might be in hundredths of a year. This column needs to be converted to actual years.

Blood Pressure (ap_hi, ap_lo): There are extreme and likely erroneous values, such as -150 for ap_hi and -70 for ap_lo. These need to be corrected or removed.

Binary Variables (smoke, alco, active): Ensure these binary variables only contain values 0 and 1, as there might be incorrect or inconsistent entries

Skewness



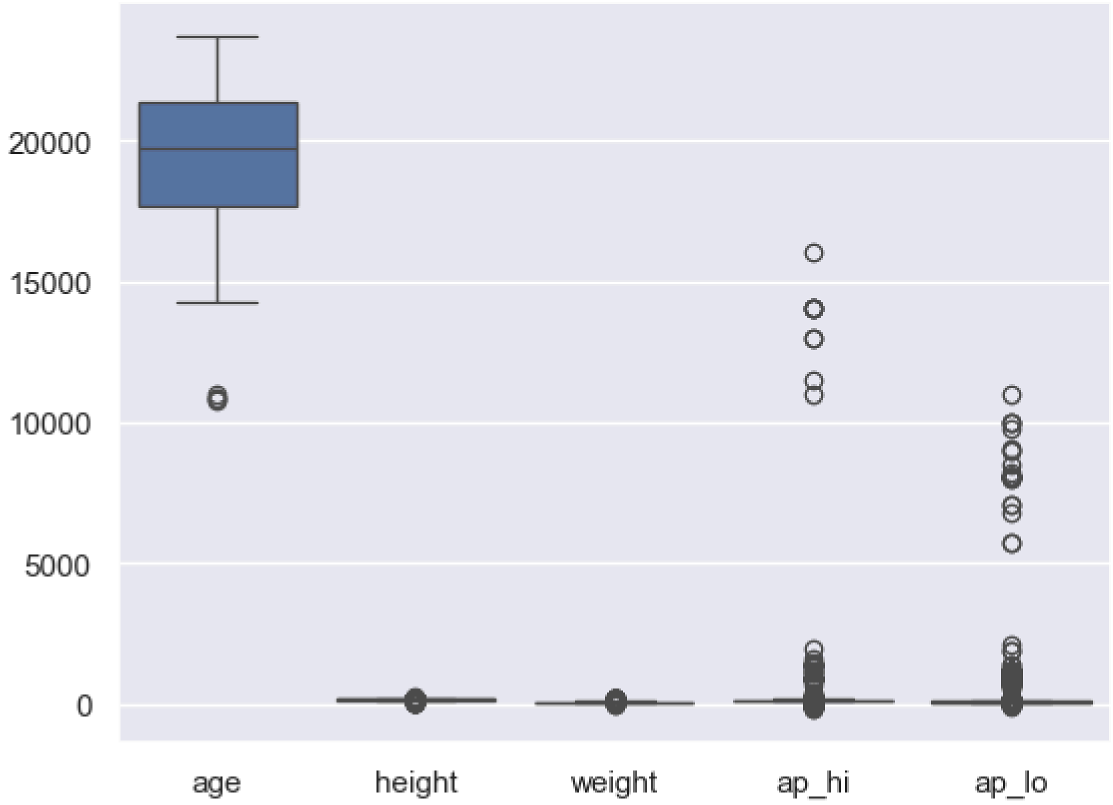
Interpretation

Columns with high variance (e.g., id, age, ap_hi, ap_lo) may have a wide spread in values, which could indicate useful features for predictive modeling, but some columns like id should be discarded as they do not hold predictive power.

Columns with low variance (e.g., gender, smoke, alco, cardio) indicate less variability in the data, which might mean they don’t contribute much in terms of differentiating between observations, especially if they are binary features.

Outliers Detection :

