Employee Attrition Analysis

Nabeel Ghalib

18-04-2024

Project overview

This Data Analysis project aims to provide insights into Attrition of employees from XYZ company. By Analyzing various aspects of the data we can identify trends, make Data-driven recommendation to improve the company.

Problem Statement

XYZ company which was established a few years back is facing around a 15% attrition rate for a couple of years. And it's majorly affecting the company in many aspects. In order to understand why employees are leaving the company and reduce the attrition rate XYZ company has approached an HR analytics consultancy for analyzing the data they have. You are playing the HR analyst role in this project and building a dashboard which can help the organization in making data-driven decisions.

ASK

The key business task is to identify the reason employees are leaving the company,

1. Finding out total employees

2. Calculating the attrition rate

3. Finding out the reason for attrition

Data Preparation

The dataset used is provided by Unified Mentor Private Limited which was provided for my Data Analytics internship program.

Note - The XYZ is a fictional company.

Tools Used

RStudio - Data cleaning, Analyzing, and Visualization
Tableau - Data Visualization

Installing required packages

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(forcats) # to reorder by values, variables etc..
library(scales) # to use percent()
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

Importing the dataset

Importing the dataset and storing it in a data frame

```
employee_attrition_data = read.csv("F:/Rprojects/Rprojects/Projects to work/Employee Attrition data.csv
```

DATA CLEANING

Finding null values and na values

```
print(paste0("There are ",nrow(employee_attrition_data)," rows" ))
```

```
## [1] "There are 4410 rows"
```

```
print(paste0("There are ",ncol(employee_attrition_data)," columns"))
```

```
## [1] "There are 29 columns"
```

```
print(paste0("There are ",n_distinct(employee_attrition_data)," distinct rows"))
```

```
## [1] "There are 4410 distinct rows"
```

```
print(paste0("There are ",sum(is.null(employee_attrition_data))," null values"))
```

```
## [1] "There are 0 null values"
```

```
print(paste0("There are ",sum(is.na(employee_attrition_data))," na values"))
```

```
## [1] "There are 111 na values"
```

```
print(paste0("There are ",sum(is.na(employee_attrition_data$EmployeeID))," na values in EmployeeID"))
```

```
## [1] "There are 0 na values in EmployeeID"
```

Removing na values

```
employee_attrition_data = employee_attrition_data %>%
  drop_na()
```

Checking Number of rows, columns and distinct values after removing na values

```
print(paste0("There are ",nrow(employee_attrition_data)," rows"))
```

```
## [1] "There are 4300 rows"
```

```r
print(paste0("There are ",ncol(employee_attrition_data)," columns"))
```

```
## [1] "There are 29 columns"
```

```r
print(paste0("There are ",n_distinct(employee_attrition_data)," distinct rows"))
```

```
## [1] "There are 4300 distinct rows"
```

```r
n_distinct(employee_attrition_data$BusinessTravel)
```

```
## [1] 3
```

```r
n_distinct(employee_attrition_data$Attrition)
```

```
## [1] 2
```

```r
n_distinct(employee_attrition_data$JobRole)
```

```
## [1] 9
```

```r
n_distinct(employee_attrition_data$Gender)
```

```
## [1] 2
```

```r
n_distinct(employee_attrition_data$JobLevel)
```

```
## [1] 5
```

the data is cleaned and ready for analysis.

DATA ANALYSIS

Total Employees

```r
total_employees = employee_attrition_data %>%
  select(EmployeeCount) %>%
  summarise(total_employees = sum(EmployeeCount))

print(paste0("There are ",total_employees," employees"))
```

```
## [1] "There are 4300 employees"
```

Employee Attrition Count and Attrition rate

```r
emp_att_count2 = employee_attrition_data %>%
  select(Attrition) %>%
  count(Attrition, name = 'total_employees') %>%
  summarise(Attrition, total_employees, attrition_rate = round(total_employees/sum(total_employees)* 10

emp_att_count2
```

```
##   Attrition total_employees attrition_rate
## 1        No            3605          83.84
## 2       Yes             695          16.16
```

The attrition count is 695 and the attrition rate is 16.16%

Active Employee

```r
 active_employee = emp_att_count2 %>%
  select(Attrition, total_employees) %>%
  filter(Attrition == "No")

print(paste0('There are : ',active_employee$total_employees ,' active employees'))
```

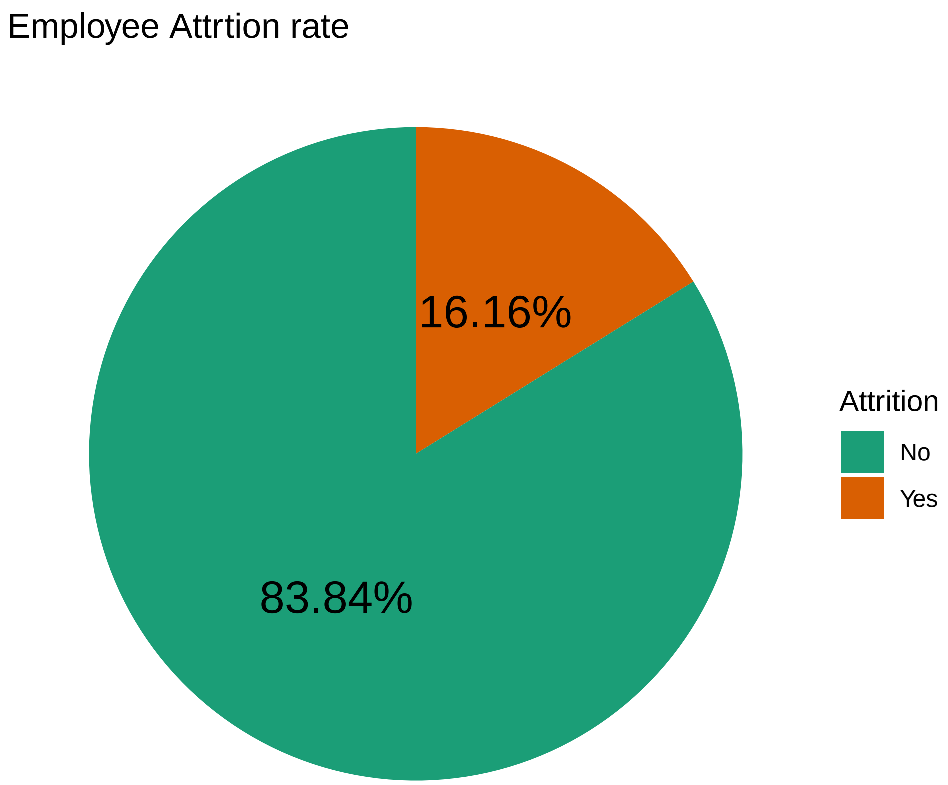
```
## [1] "There are : 3605 active employees"
```

