MAHATMA EDUCATION SOCIETY'S

PILLAI COLLEGE OF ARTS, COMMERCE & SCIENCE
(Autonomous)

NEW PANVEL

PROJECT REPORT ON

# "HR Analytics: Understanding Employee Satisfaction and Attrition"

IN PARTIAL FULFILLMENT OF

MASTERS OF DATA ANALYTICS

SEMESTER I – 2023-24

PROJECT GUIDE

Name: Prof. Sabitha Praveen

SUBMITTED BY: Ashwin Suresh

ROLL NO: 3104

# RESEARCH METHODOLOGY
# CA-2 PROJECT

## HR Analytics: Understanding Employee Satisfaction and Attrition


### Abstract

This comprehensive project involves conducting an Exploratory Data Analysis (EDA) on a dataset containing employee information, with a focus on factors that may influence employee turnover. The analysis utilizes the R programming language and various data visualization techniques to uncover insights into employee satisfaction, workload, and other attributes. It contain mostly about data preparation, data preprocessing, descriptive data analytics (what is happening?) using various statistical and data visualization techniques, analyzing the human resources data set, doing multi-variate, bi-variate and univariate analysis. Statistical tests, including Chi-Square tests, T-tests, and ANOVA, are applied to assess the significance of various factors.


### Introduction

Employee turnover can have a significant impact on an organization's productivity, morale, and bottom line. Understanding the factors that contribute to employee turnover is crucial for HR professionals and management. In this project, we explore a dataset containing employee information, including the number of projects, average monthly hours worked, salary, satisfaction level, work accidents, employee turnover, time spent at the company, department, and promotion status over the last 5 years.

### Objectives

The primary objectives of this project are as follows:
>Explore the dataset and gain a comprehensive understanding of its structure and variables.
>Preprocess the data to ensure its suitability for analysis.
>Calculate key statistical measures such as mean, median, standard deviation, and coefficient of variation for the "satisfaction_level" column.
>Visualize the relationships between variables using appropriate plots, including scatter plots, box plots, bar plots, and more.
>Perform statistical tests to assess the significance of various factors in relation to employee turnover.
>Provide meaningful insights and conclusions based on the analysis to help organizations better understand and address employee turnover issues.

The dataset has 9 attributes(variables) and 2121 observations(instances).

| Column Names | Data Types | Values | Description |
|---|---|---|---|
| number_project | int64 | Numerical | number of projects an employee is involved in. |
| average_monthly_hours | Object | Categorical | The average number of monthly work hours for an employee. |
| salary | float64 | Numerical | salary level of an employee, categorized as "low," "medium," or "high." |
| satisfaction_level | int64 | Numerical | satisfaction level of an employee, which could be a measure of job satisfaction. |
| work_accident | int64 | Numerical | Indicates whether the employee has had a work-related accident (1 for yes, 0 for no). |
| left | int64 | Numerical | Indicates whether the employee has left the company (1 for yes, 0 for no), which could be a target variable for prediction. |
| time_spend_company | int64 | Numerical | number of years the employee has spent at the company. |
| sales | Object | Categorical | The department or sales sector where the employee works. |
| promotion_last_5years | int64 | Numerical | Indicates whether the employee has been promoted in the last 5 years (1 for yes, 0 for no). |

## Research Methodology

### Data Preprocessing

- The dataset is loaded using the read.csv function, and its structure is examined using str and summary.

- Missing values are checked for using colsums(is.na(hrm)), and data types are converted as needed.
- Data is categorized and converted into appropriate data types for analysis.

**Data Analysis and Interpretation**

- Descriptive statistics such as mean, median, standard deviation, and coefficient of variation are calculated for the "satisfaction_level" column.
- Data visualizations are created to explore relationships between variables, including scatter plots, box plots, bar plots, and more.
- Statistical tests are performed to assess the significance of various factors, including Chi-Square tests, T-tests, and ANOVA.

