**Data Visualization and Analytics**

**Heart Disease Dataset**

**Presented by-**

**160121748023 – Tulasi Vullendula**

**OVERVIEW**

The following data is a collection of medical and demographic data from patients. This dataset includes 1025 rows and 14 columns.

From this analysis, healthcare professionals and researchers may find a resource in identifying patients at risk and developing early intervention plans.

Estimates suggest that millions worldwide experience Heart Disease.The prevalence of heart disease can vary between men and women, and it's influenced by various factors such as age, Blood Pressure,cholesterol level etc. Generally, men tend to have a higher incidence of heart disease at a younger age compared to women, but the risk for women increases after menopause.

We set out to achieve the following objectives:

1. Data Organization

2. Exploring Relationships

**Benefits of using this dataset:**

* **Insights into Heart Disease:** The dataset provides valuable insights into Heart Disease by analyzing factors influencing heart disease and inform the development of effective prevention and treatment strategies.
* **Early Detection:** The dataset provides valuable information that can aid in the early detection of heart disease. By analyzing medical and demographic data, healthcare professionals can identify individuals at risk of developing heart disease before symptoms manifest, enabling early intervention and preventive measures.
* **Risk Stratification:** By leveraging the dataset, healthcare professionals can stratify patients based on their risk of heart disease. This risk stratification allows for targeted interventions and personalized treatment plans tailored to each patient's individual risk profile, optimizing patient outcomes and resource allocation.
* **Education and Awareness:** The dataset can serve as a valuable educational resource for healthcare professionals, students, and the general public. By disseminating information about heart disease risk factors, symptoms, and prevention strategies, the dataset can raise awareness and empower individuals to take proactive steps to protect their heart health.

**Visualization Code**:

Libraries

# Import libraries for numerical computations

import numpy as np

import pandas as pd

# Import libraries for data visualization

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express as px

# Import libraries for data preprocessing

from sklearn.preprocessing import LabelEncoder, MinMaxScaler

from sklearn.model\_selection import train\_test\_split

# Suppress warnings (use with caution)

import warnings

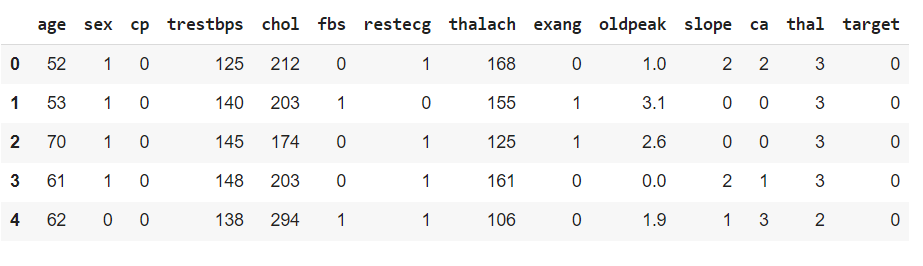
warnings.filterwarnings("ignore")

Reading Data

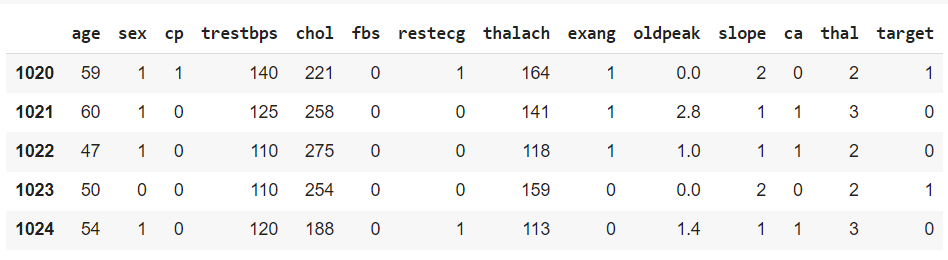
df=pd.read\_csv('/kaggle/input/heart-disease-dataset/heart.csv')

df.shape()#(1025,14)

df.head()



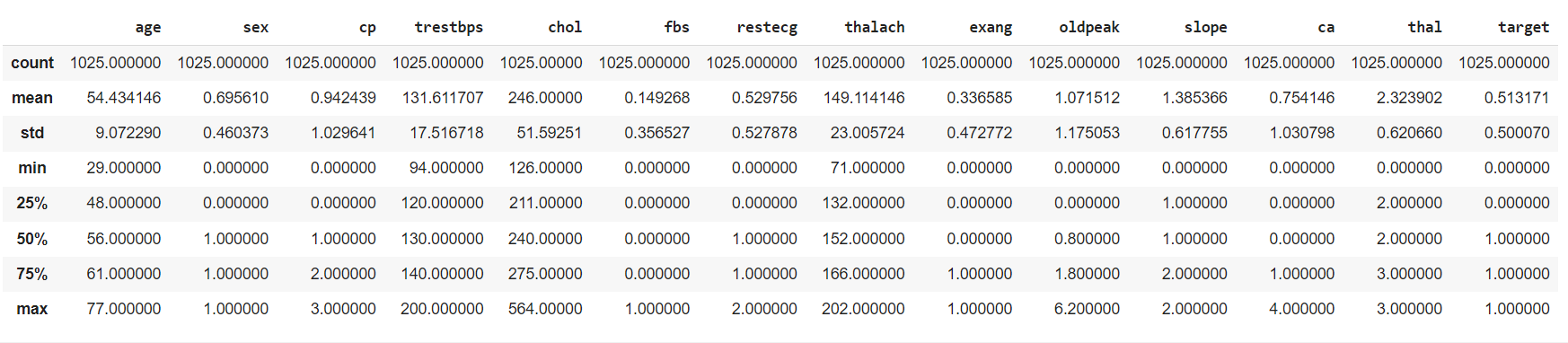
df.tail()