Attrition rate pie chart

```r
# pie chart attrition rate
# calculation to label the values in their respective positions

empatt_count_pie = emp_att_count2

empatt_count_pie = empatt_count_pie %>%
  arrange(desc(Attrition)) %>%
  mutate(prop = (total_employees / sum(empatt_count_pie$total_employees)))%>%
  mutate(ypos = cumsum(prop)- 0.5 * prop)

empatt_count_pie
```

```
##   Attrition total_employees attrition_rate      prop       ypos
## 1       Yes             695          16.16 0.1616279 0.08081395
## 2        No            3605          83.84 0.8383721 0.58081395
```

```r
ggplot(empatt_count_pie, aes(x="", y = prop , fill= Attrition)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  labs(title = 'Employee Attrtion rate') +
  theme_void() + # remove background, grid, numeric labels
  geom_text(aes(y = ypos, label = percent(prop,accuracy = 0.01)), color = 'black',size = 6)+
  scale_fill_brewer(palette="Dark2")
```



Avg age for employees

```
employee_attrition_data %>%
  select(Age) %>%
  summarise(average_age=mean(Age))
```

```
##   average_age
## 1    36.92698
```

Average age for employees Attrition wise

```
employee_attrition_data %>%
  select(Attrition,Age) %>%
  filter(Attrition == 'Yes') %>%
  summarise(attrition_average_age = mean(Age))
```

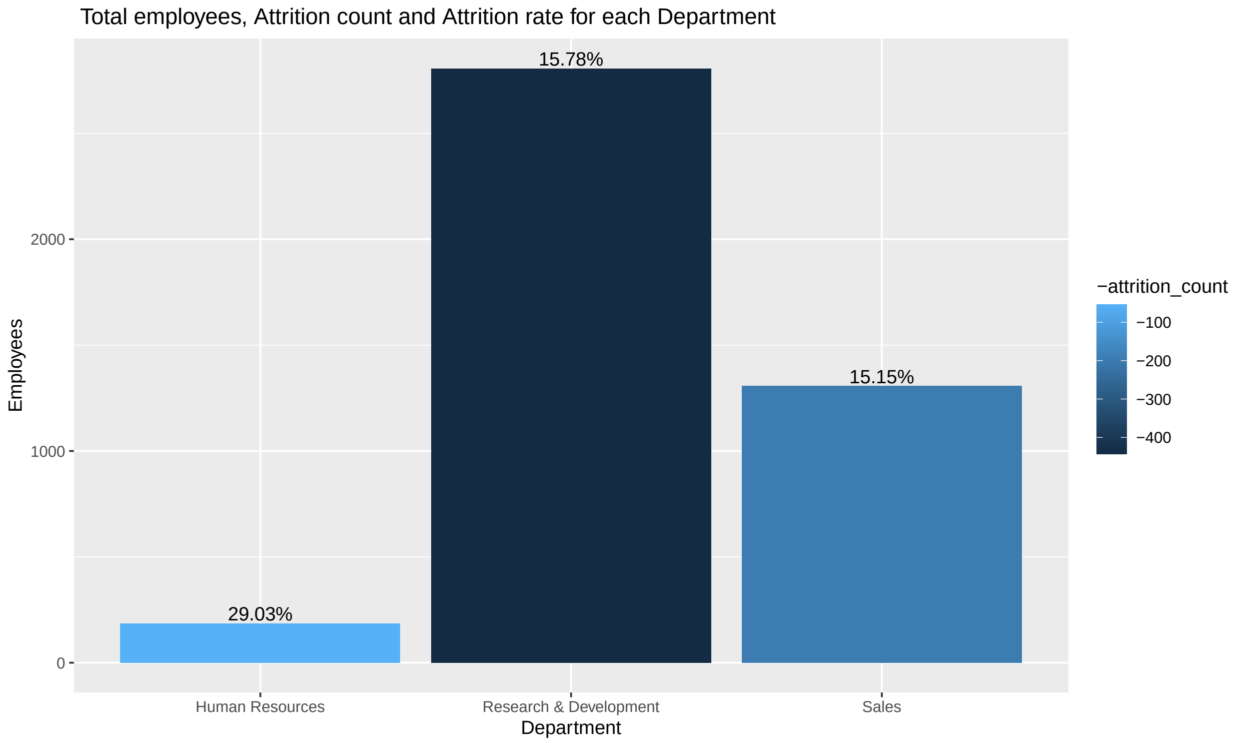
```
##   attrition_average_age
## 1              33.68633
```

Total employees and Attrition count from each department

```
# merging emp_dep , dep_att by department

dept_att = merge(emp_dep,dep_att, by = c("Department","Department"))

dept_att = dept_att %>%
  arrange(-attrition_count)

dept_att = dept_att %>%
  select(Department,total_employees,attrition_count) %>%
  mutate(attrition_rate = (attrition_count/total_employees)) %>%
  mutate(proportion_of_attrition = (attrition_count/sum(attrition_count)))

dept_att
```

```
##               Department total_employees attrition_count attrition_rate
## 1 Research & Development            2807             443      0.1578197
## 2                  Sales            1307             198      0.1514920
## 3        Human Resources             186              54      0.2903226
##   proportion_of_attrition
## 1              0.63741007
## 2              0.28489209
## 3              0.07769784
```

