IN PARTNERSHIP WITH PLYMOUTH UNIVERSITY

Name: Group No. -
Student Reference Number: 10899486

Module Code: PUSL 2076	Module Name: Data programming in R
Coursework Title: Final Coursework Report Submission	
Deadline Date: 10/01/2024	Member of staff responsible for coursework:
Program: BSc.(hons) in Data Science	

Please note that University Academic Regulations are available under Rules and Regulations on the University website www.plymouth.ac.uk/studenthandbook.

Group work: please list all names of all participants formally associated with this work and state whether the work was undertaken alone or as part of a team. Please note you may be required to identify individual responsibility for component parts.

* Fernando Naveen – Age
* Weerasinghe dissanayake – Marital status
* gosisa jinasena - Education
* sangarange sandanayke - Experience
* urulugastenne amarakone - Gender

We confirm that we have read and understood the Plymouth University regulations relating to Assessment Offences and that we are aware of the possible penalties for any breach of these regulations. We confirm that this is the independent work of the group.

Signed on behalf of the group: *S.chamara*

Individual assignment: ***I confirm that I have read and understood the Plymouth University regulations relating to Assessment Offences and that I am aware of the possible penalties for any breach of these regulations. I confirm that this is my own independent work.***

Signed:

Use of translation software: failure to declare that translation software or a similar writing aid has been used will be treated as an assessment offence.

I *have used/not used translation software.

If used, please state name of software……

UNIVERSITY OF PLYMOUTH

PUSL2076 Data Programming In R (23/AU/M)

Final Report

PUSL 2076 - Group Assignment

Group Members Details

PU Index No.	Student Name	Degree Program
10899177	Urulugastenne Amarakone	Data science
10899302	Weerasinghe Dissanayaka	Data science
10899478	Gosisa Jinasena	Data science
10899483	Fernando Naveen	Data science
10899486	Sangarange Sandanayake	Data science

CONTENTS

Introduction **5**

Correlation matrix heatmap 7

1.Age **8**

1.Data Visualization 8

1. Relationship between an employee's age group, the amount of business travel they do, and how it affects their pay each month. 9

2. Proportion of Departments in Age Group. 11

3. Distance from home Across Age Groups 12

4. Distribution of Job Levels Across Age Group 13

2.Data analytics through models 14

1.Random Forrest Classification on Age group 14

2.Naïve Bayes Classification on Age Group 16

3.Decision tree classification on Attrition 17

3.Findings and Discussion 19

2. Gender **20**

1.Data Visualizations 20

1. Gender representation of the company 20

2. Gender distribution among departments 21

3. Education distribution among Department 22

4. Salary difference between gender 23

5. Environment Satisfaction between gender 24

6. Job Satisfaction between gender 25

7. Work and life balance between gender 26

8. Job involvement between gender 27

9. Salary difference between gender 28

2.Data analytics through models 31

1.Implementing Navi bayes classification to Gender 31

2.Implementing K-means clustering for the Gender ,Age and Years at the company 32

3. Applying clustering Algorithm for the Gender 34

3.Total Working Years (Experience) 35
1.Data visualizations 35
1. Main Column Introduction 35
2. How experience vary on employee age distribution 37
3. More analysis about experience with boxplots 38
4. Employee attrition rate with their experience 42
5. How job level varies on employee experience display with age 43
6. Employee job satisfaction rate with their experience and job level 44
7. Employee monthly income with their experience and job level 45
8. More analysis about job level with boxplots 46
2. Data analytics through models 47
1. Predict attrition rate using Logistic regression model 47
4. Education 50
1. Data Visualization 50
2. Data Analytics through models 56
1. A classification model using Logistic Regression algorithm to predict Attrition based on Monthly Income and Education Level. 56
5. Marital Status 60
1. Data Visualization 60
1. Marital Status related with Age of the Employees 62
2. Marital Status based on Gender of the Employees 64
2. Data Analytics through models 67
1. Naïve Bayes model 67
2. Decision Tree Classification on Marital Status 69
3.Findings and Discussion 71
6. Appendices 72

Introduction

The objective of this project is to examine dataset using R language analytical tools and techniques. So, we have selected an HR dataset of a pharmaceutical company to conduct the analysis. Understanding HR data becomes critical for organizational success in the business, where people acquisition, retention, and development play crucial roles. The dataset used in this analysis includes employee-related data from internal databases, including training records, performance indicators, and demographics.

Finding insights that help improve decision-making, optimize HR initiatives, and lead to a more productive and efficient staff are some of our goals. This project uses R's exploratory data analysis and predictive modelling tools to deliver useful insights and recommendations for HR management.

```
> str(dataset)
'data.frame':   1480 obs. of  38 variables:
 $ EmpID                   : chr  "RM297" "RM302" "RM458" "RM728" ...
 $ Age                     : int  18 18 18 18 18 18 18 18 19 19 ...
 $ AgeGroup                : chr  "18-25" "18-25" "18-25" "18-25" ...
 $ Attrition               : chr  "Yes" "No" "Yes" "No" ...
 $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Rarely" "Trav
 $ DailyRate               : int  230 812 1306 287 247 1124 544 1431 52
 $ Department              : chr  "Research & Development" "Sales" "Sal
 $ DistanceFromHome        : int  3 10 5 5 8 1 3 14 22 3 ...
 $ Education               : int  3 3 3 2 1 3 2 3 1 1 ...
 $ EducationField          : chr  "Life Sciences" "Medical" "Marketing"
 $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ EmployeeNumber          : int  405 411 614 1012 1156 1368 1624 1839
 $ EnvironmentSatisfaction : int  3 4 2 2 3 4 2 2 4 2 ...
 $ Gender                  : chr  "Male" "Female" "Male" "Male" ...
 $ HourlyRate              : int  54 69 69 73 80 97 70 33 50 79 ...
 $ JobInvolvement          : int  3 2 3 3 3 3 3 3 3 3 ...
 $ JobLevel                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ JobRole                 : chr  "Laboratory Technician" "Sales Repres
 $ JobSatisfaction         : int  3 3 2 4 3 4 4 3 3 2 ...
 $ MaritalStatus           : chr  "Single" "Single" "Single" "Single" .
 $ MonthlyIncome           : int  1420 1200 1878 1051 1904 1611 1569 15
 $ SalarySlab              : chr  "Upto 5k" "Upto 5k" "Upto 5k" "Upto 5
 $ MonthlyRate             : int  25233 9724 8059 13493 13556 19305 184
 $ NumCompaniesWorked      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
 $ OverTime                : chr  "No" "No" "Yes" "No" ...
 $ PercentSalaryHike       : int  13 12 14 15 12 15 12 16 19 14 ...
 $ PerformanceRating       : int  3 3 3 3 3 3 3 3 3 3 ...
 $ RelationshipSatisfaction: int  3 1 4 4 4 3 3 3 4 4 ...
 $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
 $ StockOptionLevel        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TotalWorkingYears       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ TrainingTimesLastYear   : int  2 2 3 2 0 5 2 4 2 3 ...
 $ WorkLifeBalance         : int  3 3 3 3 3 4 4 1 2 3 ...
 $ YearsAtCompany          : int  0 0 0 0 0 0 0 0 0 1 ...
 $ YearsInCurrentRole      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ YearsSinceLastPromotion : int  0 0 0 0 0 0 0 0 0 0 ...
 $ YearsWithCurrManager    : int  0 0 0 0 0 0 0 0 0 0 ...
```

The structure of the HR dataset we selected is shown above.

We conducted our project by dividing our dataset into 5 main categories,

1. **Age**
2. **Gender**
3. **Experience**
4. **Education**
5. **Marital Status**

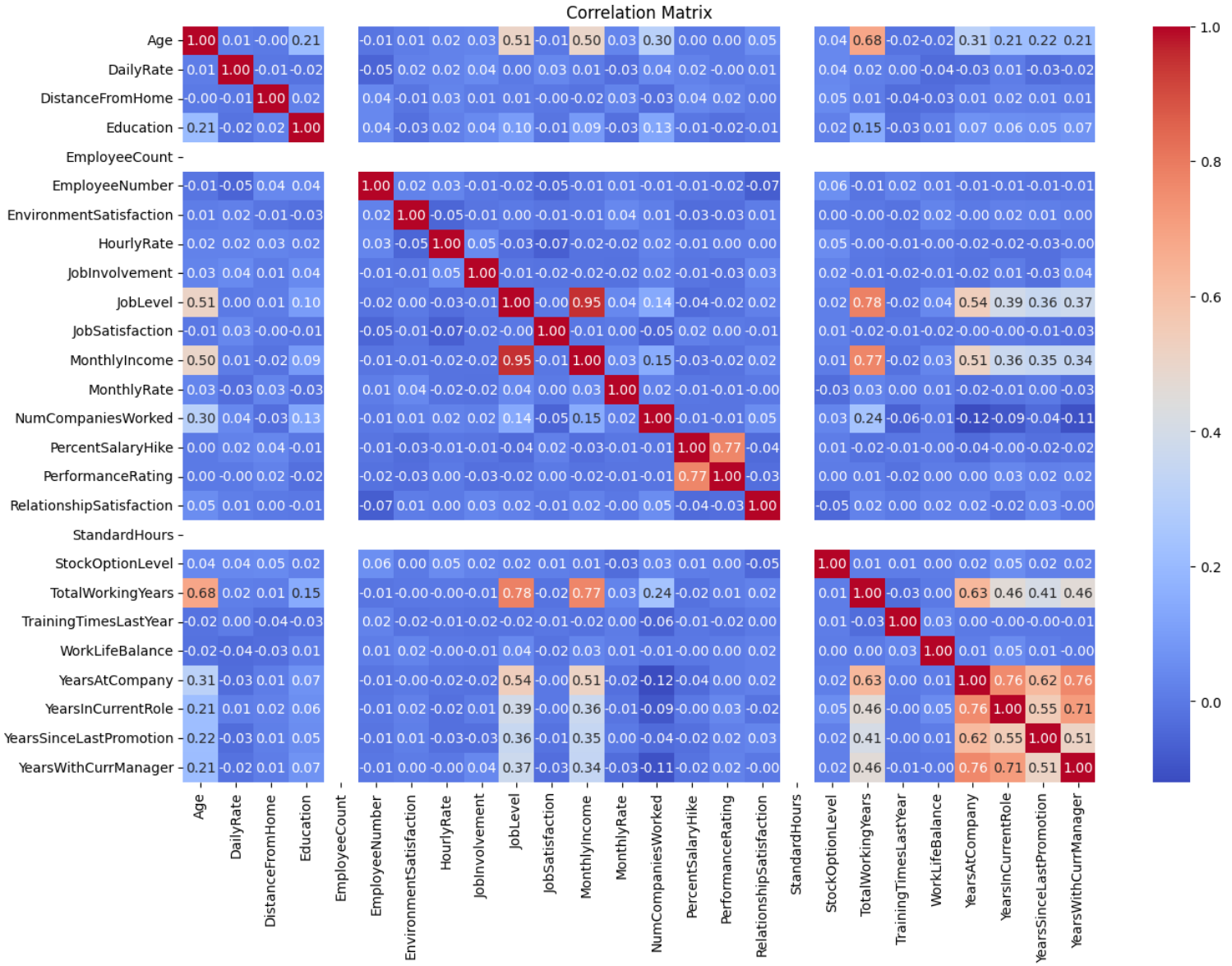
Under these categories we conducted an analysis to find valuable insights about how these 5 categories impact a person who works in a company.

To find which columns in our dataset impact these categories the most, we used a correlation matrix heatmap. With this we were able to get an initial idea about our dataset.

```r
# Load required libraries
library(reshape2)
library(ggplot2)

# Load your data from the CSV file
data <- read.csv("D:\\N learn\\2nd year\\projects\\R project\\HR_Analytics.csv")

# Check the data types
str(data)

# Remove non-numeric variables
data_numeric <- data[, sapply(data, is.numeric)]

# Compute the correlation matrix
cormat <- round(cor(data_numeric), 2)

# Melt the correlation matrix
melted_cormat <- melt(cormat)

# Create a clearer heatmap using ggplot2
ggplot(data = melted_cormat, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limits = c(-1, 1), name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, size = 10, hjust = 1)) +
  coord_fixed(ratio = 1) +
  theme(legend.position="right") +
  labs(title = "Correlation Matrix Heatmap")

# Optional: Add correlation coefficients on the heatmap
ggplot(data = melted_cormat, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limits = c(-1, 1), name = "Correlation") +
  geom_text(aes(label = value), color = "black", size = 2.5) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, size = 10, hjust = 1)) +
  coord_fixed(ratio = 1) +
  theme(legend.position="right") +
  labs(title = "Correlation Matrix Heatmap with Coefficients")

```

Figure 1 : correlation matrix heatmap source code

Correlation matrix heatmap



1.Age

Analyzing age in HR datasets provides crucial insights for understanding workforce demographics, fostering an inclusive workplace and more.

This analysis has been performed on the HR dataset we chose to identify the hidden information inside the dataset.

1.Data Visualization

In this analysis , we have used several types of plots to perform data visualizations. From this we can get hidden information on age that we cannot by looking at the dataset directly.

The graphs and the visualizations used in this analysis were accessed by using the ggplot2 libraries.

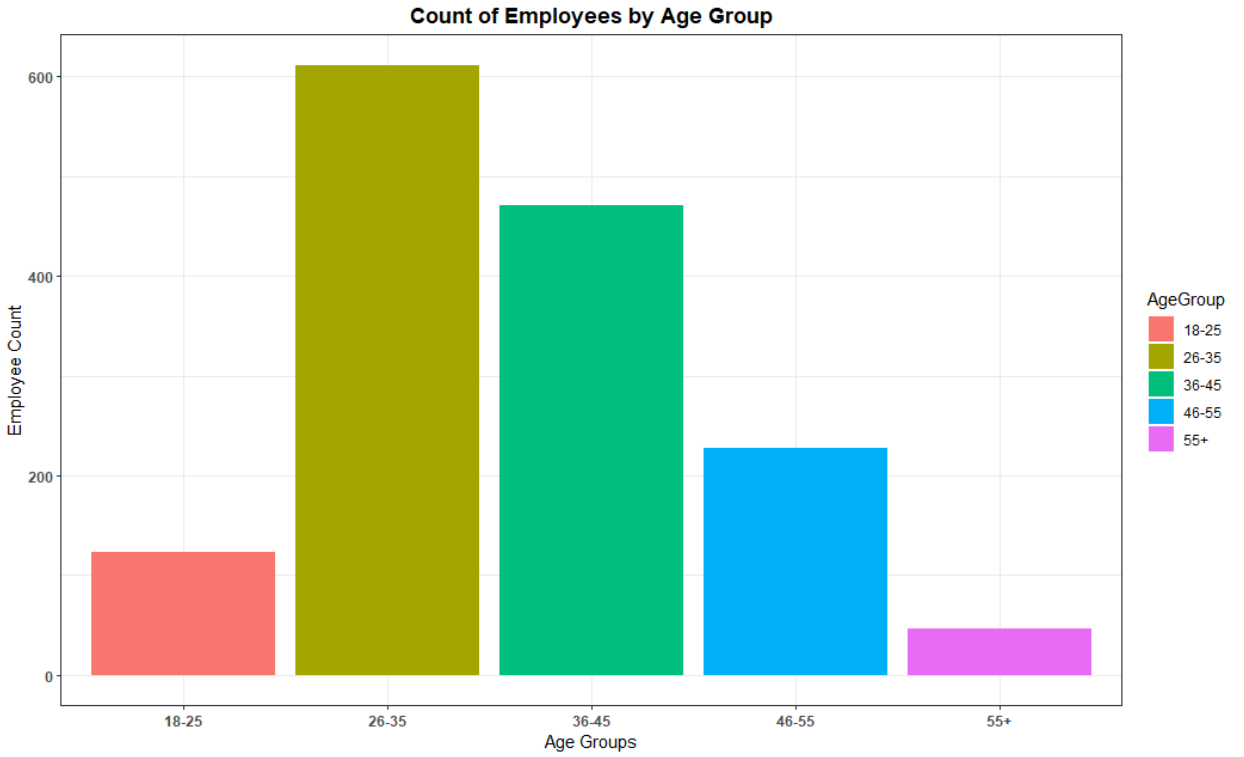
First, we need to import our data set into the R script and install the necessary libraries.

```
# Load the library
library(ggplot2)
library(dplyr)
library(tidyverse)

#loading the dataset
data <- read.csv("C:\\Users\\User\\Desktop\\R lec\\Assignment\\ProjectAge\\dataAssignment.csv")
```

After loading and installing the necessary data and the libraries we looked at our dataset and did some research using the heatmap used to get the correlation with each column of the dataset. From that we have chosen some data columns from the data set to get insights with the age.

We have used a bar plot to visualize the employee count by age group. Shown below is the R code we used to plot this bar plot.

```
#Bar plot

AgeGroup <- data$AgeGroup


ggplot(data, aes(x = AgeGroup, y = EmployeeCount , fill = AgeGroup)) +
  geom_bar(stat = "identity") +
  labs(title = "Count of Employees by Age Group", x = "Age Groups", y = "Employee Count") +
  theme_bw()+ theme(
    plot.title = element_text(face = "bold", hjust = 0.5), # Set title to bold and centered
    axis.text.x = element_text(face = "bold"), # Set x-axis labels to bold
    axis.text.y = element_text(face = "bold")
  )
```



From this plot we can say that most employees in this company are in the age group of 26-35. And this plot has a bell type shape so we can say that this company's' age distribution is a standard distribution and doing statistical calculation will give us a higher accuracy.

1. Relationship between an employee's age group, the amount of business travel they do, and how it affects their pay each month.

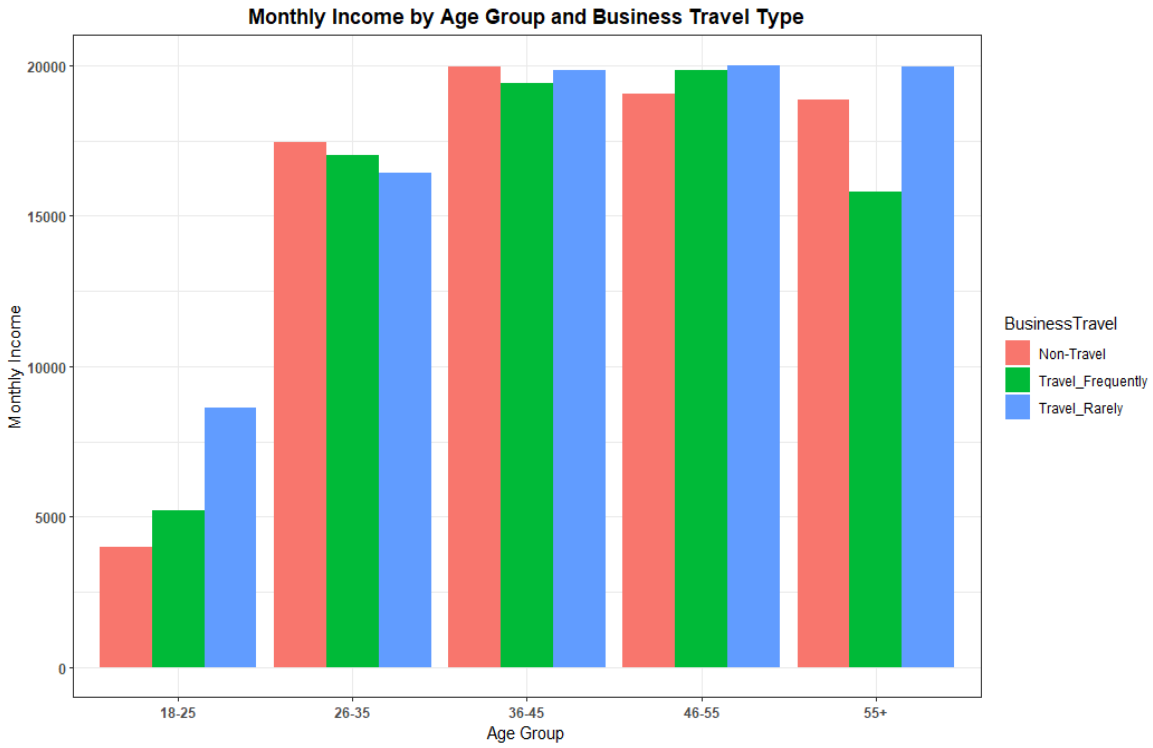
Group bar plots were used to visualize this data representation.

We observed that the dataset's business travel section had two data entries that were comparable to each other: "Travel_Rarely" and "TravelRarely." We have changed any instances of "TravelRarely" to "Travel_Rarely" to maintain consistency.

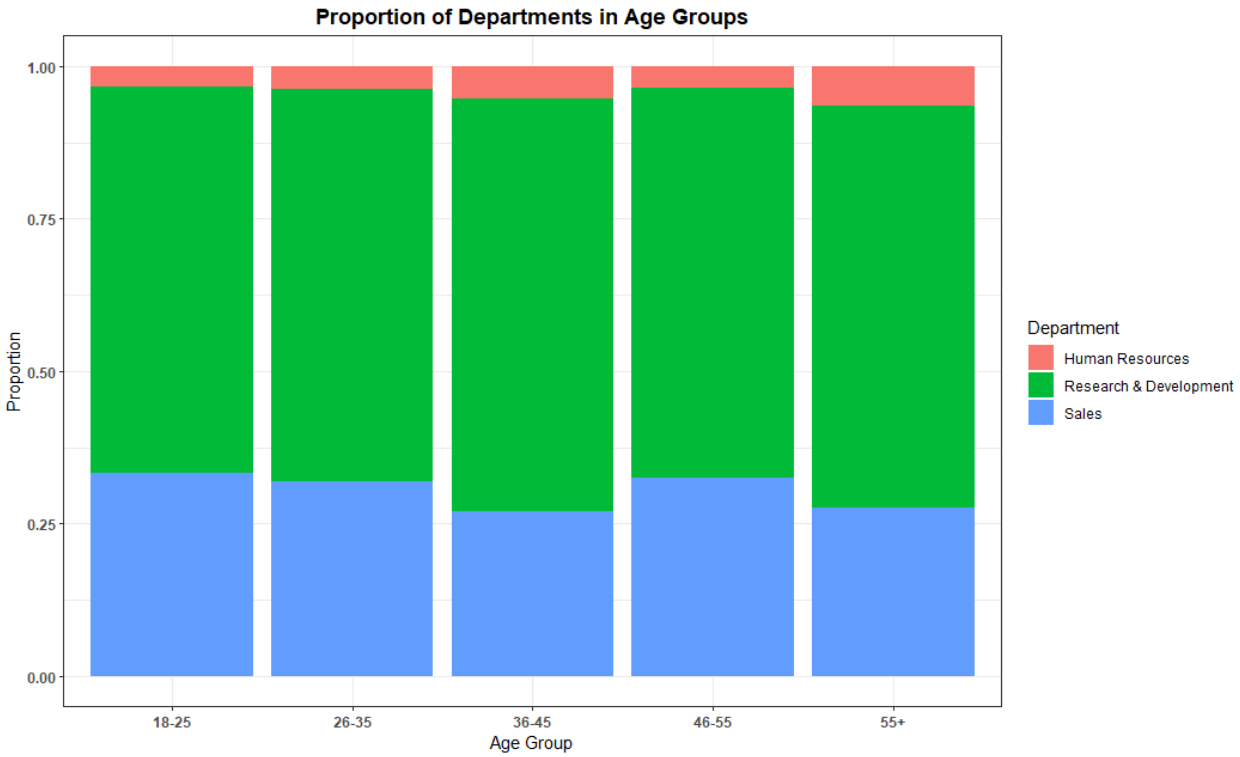
```
data$BusinessTravel[data$BusinessTravel == "TravelRarely"] <- "Travel_Rarely"
```

```r
ggplot(data, aes(x = AgeGroup, y = MonthlyIncome, fill = BusinessTravel)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Monthly Income by Age Group and Business Travel Type",
       x = "Age Group", y = "Monthly Income") +
  theme_bw() + theme(
    plot.title = element_text(face = "bold", hjust = 0.5), # Set title to bold and centered
    axis.text.x = element_text(face = "bold"), # Set x-axis labels to bold
    axis.text.y = element_text(face = "bold")
  )
```

Shown below is the plot we generated from the R code shown above.