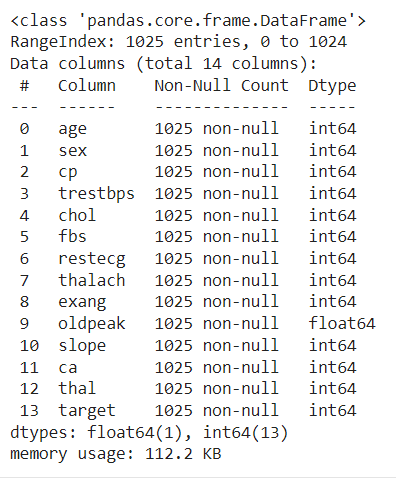
EDA

1. Gather Information About the Data

df.describe()



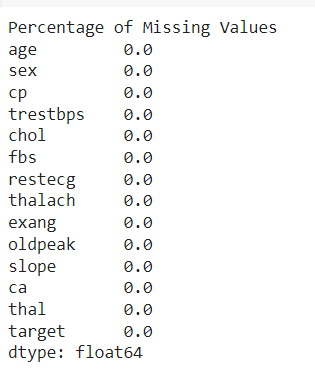
df.info()



2a) Check Missing Variables

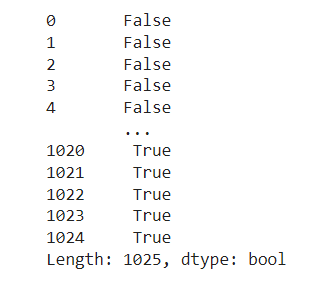
print('Percentage of Missing Values')

(df.isna().mean() \* 100).round(1)



2b) Check for Duplicate Entries, Unnecessary Columns, and Data Formatting

df.duplicated()



# Print amount of duplicates

print('Amount of duplicates: {:,}'.format(df.duplicated().sum()))

# Print percentage of duplicates

print('Percentage of duplicates: {:.1%}'.format(df.duplicated().mean()))

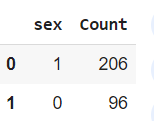
Amount of duplicates: 723

Percentage of duplicates: 70.5%

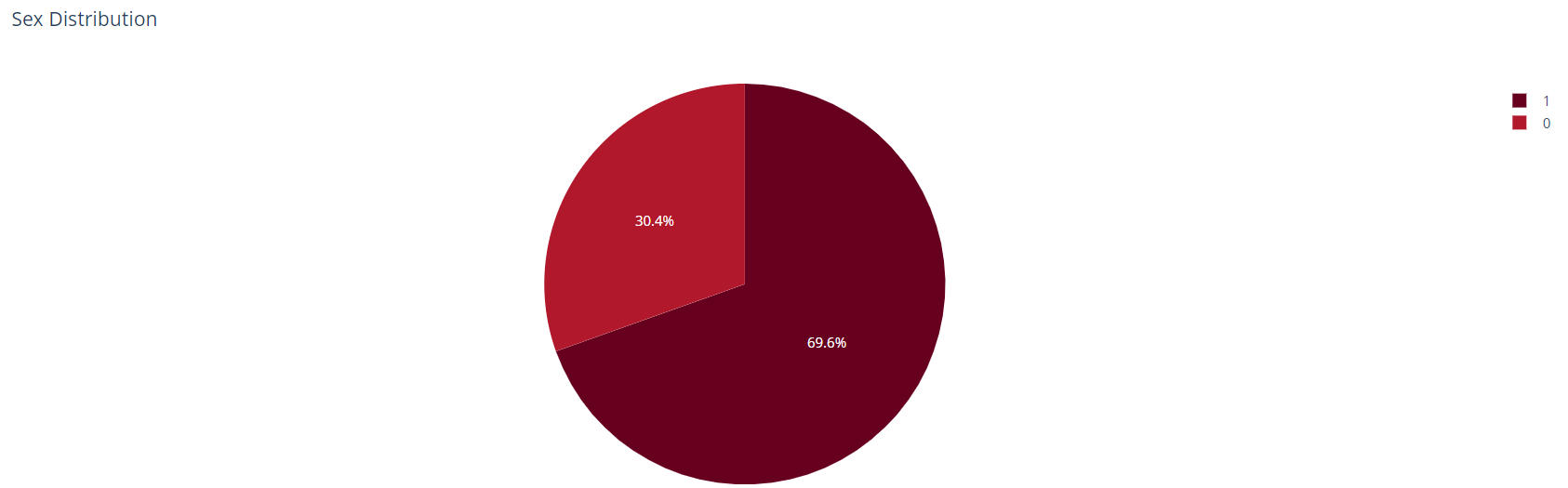
df.drop\_duplicates(inplace = True)

sex\_count=df['sex'].value\_counts().reset\_index(name='Count')

