IBM HR Analytics

Attrition in an Organization: Why Workers Quit?

Employees are the backbone of any organization. The organization's performance is directly linked to the quality and retention of its workforce. High employee attrition can have serious consequences. Some of the challenges an organization faces due to employee attrition include:

1. Expensive in Terms of Both Money and Time to Hire New Employees:

- **Recruitment Costs**: Hiring new employees involves significant expenses, including advertising job openings, conducting interviews, and onboarding new hires.
- **Training Costs**: New employees require training, which involves time and money. Additionally, training programs may not be as efficient for inexperienced workers, leading to further delays.

2. Loss of Experienced Employees:

- **Knowledge Gap**: When experienced employees leave, they take valuable knowledge and skills with them. This can lead to a loss of institutional knowledge, which can take time to rebuild.
- **Skill Gap**: The company may face difficulties in finding new employees who possess the exact skills and experience of those who left, leading to decreased efficiency.

3. Impact on Productivity:

- **Disruptions in Workflow**: The absence of experienced workers can disrupt existing workflows and lead to reduced productivity in the short term.
- **Decreased Team Morale**: High attrition rates can create a sense of instability within teams, leading to lower employee morale and, in turn, reduced productivity.

4. Impact on Profit:

- **Cost of Replacement**: The costs of replacing employees—both in terms of recruitment and training—can eat into the company's profits.
- **Lower Quality and Efficiency**: Attrition often results in a less experienced workforce, which can affect the quality and efficiency of work, further impacting the organization's profitability.

Business Task

1. What factors contribute to employee attrition?
2. What measures should the company take to retain their employees?

```r
# Install the tidyverse package (a collection of powerful data science tools)
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
# Install the dplyr package for data manipulation
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
# Install the ggplot2 package for data visualization
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
# Install the readr package for efficient reading of data files
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
# Install the caret package for machine learning models
install.packages("caret")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("gridExtra")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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 errors
```

```r
library(dplyr)
library(ggplot2)
library(readr)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
data <- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
df <- data
colnames(data)
```

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
```

```
##  [7] "Education"                "EducationField"
##  [9] "EmployeeCount"            "EmployeeNumber"
## [11] "EnvironmentSatisfaction"  "Gender"
## [13] "HourlyRate"               "JobInvolvement"
## [15] "JobLevel"                 "JobRole"
## [17] "JobSatisfaction"          "MaritalStatus"
## [19] "MonthlyIncome"            "MonthlyRate"
## [21] "NumCompaniesWorked"       "Over18"
## [23] "OverTime"                 "PercentSalaryHike"
## [25] "PerformanceRating"        "RelationshipSatisfaction"
## [27] "StandardHours"            "StockOptionLevel"
## [29] "TotalWorkingYears"        "TrainingTimesLastYear"
## [31] "WorkLifeBalance"          "YearsAtCompany"
## [33] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

```r
head(data)
```

```
##   Age Attrition    BusinessTravel DailyRate             Department
## 1  41       Yes     Travel_Rarely      1102                  Sales
## 2  49        No Travel_Frequently       279 Research & Development
## 3  37       Yes     Travel_Rarely      1373 Research & Development
## 4  33        No Travel_Frequently      1392 Research & Development
## 5  27        No     Travel_Rarely       591 Research & Development
## 6  32        No Travel_Frequently      1005 Research & Development
##   DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 1                1         2  Life Sciences             1              1
## 2                8         1  Life Sciences             1              2
## 3                2         2          Other             1              4
## 4                3         4  Life Sciences             1              5
## 5                2         1        Medical             1              7
## 6                2         2  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2 Female         94              3        2
## 2                       3   Male         61              2        2
## 3                       4   Male         92              2        1
## 4                       4 Female         56              3        1
## 5                       1   Male         40              3        1
## 6                       4   Male         79              3        1
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate
## 1       Sales Executive               4        Single          5993       19479
## 2    Research Scientist               2       Married          5130       24907
## 3 Laboratory Technician               3        Single          2090        2396
## 4    Research Scientist               3       Married          2909       23159
## 5 Laboratory Technician               2       Married          3468       16632
## 6 Laboratory Technician               4        Single          3068       11864
##   NumCompaniesWorked Over18 OverTime PercentSalaryHike PerformanceRating
## 1                  8      Y      Yes                11                 3
## 2                  1      Y       No                23                 4
## 3                  6      Y      Yes                15                 3
## 4                  1      Y      Yes                11                 3
## 5                  9      Y       No                12                 3
## 6                  0      Y       No                13                 3
##   RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears
## 1                        1            80                0                 8
```

```
## 2                       4           80               1              10
## 3                       2           80               0               7
## 4                       3           80               0               8
## 5                       4           80               1               6
## 6                       3           80               0               8
##   TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole
## 1                     0               1              6                  4
## 2                     3               3             10                  7
## 3                     3               3              0                  0
## 4                     3               3              8                  7
## 5                     3               3              2                  2
## 6                     2               2              7                  7
##   YearsSinceLastPromotion YearsWithCurrManager
## 1                       0                    5
## 2                       1                    7
## 3                       0                    0
## 4                       3                    0
## 5                       2                    2
## 6                       3                    6
```

1. Data Cleaning

```r
# dimensions of the dataset
nrow(data)
```

```
## [1] 1470
```

```r
ncol(data)
```

```
## [1] 35
```

Inference

1. **Number of Records**:
   - The dataset contains a total of **1470 rows/records**.
2. **Number of Features**:
   - The dataset includes **35 columns/features**.

