```
#####Importing the Dataset########
emp_churn <- read.csv("C:/Users/Amisha Sancheti/Desktop/MITA sem2/Multivariate Analysis/Project/WA_Fn-Us
```

```
##########Exploring the Dataset############
str(emp_churn)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ ï..Age                  : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 3 2 3 2 3 3
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
##  $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 2 1 2 2 2 ...
##  $ HourlyRate              : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...
##  $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
##  $ MonthlyRate             : int  19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...
##  $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
##  $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

```
library(data.table)
setDT(emp_churn)
class(emp_churn)
```

```
## [1] "data.table" "data.frame"
```

```
table(is.na(emp_churn)) ##The output is false, hence we don't have any null values in our data.
```

```
##
```

```
## FALSE
## 51450
```

```
####### Now we will look for erroneous data in our table, column wise.########

unique(emp_churn$Attrition) #the output is yes and no. There is no other deformed value.
```

```
## [1] Yes No
## Levels: No Yes
```

```
unique(emp_churn$BusinessTravel) #the output is 'Non-Travel, Travel_Frequently, Travel_Rarely'. There i
```

```
## [1] Travel_Rarely     Travel_Frequently Non-Travel
## Levels: Non-Travel Travel_Frequently Travel_Rarely
```

```
#####Like wise, our data has only integer values and factors with defined labels in accordance with mil

##### For th EDA
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(ggpubr)
```

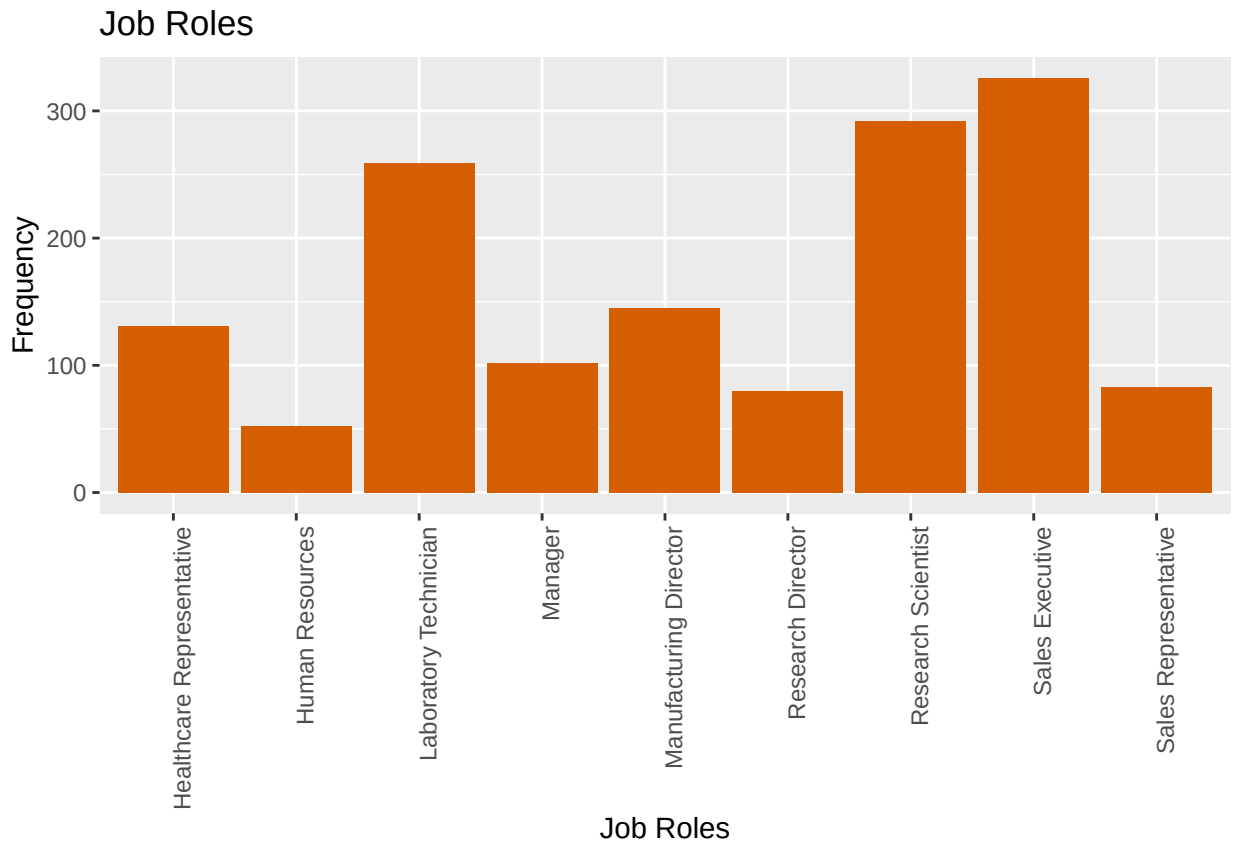
```
## Loading required package: magrittr
```

```
#Extracting JobRoles
x = table(emp_churn$JobRole)
```

```
#Converting to dataframe
x1 = as.data.frame(x)
x1
```

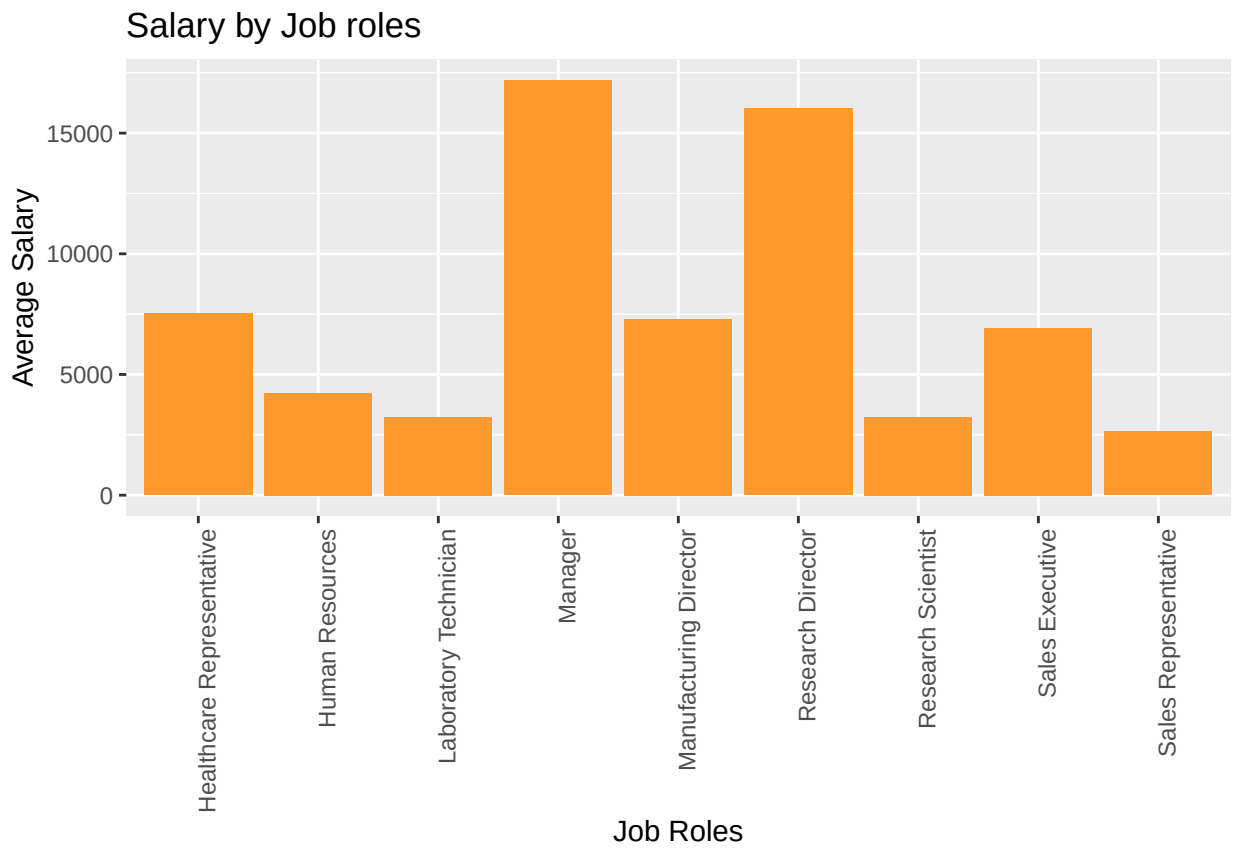
```
##                        Var1 Freq
## 1 Healthcare Representative  131
## 2           Human Resources   52
## 3     Laboratory Technician  259
## 4                   Manager  102
## 5    Manufacturing Director  145
## 6         Research Director   80
## 7        Research Scientist  292
## 8           Sales Executive  326
## 9      Sales Representative   83
```

