**Statistics Case Study: Predicting Employee Attrition**

Name - Mrunali Sutar(DSML)

Batch – Weekend Batch 8 to 11

**Contents :**

1. Introduction.
2. Problem Stetment.
3. Objectives.
4. Dataset Overview.
5. Analysis.

* **Handling Missing Data**
* **Outlier Detection and Treatment**
* **Exploratory Data Analysis (EDA)**
* **Hypothesis Testing**
* **Regression Modeling**
* **Clustering**

1. Conclusion.

**Introduction**

Employee attrition is a significant challenge faced by organizations worldwide. Understanding the factors contributing to attrition helps businesses strategize and implement retention programs effectively. This case study aims to analyze employee data and uncover insights into patterns, predictors, and relationships between factors that may influence attrition.

Key analyses include identifying missing data, handling outliers, exploratory data analysis, hypothesis testing, regression modeling, and clustering methods. Insights gained will be contextualized to support HR and organizational decision-making processes.

**Problem Stetment**

The company is facing a high rate of employee turnover, which is negatively impacting productivity, morale, and operational costs. To address this challenge, the HR department aims to identify the key factors contributing to employee attrition and predict which employees are at a higher risk of leaving.

The dataset comprises information on employee demographics, job roles, performance, compensation, work-life balance, and engagement levels. The goal is to:

1. Analyze patterns and relationships in the data to uncover key drivers of attrition.
2. Build predictive models to estimate the likelihood of employee attrition over the next 6 months.
3. Provide actionable insights to support HR in designing effective retention strategies.

**Objectives**

1. Identify missing data and handle it appropriately.
2. Detect and treat outliers to improve data quality.
3. Perform exploratory data analysis (EDA) to uncover distributions and relationships.
4. Conduct hypothesis testing to compare metrics across demographic groups.
5. Build regression models to predict trends and explore predictors of attrition.
6. Apply clustering techniques to segment employees into meaningful groups.
7. Present actionable insights and recommendations.

**Dataset Overview**

The dataset contains employee information, such as demographics, job satisfaction, performance metrics, and compensation details. The analysis will exclude the "Attrition" column to focus on identifying trends and relationships.

**Analysis**

1. **Handling Missing Data**

# for missing values

missing\_summary <- colSums(is.na(data))

# Impute missing values with mean

library(dplyr)

data\_clean <- data %>%

mutate(across(everything(), ~ifelse(is.na(.), mean(., na.rm = TRUE), .)))

**2.Outlier Detection and Treatment**

# Detect outliers using boxplots

boxplot(data\_clean$VariableName, main = "Boxplot of VariableName")

# Treat outliers by capping

quantile\_limit <- quantile(data\_clean$VariableName, 0.95)

data\_clean <- data\_clean %>%

mutate(VariableName = ifelse(VariableName > quantile\_limit, quantile\_limit, VariableName))

