# UNIVERSITY INSTITUTE OF COMPUTING

**PROJECT REPORT**

**ON**

**HEALTH CARE DATA ANALYSIS**

Program Name: BCA (Data Science)

Subject Name: R Programming

Subject Code: 24CAP-161

**Submitted by :**

**Jatin Changotra**

24BCD10135

24BCD-1(A)

**Submitted to :**

Ms. Ambika

Assistant Professor

# HEALTH CARE DATA ANALYSIS

* Using correlation analysis to determine the relationship between lifestyle factor and diseases.
* Apply regression model to predict life expectancy based on health data.

# ABSTRACT

**Objective :**

To analyze healthcare data and testing & determine how lifestyle factors (like diet, exercise smoking, etc.) correlate with diseases. Additionally, build a regression model to predict life expectancy based on various health indicators.

**Methodology:**

* **Data Collection:** Gather relevant data (e.g., blood sugar levels, smoking habits, BMI, cholesterol, life expectancy).
* **Statistical Testing:** Apply statistical tests like t-test

(t.test()) for comparing groups, correlation (cor()) for relationships, and linear regression (lm()) for predictive modeling.

* **Result Interpretation:** Analyze outputs (p-values, correlation coefficients, regression summaries) to determine significance and relationships.
* **Assumption Checking:** Ensure data meets assumptions (normality, linearity, etc.) for valid analysis (optional but important).
* **Visualization:** Use plots (histograms, scatter plots, regression lines) with ggplot2 to visually interpret data trends.

# DESCRIPTION

## 1️⃣ T-Test: Blood Sugar Levels (Diabetics vs Non-Diabetics)

This analysis checks if there's a significant difference in blood sugar levels between diabetics and non-diabetics.

* **diabetic <- c(150, 160, 170, 180, 190, 200, 210)**

Creates a vector of blood sugar levels for diabetics.

* **non\_diabetic <- c(90, 95, 100, 105, 110, 115, 120)**

Creates a vector of blood sugar levels for non-diabetics.

* **t\_test\_result <- t.test(diabetic, non\_diabetic, alternative = "greater")**

Performs a one-sided t-test to check if diabetics have significantly higher blood sugar levels than non-diabetics.

* **print(t\_test\_result)**

Displays the results of the t-test, including p-values and confidence intervals.

* **blood\_sugar <- c(diabetic, non\_diabetic)**

Combines both groups into one vector for plotting.

* **group <- c(rep("Diabetic", length(diabetic)), rep("Non-Diabetic", length(non\_diabetic)))**

Creates a group vector to label each observation as either "Diabetic" or "Non-Diabetic".

* **data <- data.frame(blood\_sugar, group)**

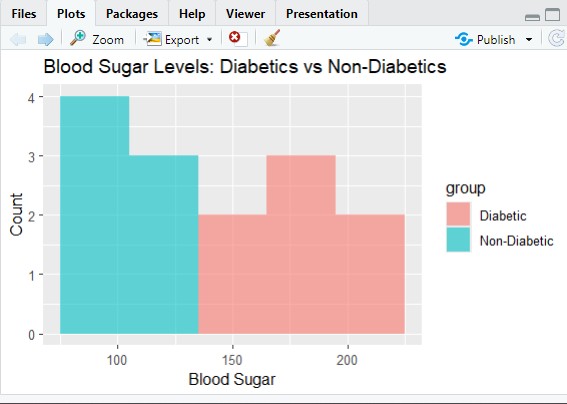
Combines blood\_sugar and group into a data frame for visualization.

* **ggplot(data, aes(x = blood\_sugar, fill = group)) + geom\_histogram(alpha = 0.6, position = "identity", bins = 5) + labs(title = "Blood Sugar Levels:**

**Diabetics vs Non-Diabetics", x = "Blood Sugar", y =**

**"Count")**

Creates a histogram comparing blood sugar levels of diabetics and non-diabetics with overlapping bars and labeled axes.



## 2️⃣ Correlation Analysis: Smoking & Heart Disease Risk

This section explores the relationship between smoking and heart disease risk.

* **smoking <- c(5, 10, 20, 15, 25, 30, 35)**

Creates a vector representing the number of cigarettes smoked per day.