2. Proportion of Departments in Age Group.



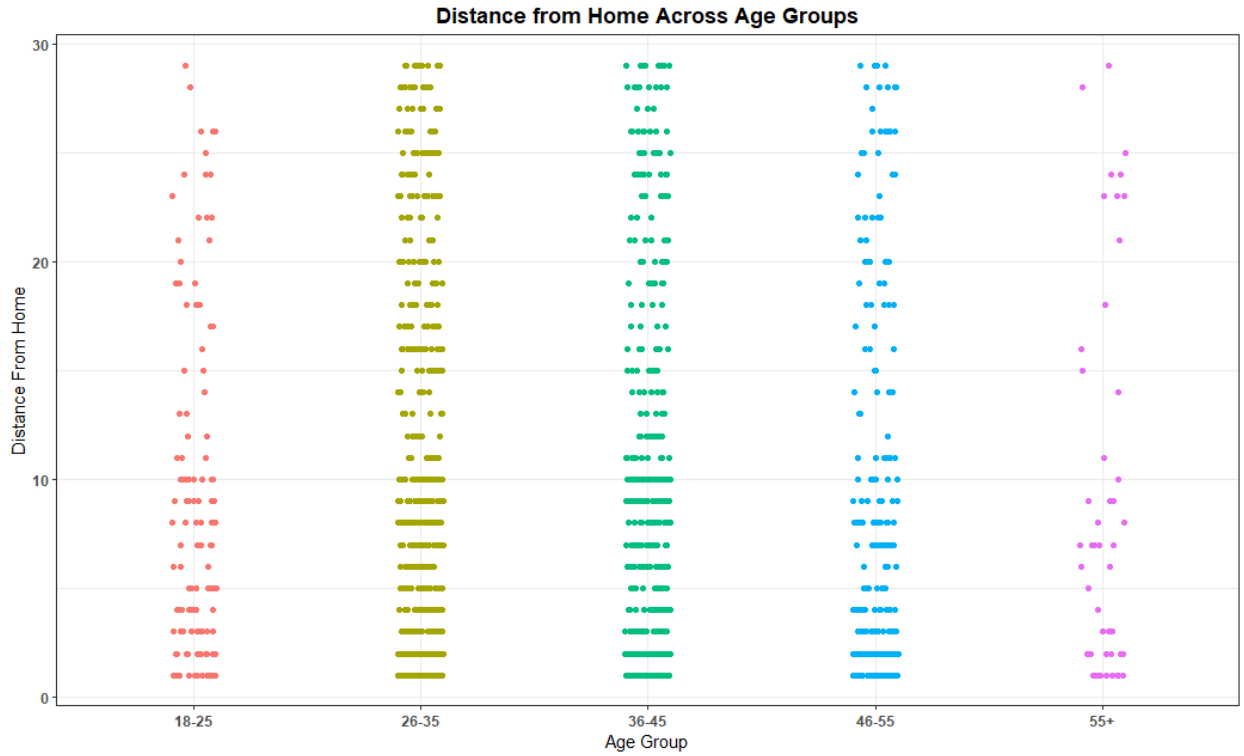
With this mosaic plot we can get a better understanding of how the employees in this company have been distributed according to the departments they work in.

So, from this we can say that this company's focus is on research and development.

```r
#Mosaic Plot

Department <- data$Department

ggplot(data, aes(x = data$AgeGroup, fill = Department)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Departments in Age Groups", x = "Age Group", y = "Proportion") +
  theme_bw()+ theme(
    plot.title = element_text(face = "bold", hjust = 0.5), # Set title to bold and centered
    axis.text.x = element_text(face = "bold"), # Set x-axis labels to bold
    axis.text.y = element_text(face = "bold")
  )
```

Shown above is the R code used to generate the mosaic plot.

3. Distance from home Across Age Groups

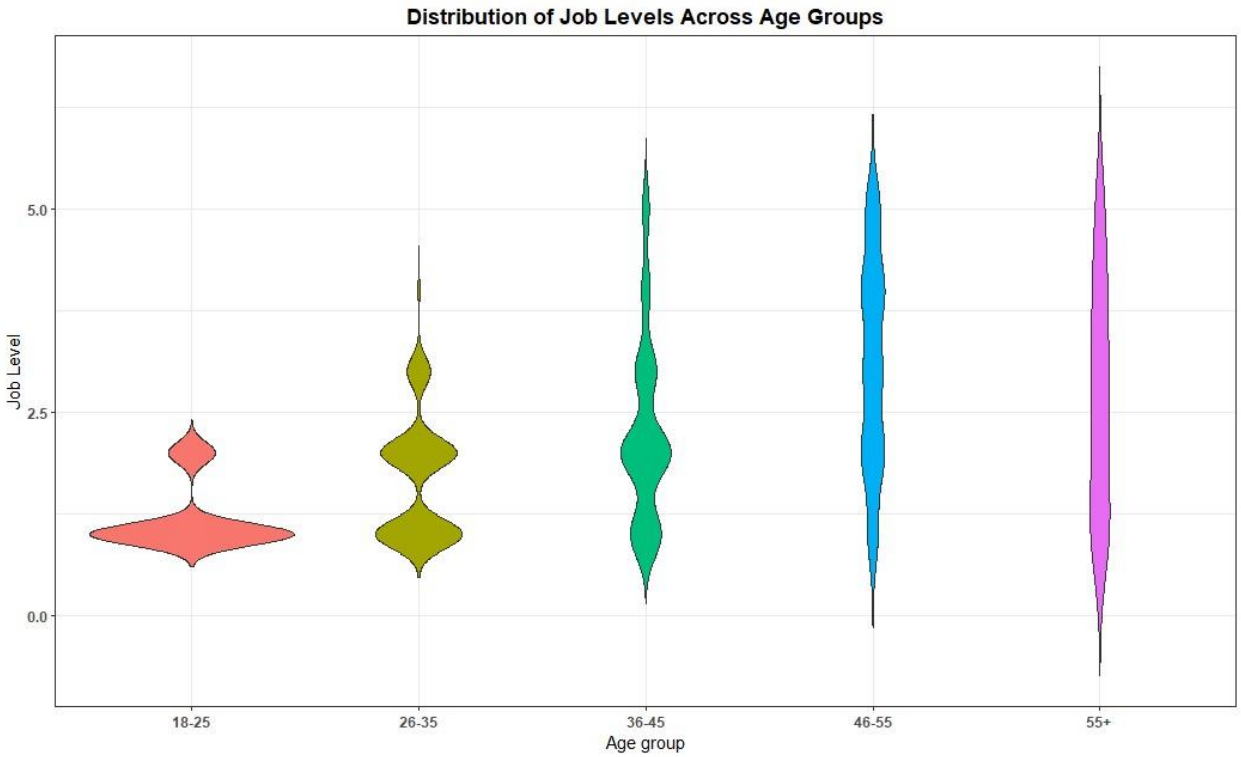


From this plot we can see almost all the employees of this company live near to the company premises.

Given below is the R code used to generate this plot.

```r
# swarm-like plot

ggplot(data, aes(x = data$AgeGroup, y = data$DistanceFromHome , color = data$AgeGroup)) +
  geom_point(position = position_jitterdodge(jitter.width = 0.2, dodge.width = 0.5)) +
  labs(title = "Distance from Home Across Age Groups", x = "Age Group", y = "Distance From Home") +
  scale_color_discrete(guide = FALSE) + # Removing the legend
  theme_bw()+ theme(
    plot.title = element_text(face = "bold", hjust = 0.5), # Set title to bold and centered
    axis.text.x = element_text(face = "bold"), # Set x-axis labels to bold
    axis.text.y = element_text(face = "bold")
  )
```

4. Distribution of Job Levels Across Age Group



This is a violin plot representing the distribution of the job level across age groups. As you can see from this plot the job level gradually increases as the age increases.

Given below is the r code used to generate this plot.

```r
#Violin Plot

ggplot(data, aes(x = data$AgeGroup, y = data$JobLevel, fill = data$AgeGroup)) +
  geom_violin(trim = FALSE) +
  labs(title = "Distribution of Job Levels Across Age Groups", x = "Age group", y = "Job Level") +
  scale_fill_discrete(guide = FALSE) + # Removing the legend
  theme_bw() + theme(
    plot.title = element_text(face = "bold", hjust = 0.5), # Set title to bold and centered
    axis.text.x = element_text(face = "bold"), # Set x-axis labels to bold
    axis.text.y = element_text(face = "bold")
  )
```

2.Data analytics through models

1.Random Forrest Classification on Age group

Using a Random Forest model to forecast age groups in an HR dataset can help create a more inclusive and productive work environment, provide insightful information about the demographics of the workforce, and assist in developing focused HR campaigns.

```
library(randomForest)

data1 <- read.csv("C:\\Users\\User\\Desktop\\R lec\\Assignment\\ProjectAge\\dataAssignment.csv")
```

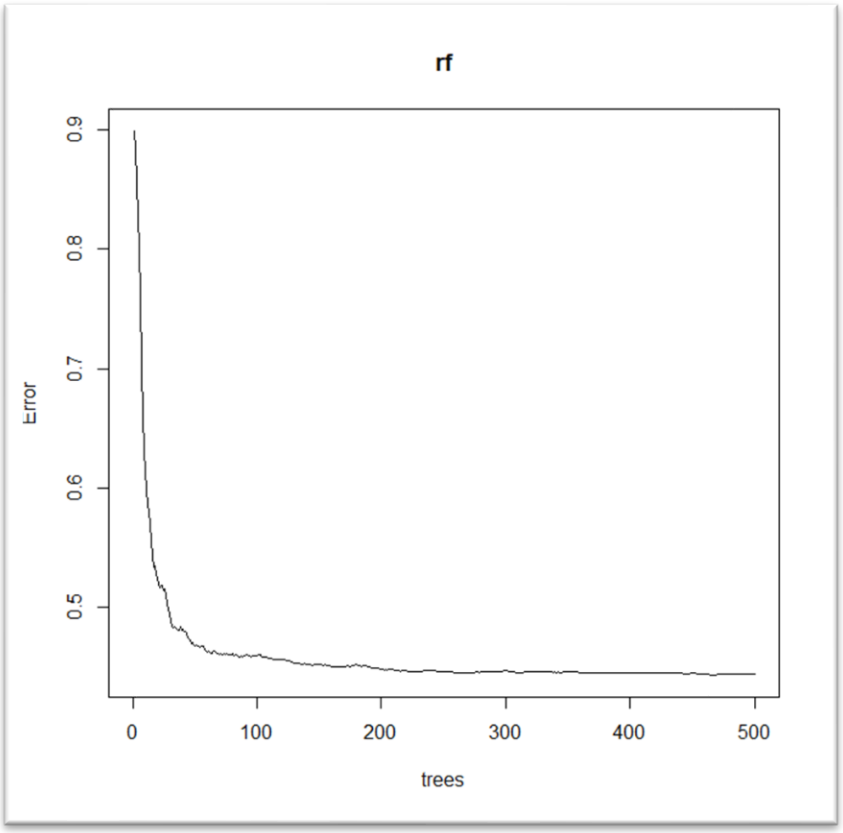
```
data1$AgeGroup <-as.integer(factor(data1$AgeGroup))
data1$Attrition <-as.integer(factor(data1$Attrition))
data1$BusinessTravel <-as.integer(factor(data1$BusinessTravel))
data1$Department <-as.integer(factor(data1$Department))
data1$EducationField <-as.integer(factor(data1$EducationField))
data1$Gender <-as.integer(factor(data1$Gender))
data1$JobRole <-as.integer(factor(data1$JobRole))
data1$MaritalStatus <-as.integer(factor(data1$MaritalStatus))
data1$SalarySlab <-as.integer(factor(data1$SalarySlab))
data1$Over18 <-as.integer(factor(data1$Over18))
data1$OverTime <-as.integer(factor(data1$OverTime))

str(data1)

data2 <- data1[, !names(data1) %in% c("EmpID")]
data2 <- data1[, !names(data1) %in% c("Age")]

sum(is.na(data2))
data2 <- na.omit(data2)
sum(is.na(data2))
```

After loading the needed libraries and dataset data cleaning and preprocessing was done.

```
# Check the data type of AgeGroup
class(train$AgeGroup)

# Check unique levels and their counts
table(train$AgeGroup)

# Check for non-numeric values in AgeGroup
non_numeric_AgeGroup <- train$AgeGroup[!is.numeric(train$AgeGroup)]
non_numeric_AgeGroup
str(data2)

plot(rf)
```

```
set.seed(222)
ind <- sample(2, nrow(data2), replace = TRUE, prob = c(0.7, 0.3))
train <- data2[ind==1,]
test <- data2[ind==2,]

rf <- randomForest(AgeGroup~., data=train, proximity=TRUE)
print(rf)
randomForest(formula = data2$AgeGroup ~ ., data = train)
```

With this we can plot an Error Rate Plot.



By evaluating this error rates, we can assess the overall performance and stability of the random tree. From the above tree the error rate has decreased meaning the model has been stabilized after 300 trees.

2.Naïve Bayes Classification on Age Group.

The Bayes theorem, which determines the probability of a label (class) given the observed features, is the foundation of naive Bayes classifiers.

```
data1 <- read.csv("C:\\Users\\User\\Desktop\\R Tec\\Assignment\\ProjectAge\\dataAssignment.csv")

str(data1)

data1$AgeGroup <-as.integer(factor(data1$AgeGroup))
data1$Attrition <-as.integer(factor(data1$Attrition))
data1$BusinessTravel <-as.integer(factor(data1$BusinessTravel))
data1$Department <-as.integer(factor(data1$Department))
data1$EducationField <-as.integer(factor(data1$EducationField))
data1$Gender <-as.integer(factor(data1$Gender))
data1$JobRole <-as.integer(factor(data1$JobRole))
data1$MaritalStatus <-as.integer(factor(data1$MaritalStatus))
data1$SalarySlab <-as.integer(factor(data1$SalarySlab))
data1$Over18 <-as.integer(factor(data1$Over18))
data1$OverTime <-as.integer(factor(data1$OverTime))

str(data1)


sum(is.na(data1))
data1 <- na.omit(data1)
sum(is.na(data1))

str(data1)
```

```
library(caTools)

set.seed(1000)
split <- sample.split(data1$AgeGroup, SplitRatio = 0.75)
train <- subset(data1, split == TRUE)
test  <- subset(data1, split == FALSE)

library(e1071)

classifier <- naiveBayes ( AgeGroup~., data = train)
pre <- predict(classifier, newdata = test)

contable <- table(pre,test$AgeGroup)
contable

library(caret)

confusionMatrix(contable)
```

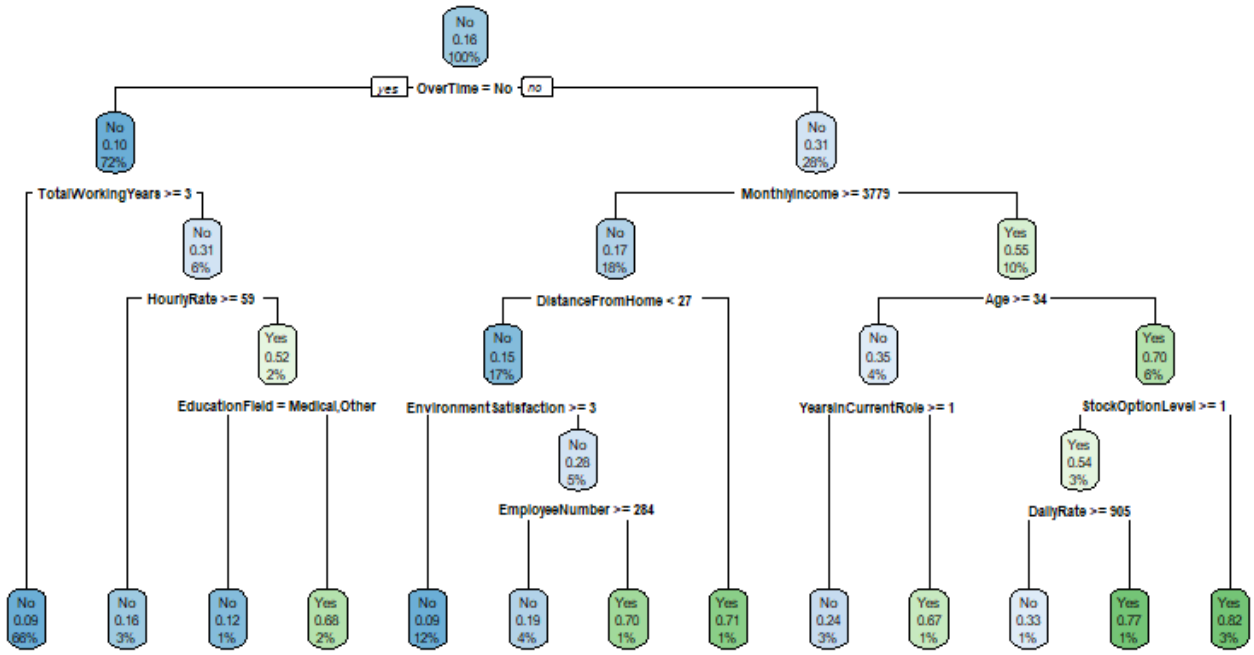
```
data1 <- data1[, !names(data1) %in% c("Age")]
```

In this classification model we were able to classify the age group into groups with an accuracy of 0.7781.

```
> confusionMatrix(contable)
Confusion Matrix and Statistics


pre   1   2   3   4   5
  1  25  18   0   0   0
  2   4 118  21   0   0
  3   0  12  78   6   0
  4   0   0  13  50   5
  5   0   0   0   0   6

Overall Statistics

               Accuracy : 0.7781
                 95% CI : (0.7313, 0.8202)
    No Information Rate : 0.4157
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6865

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.86207   0.7973   0.6964   0.8929  0.54545
Specificity           0.94495   0.8798   0.9262   0.9400  1.00000
Pos Pred Value        0.58140   0.8252   0.8125   0.7353  1.00000
Neg Pred Value        0.98722   0.8592   0.8692   0.9792  0.98571
Prevalence            0.08146   0.4157   0.3146   0.1573  0.03090
Detection Rate        0.07022   0.3315   0.2191   0.1404  0.01685
Detection Prevalence  0.12079   0.4017   0.2697   0.1910  0.01685
Balanced Accuracy     0.90351   0.8386   0.8113   0.9164  0.77273
```

3.Decision tree classification on Attrition

Decision tree classification is a useful tool in many applications because it provides comprehensible results and a plethora of insights into data patterns, feature relevance, and predictive capabilities.

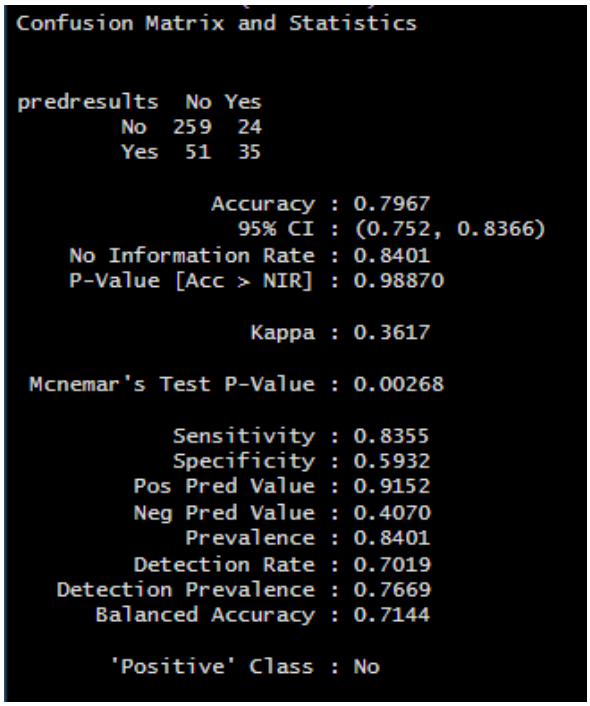


After evaluating this model though, a confusion matrix has an accuracy of 0.7967.

```
library(caret)
library(rpart.plot)
library(tidyverse)                #workflow
library(skimr)
library(caTools)


#setting dataset

dataset <- read.csv("C:\\Users\\User\\Desktop\\R lec\\Assignment\\ProjectAge\\dataAssignment.csv")
str(dataset)

dataset <- dataset[, !names(dataset) %in% c("EmpID")]
dataset <- dataset[, !names(dataset) %in% c("JobRole")] # by this we can  increase the accuracy of this analysis
```

```
split <- createDataPartition(y=dataset$Attrition , p = 0.75, list = FALSE)
train <- dataset[split, ]
test <- dataset[-split, ]

#split <- sample.split(dataset$Purchase, SplitRatio = 0.75)
#train <- subset(dataset, split == TRUE)
#test <- subset(dataset, split == FALSE)

dim(train)
dim(test)

#the classifier
set.seed(150)

dec_tree <- rpart(formula = Attrition   ~.,
                  data = train,
                  method = "class",
                  xval = 10)

# drawing the tree
rpart.plot(dec_tree, yesno = TRUE)
```

```
library(e1071)

naiveclassifier <- naiveBayes(Attrition   ~ ., data=train)
predresults <- predict(naiveclassifier , newdata = test)

#the confusion matrix

conftable <- table(predresults, test$Attrition  )
conftable

confusionMatrix(conftable)
```



In this model removing the 'Joblevel' from the dataset gives us an increased accuracy for this dataset

3.Findings and Discussion

From the employee count bar plot we can derive that most of the employees work in this are in the age group 26-35 or above so when the company give benefits to the employees' customization of benefits and programs that cater to the specific needs according to the nature of the employees will more appreciated. As an example, giving more weight to retirement planning or healthcare benefits will be more appropriate because most of the employees are in the older age groups.

From the plot that shows how the distance from home across the age group we can say that most of the employees who works in this company have a distance between 10 – 20 Kms. So, arranging a daily commute for them will be more beneficial because it will increase employee satisfaction, enhance productivity, improve punctuality and the attendance of the employees.

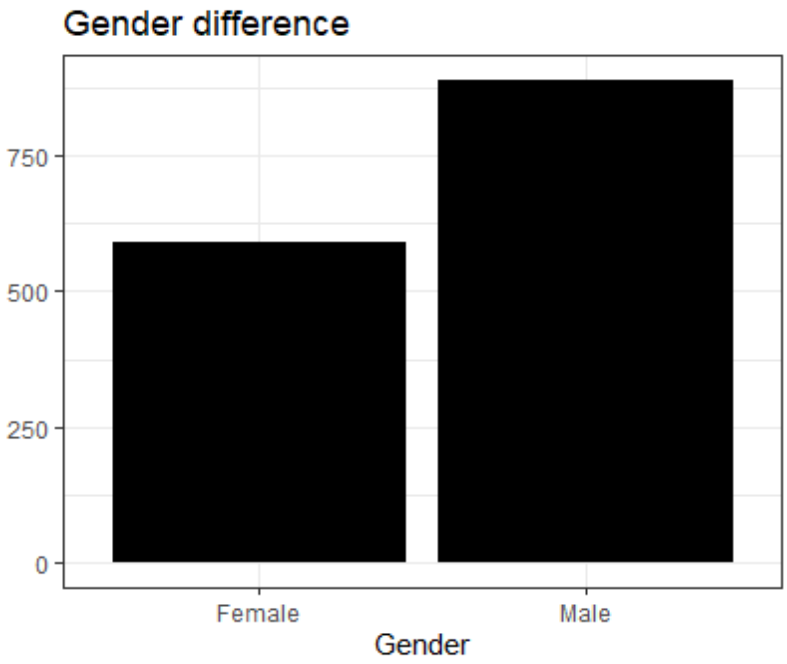
When we look at the violine plot demonstrates distribution of job levels across age group we can say that the youngest and oldest age groups have the widest distribution in job levels. As a result, there is a greater variety of jobs among these age groups, with some members of these groups having extremely high or extremely low job levels.

In conclusion, a thorough examination of the dataset's age-related features was made possible by the combination of modeling and data visualization techniques, which provided useful insights and laid an outline for future in-depth research in the relevant field.

2. Gender

1.Data Visualizations

1. Gender representation of the company

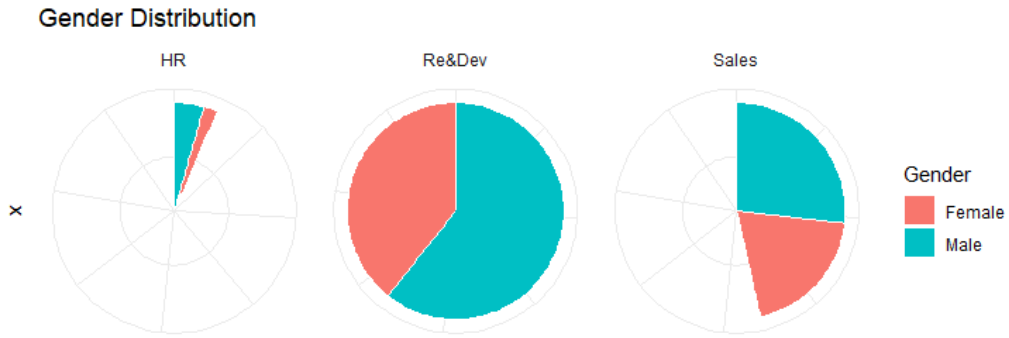


This bar chart represents the gender difference representation of the above HR data set. According to this bar chart clearly, we can see there are males working in this company More than female employees.