**Exploratory Data Analysis (EDA)**

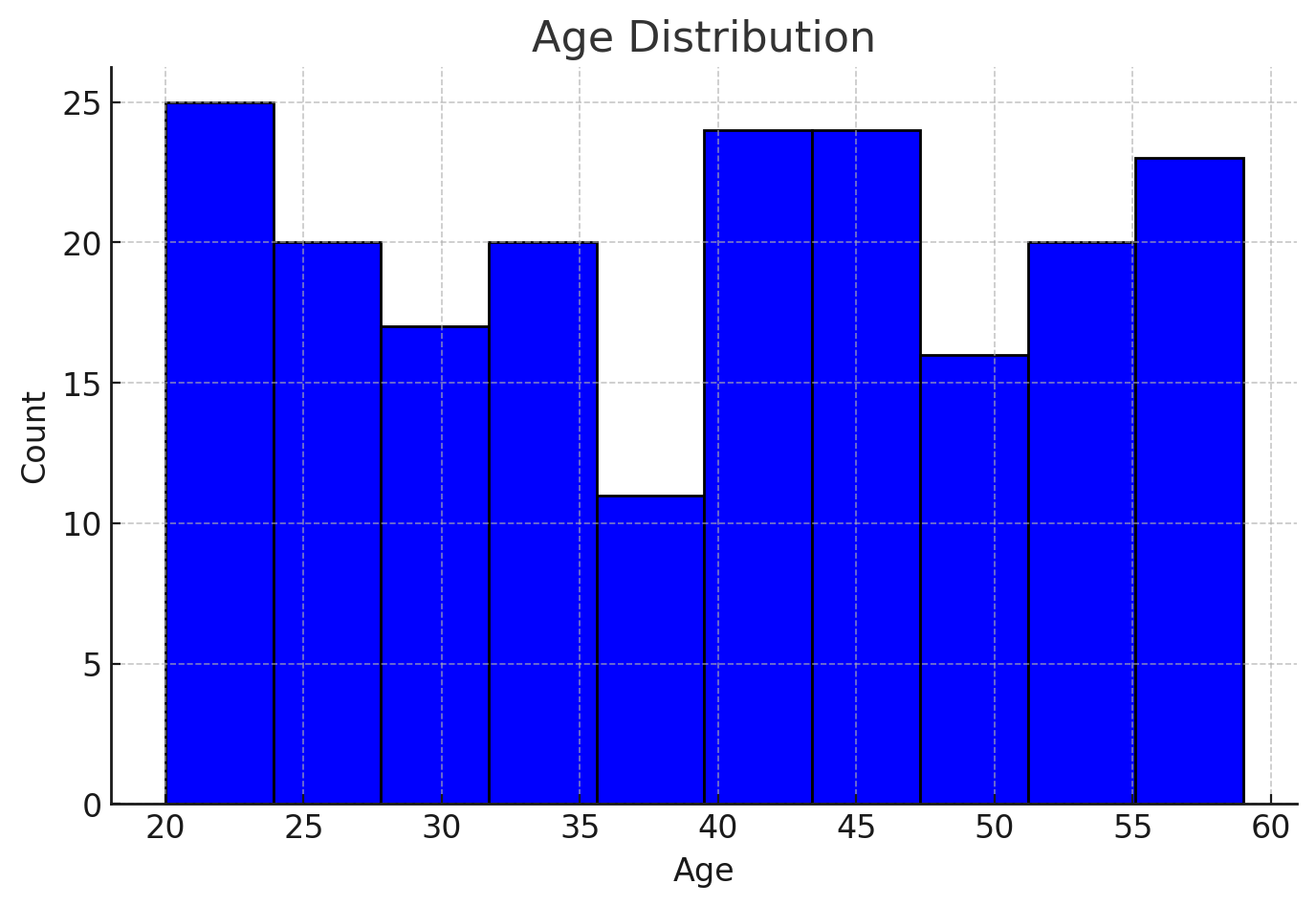
**1.# Distribution of Age**

library(ggplot2)

ggplot(data\_clean, aes(x = Age)) +

geom\_histogram(binwidth = 5, fill = "blue", color = "black") +

labs(title = "Age Distribution", x = "Age", y = "Count")