# CODE AND INPUT:

### #Installing and importing the libraries

```
install.packages("ggplot2")
library(ggplot2)
library(dplyr)
library(tidyr)

#loading the Dataset
hrm<-read.csv('C:\\Users\\Comp\\Desktop\\New folder (3)\\Book1.csv',
              header = T, stringsAsFactors = F)
```

### #converting variables to factor

```
> hrm$left <- as.factor(hrm$left)
> hrm$salary <- as.factor(hrm$salary)
> hrm$sales <- as.factor(hrm$sales)
> hrm$work_accident <- as.factor(hrm$work_accident)
> hrm$promotion_last_5years <- as.factor(hrm$promotion_last_5years)
```

### #Structure of the Dataset

```
> str(hrm)
'data.frame':   2121 obs. of  9 variables:
 $ number_project       : int  5 6 4 6 6 3 4 4 4 6 ...
 $ average_monthly_hours: int  208 223 181 219 229 170 200 200 171 227 ...
 $ salary               : Factor w/ 3 levels "high","low","medium": 2 2 2 2 2 3 3 3 3 2 ...
 $ satisfaction_level   : num  0.27 0.19 0.43 0.2 0.14 0.44 0.38 0.32 0.55 0.21 ...
 $ work_accident        : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 2 1 ...
 $ left                 : Factor w/ 2 levels "0","1": 2 2 2 2 2 1 1 2 2 2 ...
 $ time_spend_company   : int  2 2 1 2 2 2 3 2 3 2 ...
 $ sales                : Factor w/ 6 levels "hr","management",..: 6 1 5 1 3 4 4 2 2 1 ...
 $ promotion_last_5years: Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 2 1 1 ...
>
```

## #Summary Statistics of the dataset

```
> summary(hrm)
 number_project  average_monthly_hours    salary    satisfaction_level
 Min.   :2.000   Min.   :130.0         high  :426   Min.   :0.1000
 1st Qu.:3.000   1st Qu.:154.0         low   :633   1st Qu.:0.3700
 Median :4.000   Median :182.0         medium:741   Median :0.5000
 Mean   :3.917   Mean   :180.7                      Mean   :0.4957
 3rd Qu.:5.000   3rd Qu.:206.0                      3rd Qu.:0.6300
 Max.   :6.000   Max.   :242.0                      Max.   :0.9000
 work_accident left     time_spend_company        sales
 0:1401        0:1033   Min.   :1.000      hr         :265
 1: 399        1: 767   1st Qu.:1.000      management:333
                        Median :2.000      marketing :319
                        Mean   :2.342      sales     :292
                        3rd Qu.:3.000      support   :292
                        Max.   :5.000      technical :299
 promotion_last_5years
 0:1658
 1: 142
```

## #Checking the null values in columns

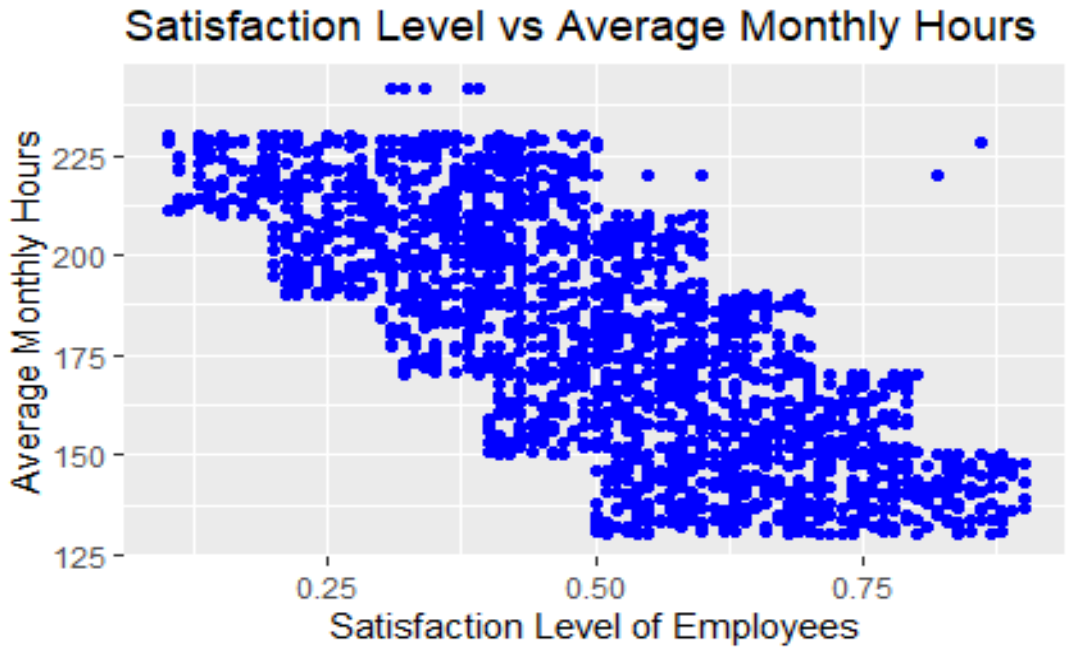
```
> colSums((is.na(hrm)))
       number_project average_monthly_hours                 salary
                    0                     0                      0
   satisfaction_level         work_accident                   left
                    0                     0                      0
   time_spend_company                 sales promotion_last_5years
                    0                     0                      0
```

