

In [23]:



Data visualization

Data visualization is a powerful way to represent information visually, making it easier to analyze, understand, and communicate insights. Depending on the type of data and the story you want to tell, different types of visualizations are more appropriate.

Here are some of the most common types of data visualization:

1. Correlation heatmap
2. Bar Chart
3. Pie Chart
4. Box Plot
5. Scatter Plot
6. Line graph

```
In [9]: import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
```

```
In [10]: df = pd.read_csv('heart.csv')
df.head()
```

```
Out[10]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	t
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	

Features of dataset

- age (#)
- sex : 1= Male, 0= Female (Binary)
- (cp)chest pain type (4 values -Ordinal):Value 1: typical angina ,Value 2: atypical angina, Value 3: non-anginal pain , Value 4: asymptomatic
- (trestbps) resting blood pressure (#)
- (chol) serum cholesterol in mg/dl (#)
- (fbs)fasting blood sugar > 120 mg/dl(Binary)(1 = true; 0 = false)
- (restecg) resting electrocardiography results(values 0,1,2)
- (thalach) maximum heart rate achieved (#)
- (exang) exercise induced angina (binary) (1 = yes; 0 = no)
- (oldpeak) = ST depression induced by exercise relative to rest (#)
- (slope) of the peak exercise ST segment (Ordinal) (Value 1: up sloping , Value 2: flat , Value 3: down sloping)
- (ca) number of major vessels (0–3, Ordinal) colored by fluoroscopy
- (thal) maximum heart rate achieved — (Ordinal): 3 = normal; 6 = fixed defect; 7 = reversible defect