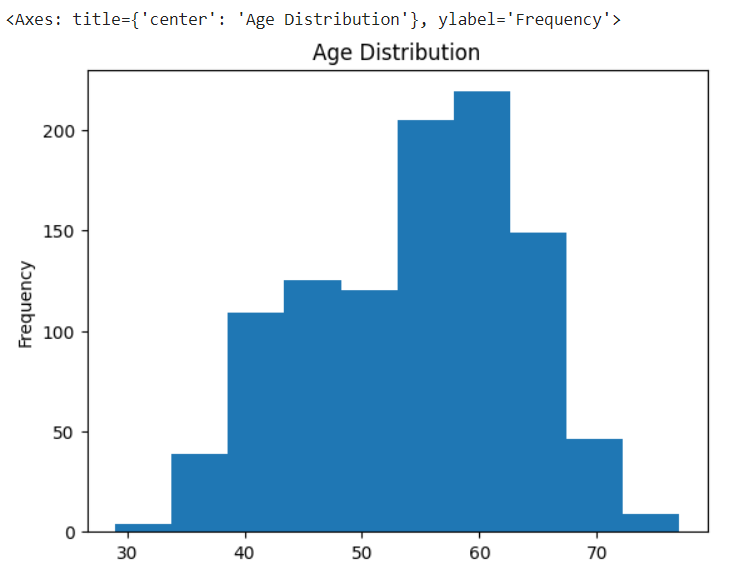
sex\_count



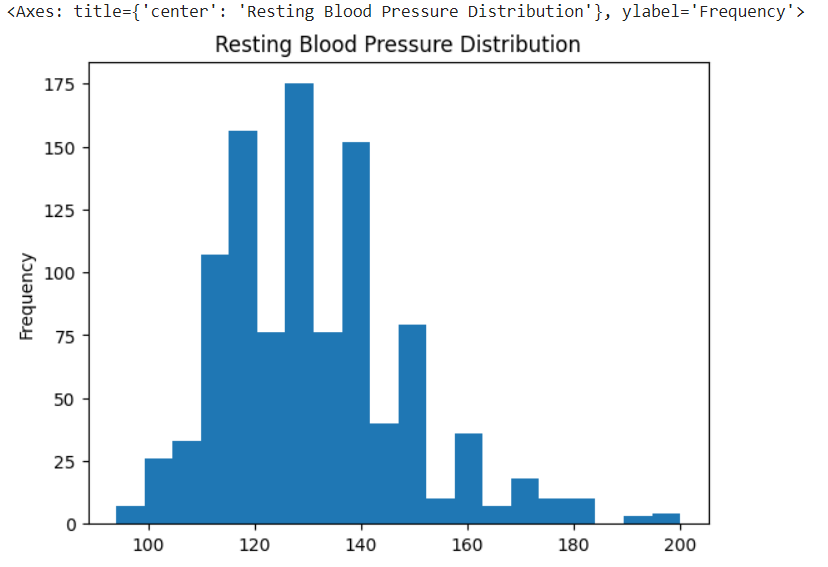
px.pie(sex\_count,names='sex',title='Sex Distribution',values='Count',color\_discrete\_sequence=px.colors.sequential.RdBu)



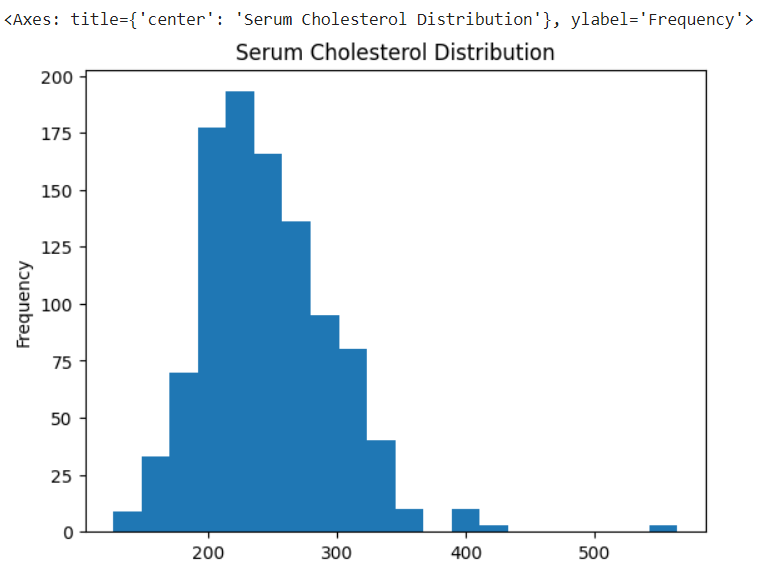
df['age'].plot(kind='hist',title='Age Distribution')



df['trestbps'].plot(kind='hist', bins=20, title='Resting Blood Pressure Distribution')



df['chol'].plot(kind='hist', bins=20, title='Serum Cholesterol Distribution')



def top\_10\_bar(category, x\_rotation = 45):

*# Get the top 10 categories and their counts in a DataFrame*

top\_10\_df = df[category].value\_counts().head(10).reset\_index()

# Plotting

sns.barplot(data = top\_10\_df, x = category, y = "count")

plt.title(category)

plt.xticks(rotation = x\_rotation)

plt.xlabel(category)