## #satisfaction_level column is analyzed, and the mean, median, standard deviation, and coefficient of variation are calculated.

```
> satisfaction_stats <- summary(hrm$satisfaction_level)
> mean_satisfaction <- satisfaction_stats["Mean"]
> median_satisfaction <- satisfaction_stats["Median"]
> sd_satisfaction <- sd(hrm$satisfaction_level, na.rm = TRUE)
> cv_satisfaction <- (sd_satisfaction / mean_satisfaction) * 100
> cat(paste("Mean Satisfaction Level:", mean_satisfaction, "\n"))
Mean Satisfaction Level: 0.498694012258369
> cat(paste("Median Satisfaction Level:", median_satisfaction, "\n"))
Median Satisfaction Level: 0.5
> cat(paste("Standard Deviation of Satisfaction Level:", sd_satisfaction, "\n"))
Standard Deviation of Satisfaction Level: 0.179250907755334
> cat(paste("Coefficient of Variation (CV) of Satisfaction Level:", cv_satisfaction,
"%\n"))
Coefficient of Variation (CV) of Satisfaction Level: 35.9440665717208 %
```

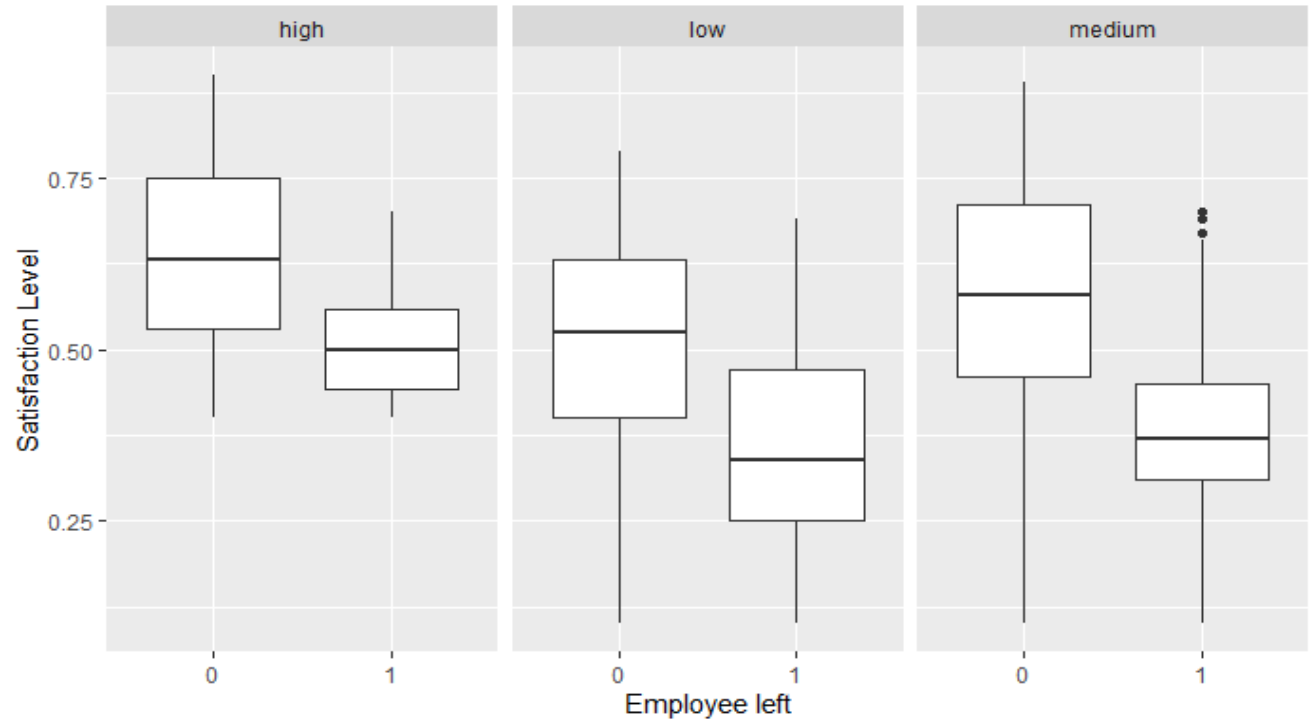
## #Box plot for Satisfaction Level vs Average Monthly Hours

```
ggplot(aes(x = satisfaction_level, y = average_monthly_hours), data = hrm) +
  geom_point(color = 'blue') +
  labs(title = "Satisfaction Level vs Average Monthly Hours",
       x = 'Satisfaction Level of Employees',
       y = 'Average Monthly Hours')
```