Code:

```
library(dplyr)
library(ggplot2)
library(tidyverse)



data <- read.csv("C:\\Users\\NMC\\Downloads\\archive (4)\\HR_Analytics.csv")


print(data)



ggplot(data,aes(x=Gender))+
  geom_bar(fill="black")+

theme_bw()+
labs(x="Gender",
     y='',
     title="Gender difference")
```

2. Gender distribution among departments



HR-Human Resources Department

Re & Dev- Research and Development Department

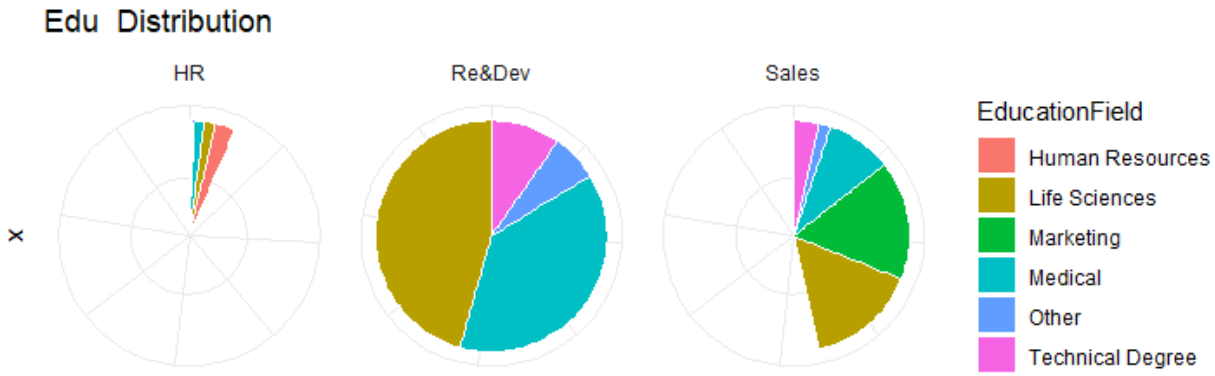
Sales-Sales Department

This is a pie chart crated using "Gender" column and department column . There are three different departments inside the above company . According to the data set all the employees divided in to these three departments. The least number of employees are working in "HR" department. The second largest number of employees are working in "sales department". And the largest number of employees are working in research and development department. In this each and every department we can see the number of male employees is higher than the number of female employees is working in.

Code:

```
Hr<-data
Hr
Hr %>%
  filter(DailyRate>1000) %>%
  select(EducationField,DailyRate) %>%
  arrange(DailyRate)->edu_rate

print(edu_rate)
  library(dplyr)

Hr <- data

Hr %>%
  select(Department, EducationField,Gender) %>%
  mutate(Department=recode(Department,
                           "Human Resources"="HR",
                           "Research & Development"="Re&Dev")) %>%
  arrange(Department)->gender_edu

print(gender_edu)
tail(gender_edu)
gender_edu %>%
  group_by(Department, Gender) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = "", y = count, fill = Gender)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  facet_wrap(~Department) +
  coord_polar("y") +
  ggtitle("Gender Distribution ") +
  theme_minimal() +
  theme(axis.text.x=element_blank(),
        axis.title.x=element_blank(),
        axis.ticks.x=element_blank())
```

3. Education distribution among Department



HR-Human Resources Department

Re&Dev- Research and Development Department

Sale – Sale Department

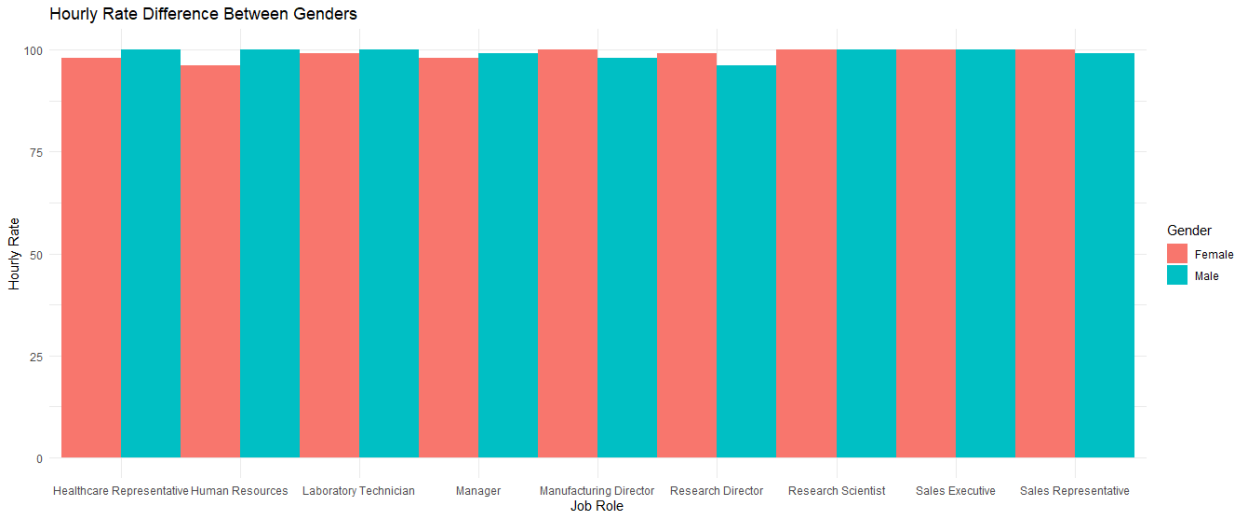
This above three pie charts are created using "Department column" and "Educational field column". Using these three plots we and understand the diversity of education fields among departments. According to these three pie charts " sales department" shows the highest diversity.

And "HR Department" shows the lowest diversity .In this company the highest number of employees are from "Life Science" stream . The lowest number is the employees who followed other education streams. According to this data we can see there was a high demand for Life science stream employees.

Code:

```
gender_edu %>%
  group_by(Department, EducationField) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = "", y = count, fill = EducationField)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  facet_wrap(~Department) +
  coord_polar("y") +
  ggtitle("Edu  Distribution ") +
  theme_minimal() +
  theme(axis.text.x=element_blank(),
        axis.title.x=element_blank(),
        axis.ticks.x=element_blank())
```

4. Salary difference between gender

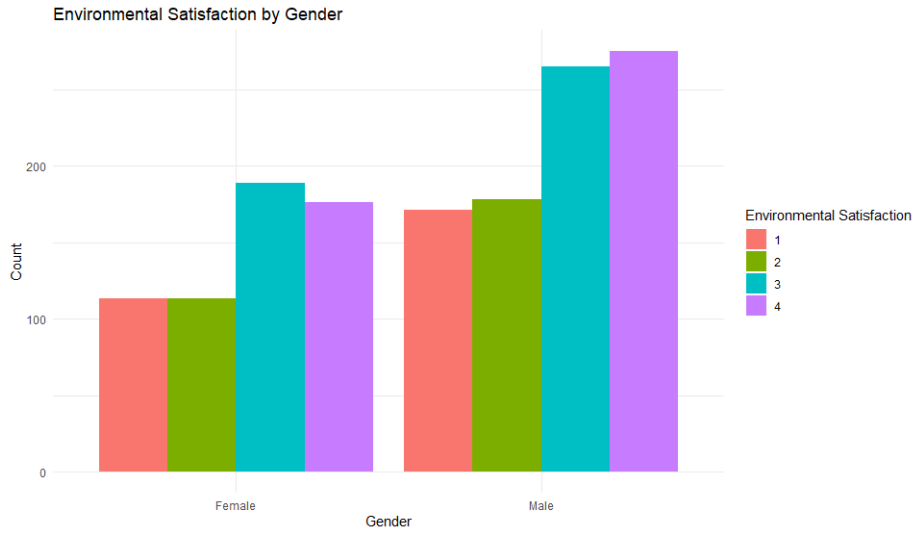


This bar chart was plotted using "Hourly Rate", "Gender" and "job role ".According to this bar chart we can see there are no big salary gap between "Genders" in different job roles. This Is a good sign about this company for the female employees who willing to join this company.

Code:

```
Hr<-data
Hr %>%
  select(HourlyRate,Gender,JobRole) %>%

  arrange(Gender)->hourly_rate_gender

print(hourly_rate_gender)

print(ggplot(hourly_rate_gender, aes(x = JobRole, y = HourlyRate, fill = Gender)) +
        geom_bar(stat = "identity", position = "dodge", width = 1.0) +
        labs(title = "Hourly Rate Difference Between Genders ",
             x = "Job Role",
             y = "Hourly Rate") +
        theme_minimal())
```

5. Environment Satisfaction between gender



1-Poor

2-Good

3-Great

4-Exelent

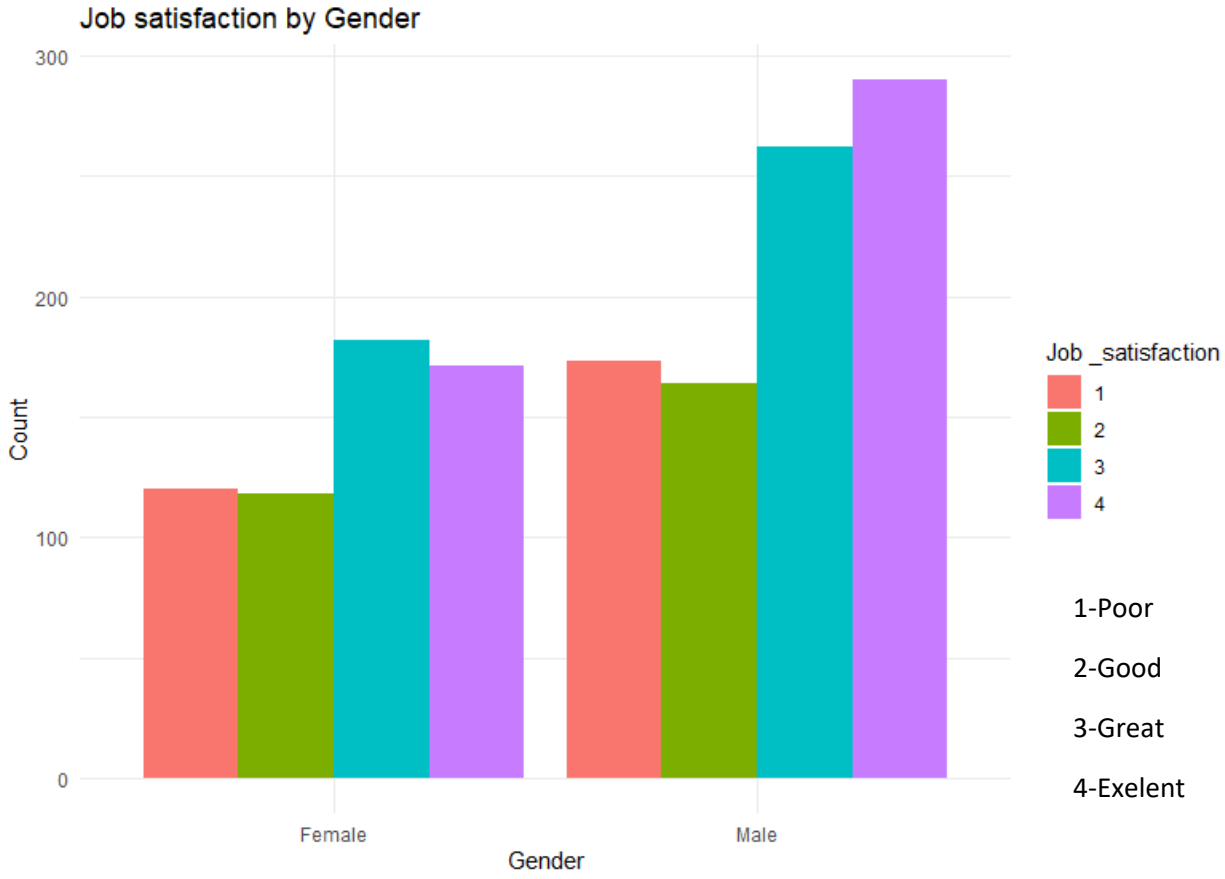
This above bar chart was plotted to visualize the environment satisfaction between gender. According to this bar chart overall we can identify they is an excellent environment satisfaction about the company among males. The environment satisfaction of females is great but comparing to males it is low .And we can identify the is an environment issue among females.

As a suggestion this company needs to maintain their environment as a female friendly environment .

Code:

```
#  evironment satisfaction

Hr<-data
Hr %>%
  select(Gender,EnvironmentSatisfaction) %>%
  arrange(Gender)->environment_satisfaction

print(environment_satisfaction)

ggplot(environment_satisfaction, aes(x = Gender, fill = factor(EnvironmentSatisfaction))) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Environmental Satisfaction by Gender",
       x = "Gender",
       y = "Count",
       fill = "Environmental Satisfaction") +
  theme_minimal()
```

6. Job Satisfaction between gender

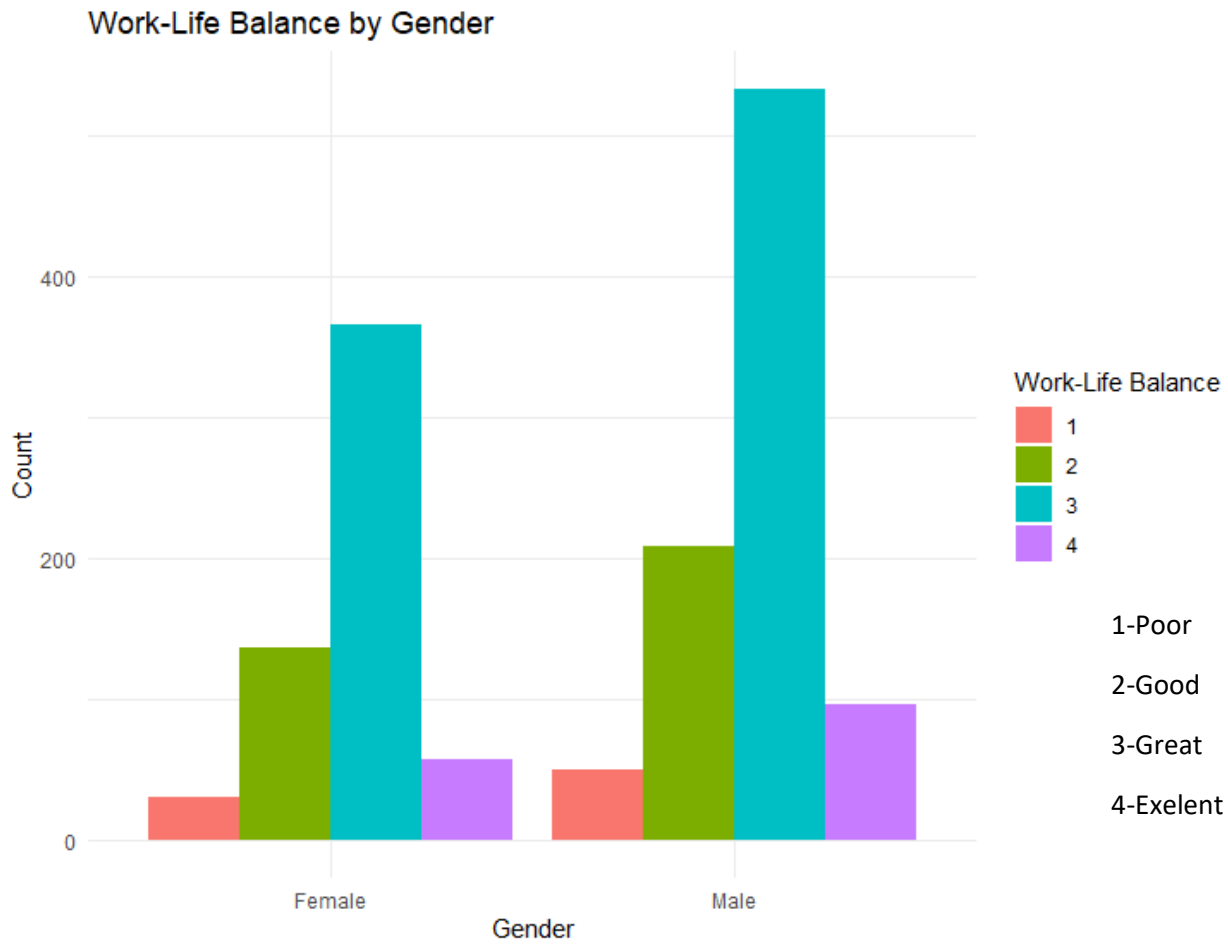


This bar chart was plotted using “Gender ”and “Job satisfaction ” columns . According to this bar chart we can identify the overall job satisfaction of the male is excellent .But the overall average job satisfaction of females is Good but comparing to males it is low .The environment satisfaction from above bar chart there were also an issue from females about their environment satisfaction . As a company this company needs to find a quick solution for this problem.

Code:

```
#Job_satisfacgion
Hr<-data
Hr %>%
  select(Gender,JobSatisfaction) %>%
  arrange(Gender)->Job_satisfacgion

print(Job_satisfacgion)

ggplot(Job_satisfacgion, aes(x = Gender, fill = factor(JobSatisfaction))) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Job satisfaction by Gender",
       x = "Gender",
       y = "Count",
       fill = "Job _satisfaction") +
  theme_minimal()
```

7. <u>Work and life balance between gender</u>

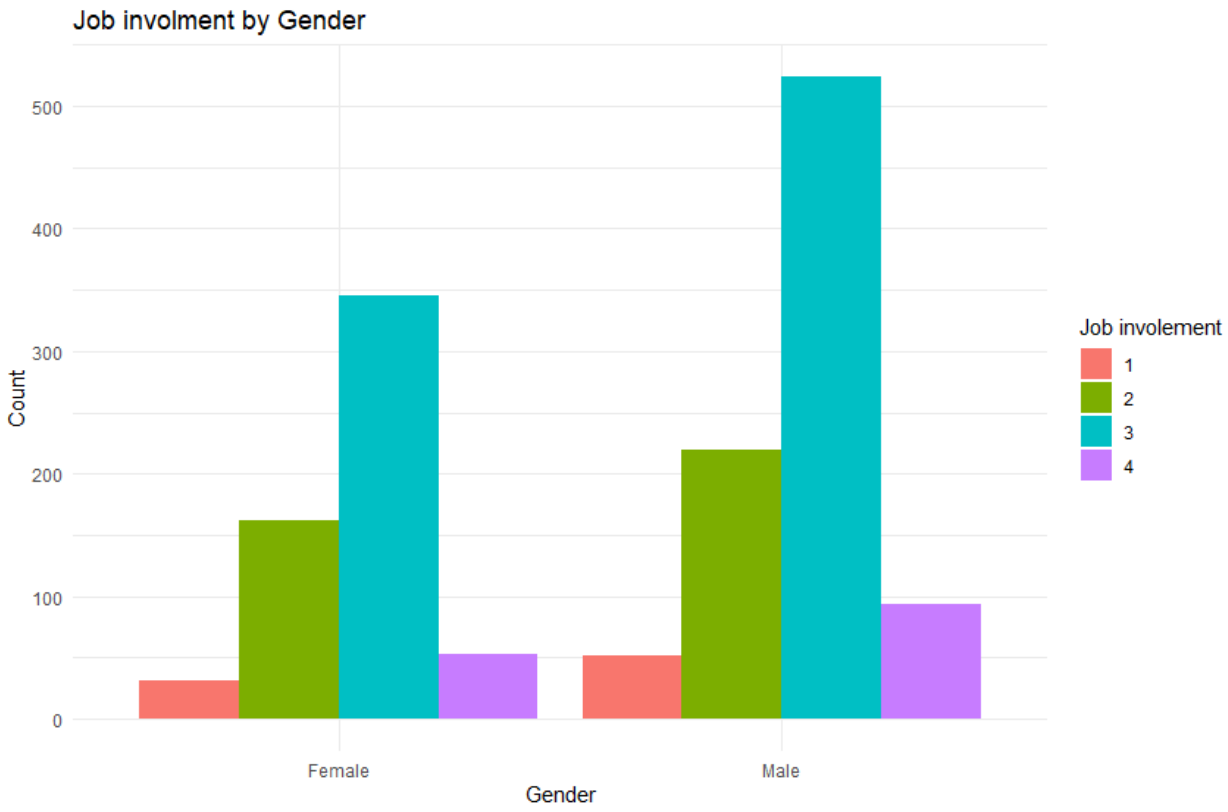


According to this bar chart we can clearly see that there is a Great work and life balance between their employees both males and females have very good work and life balance .The number of employees who have poor work and life balance between both male and female employees are very low . And this is a big green light for the company and company health.

Code:

```
#work life balance

Hr <- data
Hr %>%
  select(Gender, WorkLifeBalance) %>%
  arrange(Gender) -> work_lifebalance

print(work_lifebalance)

ggplot(work_lifebalance, aes(x = Gender, fill = factor(WorkLifeBalance))) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Work-Life Balance by Gender",
       x = "Gender",
       y = "Count",
       fill = "Work-Life Balance") +
  theme_minimal()
```

8. Job involvement between gender

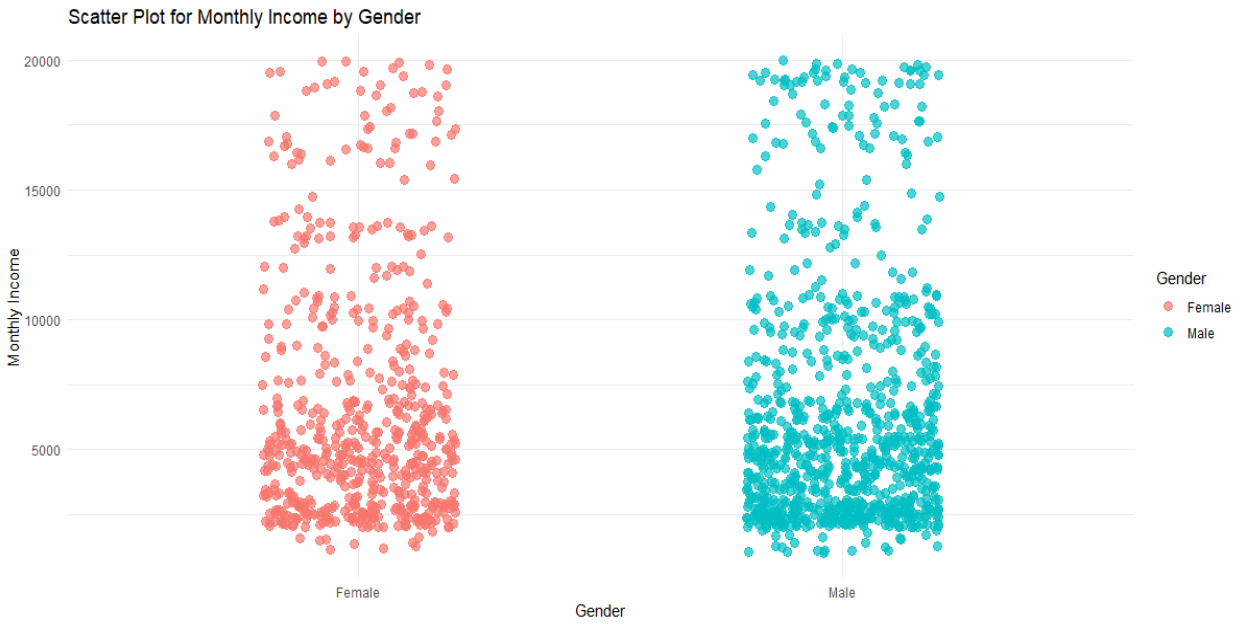


According to this bar chart we can see the job involvement of both male and female is good. And as an average It also very satisfying measure . This is a good signal for a company . And the number of poor jobs satisfied people also very low number .

Code:

```
#job enolement

Hr <- data
Hr %>%
  select(Gender, JobInvolvement) %>%
  arrange(Gender) -> Job_involement

print(Job_involement)

ggplot(Job_involement, aes(x = Gender, fill = factor(JobInvolvement))) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Job involment by Gender",
       x = "Gender",
       y = "Count",
       fill = "Job involement") +
  theme_minimal()
```

9. Salary difference between gender



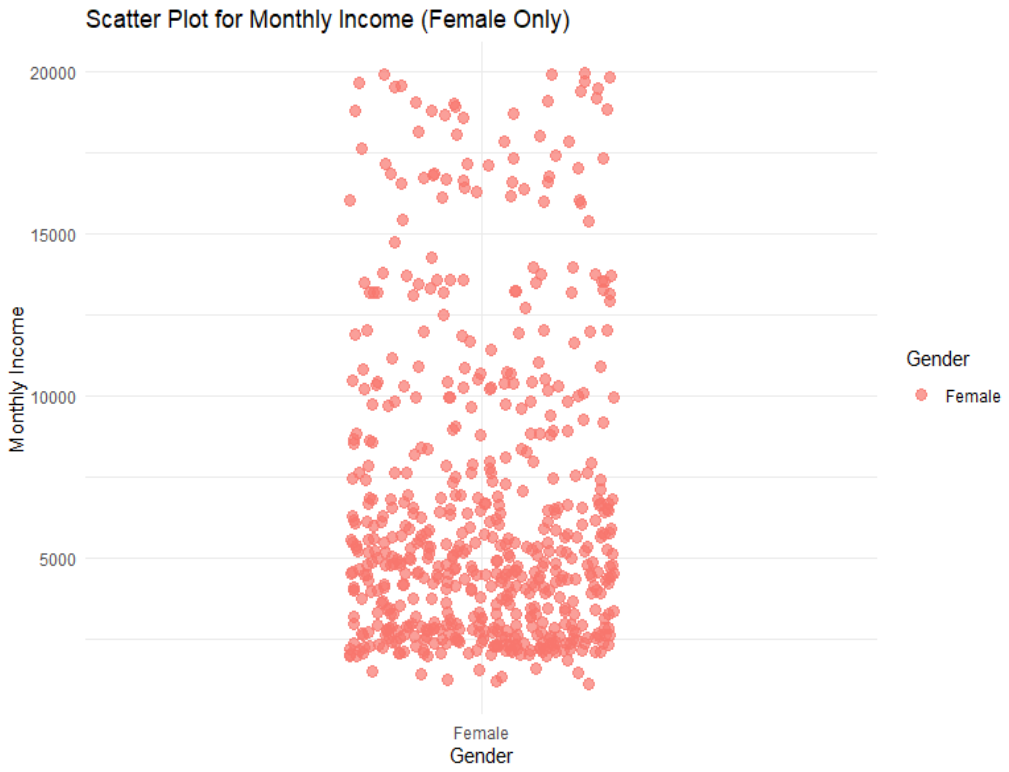
According to this scatter plot we can see most employees having salary among 5000 to 6000. The number of female employees having more than that salary is lower than the number of male employees having more than 6000 salaries. The number of male employees who are having near 20000 salary is more than the number of female employees having that salary. According to this bar chart we can identify the number of male employees who having a top post I the company are greater than the number of female employees who have top posts.