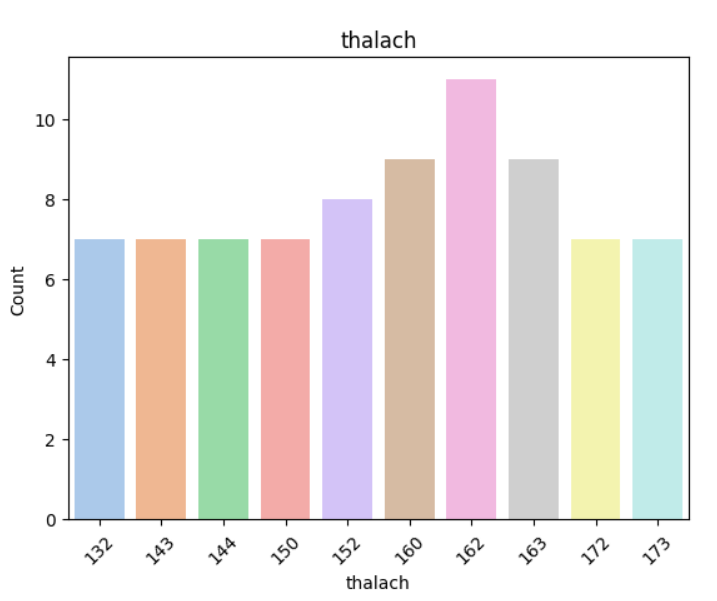
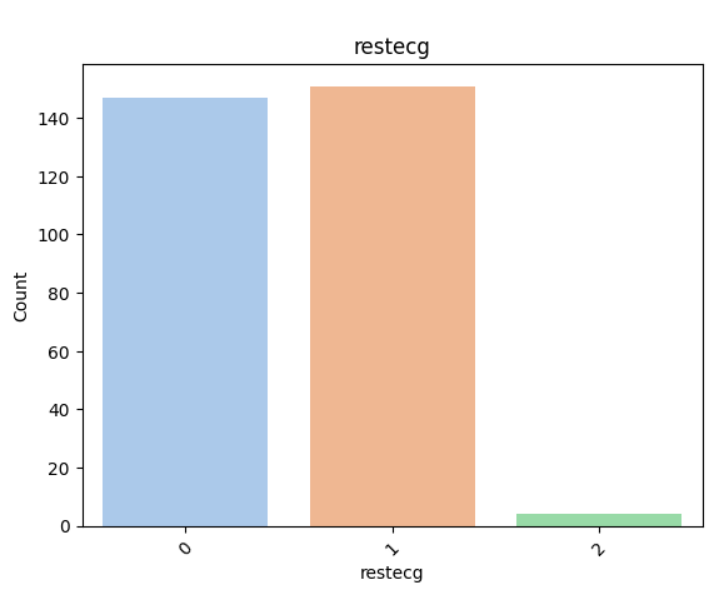
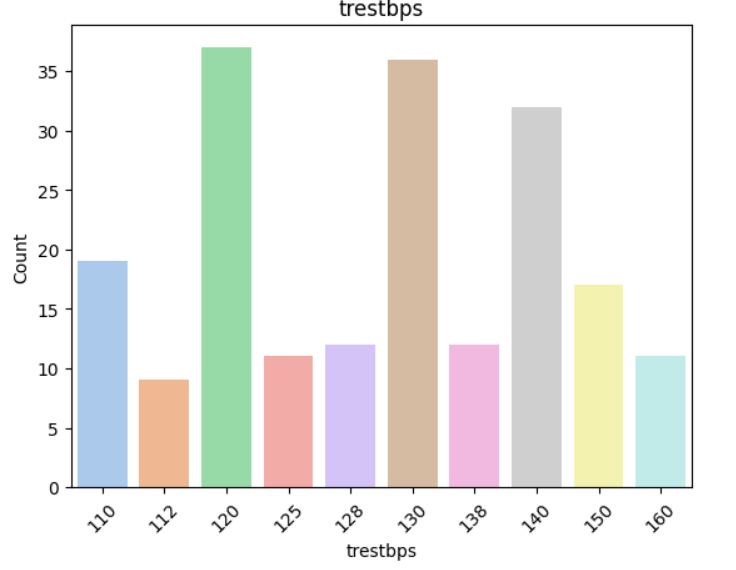
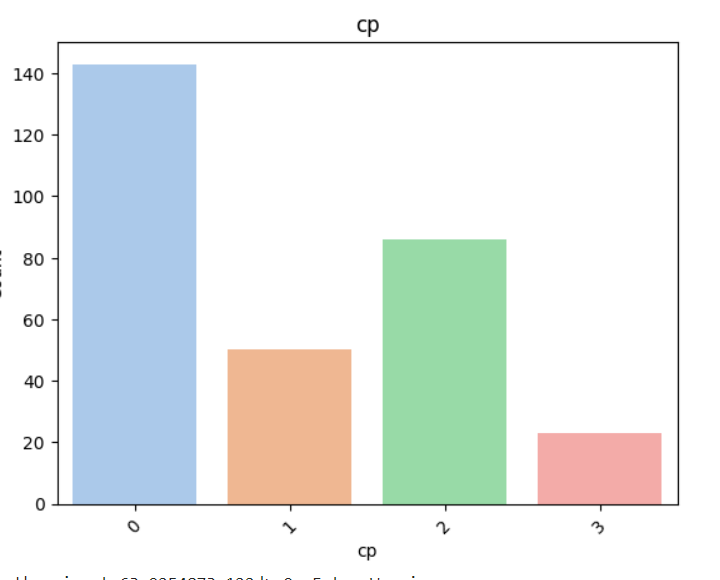
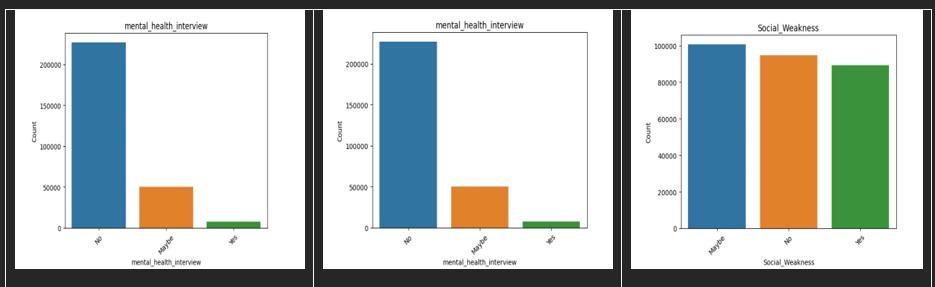
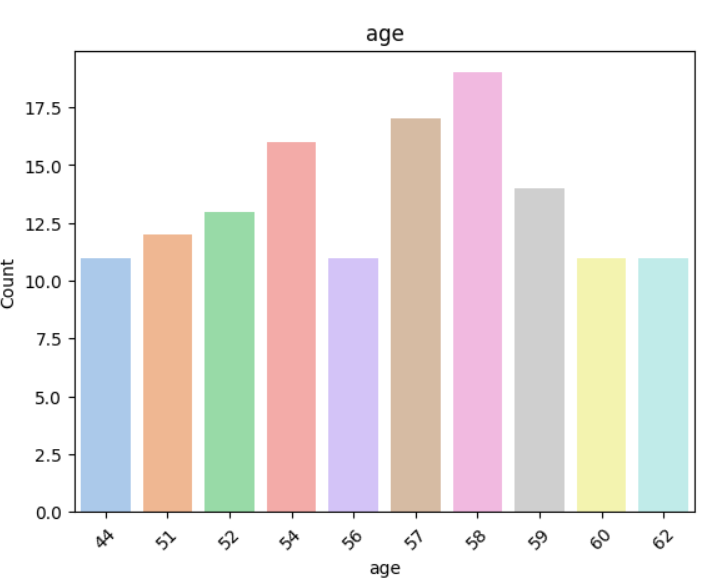
plt.ylabel("Count")