BoxPlot

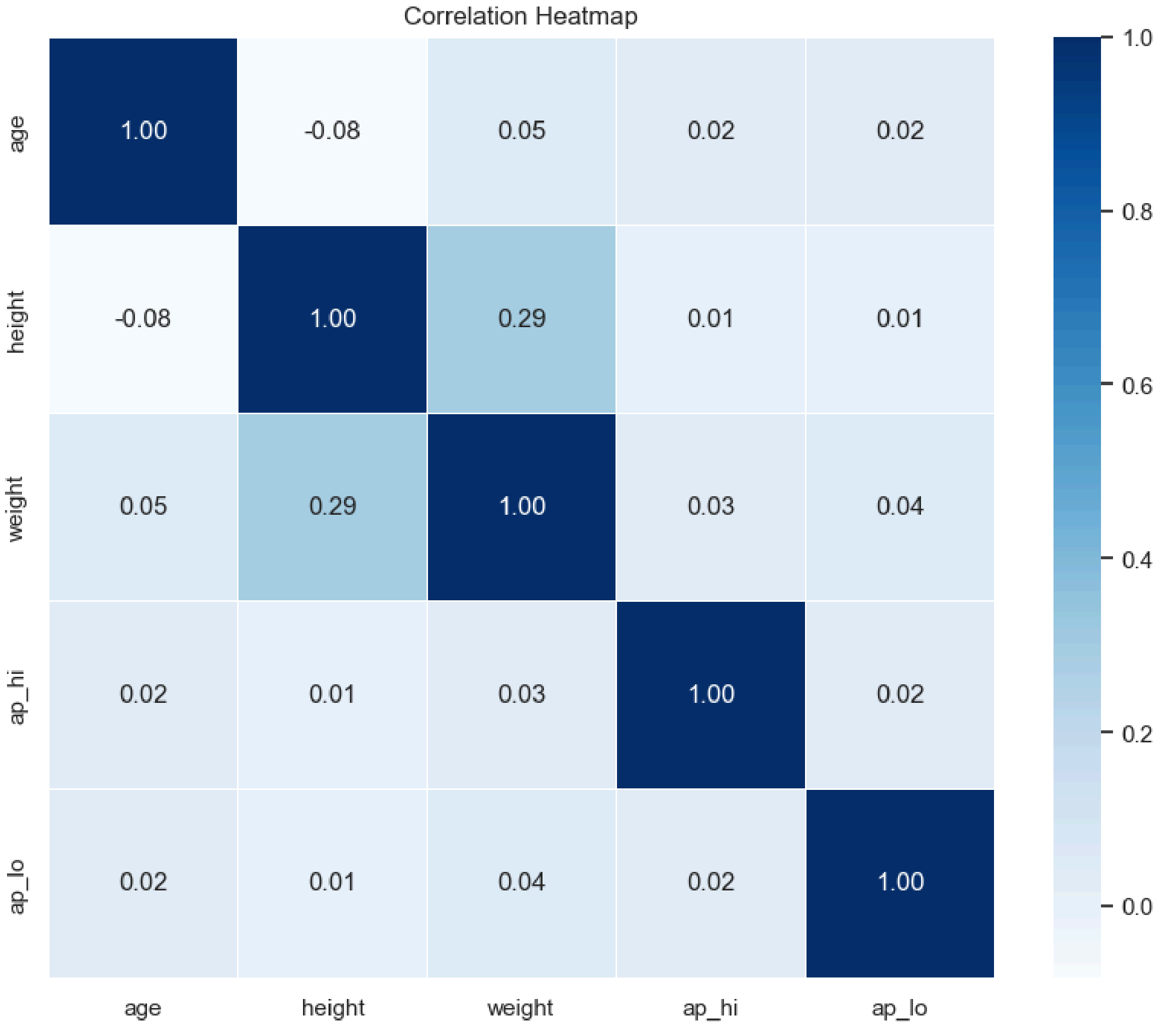


IQR Method :

```
         age  height  weight  ap_hi  ap_lo
0      False   False   False  False  False
1      False   False   False  False  False
2      False   False   False  False  False
3      False   False   False  False  False
4      False   False   False  False   True
...      ...     ...     ...    ...    ...
69995  False   False   False  False  False
69996  False   False    True  False  False
69997  False   False   False   True  False
69998  False   False   False  False  False
69999  False   False   False  False  False

[70000 rows x 5 columns]
```