This graph indicates that employee with higher satisfaction level, the more they have to work monthly.

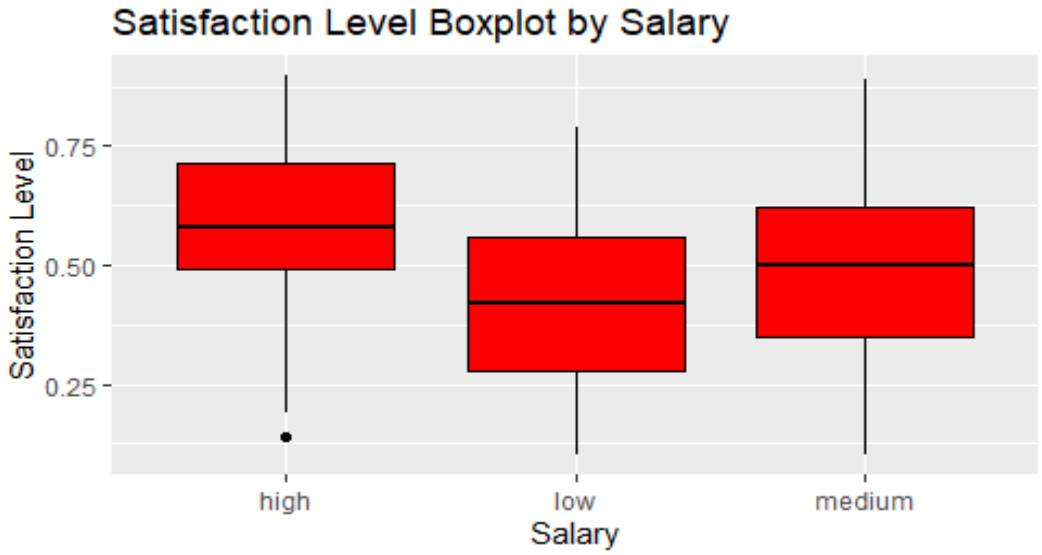
```
#Boxplot for Satisfaction level vs left facetted by Salary Ranges
ggplot(aes(x = left,y=satisfaction_level),data= hrm) +
  geom_boxplot() +
  ylab('Satisfaction Level') +
  xlab("Employee left") +
  facet_wrap(~salary)
```




This graph indicates that employee with higher satisfaction level stayed in the company and with lower satisfaction level left the company