Code:

```
library(ggplot2)


ggplot(Monthly_inome, aes(x = Gender, y = MonthlyIncome, color = Gender)) +
  geom_point(position = position_jitter(width = 0.2, height = 0), size = 3, alpha = 0.7) +
  labs(title = "Scatter Plot for Monthly Income by Gender",
       x = "Gender",
       y = "Monthly Income",
       color = "Gender") +
  theme_minimal()
```

Female:

With the below scatter plot we can see the largest number of employees are getting salary among 2500$ to 7500$ the number of female employees who are getting a salary greater than 7500$ s very less when compared to the majority .And the number of female employees who are getting top range salary like 20000$ are very less when it compared to the number of male employees who are getting that salary. According to this chart we can identify there is a lack of female representation in top job positions . In the above bar charts, there were also a lack of job involvement among female employees when it compared to the male employees .And as a solution for that I can suggest the company needs to encourage female employees for a positive job involvement and need to create a positive environment and job satisfaction among them.



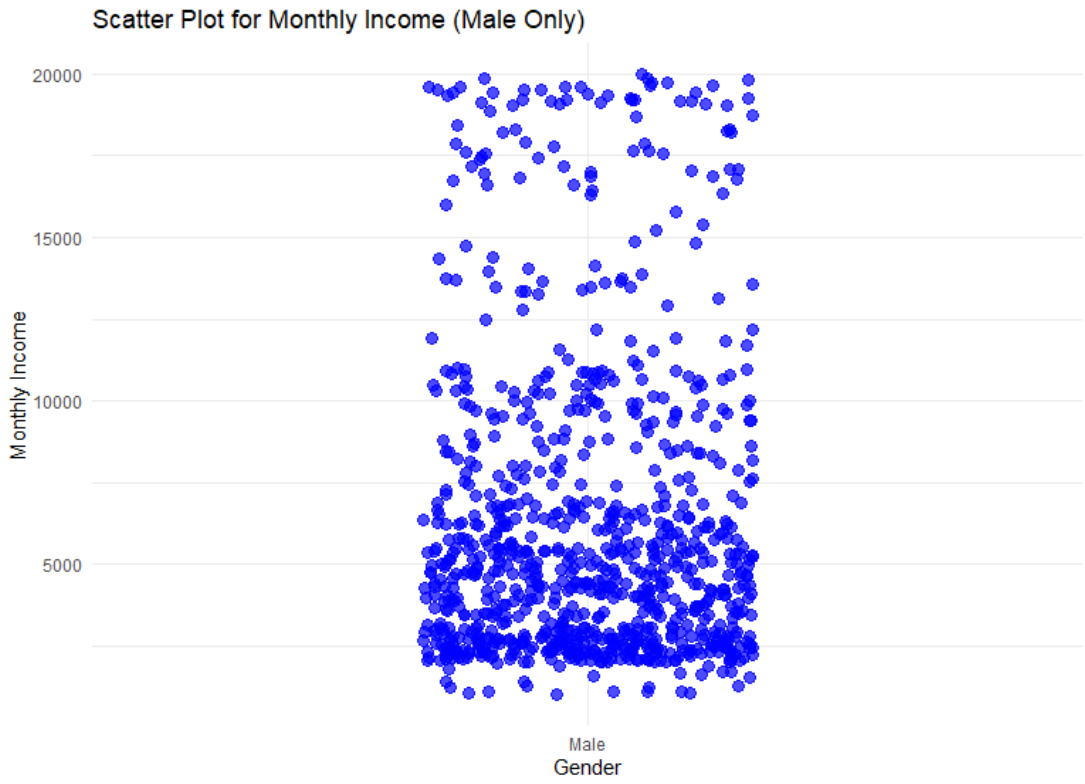
Code:

```
library(ggplot2)
library(dplyr)


female_data <- Monthly_inome %>% filter(Gender == "Female")

ggplot(female_data, aes(x = Gender, y = MonthlyIncome, color = Gender)) +
  geom_point(position = position_jitter(width = 0.2, height = 0), size = 3, alpha = 0.7) +
  labs(title = "Scatter Plot for Monthly Income (Female Only)",
       x = "Gender",
       y = "Monthly Income",
       color = "Gender") +
  theme_minimal()
```

Male:

The below scatter plot shows the salary distribution of female employees .Majority of female employees are getting among 2500$ to 5000$ .And in there are a greater number of male employees who are getting to range salary when it compared to the female employees.

Also, in above charts there was a good job involvement ,good job satisfaction and good environment satisfaction among male employees .According to these data we can decide male employees are more productive than male employees .



Code:

```
male_data <- Monthly_inome %>% filter(Gender == "Male")

ggplot(male_data, aes(x = Gender, y = MonthlyIncome)) +
  geom_point(position = position_jitter(width = 0.2, height = 0), size = 3, alpha = 0.7, color = "blue") +
  labs(title = "Scatter Plot for Monthly Income (Males Only)",
       x = "Gender",
       y = "Monthly Income") +
  theme_minimal()
```

2.Data analytics through models

1.Implementing Navi bayes classification to Gender

The expected outcomes of this model are correctly categorizing workers as male or female, find the important variables that have a significant impact on gender prediction, providing information about trends in HR and possible prejudices. And provide insight for HR choices and plans pertaining to personnel management, recruiting, inclusion, and diversity.

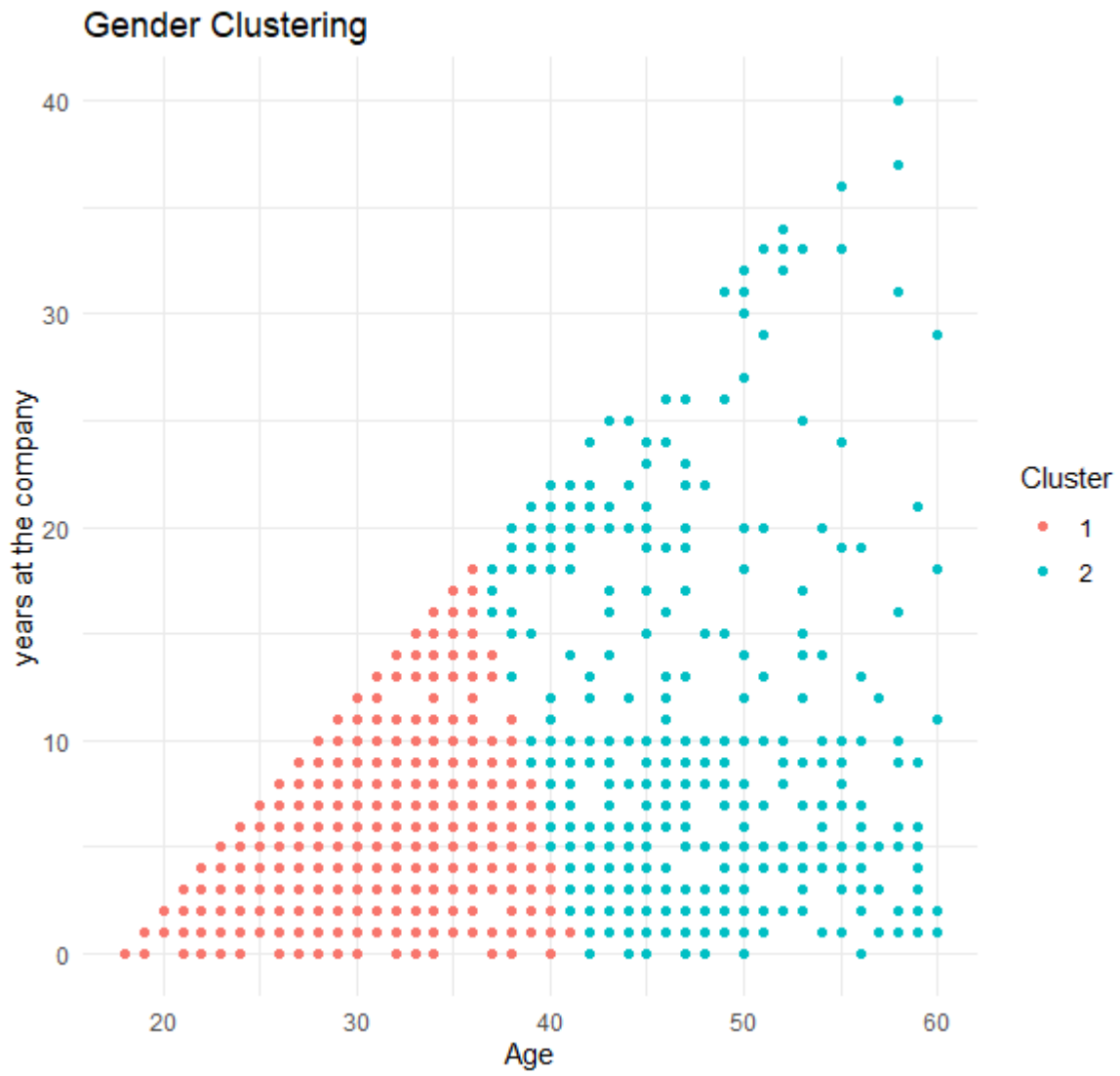
Code:

```
data2 <- read.csv("C:\\Users\\NMC\\Downloads\\archive (4)\\HR_Analytics.csv")

str(data2)

data2$AgeGroup <- as.integer(factor(data2$AgeGroup))
data2$Attrition <- as.integer(factor(data2$Attrition))
data2$BusinessTravel <- as.integer(factor(data2$BusinessTravel))
data2$Department <- as.integer(factor(data2$Department))
data2$EducationField <- as.integer(factor(data2$EducationField))
data2$Gender <- as.integer(factor(data2$Gender))
data2$JobRole <- as.integer(factor(data2$JobRole))
data2$MaritalStatus <- as.integer(factor(data2$MaritalStatus))
data2$SalarySlab <- as.integer(factor(data2$SalarySlab))
data2$Over18 <- as.integer(factor(data2$Over18))
data2$OverTime <- as.integer(factor(data2$OverTime))
data2$JobLevel <- as.integer(factor(data2$JobLevel))

str(data2)

sum(is.na(data2))
data2 <- na.omit(data2)
sum(is.na(data2))

str(data2)
```

```
library(caTools)

set.seed(150)
split <- sample.split(data2$Gender, SplitRatio = 0.75)
train <- subset(data2, split == TRUE)
test <- subset(data2, split == FALSE)

library(e1071)

classifier <- naiveBayes(Gender ~ ., data = train)
pre <- predict(classifier, newdata = test)

contable <- table(pre, test$Gender)
contable

library(caret)

confusionMatrix(contable)
```

```
pre   1   2
  1  35  53
  2 107 161

               Accuracy : 0.5506
                 95% CI : (0.4972, 0.603)
    No Information Rate : 0.6011
    P-Value [Acc > NIR] : 0.9769

                  Kappa : -0.0013

 Mcnemar's Test P-Value : 2.789e-05

            Sensitivity : 0.24648
            Specificity : 0.75234
         Pos Pred Value : 0.39773
         Neg Pred Value : 0.60075
             Prevalence : 0.39888
         Detection Rate : 0.09831
   Detection Prevalence : 0.24719
      Balanced Accuracy : 0.49941

       'Positive' Class : 1
```

2.Implementing K-means clustering for the Gender ,Age and Years at the company

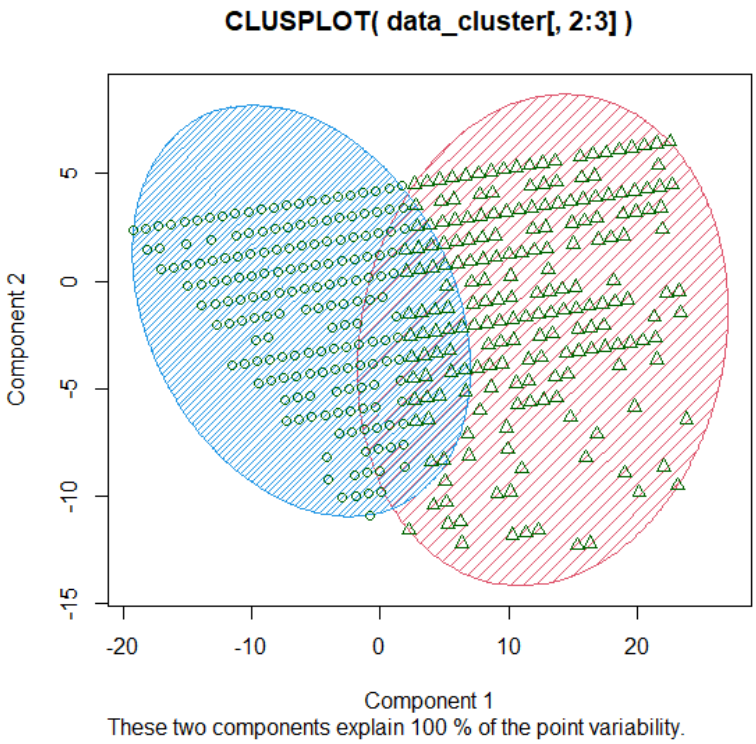


In this scatter plot here are two clusters the cluster 1 represents younger employees and the cluster 2 represents older employees in the company . According to the clustering the most employees have less than 10 years of experience .The number of employees who have a great work experience in that company is low when compared to the total number of employees.

Code:

```
install.packages(c("dplyr", "ggplot2"))
library(dplyr)
library(ggplot2)


data_cluster <- Hr %>%
  select(Gender, Age, YearsAtCompany)


data_cluster$Gender <- as.numeric(factor(data_cluster$Gender))


set.seed(123)
kmeans_model <- kmeans(data_cluster[, 2:3], centers = 2)


data_cluster$cluster <- as.factor(kmeans_model$cluster)


ggplot(data_cluster, aes(x = Age, y =YearsAtCompany , color = cluster)) +
  geom_point() +
  labs(title = "Gender Clustering",
       x = "Age",
       y = "years at the company",
       color = "Cluster") +
  theme_minimal()
```

3. Applying clustering Algorithm for the Gender



This cluster plot visually represents the gender distribution among clusters. And Plotting illustrates the variation in years worked for the firm across genders and maybe between clusters. In this plot each point represents an employee. .The x axis of the plot represent Gender and the y axis of the plot represent years at the company.

Code:

```
install.packages(c("dplyr", "cluster"))
library(dplyr)
library(cluster)


data_cluster <- Hr %>%
  select(Gender, Age, YearsInCurrentRole)


data_cluster$Gender <- as.numeric(factor(data_cluster$Gender))


set.seed(123)
kmeans_model <- kmeans(data_cluster[, 2:3], centers = 2)


data_cluster$cluster <- as.factor(kmeans_model$cluster)


clusplot(data_cluster[, 2:3], kmeans_model$cluster, color = TRUE, shade = TRUE, labels = 0, lines = 0)
```

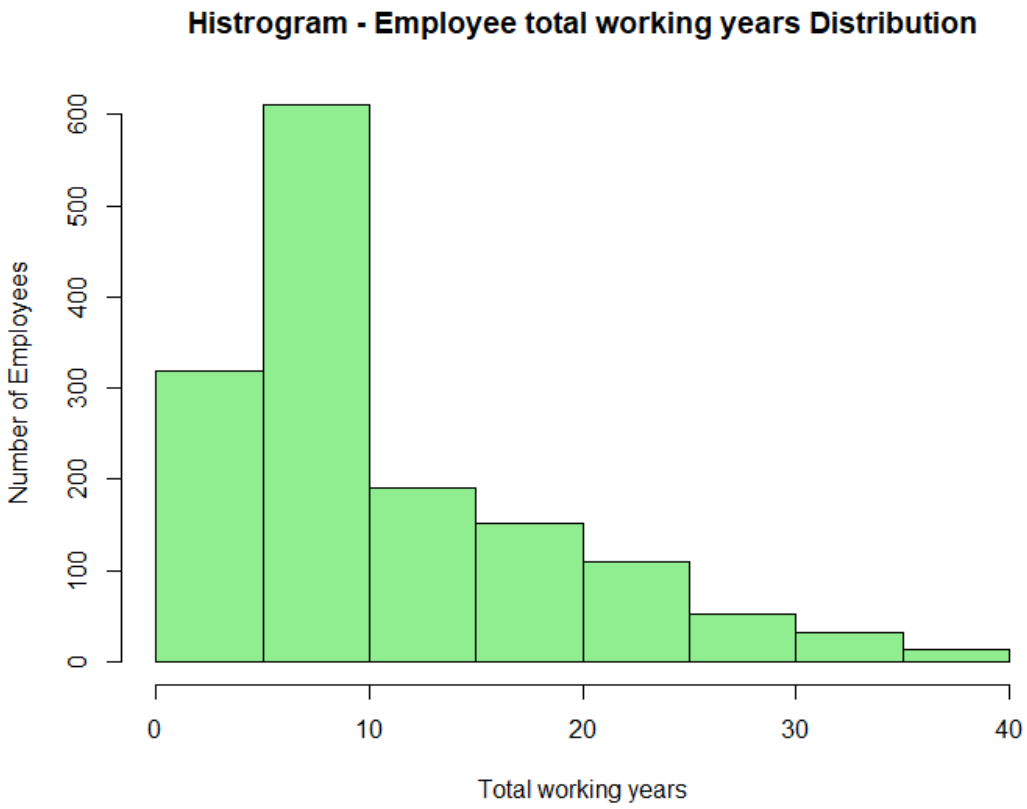
3.Total Working Years (Experience)

This is the 3rd category that we are selected to conduct an analysis to find out how it affects to a business. We try to identify patterns, trends and finally get valuable insights using this column.

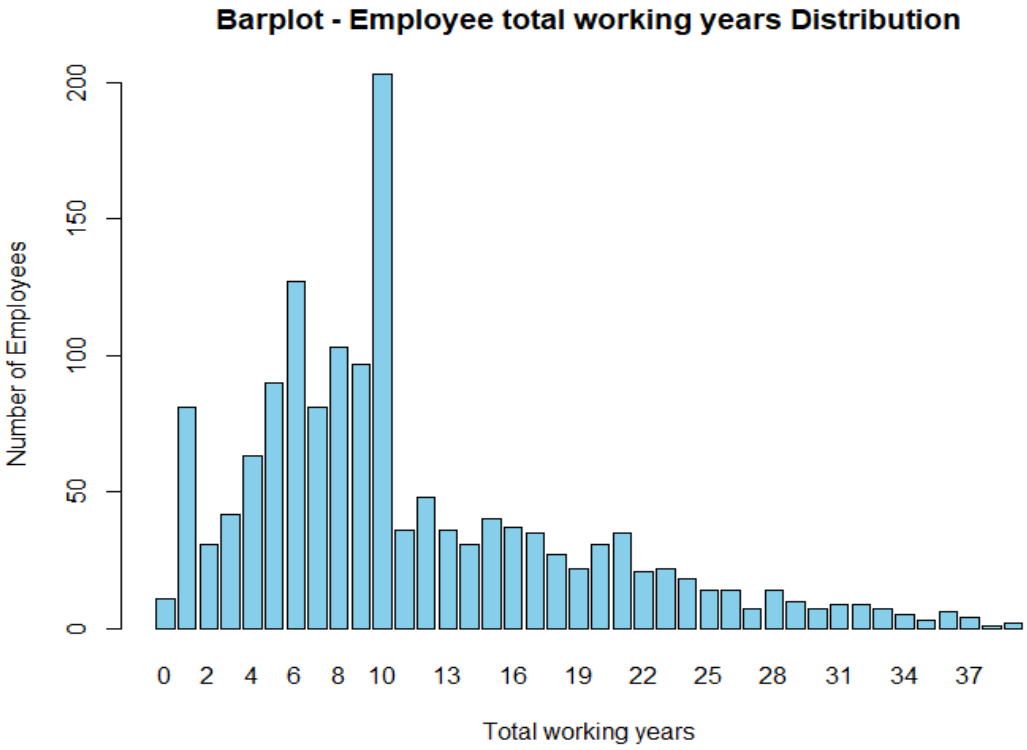
1.Data visualizations

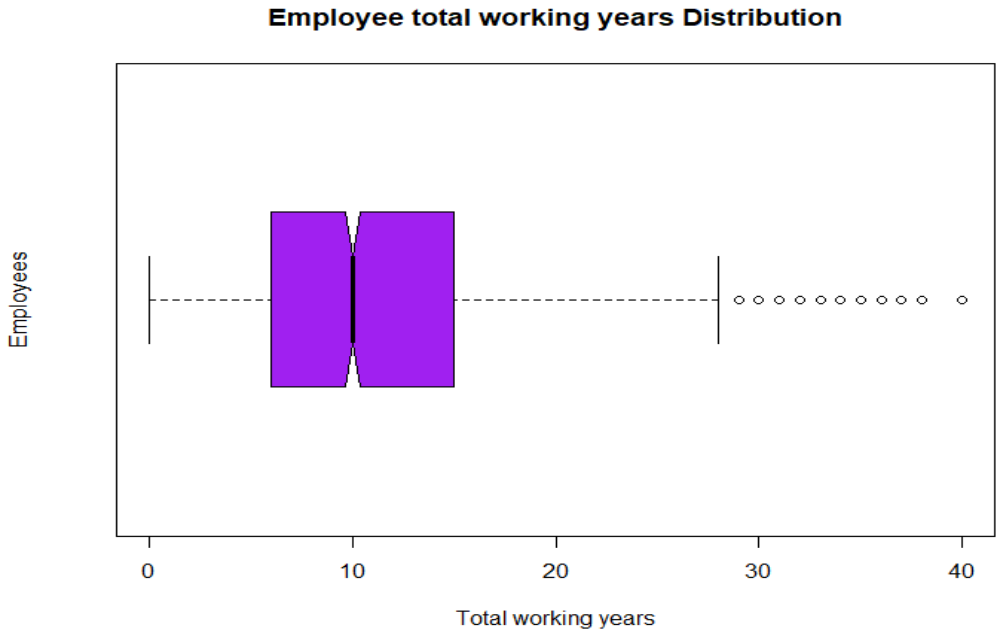
1. Main Column Introduction

This is an introduction about **"Total working Years"** column, **"Experience"** of an employee. This column includes numerical data. So, the most suitable visualization methods are Boxplots, Histograms, violin plots and Whisker plots.



This histogram represents distribution of "total working years" among employees. A larger proportion of employees(approximately 950) included in 0-10 years range. Upon that we can assume this company recruits more fresh graduates and young employees.

