LAB 1

Subham Beura
B521060
CE

1. Import the libraries required for data preprocessing and data visualization.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

This section imports the required libraries for data analysis and visualization:

- **pandas (pd):** Used for data manipulation and analysis, particularly working with DataFrames.
- **numpy (np):** Provides support for numerical operations and array manipulation.
- **matplotlib.pyplot (plt):** Enables the creation of static, interactive, and animated visualizations in Python.
- **seaborn (sns):** Built on top of matplotlib, seaborn provides a high-level interface for creating informative and attractive statistical graphics. Use code with caution

Q.2) Load the dataset and read it.

```
# df = pd.read_csv("/content/drive/MyDrive/ML Lab/Subham Beura - dataset.csv")
df = pd.read_csv("./Subham Beura - dataset.csv")
df.head( )
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	targe
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	

3. **Show the head (first five) and tail (last five) instances with features.**

```
df.head(5)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
0	63	1	3	145	233	1	0	150	0	2.3	0	0
1	37	1	2	130	250	0	1	187	0	3.5	0	0
2	41	0	1	130	204	0	0	172	0	1.4	2	0
3	56	1	1	120	236	0	1	178	0	0.8	2	0
4	57	0	0	120	354	0	1	163	1	0.6	2	0

```
df.tail(5)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
298	57	0	0	140	241	0	1	123	1	0.2	1
299	45	1	3	110	264	0	1	132	0	1.2	1
300	68	1	0	144	193	1	1	141	0	3.4	1
301	57	1	0	130	131	0	1	115	1	1.2	1
302	57	0	1	130	236	0	0	174	0	0.0	1

Q4 Correlation Matrix and Heatmap

This code calculates the correlation matrix for the DataFrame `df` and visualizes it using a heatmap.

1. Calculating Correlation Matrix:

```
correlation_matrix=df.corr()
```

The `df.corr()` method computes the pairwise correlation of columns in the DataFrame, excluding null values.

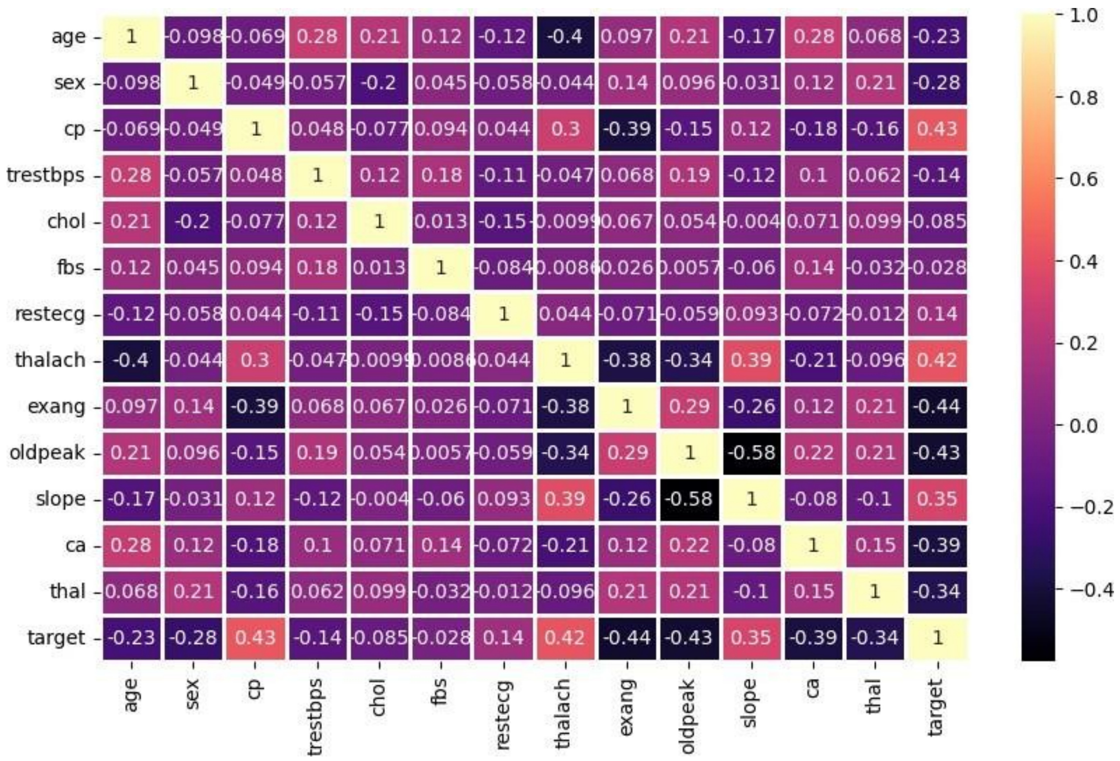
2. Creating Heatmap:

- `plt.figure(figsize=(10, 6))` sets the size of the figure for the heatmap.
- `sns.heatmap(...)` generates the heatmap using the calculated correlation matrix.
  - `annot=True` displays the correlation values on the heatmap.

  - `cmap='magma'` sets the color map for the heatmap.
  - `linewidths=1` adds lines between the cells for better visual separation.
- `plt.show()` displays the generated heatmap.

This visualization helps identify patterns and relationships between different variables in the dataset.

