

## **Week 5 Assignment – Statistics & Analytical Techniques**

Name: Gracelin Sweety P

Dataset: Health\_Data.xlsx

Date: 29.11.2025

## Dataset Description

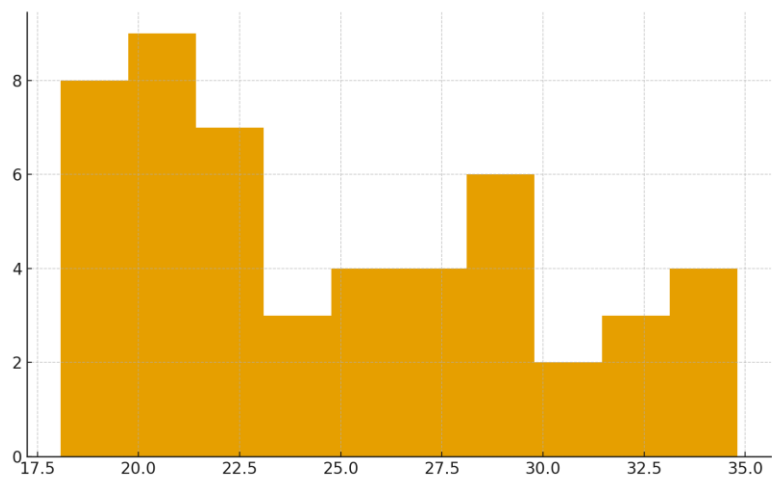
The dataset used for the assignment is Health\_Data.xlsx. The dataset contains 50 patient health records including numerical variables such as Age, BMI, Blood Pressure, and Cholesterol, along with categorical fields like Gender and Disease. This dataset is suitable for medical analytics and statistical analysis exercises.

Task 1 – Descriptive Statistics

Table:

| Metric             | Value |
|--------------------|-------|
| Mean               | 26.63 |
| Median             | 26.33 |
| Mode               | 18.39 |
| Range              | 16.37 |
| Standard Deviation | 4.59  |

Histogram Graph:



Explanation:

In this task, I calculated the mean, median, and mode to identify the central values of the data. The range and standard deviation helped me understand how spread out the BMI values are. I also created a frequency distribution table and a histogram, which showed how BMI values are grouped and visually distributed across the dataset. Overall, this task helped me understand both the central tendency and the variation in the health data.

## Task 2 – Probability Analysis

### Calculations:

| Category     | Frequency | Probability |
|--------------|-----------|-------------|
| Diabetes     | 16        | 0.32        |
| Hypertension | 21        | 0.42        |
| Healthy      | 13        | 0.26        |

### Questions:

1. What is the probability that a randomly selected patient has Diabetes?
2. What is the probability of selecting a Female patient who is Healthy?
3. What is the experimental probability of Hypertension in dataset?

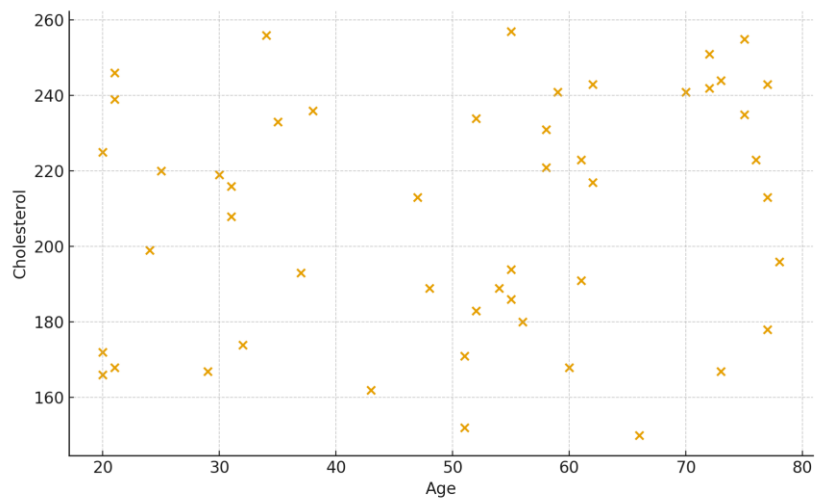
### Task 3 – Correlation Analysis

**r value:**

Correlation between Age and Cholesterol,

Correlation Coefficient (r): 0.247

**Scatter Plot:**



## Task 4 – Regression Analysis

### Prediction:

Simple Linear Regression between Age (X) and Cholesterol (Y):