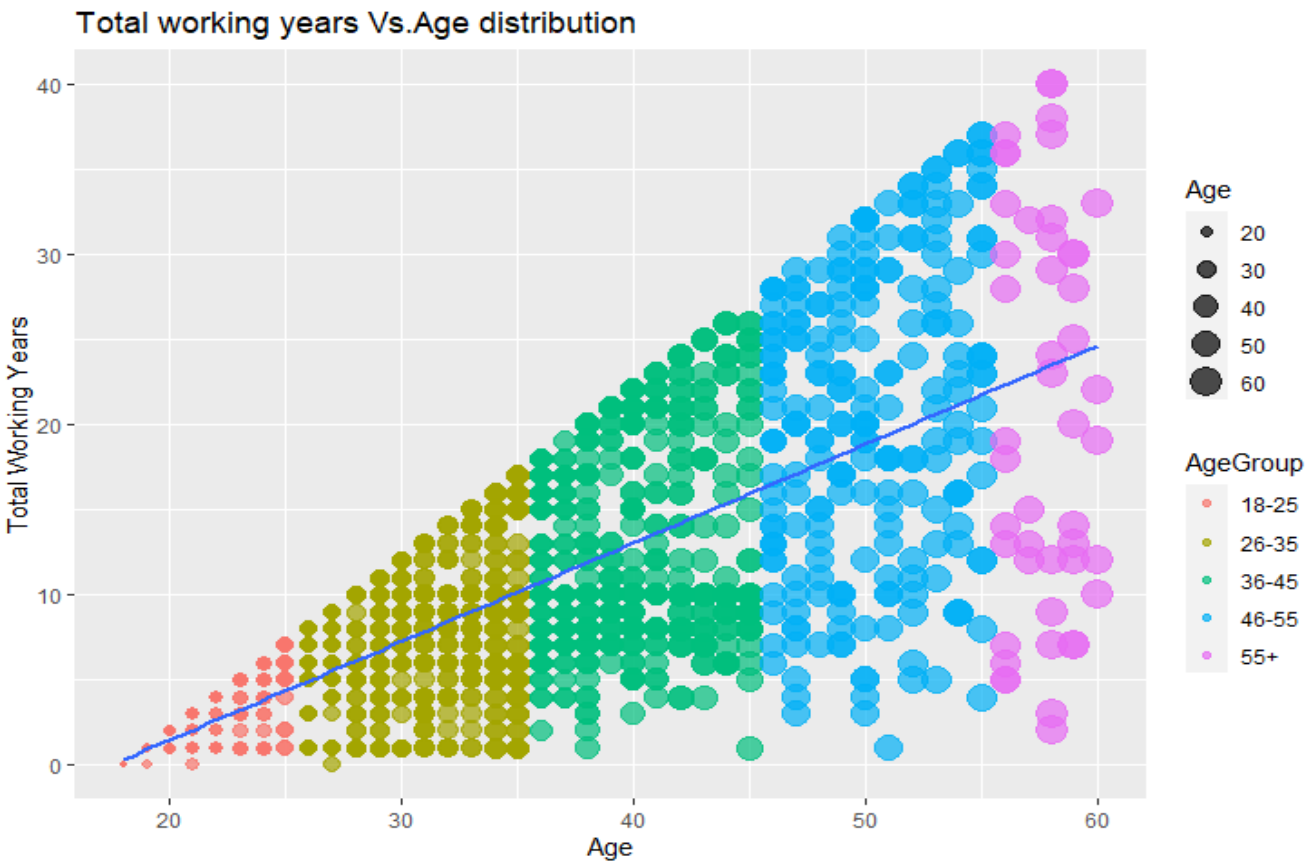




Above bar plot shows employee count that belongs to each experience year. And the box plot show min value is '0 years', 1st quartile is '5 years', median experience is '10 years', 3rd quartile is '15 years', max value is '28 years' and other years are as outliers. The graphs and the visualizations are generated by using the ggplot2 libraries. From now **"Total working Years"** is called as, **"Experience"** of an employee.

```
Experience <- table(data$TotalWorkingYears)
Experience
```

2. How experience vary on employee age distribution



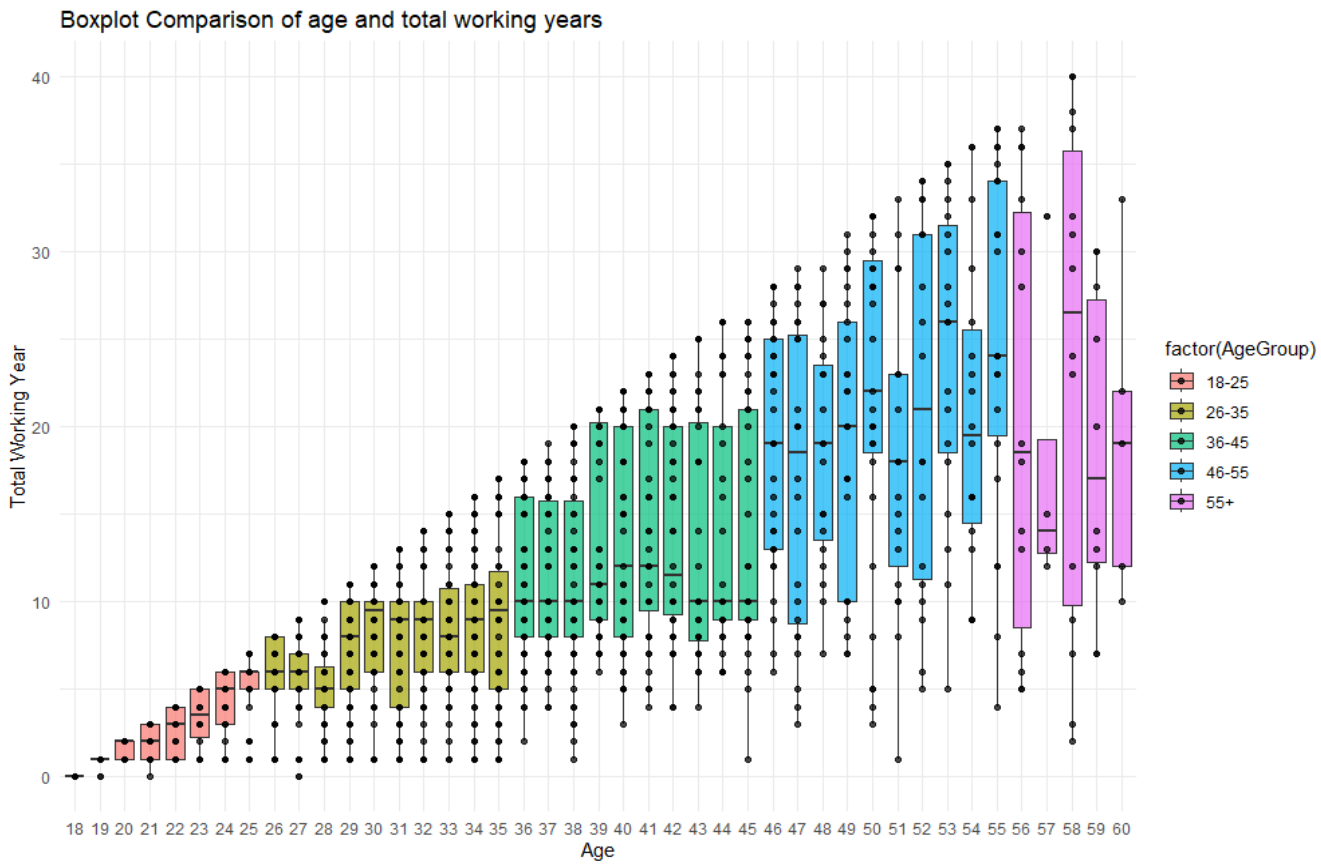
Above scatter plot shows the total experience of employees against age distribution. It has a positive correlation. The average number of working years increases with age. This is because people typically start working in their 20's and continue working until, they retire in 60's.

There is significant spread of variations in the data around the trendline. Based on the observations we can say, the average number of working years will continue to increase over time the employees ages, this can be helpful for get discissions about future workforce, and making future business policies.

code:

```
ggplot(data = data,aes(x=data$TotalWorkingYears,y=data$Age))+
  geom_point(alpha = 0.7,
             aes(size= Age,
                 colour = AgeGroup))+
  geom_smooth(method = lm, se=F)+
  labs(
    title = "Total working years Vs.Age distribution",
    x = "Total Working Years",
    y = "Age"
  )
```

3. More analysis about experience with boxplots



This is a Bar-plot comparison of total working years of employees with age, the data are divided into 5 age groups. The box plot shows the spread of the data in each age group. The median age in each group increases as the age group gets older.

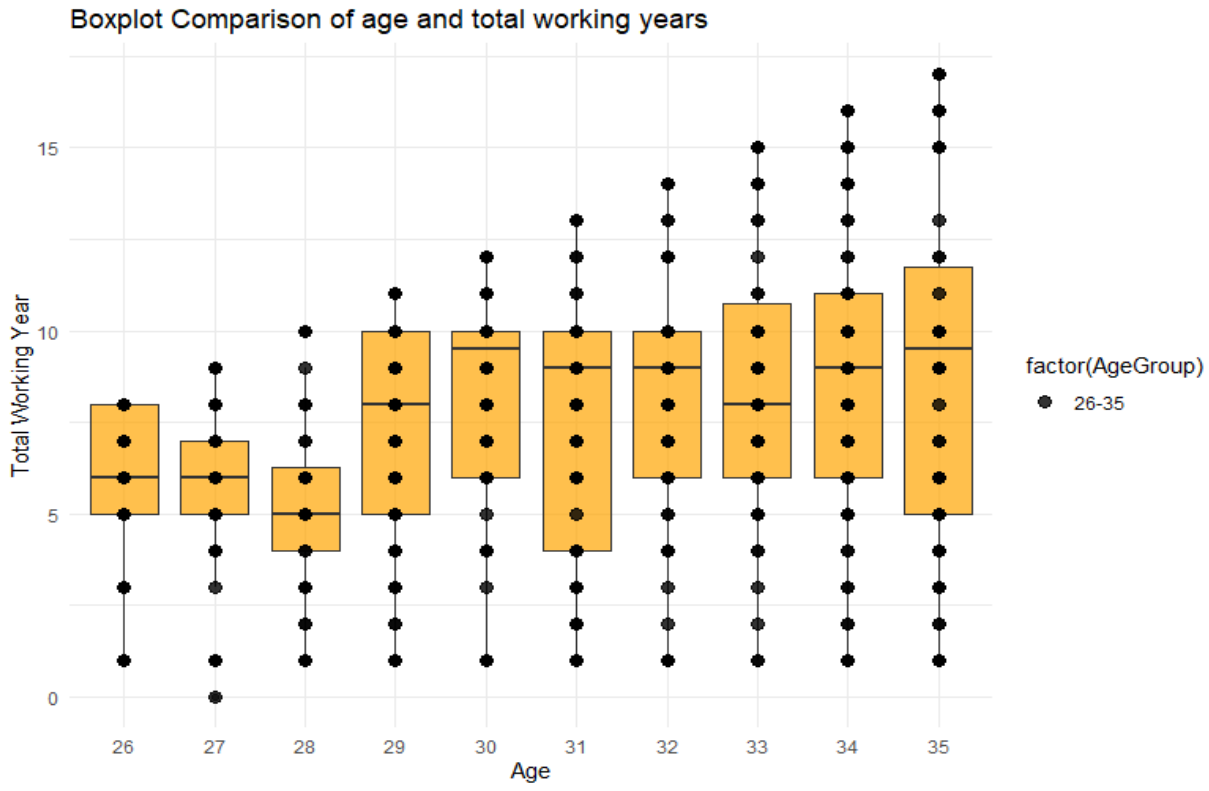
The spread of experience is greater in older age groups. We can assume that there is more variation in the number of years that people work in later life than in earlier life. There may be factors other than age that influence the experience of employees. Such as, education level, health, economic conditions and monthly income.

Code:

```
#Boxplots for all age groups with age distribution
ggplot(data, aes(x = factor(Age), y =TotalWorkingYears, fill = factor(AgeGroup))) +
  geom_boxplot(alpha = 0.7) +
  geom_point(alpha = 0.7)+
  labs(title = "Boxplot Comparison of age and total working years",
       x = "Age",
       y = "Total Working Year") +
  theme_minimal()
```

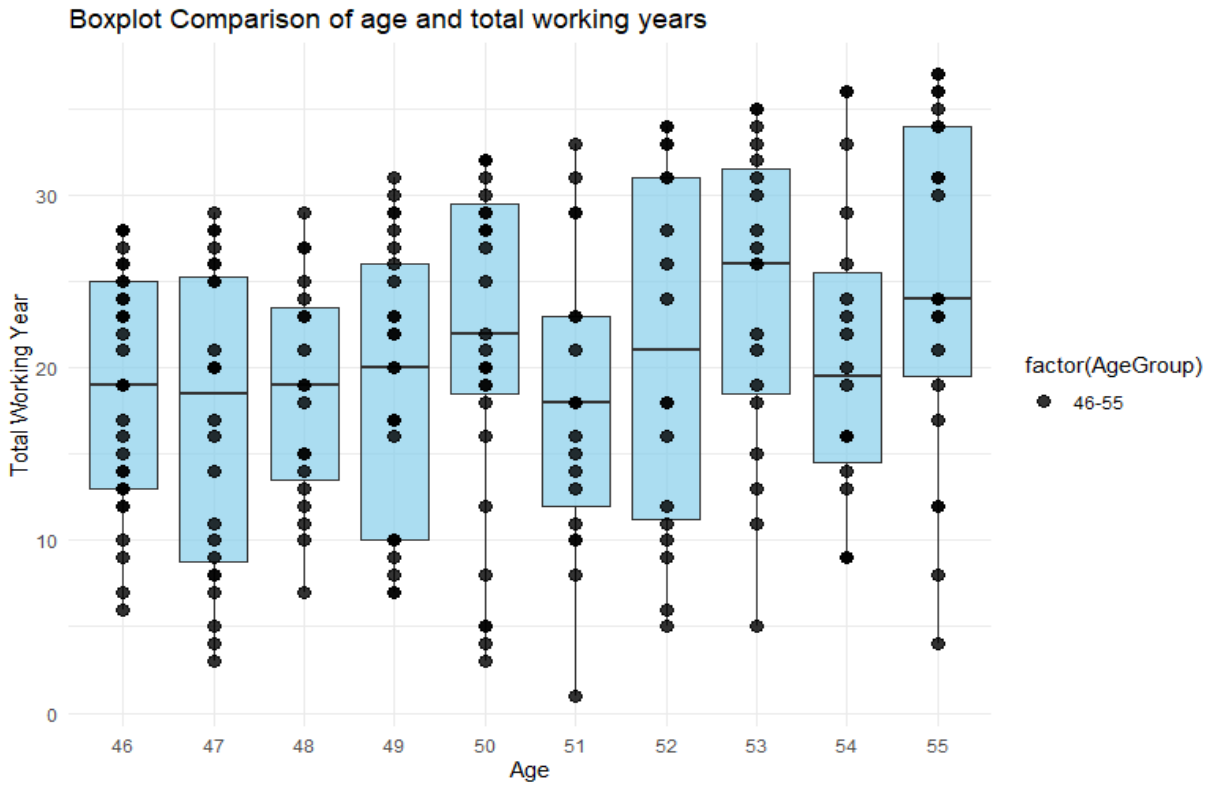
Age Group 18-25

Age Group 26-35

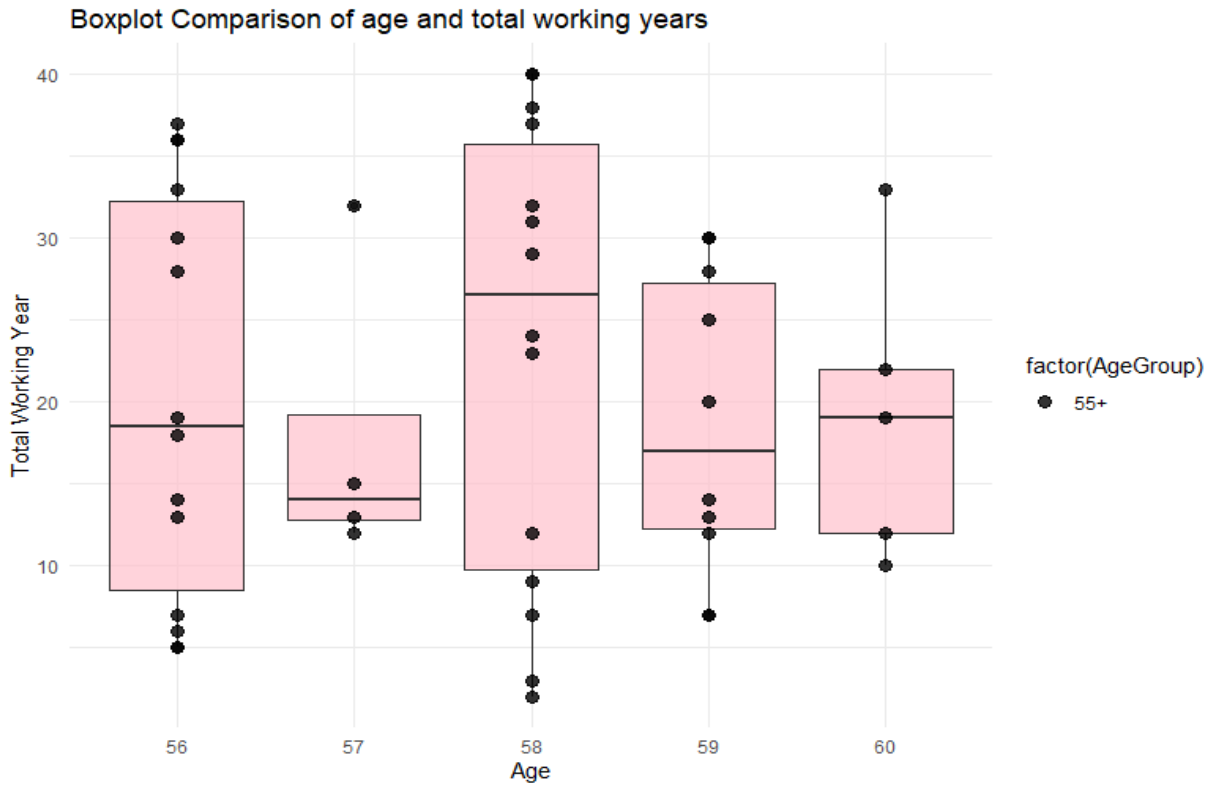


Age Group 36-45

Age Group 46-55



Age Group 55+

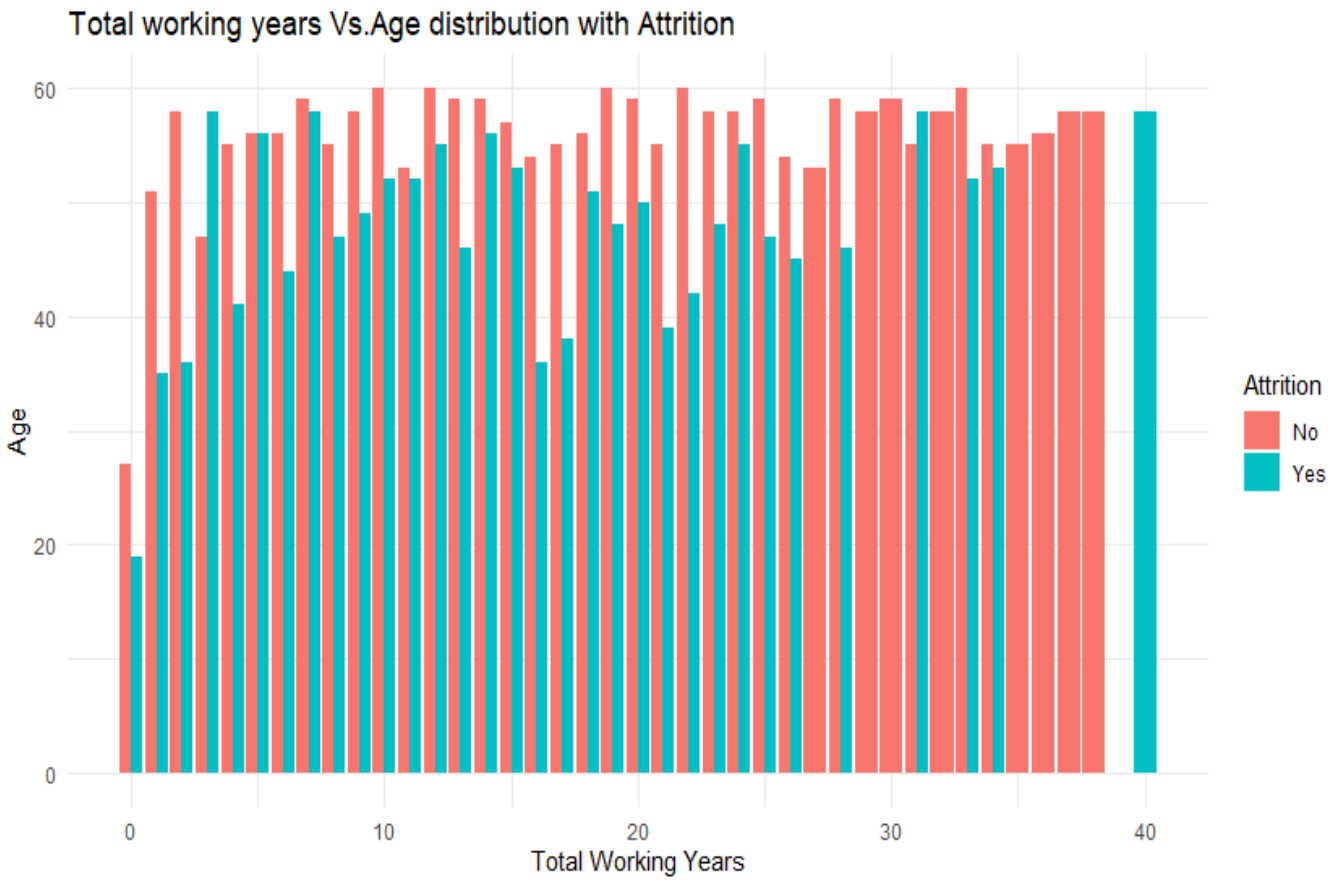


By observing each box-plot that plotted to age groups. We can get an idea about minimum value, 1st quartile, median value, 3rd quartile of total working years (experience) of employees in each age.

Code:

```
#Boxplots for <=25 age groups with age distribution
data %>%
  filter(Age<=25) %>%
ggplot(aes(x = factor(Age), y =TotalWorkingYears, fill = factor(AgeGroup))) +
  geom_boxplot(alpha = 0.6,fill= "red") +
  geom_point(alpha = 0.8,size = 3)+
  labs(title = "Boxplot Comparison of age and total working years",
       x = "Age",
       y = "Total Working Year") +
  theme_minimal()

#Boxplots for 25-35 age groups with age distribution
data %>%
  filter(Age > 25 & Age <= 35) %>%
  ggplot(aes(x = factor(Age), y = TotalWorkingYears, fill = factor(AgeGroup))) +
  geom_boxplot(alpha = 0.7,fill= "orange") +
  geom_point(alpha = 0.8, size = 3) +
  labs(title = "Boxplot Comparison of age and total working years",
       x = "Age",
       y = "Total Working Year") +
  theme_minimal()
```

4. Employee attrition rate with their experience



This plot shows the experience by age distribution with attrition. There are two sets of bars for each age group. Red implements employees who haven't left the company (Attrition = "No"), Green shows employees who have left (Attrition = "Yes"). When the experience increases the attrition rate gets low. Employees who are having 27 or more years of experience tends work until they become 60 and they retire in 60. There are lot of variations below 26 years of experience.

code:

```
#bar plot total working years (experience) vs Age with Attrition
ggplot(data = data, aes(x =data$Age, y = data$TotalWorkingYears, fill = Attrition)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total working years Vs.Age distribution with Attrition",
       x = "Total Working Years", y = "Age") +
  theme_minimal()
```

5. How job level varies on employee experience display with age

The scatter plot shows the relationship between total working years and job level distribution, with age groups as a secondary factor. Here are some observations and insights we can get from this. There is positive correlation between these two variables. As employee gain more experience and become more older, they tend to move up in job level. But we can observe some variations like young employees get into higher job levels but, it may affect various factors. There are lot of variability in the data. Even within the same age group and total working years, there are wide range of job levels spread.

Code:

```
ggplot(data = data,aes(x=data$TotalWorkingYears,y=data$JobLevel))+
  geom_point(alpha = 0.7,            #here we mapping the geom points only
             aes(size= Age,
                 colour = AgeGroup))+
  geom_smooth(method = lm, se=F)+
  labs(
    title = "Total working years Vs.Job level distribution with age groups",
    x = "Total Working Years",
    y = "Job Level"
  )
```

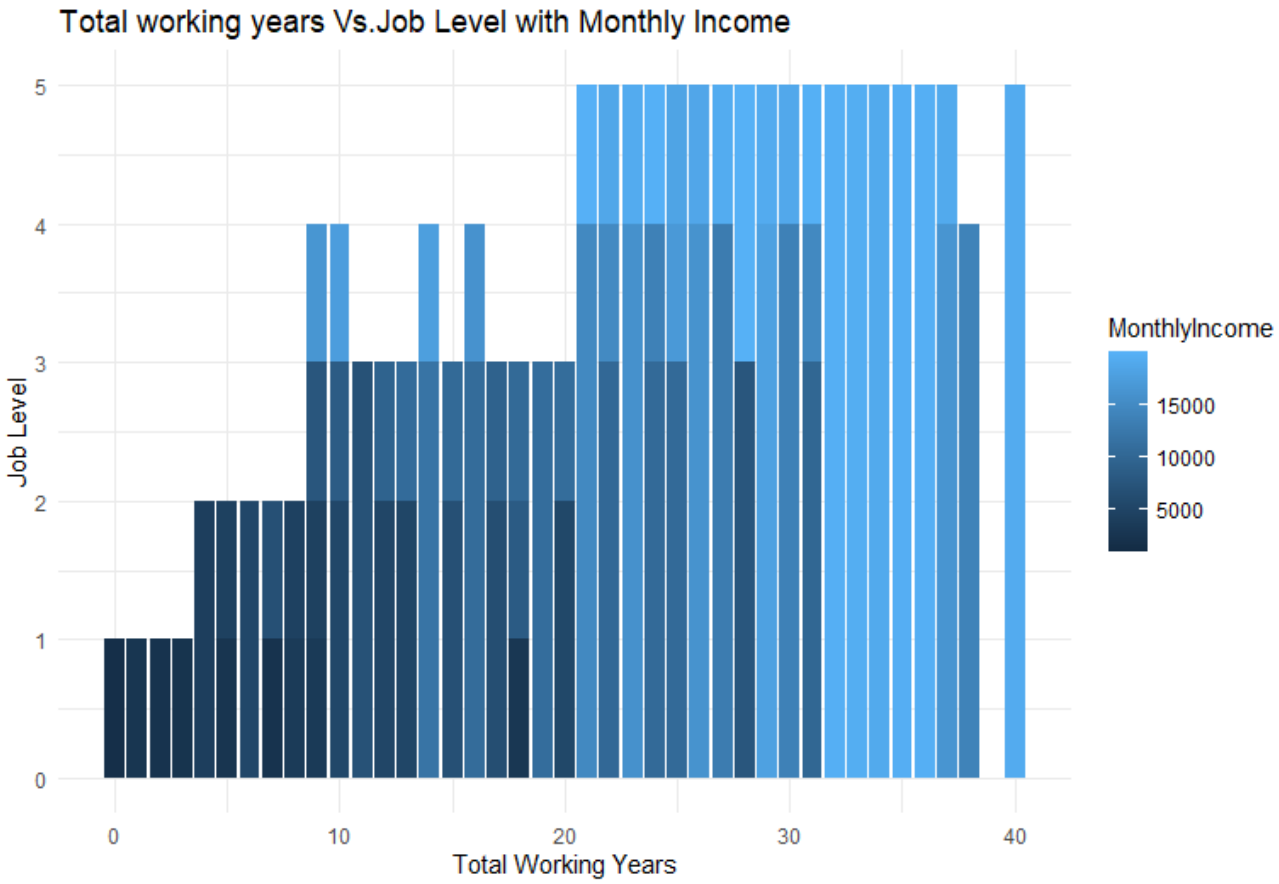
6. Employee job satisfaction rate with their experience and job level

Above graph shows the level of job satisfaction for total working years. Here are some observations and insights we found out. New employees tend to be more satisfied with their jobs. The highest level of job satisfaction is for employees with 10-30 years of experience. After that, job satisfaction starts to decline slightly. Employees who have been working longer and have higher job level seems to have low job satisfaction rate. This could be due to a number of factors such as people in this age group are approaching retirement and may be starting to think about slowing down or changing careers.