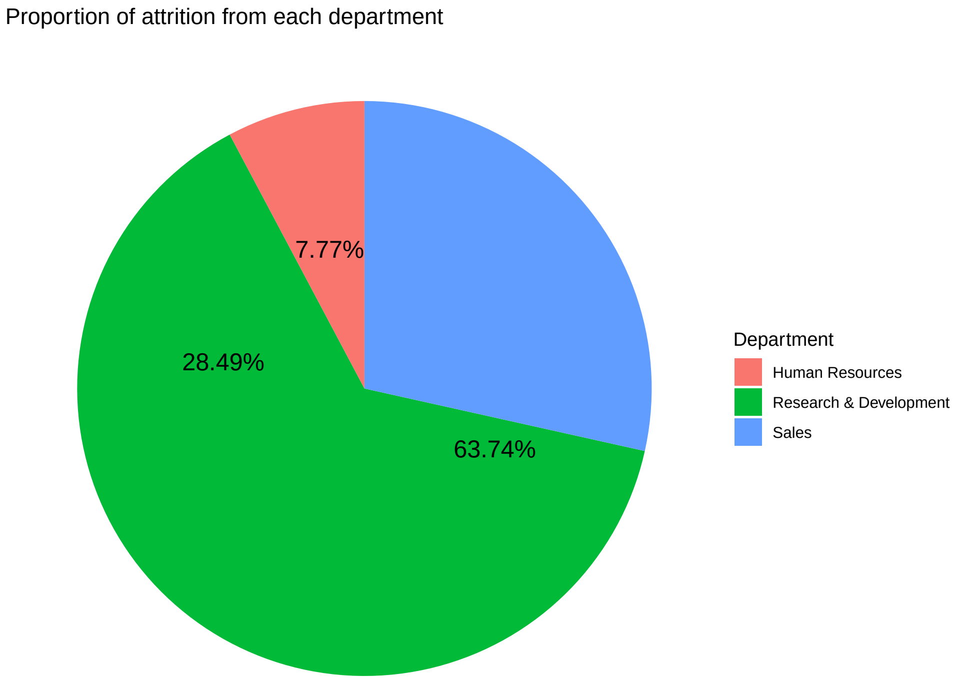
```
# Bar graph


ggplot(data = dept_att, aes(x=Department, y = total_employees, attrition_count, fill = - attrition_coun
  geom_col(position = "dodge") + labs(title = " Total employees, Attrition count and Attrition rate for
  geom_text(aes(label = percent(attrition_rate)), vjust = -0.2)
```



```
# pie chart



dept_att_pie = dept_att

dept_att_pie = dept_att_pie %>%
  arrange(-proportion_of_attrition) %>%
  mutate(prop = attrition_count / sum(attrition_count)) %>%
  mutate(ypos = cumsum(prop)- 0.5 * prop)

dept_att_pie
```

```
##               Department total_employees attrition_count attrition_rate
## 1 Research & Development            2807             443      0.1578197
## 2                  Sales            1307             198      0.1514920
## 3        Human Resources             186              54      0.2903226
##   proportion_of_attrition       prop      ypos
## 1              0.63741007 0.63741007 0.3187050
## 2              0.28489209 0.28489209 0.7798561
## 3              0.07769784 0.07769784 0.9611511
```

```
ggplot(data = dept_att_pie, aes (x=" ", y = prop, fill = Department))+
  geom_bar(stat= "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Proportion of attrition from each department") +
  theme_void() +
  geom_text(aes(y = ypos, label = percent(prop, accuracy = 0.01)), color = "black", size = 5)
```


Proportion of attrition from each department

- Highest attrition count is from **Research & Development Department**, Out of 2807 employees 443 left **(63.74%)**
- Highest attrition rate (%) is from **Human Resources Department**, Out of 186 employees 54 left **(29.03%)**
- Highest proportion of attrition is 64% from **Research & Development Department**

Employees average, max, and min age depatment wise

```
employee_attrition_data %>%
  select(Age,Department) %>%
  group_by(Department) %>%
  summarise(average_age=mean(Age),min_age=min(Age),max_age=max(Age))
```

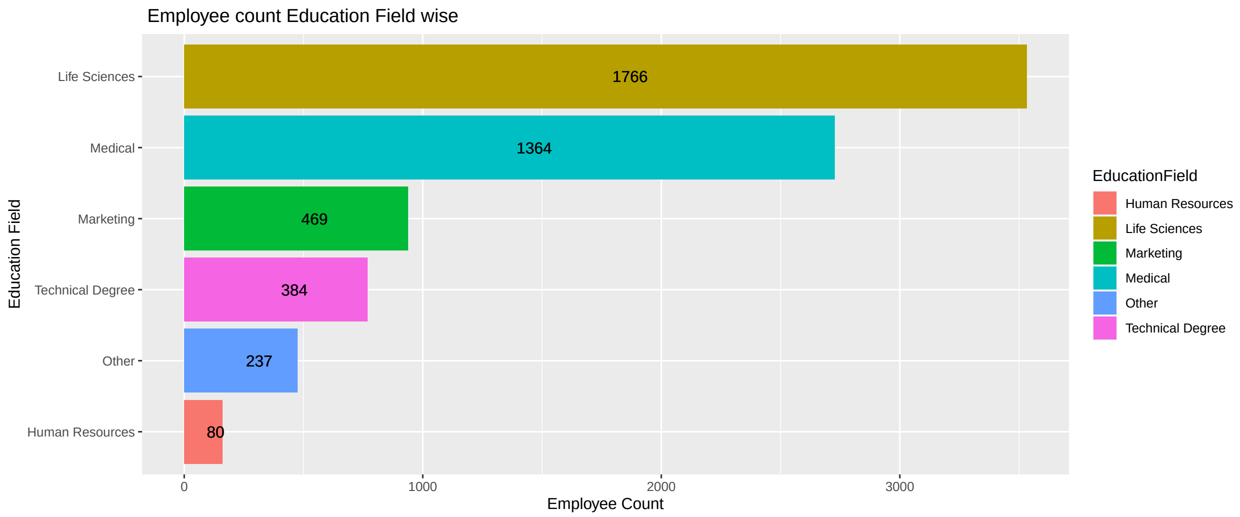
```
## # A tibble: 3 x 4
##   Department             average_age min_age max_age
##   <chr>                        <dbl>   <int>   <int>
## 1 Human Resources               36.7      21      56
## 2 Research & Development        37.1      18      60
## 3 Sales                         36.7      18      60
```