```
In [11]: df.describe()
```

```
Out[11]:
```

	age	sex	cp	trestbps	chol	fbs	r
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.000000	0.149268	0.500000
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.500000
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.000000	0.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000

Correlation heatmap

The image you provided appears to be a correlation heatmap, which visualizes the correlation coefficients between different variables in a dataset. Color Coding:

- The color gradient ranges from blue to red, indicating the strength and direction of the correlation:
 - Blue shades suggest negative correlations, meaning that as one variable increases, the other tends to decrease.
 - Red shades indicate positive correlations, meaning that as one variable increases, the other also tends to increase.
- The intensity of the color reflects the strength of the correlation: Darker shades indicate stronger correlations (close to -1 or 1), while lighter shades indicate weaker correlations (close to 0).

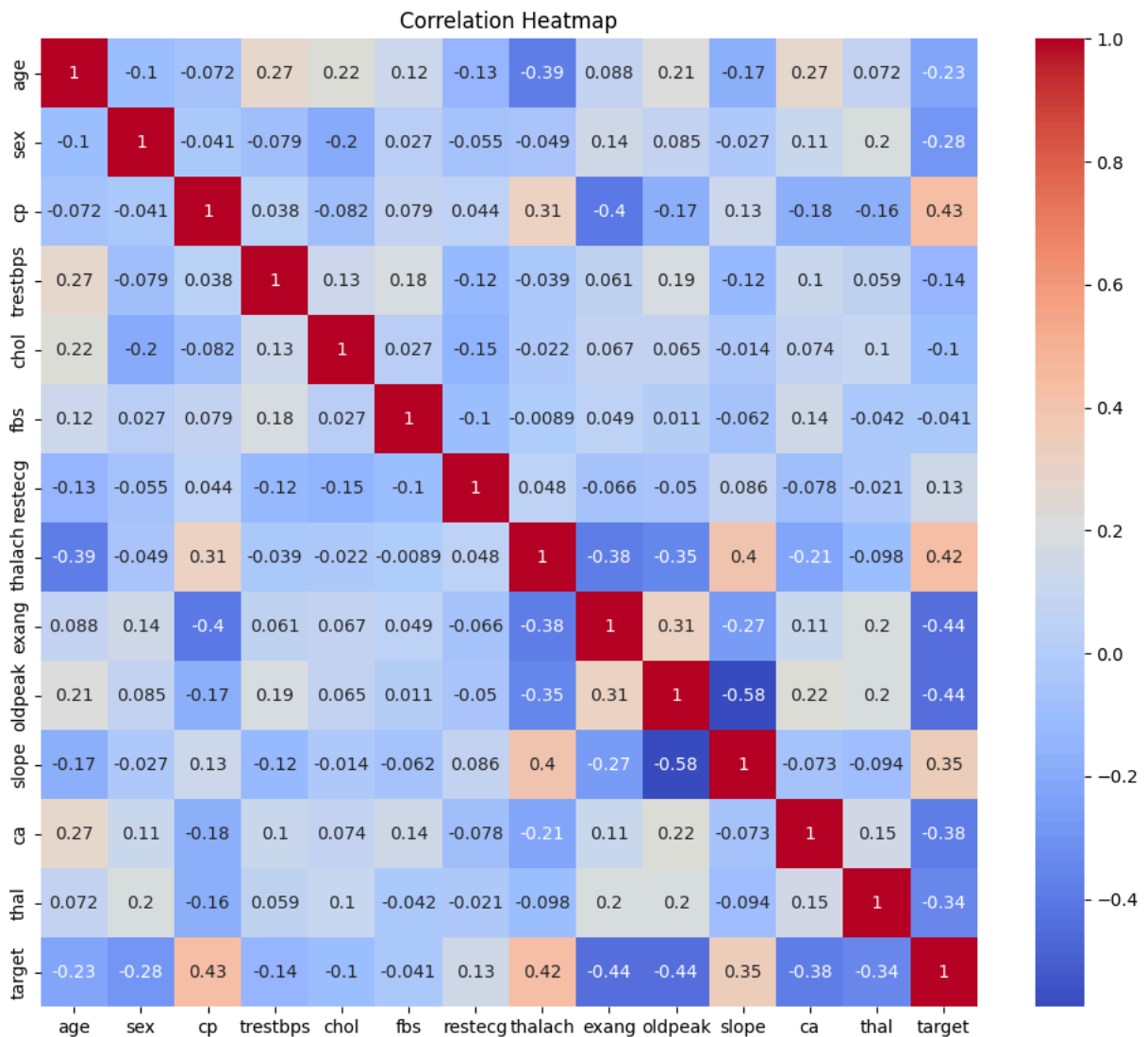
Correlation Coefficients:

- Each cell in the heatmap shows the correlation coefficient (ranging from -1 to 1) between the variables corresponding to that cell's row and column:
 - A value of 1 indicates a perfect positive correlation.
 - A value of -1 indicates a perfect negative correlation.
 - A value of 0 indicates no correlation.

Observations:

- There is a strong negative correlation between oldpeak and thalach, suggesting that higher levels of oldpeak (depression induced by exercise) are associated with lower maximum heart rates achieved.
- The correlation between sex and chol appears to be weak.
- The variable age shows a relatively strong negative correlation with thalach, indicating that as age increases, the maximum heart rate achieved tends to decrease.

```
In [12]: plt.figure(figsize=(12, 10))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



Bar Chart

A bar chart (or bar graph) is a graphical representation of data using rectangular bars. Each bar's length (or height) is proportional to the value it represents. Bar charts are commonly used to compare different categories or groups.

Axes:

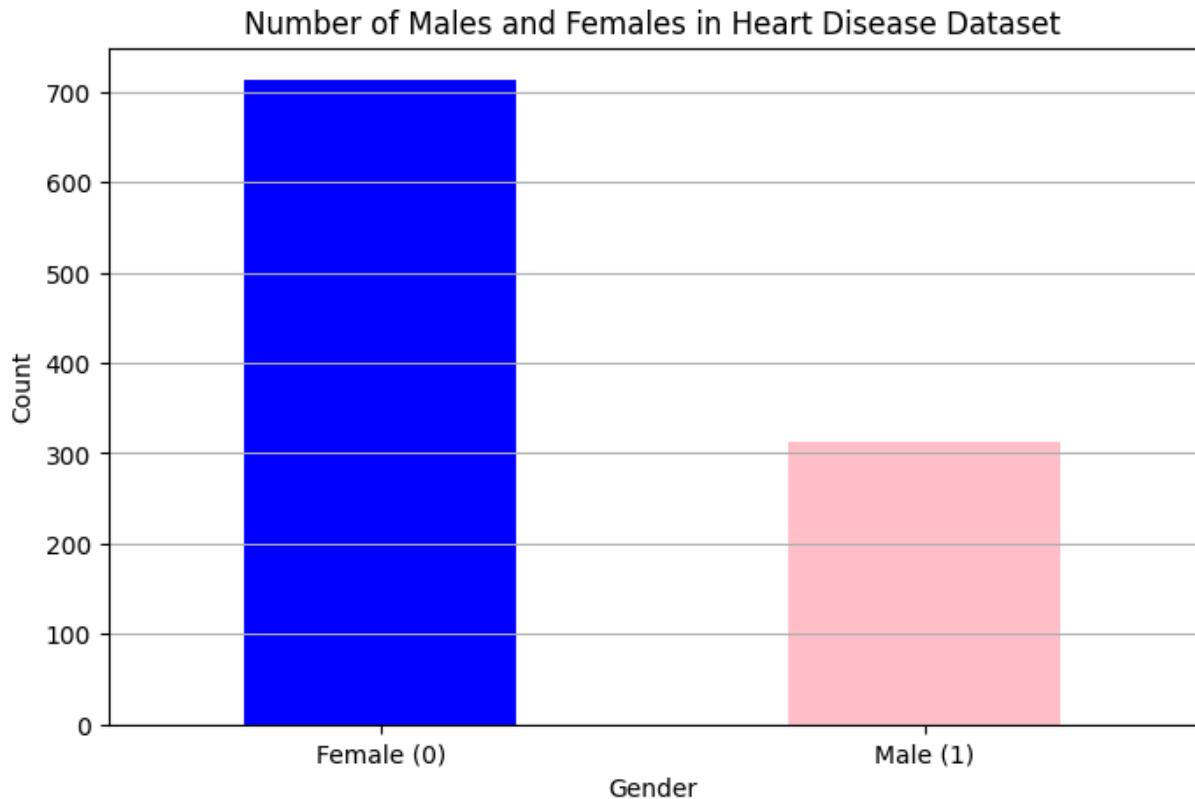
The x-axis typically represents the categories being compared.
The y-axis represents the numerical values or frequencies.

```
In [13]: # Check the gender column
gender_counts = df['sex'].value_counts() # 'sex' is typically coded as 0 (female)

# Display the counts
print("Number of males:", gender_counts[1]) # Assuming 1 is for male
print("Number of females:", gender_counts[0]) # Assuming 0 is for female
```

Number of males: 713
Number of females: 312

```
In [14]: # Create a bar chart
plt.figure(figsize=(8, 5))
gender_counts.plot(kind='bar', color=['blue', 'pink'])
plt.xticks([0, 1], ['Female (0)', 'Male (1)'], rotation=0)
plt.title('Number of Males and Females in Heart Disease Dataset')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.grid(axis='y')
plt.show()
```