Code:

```
#section 3 (main insights)
#bar plot total working years (experience) vs Job Level with Job Satisfaction
ggplot(data = data, aes(x =data$TotalWorkingYears, y = data$JobLevel, fill = JobSatisfaction)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total working years Vs.Job Level with Job Satisfaction",
       x = "Total Working Years", y = "Job Level") +
  theme_minimal()
```

7. Employee monthly income with their experience and job level

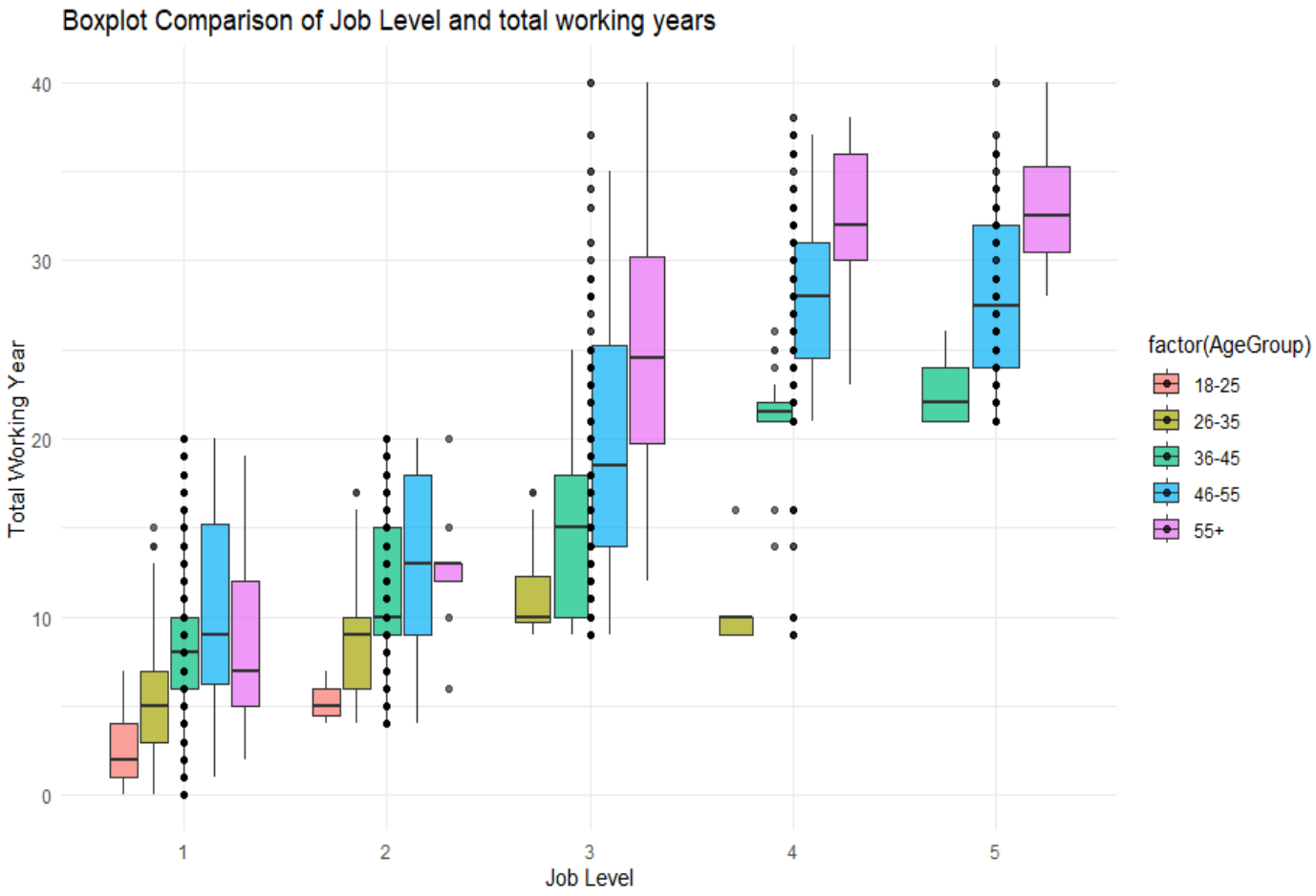


Above graph shows how monthly income will differ with experience. Here are some observations and insights we found out. New employees seem to have low income. Employees who are having more work experience have higher monthly income. So, there is a positive correlation. Finally, we identified when your work experience and job level increase you paid more amount monthly. Factors such as, job role, education level and responsibility can be affected to your monthly income.

Code:

```
#bar plot total working years (experience) vs Job Level with Monthly Income
ggplot(data = data, aes(x =data$TotalWorkingYears, y = data$JobLevel, fill = MonthlyIncome)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total working years Vs.Job Level with Monthly Income",
       x = "Total Working Years", y = "Job Level") +
  theme_minimal()
```

8. More analysis about job level with boxplots



This is a comparison of job level and total working years, with each box representing a different age group. Median job level increases with total working years across all age groups. The median job level tends to be higher for those with more total working years. This suggests that, when employees move up in job level as they gain experience. We can get more information using these boxplots.

Code:

```
#Boxplots for all Job Levels with age distribution
ggplot(data, aes(x = factor(JobLevel), y =TotalWorkingYears, fill = factor(AgeGroup))) +
  geom_boxplot(alpha = 0.7) +
  geom_point(alpha = 0.7)+
  labs(title = "Boxplot Comparison of Job Level and total working years",
       x = "Job Level",
       y = "Total Working Year") +
  theme_minimal()
```

2. Data analytics through models

1. Predict attrition rate using Logistic regression model

Logistic Regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome (Yes or No). In this data set there are two columns (Attrition & Over Time) that compatible for a binary classification. So, we decided to choose attrition as our dependent variable (Y) and total working years plus Job level as independent variables (X).

At first, we load the necessary libraries. Then did the data cleaning and preprocessing. We checked for null values in the dataset and removed them. In logistic regression the dependent variable always must be a binary outcome. So, we converted “Yes” & “No” to “1” & “0” in Attrition column.

Code:

```
dataset <- read.csv("D:\\N learn\\2nd year\\projects\\R project\\HR_A
summary(dataset)

#data cleaning and pre-processing
sum(is.na(dataset))
dataset <- na.omit(dataset)

class(dataset$Attrition)
unique(dataset$Attrition)
dataset <- dataset %>%
  mutate(Attrition = ifelse(Attrition == "No",0,1))
str(dataset)
```

After the cleaning & preprocessing we continued to the model building process. First of all, we split the dataset in to training and testing sets. Then trained the logistic regression model.

```
# Splitting dataset
set.seed(123)
split <- sample.split(dataset, SplitRatio = 0.7) # 30/70
split

train_reg <- subset(dataset, split == "TRUE")
test_reg <- subset(dataset, split == "FALSE")

# Training model
logistic_model <- glm(Attrition ~ JobLevel + TotalWorkingYears,
                      data = train_reg,
                      family = "binomial")
logistic_model

# Summary
summary(logistic_model)

predict_reg <- predict(logistic_model,
                       test_reg, type = "response")
predict_reg
```

We trained the model using two independent variables so, it called as multivariate logistic regression model. Then we did the model evaluation using confusion matrix and ROC curve. Finally, we got the predicted output for attrition.

```
# Changing probabilities for prediction
predict_reg <- ifelse(predict_reg >0.5, 1, 0)

# Keep probabilities for ROC curve
predict1_reg <- predict(logistic_model, test_reg, type = "response")


# Evaluating model accuracy using confusion matrix
confusion_matrix <-table(test_reg$Attrition, predict_reg)
print(confusion_matrix)

missing_classerr <- mean(predict_reg != test_reg$Attrition)
print(paste('Accuracy =', 1 - missing_classerr))

contable <- confusionMatrix(as.factor(predict_reg), as.factor(test_reg$Attrition))
print(contable)


# ROC-AUC Curve
ROCPred <- prediction(predict1_reg, test_reg$Attrition)
ROCPer <- performance(ROCPred, measure = "tpr",
                      x.measure = "fpr")

auc <- performance(ROCPred, measure = "auc")
auc <- auc@y.values[[1]]
auc

# Plotting curve
plot(ROCPer)
plot(ROCPer, colorize = TRUE,
     print.cutoffs.at = seq(0.1, by = 0.1),
     main = "ROC CURVE")
abline(a = 0, b = 1)

auc <- round(auc, 4)
legend(.6, .4, auc, title = "AUC", cex = 1)
```

Final outputs: (model summery)

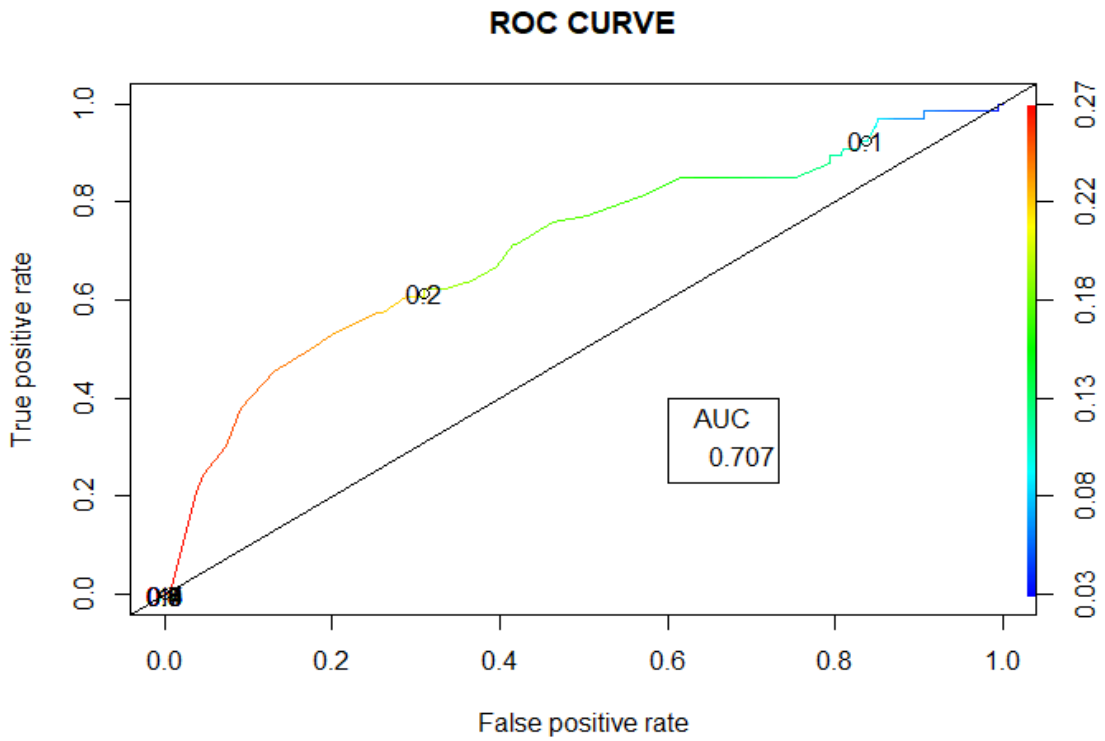
```
> # Summary
> summary(logistic_model)

Call:
glm(formula = Attrition ~ JobLevel + TotalWorkingYears, family = "binomial",
    data = train_reg)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.64158    0.19176  -3.346  0.00082 ***
JobLevel          -0.35777    0.13991  -2.557  0.01055 *
TotalWorkingYears -0.02753    0.01955  -1.408  0.15907
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 882.67  on 972  degrees of freedom
Residual deviance: 849.45  on 970  degrees of freedom
AIC: 855.45

Number of Fisher Scoring iterations: 5
```

```
> print(contable)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 384  66
         1   0   0

               Accuracy : 0.8533
                 95% CI : (0.8172, 0.8847)
    No Information Rate : 0.8533
    P-Value [Acc > NIR] : 0.5328

                  Kappa : 0

 Mcnemar's Test P-Value : 1.235e-15

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8533
         Neg Pred Value :    NaN
             Prevalence : 0.8533
         Detection Rate : 0.8533
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

```
> confusion_matrix <-table(test_reg$Attrition, predict_reg)
> print(confusion_matrix)
   predict_reg
     0
  0 384
  1  66
>
> missing_classerr <- mean(predict_reg != test_reg$Attrition)
> print(paste('Accuracy =', 1 - missing_classerr))
[1] "Accuracy = 0.853333333333333"
```

As shown in the above the model accuracy is 0.853. and our test data set have 450 observations. The confusion matrix shows out of 450, 384 are stay at the company and 66 of them will leave the company. So, that is the prediction about attrition.

ROC curve image is shown below. ROC curve provides a more evaluation of a model's performance. Area under curve is 0.707.



4. Education

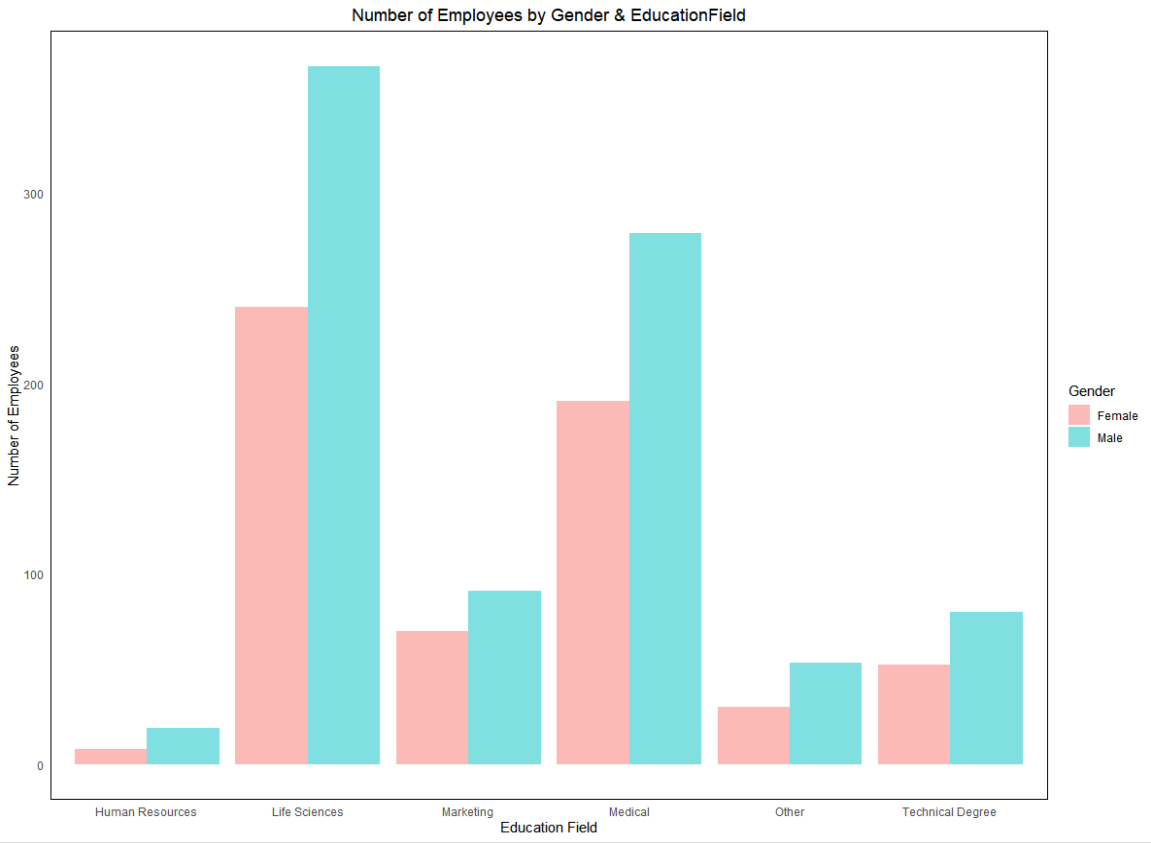
Education is one of the main components of a HR dataset. If we take look at the dataset, we can see several columns are related to education. In this part, our goal is to identify those relations and how education is affecting in other columns.

1. Data Visualization

We have got Education Field and Education Level as the columns regarding to Education. Both of them are qualitative data. So, we used Bar Plots, Pie Charts, Bee Swarm Plots, Strip Charts & Violin Plots as our techniques, which are commonly used to plot qualitative data in R.

First of all, we imported the required libraries and the data set to start the visualization process. Below code shows that process. After that process we started to plot different columns with different plotting tools and methods. We are going to mention all of them below.

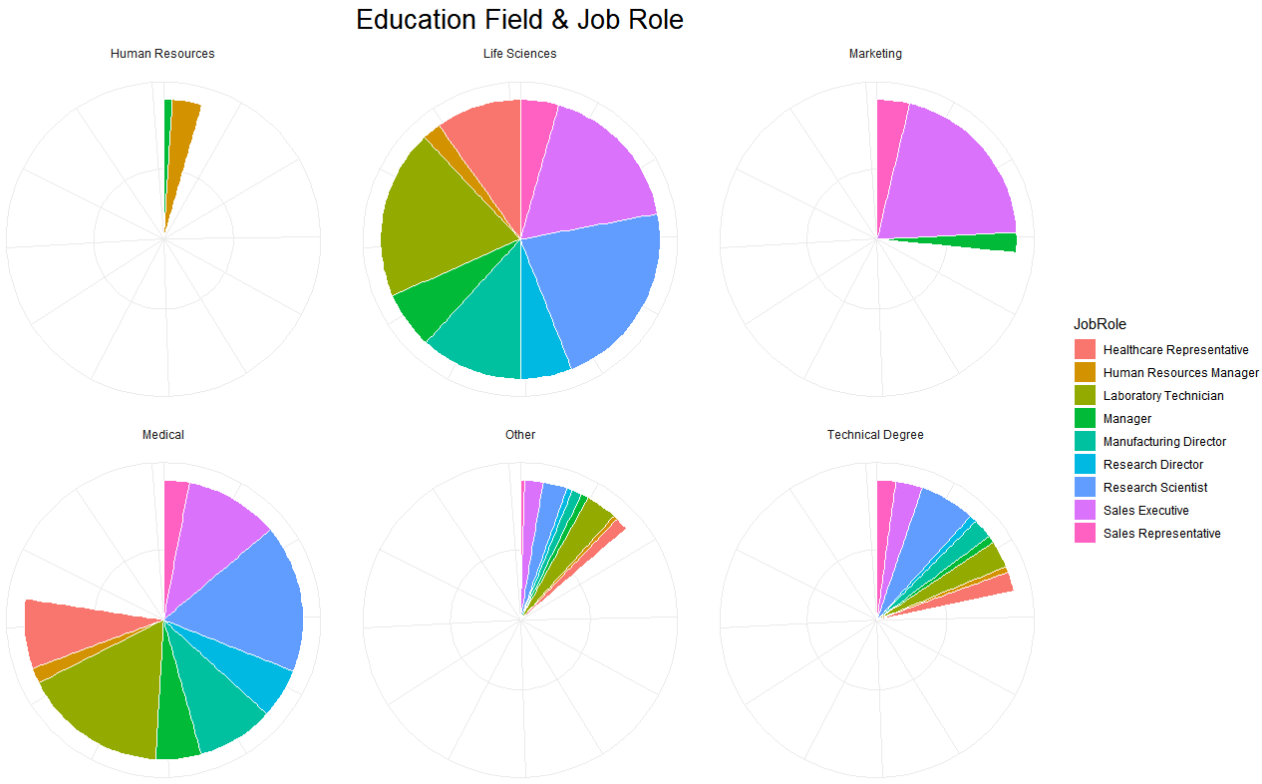
```r
# Importing the required libraries and the data set

library(tidyverse)
library(ggplot2)
library(dplyr)

data <- read.csv("C:\\Users\\Gosisa Jinasena\\Desktop\\R Project Education\\Education Viz\\HR_Analytics.csv")
```



This bar-plot was used to check whether how employees distributed on their Education Field & Gender. According to the bar-plot Life Sciences takes a huge number of employees compared to other fields.

This bar-plot was generated by below code

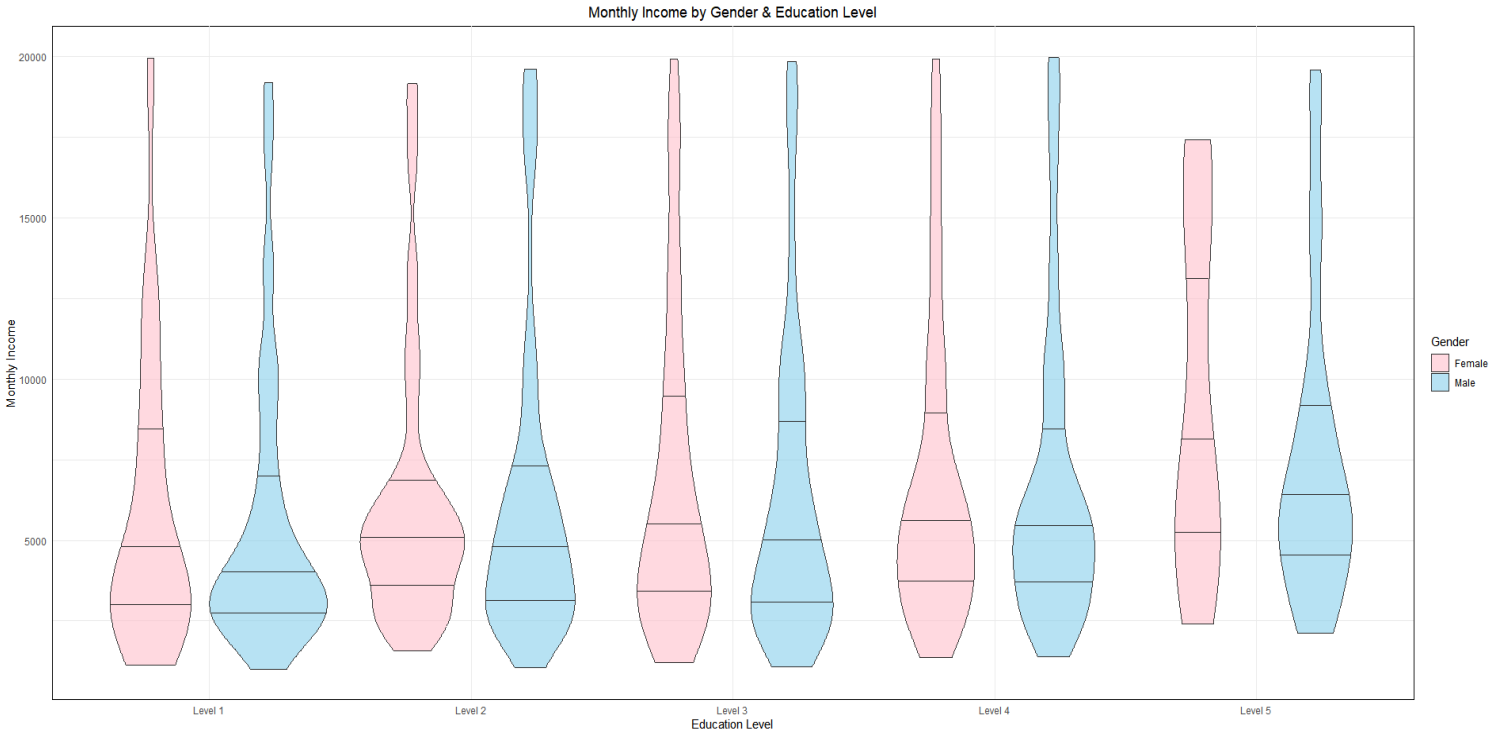
```
# Number of Employees by Gender & Education Field

ggplot(data, aes(x = EducationField, fill = Gender)) +
  geom_bar(position = "dodge",
           alpha = 0.5) +
  labs(title = "Number of Employees by Gender & Education Field",
       x = "Education Field",
       y = "Number of Employees") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_rect(fill = "white", color = "black", linetype = "solid"),
        axis.line = element_line(color = "black"))
```



This pie charts were used to take a proper idea about how job roles are distributed with different education fields. Number of employees are shown as percentages.

