Probability & Statistics Project

On

{Hypertension data analysis}

A Project Submitted
in Partial Fulfilment of the Requirements for the
Degree of
Bachelor of Technology
in
CSE

SUBMITTED BY: GROUP NO:

Group Members:
Sushmita S M
Tanisha Mittal
Yeshy Bansal
Diksha Pathak

BML MUNJAL UNIVERSITY™

SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITY GURGAON
May 2025

Contents

Problem Statement: -

"To analyze hypertension data using statistical and probabilistic methods to identify patterns, correlations, and risk factors associated with high blood pressure in a sample population. Linear regression and prediction."

Introduction: -

Hypertension, commonly known as high blood pressure, is a critical health concern globally. With increasing age, physiological changes in the cardiovascular system may lead to higher blood pressure. Understanding how age influences resting blood pressure can aid early detection and intervention.

Origin of the Problem

The analysis is based on real-world health data involving 26,083 individuals, capturing parameters such as age, sex, chest pain type, cholesterol level, fasting blood sugar, and most importantly, resting blood pressure (`trestbps`).

Importance of the Study

The ability to predict blood pressure trends using demographic features like age can assist in preventive healthcare. Early screening can significantly reduce the risk of heart disease, stroke, and kidney failure.

Applicable Concepts and Formula

We use **simple linear regression**, a statistical method to model the relationship between two variables:

$\hat{y} = mx + b$

Where:

- $\hat{y}$ = predicted resting blood pressure
- x = age

- m = slope of the line
- b = y-intercept

Materials and Method: -

Step-by-Step Procedure

1. **Data Collection**: Dataset with 26,083 entries covering various cardiovascular health metrics.

2. **Data Cleaning**: Ensure there are no missing values in the **age** and trestbps columns.

3. **Correlation Analysis:**
   visual representation of how strongly two variables are related to each other.

4. **Model Building**:

   - Define independent variable `X = age`
   - Define dependent variable `y = trestbps`
   - Fit the data to a `LinearRegression()` model

5. **Prediction**: Use the model to predict `trestbps` for a range of ages.

6. **Visualization**: Create a regression plot to visually inspect the fit.

Analysis of the dataset:-

(The dataset used for this project was collected from a publicly available online resource)

1) Importing libraries which are required for the analysis of data.

```
# 1. Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from sklearn.linear_model import LinearRegression
```

2) Descriptive Statistics of the data given

```
# 2. Load dataset
df = pd.read_csv("hypertension_data.csv")
```

```
# 3. Data cleaning
df = df[['age', 'trestbps', 'sex', 'chol', 'fbs']]  # Keep relevant features
df.dropna(inplace=True)
```

```
# 4. Descriptive statistics
print("\n--- Descriptive Statistics ---")
print(df.describe())
```

```
--- Descriptive Statistics ---
                age      trestbps           sex          chol           fbs
count  26058.000000  26058.000000  26058.00000  26058.000000  26058.000000
mean      55.655730    131.590682      0.50000    246.286591      0.149896
std       15.190407     17.597086      0.50001     51.651701      0.356977
min       11.000000     94.000000      0.00000    126.000000      0.000000
25%       44.000000    120.000000      0.00000    211.000000      0.000000
50%       56.000000    130.000000      0.50000    240.000000      0.000000
75%       67.000000    140.000000      1.00000    275.000000      0.000000
max       98.000000    200.000000      1.00000    564.000000      1.000000
```