Education field wise total employees and attrition

```
eduf_att_tot = employee_attrition_data %>%
  select(EducationField,Attrition) %>%
  group_by(EducationField) %>%
  count(Attrition, name ='attrition_count') %>%
  reframe(EducationField,Attrition, attrition_count, total_employees=sum(attrition_count)) %>%
  arrange(-total_employees,EducationField)

eduf_att_tot
```

```
## # A tibble: 12 x 4
##    EducationField   Attrition attrition_count total_employees
##    <chr>            <chr>               <int>           <int>
##  1 Life Sciences    No                   1471            1766
##  2 Life Sciences    Yes                   295            1766
##  3 Medical          No                   1145            1364
##  4 Medical          Yes                   219            1364
##  5 Marketing        No                    395             469
##  6 Marketing        Yes                    74             469
##  7 Technical Degree No                    339             384
##  8 Technical Degree Yes                    45             384
##  9 Other            No                    207             237
## 10 Other            Yes                    30             237
## 11 Human Resources  No                     48              80
## 12 Human Resources  Yes                    32              80
```

```
# Horizontal bar chart for education field employee count

ggplot(data = eduf_att_tot,aes(x = reorder(EducationField, total_employees) ,y = total_employees, fill
  geom_bar(stat = "identity") +
  coord_flip()+
  labs(title = " Employee count Education Field wise", x= 'Education Field', y = 'Employee Count')+
  geom_text(aes(label = total_employees), hjust = -0.2)
```



Total employees and attrition count Business Travel wise

```r
bus_emp = employee_attrition_data %>%
  select(BusinessTravel) %>%
  count(BusinessTravel, name = 'total_employees')



bus_att = employee_attrition_data %>%
  select(BusinessTravel, Attrition) %>%
  filter(Attrition == "Yes") %>%
  group_by(BusinessTravel) %>%
  count(Attrition, name = 'attrition_count') %>%
  summarise(BusinessTravel, attrition_count)

# merging df's with businesstravel wise employees and attrition count into another df

bus_emp_att = merge(bus_emp,bus_att, by = "BusinessTravel")

bus_emp_att  =
  bus_emp_att %>%
  mutate(attrition_rate = percent(attrition_count/total_employees)) %>%
  mutate(percent(attrition_count/sum(attrition_count)))

bus_emp_att
```

```
##      BusinessTravel total_employees attrition_count attrition_rate
## 1        Non-Travel             440              36           8.2%
## 2 Travel_Frequently             809             199          24.6%
## 3     Travel_Rarely            3051             460          15.1%
##   percent(attrition_count/sum(attrition_count))
## 1                                             5%
## 2                                            29%
## 3                                            66%
```

```r
bus_emp_att %>%
  mutate(attrition_rate = percent(attrition_count/total_employees)) %>%
  mutate(percent(attrition_count/sum(attrition_count)))
```

```
##      BusinessTravel total_employees attrition_count attrition_rate
## 1        Non-Travel             440              36           8.2%
## 2 Travel_Frequently             809             199          24.6%
## 3     Travel_Rarely            3051             460          15.1%
##   percent(attrition_count/sum(attrition_count))
## 1                                             5%
## 2                                            29%
## 3                                            66%
```

Employee count and attrition count Gender wise

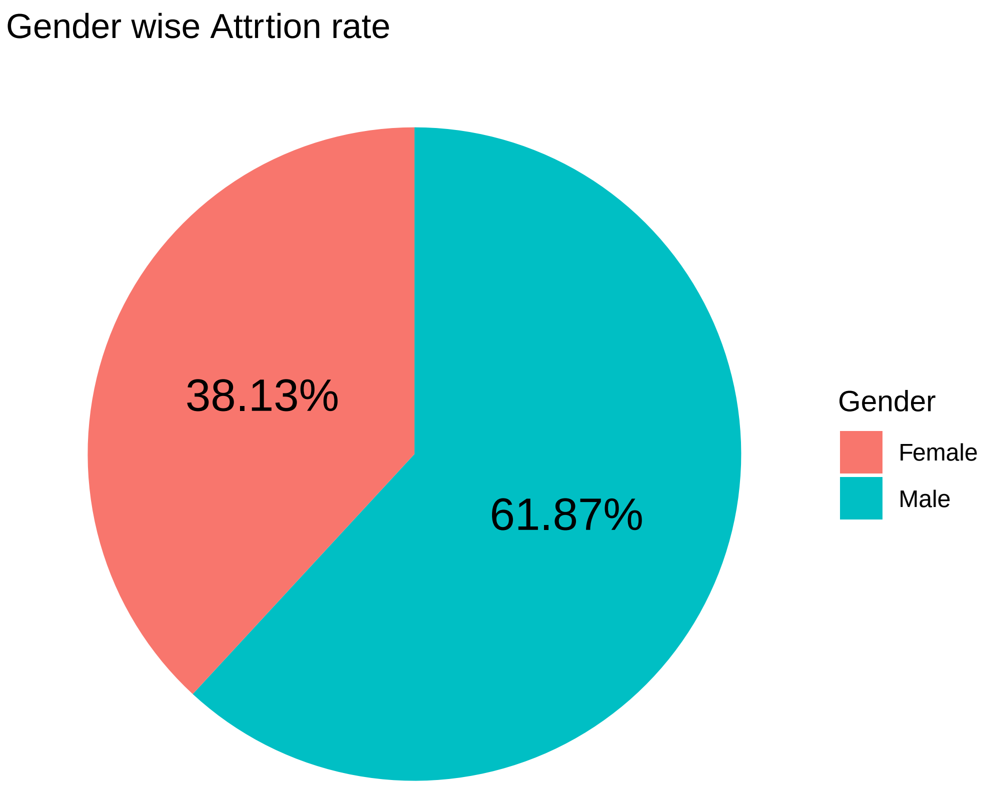
```
gend_tot_att = merge(gend_tot,gend_att, by = c("Gender","Gender"))

gend_tot_att
```

```
##   Gender total_employees attrition_count attrition_rate
## 1 Female            1729             265          38.13
## 2   Male            2571             430          61.87
```

```
gend_pie = gend_tot_att

gend_pie = gend_pie %>%
  arrange(-attrition_rate) %>%
  mutate(prop = (attrition_count/sum(attrition_count))) %>%
  mutate(ypos= cumsum(prop) - 0.5 * prop)

gend_pie
```

```
##   Gender total_employees attrition_count attrition_rate     prop      ypos
## 1   Male            2571             430          61.87 0.618705 0.3093525
## 2 Female            1729             265          38.13 0.381295 0.8093525
```

```
ggplot( data = gend_pie , aes(x= "", y = prop, fill = Gender)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = 'Gender wise Attrtion rate') +
  theme_void() + # remove background, grid, numeric labels
  geom_text(aes(y = ypos, label = percent(prop, accuracy = 0.01)), color = 'black',size = 6)
```