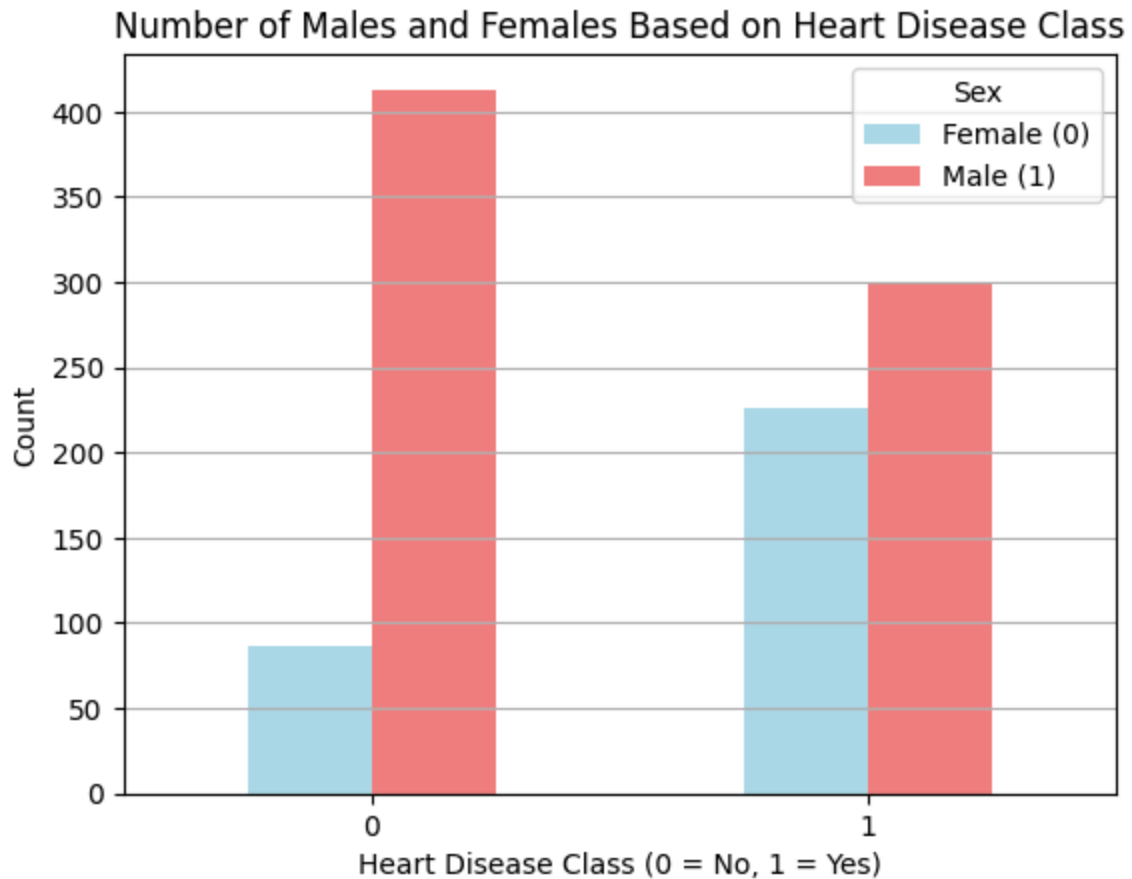
```
In [15]: # Group by target (heart disease presence) and sex, then count
sex_class_counts = df.groupby(['target', 'sex']).size().unstack(fill_value=0)

# Display the result
print(sex_class_counts)
```

sex	0	1
target		
0	86	413
1	226	300

```
In [16]: # Create a bar chart
sex_class_counts.plot(kind='bar', color=['lightblue', 'lightcoral'])
plt.title('Number of Males and Females Based on Heart Disease Class')
plt.xlabel('Heart Disease Class (0 = No, 1 = Yes)')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.legend(title='Sex', labels=['Female (0)', 'Male (1)'])
```

```
plt.grid(axis='y')
plt.show()
```



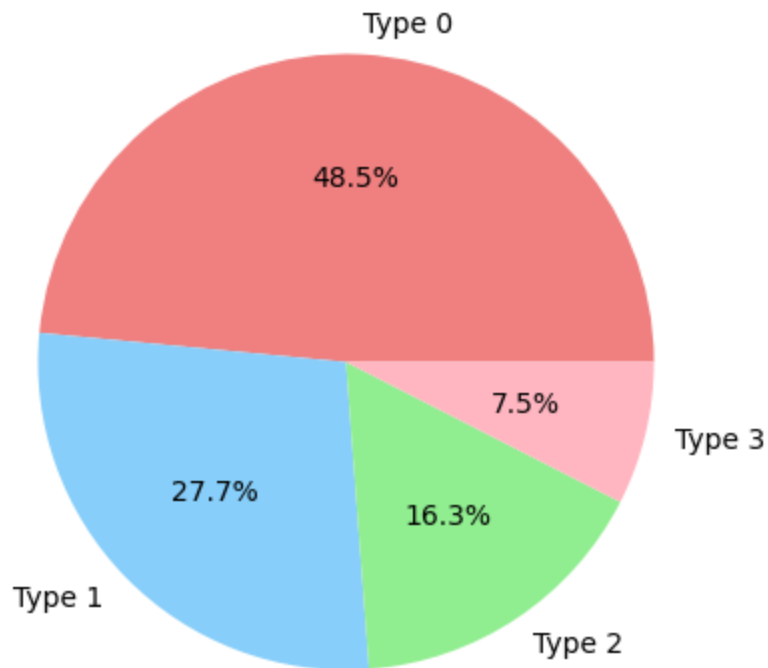
Pie chart

A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions. Each slice's arc length is proportional to the quantity it represents.

```
In [17]: cp_counts = df['cp'].value_counts()

# Create a pie chart
plt.figure(figsize=(8, 5))
plt.pie(cp_counts, labels=['Type 0', 'Type 1', 'Type 2', 'Type 3'], autopct='%1.1f%%')
plt.title('Distribution of Chest Pain Types (cp)')
plt.show()
```