```
plt.figure(figsize=(10,6))
sns.heatmap(correlation_matrix,annot=True,cmap='magma',linewidths=1
plt.show()
```



Q.5) Do the exploratory data analysis by generating graphs and plots.

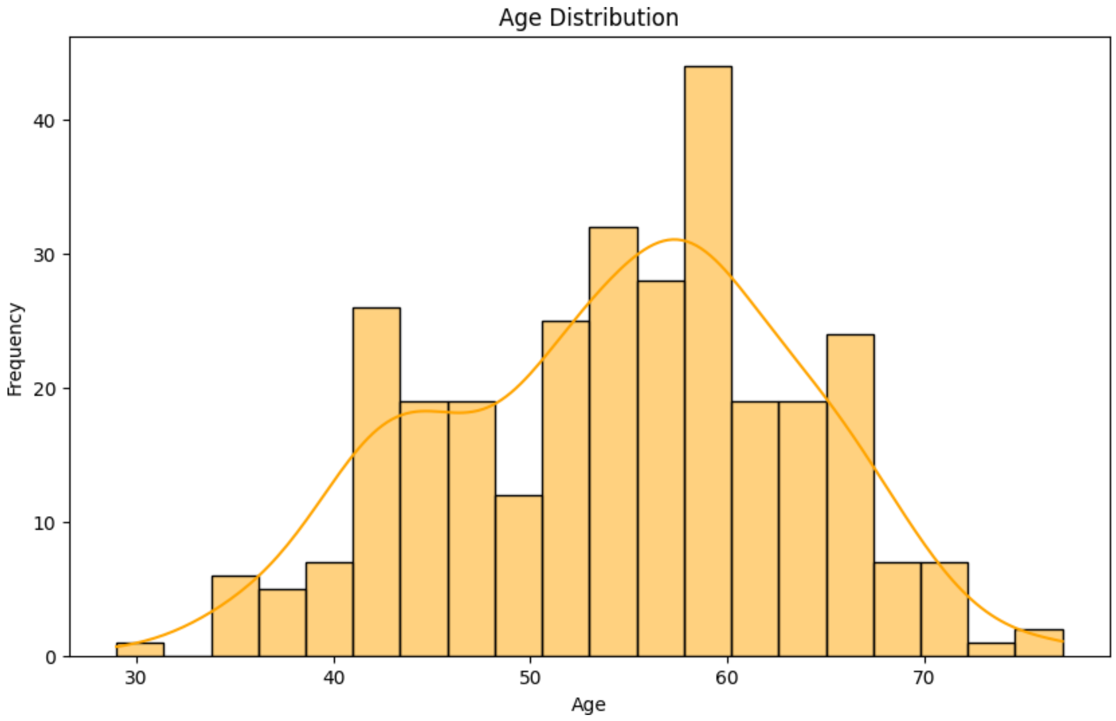
Explain your observations.