```
#plotting the barplot
ggplot(x1, aes(x=Var1, y=Freq)) + geom_bar(stat="identity",fill="#D55E00") +
  labs(x="Job Roles", y="Frequency", title="Job Roles")+theme(axis.text.x = element_text(angle = 90, hj
```



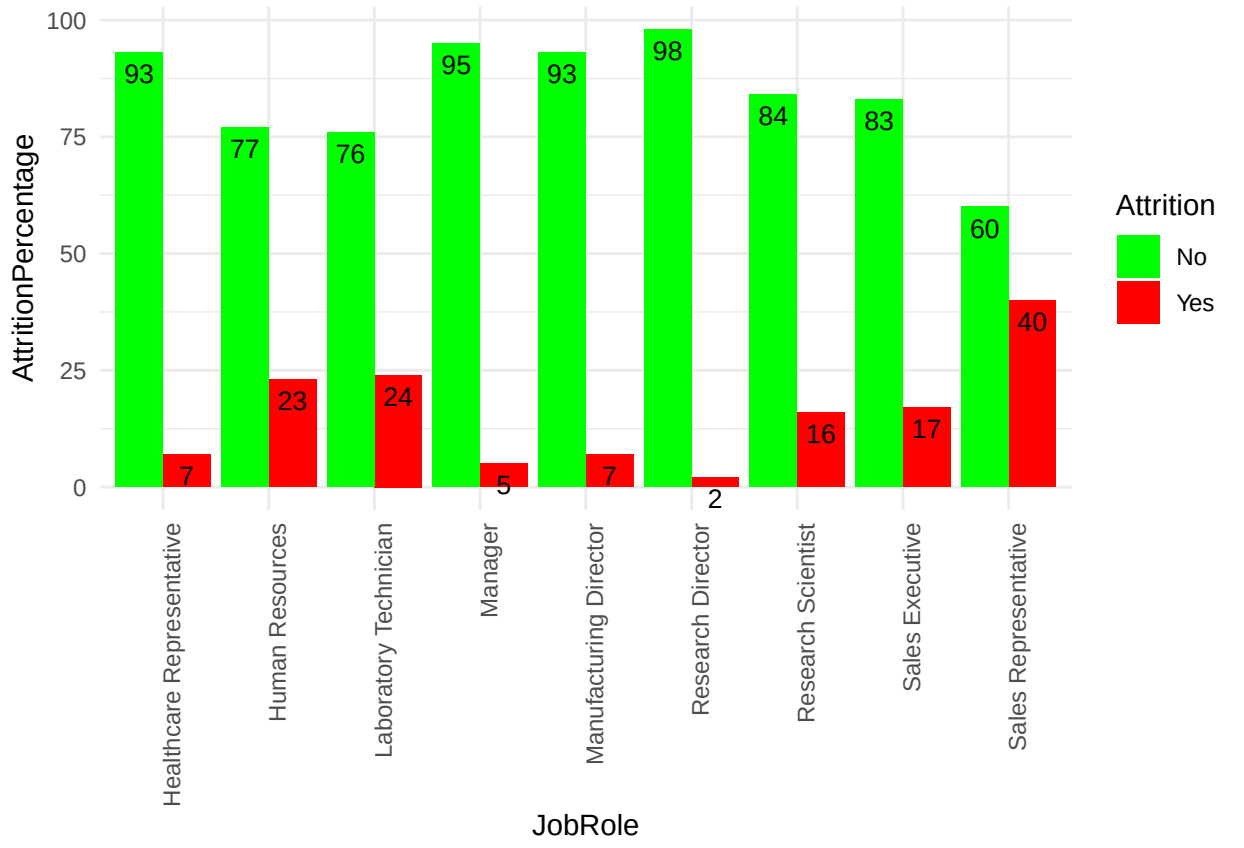

```
#Extracting the average salary by jobroles
job_sal = emp_churn %>% select(JobRole, MonthlyIncome) %>% group_by(JobRole) %>% summarize(avg=mean(Mon
```

```
#Converting to dataframe
x2 = as.data.frame(job_sal)

#Barplot
ggplot(x2, aes(x=JobRole, y=avg)) + geom_bar(stat="identity",fill="#FE9A2E") +
  labs(x="Job Roles", y="Average Salary", title="Salary by Job roles")+theme(axis.text.x = element_text
```



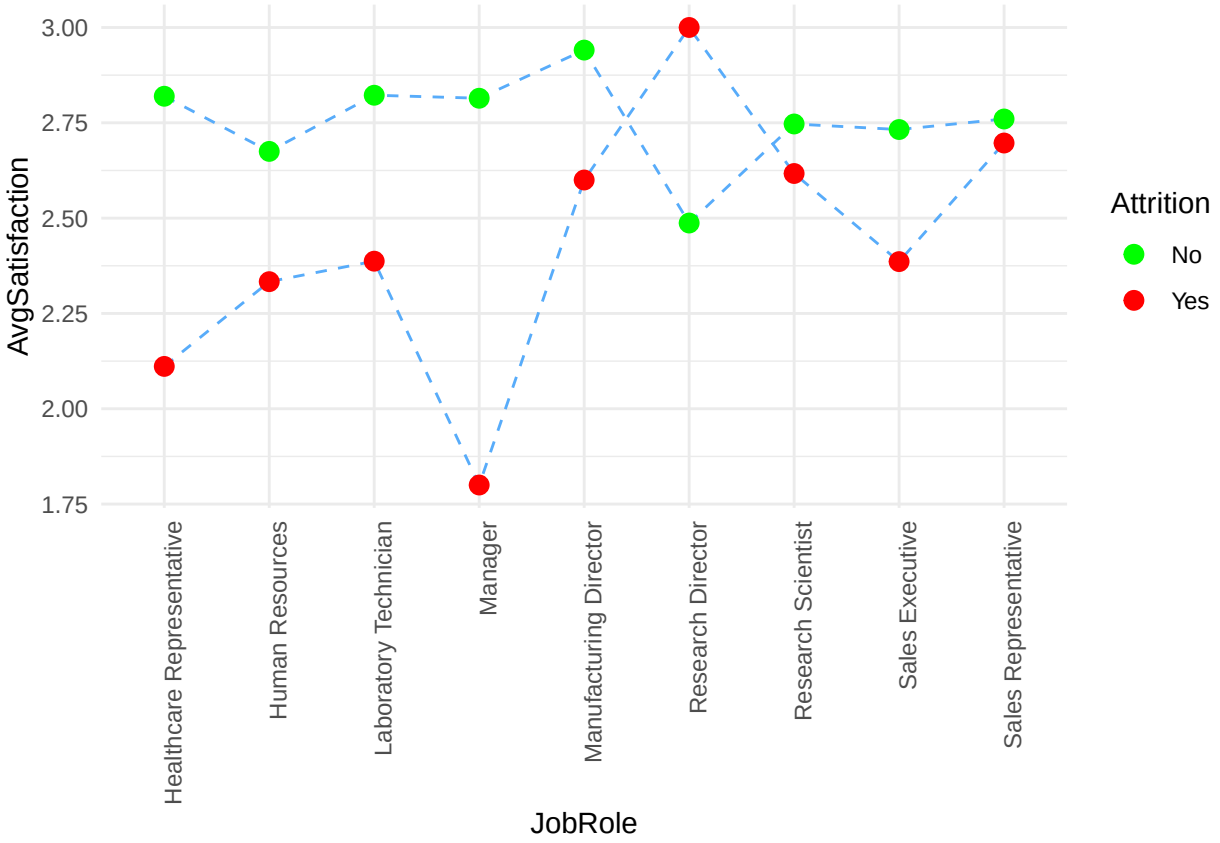

```
#Managers and Research directors have a very higher salary


#Extracting the attrition by job roles
attr_job <- emp_churn %>% select(JobRole, Attrition) %>% group_by(JobRole, Attrition) %>% summarize(amou
  mutate(AttritionPercentage=round(prop.table(amount),2) * 100) %>% arrange(AttritionPercentage)
```

```
#Converting to dataframe
x3 = as.data.frame(attr_job)

#Barplot
ggplot(data=x3, aes(x=JobRole, y=AttritionPercentage, fill=Attrition)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=AttritionPercentage), vjust=1.6, color="black", position = position_dodge(0.9), si
  theme_minimal() + scale_fill_manual(values=c("green","red")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
#Sales representatives show a higher attrition rate

#Extracting attrition by job role and environment satisfaction
env_attr <- emp_churn %>% select(EnvironmentSatisfaction, JobRole, Attrition) %>% group_by(JobRole, Att
  summarize(AvgSatisfaction=mean(EnvironmentSatisfaction))
```

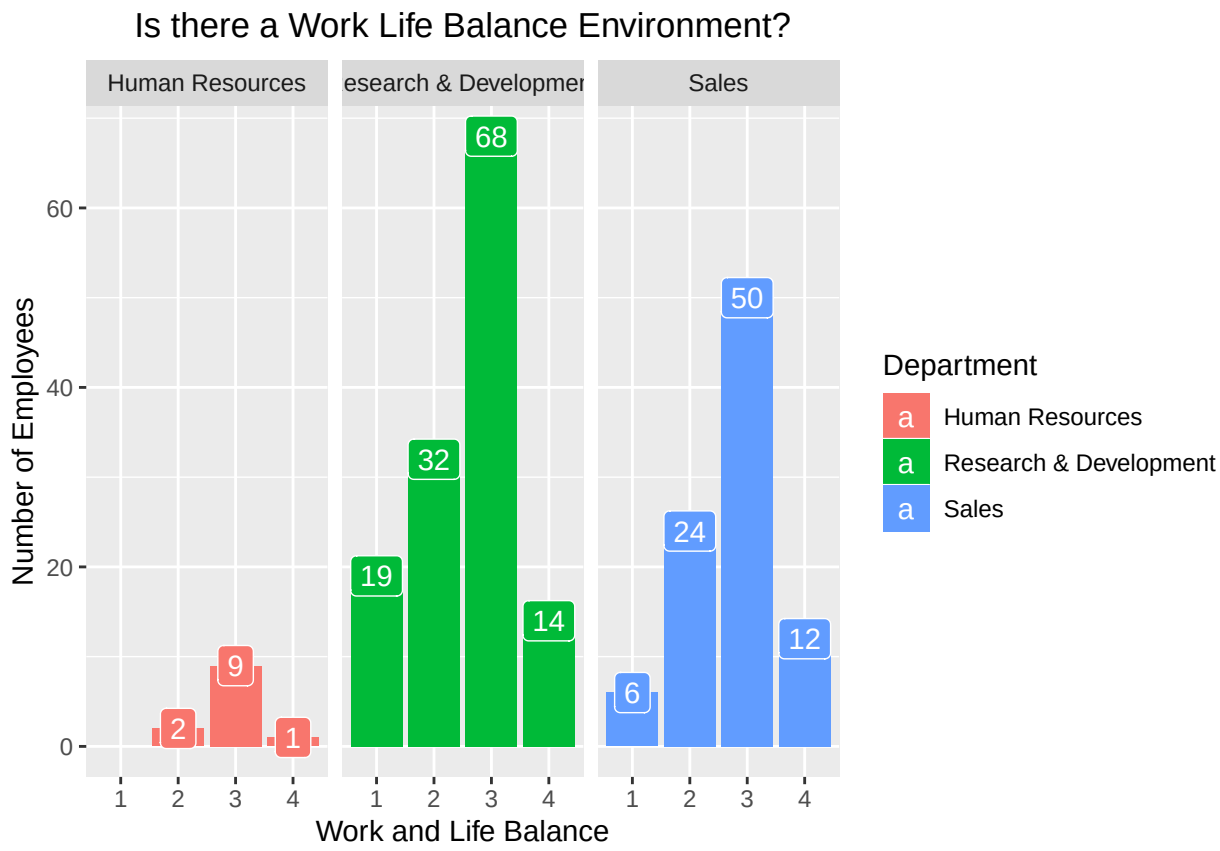
```
#Converting to dataframe
x4 = as.data.frame(env_attr)

