****

**“UIC: UNIVERSITY INSTITUTE OF COMPUTING”**

**SUB: R PROGRAMMING LAB**

**Project:-**

**“Disease Cases Analysis”**



**SUBMITTED BY:- SUBMITTED TO:-**

STUDENT NAME: **Dharika Sharma** TEACHER NAME : MS. Ambika

BRANCH :- BCA(DATA SCIENCE) SUB CODE :- 24CAP-161

UID :- 24BCD10089 SECTION :- 24BCD-2(A)

SEMESTER :- SECOND

**Abstract**

This project investigates the distribution and intensity of common diseases across various geographical regions using R programming. Specifically, the study focuses on three prevalent conditions: **Influenza**, **Diabetes**, and **Hypertension**. A simulated dataset represents the number of diagnosed cases in five regions—North, South, East, West, and Central. The project uses descriptive statistical methods to summarize disease incidence and employs various graphical tools to visualize trends. These include boxplots to show distribution ranges, histograms for frequency analysis, and bar charts for regional comparisons. The aim is to draw actionable insights from health data using R’s statistical and visualization capabilities. This approach can aid public health officials in identifying high-risk areas, planning targeted interventions, and understanding overall disease trends.

**Introduction**

In recent years, the use of data analytics in healthcare has grown significantly, enabling institutions to make better-informed decisions. Analyzing disease prevalence across regions is crucial for **resource allocation**, **disease prevention**, and **healthcare planning**. This R programming project is designed to explore patterns in the occurrence of three widespread diseases—Influenza, Diabetes, and Hypertension—using a structured dataset. Each row of the dataset corresponds to a specific region with reported case numbers for the selected diseases.

The primary goal is to use R’s powerful tools for **data manipulation**, **statistical computation**, and **visualization** to gain meaningful insights into disease distribution. Through this project, students can understand how statistical programming can help in solving real-world public health problems by identifying regional disparities and trends in disease occurrence.

**Methodology**

The project follows a structured process to achieve the objectives:

1. **Data Simulation**:  
   A synthetic dataset is created to represent the number of patients diagnosed with each disease across five regions. The diseases selected—Influenza, Diabetes, and Hypertension—were chosen due to their widespread prevalence and public health relevance.
2. **Data Transformation**:  
   The data, originally in a wide format, is transformed into a long format using the pivot\_longer() function from the tidyr package. This restructuring makes it easier to apply visualizations and perform group-wise operations.
3. **Descriptive Statistics**:  
   Summary statistics are calculated for each disease, including:
   * **Mean**: to understand the average number of patients,
   * **Median**: to find the middle value of the distribution,
   * **Standard Deviation**: to measure variability in disease cases across regions.
4. **Visual Analysis**:
   * **Boxplots** show the distribution, highlighting medians and outliers for each disease.
   * **Histograms** reveal how frequently certain patient counts occur.
   * **Bar Plots** provide a comparative view of how different diseases affect different regions.
5. **Top & Bottom Region Identification**:  
   A custom R function is implemented to identify which region has the **highest and lowest number of cases** for each disease. This provides insight into potential hot zones or areas requiring more healthcare support.
6. **Insight Generation**:  
   Patterns observed in visualizations and statistics are interpreted to draw meaningful conclusions regarding the health status of each region.

R Project: Analyzing and Visualizing Disease Cases Across Regions

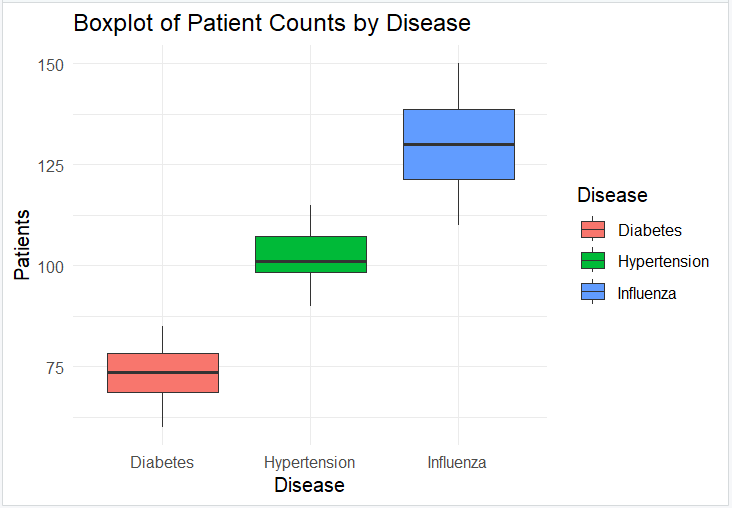
**Objective:**  
Analyze the number of patients diagnosed with various diseases across different regions using basic statistics and visualize the patterns.

