**Probability and Statistics**

**(MT-2005)**



**UNIVERSITY OF COMPUTER AND EMERGING SCIENCES**

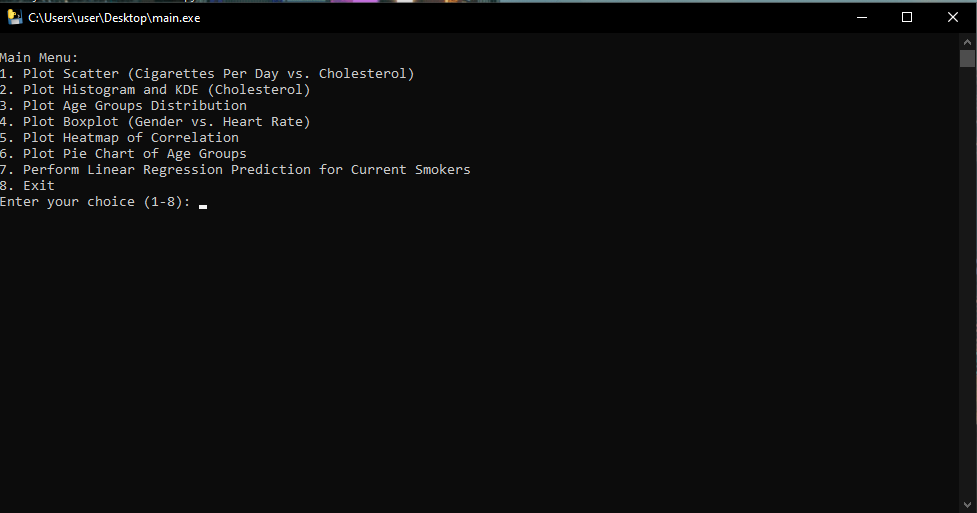
**Problem Statement:**

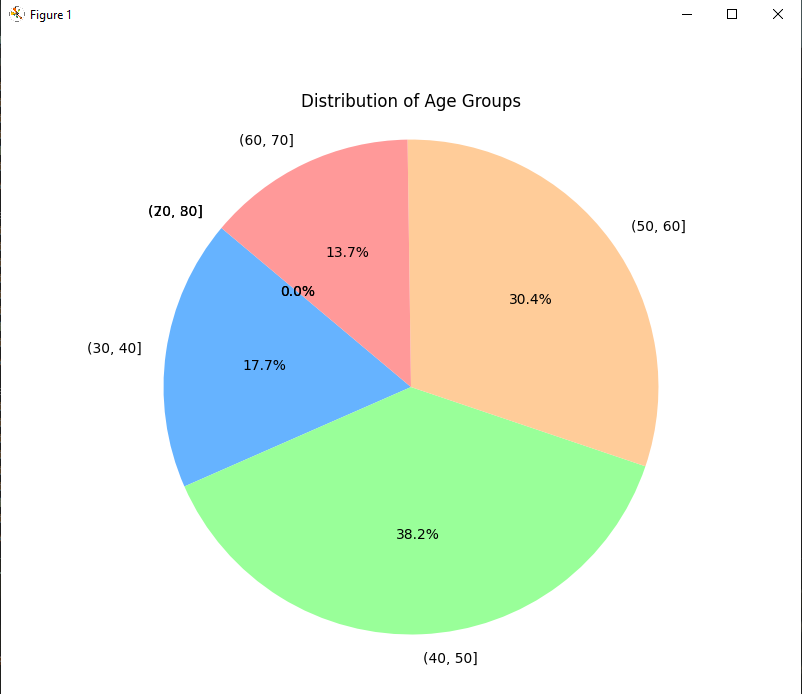
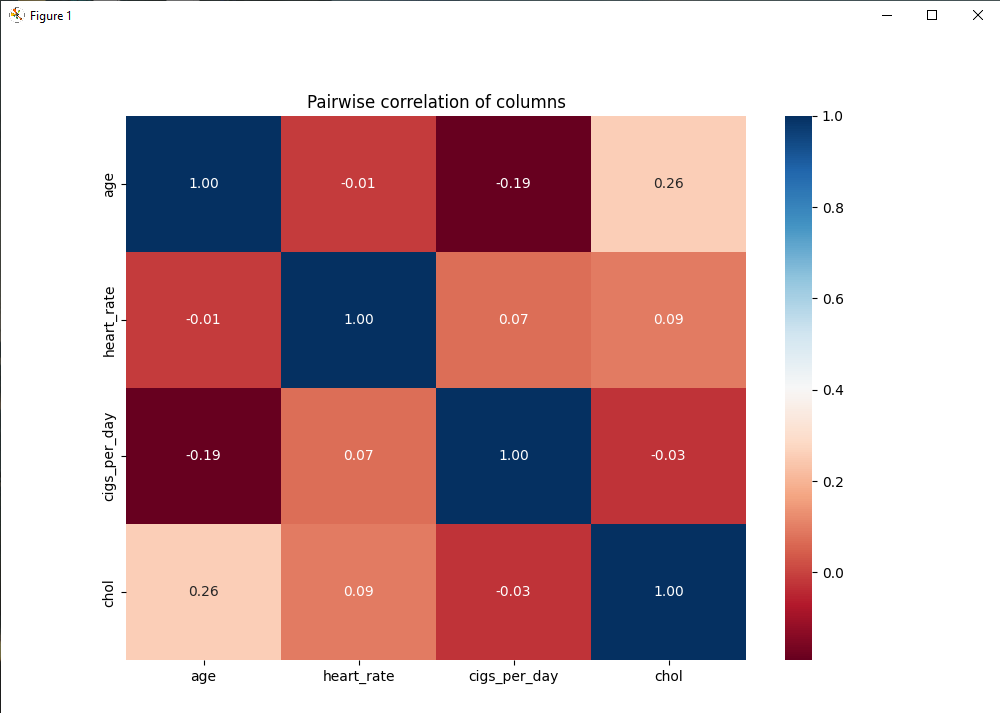
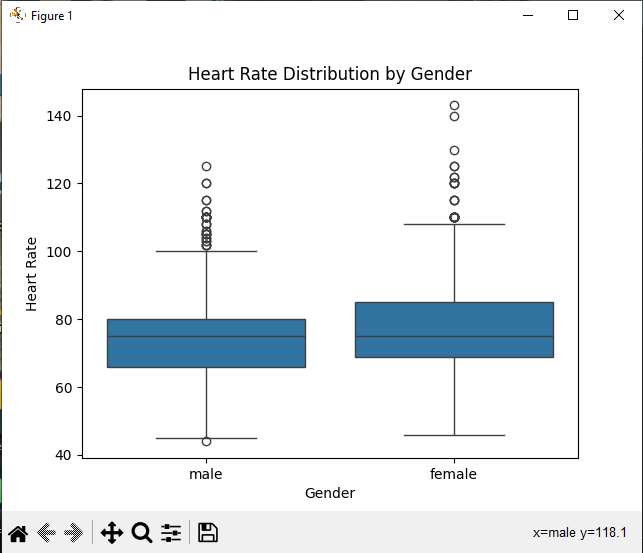
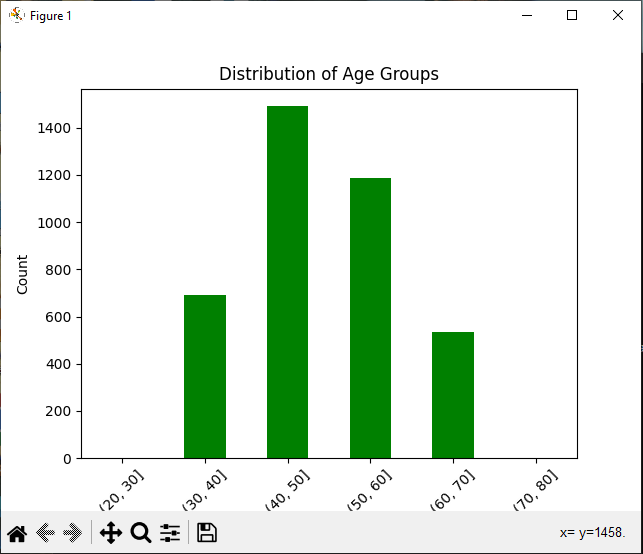
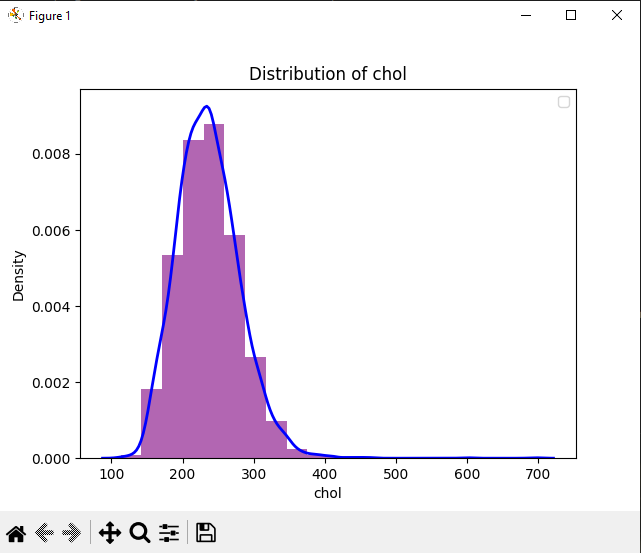
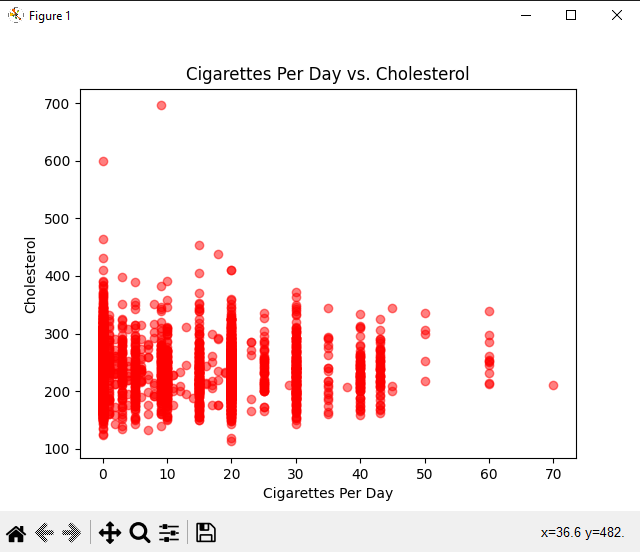
Understanding the impact of smoking on cardiovascular health is crucial for public health initiatives. Our problem revolves around analyzing a dataset that includes information on various health metrics of individuals, particularly focusing on their smoking habits and cholesterol levels. By exploring this data, we aim to identify patterns, correlations, and potentially predictive relationships to better understand the health implications of smoking and develop insights for preventive healthcare strategies.

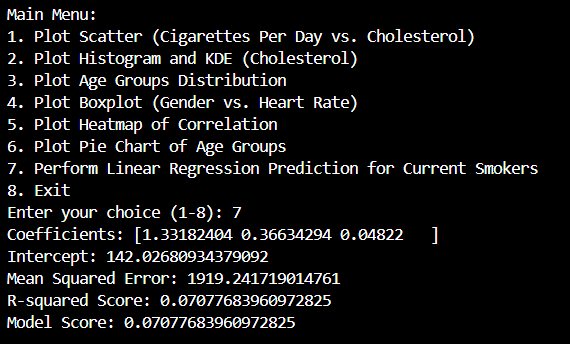
**Project Description:**

Our project utilizes a dataset containing health-related attributes of individuals, emphasizing smoking habits and cholesterol levels. Through Python programming and data visualization techniques, we delve into the dataset to extract meaningful insights. Here's an outline of our project's key components:

1. **Data Exploration and Visualization:** We start by exploring the dataset, visualizing distributions of variables like cholesterol levels, age groups, and heart rates. Visualizations such as scatter plots, histograms, box plots, and heatmaps are employed to gain an initial understanding of the data.
2. **Exploring Smoking Habits:** We investigate the relationship between smoking habits (cigarettes per day) and cholesterol levels using scatter plots and regression analysis. This step aims to discern any discernible patterns or correlations between smoking intensity and cholesterol.
3. **Analyzing Age Groups:** Understanding how age influences health metrics is crucial. We group individuals into age brackets and analyse the distribution of health parameters within each group, shedding light on age-related health trends.
4. **Examining Gender Differences:** We explore gender differences in heart rates, utilizing box plots to visualize distributions and discern any disparities.
5. **Correlation Analysis:** We investigate correlations among various health metrics using a heatmap, allowing us to identify potential relationships between different variables.
6. **Predictive Modelling:** Leveraging linear regression, we attempt to predict cholesterol levels based on age, heart rate, and smoking intensity. This predictive analysis offers insights into how these factors interplay and impact cholesterol levels.
7. **Concluding Insights:** We summarize our findings, highlighting key insights derived from the data analysis. We discuss the implications of our findings for public health initiatives, emphasizing the importance of smoking cessation and preventive healthcare measures.







Codes  
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

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

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

def plot\_scatter(df, x, y, xlabel, ylabel, title):

plt.scatter(df[x], df[y], alpha=0.5, color='red')

plt.xlabel(xlabel)

plt.ylabel(ylabel)

plt.title(title)

plt.show()

def plot\_histogram\_kde(df, column, bins=20):

plt.hist(df[column], color='purple', bins=bins, density=True, alpha=0.6)

sns.kdeplot(df[column], color='blue', linestyle='-', linewidth=2)

plt.xlabel(column)

plt.ylabel('Density')

plt.title(f'Distribution of {column}')

plt.legend()

plt.show()

def plot\_age\_groups(df):

age\_groups = pd.cut(df['age'], bins=range(20, 81, 10))

age\_groups.value\_counts().sort\_index().plot(kind='bar', color='green')

plt.xlabel('Age Group')

plt.ylabel('Count')

plt.title('Distribution of Age Groups')

plt.xticks(rotation=45)

plt.show()

def plot\_boxplot(df, x, y, xlabel, ylabel, title):

sns.boxplot(x=df[x], y=df[y])

plt.xlabel(xlabel)

plt.ylabel(ylabel)

plt.title(title)

plt.show()

def plot\_heatmap\_correlation(df):

plt.figure(figsize=(10, 10))

numeric\_df = df.\_get\_numeric\_data()

sns.heatmap(numeric\_df.corr(), annot=True, cmap='RdBu', fmt=".2f")

plt.title('Pairwise correlation of columns')

plt.show()

def plot\_pie\_chart\_age\_groups(df):

plt.figure(figsize=(8, 8))

age\_groups = pd.cut(df['age'], bins=range(20, 81, 10))

age\_groups\_counts = age\_groups.value\_counts().sort\_index()

plt.pie(age\_groups\_counts, labels=age\_groups\_counts.index, autopct='%1.1f%%', startangle=140, colors=['#ff9999','#66b3ff','#99ff99','#ffcc99'])

plt.title('Distribution of Age Groups')

plt.axis('equal')

plt.show()

def perform\_linear\_regression(df):

# Fill NaN values as zero

df.fillna(0, inplace=True)

X = df[['age', 'heart\_rate', 'cigs\_per\_day']]

y = df['chol']

model = LinearRegression()

model.fit(X, y)

y\_pred = model.predict(X)

mse = mean\_squared\_error(y, y\_pred)

r2 = r2\_score(y, y\_pred)

model\_score = model.score(X, y)

print("Coefficients:", model.coef\_)

print("Intercept:", model.intercept\_)

print("Mean Squared Error:", mse)

print("R-squared Score:", r2)

print("Model Score:", model\_score)

def main\_menu(df):

while True:

print("\nMain Menu:")

print("1. Plot Scatter (Cigarettes Per Day vs. Cholesterol)")

print("2. Plot Histogram and KDE (Cholesterol)")

print("3. Plot Age Groups Distribution")

print("4. Plot Boxplot (Gender vs. Heart Rate)")

print("5. Plot Heatmap of Correlation")

print("6. Plot Pie Chart of Age Groups")

print("7. Perform Linear Regression Prediction for Current Smokers")

print("8. Exit")

choice = input("Enter your choice (1-8): ")

if choice == '1':

plot\_scatter(df, 'cigs\_per\_day', 'chol', 'Cigarettes Per Day', 'Cholesterol', 'Cigarettes Per Day vs. Cholesterol')

elif choice == '2':

plot\_histogram\_kde(df, 'chol')

elif choice == '3':

plot\_age\_groups(df)

elif choice == '4':

plot\_boxplot(df, 'sex', 'heart\_rate', 'Gender', 'Heart Rate', 'Heart Rate Distribution by Gender')

elif choice == '5':

plot\_heatmap\_correlation(df)

elif choice == '6':

plot\_pie\_chart\_age\_groups(df)

elif choice == '7':

perform\_linear\_regression(df)

elif choice == '8':

print("Exiting the program...")

break

else:

print("Invalid choice. Please enter a number from 1 to 8.")

# Read data

url = "https://github.com/umairaltaf982/Smokers-\_Health\_Analysis/raw/main/smoking\_health\_data\_final.csv"

df = pd.read\_csv(url)

# Main menu

main\_menu(df)

**Conclusion:**

In conclusion, our project provides valuable insights into the relationship between smoking habits and cardiovascular health. Through comprehensive data analysis and visualization, we've identified correlations, trends, and predictive relationships that underscore the adverse effects of smoking on cholesterol levels and overall health. These findings emphasize the importance of smoking cessation programs and preventive healthcare interventions to mitigate the risks associated with smoking-related cardiovascular diseases. Our project contributes to the broader goal of promoting public health awareness and fostering evidence-based strategies for disease prevention and health promotion.