#Lineplot
ggplot(data=x4, aes(x=JobRole, y=AvgSatisfaction, fill=Attrition)) +
  geom_line(aes(group=Attrition), color="#58ACFA", linetype="dashed") +
  geom_point(aes(color=Attrition), size=3) +
  theme_minimal() + scale_color_manual(values=c("green","red")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
#It is quite evident that employess with less job satisfaction have high attrition rate


#Filtering employess who left
attritions <- emp_churn %>% filter(Attrition == "Yes")
```

```
#Extracting the employess who left along with worklife balance
attritions$WorkLifeBalance <- as.factor(attritions$WorkLifeBalance)
```

```
#Barplot
attr_wlb_dpt <- attritions %>% select(Department, WorkLifeBalance) %>% group_by(Department, WorkLifeBala
  summarize(count=n()) %>%
  ggplot(aes(x=WorkLifeBalance, y=count, fill=Department)) + geom_bar(stat='identity') + facet_wrap(~Dep
  theme(plot.title=element_text(hjust=0.5)) +
  scale_color_manual(values=c("Pink", "Orange", "Blue")) +
  geom_label(aes(label=count, fill = Department), colour = "white") +
  labs(title="Is there a Work Life Balance Environment?", x="Work and Life Balance", y="Number of Emplo

attr_wlb_dpt
```

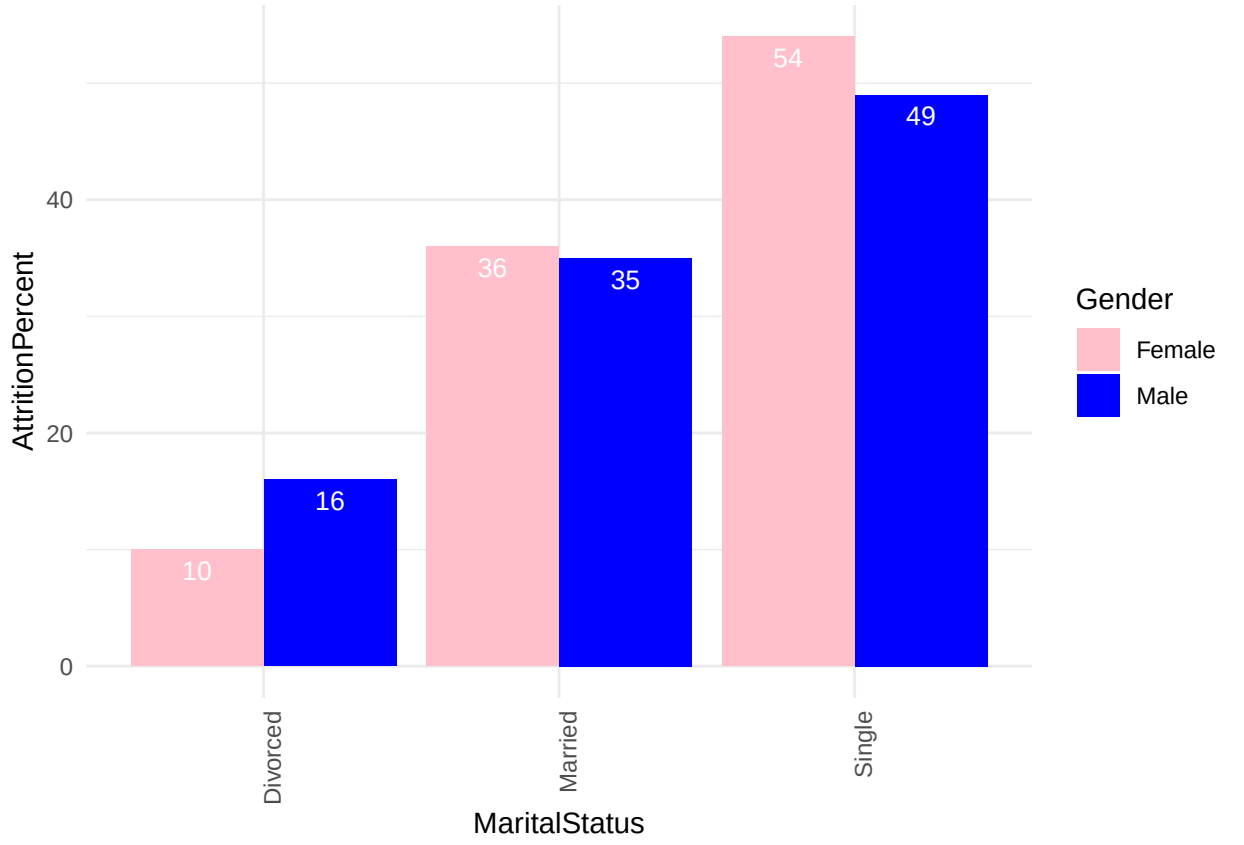



```
#Worklife balance is not a major reason for employee attrition


#Extracting employees who left by gender and marital status
attr_mrg_gdr <- attritions %>% select(Gender, MaritalStatus) %>% group_by(Gender, MaritalStatus) %>%
  summarize(countn=n())%>%
  mutate(AttritionPercent=round(prop.table(countn),2) * 100) %>% arrange(AttritionPercent)
attr_mrg_gdr
```

```
## # A tibble: 6 x 4
## # Groups:   Gender [2]
##   Gender MaritalStatus countn AttritionPercent
##   <fct>  <fct>          <int>             <dbl>
## 1 Female Divorced           9                10
## 2 Male   Divorced          24                16
## 3 Male   Married           53                35
## 4 Female Married           31                36
## 5 Male   Single            73                49
## 6 Female Single            47                54
```

```
#Converting to dataframe
x5 = as.data.frame(attr_mrg_gdr)

#Barplot
ggplot(data=x5, aes(x=MaritalStatus, y=AttritionPercent, fill=Gender)) +
```

```
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=AttritionPercent), vjust=1.6, color="white", position = position_dodge(0.9), size=
  theme_minimal() + scale_fill_manual(values=c("pink","blue")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



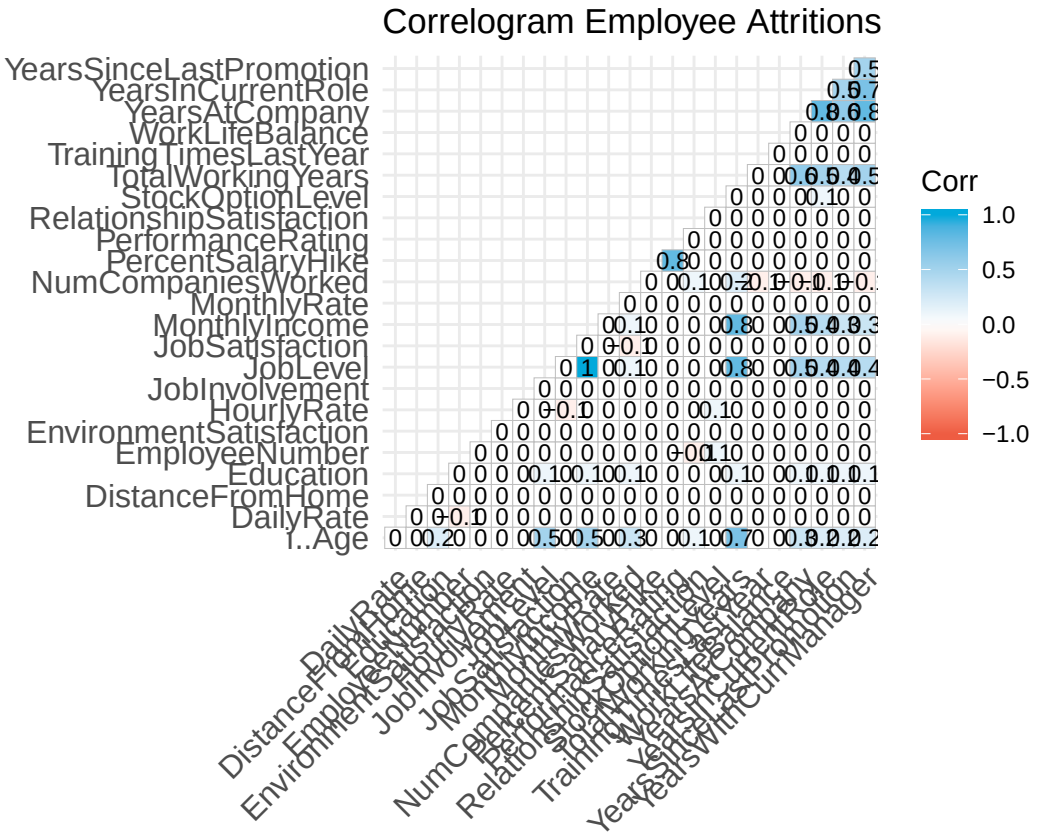

```
#Correlation plot
library(corrplot)
```

```
## corrplot 0.84 loaded
```

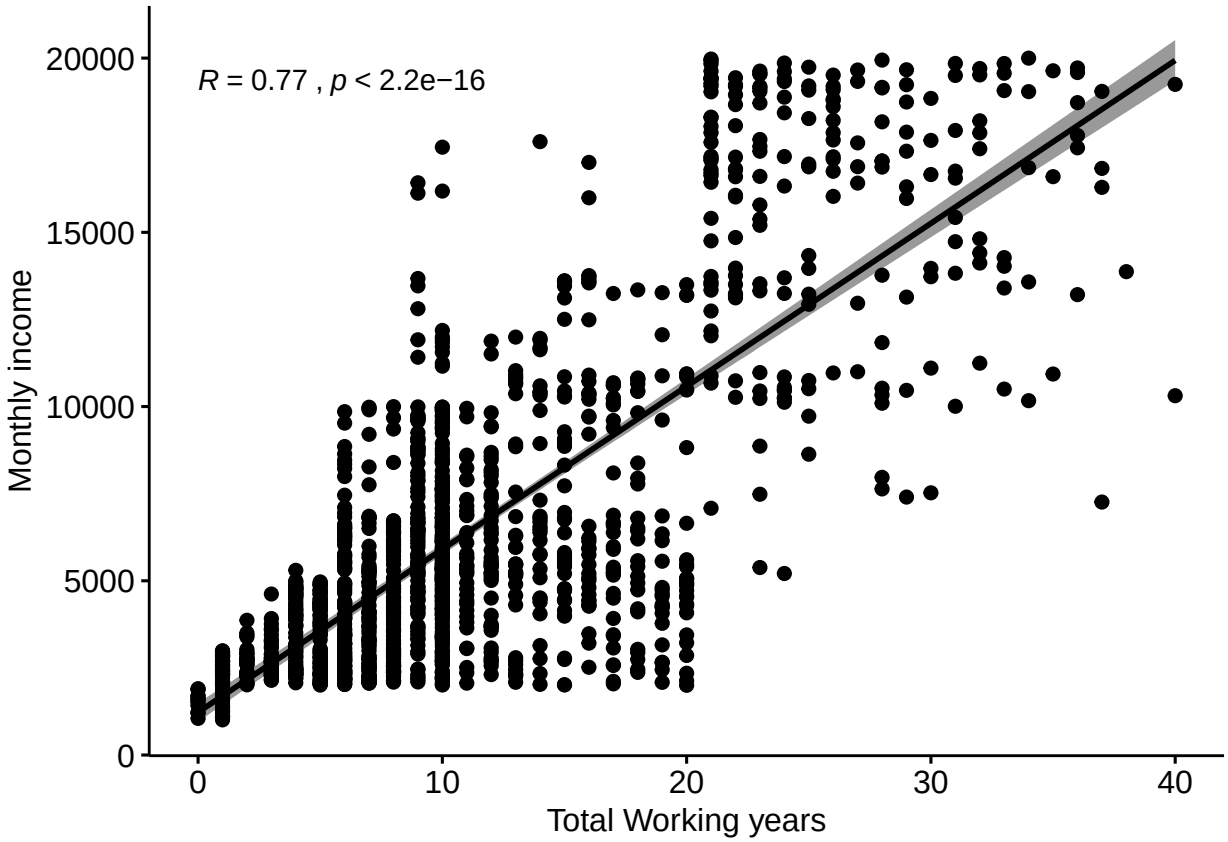
```
library(ggcorrplot)

nums <- select_if(emp_churn, is.numeric)
corr <- round(cor(nums), 1)
```

```
## Warning in cor(nums): the standard deviation is zero
```