```r
# Education Field & Job Role

data %>%
  select(JobRole, EducationField) %>%
  mutate(JobRole = recode(JobRole,"Human Resources" = "Human Resources Manager")) %>%
  arrange(JobRole) -> Job_Edu

Job_Edu %>%
  ggplot(aes(x = "", fill = JobRole)) +
  theme_bw() +
  geom_bar(width = 1, color = "white") +
  facet_wrap(~EducationField) +
  coord_polar("y", start = 0) +
  ggtitle("Education Field & Job Role") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank())
```

We found a job role called “Human Resources” in our dataset. Human Resources is no a job role. So, we modified that data in to “Human Resource Manager”. Top of the code snippet is mentioning the process.



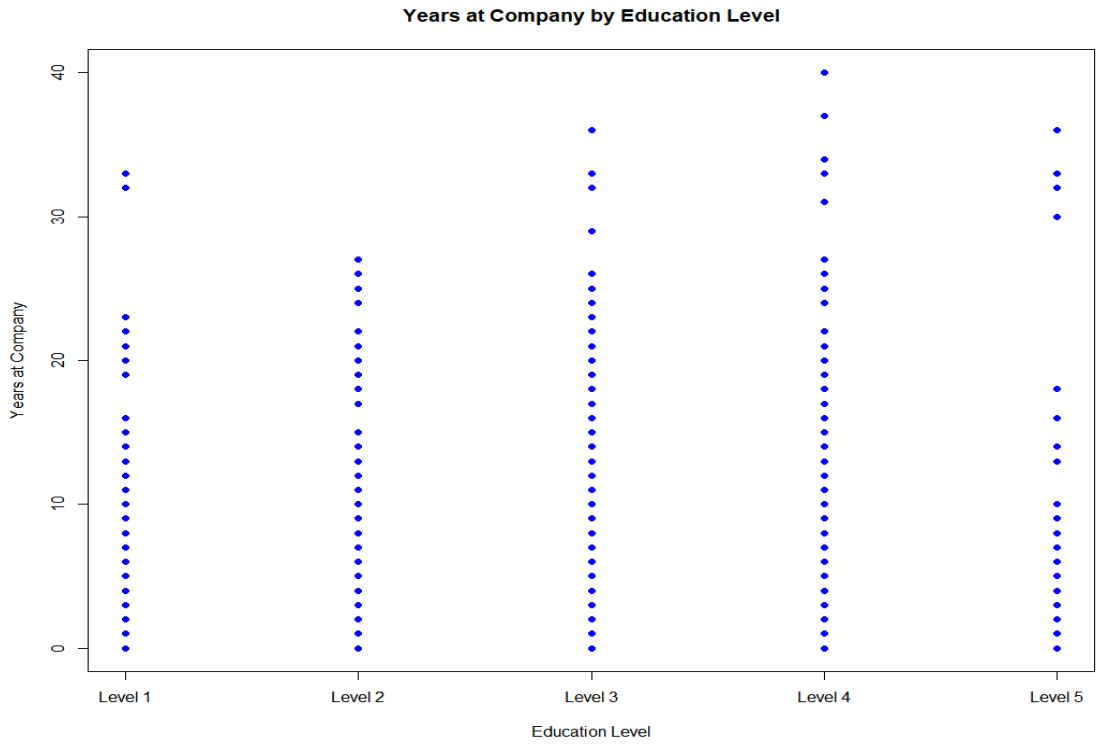
This is the violin plot for Monthly Income by Gender & Education Level. According to the 2nd quantile of the plot, we can say that levels 1,2,3, and 4 have the same behavior and each of them (50% of each level) has got around 5000 as their monthly income. Compared to the other levels level 5 has a different behavior. It has more density on the top (that means most of them have higher monthly incomes) and the beginning of monthly income for level 5 is around 2500(while other levels are beginning from 1250).

According to this plot, we can consider that when the Education Level is higher level monthly income also increases relatively.

We generated this violin plot from below code.

```r
# Monthly Income by Gender & Education Level

data %>%
  select(Education, MonthlyIncome, Gender) %>%
  mutate(Education = recode(Education,
                            "1" = "Level 1",
                            "2" = "Level 2",
                            "3" = "Level 3",
                            "4" = "Level 4",
                            "5" = "Level 5")) %>%
  arrange(Education) -> Monthly_Edu

ggplot(Monthly_Edu, aes(x = Education, y = MonthlyIncome, fill = Gender)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75), alpha = 0.6) +
  labs(title = "Monthly Income by Gender & Education Level",
       x = "Education Level",
       y = "Monthly Income")+
  scale_fill_manual(values = c("pink","skyblue")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        panel.background = element_rect(fill = "white", color = "black", linetype = "solid"),
        axis.line = element_line(color = "black"))
```

In our Education Level column, we received the levels as 1,2,3,4, and 5. For better understanding and visualization we changed them to Level 1, Level 2, Level 3, Level 4, and Level 5. At the top of the code snippet, we can see the code operations.



This is the strip chart for Years at company by Education level. We can notice that the majority of employees who have level 3 and 4 level education are have worked many years at the company. It means for that employees who are at a higher education level there will be a high demand in the company.

Code for the strip chart is given below

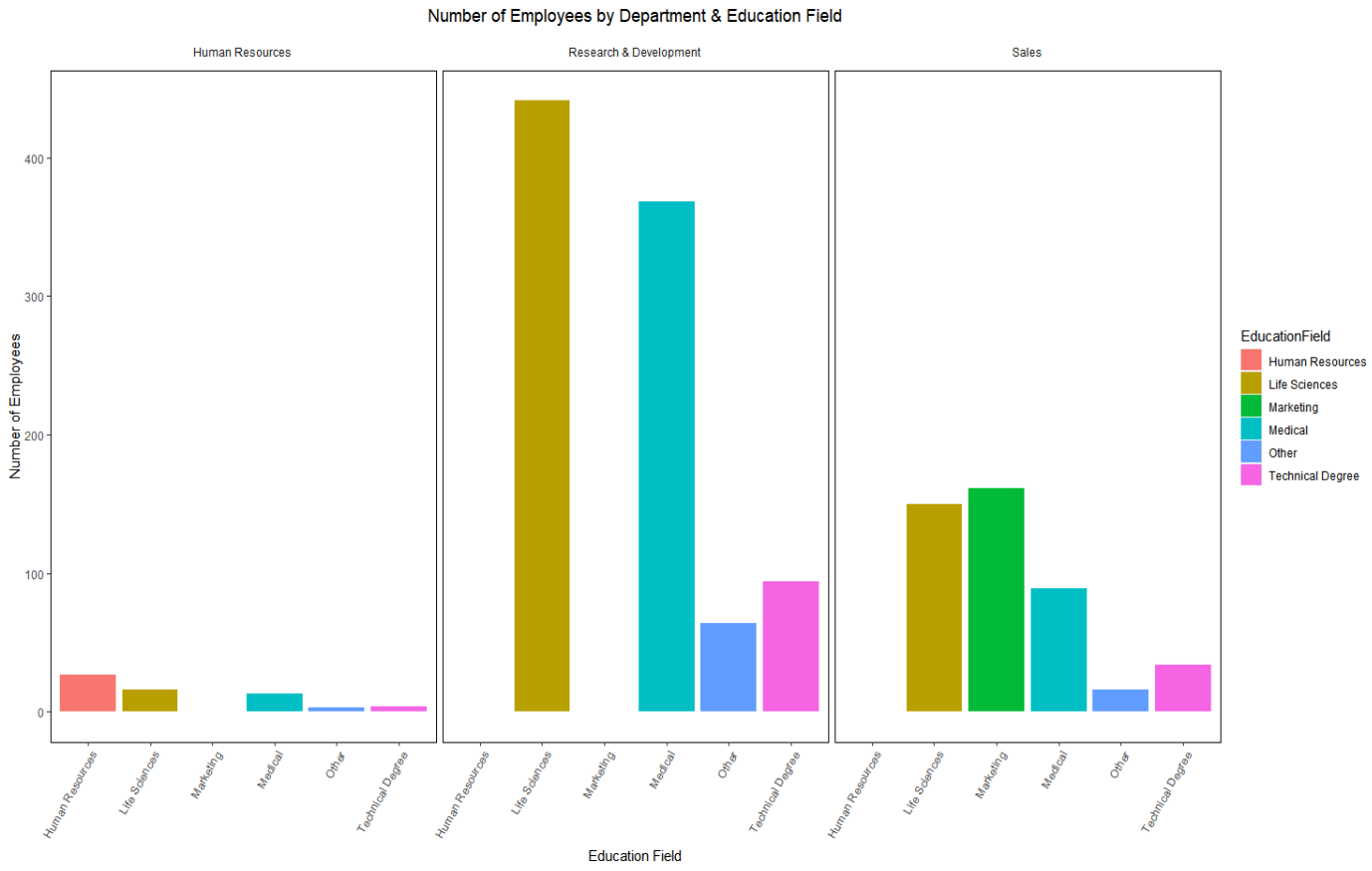
```
# Years at Company by Education Level

stripchart(YearsAtCompany~Education,
           data=data,
           main="Years at Company by Education Level",
           xlab="Education Level",
           ylab="Years at Company",
           col="blue",
           group.names=c("Level 1","Level 2","Level 3","Level 4","Level 5"),
           vertical=TRUE,
           pch=16)
```

This is the bee swarm plot of Monthly income by education field. We can say that majority of employees who choose their education field as life sciences and medical are having 2500 to 7500 as their monthly income.

Code snippet for the above bee swarm plot is given below

```
# Monthly Income by Education Field

install.packages("beeswarm")
library(beeswarm)

beeswarm(MonthlyIncome~EducationField,
         data=data,
         main="Monthly Income by Education Field",
         xlab="Education Field",
         ylab="Monthly Income",
         col="skyblue",
         vertical=TRUE,
         pch=16)
```

We need to install the library called “bee swarm” to generate this plot.



This facet wrapped bar plots were used to get an idea about how the employees distributed to the departments by their education field.

Code of the above facet wrap ped bar plots are given below

```
# Number of Employees by Department & Education Field

ggplot(data, aes(x = EducationField, fill = EducationField)) +
  geom_bar() +
  facet_wrap(~ Department) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        plot.title = element_text(hjust = 0.5),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_rect(fill = "white", color = "black", linetype = "solid"),
        axis.line = element_line(color = "black"),
        strip.text = element_text(margin = margin(10, 10, 10, 10))) +
  theme(strip.background = element_blank()) +
  labs(title = "Number of Employees by Department & Education Field",
         x = "Education Field",
         y = "Number of Employees")
```

2. Data Analytics through models

1. A classification model using Logistic Regression algorithm to predict Attrition based on Monthly Income and Education Level.

Logistic Regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome. In the data set there were two columns (Attrition & Over Time) compatible for a binary classification. So, we decided to choose attrition as our dependent variable.

```
library(caTools)
library(ROCR)
library(dplyr)
library(caret)

dataset <- read.csv("C:\\Users\\Gosisa Jinasena\\Desktop\\R Project Education\\Education Viz\\HR_Analytics.csv")
```

We loaded some needed libraries and the dataset to start the process. After that we started the data cleaning and the preprocessing. We checked for null values in the dataset and removed them. In logistic regression the dependent variable always must be a binary outcome. So, we converted “Yes” & “No” to “1” & “0” in Attrition column. Also converted other categorical variables into integer variables.

Code of the Cleaning & Preprocessing is given below

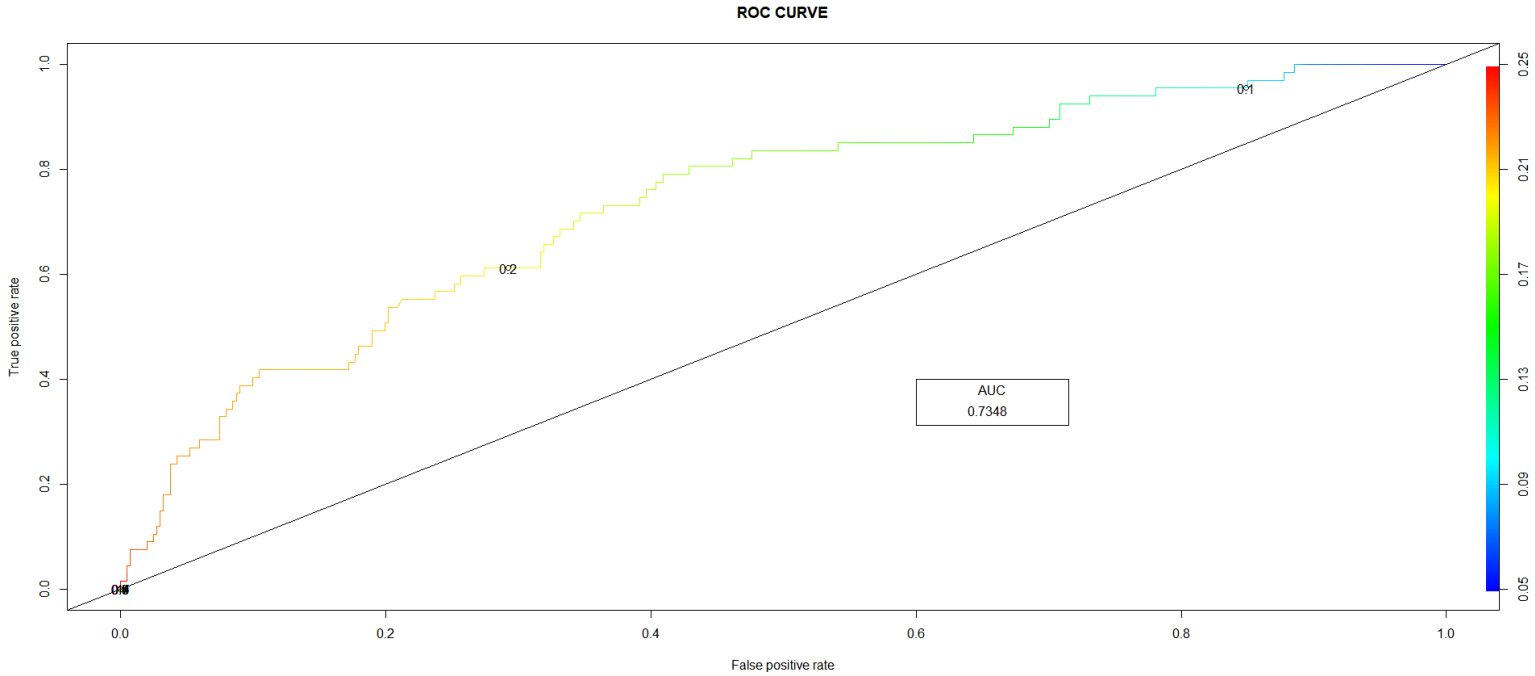
```
# Data Cleaning & Data Preprocessing

dataset$AgeGroup <-as.integer(factor(dataset$AgeGroup))
dataset$BusinessTravel <-as.integer(factor(dataset$BusinessTravel))
dataset$Department <-as.integer(factor(dataset$Department))
dataset$EducationField <-as.integer(factor(dataset$EducationField))
dataset$Gender <-as.integer(factor(dataset$Gender))
dataset$JobRole <-as.integer(factor(dataset$JobRole))
dataset$MaritalStatus <-as.integer(factor(dataset$MaritalStatus))
dataset$Salaryslab <-as.integer(factor(dataset$Salaryslab))
dataset$Over18 <-as.integer(factor(dataset$Over18))
dataset$OverTime <-as.integer(factor(dataset$OverTime))

dataset <- dataset %>%
  mutate(Attrition = ifelse(Attrition == "No",0,1))

sum(is.na(dataset))
dataset <- na.omit(dataset)
```

After the cleaning process done, we continued to the model building process. First of all, we split the dataset in to training and testing. After that trained a logistic regression model (used attrition as dependent variable and monthly income and education level as independent variables). After that moved to the evaluation process. We used a confusion matrix for that process. It showed the model’s accuracy as 0.8568. Overall coding and the confusion matrix are given below

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 401  67
         1   0   0

               Accuracy : 0.8568
                 95% CI : (0.8218, 0.8873)
    No Information Rate : 0.8568
    P-Value [Acc > NIR] : 0.5325

                  Kappa : 0

 Mcnemar's Test P-Value : 7.433e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8568
         Neg Pred Value :    NaN
             Prevalence : 0.8568
         Detection Rate : 0.8568
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

```r
# Splitting Dataset

set.seed(123)
split <- sample.split(dataset, SplitRatio = 0.7)
split

train_reg <- subset(dataset, split == "TRUE")
test_reg <- subset(dataset, split == "FALSE")

# Training Model

logistic_model <- glm(Attrition ~ MonthlyIncome + Education,
                      data = train_reg,
                      family = "binomial")
logistic_model

# Summary

summary(logistic_model)

predict_reg <- predict(logistic_model,
                       test_reg, type = "response")
predict_reg


# Changing probabilities for prediction

predict_reg <- ifelse(predict_reg >0.5, 1, 0)

# Keep probabilities for ROC curve

predict1_reg <- predict(logistic_model, test_reg, type = "response")


# Evaluating model accuracy using confusion matrix

confusion_matrix <-table(test_reg$Attrition, predict_reg)
print(confusion_matrix)

missing_classerr <- mean(predict_reg != test_reg$Attrition)
print(paste('Accuracy =', 1 - missing_classerr))

contable <- confusionMatrix(as.factor(predict_reg), as.factor(test_reg$Attrition))
print(contable)
```

```r
# ROC-AUC Curve

ROCPred <- prediction(predict1_reg, test_reg$Attrition)
ROCPer <- performance(ROCPred, measure = "tpr",
                      x.measure = "fpr")

auc <- performance(ROCPred, measure = "auc")
auc <- auc@y.values[[1]]
auc

# Plotting curve

plot(ROCPer)
plot(ROCPer, colorize = TRUE,
     print.cutoffs.at = seq(0.1, by = 0.1),
      main = "ROC CURVE")
abline(a = 0, b = 1)

auc <- round(auc, 4)
legend(.6, .4, auc, title = "AUC", cex = 1)
```



ROC AUC score is using to check the efficiency of a logistic regression model. We have got our AUC score as 0.7348. It is closer to 1 which means model is a good classifier.

5. Marital Status

Marital Status of the employees can be a determining factor regarding a lot of aspects in this HR analytics dataset. At first glance, u can notice from the columns in the dataset that it could be related to a lot of the columns (areas) that are mentioned.

Therefore, we will try to analyze these data so we can get more insights, information, identify patterns etc.

1. Data Visualization

We've used several types of plots in trying to gain insights from the data analyzed related to marital status. Although since marital status is a categorical data variable, we've used more types of plots like boxplots and bar-plots as they are the ones that are mostly used to visualize categorical data type variables.

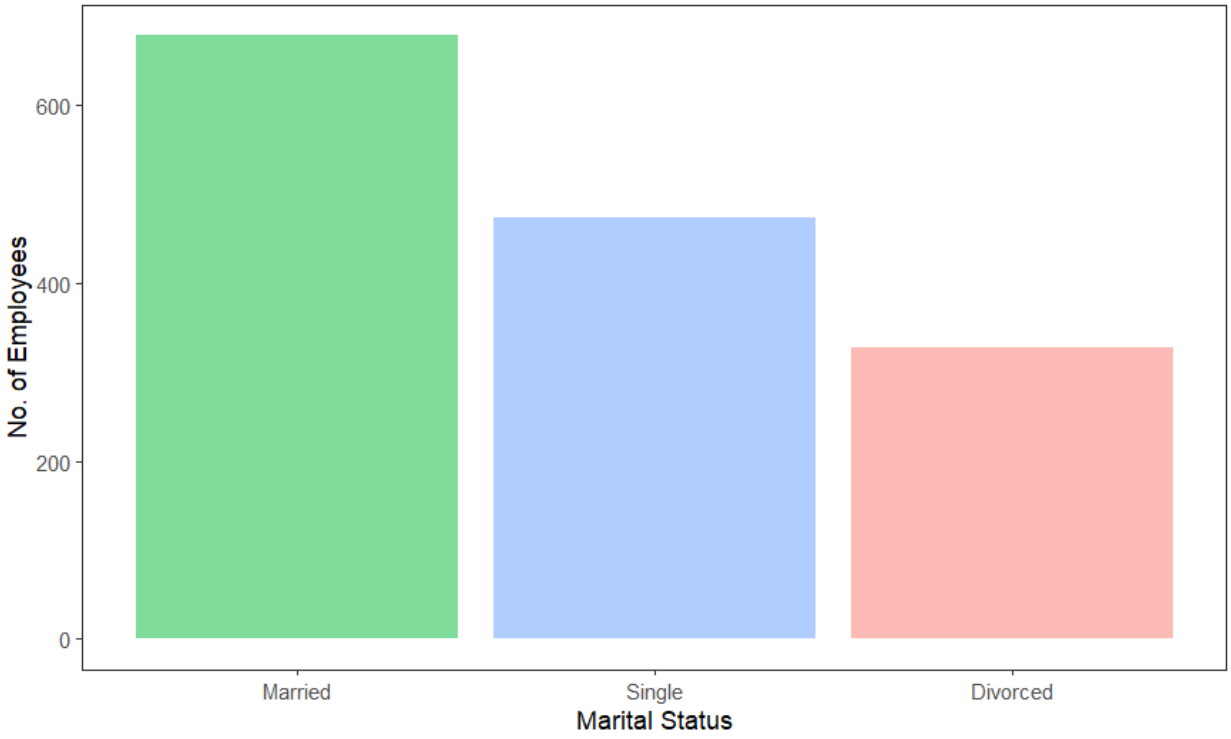
Although we can also convert these categorical variables to numerical data type variables by encoding them into a factor variable and then use mapping to convert them into a numerical variable. The following code can be used for this process but, we haven't used it in data visualization as it wasn't necessary.

```
#ENCODING CATEGORICAL VARIABLE

data$MaritalStatus = factor(data$MaritalStatus, level = c("Single","Divorced","Married"),
                            labels = c("S","D","M"))

MStatus_mapping <- c("S"=0,"D"=1,"M"=2)

#CONVERT THE FACTOR VARIABLE TO NUMERIC USING THE MAPPING

data$NumericMStatus <- MStatus_mapping[data$MaritalStatus]
```

Now we are going to look to gain some insights from the plots we've generated using R. But first we must install the necessary libraries and load the dataset we are using into a data frame in a R script file.

```
library(ggplot2)
library(dplyr)
library(tidyverse)

data <- read.csv("D:\\2nd Year\\Semester 1\\Data Programming in R\\Assignments\\HR_Analytics.csv")

View(data)
str(data)
```

We've used the above code for that process and to view and analyze the structure of the dataset we will be using.



The above bar-plot was created to get an idea of the number of employees based on their marital status.

From this we can see that most of the employees are married (more than 600). There're about 500 employees that are still single. And there're around 300 employees that are divorced.

The following code was used to generate the above plot.

```
data %>%
  drop_na(MaritalStatus) %>%  #getting rid of missing data
  ggplot(aes(fct_infreq(MaritalStatus), fill = MaritalStatus))+
  geom_bar(position= "dodge", alpha = 0.5)+
  theme_bw()+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())+
  labs(x = "Marital Status", y = "No. of Employees")+
  guides(fill = FALSE)
```

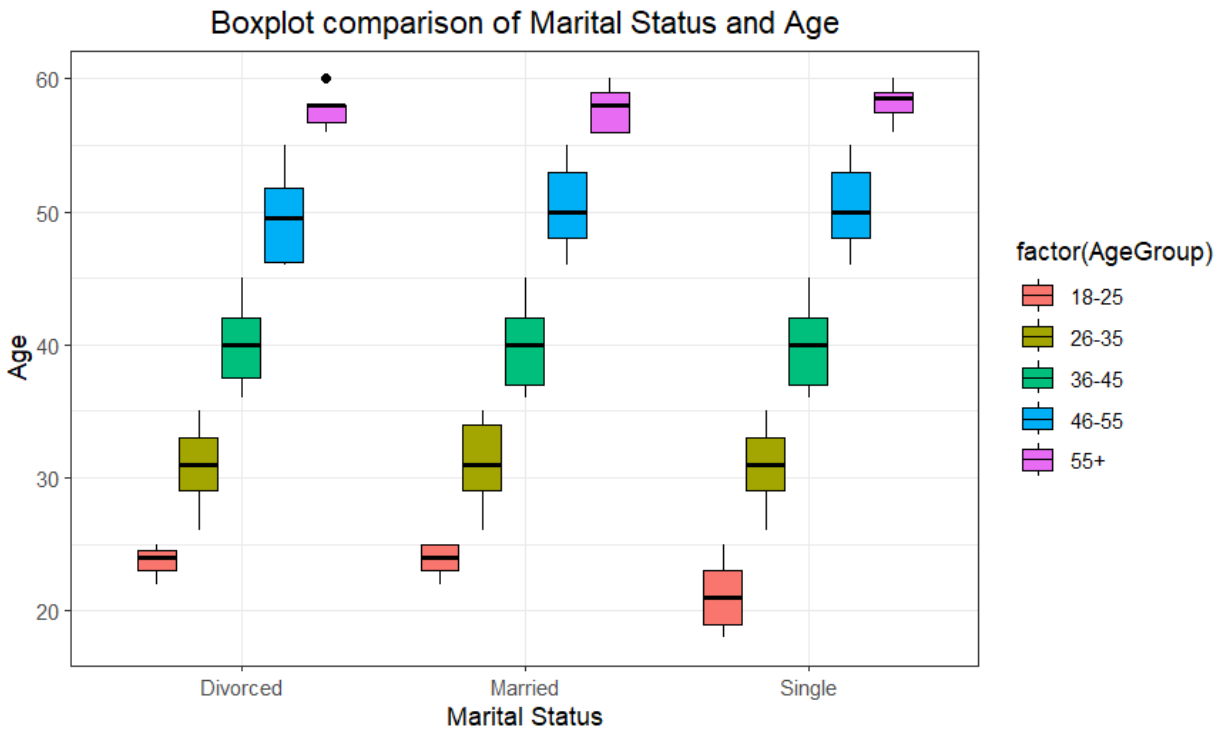
And we've also used the following code to create a table of the percentages of the marital status of the employees.