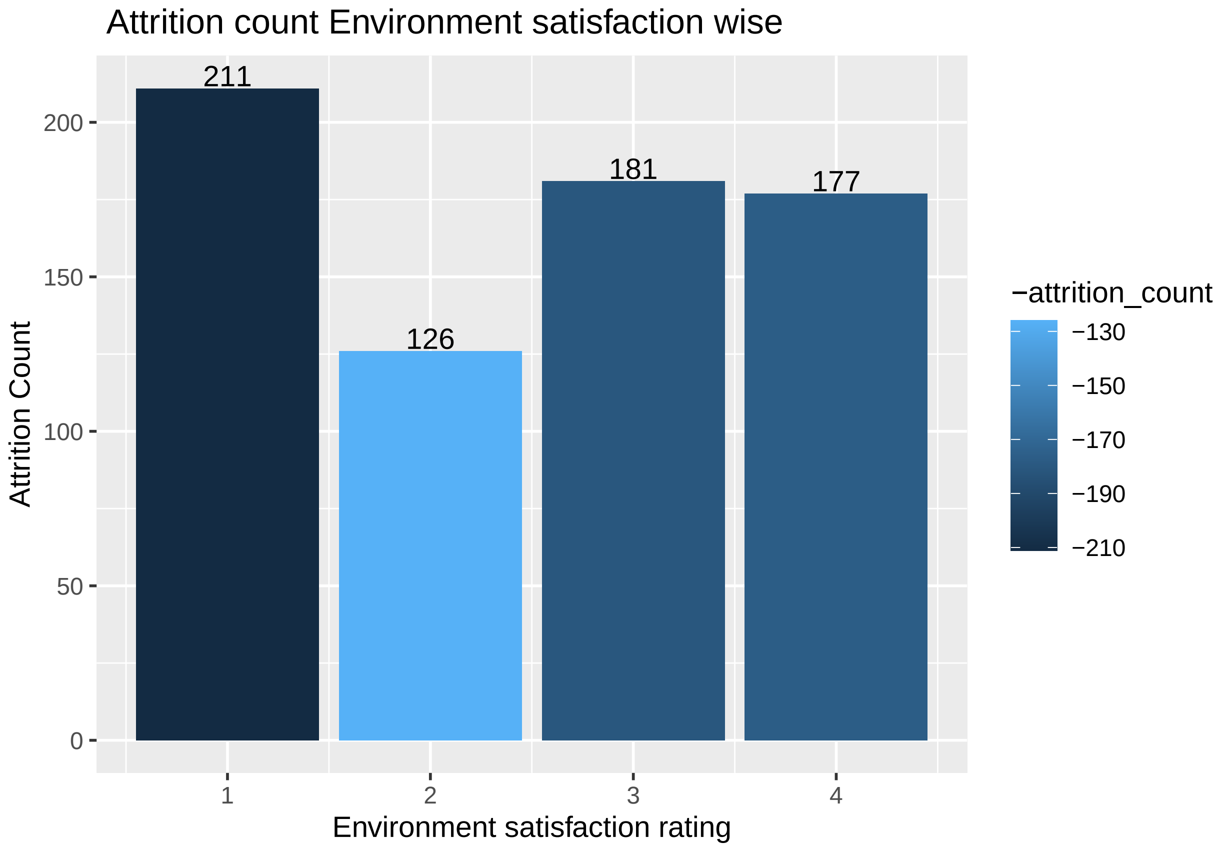


Attrition count for Environment Satisfaction

```
en_sat = employee_attrition_data %>%
  select(Attrition, EnvironmentSatisfaction) %>%
  filter(Attrition == 'Yes') %>%
  group_by(EnvironmentSatisfaction) %>%
  count(Attrition, name = 'attrition_count') %>%
  arrange(-attrition_count)

en_sat
```

```
## # A tibble: 4 x 3
## # Groups:   EnvironmentSatisfaction [4]
##   EnvironmentSatisfaction Attrition attrition_count
##                     <int> <chr>               <int>
## 1                       1 Yes                   211
## 2                       3 Yes                   181
## 3                       4 Yes                   177
## 4                       2 Yes                   126
```

```
  ggplot(data = en_sat, aes(x= EnvironmentSatisfaction, y = attrition_count, fill = - attrition_count))
  geom_col(position = "dodge")+
  labs(title = " Attrition count Environment satisfaction wise", x= 'Environment satisfaction rating', 
  geom_text(aes(label = attrition_count), vjust = -0.1)
```