**#Satisfaction Level Boxplot by Salary**

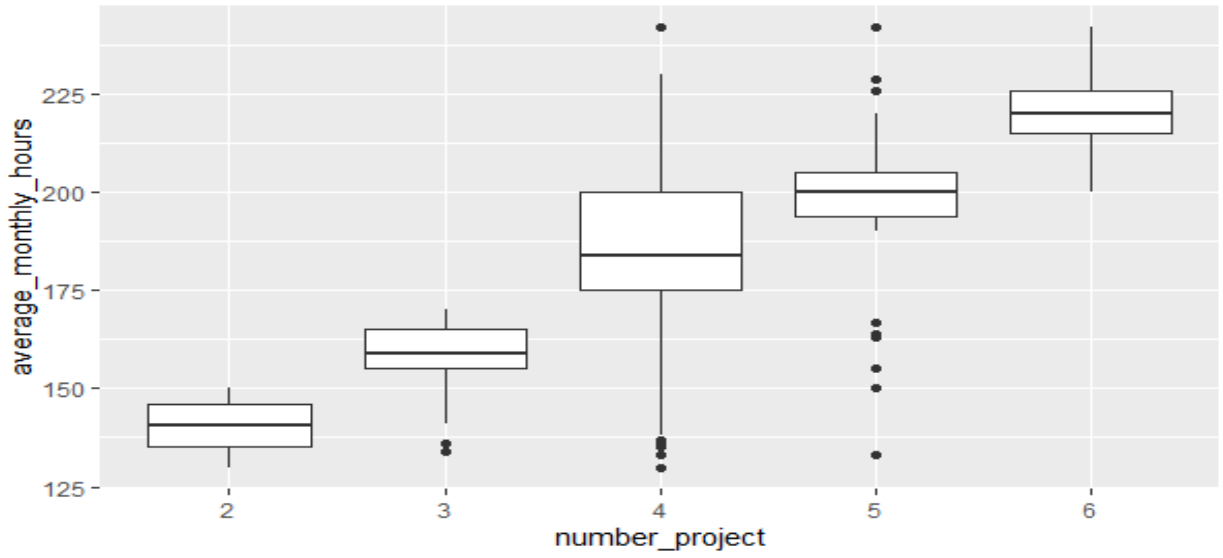
```
p1 <- ggplot(aes(x = salary, y = satisfaction_level), data = hrm) +
  geom_boxplot(color = "black", fill = "red") +
  labs(title = "Satisfaction Level Boxplot by Salary", x = "Salary",
       y = "Satisfaction Level")
p1
```




The above graph indicates that employee with high salary have higher satisfaction level and vice versa. So hike in their salary may make them more satisfied.

**#boxplot of number of projects vs Average monthly hours**

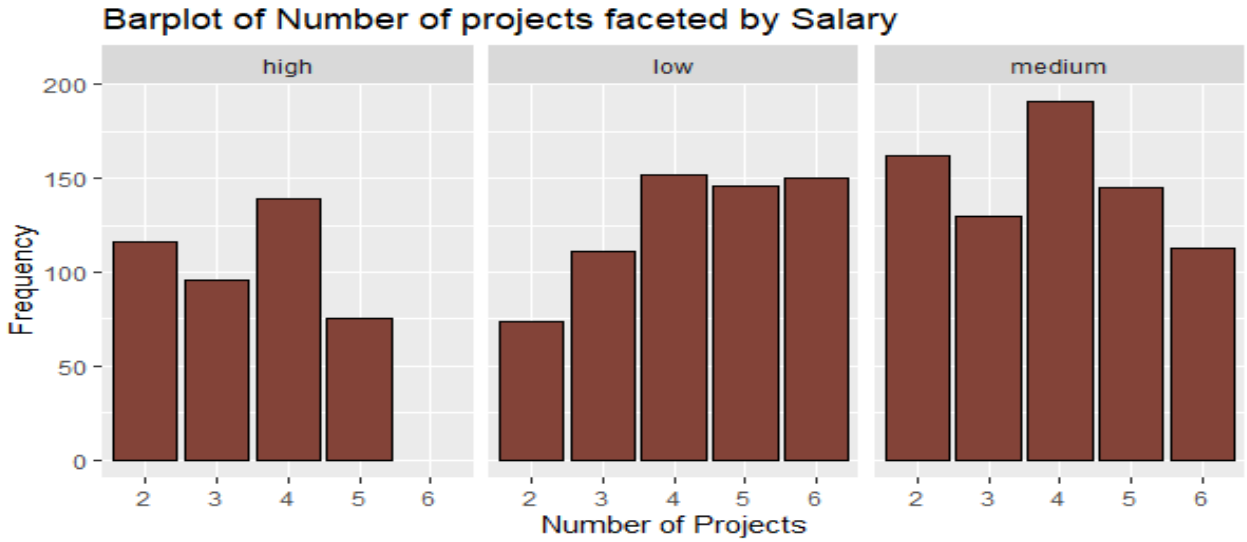
```
hrm$number_project<-factor(hrm$number_project)
p3 <- ggplot(aes(x = average_monthly_hours, y = number_project), data = hrm) +
  geom_boxplot() +
  labs(title = "Boxplot of Number of Projects vs Average Monthly Hours",
       x = "Average Monthly Hours",
       y = "Number of Projects")
p3
```



We can see that increase in number of projects takes employee to work for more time. So distributing the projects equally may help the employees feel less pressurized.