```
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restec
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.00000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.52805
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.52586
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.00000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.00000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.00000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.00000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.00000

Age Distribution Visualization

```
plt.figure(figsize=(10, 6))
sns.histplot(df['age'],kde=True,bins=20,color='orange')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



This plot illustrates the distribution of ages within the dataset.

X-axis (Age): Shows the range of ages present in the data, spanning from around 30 to beyond 70.

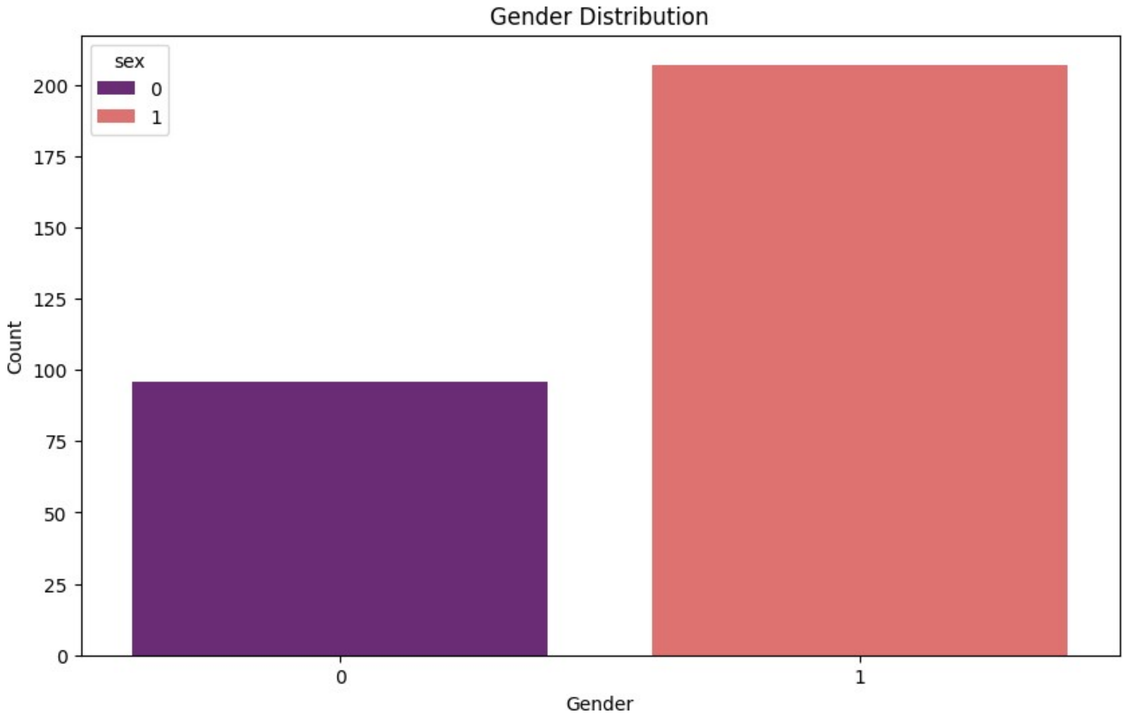
Y-axis (Frequency): Indicates how often each age value appears in the dataset. The height of each bar corresponds to the count or frequency of individuals within a specific age range.

Kernel Density Estimate (KDE) Curve: This smooth curve provides an estimated probability density function, offering a visual representation of the overall age distribution. It helps identify areas of higher and lower age concentration within the dataset.

Observation: The curve reveals a notable concentration of individuals within the age range of 55 to 65, indicating a significant portion of the dataset falls within this age bracket.

Gender Distribution

```
plt.figure(figsize=(10, 6))
sns.countplot(x='sex',data=df,palette='magma',hue="sex")
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