Marital status wise employees and attrition rate

```
mar_stat_tot = employee_attrition_data %>%
  select(MaritalStatus) %>%
  count(MaritalStatus, name = "total_employees")
mar_stat_tot
```

```
##   MaritalStatus total_employees
## 1      Divorced             949
## 2       Married            1969
## 3        Single            1382
```

```
marstatfull = employee_attrition_data %>%
  select(MaritalStatus, Attrition) %>%
  filter(Attrition == "Yes") %>%
  count(MaritalStatus,Attrition, name = "attrition_count") %>%
  summarise(MaritalStatus, attrition_count, attrition_rate = percent(attrition_count/sum(attrition_coun
```

```
## Warning: Returning more (or less) than 1 row per ‘summarise()‘ group was deprecated in
## dplyr 1.1.0.
## i Please use ‘reframe()‘ instead.
## i When switching from ‘summarise()‘ to ‘reframe()‘, remember that ‘reframe()‘
##   always returns an ungrouped data frame and adjust accordingly.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
marstatfull
```

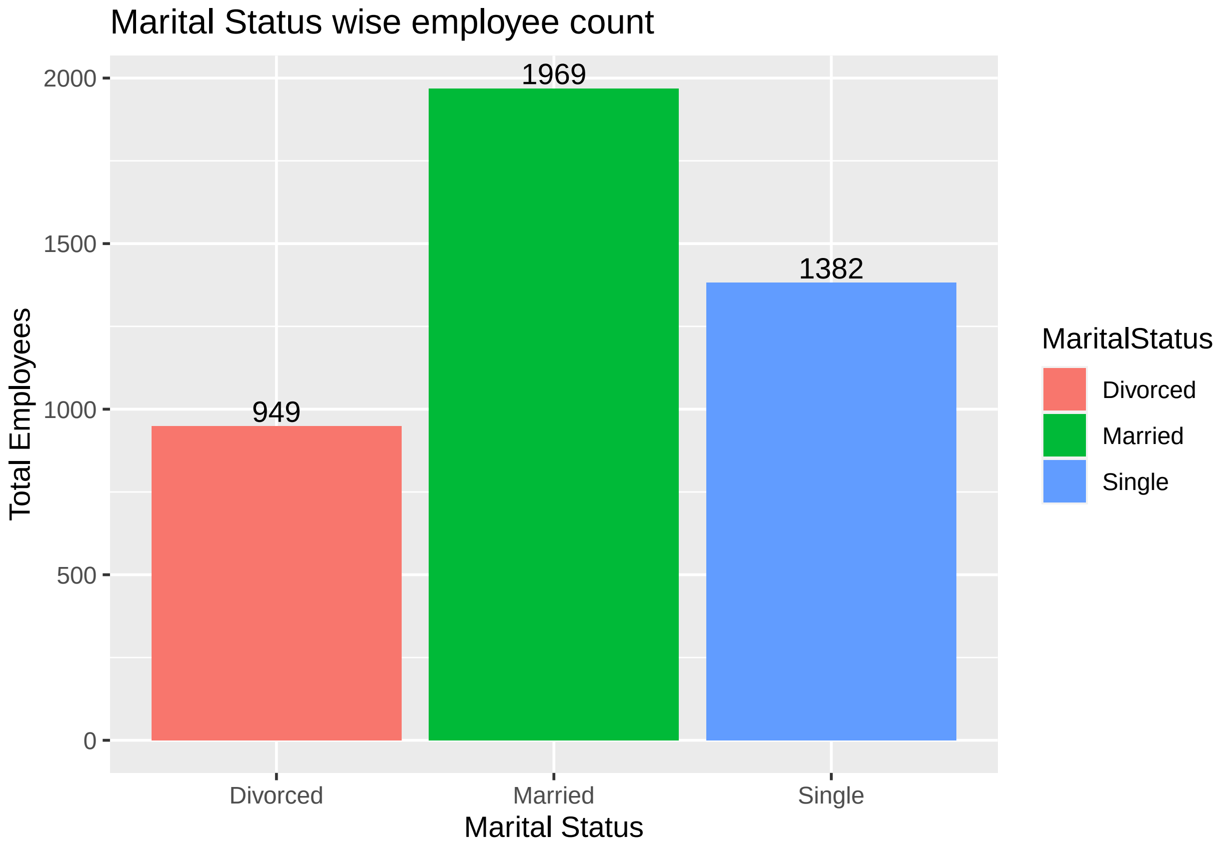
```
##   MaritalStatus attrition_count attrition_rate
## 1      Divorced              94         13.53%
## 2       Married             251         36.12%
## 3        Single             350         50.36%
```

```
mar_stat_full = merge(mar_stat_tot,marstatfull, by = "MaritalStatus","MaritalStatus")

mar_stat_full
```

```
##   MaritalStatus total_employees attrition_count attrition_rate
## 1      Divorced             949              94         13.53%
## 2       Married            1969             251         36.12%
## 3        Single            1382             350         50.36%
```

```
ggplot(data = mar_stat_full,aes(x=MaritalStatus , y = total_employees, fill = MaritalStatus)) +
  geom_col(position = "dodge",stat = "identity")+
  labs(title = "Marital Status wise employee count", x = "Marital Status", y = "Total Employees") +
  geom_text(aes(label = total_employees, vjust = -0.2))
```