**#barplot of number of projects facetted by salary**

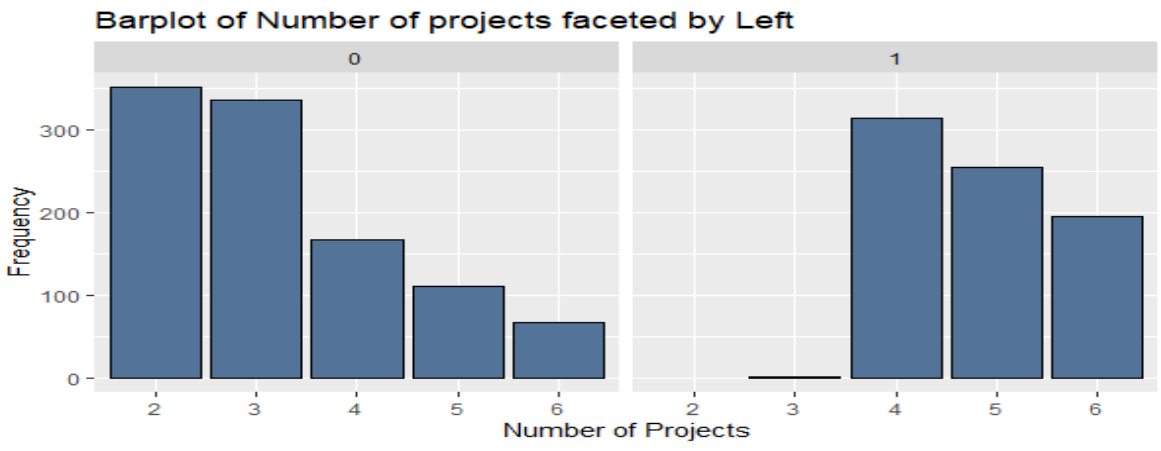
```
#facetted by salary
ggplot(aes(x=number_project),data = hrm) +
  geom_bar(color='black',fill='#834338') +
  xlab("Number of Projects") +
  ylab("Frequency") +
  labs(title="Barplot of Number of projects faceted by Salary") +
  facet_wrap(~salary)
```




Salary: low , medium, high . Employees with higher salary gets less number of projects and vice versa

**#barplot of number of projects faceted by left**

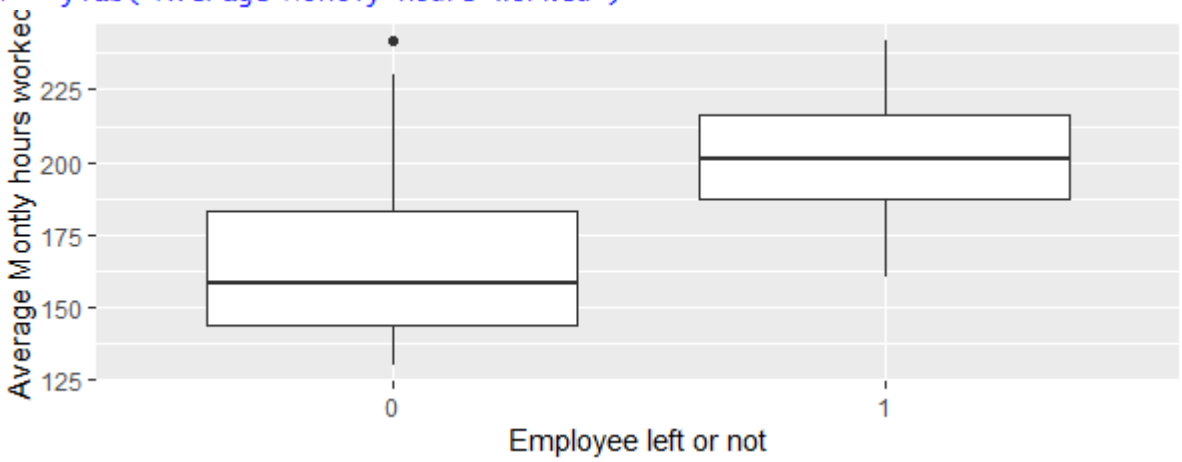
```
#faceted by If a employee left or not
ggplot(aes(x=number_project),data = hrm) +
  geom_bar(color='black',fill='#547398') +
  xlab("Number of Projects") +
  ylab("Frequency") +
  labs(title="Barplot of Number of projects faceted by Left")+
  facet_wrap(~left)
```