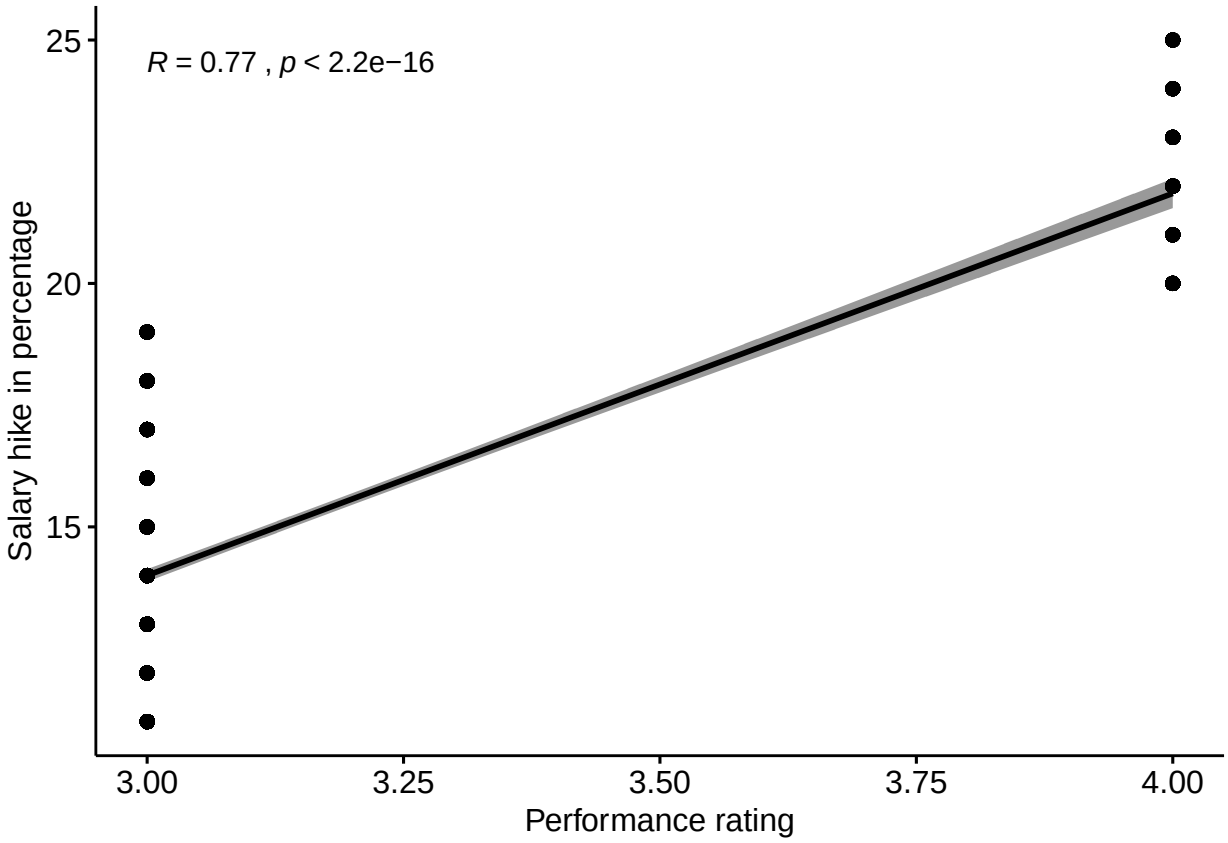
```
ggcorrplot(corr, type = "lower", lab = TRUE, lab_size = 3, colors = c("tomato2", "white", "#01A9DB"),
           title="Correlogram Employee Attritions", ggtheme=theme_minimal())
```



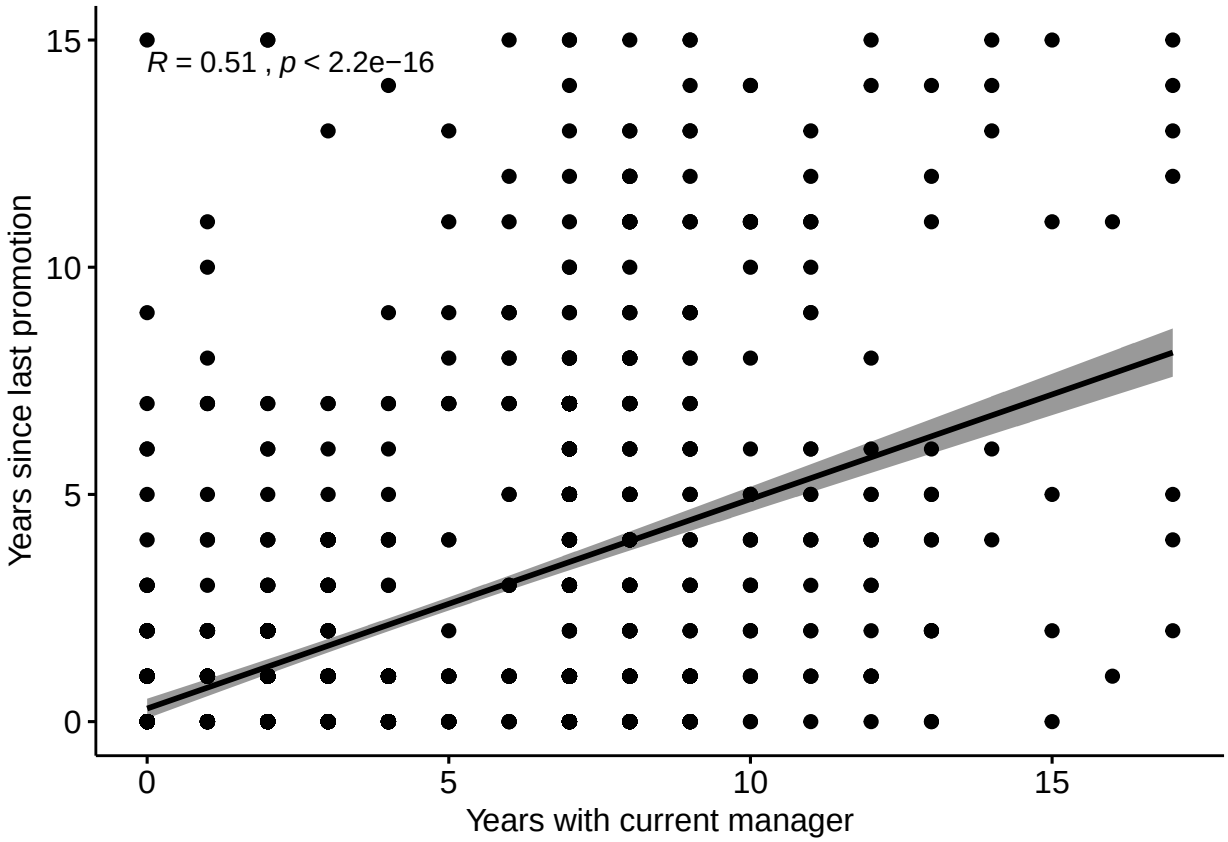

```r
#Bivariate Analysis
library(ggpubr)

ggscatter(emp_churn, x = "TotalWorkingYears", y = "MonthlyIncome",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Total Working years", ylab = "Monthly income")
```