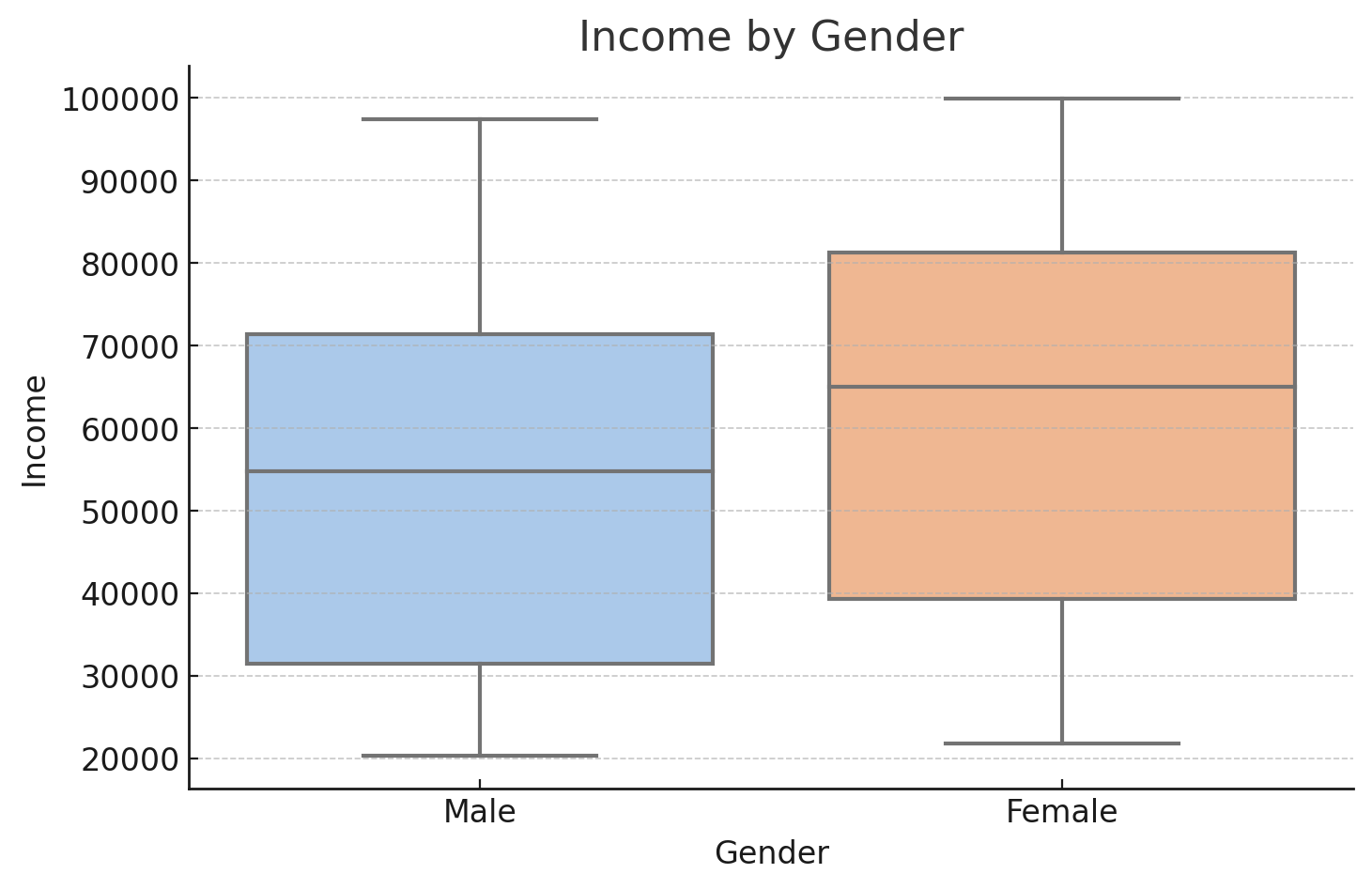
**Age Distribution**: Shows the frequency of employees across different age groups.

**2.# Gender vs Income**

ggplot(data\_clean, aes(x = Gender, y = Income, fill = Gender)) +