plt.show()

for column **in** df.columns:

if df[column].nunique() > 2:

top\_10\_bar(column)



*# Group data by Sex then filter by CP and count occurrences*

*gender\_data = (*

*df*

*.groupby("sex")["cp"]*

*.value\_counts()*

*.reset\_index(name="Count")*

*.sort\_values(by="Count", ascending=False)*

*)*

*import seaborn as sns*

*import matplotlib.pyplot as plt*

*# Creating the bar plot*

*sns.barplot(data=gender\_data, x="cp", y="Count", hue="sex")*

*plt.title("Chest Pain by Sex")*

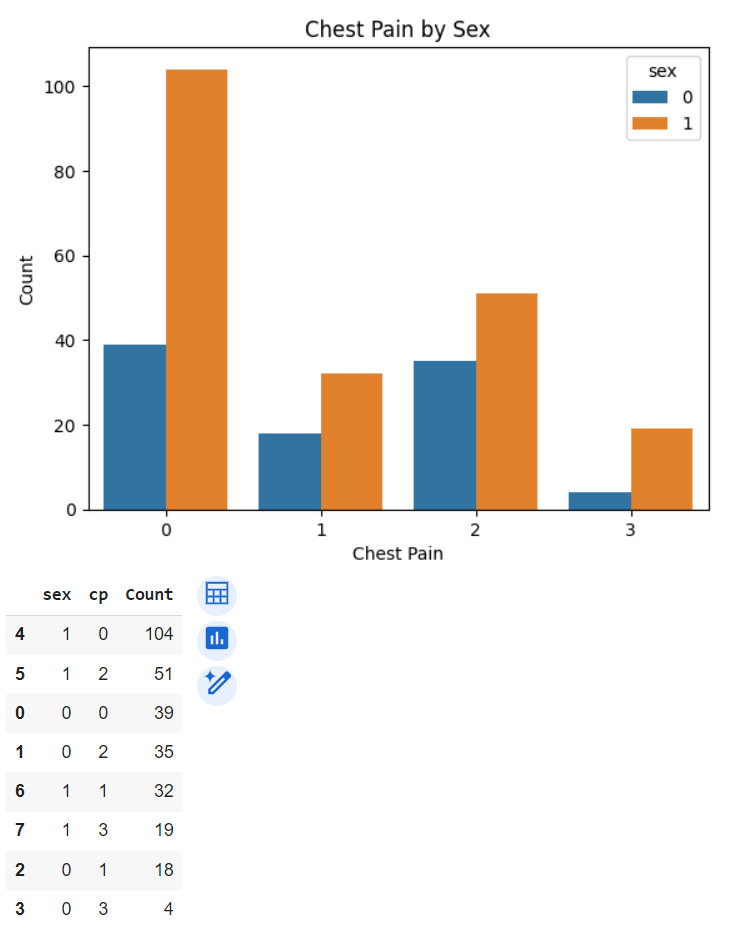
*plt.xlabel("Chest Pain")*

*plt.ylabel("Count")*

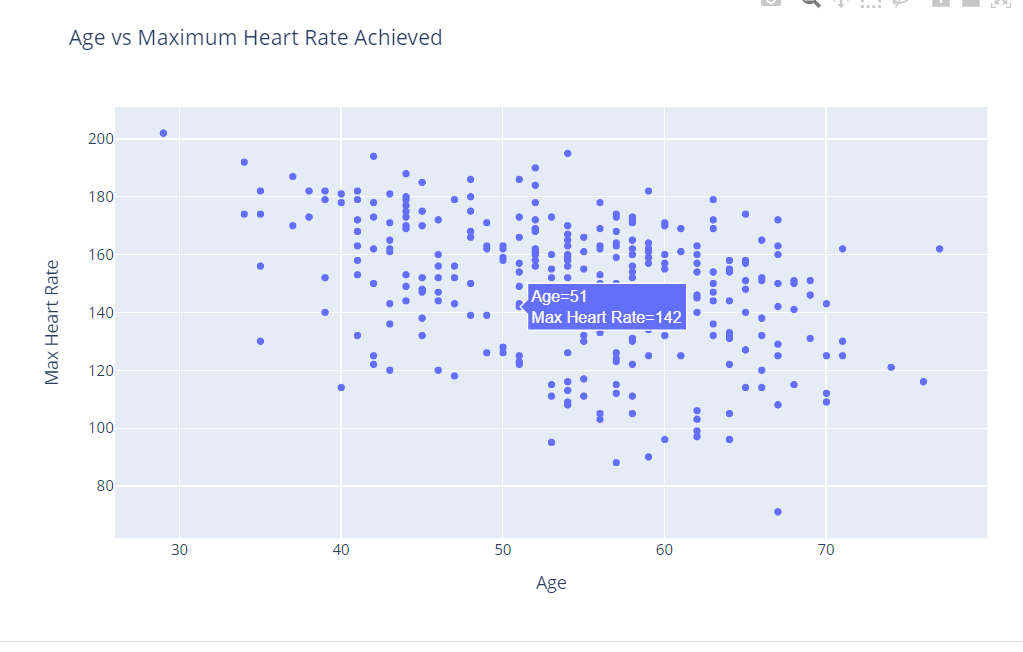
*plt.show()*

*# Displaying the DataFrame*