Distribution of Chest Pain Types (cp)

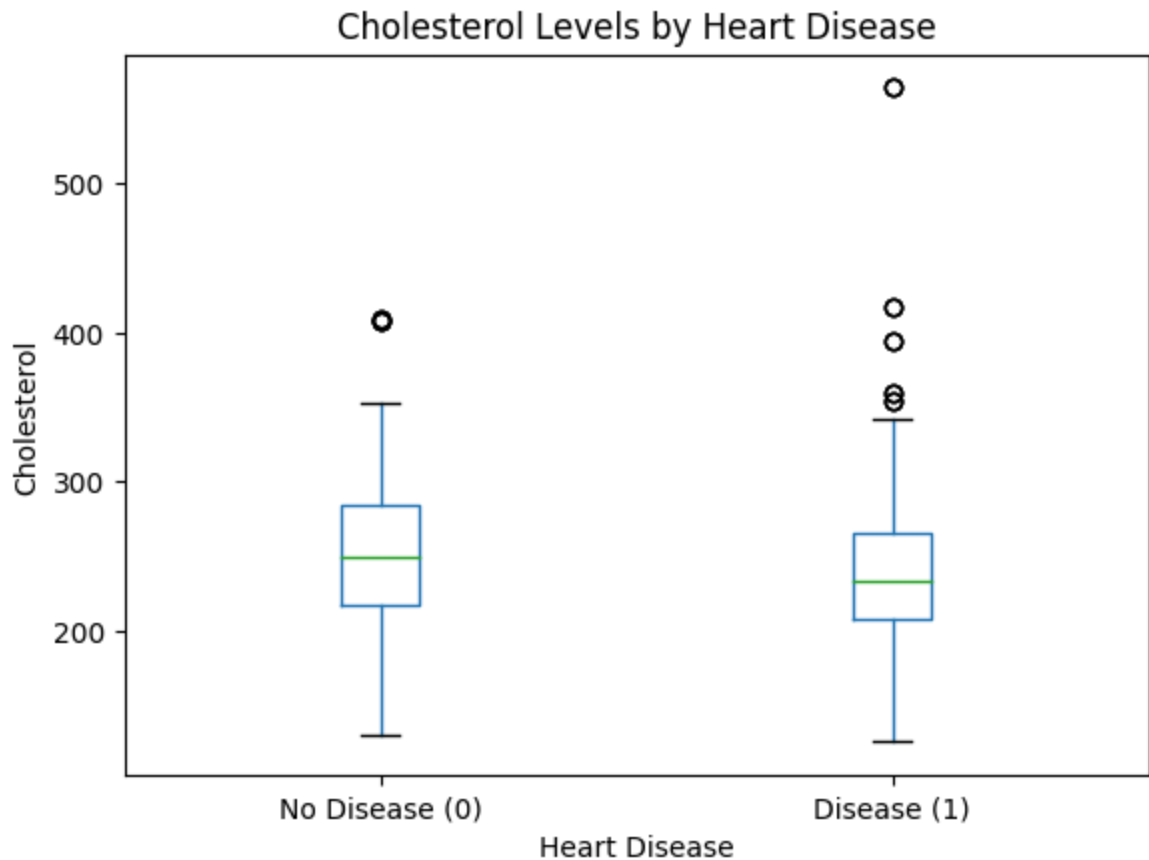


Box Plot

Cholesterol levels are often associated with heart disease risk. We can visualize the distribution of cholesterol levels among patients with and without heart disease using a box plot.

```
In [18]: # Create the box plot
plt.figure(figsize=(8, 5))
df.boxplot(column='chol', by='target', grid=False)
plt.title('Cholesterol Levels by Heart Disease')
plt.suptitle('')
plt.xlabel('Heart Disease')
plt.ylabel('Cholesterol')
plt.xticks([1, 2], ['No Disease (0)', 'Disease (1)'])
plt.show()
```