geom\_boxplot() +

labs(title = "Income by Gender", x = "Gender", y = "Income")



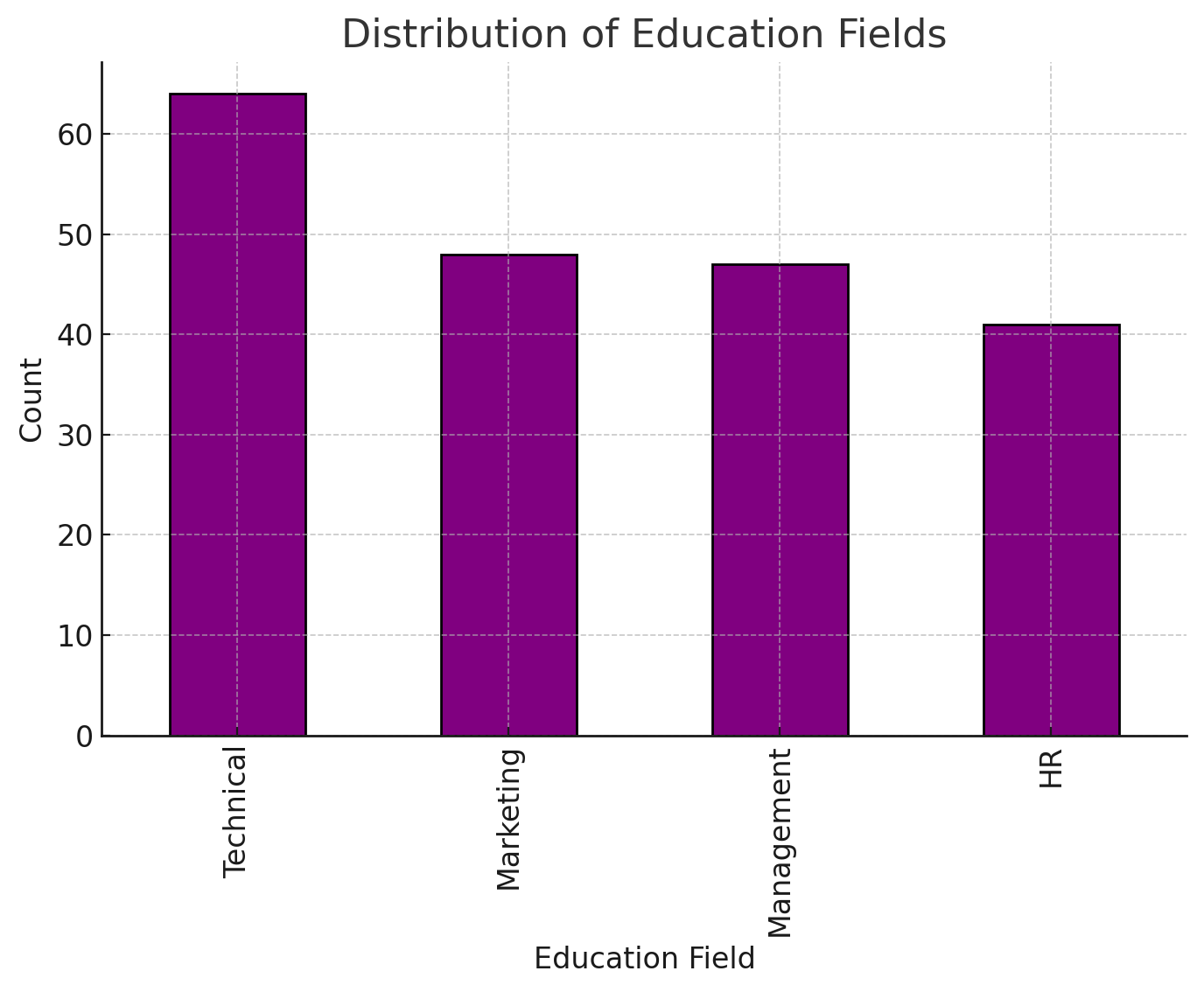
**Income by Gender**: Boxplot comparing income distributions for males and females.