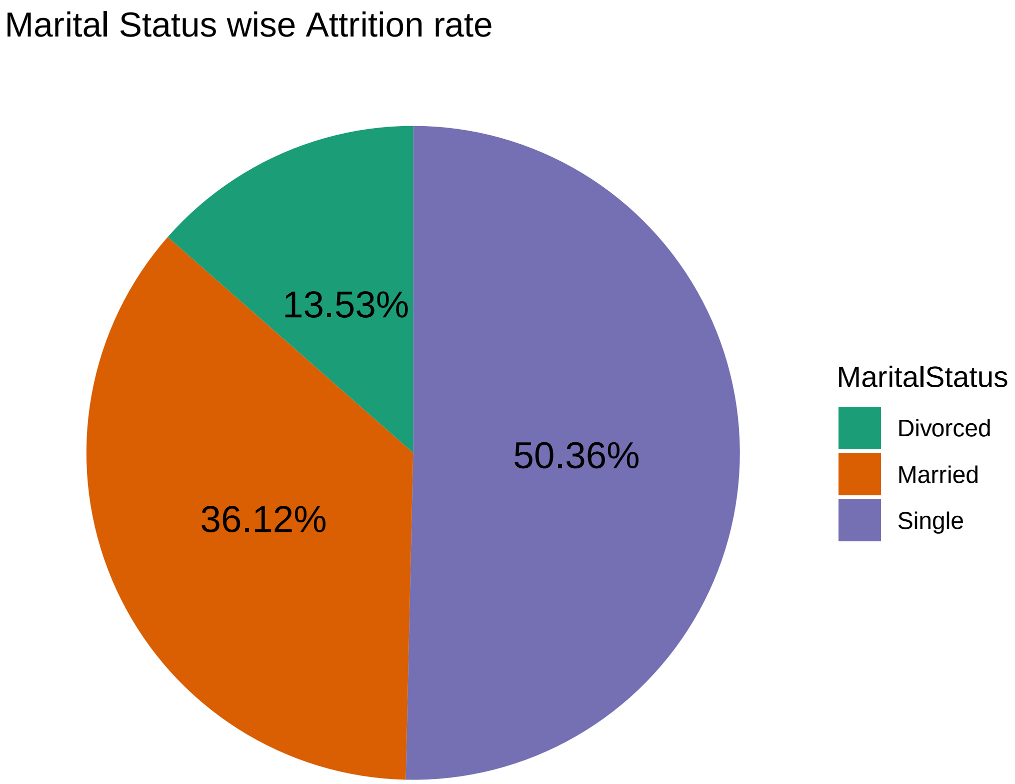
```
## Warning in geom_col(position = "dodge", stat = "identity"): Ignoring unknown
## parameters: 'stat'
```



```r
mar_stat_full_pie = mar_stat_full

mar_stat_full_pie = mar_stat_full_pie %>%
  arrange(-attrition_count) %>%
  mutate(prop = (attrition_count/sum(attrition_count))) %>%
  mutate(ypos = (cumsum(prop) - 0.5 * prop))

mar_stat_full_pie
```

```
##   MaritalStatus total_employees attrition_count attrition_rate      prop
## 1        Single            1382             350         50.36% 0.5035971
## 2       Married            1969             251         36.12% 0.3611511
## 3      Divorced             949              94         13.53% 0.1352518
##        ypos
## 1 0.2517986
## 2 0.6841727
## 3 0.9323741
```

```r
ggplot(data = mar_stat_full_pie, aes(x="",y = prop , fill = MaritalStatus))+
  geom_bar(stat= "identity", width = 1)+
  coord_polar("y", start = 0)+
  labs(title = "Marital Status wise Attrition rate")+
  theme_void()+
  geom_text(aes(y = ypos  , label = percent(prop,accuracy = 0.01)), color = "Black",size = 5) +
  scale_fill_brewer(palette = "Dark2")
```