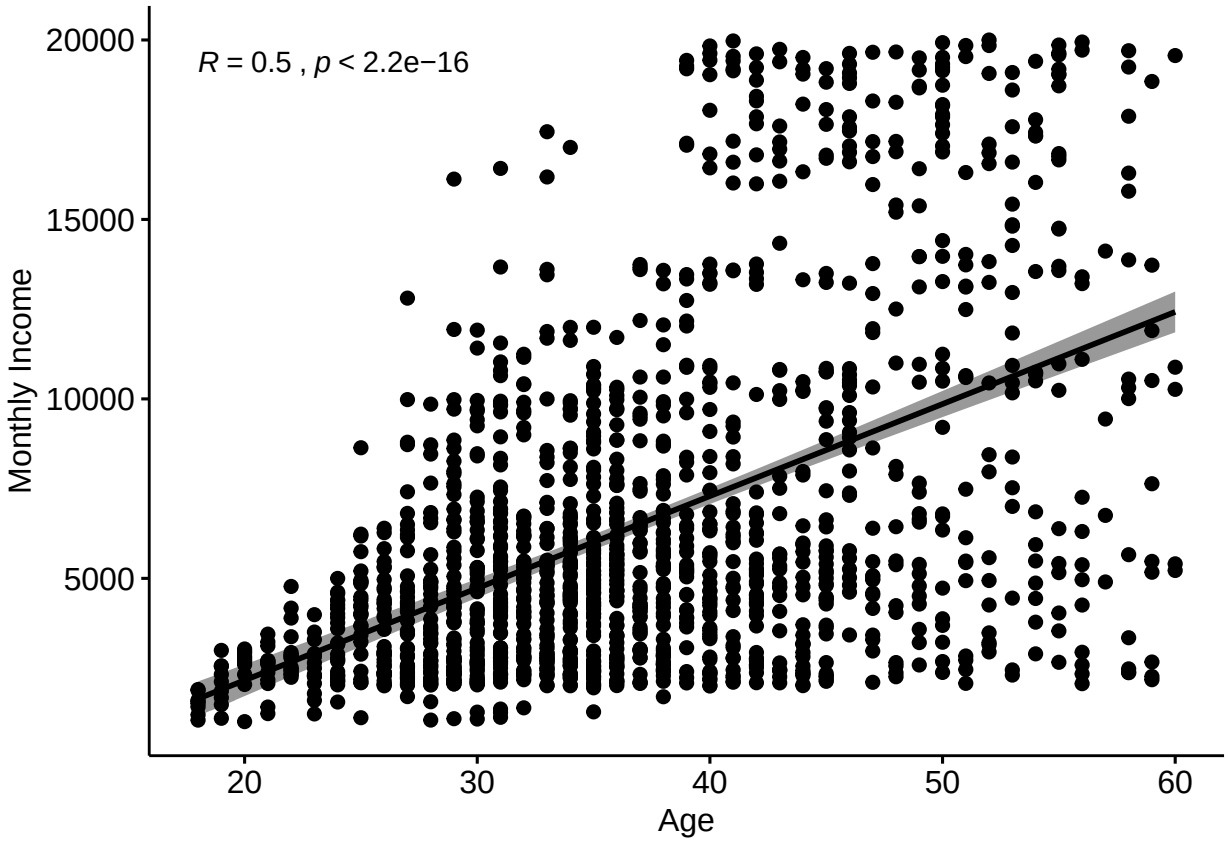

```
ggscatter(emp_churn, x = "PerformanceRating", y = "PercentSalaryHike",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",xlab = "Performance rating", ylab = "Salary hike in p
```




```
ggscatter(emp_churn, x = "YearsWithCurrManager", y = "YearsSinceLastPromotion",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Years with current manager", ylab = "Years since last promotion")
```

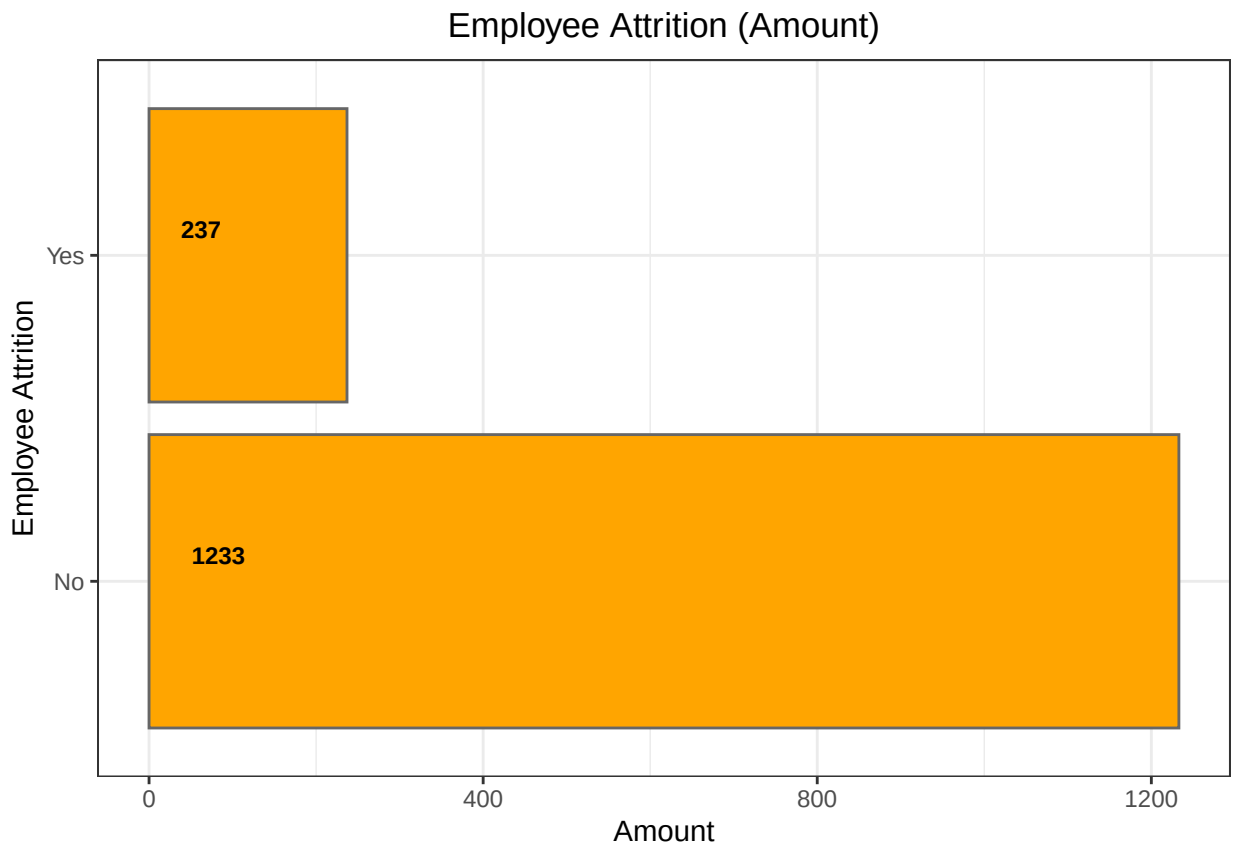



```
ggscatter(emp_churn, x = "ï..Age", y = "MonthlyIncome",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Age", ylab = "Monthly Income")
```

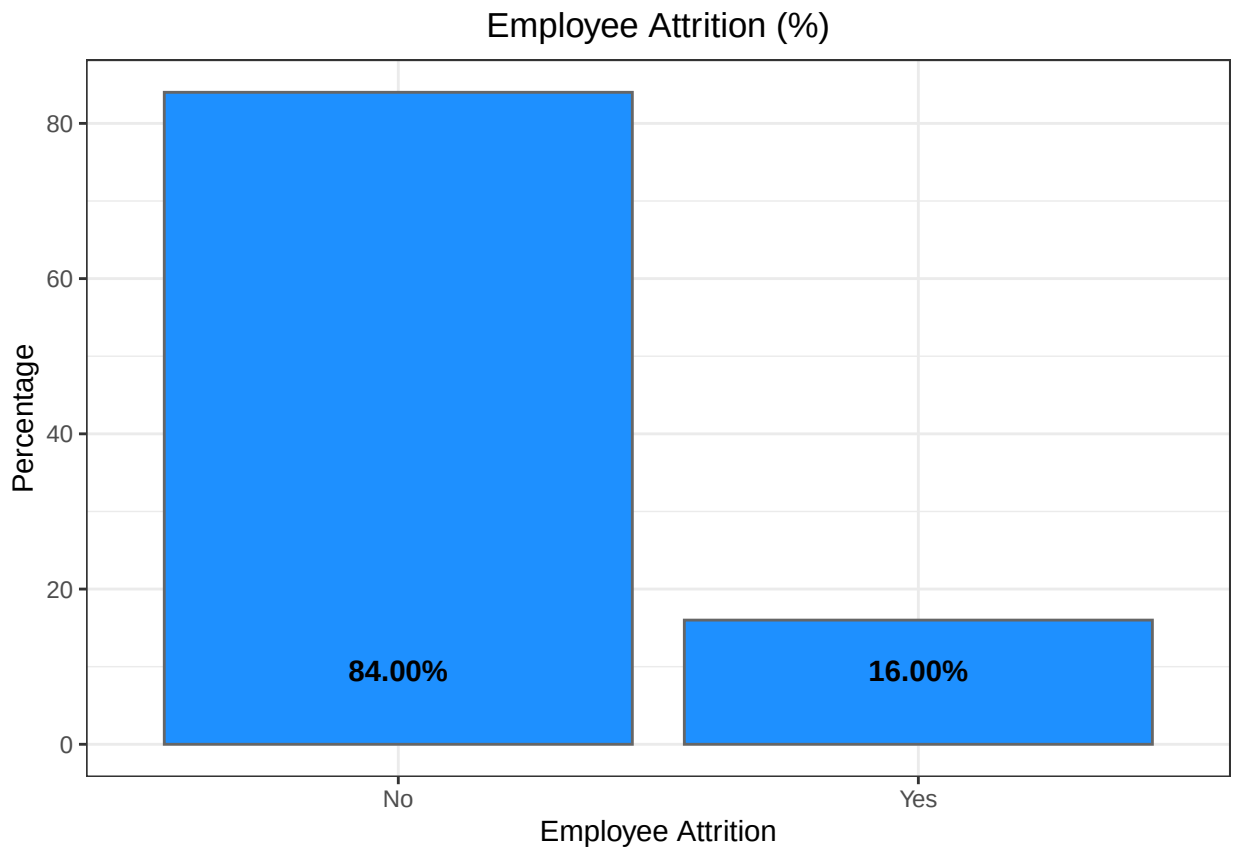



```
library(lattice)
library(ggplot2)
library(dplyr)
```

```
options(repr.plot.width=8, repr.plot.height=4)

attritions_number <- emp_churn %>% group_by(Attrition) %>% summarise(Count=n()) %>%
  ggplot(aes(x=Attrition, y=Count)) + geom_bar(stat="identity", fill="orange", color="grey40") + theme_
  geom_text(aes(x=Attrition, y=0.01, label= Count),
            hjust=-0.8, vjust=-1, size=3,
            colour="black", fontface="bold",
            angle=360) + labs(title="Employee Attrition (Amount)", x="Employee Attrition",y="Amount") +
attritions_number
```