* **heart\_disease\_risk <- c(10, 20, 40, 30, 50, 60, 70)**

Creates a vector representing corresponding heart disease risk scores.

* **correlation <- cor(smoking, heart\_disease\_risk)**

Calculates the Pearson correlation coefficient to measure the strength and direction of the linear relationship.

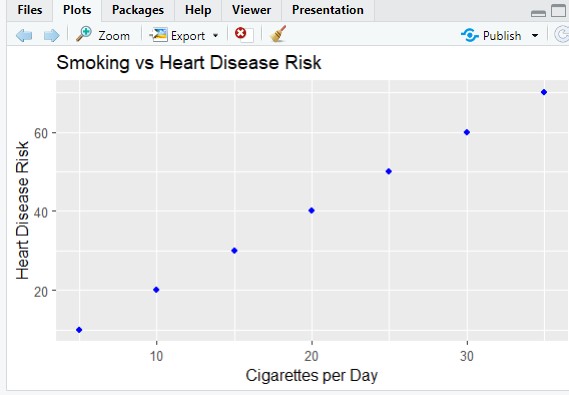
* **print(correlation)**

Displays the correlation coefficient.

* **ggplot(data.frame(smoking, heart\_disease\_risk), aes(x = smoking, y = heart\_disease\_risk)) + geom\_point(color = "blue") + labs(title = "Smoking vs Heart Disease Risk", x = "Cigarettes per Day", y**

**= "Heart Disease Risk")**

Creates a scatter plot to visually examine the relationship between smoking and heart disease risk with blue points and labeled axes.



## 3️⃣ Regression Model: BMI, Cholesterol & Life Expectancy

This analysis models how BMI and cholesterol levels predict life expectancy.

* **bmi <- c(22, 25, 28, 30, 35, 40, 45)**

Creates a vector representing Body Mass Index (BMI) values.

* **cholesterol <- c(180, 190, 200, 220, 250, 270, 300)**

Creates a vector representing cholesterol levels.

* **life\_expectancy <- c(80, 78, 76, 74, 70, 68, 65)**

Creates a vector representing life expectancy values.

* **model <- lm(life\_expectancy ~ bmi + cholesterol)**

Builds a multiple linear regression model to predict life expectancy based on BMI and cholesterol.

* **summary(model)**

Displays the regression model’s summary, including coefficients, Rsquared, and statistical significance.

* **ggplot(data.frame(bmi, life\_expectancy), aes(x = bmi, y = life\_expectancy)) + geom\_point(color = "green") + geom\_smooth(method = "lm", col = "red") + labs(title = "BMI vs Life Expectancy", x = "BMI", y**

**= "Life Expectancy")**

Creates a scatter plot with a red regression line to visualize the relationship between BMI and life expectancy.



**Technologies Used:**

* **R Programming-** Used for data analysis, statistical modeling, and visualization.
* **R Studio-** Integrated Development Environment (IDE) for writing and executing Rcode.
* **ggplot2-** Used for creating visualizations like histograms and scatter plots.
* **Base R Functions:** Functions like t.test(), cor(), and lm() for statistical testing, correlation analysis, and linear regression modeling.
* **Statistical Analysis Techniques:** Application of hypothesis testing, correlation, and regression analysis for data-driven decision-making.

**Conclusion**

* **Relationship Analysis-** Identified significant correlations between lifestyle factors and life expectancy.
* **Predictive Modeling -** Developed a regression model that effectively predicted life expectancy based on key health indicators.
* **Data Visualization-** Used graphs and heat maps to better understand trends, distributions and relationships.
* **Key Insights-** Highlighted the importance of healthy life style choices and accessible health care in improving life expectancy.
* **Overall Impact-** Demonstrate how data-driven analysis help in making informed health-related decisions and policies.

# References/ Tools Used

* **R Studio Documentation-** Official guide for R programming and development.
* **R Programming Language** - Core language for statistical computing and data analysis.
* **ggplot2 Package** - For creating advanced data visualizations like histograms and scatter plots.
* **Base R Functions** - Functions like t.test(), cor(), and lm() for statistical analysis.
* **Statistical Analysis Techniques** - Methods for hypothesis testing, correlation, and regression analysis.