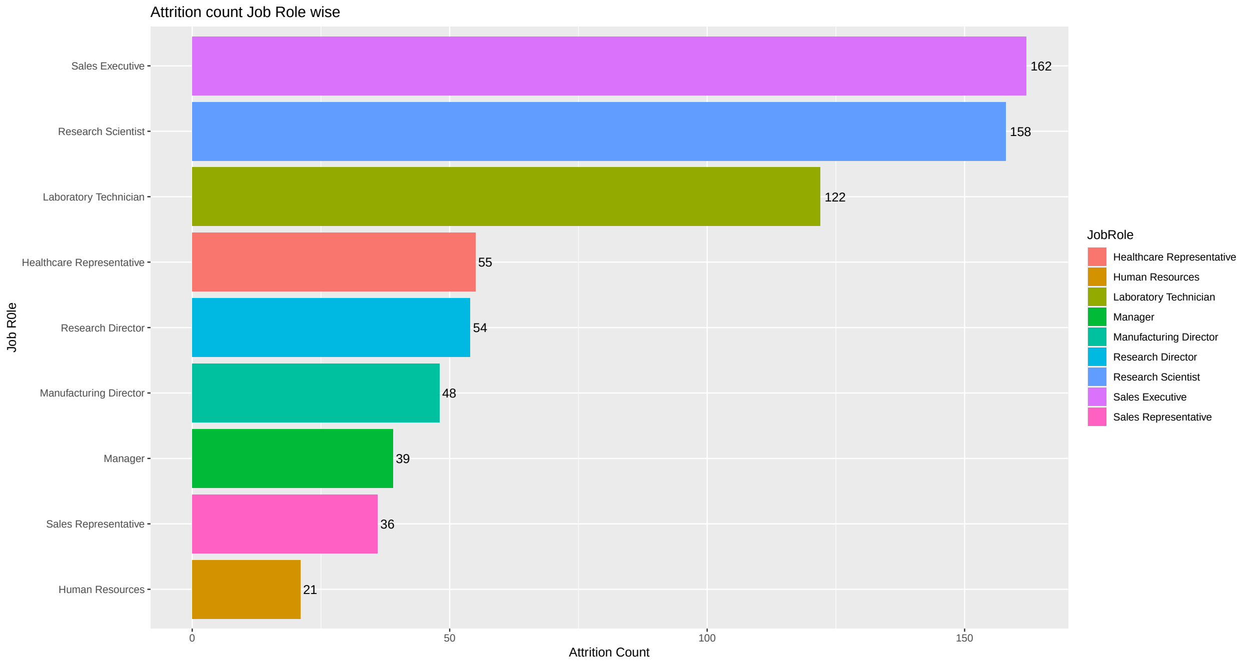


Job Role

```
# total employees

jr_emp = employee_attrition_data %>%
  select(JobRole) %>%
  count(JobRole, name = 'total_employees')

# attrition count

jr_att = employee_attrition_data %>%
  select(JobRole, Attrition) %>%
  filter(Attrition == 'Yes') %>%
  count(JobRole, name = 'attrition_count')

# merged

jr_emp_att = merge(jr_emp, jr_att, by = "JobRole")

jr_emp_att = jr_emp_att %>%
  select(JobRole,total_employees,attrition_count) %>%
  mutate(attrition_rate = (attrition_count / total_employees)*100) %>%
  mutate(prop_of_att= attrition_count / sum(attrition_count)* 100) %>%
  arrange(- attrition_count)

jr_emp_att
```

```
##                      JobRole total_employees attrition_count attrition_rate
## 1          Sales Executive             956             162       16.94561
## 2       Research Scientist             859             158       18.39348
## 3    Laboratory Technician             757             122       16.11625
## 4 Healthcare Representative             377              55       14.58886
## 5        Research Director             235              54       22.97872
## 6   Manufacturing Director             422              48       11.37441
## 7                  Manager             299              39       13.04348
## 8     Sales Representative             241              36       14.93776
## 9          Human Resources             154              21       13.63636
##   prop_of_att
## 1   23.309353
## 2   22.733813
## 3   17.553957
## 4    7.913669
## 5    7.769784
## 6    6.906475
## 7    5.611511
## 8    5.179856
## 9    3.021583
```

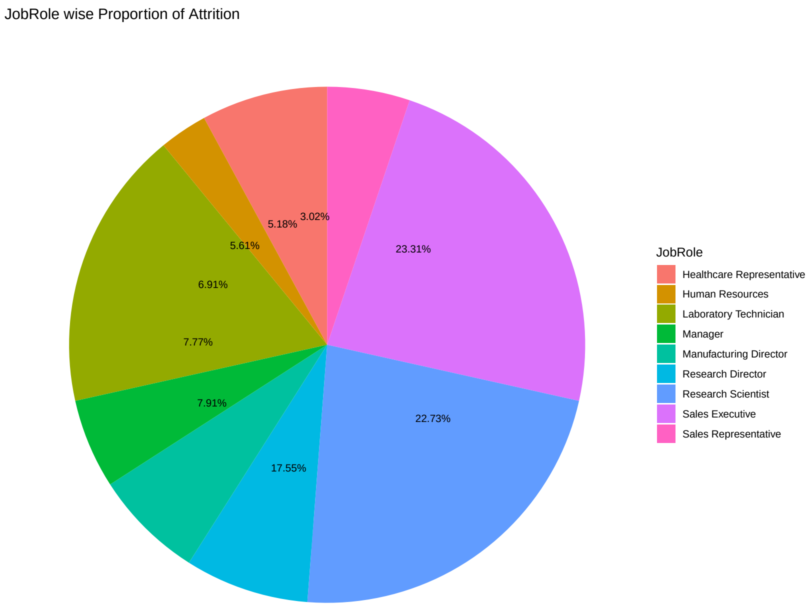
```
# bar chart attrition count

ggplot(data = jr_emp_att,aes(x= reorder(JobRole,attrition_count), y = attrition_count, fill = JobRole))
  geom_col(position = "dodge") +
  coord_flip()+
  labs(title = "Attrition count Job Role wise", x = "Job R0le" , y = "Attrition Count")+
  geom_text(aes(label = attrition_count, hjust= -0.2))
```



```
# Pie Chart attrition rate

jr_emp_att_pie = jr_emp_att

jr_emp_att_pie = jr_emp_att_pie %>%
  arrange(-prop_of_att) %>%
  mutate(prop = attrition_count/sum(attrition_count)) %>%
  mutate(ypos = cumsum(prop) -0.5 * prop)

jr_emp_att_pie
```

```
##                    JobRole total_employees attrition_count attrition_rate
## 1          Sales Executive             956             162       16.94561
## 2       Research Scientist             859             158       18.39348
## 3    Laboratory Technician             757             122       16.11625
## 4 Healthcare Representative            377              55       14.58886
## 5        Research Director             235              54       22.97872
## 6   Manufacturing Director             422              48       11.37441
## 7                  Manager             299              39       13.04348
## 8     Sales Representative             241              36       14.93776
## 9          Human Resources             154              21       13.63636
##   prop_of_att       prop      ypos
## 1   23.309353 0.23309353 0.1165468
## 2   22.733813 0.22733813 0.3467626
## 3   17.553957 0.17553957 0.5482014
## 4    7.913669 0.07913669 0.6755396
## 5    7.769784 0.07769784 0.7539568
## 6    6.906475 0.06906475 0.8273381
## 7    5.611511 0.05611511 0.8899281
## 8    5.179856 0.05179856 0.9438849
## 9    3.021583 0.03021583 0.9848921
```

```
ggplot(data = jr_emp_att_pie,aes(x="" , y = prop, fill = JobRole)) +
  geom_bar(stat = "identity" , width = 1)+
  coord_polar("y" , start = 0) +
  labs(title = "JobRole wise Proportion of Attrition")+
  theme_void()+
  geom_text(aes(y = ypos  , label = percent(prop,accuracy = 0.01)), color = "Black",size = 3)
```


JobRole wise Proportion of Attrition
JobRole
Healthcare Representative
Human Resources
Laboratory Technician
Manager
Manufacturing Director
Research Director
Research Scientist
Sales Executive
Sales Representative