Employees left the company when they are given higher no.of projects so we should minimize the no.of projects so that we can control the employee leaving the company

**#boxplot of average_monthly_hours vs left**

```
> ggplot(aes(y = average_monthly_hours, x = hrm$left),data=hrm)+
+   geom_boxplot() +
+   xlab("Employee left or not") +
+   ylab("Average Montly hours worked")
```



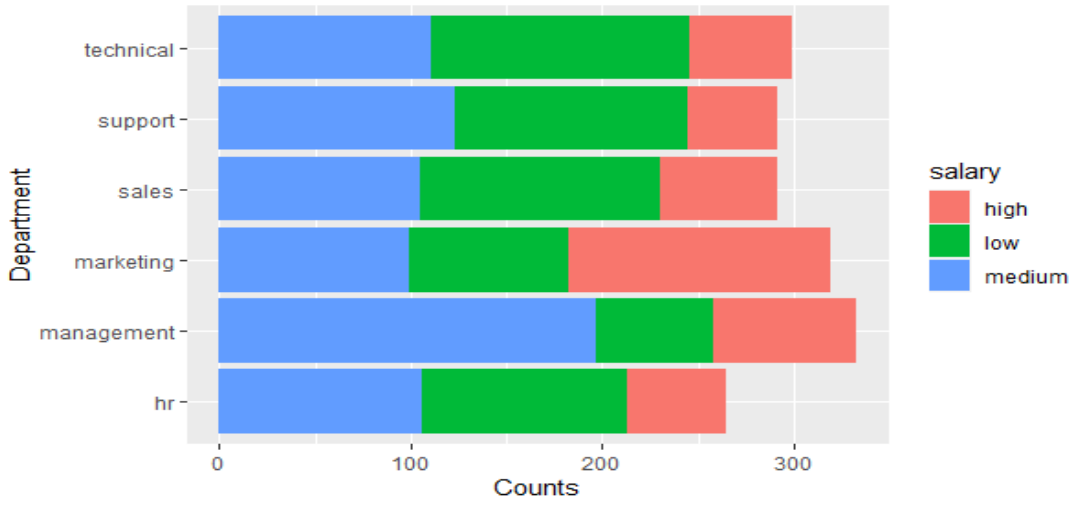

A thing to notice is that employee who left the company worked more hours than those who did not leave, hence it might be possible that they left because they were over pressurized by their peers or bosses or over worked or stressed with lots of work.

```
#Time spend at company vs Left or not
ggplot(aes(x = factor(time_spend_company)),data = hrm) +
  geom_bar(fill = 'grey',color='black') +
  xlab("Time spend at compnay in years") +
  ylab("Frequency")+
  labs(title = "Barplot of Time spend at Company faceted by Left") +
  facet_wrap(~left)
```


Employees leaving the company within starting years is very high. So the company should keep the environment peaceful for the employees joining the company

**#Barplot of 'sales' faceted by 'salary'**

```
> ggplot(aes(x =sales),data = hrm ) +
+   geom_bar(aes(fill=salary)) +
+   xlab('Department') +
+   ylab('Counts') +
+   coord_flip()
```




Management and Marketing department has the most no.of employees having high salary among other departments

# Performing Hypothesis Tests:

## Chi-Square Test

Null Hypothesis (H0): There is no association between number_project and left.
Alternative Hypothesis (H1): There is an association between number_project and left.

```
> # Create a contingency table
> contingency_table <- table(hrm$number_project, hrm$left)
> # Perform Chi-Square Test
> chi_square_result <- chisq.test(contingency_table)
> # Print the Chi-Square Test result along with hypotheses
> cat("Chi-Square Test Result:\n")
Chi-Square Test Result:
> print(chi_square_result)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 986.7, df = 4, p-value < 2.2e-16

> cat("\nHypotheses:\n")

Hypotheses:
> if (chi_square_result$p.value < 0.05) {
+   cat("Reject H0: There is an association between
+       number_project and left.\n")
+ } else {
+   cat("Fail to reject H0: There is no association between
+       number_project and left.\n")
+ }
Reject H0: There is an association between
      number_project and left.
```

The Chi-Square Test results strongly suggest that there is a significant association between the number of projects an employee is assigned to ("number_project") and whether or not they left the company ("left"). The extremely low p-value (less than 0.05) indicates that we can reject the null hypothesis, which means that the number of projects and employee attrition are not independent of each other. In other words, the number of projects an employee is involved in may have a statistically significant impact on their decision to leave or stay in the company.
This finding implies that HR should carefully monitor and manage the workload and project allocation for employees to reduce the likelihood of attrition.