<Figure size 800x500 with 0 Axes>

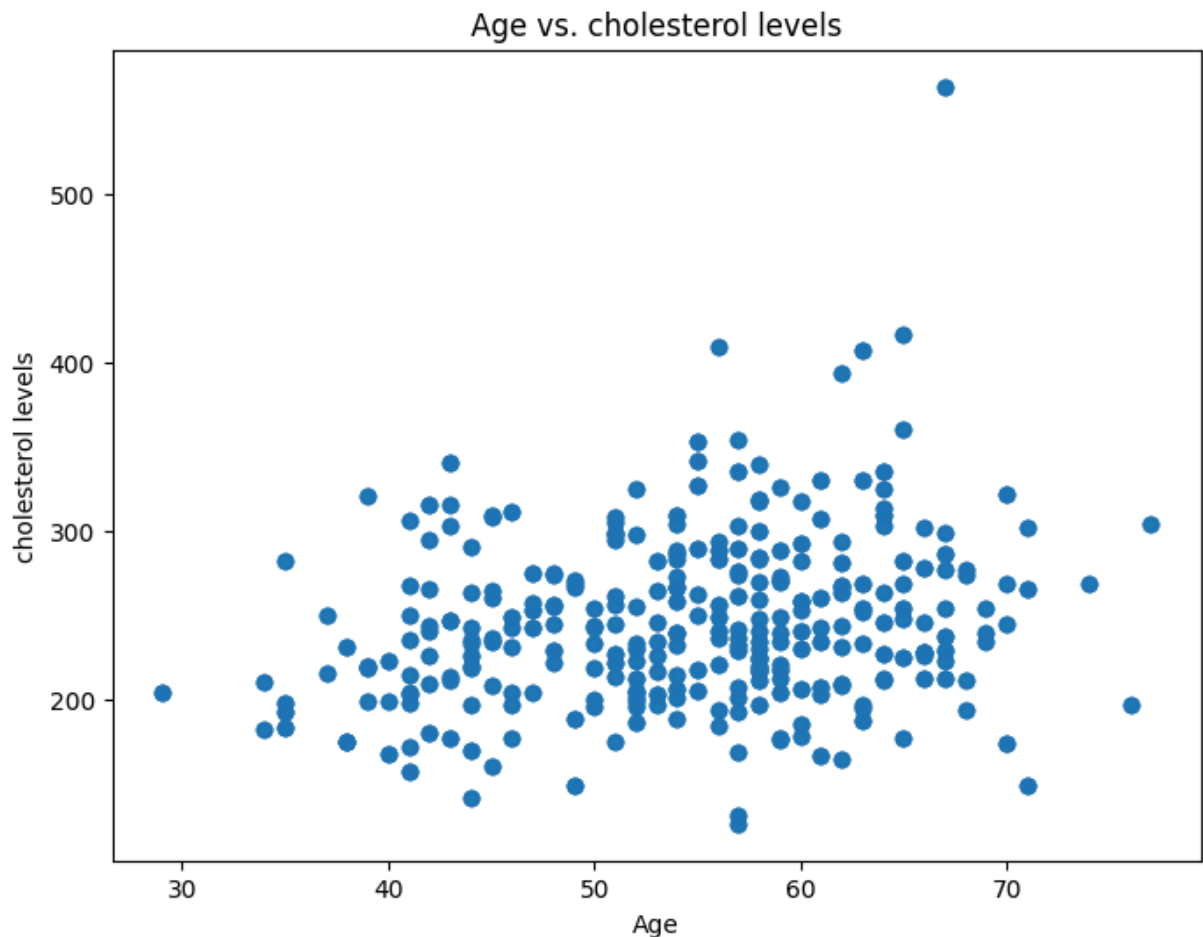


Scatter Plot

A scatter plot is a type of graph used to display the relationship or correlation between two numerical variables. It consists of points (dots), where each point represents an observation in the dataset. The position of each point is determined by the values of two variables: one plotted along the x-axis and the other along the y-axis.