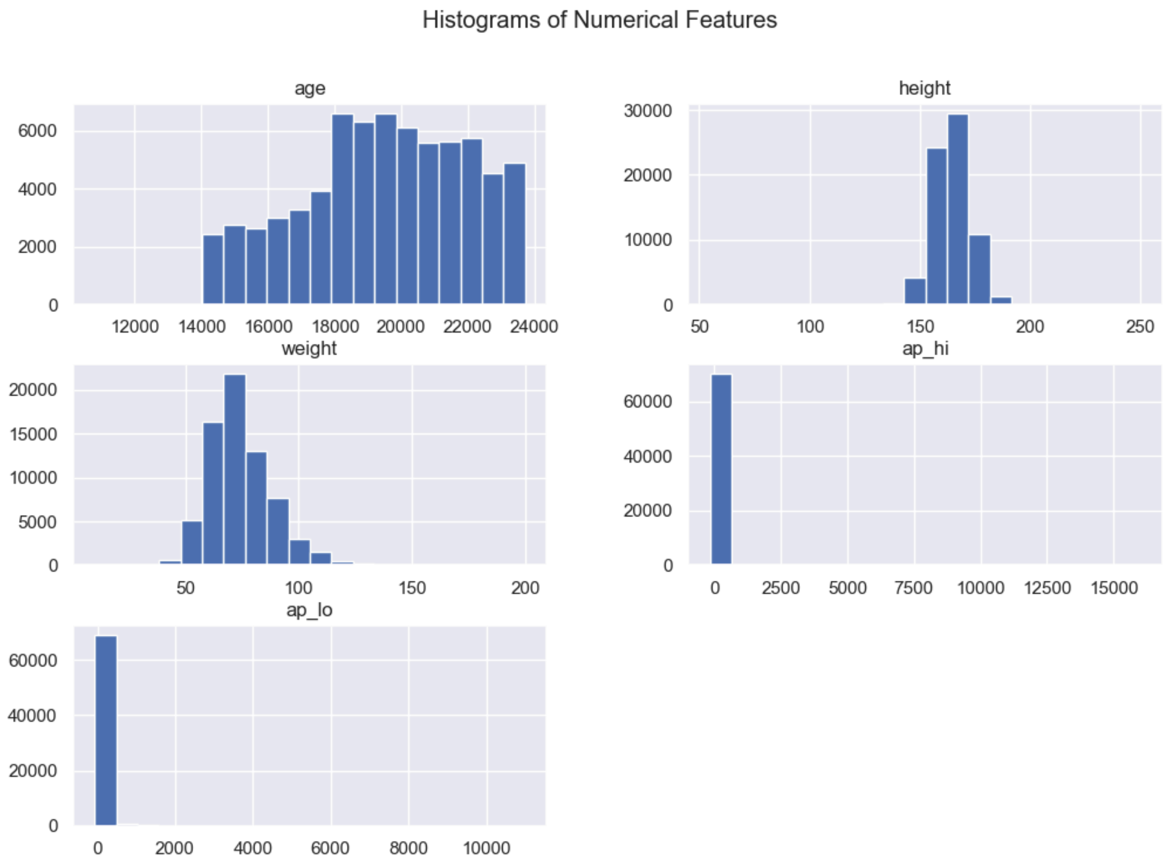
📊 Visualization Summary



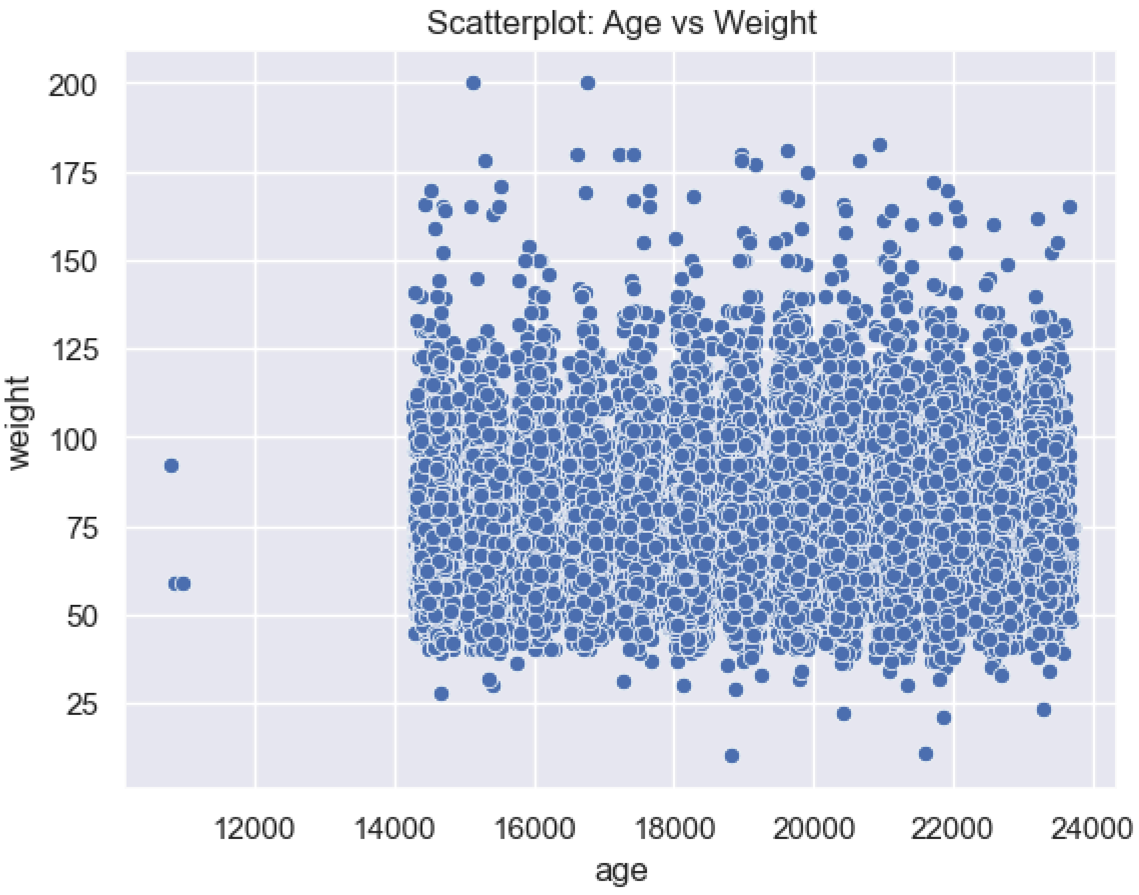
From this correlation matrix, we can understand that all data is unique and not strongly correlated with each other.