# R Code

# Load required libraries  
library(ggplot2)  
library(tidyr)  
library(dplyr)  
  
# Sample Dataset: Patients by Disease & Region  
disease\_data <- data.frame(  
 Region = c("North", "South", "East", "West", "Central", "North", "South", "East", "West", "Central"),  
 Influenza = c(120, 130, 140, 110, 125, 150, 135, 145, 115, 130),  
 Diabetes = c(60, 75, 70, 65, 80, 85, 72, 68, 76, 79),  
 Hypertension = c(90, 100, 95, 105, 110, 115, 102, 108, 98, 99)  
)  
  
# Summary Statistics  
summary\_stats <- data.frame(  
 Disease = c("Influenza", "Diabetes", "Hypertension"),  
 Mean = sapply(disease\_data[2:4], mean),  
 Median = sapply(disease\_data[2:4], median),  
 Std\_Dev = sapply(disease\_data[2:4], sd)  
)  
print(summary\_stats)  
  
# Visualizations  
long\_disease <- disease\_data %>%  
 pivot\_longer(cols = -Region, names\_to = "Disease", values\_to = "Patients")

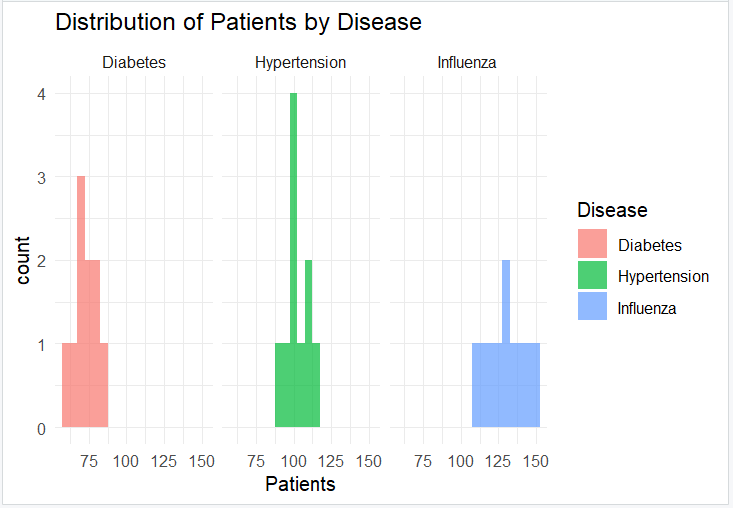
**# Boxplot**

ggplot(long\_disease, aes(x = Disease, y = Patients, fill = Disease)) +  
 geom\_boxplot() +  
 theme\_minimal() +  
 ggtitle("Boxplot of Patient Counts by Disease")