We can create a scatter plot to visualize the relationship between age and cholesterol levels. The scatter plot shows that cholesterol levels vary widely at almost all age ranges, meaning age alone may not be a strong predictor of cholesterol levels based on this data.

```
In [19]: plt.figure(figsize=(8, 6))
plt.scatter(df['age'], df['chol'], alpha=0.8)
plt.title('Age vs. cholesterol levels')
plt.xlabel('Age')
plt.ylabel('cholesterol levels')
plt.show()
```

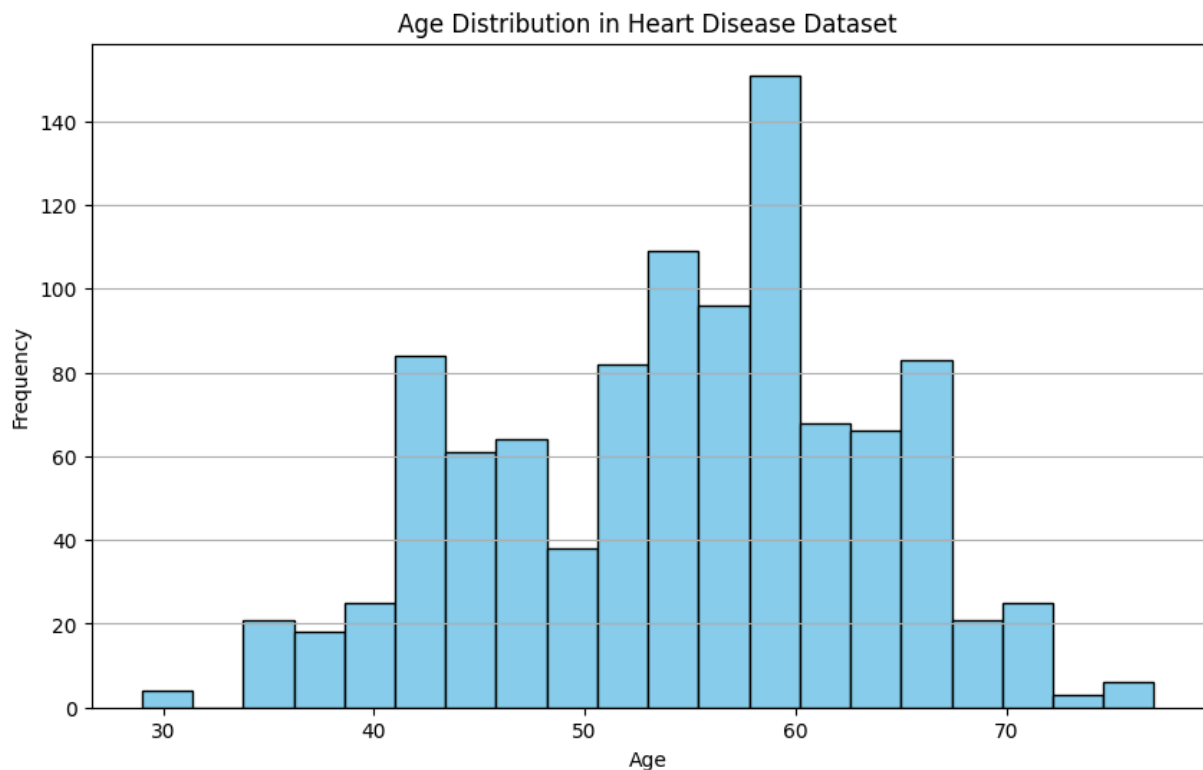


Histogram

A histogram is a graphical representation of the distribution of numerical data, showing the frequency of data points within specified ranges (bins). It helps visualize the underlying frequency distribution of a dataset, making it easier to identify patterns, trends, and outliers.