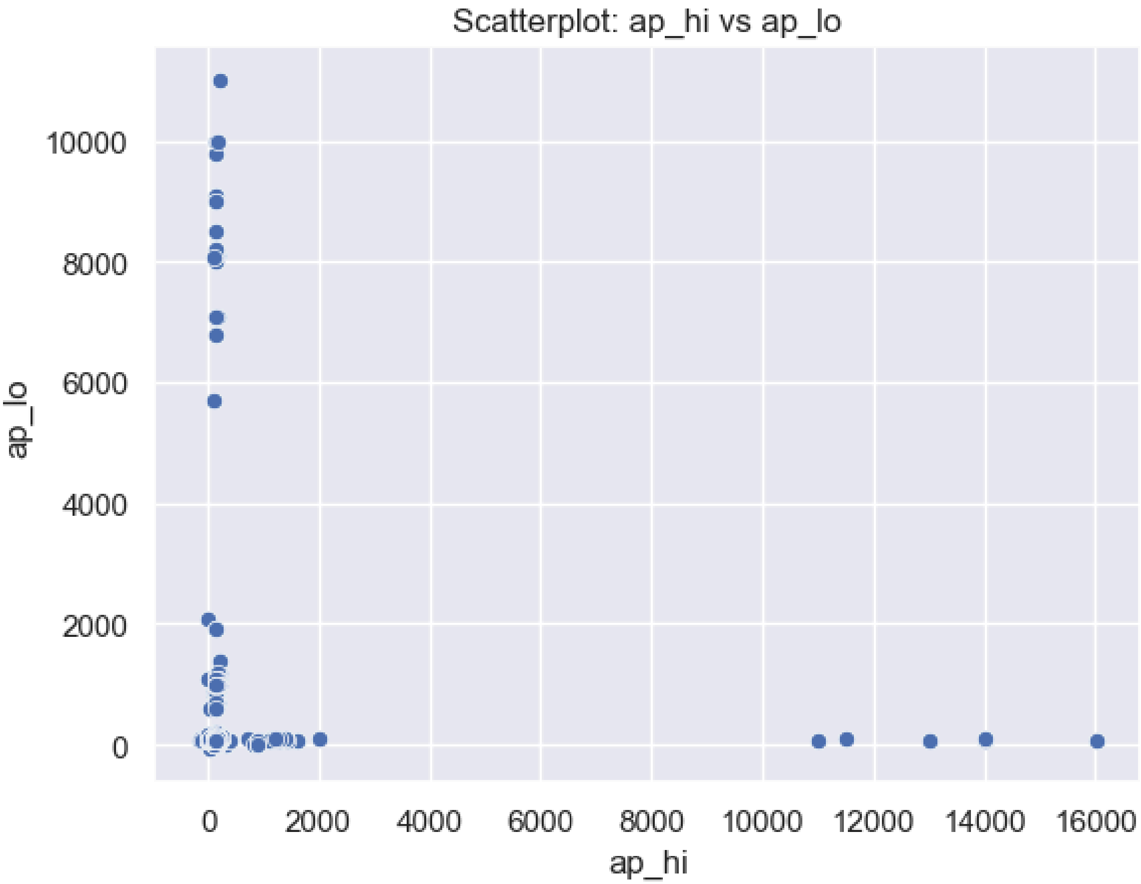
It’s also worth noting that some attributes are highly correlated with the target