## Welch Two-Sample T-Test

Null Hypothesis (H0): There is no significant difference in satisfaction_level between employees who left and those who did not.
Alternative Hypothesis (H1): There is a significant difference in satisfaction_level between employees who left and those who did not.

```
> # Perform T-Test
> t_test_result <- t.test(hrm$satisfaction_level ~ hrm$left)
> # Print the T-Test result along with hypotheses
> cat("T-Test Result:\n")
T-Test Result:
> print(t_test_result)

        Welch Two Sample t-test

data:  hrm$satisfaction_level by hrm$left
t = 26.52, df = 2101.9, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is n
ual to 0
95 percent confidence interval:
 0.1638123 0.1899739
sample estimates:
mean in group 0 mean in group 1
      0.5742551       0.3973620

> cat("\nHypotheses:\n")

Hypotheses:
> if (t_test_result$p.value < 0.05) {
+   cat("Reject H0: There is a significant difference
+       in satisfaction_level between employees who left
+       and those who did not.\n")
+ } else {
+   cat("Fail to reject H0: There is no significant difference
+       in satisfaction_level between employees who left and
+       those who did not.\n")
+ }
Reject H0: There is a significant difference
      in satisfaction_level between employees who left
      and those who did not.
```

The above Welch Two-Sample t-test provides strong evidence to conclude that there is a significant difference in satisfaction levels between employees who left the company and those who did not. In other words, employees who left had, on average, significantly lower satisfaction levels compared to those who stayed. This finding suggests that employee satisfaction plays a crucial role in employee retention. Companies should pay attention to the satisfaction levels of their employees and take measures to improve job satisfaction to reduce employee turnover.

## ANOVA Test

Null Hypothesis (H0): There is no significant difference in average_monthly_hours among different number_project groups.
Alternative Hypothesis (H1): There is a significant difference in average_monthly_hours among different number_project groups.

```
> # Perform ANOVA Test
> anova_result <- aov(average_monthly_hours ~ number_project, data = hrm)
> # Extract the p-value from the ANOVA result
> p_value <- anova(anova_result)$'Pr(>F)'[1]
> # Print the ANOVA Test result along with hypotheses
> cat("ANOVA Test Result:\n")
ANOVA Test Result:
> print(summary(anova_result))
                Df  Sum Sq Mean Sq F value Pr(>F)
number_project    4 1543470  385867    2706 <2e-16 ***
Residuals      2116  301704     143
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> cat("\nHypotheses:\n")

Hypotheses:
> if (p_value < 0.05) {
+   cat("Reject H0: There is a significant difference in
+       average_monthly_hours among different number_project groups.\n")
+ } else {
+   cat("Fail to reject H0: There is no significant difference
+       in average_monthly_hours among different number_project groups.\n")
+ }
Reject H0: There is a significant difference in
      average_monthly_hours among different number_project groups.
```

This finding implies that the number of projects assigned to employees has a significant impact on the average number of monthly hours they work. It may suggest that employees with different project loads have different workloads or time demands. Organizations should consider workload management strategies and employee productivity when assigning projects to optimize work hours and employee satisfaction.

## Conclusion

In summary, the analysis suggests that employee satisfaction, workload, salary, and tenure are all factors that may influence employee attrition within the organization. Specifically, lower satisfaction levels, higher workloads, lower salaries, and shorter tenures are associated with higher turnover rates. These findings can guide HR and management in making data-driven decisions to improve employee retention and satisfaction.

## References

Dataset Source: Github
Additional References:
Visualization: https://www.geeksforgeeks.org/
Hypothesis Testing: https://data-flair.training/blogs/hypothesis-testing-in-r/
https://www.geeksforgeeks.org/anova-test-in-r-programming/
ChatGPT: https://chat.openai.com/