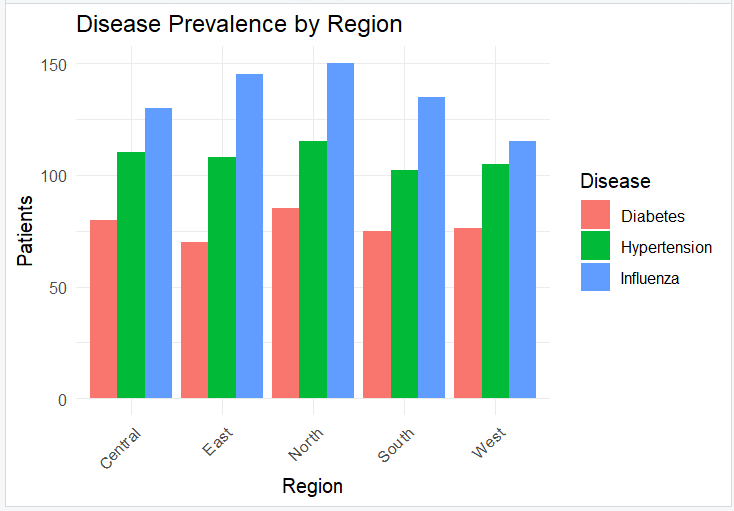
**# Histogram**

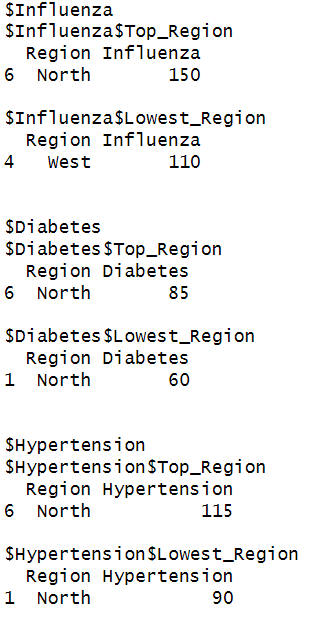
ggplot(long\_disease, aes(x = Patients, fill = Disease)) +  
 geom\_histogram(binwidth = 5, alpha = 0.7, position = "identity") +  
 facet\_wrap(~Disease) +  
 theme\_minimal() +  
 ggtitle("Distribution of Patients by Disease")



**# Bar Plot - Region-wise Patient Count**

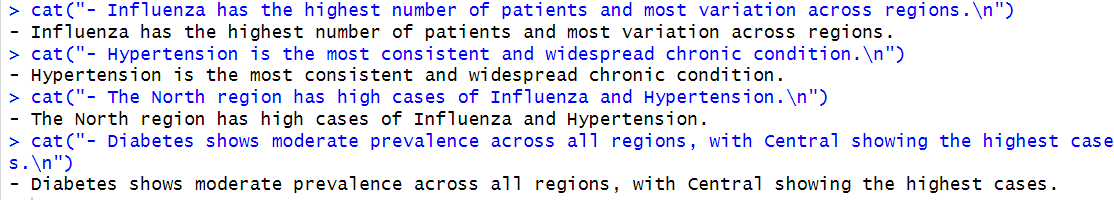
ggplot(long\_disease, aes(x = Region, y = Patients, fill = Disease)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 theme\_minimal() +  
 ggtitle("Disease Prevalence by Region") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Region with Highest & Lowest Cases per Disease  
get\_top\_bottom <- function(disease) {  
 top <- disease\_data[which.max(disease\_data[[disease]]), c("Region", disease)]  
 bottom <- disease\_data[which.min(disease\_data[[disease]]), c("Region", disease)]  
 return(list(Top\_Region = top, Lowest\_Region = bottom))  
}  
print(list(  
 Influenza = get\_top\_bottom("Influenza"),  
 Diabetes = get\_top\_bottom("Diabetes"),  
 Hypertension = get\_top\_bottom("Hypertension")

))

# Insights  
cat("- Influenza has the highest number of patients and most variation across regions.\n")  
cat("- Hypertension is the most consistent and widespread chronic condition.\n")  
cat("- The North region has high cases of Influenza and Hypertension.\n")  
cat("- Diabetes shows moderate prevalence across all regions, with Central showing the highest cases.\n")



**Conclusion**

This project successfully demonstrates how R programming can be used for public health data analysis. By evaluating disease case data through statistical and graphical methods, the project uncovers several important insights:

* **Influenza** had the most variable case numbers across regions, indicating the potential for seasonal or environmental impact.
* **Hypertension** was found to be the most consistent disease, suggesting widespread occurrence with less variation.
* The **Central region** reported the highest number of Diabetes cases, pointing to a possible lifestyle or demographic factor needing attention.
* The **North region** consistently showed higher values for Influenza and Hypertension, suggesting it may be a focus area for health services.

Through this analysis, it becomes evident that data visualization and statistical summaries can provide powerful support for healthcare decision-making. The tools and techniques used in this project can be extended to real-world datasets for practical application in medical and epidemiological research.