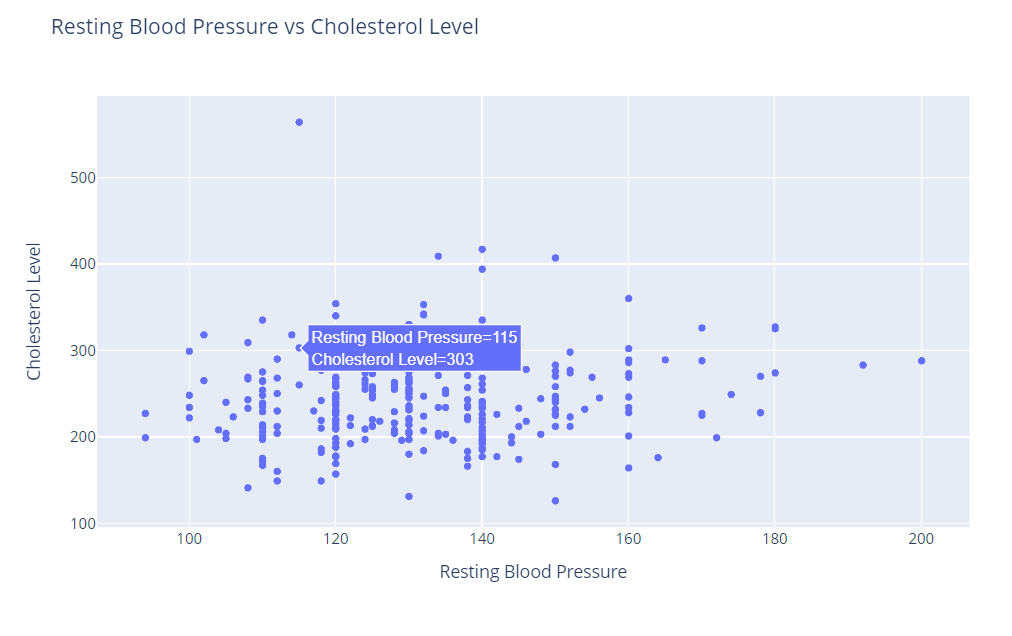
*gender\_data*



px.scatter(df, x='age', y='thalach', title='Age vs Maximum Heart Rate Achieved',labels={'age': 'Age', 'thalach': 'Max Heart Rate'})



px.scatter(df, x='trestbps', y='chol', title='Resting Blood Pressure vs Cholesterol Level',labels={'trestbps': 'Resting Blood Pressure', 'chol': 'Cholesterol Level'})



Data Preprocessing:

# Create a LabelEncoder object

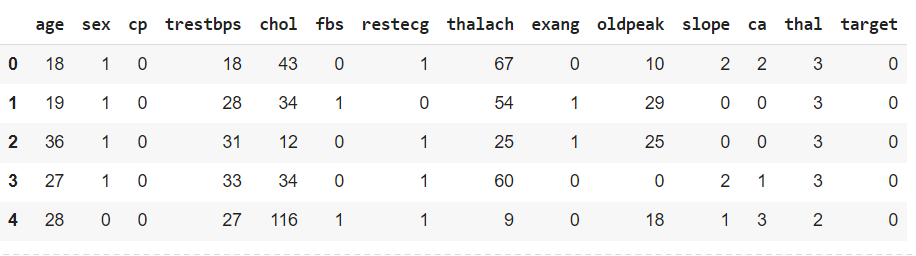
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

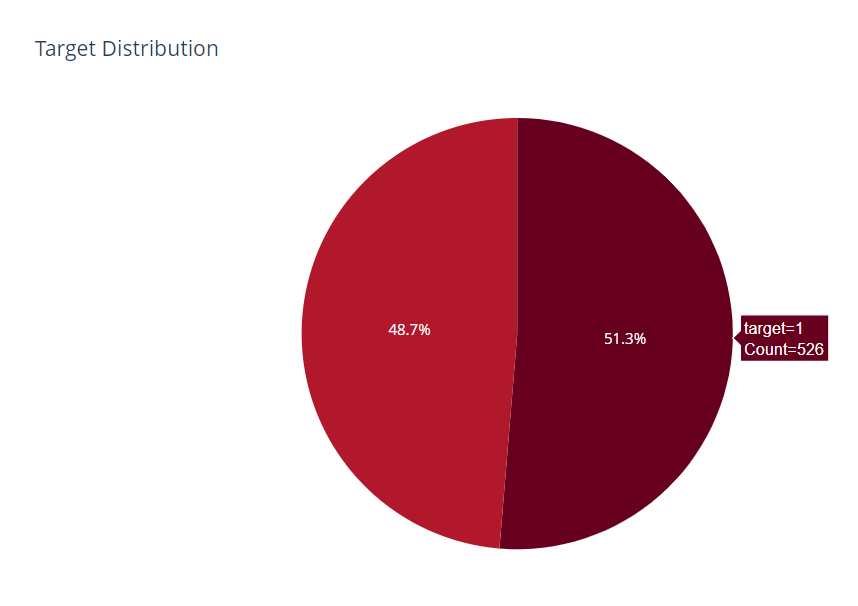
# Apply LabelEncoder to each column

encoded\_df = df.apply(le.fit\_transform)

encoded\_df.head()