Histograms for numerical features

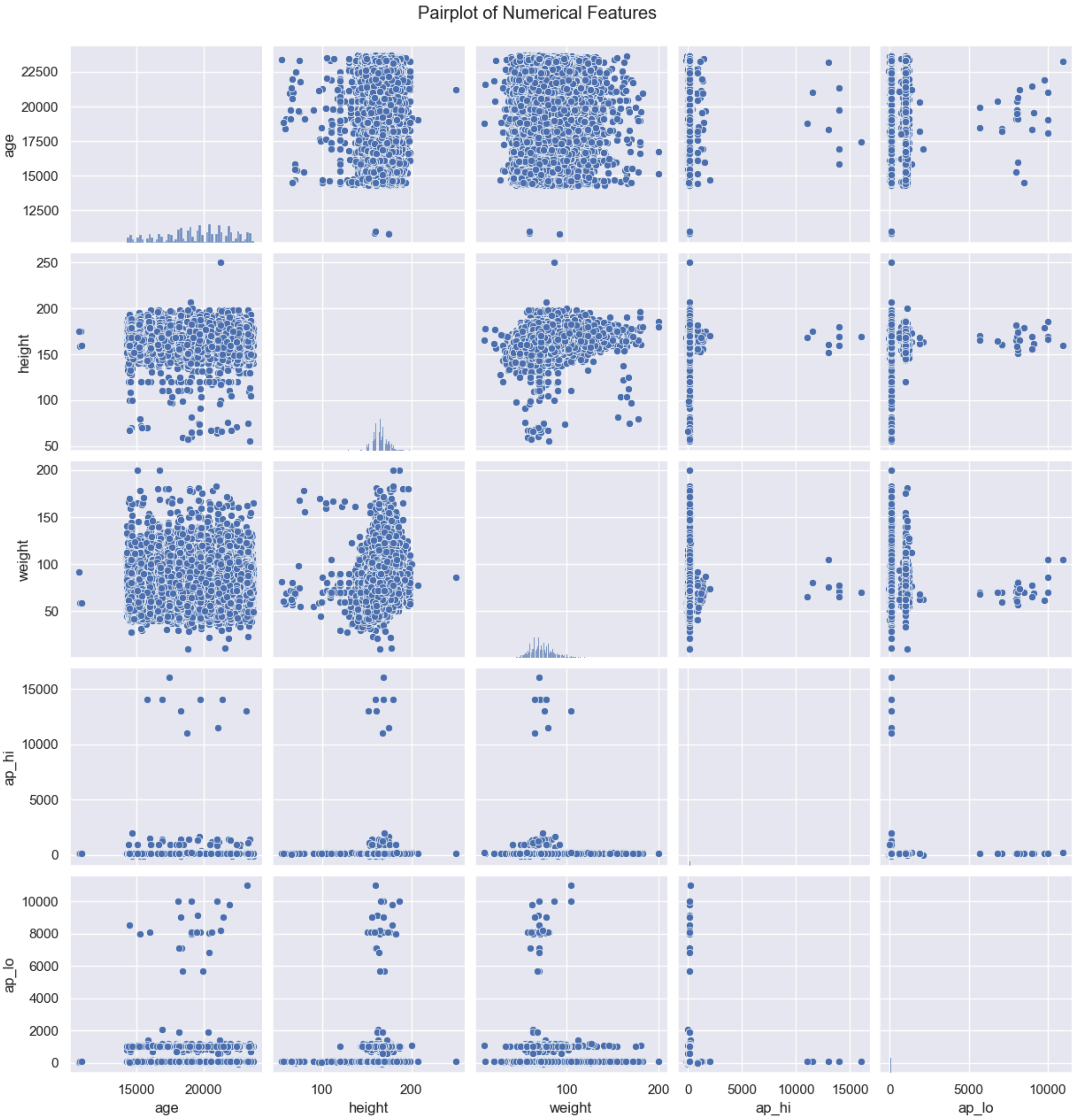


Scatterplots:

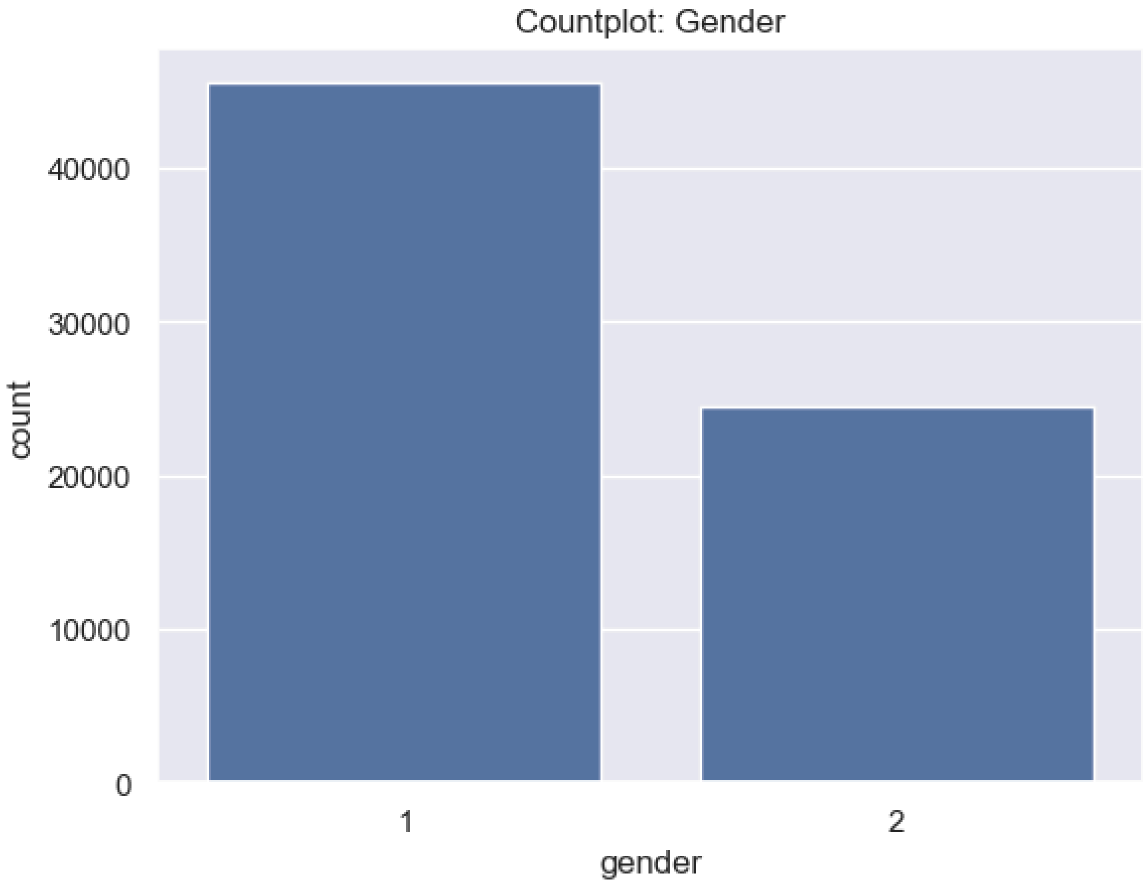


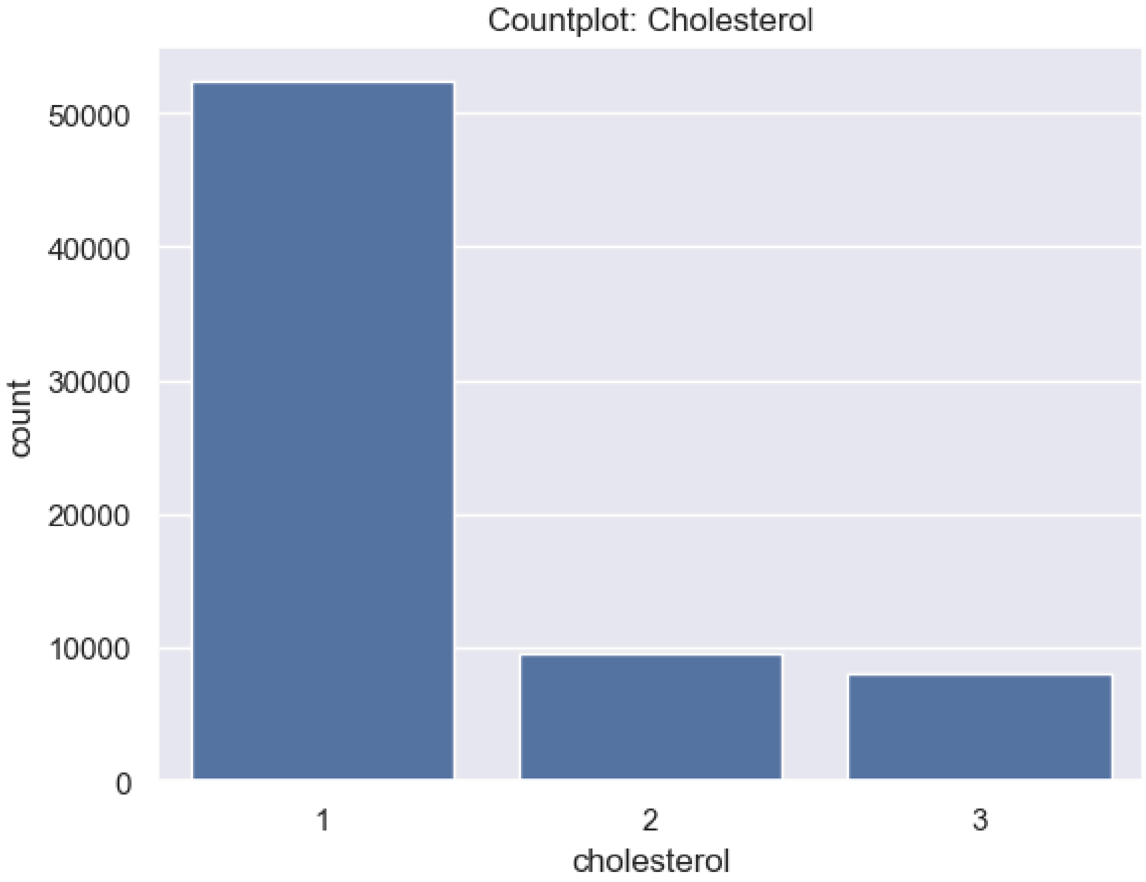

Scatterplot: ap_hi vs ap_lo
ap_lo
10000
8000
6000
4000
2000
0
ap_hi
0
2000
4000
6000
8000
10000
12000
14000
16000