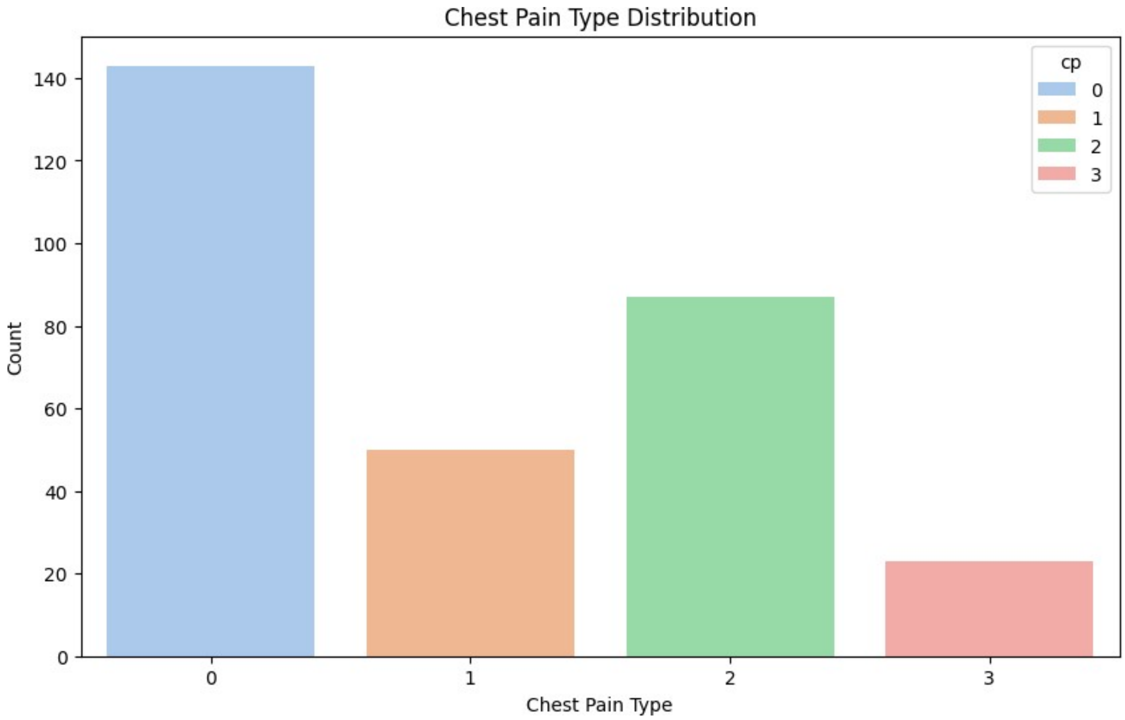
This countplot visually represents the distribution of genders within the dataset.

- **0 represents Female.**
- **1 represents Male.**

The plot clearly indicates a higher count of males compared to females in the dataset. This suggests an imbalance in gender representation, with males being more prevalent.

Chest Pain Type Distribution

```
plt.figure(figsize=(10, 6))
sns.countplot(x='cp',data=df,palette='pastel',hue="cp")
plt.title('Chest Pain Type Distribution')
plt.xlabel('Chest Pain Type')
plt.ylabel('Count')
plt.show()
```