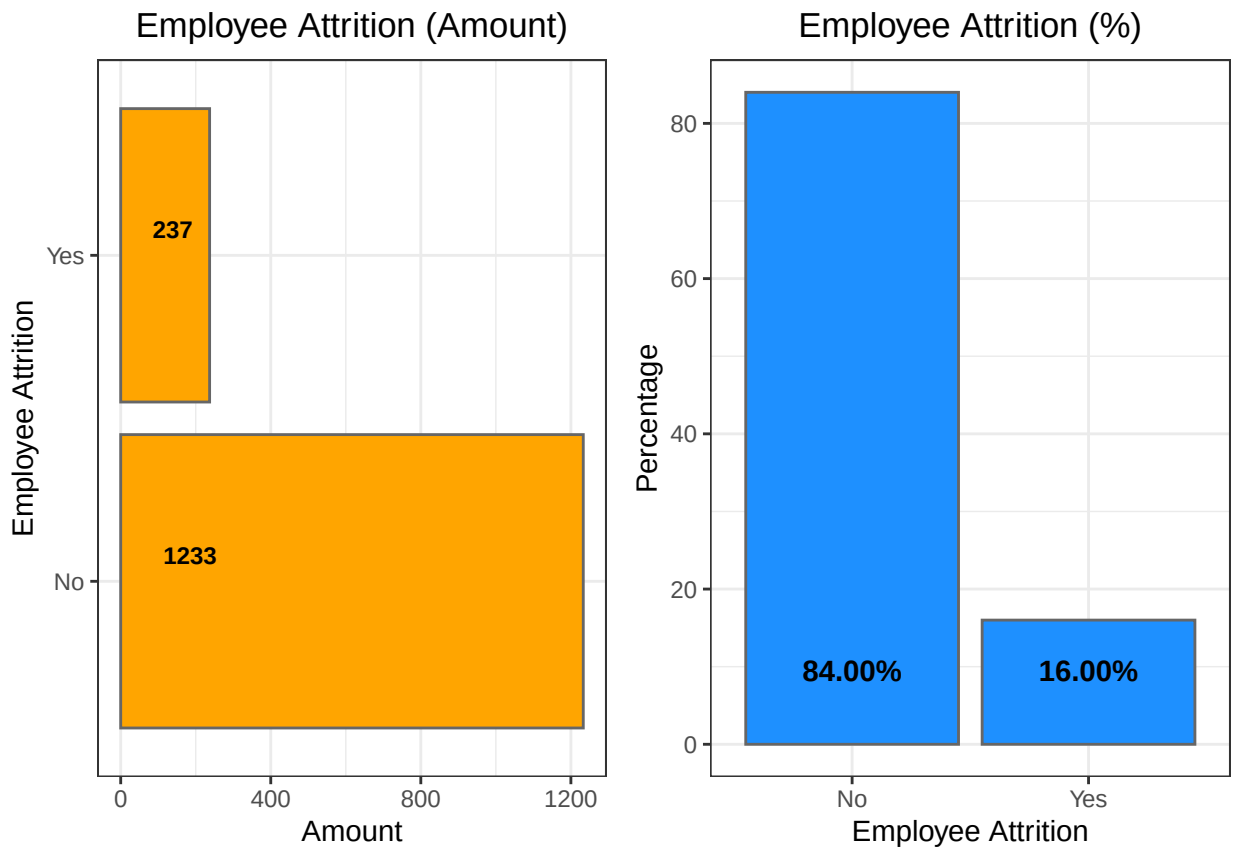

```
attrition_percentage <- emp_churn %>% group_by(Attrition) %>% summarise(Count=n()) %>%
  mutate(pct=round(prop.table(Count),2) * 100) %>%
  ggplot(aes(x=Attrition, y=pct)) + geom_bar(stat="identity", fill = "dodgerblue", color="grey40") +
  geom_text(aes(x=Attrition, y=0.01, label= sprintf("%.2f%%", pct)),
            hjust=0.5, vjust=-3, size=4,
            colour="black", fontface="bold") + theme_bw() + labs(x="Employee Attrition", y="Percentage")
  labs(title="Employee Attrition (%)") + theme(plot.title=element_text(hjust=0.5))
attrition_percentage
```



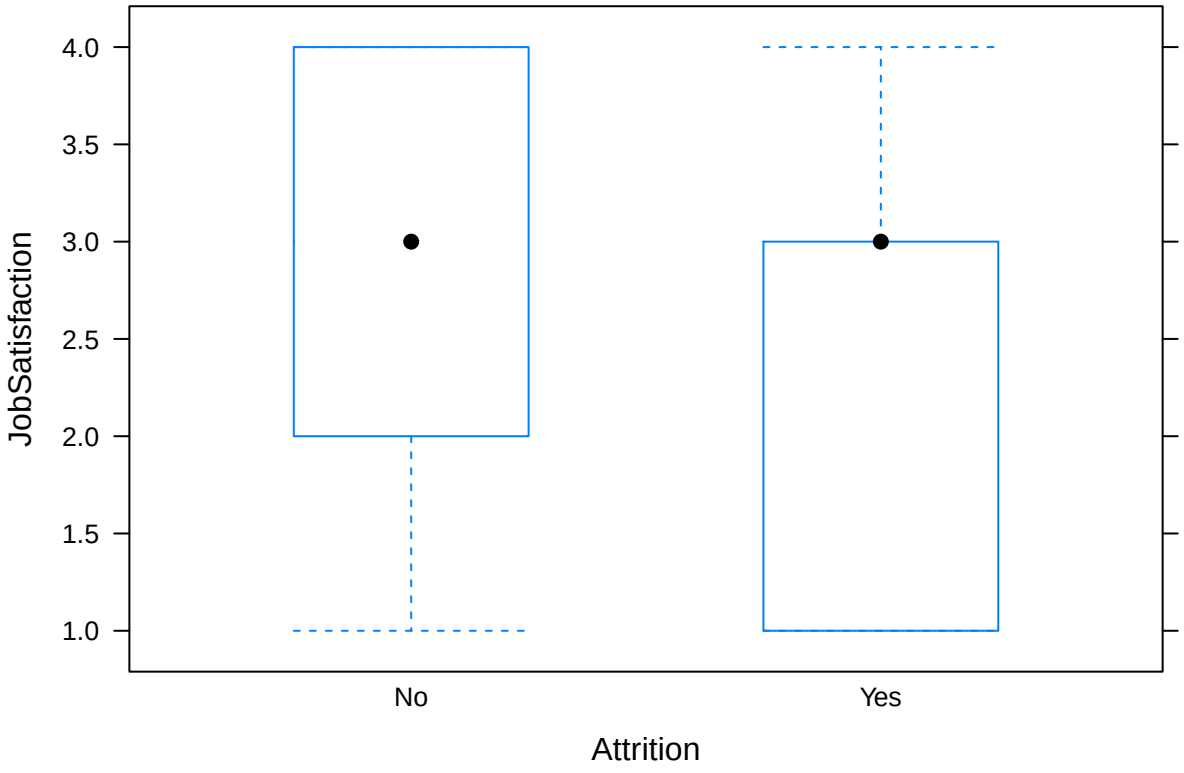

```
library(cowplot)
```

```
##
## ************************************************************

## Note: As of version 1.0.0, cowplot does not change the

##   default ggplot2 theme anymore. To recover the previous

##   behavior, execute:
##   theme_set(theme_cowplot())

## ************************************************************

##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggpubr':
##
##     get_legend
```

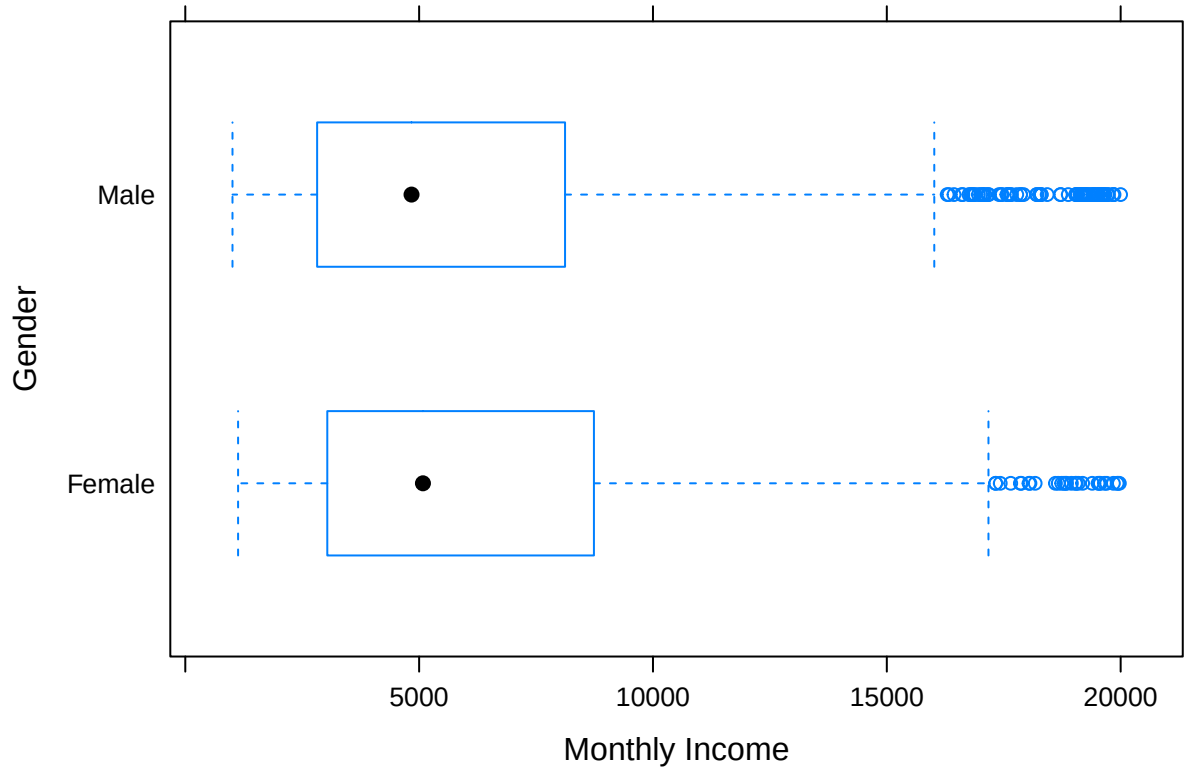
```
plot_grid(attritions_number, attrition_percentage, align="h", ncol=2)
```