```
34  percentage_table <- prop.table(table(data$MaritalStatus)) * 100
35  view(percentage_table)
```

The calculated percentages of the marital status of the employees are,

	Var1	Freq
3	Single	31.95946
2	Married	45.87838
1	Divorced	22.16216

1. <u>Marital Status related with Age of the Employees</u>



We can use this boxplot to say a lot of things about employees' ages and marital status.

From this plot as we have factored age groups, we can clearly see that there are a smaller number of people who are divorced between ages 18-25. And there's a larger number of people who are divorced from ages 46 to 55. We can also see there are some existing outliers around the age of 60 which means there are a few people who are still divorced around that age.

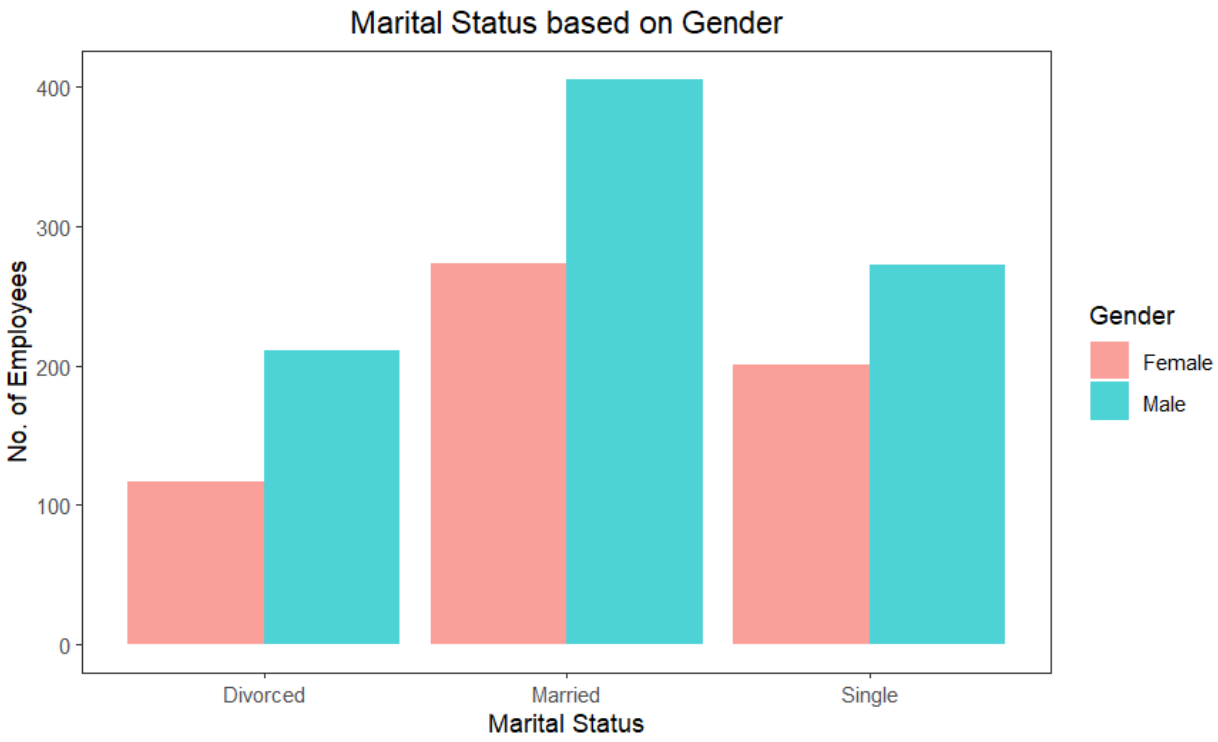
Out of the married employees, we can see there's only a few numbers of employees who are married between the ages of 18-25. And with the rest of the age groups, it's kind of divided evenly.

As for the single employees, we can see there are lots of employees who are still single around the ages 18-25; whereas there were a lesser number of people who were married or divorced in that age group. Which only makes sense as that is the youngest age group and it's more likely that they are mostly single and only a few employees are divorced because that is like the youngest age one would get married. And we can also see how there's only a few numbers of employees who are still single within ages of 46-55, and even more less in 55+ aged employees; it is kind of self-explanatory as it is very unlikely that people are still single around older ages (not that it can't be the case but by nature it is not how it is).

We've used the following code to generate the above plot,

```
data %>%
  ggplot(aes(x = factor(MaritalStatus), y = Age, fill = factor(AgeGroup))) +
  geom_boxplot(color = "black") +
  labs(title = "Boxplot comparison of Marital Status and Age",
       x = "Marital Status",
       y = "Age") +
  theme_bw()+
  theme(
    plot.title = element_text(hjust = 0.5)  # Center the title
    )
```

2. Marital Status based on Gender of the Employees



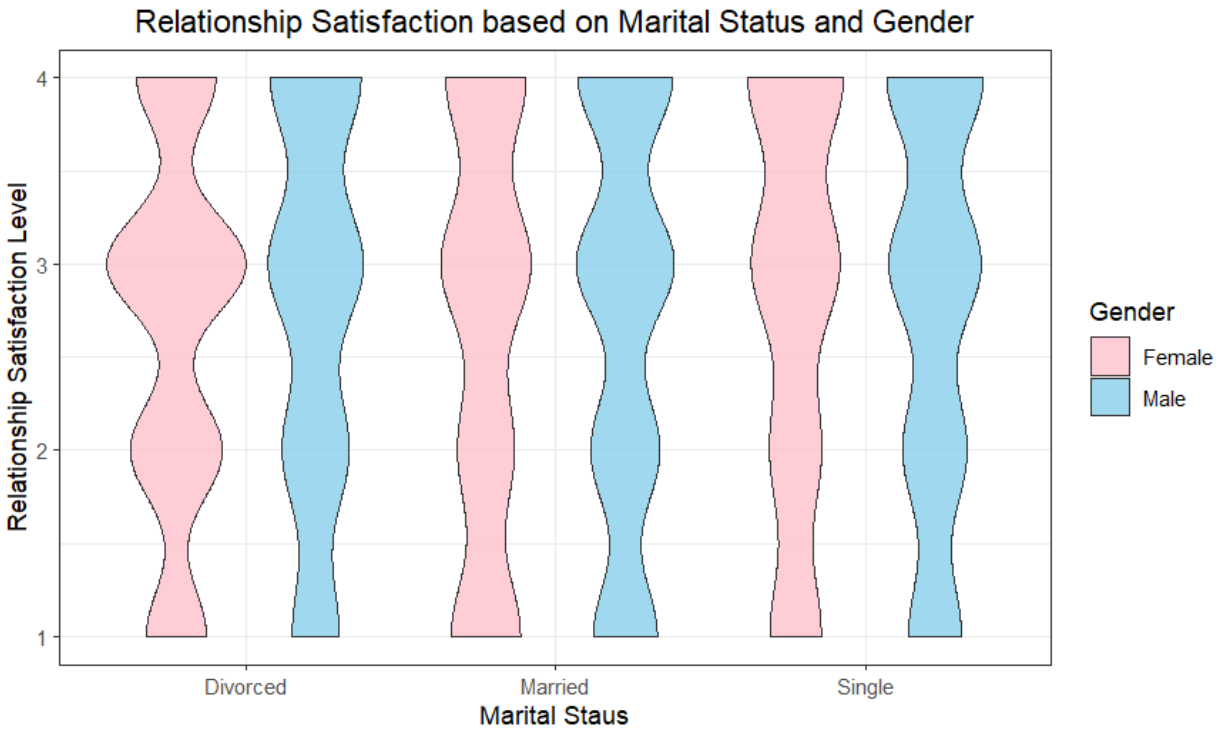
With data visualization using R, we can easily get an understanding of the distribution of marital status of the employees according to their respective genders from the above bar graph.

We can see most of the employees that are married are male employees. Or with any category the number of male employees outnumber the female employees; it would make sense as there's an exceedingly large number of male employees in this HR analytics dataset. However, we can still make observations such as there's a 1:2 ratio between male and female employees respectively who are divorced. We can see there's not that much of a difference in numbers between still single male or female employees.

The following code was used to generate the above bar-plot.

```
data %>%
ggplot(aes(x = MaritalStatus, fill = Gender)) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title = "Marital Status based on Gender",
       x = "Marital Status",
       y = "No. of Employees")+
    theme_bw()+
  theme(
    plot.title = element_text(hjust = 0.5)
    )+
theme(panel.grid.major = element_blank(),
      panel.grid.minor = element_blank())
```

Here are a few other plots we've generated to get an understanding of **Relationship Satisfaction**.

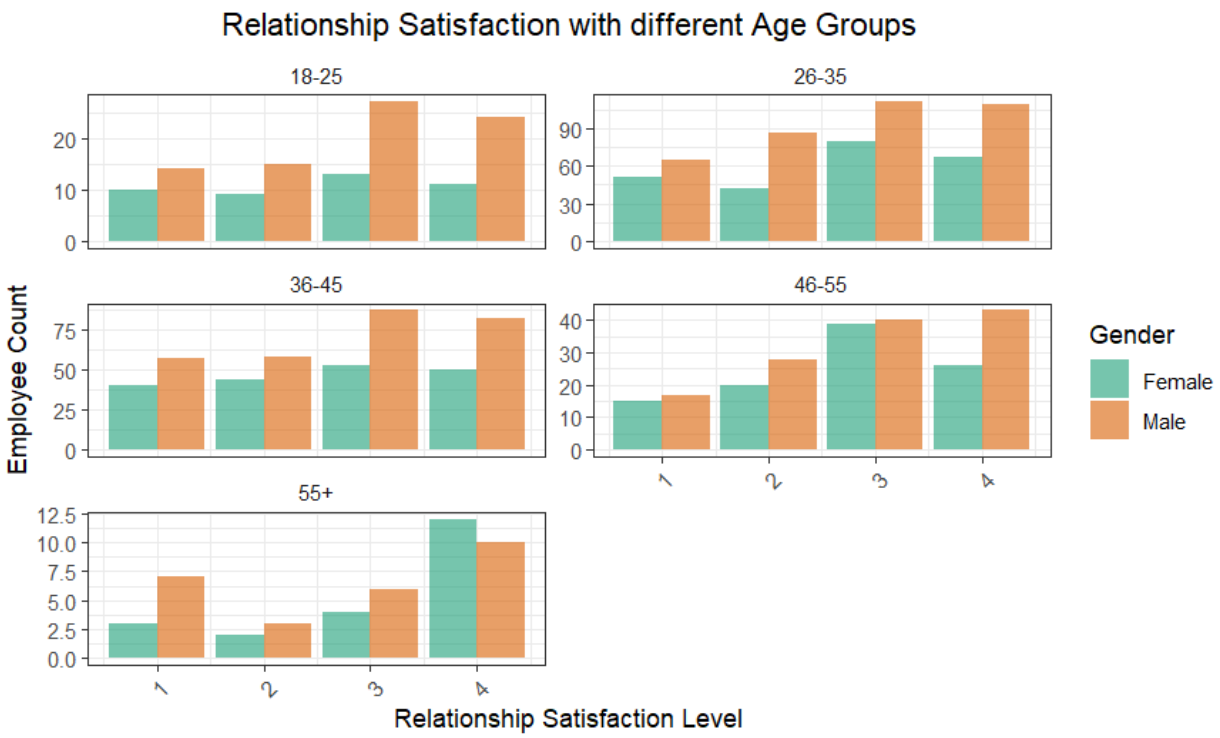


This graph shows us the density of the number of employees with their relative Relationship Satisfaction level with gender as a factor. We can see that there're a lot of divorced female employees with a Relationship Satisfaction level of 3. There're only a few observations like that you can make from this graph.

We've used the following code to generate the above violin plot.

```
data %>%
ggplot(aes(x = MaritalStatus, y = RelationshipSatisfaction, fill = Gender)) +
  geom_violin(draw_quantiles = TRUE, alpha = 0.8) +
  labs(title = "Relationship Satisfaction based on Marital Status and Gender",
       x = "Marital Staus",
       y = "Relationship Satisfaction")+
  scale_fill_manual(values = c("pink","skyblue")) +
  theme_bw()+
theme(
  plot.title = element_text(hjust = 0.5)
)
```

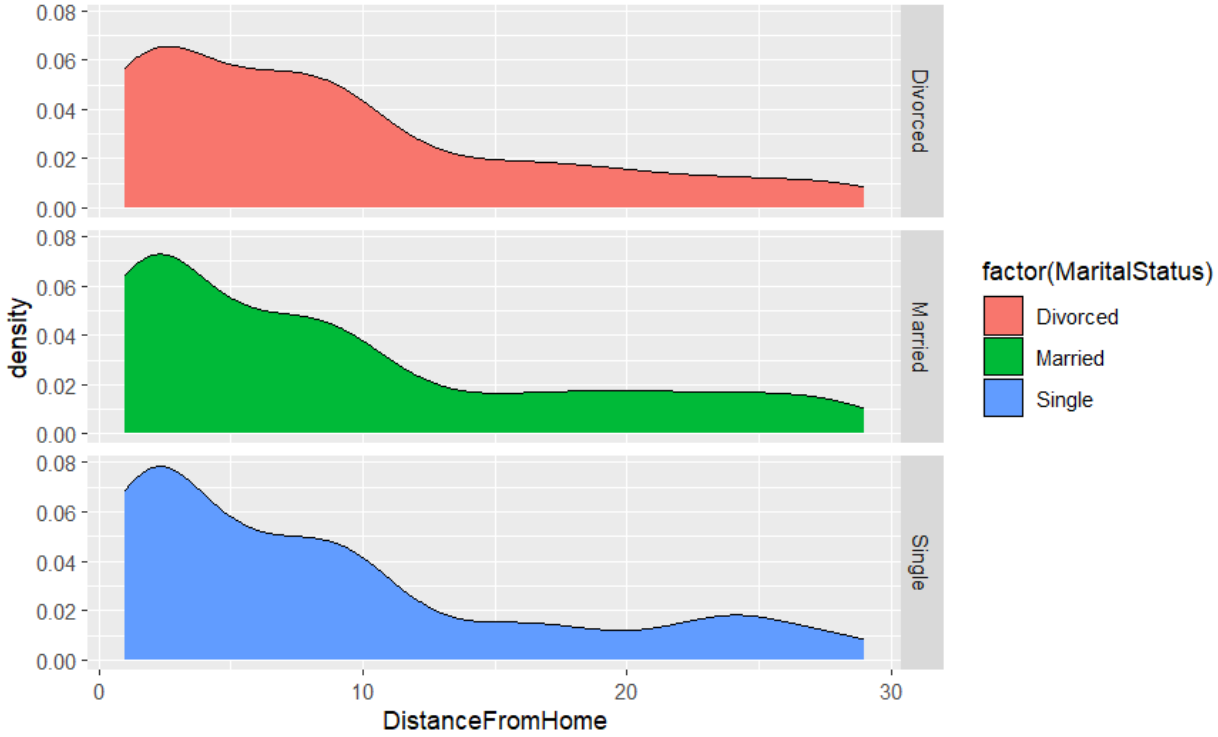
The following graph shows us the distribution the of number of employees in different age groups according to their relationship satisfaction level with gender also as a factor.



The following code was used to generate the above bar graphs.

```
data %>%
  ggplot(aes(x = RelationshipSatisfaction, fill= Gender)) +
  geom_bar(position = "dodge", alpha = 0.6) +
  facet_wrap(~ AgeGroup, scales = "free_y", ncol = 2) +  # Customize the number of columns and scales
  labs(title = "Relationship Satisfaction by Age Groups with Gender factor",
       x = "Relationship Satisfaction",
       y = "Employee Count") +
  theme_bw() +
  theme(
    plot.title = element_text(hjust = 0.5),
    strip.background = element_blank(),
    strip.text = element_text(size = 9),  # Customize facet label appearance
    axis.text.x = element_text(angle = 45, hjust = 1)  # Rotate x-axis labels
  )+
  scale_fill_brewer(palette = "Dark2")
```

The following plot is a density plot between the Distance to Work from Home of the employees and their Marital Status.



2. Data Analytics through models

1. Naïve Bayes model

Naïve bayes is a probabilistic machine learning algorithm based on Bayes' theorem. It assumes that features are conditionally independent given the class label, which simplifies computations. So, we are going to use this model to calculate the predictability of the variable – Marital Status.

```r
library(caTools)
library(e1071)
library(lattice)
library(caret)


data1 <- read.csv("D:\\2nd Year\\Semester 1\\Data Programming in R\\Assignments\\HR_Analytics.csv")

str(data1)

data1$AgeGroup <-as.integer(factor(data1$AgeGroup))
data1$Attrition <-as.integer(factor(data1$Attrition))
data1$BusinessTravel <-as.integer(factor(data1$BusinessTravel))
data1$Department <-as.integer(factor(data1$Department))
data1$EducationField <-as.integer(factor(data1$EducationField))
data1$Gender <-as.integer(factor(data1$Gender))
data1$JobRole <-as.integer(factor(data1$JobRole))
data1$MaritalStatus <-as.integer(factor(data1$MaritalStatus))
data1$SalarySlab <-as.integer(factor(data1$SalarySlab))
data1$Over18 <-as.integer(factor(data1$Over18))
data1$OverTime <-as.integer(factor(data1$OverTime))

str(data1)

data1 <- na.omit(data1)
sum(is.na(data1))

data1$EmpID <- NULL
data1$Over18 <- NULL

set.seed(900)
split <- sample.split(data1$MaritalStatus, SplitRatio = 0.75)
train <- subset(data1, split == TRUE)
test  <- subset(data1, split == FALSE)


classifier <- naiveBayes ( MaritalStatus~., data = train)
pre <- predict(classifier, newdata = test)

contable <- table(pre,test$MaritalStatus)
contable

confusionMatrix(contable)
```

First, we install the necessary libraries and load the dataset we are going to use. Then we identify the structure of the dataset and convert all the categorical variables into factorial variables and then convert them to integer variables.

Then we get rid of missing values in the dataset and check to see if there are any left.

Then we get rid of the columns which shouldn't be a part of the naïve bayes model.

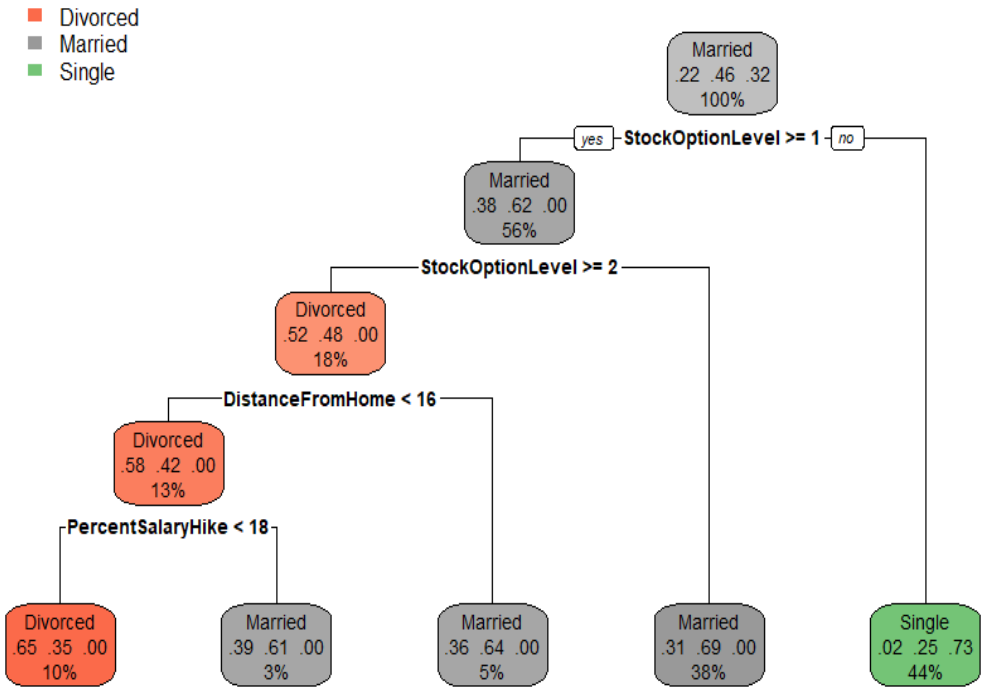
We get the following output from the above code.

```
> confusionMatrix(contable)
Confusion Matrix and Statistics


pre   1   2   3
  1  23  20   0
  2  55 113   1
  3   1  31 112

Overall Statistics

               Accuracy : 0.6966
                 95% CI : (0.646, 0.744)
    No Information Rate : 0.4607
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5155

 Mcnemar's Test P-Value : 7.394e-10

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.29114   0.6890   0.9912
Specificity           0.92780   0.7083   0.8683
Pos Pred Value        0.53488   0.6686   0.7778
Neg Pred Value        0.82109   0.7273   0.9953
Prevalence            0.22191   0.4607   0.3174
Detection Rate        0.06461   0.3174   0.3146
Detection Prevalence  0.12079   0.4747   0.4045
Balanced Accuracy     0.60947   0.6987   0.9297
```

From this, we can say that we can predict the outcome of the variable Marital Status with an accuracy of 0.6966 (≈ 0.7).

2. Decision Tree Classification on Marital Status

A machine learning approach known as a decision tree classification creates a tree-like structure by iteratively dividing the data according to its attributes. Based on a particular feature, a decision is taken at each node of the tree, resulting in a branch that eventually generates a classification or prediction for a given input. So, we are going to generate a decision tree for the variable – Marital Status and see what kind of information we can gather.



The following code was used to obtain the above decision tree and create a naïve bayes model.

```
library(caret)
library(rpart)
library(rpart.plot)
library(tidyverse)
library(skimr)
library(caTools)
library(e1071)

data <- read.csv("D:\\2nd Year\\Semester 1\\Data Programming in R\\Group Project\\HR_Analytics.csv")
str(data)

data$EmpID <- NULL
data$Over18 <- NULL
data$StandardHours <- NULL
data$EmployeeCount <- NULL
data$JobLevel <- NULL

split <- createDataPartition(y=data$MaritalStatus , p = 0.75, list = FALSE)
train <- data[split, ]
test <- data[-split, ]

dim(train)
dim(test)

set.seed(900)

dec_tree <- rpart(formula = MaritalStatus   ~.,
                  data = train,
                  method = "class",
                  xval = 10)

rpart.plot(dec_tree, yesno = TRUE)

naiveclassifier <- naiveBayes(MaritalStatus   ~ ., data=train)
predresults <- predict(naiveclassifier , newdata = test)

conftable <- table(predresults, test$MaritalStatus  )
conftable

confusionMatrix(conftable)
```

This was the output we got from the naïve bayes model we trained under it.

```
> confusionMatrix(conftable)
Confusion Matrix and Statistics


predresults Divorced Married Single
   Divorced       15      21      0
   Married        67     115      0
   Single          0      33    118

Overall Statistics

               Accuracy : 0.6721
                 95% CI : (0.6216, 0.7198)
    No Information Rate : 0.458
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4724

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Divorced Class: Married Class: Single
Sensitivity                  0.18293         0.6805        1.0000
Specificity                  0.92683         0.6650        0.8685
Pos Pred Value               0.41667         0.6319        0.7815
Neg Pred Value               0.79880         0.7112        1.0000
Prevalence                   0.22222         0.4580        0.3198
Detection Rate               0.04065         0.3117        0.3198
Detection Prevalence         0.09756         0.4932        0.4092
Balanced Accuracy            0.55488         0.6727        0.9343
```

3.Findings and Discussion

*They are written under the relevant plots and topics.

6. Appendices

Git hub repository link (Full project) –

https://github.com/scssandanayake/R-coursework