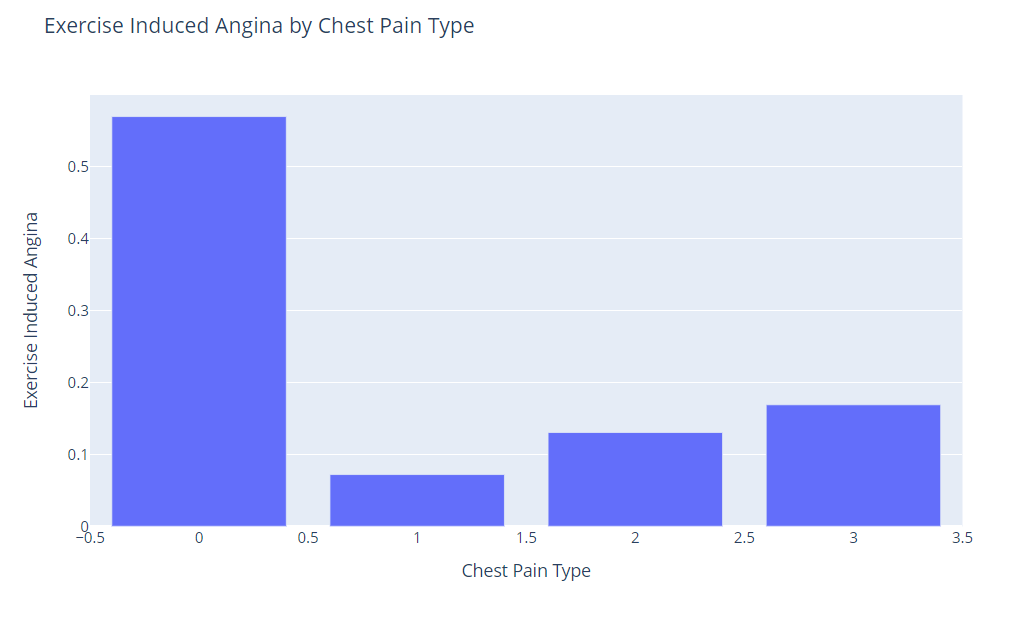


px.pie(target\_count,names='target',values='Count', title='Target Distribution',color\_discrete\_sequence=px.colors.sequential.RdBu)

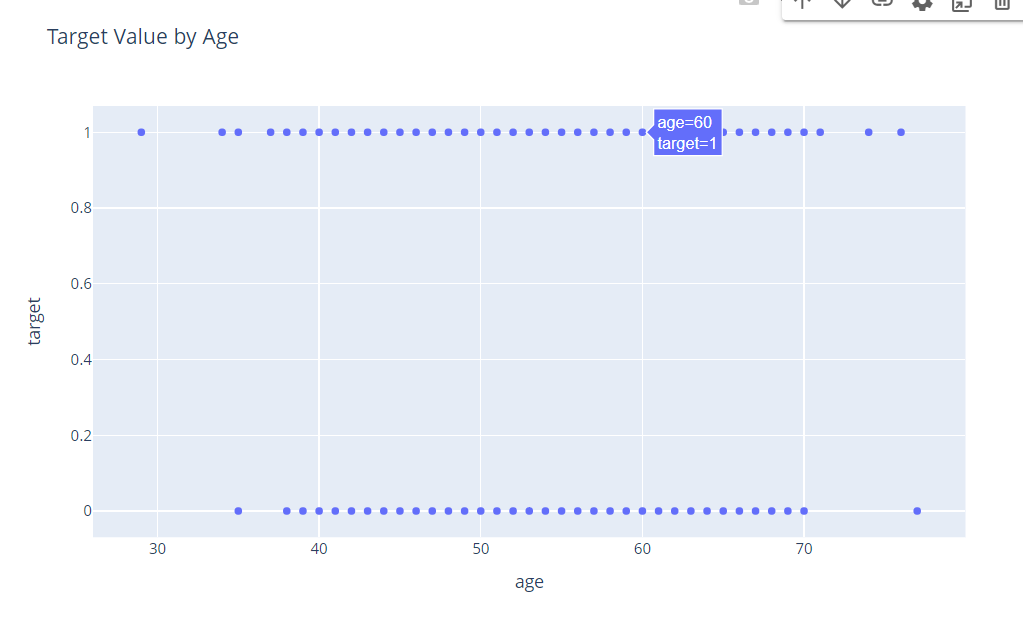


cp\_exang=df.groupby('cp')['exang'].mean().reset\_index()

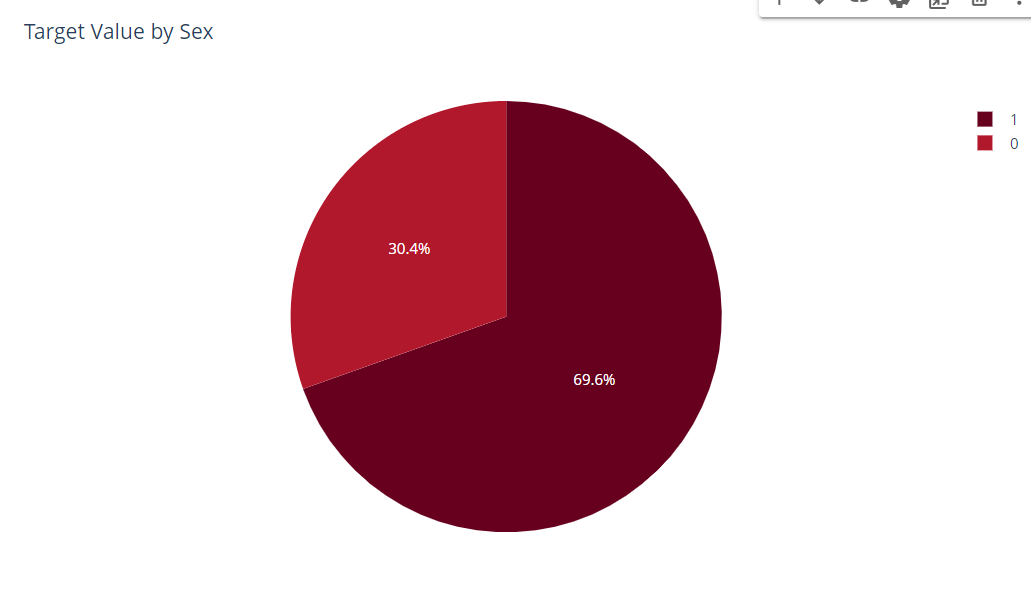
px.bar(cp\_exang, x='cp', y='exang',title='Exercise Induced Angina by Chest Pain Type',labels={'cp': ' Chest Pain Type', 'exang': 'Exercise Induced Angina'})



px.scatter(age\_target, x='age', y='target', title='Target Value by Age')

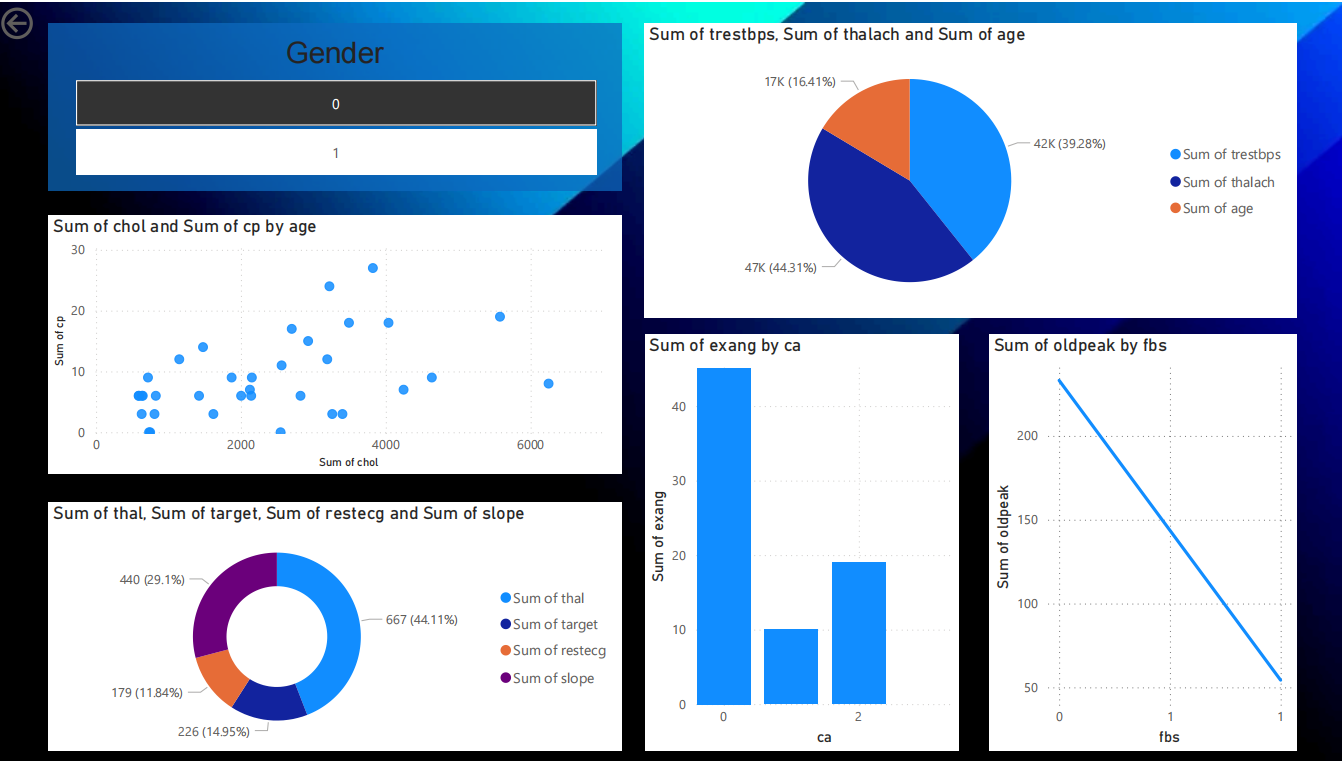


px.pie( df.groupby('sex')['target'].size().reset\_index(), names='sex', values='target', title='Target Value by Sex',color\_discrete\_sequence=px.colors.sequential.RdBu)



**Dashboard Creation for Heart Disease Dataset**



****