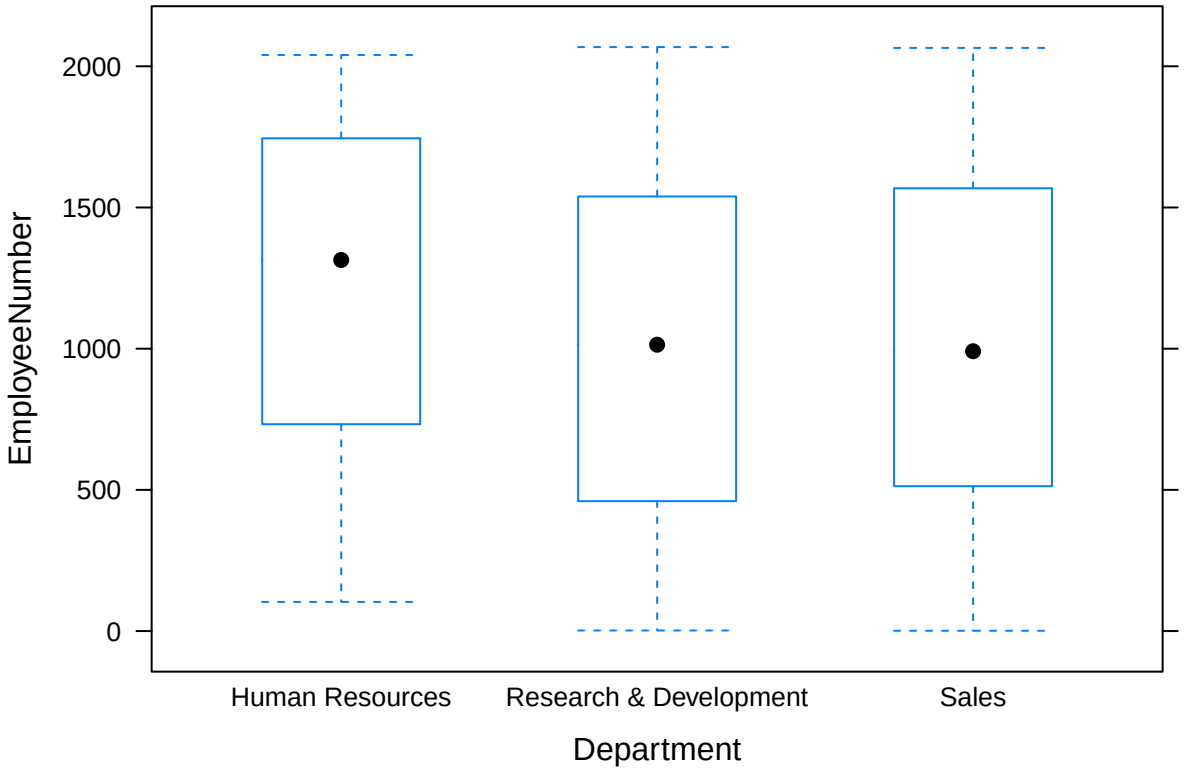

```
#4 Distribution of Job Satisfaction:
bwplot(emp_churn$JobSatisfaction ~ emp_churn$Attrition, data=emp_churn, ylab='JobSatisfaction',xlab='At
```



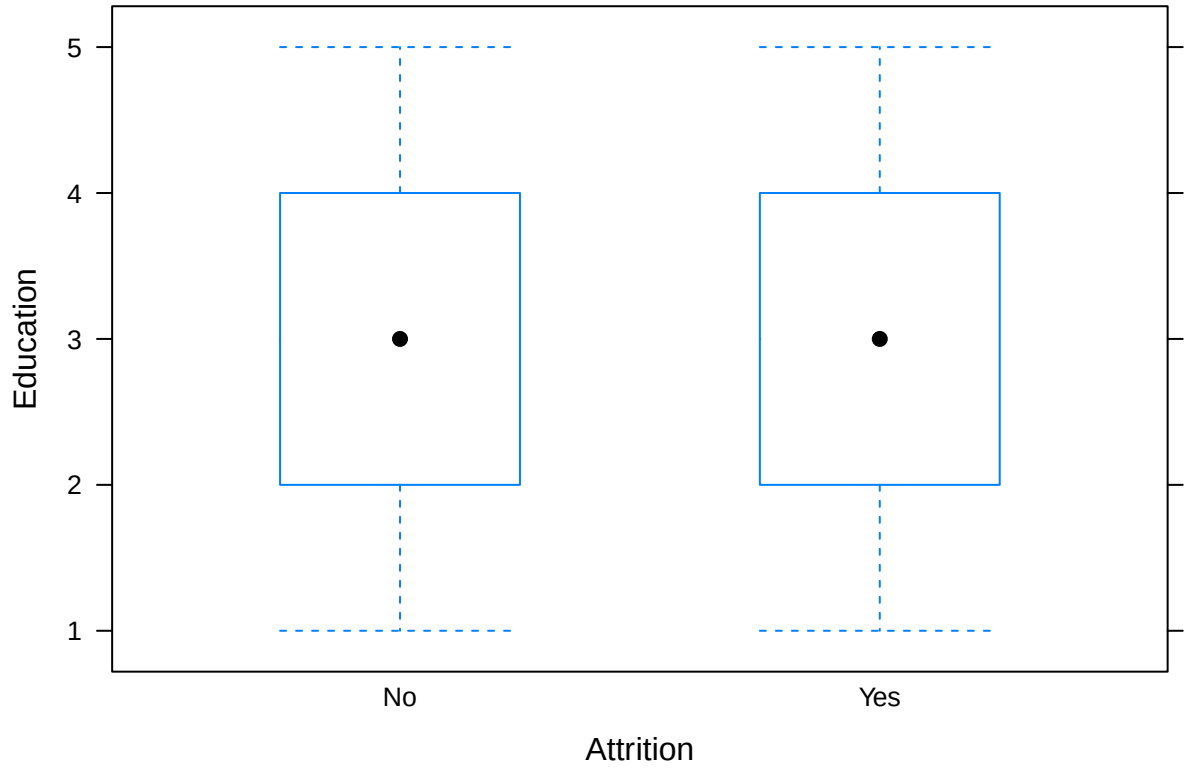

```
##5 Monthly Income by Gender
bwplot(emp_churn$Gender ~ emp_churn$MonthlyIncome, data=emp_churn, ylab='Gender',xlab='Monthly Income')
```




```
##6 Number of employee
bwplot(emp_churn$EmployeeNumber ~ emp_churn$Department, data=emp_churn, ylab='EmployeeNumber',xlab='Dep
```

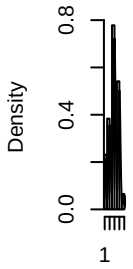



```
bwplot(emp_churn$Education ~ emp_churn$Attrition, data=emp_churn, ylab='Education',xlab='Attrition')
```




```
##7 education-attrition
par(mfrow=c(2,7))
par(mfrow = c(2,7))
hist(emp_churn$Education,xlab='',main = 'Attrition by Education level',freq = FALSE)
lines(density(emp_churn$Education,na.rm = T))
```

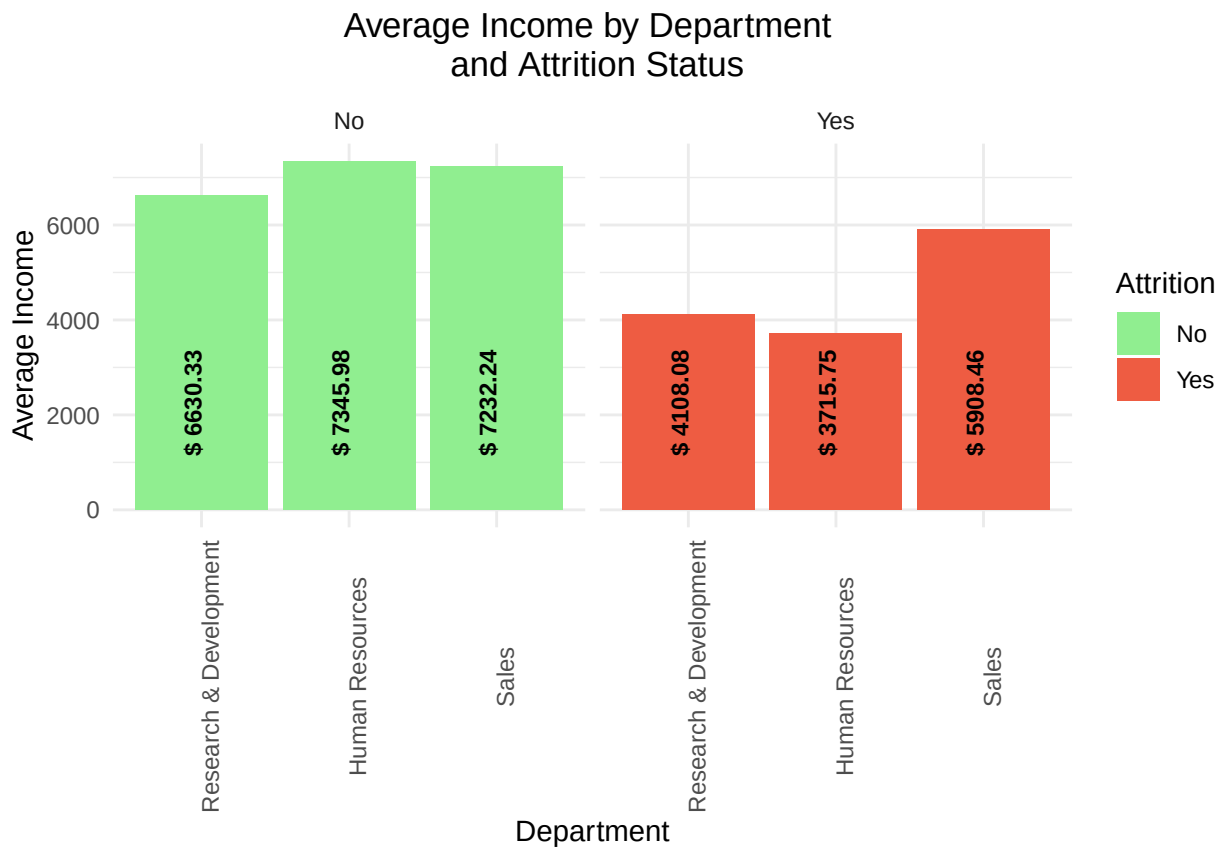
tion by Educat




```
##9


options(repr.plot.width=8, repr.plot.height=5)
emp_churn$JobSatisfaction <- as.factor(emp_churn$JobSatisfaction)
```