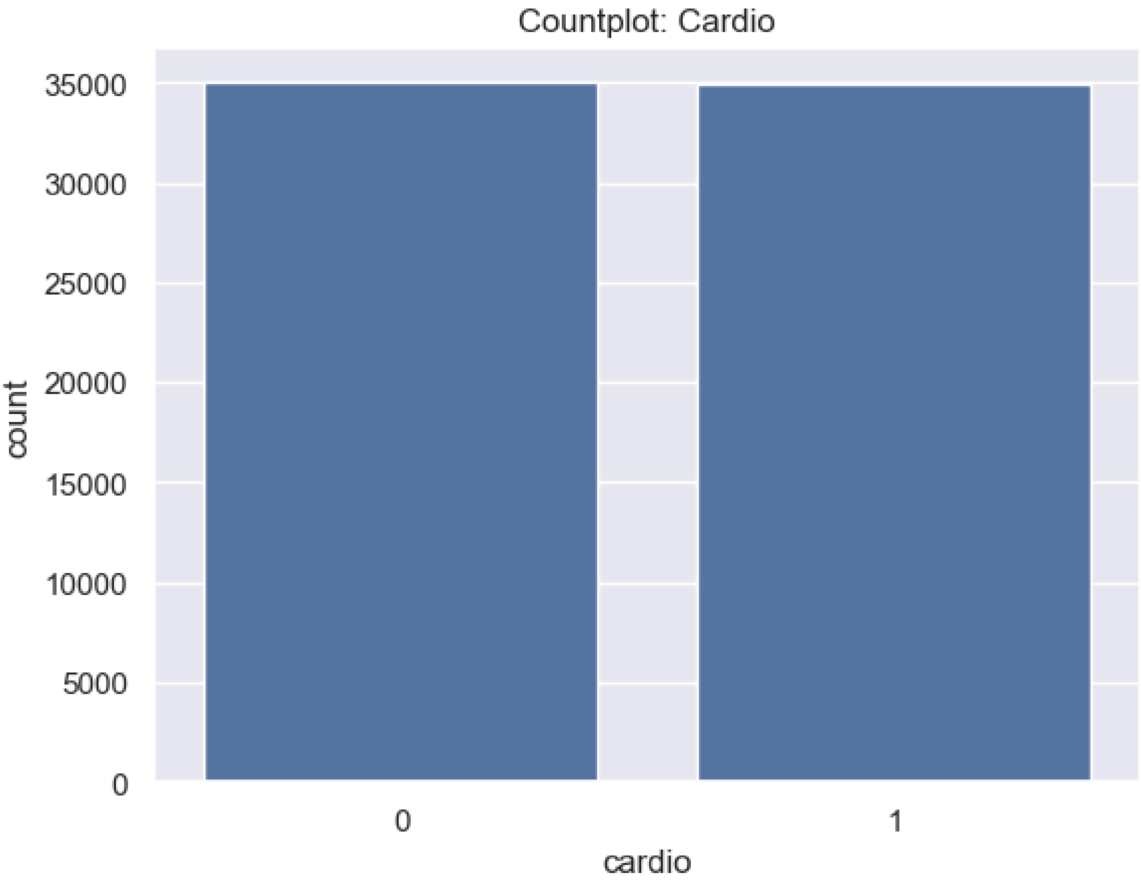
Pairplot


Pairplot of Numerical Features
age
height
weight
ap_hi
ap_lo
22500
20000
17500
15000
12500
250
200
150
100
50
200
150
100
50
15000
10000
5000
0
10000
8000
6000
4000
2000
0
15000
20000
100
200
100
200
0
5000
10000
15000
0
5000
10000