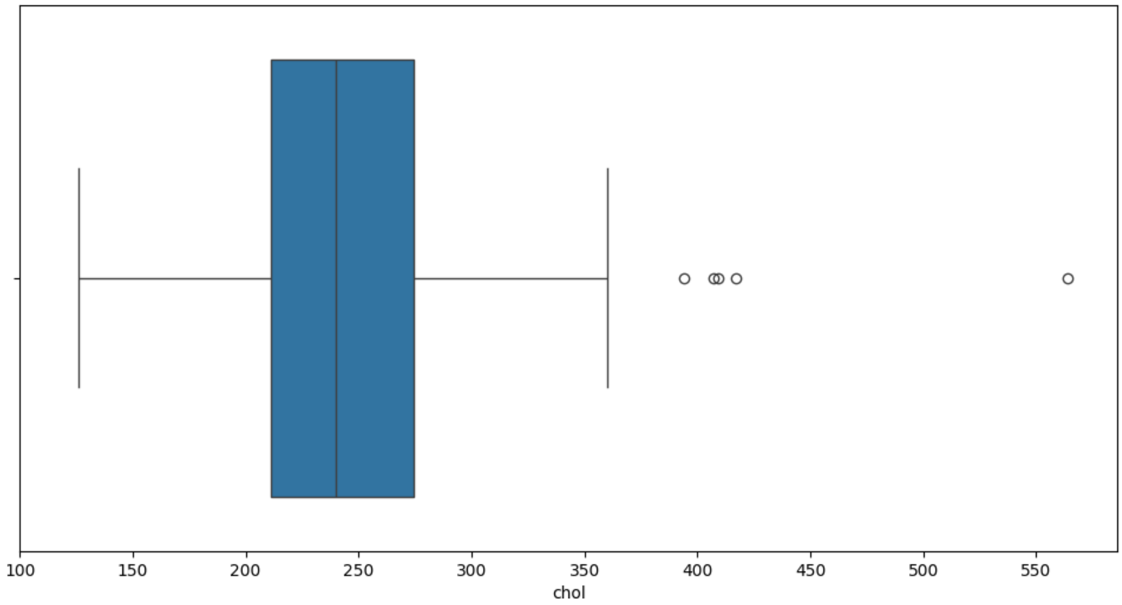
This plot illustrates the distribution of different chest pain types within the dataset.

- **Type 0:** The most prevalent type, with approximately 140 occurrences.
- **Type 2:** The second most common type, observed around 85 times.
- **Type 1:** Exhibits a moderate presence, with roughly 50 occurrences.
- **Type 3:** The least frequent type, appearing around 20 times.

This visualization provides a clear overview of the relative frequencies of each chest pain type within the dataset.

Box Plot for Outlier Detection and Distribution Analysis

```
plt.figure(figsize=(12, 6))
sns.boxplot(x=df['chol'])
plt.xticks(np.arange(100,600,50))
plt.show()
```



This box plot serves as a visual tool for identifying potential outliers and understanding the distribution of a continuous variable.

Key Features:

- **Box:** The central box represents the interquartile range (IQR), containing the middle 50% of the data.
- **Whiskers:** Lines extending from the box indicate the range of data within 1.5 times the IQR.
- **Outliers:** Points beyond the whiskers are considered potential outliers.
- **Median Line:** The line within the box represents the median value.

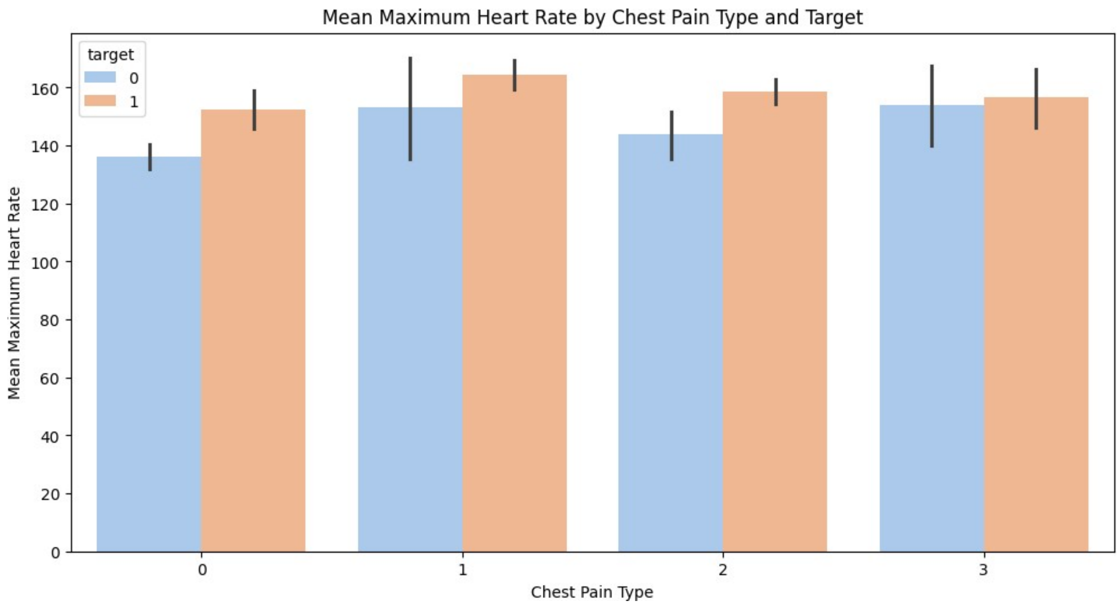
Observations:

- **Median Cholesterol Level:** The median cholesterol level appears to be slightly above 250, indicating that half of the individuals in the dataset have cholesterol levels below this value, while the other half have levels above it.
- **Interquartile Range (IQR):** The IQR spans from approximately 200 to 300, encompassing the middle 50% of cholesterol values within the dataset.
- **Whisker Range:** The whiskers extend from just below 150 to around 350, suggesting that the majority of cholesterol values fall within this range.
- **Outliers:** Cholesterol levels exceeding 350 are considered potential outliers, representing individuals with unusually high cholesterol levels compared to the rest of the dataset.

This plot helps visualize the spread and central tendency of the data, while also highlighting any extreme values that may warrant further investigation.

Bar Plot: Mean Maximum Heart Rate by Chest Pain Type and Target

```
lt.figure(figsize=(12, 6))
sns.barplot(x='cp', y='thalach', hue='target', data=df, palette='pastel')
plt.title('Mean Maximum Heart Rate by Chest Pain Type and Target')
plt.xlabel('Chest Pain Type')
plt.ylabel('Mean Maximum Heart Rate')
plt.show()
```



This bar plot illustrates the relationship between mean maximum heart rate, chest pain type, and the target variable.

X-axis (Chest Pain Type): Represents different categories of chest pain, numerically encoded as 0, 1, 2, and 3.

Y-axis (Mean Maximum Heart Rate): Represents the average maximum heart rate achieved by individuals within each chest pain type and target group.

Bars:

- **Red Bars:** Indicate the mean maximum heart rate for individuals without heart disease (target = 0).
- **Blue Bars:** Indicate the mean maximum heart rate for individuals with heart disease (target = 1).

Error Bars: The vertical lines on top of each bar represent confidence intervals for the mean maximum heart rate, providing an indication of the variability or uncertainty associated with the estimated means.

Key Insights:

- **Chest Pain Type 0:** Individuals with chest pain type 0 and no heart disease (target 0) exhibit a lower mean maximum heart rate compared to those with heart disease (target 1).
- **Chest Pain Types 1, 2, & 3:** For these chest pain types, the mean maximum heart rate remains relatively similar between target groups. However, individuals with heart disease (target 1) consistently show a slightly higher or comparable mean heart rate compared to those without heart disease (target 0).

This visualization provides a comprehensive view of the interplay between chest pain type, mean maximum heart rate, and the target variable.