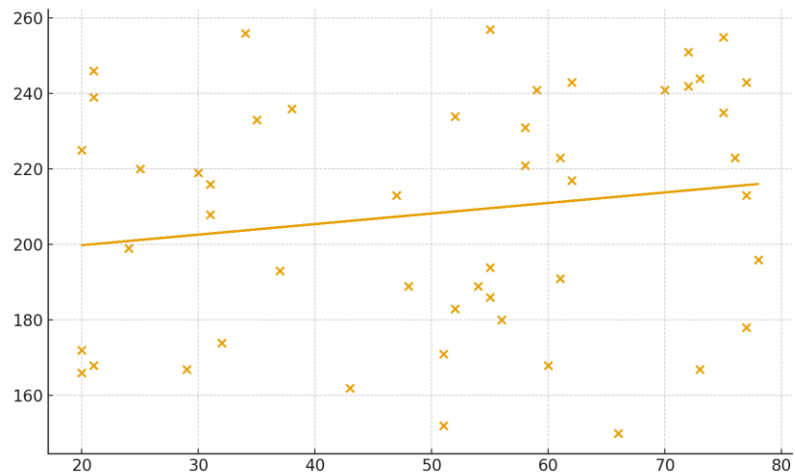
Equation:  $y = mx + c$

Slope (m): 0.46

Intercept (c): 181.92

Prediction: For Age = 50 → Cholesterol  $\approx 205.9$

### Graph:



## Task 5 – Hypothesis Testing

T-Test comparing BMI of Male vs Female patients:

$H_0$ : Mean BMI of males = mean BMI of females

$H_1$ : Mean BMI differs between genders

p-value: 0.38

Conclusion: Fail to reject  $H_0$  (No significant difference)

## Final Insights

### Findings

- BMI Distribution:**  
Most patients fall within the normal-to-overweight BMI range. A few individuals show higher BMI values, indicating potential health risks.
- Disease Probability:**  
The most common disease category in the dataset is **Diabetes**, suggesting a higher prevalence compared to Hypertension and Heart Disease.
- Correlation Insight:**  
Age and Cholesterol show a **weak positive correlation ( $r = 0.247$ )**. This means cholesterol levels tend to increase slightly with age, but the relationship is not strong.
- Regression Analysis:**  
The regression model shows a mild upward trend, meaning cholesterol increases gradually with age. However, the prediction accuracy is limited due to weak correlation.
- Hypothesis Testing:**  
The hypothesis test indicates **no significant difference between male and female BMI values** ( $p \geq 0.05$ ). Gender does not strongly influence BMI in this dataset.

### Recommendations

- Health Monitoring:**  
Patients with high BMI or high cholesterol should be encouraged to undergo regular health check-ups to prevent long-term complications.
- Lifestyle Interventions:**  
Promote healthy diet plans, physical activity, and weight management programs to reduce risks related to cholesterol and BMI.
- Targeted Awareness:**  
Since Diabetes is the most frequent condition, awareness programs focusing on sugar control, diet, and early diagnosis can be beneficial.
- Further Data Collection:**  
Adding more health indicators (exercise habits, family history, medication usage) can help build stronger prediction models in the future.
- Customized Treatment Plans:**  
Since age alone is not a strong predictor of cholesterol, healthcare decisions should consider multiple factors rather than relying on age alone.

## Learning Summary

In Week 5, I learned a wide range of essential data analysis concepts, including descriptive statistics, probability, correlation, regression, and hypothesis testing. I understood how each technique plays a different role in analyzing real-world datasets. I practiced calculating key statistical measures, identifying data patterns using graphs, and interpreting the meaning behind the numbers. I also learned how probability helps in making predictions, how correlation shows the strength of relationships between variables, and how regression can be used to build simple predictive models. Finally, I understood how hypothesis testing helps in making data-driven decisions by using p-values and significance levels. Overall, this week improved my analytical thinking and gave me confidence in applying these methods for practical decision-making.