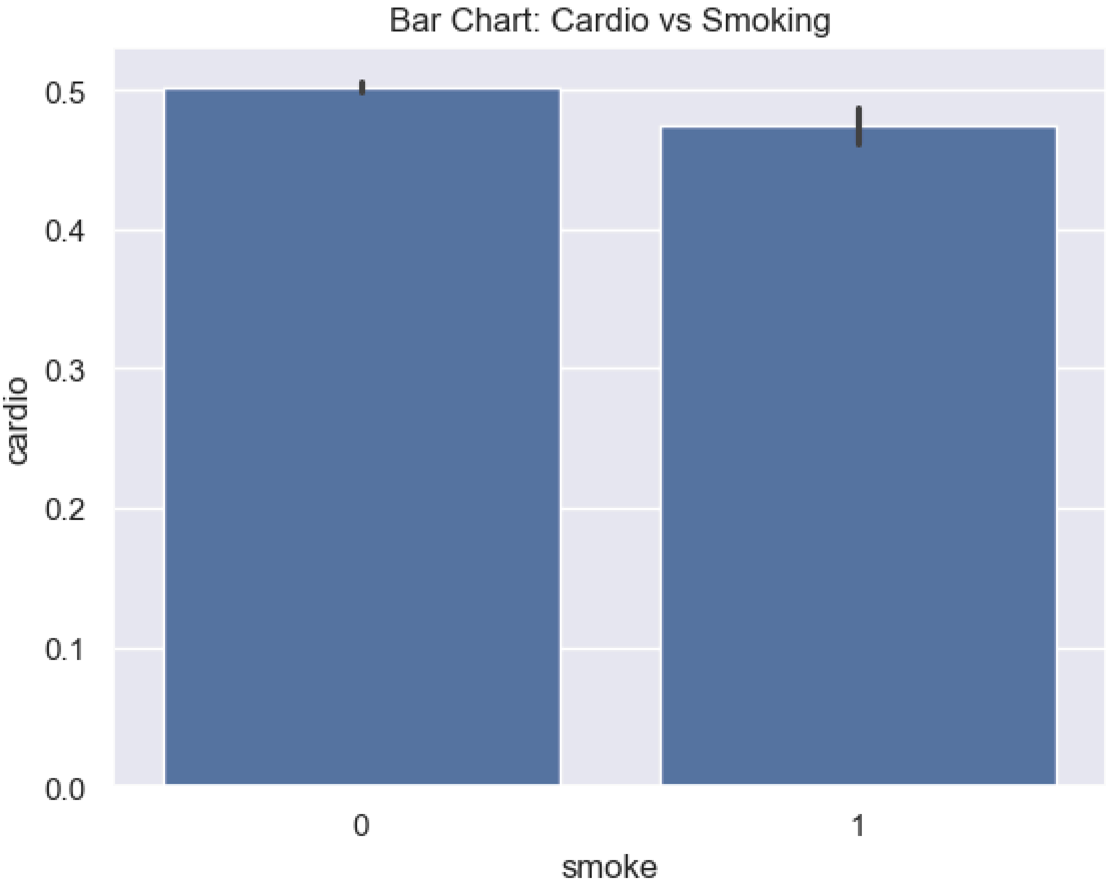
Countplot

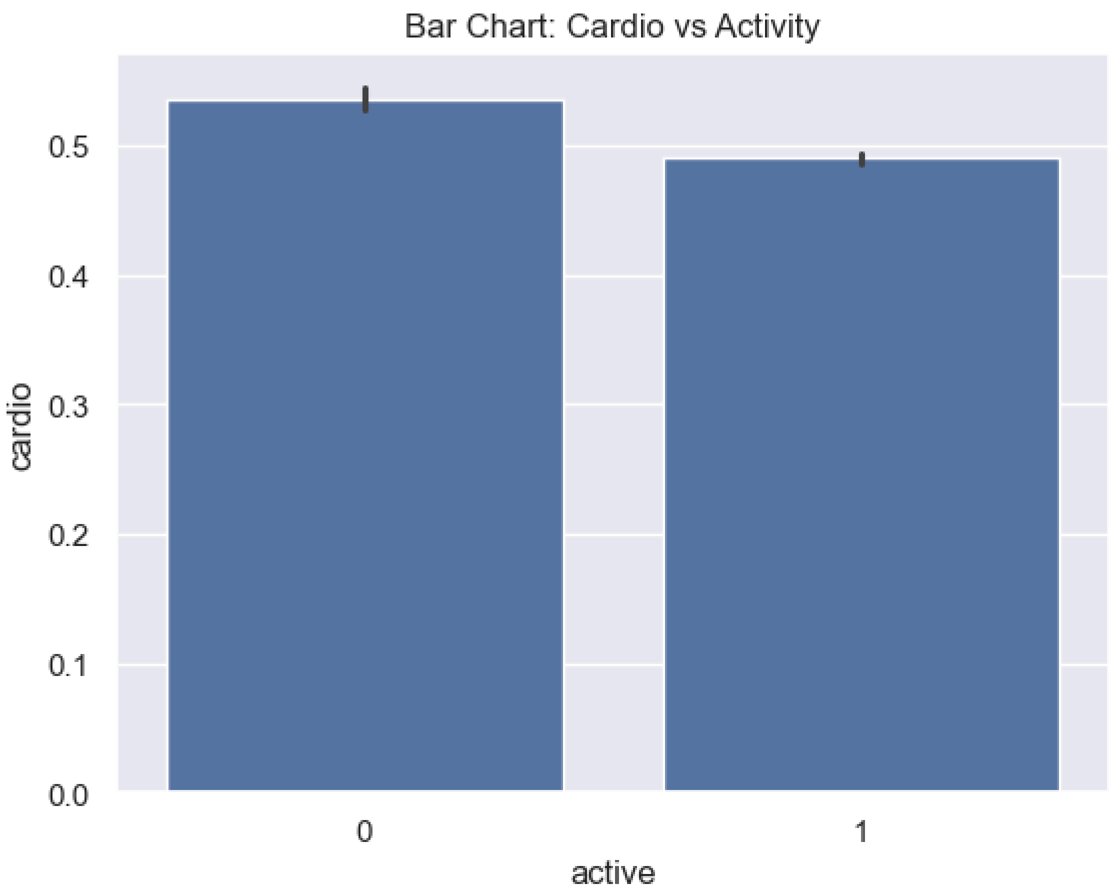

Countplot: Gender
count
40000
30000
20000
10000
0
1
2
gender


Countplot: Cholesterol
count
50000
40000
30000
20000
10000
0
1
2
3
cholesterol


Countplot: Cardio
count
35000
30000
25000
20000
15000
10000
5000
0
0
1
cardio

Barcharts


Bar Chart: Cardio vs Smoking
cardio
0.5
0.4
0.3
0.2
0.1
0.0
0
1
smoke



📊 Summary of Visualizations

🔹 **Correlation Heatmap** – Examines relationships between numerical variables.

🔹 **Scatterplots** – Explores relationships between pairs of numerical features.

🔹 **Pairplot** – Provides insights into multiple feature relationships simultaneously.

🔹 **Count Plot** – Displays the frequency distribution of categorical features.

🔹 **Bar Charts** – Analyzes relationships between categorical variables and the target variable.

💡 *These visualizations help in understanding patterns, correlations, and feature importance for better model performance!* 🚀