Scatter Plot: Age vs. Resting Blood Pressure

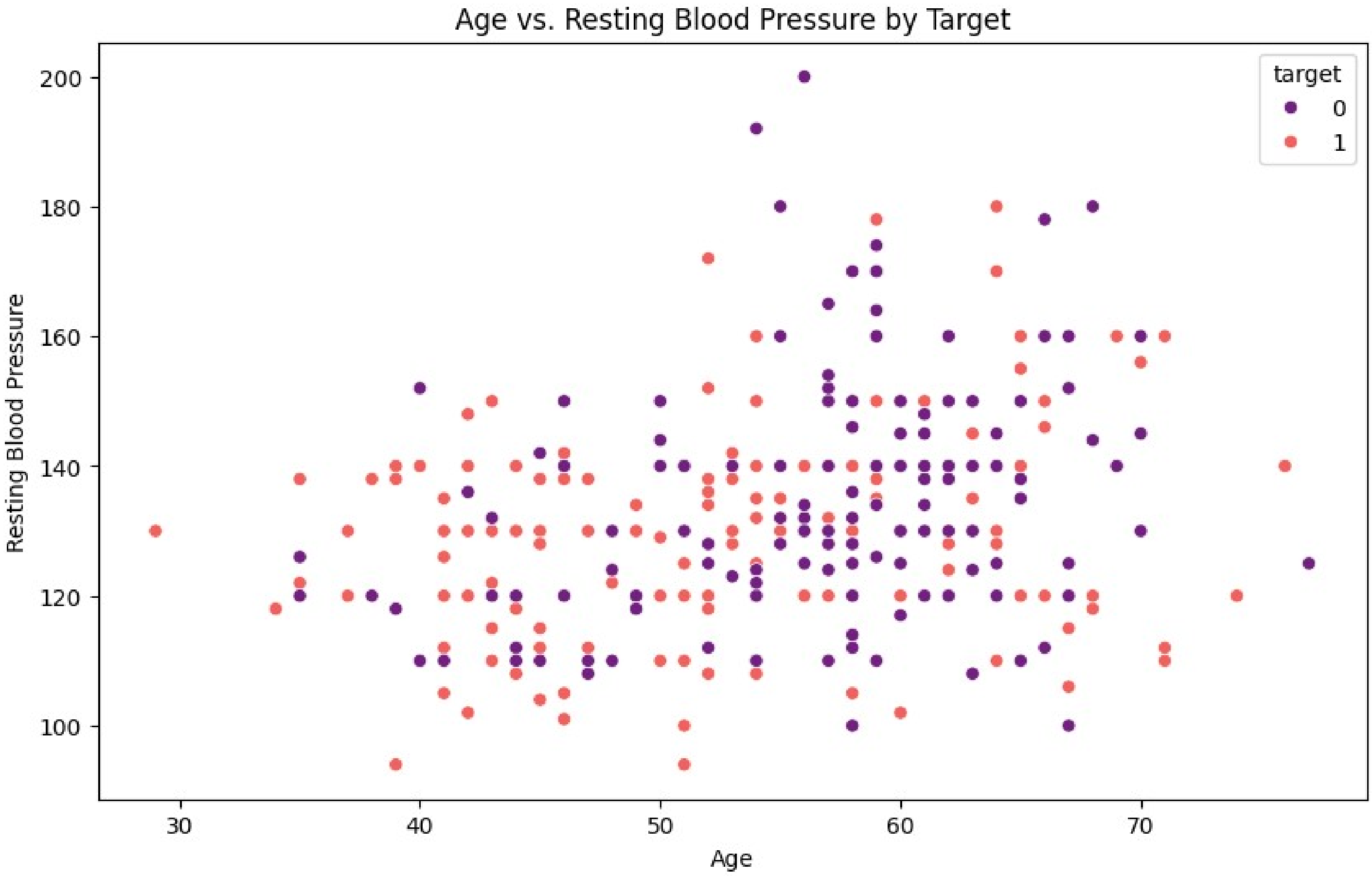
```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='age', y='trestbps', hue='target', data=df, palette='magma')
plt.title('Age vs. Resting Blood Pressure by Target')
plt.xlabel('Age')
plt.ylabel('Resting Blood Pressure')
plt.show()
```

This scatter plot visually depicts the relationship between age ('age') and resting blood pressure ('trestbps'), with data points colored according to the 'target' variable.

Observations:

Age Range: The majority of individuals in the dataset appear to fall within the age range of 40 to 70 years old.
Resting Blood Pressure Distribution: While resting blood pressure values exhibit considerable variation across age groups without a clear trend, most individuals seem to have values concentrated between 120 and 160.
Potential Outliers: Resting blood pressure values exceeding 180 could be considered potential outliers, warranting further investigation.

Relationship Exploration: The scatter plot allows for visual exploration of potential correlations or patterns between age and resting blood pressure, taking into account the 'target' variable.
This visualization provides insights into how these two variables might be associated and whether the 'target' variable influences this relationship.



Submitted by

Subham Beura
B521060
CE
7th Sem