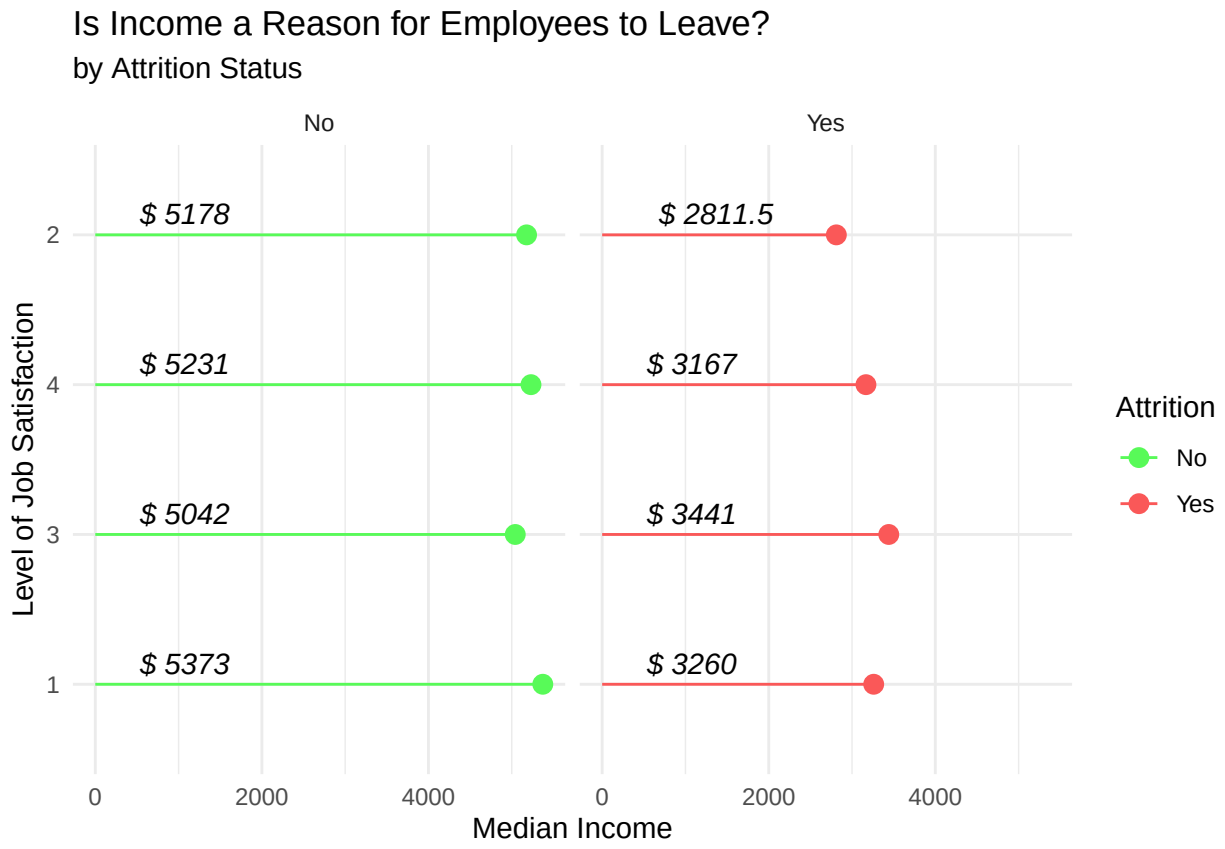
```
options(repr.plot.width=8, repr.plot.height=5)
```

```
avg.income <- emp_churn %>% select(Department, MonthlyIncome, Attrition) %>% group_by(Attrition, Departm
  summarize(avg.inc=mean(MonthlyIncome)) %>%
  ggplot(aes(x=reorder(Department, avg.inc), y=avg.inc, fill=Attrition)) + geom_bar(stat="identity", pos
  theme_minimal() + theme(axis.text.x = element_text(angle = 90), plot.title=element_text(hjust=0.5)) +
  scale_fill_manual(values=c("lightgreen", "tomato2")) +
  labs(y="Average Income", x="Department", title="Average Income by Department \n and Attrition Status"
  geom_text(aes(x=Department, y=0.01, label= paste0("$ ", round(avg.inc,2))),
            hjust=-0.5, vjust=0, size=3,
            colour="black", fontface="bold",
            angle=90)
avg.income
```

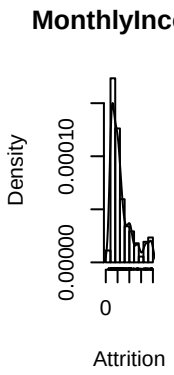



```
#### 10
options(repr.plot.width=8, repr.plot.height=5)

emp_churn$JobSatisfaction <- as.factor(emp_churn$JobSatisfaction)

high.inc <- emp_churn %>% select(JobSatisfaction, MonthlyIncome, Attrition) %>% group_by(JobSatisfactio
  summarize(med=median(MonthlyIncome)) %>%
  ggplot(aes(x=reorder(JobSatisfaction, -med), y=med, color=Attrition)) +
  geom_point(size=3) +
  geom_segment(aes(x=JobSatisfaction,
                   xend=JobSatisfaction,
                   y=0,
                   yend=med)) + facet_wrap(~Attrition)+
  labs(title="Is Income a Reason for Employees to Leave?",
       subtitle="by Attrition Status",
       y="Median Income",
       x="Level of Job Satisfaction") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6), plot.title=element_text(hjust=0.5), strip.backg
        strip.text = element_blank()) +
  coord_flip() + theme_minimal() + scale_color_manual(values=c("#58FA58", "#FA5858")) +
  geom_text(aes(x=JobSatisfaction, y=0.01, label= paste0("$ ", round(med,2))),
            hjust=-0.5, vjust=-0.5, size=4,
            colour="black", fontface="italic",
            angle=360)
```

```
high.inc
```

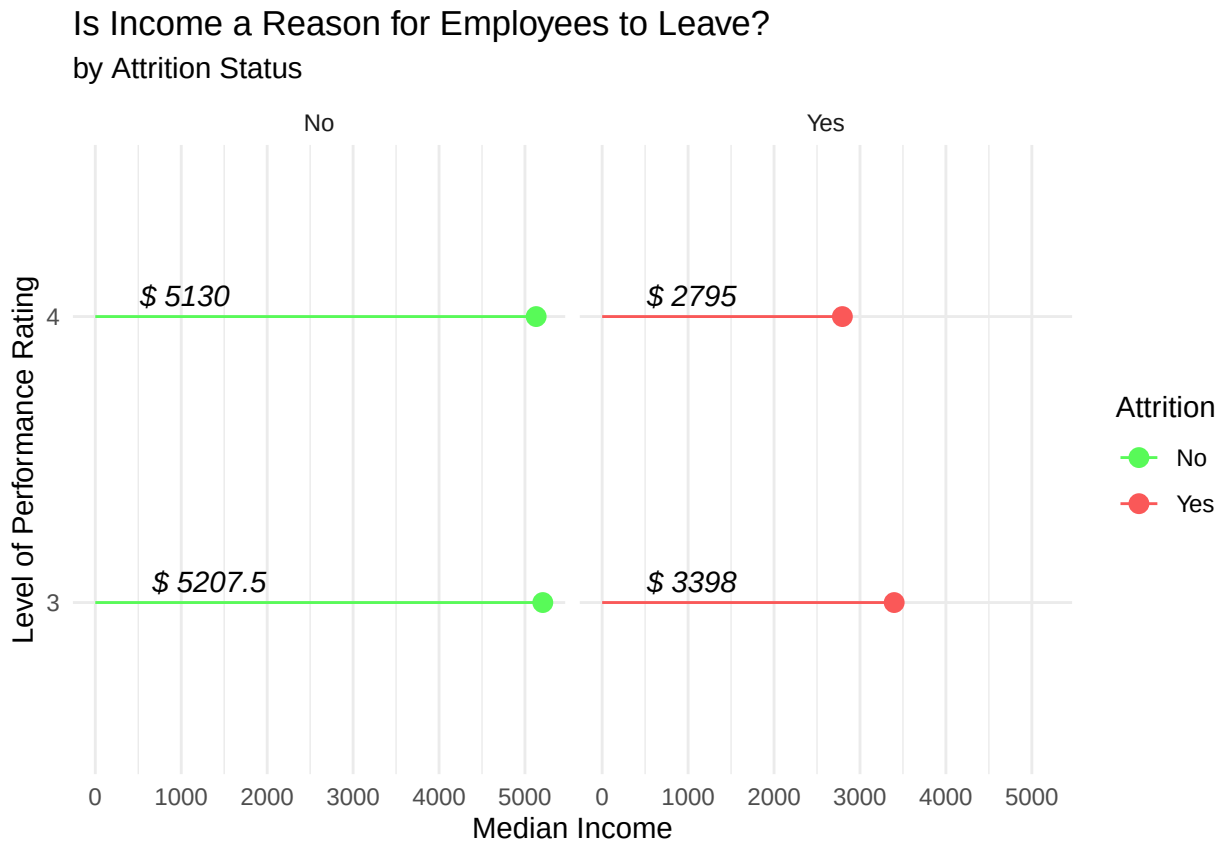



```
##11
par(mfrow=c(2,6))
par(mfrow = c(2,6))
hist(emp_churn$MonthlyIncome,xlab='Attrition',main = 'MonthlyIncome',freq = FALSE)
lines(density(emp_churn$MonthlyIncome,na.rm = T))
rug(jitter(emp_churn$MonthlyIncome))
```




```
#12
options(repr.plot.width=8, repr.plot.height=5)

emp_churn$PerformanceRating <- as.factor(emp_churn$PerformanceRating)

high1.inc <- emp_churn%>% select(PerformanceRating, MonthlyIncome, Attrition) %>% group_by(PerformanceRa
  summarize(med=median(MonthlyIncome)) %>%
  ggplot(aes(x=reorder(PerformanceRating, -med), y=med, color=Attrition)) +
  geom_point(size=3) +
  geom_segment(aes(x=PerformanceRating,
                   xend=PerformanceRating,
                   y=0,
                   yend=med)) + facet_wrap(~Attrition)+
  labs(title="Is Income a Reason for Employees to Leave?",
       subtitle="by Attrition Status",
       y="Median Income",
       x="Level of Performance Rating") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6), plot.title=element_text(hjust=0.5), strip.backg
        strip.text = element_blank()) +
  coord_flip() + theme_minimal() + scale_color_manual(values=c("#58FA58", "#FA5858")) +
  geom_text(aes(x=PerformanceRating, y=0.01, label= paste0("$ ", round(med,2))),
            hjust=-0.5, vjust=-0.5, size=4,
            colour="black", fontface="italic",
            angle=360)
```

```
high1.inc
```


Is Income a Reason for Employees to Leave?
by Attrition Status
No
Yes
Level of Performance Rating
4
3
$ 5130
$ 2795
$ 5207.5
$ 3398
0
1000
2000
3000
4000
5000
Median Income
Attrition
No
Yes