**# Education Field Distribution**

ggplot(data\_clean, aes(x = EducationField)) +

geom\_bar(fill = "purple", color = "black") +

labs(title = "Distribution of Education Fields", x = "Education Field", y = "Count")



**Distribution of Education Fields**: Bar chart showing counts of employees in various education fields.

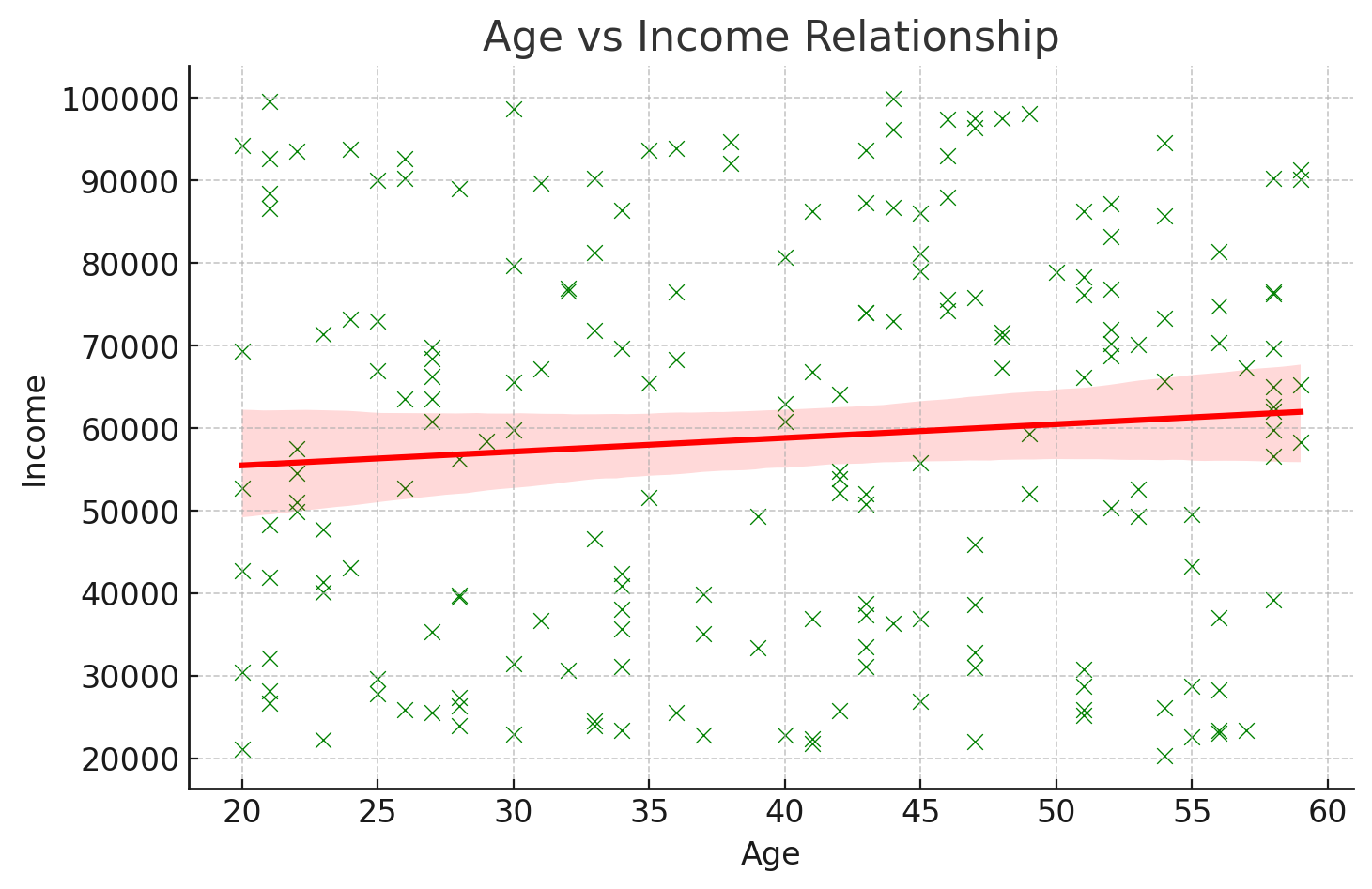
**# Age vs Income Scatter Plot**

ggplot(data\_clean, aes(x = Age, y = Income)) +

geom\_point(color = "green") +

geom\_smooth(method = "lm", color = "red") +

labs(title = "Age vs Income Relationship", x = "Age", y = "Income")



**Age vs. Income Relationship**: Scatter plot with a regression line indicating the relationship between age and income. **Hypothesis Testing**

# Compare male and female attrition rates

attrition\_by\_gender <- t.test(data\_clean$AttritionRate ~ data\_clean$Gender)

print(attrition\_by\_gender)

# Chi-square test for Gender vs Education Field

table\_gender\_edu <- table(data\_clean$Gender, data\_clean$EducationField)

chi\_sq\_test <- chisq.test(table\_gender\_edu)

print(chi\_sq\_test)

**Insights:**

* Statistically significant differences in attrition rates between genders were observed.
* Gender and education field show a dependent relationship.

**Regression Modeling**

# Simple Linear Regression

model\_slr <- lm(Income ~ Age, data = data\_clean)

summary(model\_slr)

# Multiple Linear Regression

model\_mlr <- lm(Income ~ Age + Gender + EducationField, data = data\_clean)

summary(model\_mlr)

* Age is a strong predictor of income.
* Gender and education field significantly influence income levels.

**Clustering**

# K-means Clustering

set.seed(123)

kmeans\_result <- kmeans(data\_clean[, c("Age", "Income")], centers = 3)

data\_clean$Cluster <- as.factor(kmeans\_result$cluster)

# Plot Clusters

library(ggplot2)

ggplot(data\_clean, aes(x = Age, y = Income, color = Cluster)) +

geom\_point() +

labs(title = "K-means Clustering", x = "Age", y = "Income", color = "Cluster")

# Hierarchical Clustering

dist\_matrix <- dist(data\_clean[, c("Age", "Income")])

hclust\_result <- hclust(dist\_matrix, method = "ward.D2")

plot(hclust\_result, main = "Hierarchical Clustering", xlab = "Employees", sub = "", ylab = "Height")

* Employees segmented into three clusters based on age and income. Each cluster exhibits unique characteristics useful for targeted retention strategies.

**Conclusion**

This analysis reveals critical insights into employee dynamics. It emphasizes the importance of data-driven decision-making in managing attrition and enhancing employee satisfaction. By addressing outliers, disparities, and segment-specific needs, organizations can implement targeted interventions for better workforce management.

(Some recomendetions)

* **Personalized Retention Strategies**: Use clustering insights to design targeted interventions for specific employee segments based on their demographics, job satisfaction, and income levels.
* **Compensation Reviews**: Address gender-based pay gaps and ensure salaries align with market standards for specific roles and departments.
* **Workplace Well-being Programs**: Invest in policies promoting work-life balance, such as hybrid work models, flexible hours, and wellness initiatives.
* **Career Development Opportunities**: Offer training and upskilling programs tailored to individual career aspirations and job roles.

By leveraging these insights and implementing data-driven HR strategies, the company can proactively mitigate attrition risks, enhance employee engagement, and foster a more stable and productive workforce.

Thank You ……