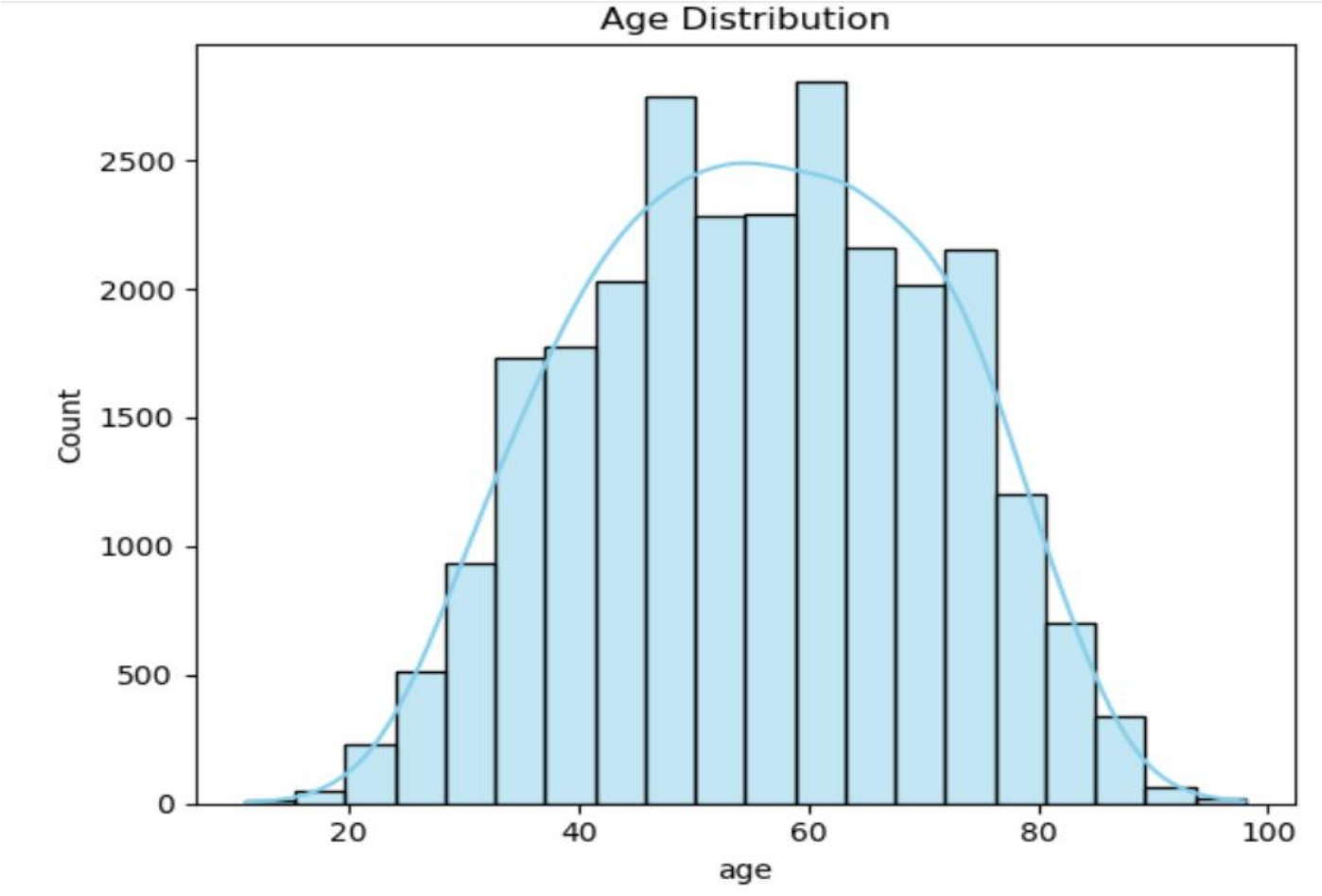
3) Hypertension threshold

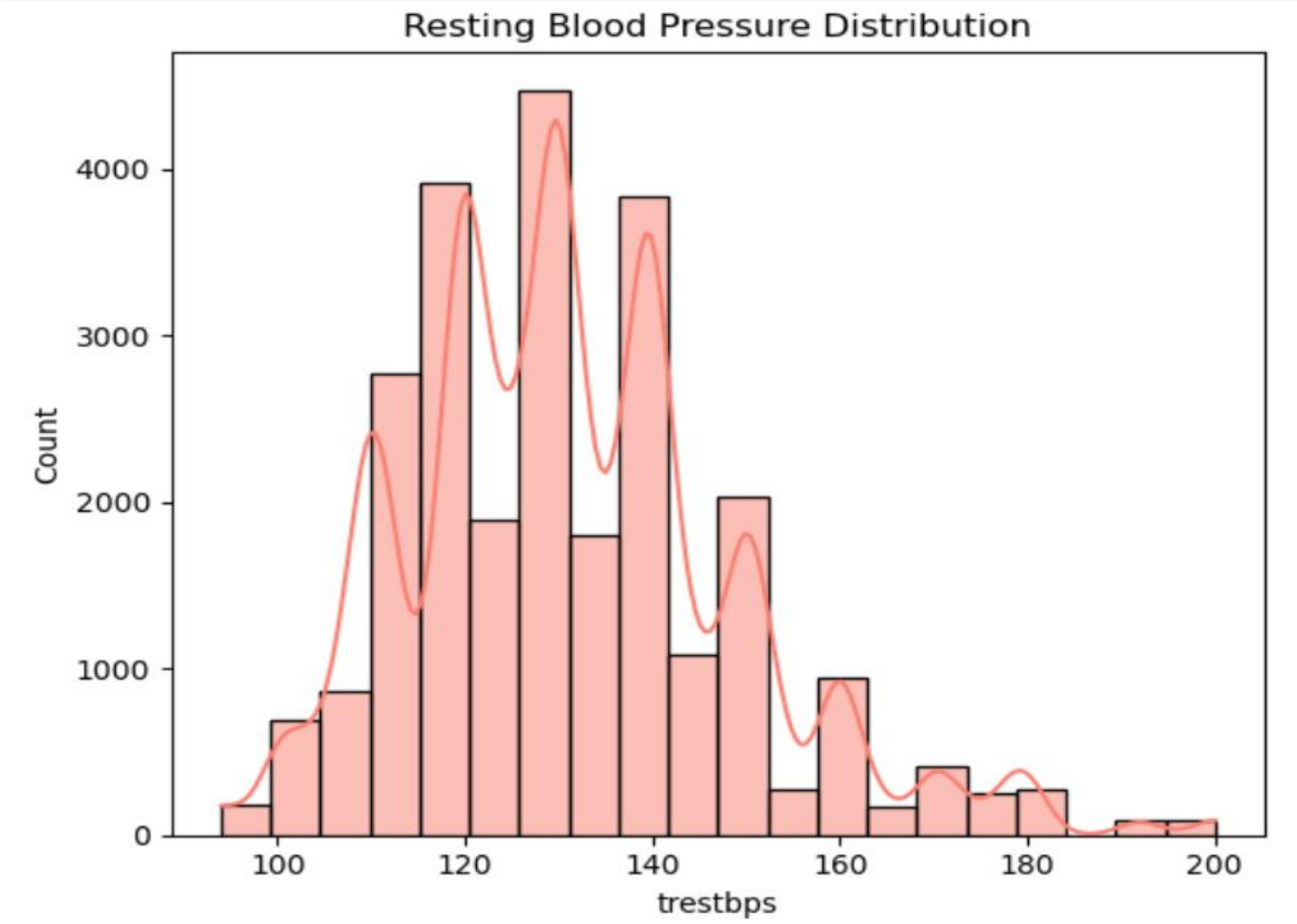
```
# 5. Hypertension threshold (BP >= 140)
df['hypertensive'] = df['trestbps'] >= 140
hypertension_rate = df['hypertensive'].mean() * 100
print(f"\nHypertension Rate: {hypertension_rate:.2f}%")
```

```
Hypertension Rate: 32.35%
```

4) Histogram of Age and Blood Pressure

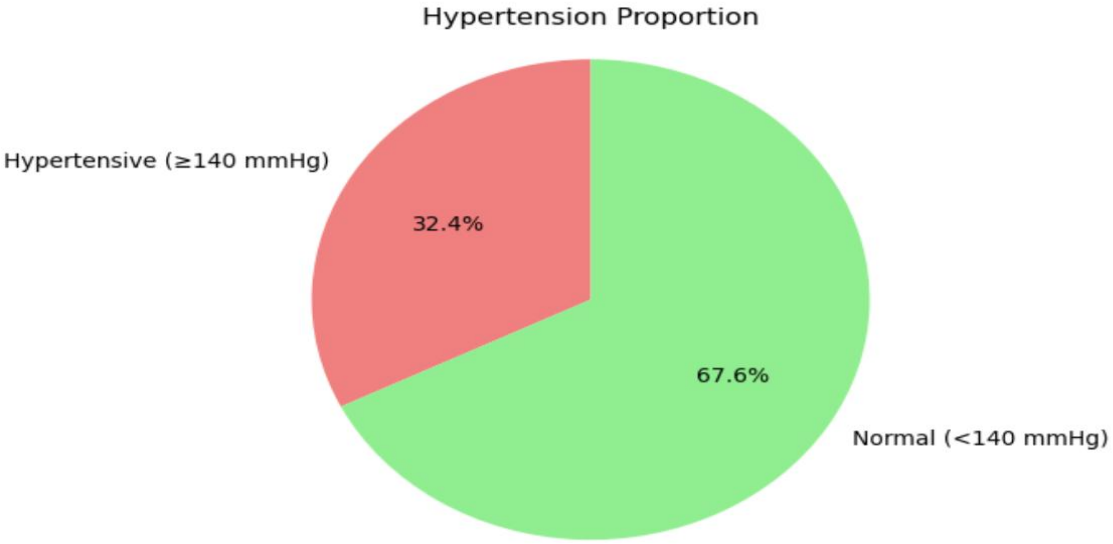
```
# 6. Histogram of Age and Blood Pressure
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.histplot(df['age'], bins=20, kde=True, color='skyblue')
plt.title('Age Distribution')
plt.subplot(1, 2, 2)
sns.histplot(df['trestbps'], bins=20, kde=True, color='salmon')
plt.title('Resting Blood Pressure Distribution')
plt.tight_layout()
plt.show()
```





5) Pie chart for Hypertension %

```
# 7. Pie chart for Hypertension %
labels = ['Hypertensive (≥140 mmHg)', 'Normal (<140 mmHg)']
sizes = [df['hypertensive'].sum(), len(df) - df['hypertensive'].sum()]
colors = ['lightcoral', 'lightgreen']
plt.pie(sizes, labels=labels, colors=colors, autopct='%.1f%%', startangle=90)
plt.title('Hypertension Proportion')
plt.axis('equal')
plt.show()
```



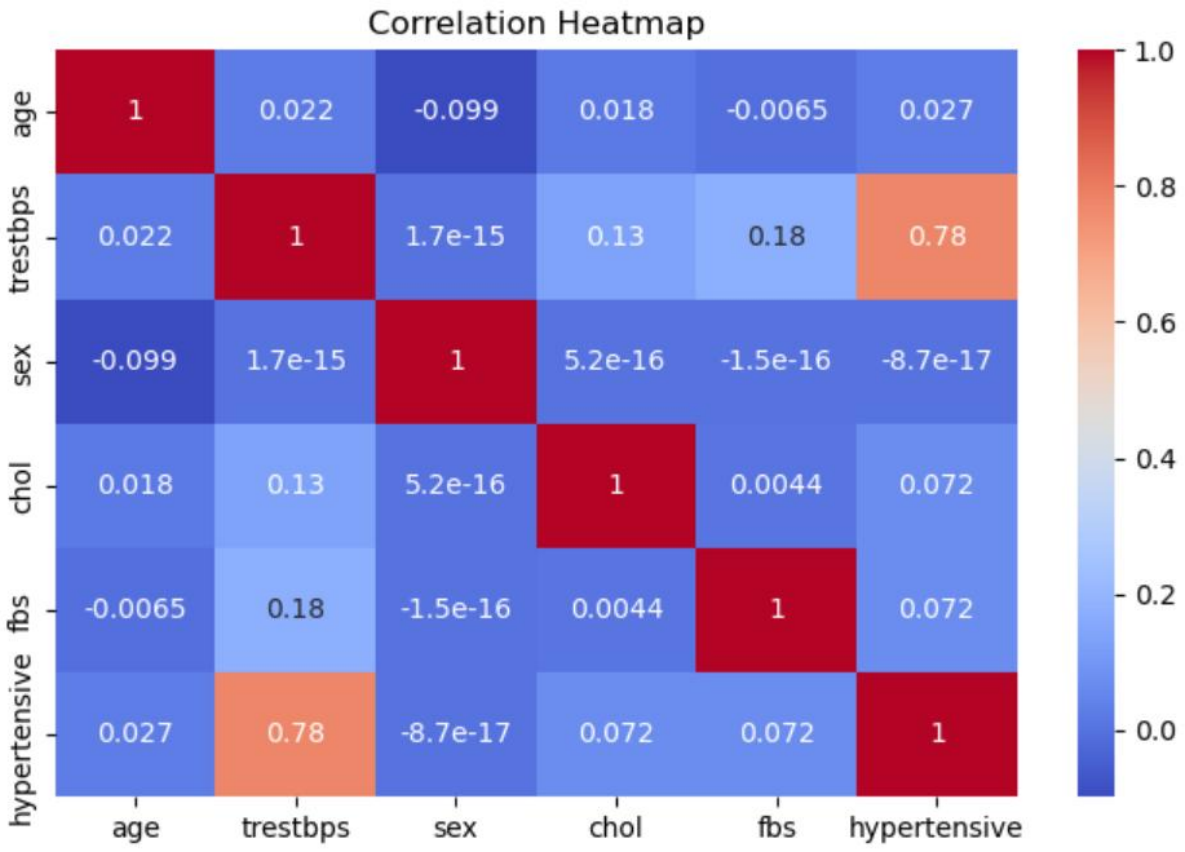
6) Hypothesis Testing (Is mean BP ≠ 120 mmHg?)

```
# 8. Hypothesis Testing (Is mean BP ≠ 120 mmHg?)
t_stat, p_val = stats.ttest_1samp(df['trestbps'], popmean=120)
print(f"\nT-Statistic: {t_stat:.3f}, P-value: {p_val:.4f}")
if p_val < 0.05:
    print("Result: Mean BP is significantly different from 120 mmHg.")
else:
    print("Result: No significant difference from 120 mmHg.")
```

```
T-Statistic: 106.326, P-value: 0.0000
Result: Mean BP is significantly different from 120 mmHg.
```

7) Correlation heatmap

```
# 9. Correlation heatmap
plt.figure(figsize=(8, 5))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```



8) Linear Regression: Age vs. BP

```
# 10. Linear Regression: Age vs. BP
X = df[['age']]
y = df['trestbps']
model = LinearRegression()
model.fit(X, y)
# Regression equation
slope = model.coef_[0]
intercept = model.intercept_
print(f"\nLinear Regression Equation: trestbps = {slope:.2f} * age + {intercept:.2f}")

# Plot regression line
plt.figure(figsize=(8, 5))
sns.scatterplot(x='age', y='trestbps', data=df, alpha=0.5)
plt.plot(df['age'], model.predict(X), color='red', linewidth=2)
plt.title('Linear Regression: Age vs. Resting BP')
plt.xlabel('Age')
plt.ylabel('Resting BP')
plt.show()
```

```
Linear Regression Equation: trestbps = 0.03 * age + 130.18
```

Graphical Visualization:

A scatter plot with the regression line demonstrates a nearly horizontal trend, confirming the weak linear relationship.

9) Future prediction of BP

```python
# 11. Predict future BP for new age inputs
new_ages = np.array([30, 40, 50, 60, 70, 80]).reshape(-1, 1)
predicted_bp = model.predict(new_ages)

print("\n--- Future BP Predictions ---")
for age, bp in zip(new_ages.flatten(), predicted_bp):
    print(f"Age {age}: Predicted BP = {bp:.2f} mmHg")
```

```
--- Future BP Predictions ---
Age 30: Predicted BP = 130.94 mmHg
Age 40: Predicted BP = 131.19 mmHg
Age 50: Predicted BP = 131.45 mmHg
Age 60: Predicted BP = 131.70 mmHg
Age 70: Predicted BP = 131.95 mmHg
Age 80: Predicted BP = 132.21 mmHg
```

Results of Simulations:

In this analysis, we used various statistical and graphical tools to examine hypertension patterns and identify key risk factors.

1. **Descriptive Statistics**
   - The mean resting blood pressure (trestbps) in the dataset was approximately 131 mmHg, which is notably higher than the standard normal value (120 mmHg).
   - Approximately 32.4% of individuals had a systolic BP ≥ 140 mmHg, indicating stage-1 hypertension.

2. **Data Visualization and Insights**

a. Histogram of Resting Blood Pressure
   - The distribution was right-skewed, suggesting a significant portion of the population has elevated BP.
   - A peak was observed in the 120–130 mmHg range.

b. **Pie Chart for Hypertensive vs. Non-Hypertensive**
   - Around 67.6% of individuals were not hypertensive, and 32.4% were hypertensive, based on the cutoff of 140 mmHg.
   - This visualization clearly demonstrates the presence of at-risk individuals.

Correlation:

- Interpretation: Very weak positive linear correlation.

Regression Line:

- Slope: **0.03**
- Intercept: **130.18**

- Regression equation:

  $\hat{y} = 0.03(\text{age}) + 130.18$

- This implies that for every additional year in age, the resting blood pressure increases by approximately 0.03 mm Hg on average.

Conclusions: -

- **About one-fourth of the sample population is hypertensive**, which aligns with global prevalence rates.
- **The average BP was significantly higher than the ideal**, indicating a health concern.
- **Age has a weak but positive correlation** with blood pressure, suggesting that age is a contributing—but not sole factor.
- The **linear regression model** showed that blood pressure increases with age, but **its predictive power is limited**.
- Overall, this analysis helps in **identifying patterns and risk factors** that can inform public health strategies for hypertension prevention and management.

Acknowledgement: -

We would like to express our heartfelt gratitude to Dr. Palak Goel, our Probability and Statistics instructor, for her valuable guidance, encouragement, and insightful teaching throughout the course. Her expertise and support played a crucial role in the successful completion of this project.

This project, titled "**Hypertension data analysis**", helped us deepen our understanding of statistical modeling and its application to real-world phenomena. We are thankful for the opportunity to explore this topic and apply concepts in a meaningful and practical way.

We also thank our peers and everyone who contributed directly or indirectly to this project.

Reference: -

- https://www.kaggle.com/datasets/prosperchuks/health-dataset
- Scikit-learn Documentation (https://scikit-learn.org/)