The histogram illustrates the age distribution of individuals in the heart disease dataset. It consists of 20 bins that range from approximately 30 to 80 years old. The frequency of individuals is represented on the y-axis. The highest frequency is around 60 years, indicating that this age group has the most occurrences in the dataset. This visualization helps identify age patterns related to heart disease, suggesting that middle-aged individuals are more prevalent in this dataset.

```
In [20]: plt.figure(figsize=(10, 6))
plt.hist(df['age'], bins=20, color='skyblue', edgecolor='black')
plt.title('Age Distribution in Heart Disease Dataset')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.grid(axis='y')
plt.show()
```



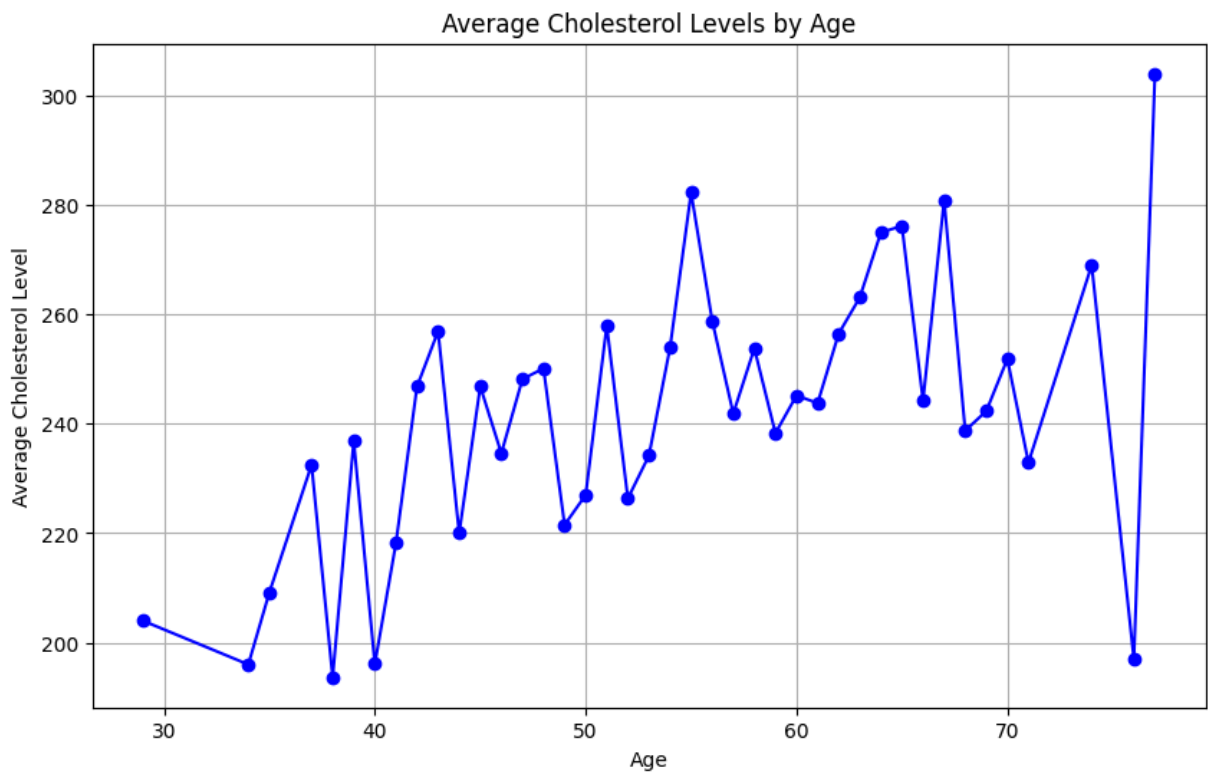
Line graph

A line graph (or line chart) is a type of chart used to visualize data points that are connected by straight lines.

Line graph presents the average of cholesterol levels by Age. The x-axis represents age (starting from around 30 years and increasing up to about 75), while the y-axis represents the average cholesterol level, with values ranging from 200 to 300.

```
In [22]: age_grouped = df.groupby('age')['chol'].mean().reset_index()

# Create a line chart
plt.figure(figsize=(10, 6))
plt.plot(age_grouped['age'], age_grouped['chol'], marker='o', linestyle='--', color=
plt.title('Average Cholesterol Levels by Age')
plt.xlabel('Age')
plt.ylabel('Average Cholesterol Level')
plt.grid()
plt.show()
```



In []: