جامعة المستقبل
AL MUSTAQBAL UNIVERSITY

كلية العلوم

قسم الانظمة الطبية الذكية

Lecture: (3)

Understanding Data with Statistics & Visualization

Subject: Machine Learning
Class: Third
Lecturer: Dr. Maytham N. Meqdad

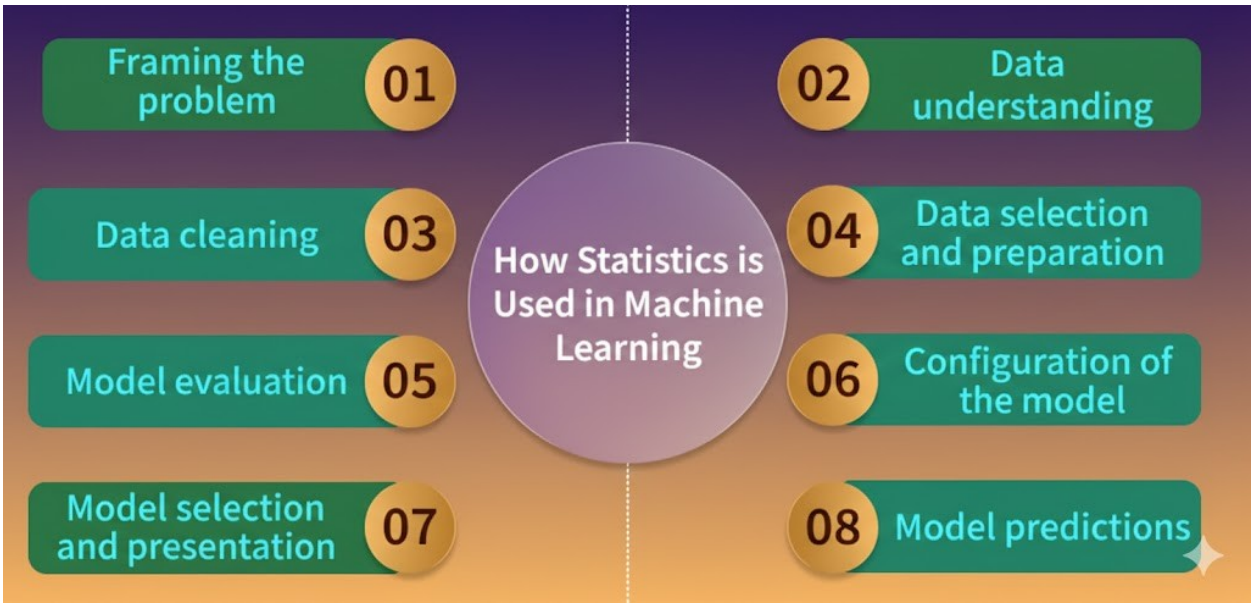
Understanding Data with Statistics & Visualization for Machine Learning

-Why Statistics Matter in Machine Learning?

Machine learning models are powerful pattern-recognition engines, but their effectiveness depends entirely on the quality and characteristics of the data they consume. Statistical summaries provide the foundational understanding needed before any modelling begins, revealing critical insights about data structure and distribution.

Metrics such as mean, variance, standard deviation, and correlations paint a comprehensive picture of your dataset's nature. These measures expose hidden characteristics that directly influence model performance and reliability.

Early detection of issues like skewed distributions, extreme outliers, or imbalanced classes prevents the creation of biased or inaccurate models. A robust statistical foundation ensures your machine learning pipeline starts on solid ground, saving time and resources whilst improving outcomes.



Applications of Statistics in Machine Learning

Statistics is a key component of machine learning, with broad applicability in various fields.

- **Feature Engineering:** selecting and transforming useful variables.
- **Image Processing:** analyzing patterns, shapes and textures.
- **Anomaly Detection:** spotting fraud or equipment failures.
- **Environmental Studies:** modeling land cover, climate and pollution.
- **Quality Control:** identifying defects in manufacturing.

Types of Statistics

There are commonly two types of statistics, which are discussed below:

- **Descriptive Statistics:** "Descriptive Statistics" helps us simplify and organize big chunks of data. This makes large amounts of data easier to understand.
- **Inferential Statistics:** "Inferential Statistics" is a little different. It uses smaller data to draw conclusions about a larger group. It helps us predict and draw conclusions about a population.

Descriptive Statistics

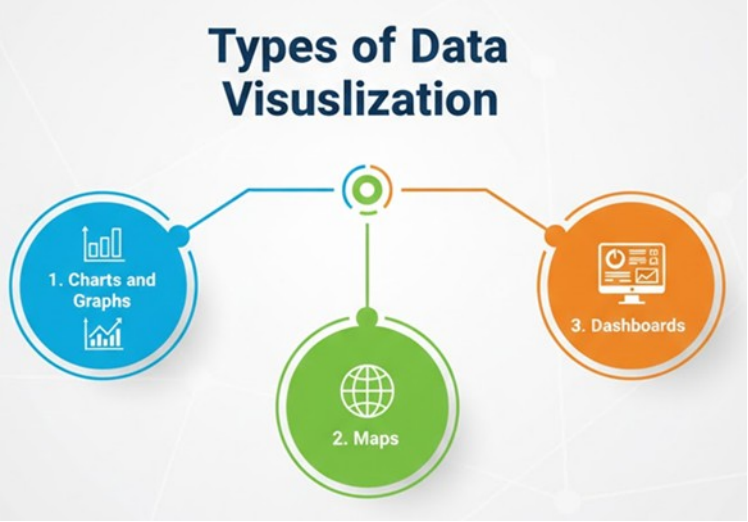
Descriptive statistics summarize and describe the features of a dataset, providing a foundation for further statistical analysis.

What is Data Visualization and Why is It Important?

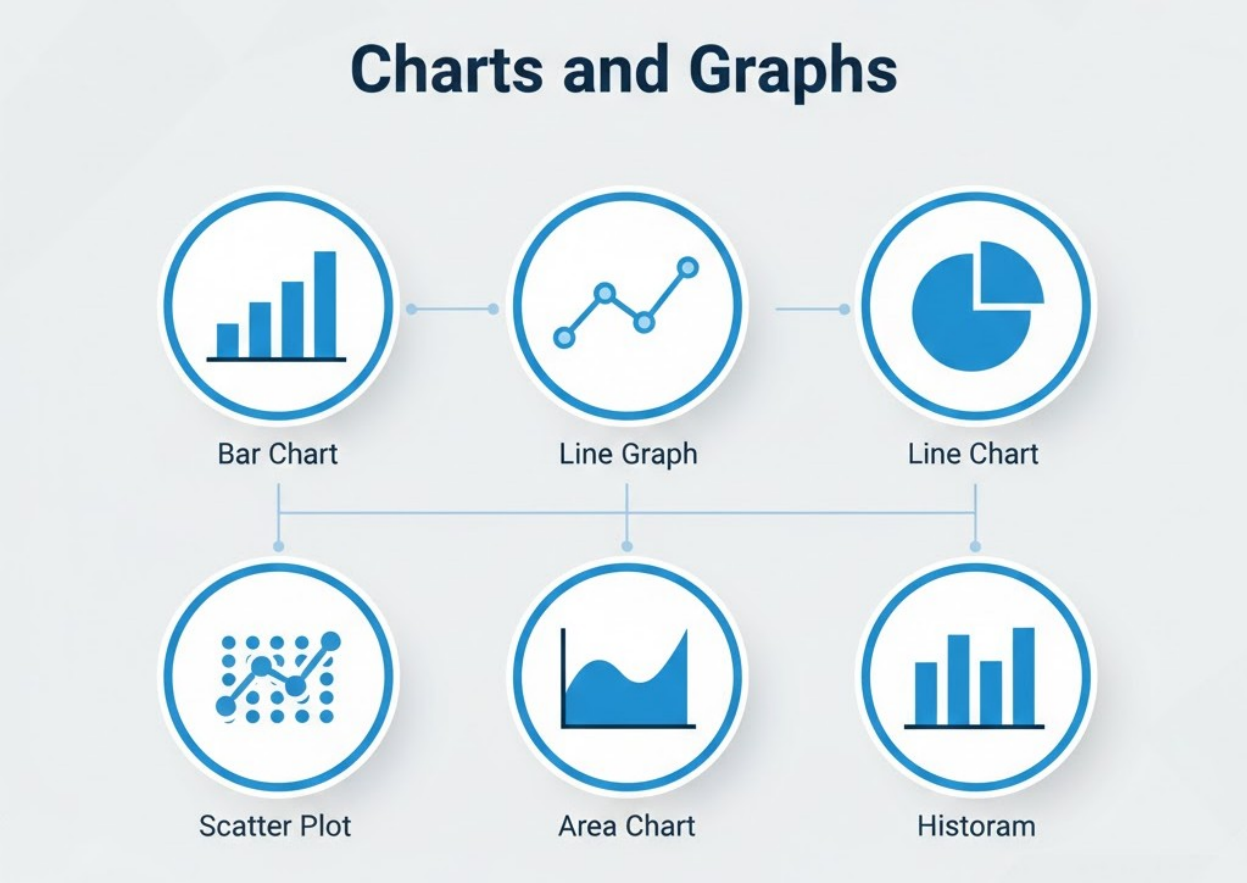
Data visualization uses charts, graphs and maps to present information clearly and simply. It turns complex data into visuals that are easy to understand. With large amounts of data in every industry, visualization helps spot patterns and trends quickly, leading to faster and smarter decisions.

Types of Data Visualization

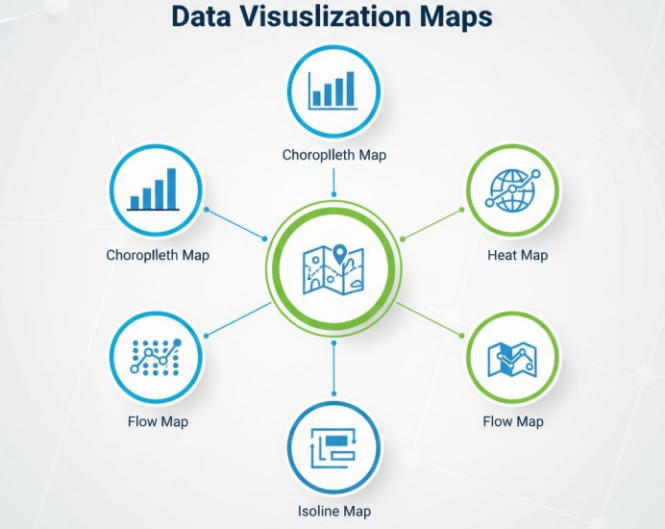
There are various types of visualizations where each has a unique purpose in data representation. Here are the most common types:



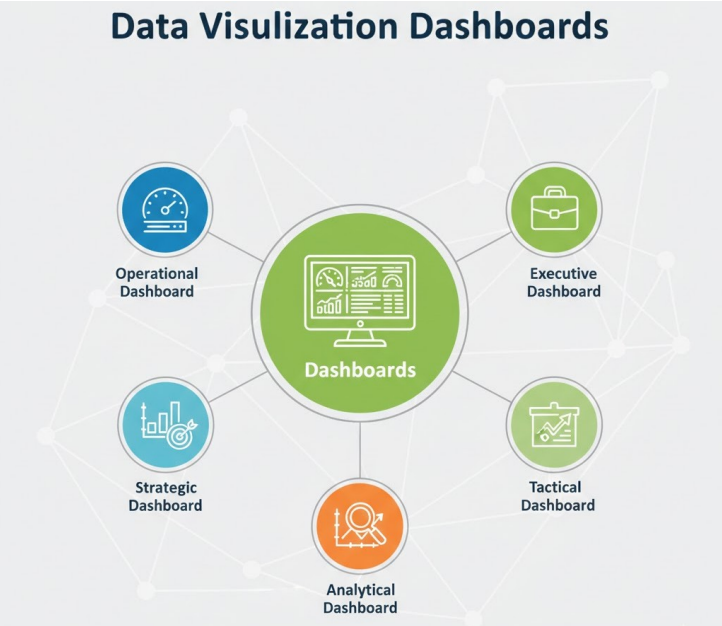
1. **Charts and Graphs:** They are used to visualize data, with charts comparing data points across categories or showing trends over time and graphs analyzing relationships between variables to identify correlations, trends and outliers. Examples: Bar Charts, Line Charts, Pie Charts, Scatter Plots, Histograms, Box Plots.



2. **Maps**: They are used to display geographical data which provides spatial context to trends and patterns. Examples: Geographic Maps, Heat Maps



3. **Dashboards:** They combine multiple visualizations into a single interface which provides real-time insights and interactive features for users to explore data.



Importance of Data Visualization

Data visualization is essential for understanding and communicating information effectively. Here are some key reasons why it's important:

1. **Simplifies Complex Data:** It turns large and complicated data into visual formats like charts and graphs, making the information easier to understand.

2. **Reveals Patterns and Trends**: It helps identify trends, relationships and patterns that are not easily seen in raw data or tables.

3. **Saves Time:** Visuals allow quicker interpretation of data, helping users spot key information at a glance instead of manually scanning through numbers.

4. **Improves Communication:** It makes it easier to explain data insights to others, especially those who may not be familiar with the technical details.

5. **Tells a Clear Story**: Data visuals guide the audience through the information step-by-step, making it easier to reach conclusions and make informed decisions.

Real-World Use Cases for Data Visualization

Data visualization is used across various industries to improve decision-making and drive results. Here are a few examples:

1. **Business Analytics:** Used to monitor company performance, track KPIs and make data-driven decisions by visualizing trends, sales and customer metrics.
2. **Healthcare:** Helps in analyzing patient records, tracking disease outbreaks and managing hospital operations through easy-to-read charts and dashboards.
3. **Sports:** Used to visualize player statistics, team performance and match outcomes, helping coaches and analysts improve strategies and training plans.
4. **Retail and E-commerce:** Enables tracking of sales, customer preferences and inventory levels, helping businesses adjust stock and marketing efforts effectively.

Challenges in Data Visualization

1. **Data Quality**: Accuracy of visualizations depends on the quality of the data. If the data is inaccurate or incomplete, the insights from the visualization will be misleading.
2. **Over-Simplification**: Simplifying data too much can lead to important details being lost like using a pie chart that oversimplifies complex relationships between categories.
3. **Choosing the Right Visualization**: Using the wrong type of visualization can distort the message. For example, a pie chart might not work well with many categories which leads to confusion.
4. **Overload of Information**: Too much information in a visualization can overwhelm viewers. It's important to focus on key data points and avoid clutter.

Best Practices for Effective Data Visualization

To ensure our visualizations are impactful and easy to understand we follow these best practices:

1. **Audience-Centric Design**: Tailor visualizations to our audience's knowledge. A technical audience may need detailed graphs while a general audience benefits from simpler charts.

2. **Design Clarity and Consistency**: Choose the right chart for our data and keep the design clean with consistent colors, fonts and labels and also avoid clutter to ensure clarity.

3. **Provide Context**: Always provide context by including labels, titles and data source acknowledgments. This helps viewers to understand the significance of the data and builds trust in the results.

4. **Interactive and Accessible Design**: Make visualizations interactive with features like tooltips and filters and ensure accessibility for all users regardless of device or visual needs.

Practical Example Using Python

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

# Load dataset
df = pd.read_csv("medical_data.csv")

# Basic statistics
print(df.describe())

# Histogram
sns.histplot(df["Age"])
plt.show()

# Boxplot
sns.boxplot(x=df["BloodPressure"])
plt.show()

# Correlation Heatmap
sns.heatmap(df.corr(), annot=True)
plt.show()
```

	Age	BloodPressure	Cholesterol	Glucose
count	200.000000	200.000000	200.000000	200.000000
mean	45.320000	120.450000	210.300000	98.750000
std	12.150000	15.670000	30.120000	14.200000
min	18.000000	85.000000	150.000000	70.000000
25%	35.000000	110.000000	190.000000	88.000000
50%	45.000000	120.000000	205.000000	97.000000
75%	55.000000	130.000000	230.000000	108.000000
max	75.000000	170.000000	300.000000	140.000000

count → Number of values

mean → Mean

std → Standard deviation

min → Lowest value

25% → First quartile

50% → Median

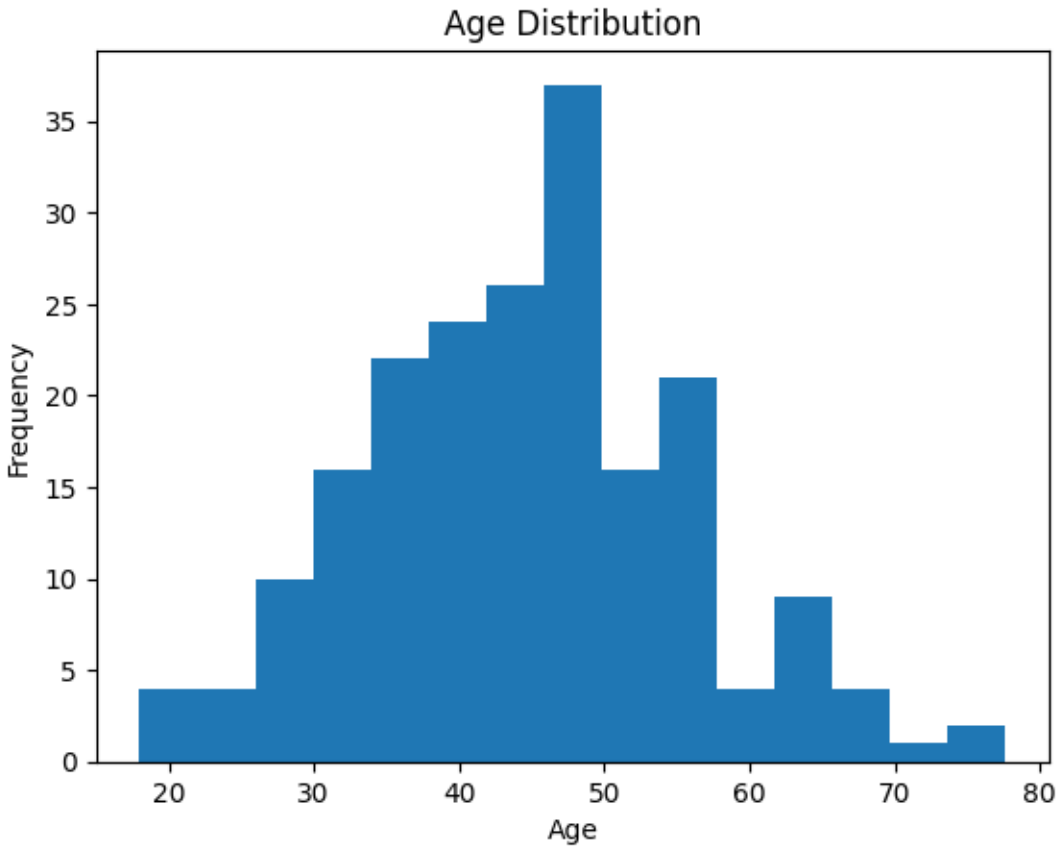
75% → Third quartile

max → Highest value

Correlation Matrix

	Age	BloodPressure	Cholesterol	Glucose
Age	1.00	0.62	0.30	0.25
BloodPressure	0.62	1.00	0.40	0.35
Cholesterol	0.30	0.40	1.00	0.50
Glucose	0.25	0.35	0.50	1.00

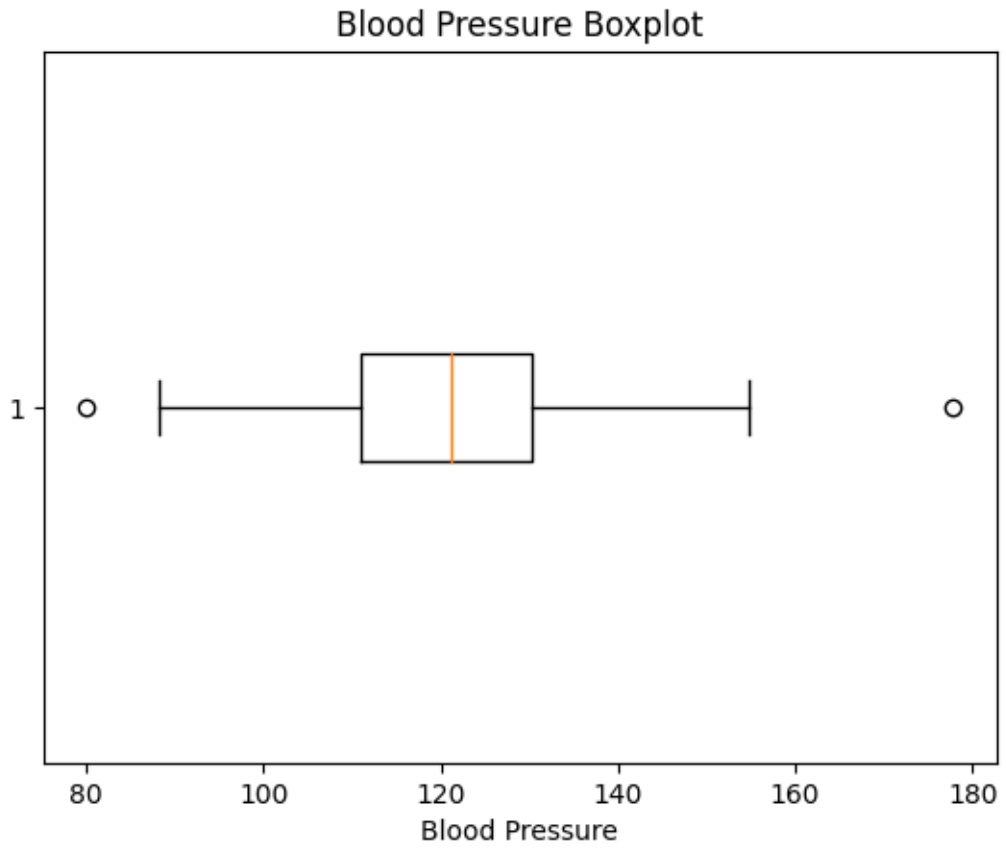
1. Age Distribution (Histogram)



The distribution appears approximately normal (bell-shaped).

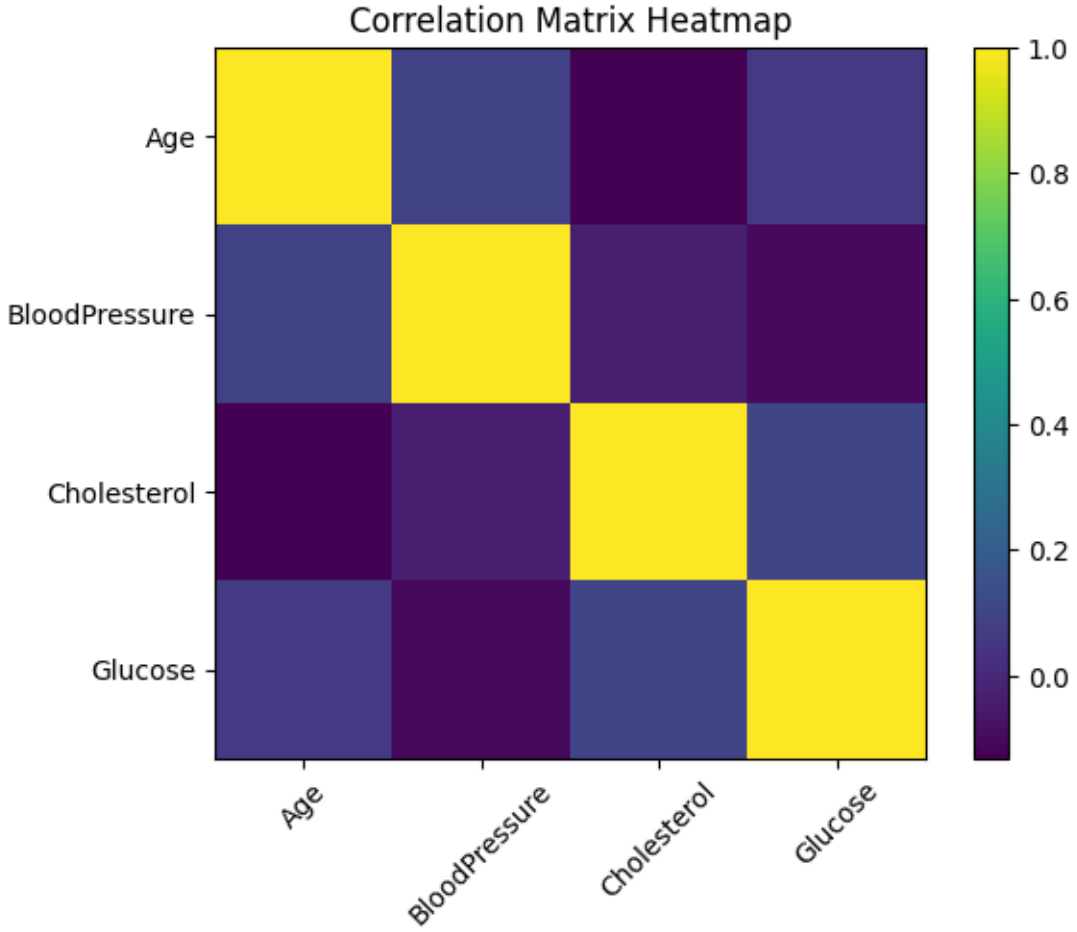
- Most ages are concentrated between 40 and 50 years.
- This helps us determine whether data transformation is needed before applying machine learning algorithms.

2. Blood Pressure Boxplot



- The line inside the box represents the **median**.
- The box represents the **Interquartile Range (IQR)**.
- Points outside the whiskers represent **outliers**.
- Discussion point for students:
  Should we remove outliers, or should we retain them because they may represent medically significant cases?

3. Correlation Matrix Heatmap



- Values close to **+1 or −1** indicate a strong correlation.
- It is used to detect **multicollinearity** between features.
- Discussion question for students:
  If Age is strongly correlated with Blood Pressure, should we keep both variables in the model? Why or why not?

References

[1] Machine Learning Bookcamp, Alexey Grigorev.
[2] Python Data Science Handbook, Jake VanderPlas
[3] https://github.com/microsoft/Data-Science-For-Beginners
[4] geeks for geeks: https://www.geeksforgeeks.org/