2. Basic Information of Attributes

```r
str(data)
```

```
## 'data.frame':    1470 obs. of  35 variables:
##  $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : chr  "Yes" "No" "Yes" "No" ...
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel_Freque
##  $ DailyRate               : int  1102 279 1373 1392 591 1005 1324 1358 216 1299 ...
##  $ Department              : chr  "Sales" "Research & Development" "Research & Development" "Research
##  $ 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          : chr  "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
##  $ 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                  : chr  "Female" "Male" "Male" "Female" ...
```

```
##  $ 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                 : chr  "Sales Executive" "Research Scientist" "Laboratory Technician" "Res
##  $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : chr  "Single" "Married" "Single" "Married" ...
##  $ 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                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : chr  "Yes" "No" "Yes" "Yes" ...
##  $ 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 ...
```

Inference

1. **Categorical Attributes**:
   - There are **9 categorical attributes** in the dataset.
2. **Numerical Attributes**:
   - There are **26 numerical attributes** in the dataset.
3. **Numerical Features with Categorical Data**:
   - Some of the **numerical features** are storing **categorical data** as numbers.

   - For better analysis, we will **replace** those numerical values with appropriate **categorical values** to ensure clarity and improve the quality of analysis.

```
# Replace numerical values with categorical labels
data$Education <- factor(data$Education, levels = 1:5,
                      labels = c("Below College", "College", "Bachelor", "Master", "Doctor"))

data$EnvironmentSatisfaction <- factor(data$EnvironmentSatisfaction, levels = 1:4,
                              labels = c("Low", "Medium", "High", "Very High"))

data$JobInvolvement <- factor(data$JobInvolvement, levels = 1:4,
                         labels = c("Low", "Medium", "High", "Very High"))

data$JobLevel <- factor(data$JobLevel, levels = 1:5,
                      labels = c("Entry Level", "Junior Level", "Mid Level", "Senior Level", "Executive

data$JobSatisfaction <- factor(data$JobSatisfaction, levels = 1:4,
                          labels = c("Low", "Medium", "High", "Very High"))

data$PerformanceRating <- factor(data$PerformanceRating, levels = 1:4,
                            labels = c("Low", "Good", "Excellent", "Outstanding"))
```

```r
data$RelationshipSatisfaction <- factor(data$RelationshipSatisfaction, levels = 1:4,
                                        labels = c("Low", "Medium", "High", "Very High"))

data$WorkLifeBalance <- factor(data$WorkLifeBalance, levels = 1:4,
                               labels = c("Bad", "Good", "Better", "Best"))
```

```r
head(data)
```

```
##   Age Attrition    BusinessTravel DailyRate             Department
## 1  41       Yes     Travel_Rarely      1102                  Sales
## 2  49        No Travel_Frequently       279 Research & Development
## 3  37       Yes     Travel_Rarely      1373 Research & Development
## 4  33        No Travel_Frequently      1392 Research & Development
## 5  27        No     Travel_Rarely       591 Research & Development
## 6  32        No Travel_Frequently      1005 Research & Development
##   DistanceFromHome     Education EducationField EmployeeCount EmployeeNumber
## 1                1       College  Life Sciences             1              1
## 2                8 Below College  Life Sciences             1              2
## 3                2       College          Other             1              4
## 4                3        Master  Life Sciences             1              5
## 5                2 Below College        Medical             1              7
## 6                2       College  Life Sciences             1              8
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement     JobLevel
## 1                  Medium Female         94           High Junior Level
## 2                    High   Male         61         Medium Junior Level
## 3               Very High   Male         92         Medium  Entry Level
## 4               Very High Female         56           High  Entry Level
## 5                     Low   Male         40           High  Entry Level
## 6               Very High   Male         79           High  Entry Level
##                 JobRole JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate
## 1       Sales Executive       Very High        Single          5993       19479
## 2    Research Scientist          Medium       Married          5130       24907
## 3 Laboratory Technician            High        Single          2090        2396
## 4    Research Scientist            High       Married          2909       23159
## 5 Laboratory Technician          Medium       Married          3468       16632
## 6 Laboratory Technician       Very High        Single          3068       11864
##   NumCompaniesWorked Over18 OverTime PercentSalaryHike PerformanceRating
## 1                  8      Y      Yes                11         Excellent
## 2                  1      Y       No                23       Outstanding
## 3                  6      Y      Yes                15         Excellent
## 4                  1      Y      Yes                11         Excellent
## 5                  9      Y       No                12         Excellent
## 6                  0      Y       No                13         Excellent
##   RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears
## 1                      Low            80                0                 8
## 2                Very High            80                1                10
## 3                   Medium            80                0                 7
## 4                     High            80                0                 8
## 5                Very High            80                1                 6
## 6                     High            80                0                 8
##   TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole
## 1                     0             Bad              6                  4
## 2                     3          Better             10                  7
## 3                     3          Better              0                  0
```

```
## 4                         3          Better            8                  7
## 5                         3          Better            2                  2
## 6                         2            Good            7                  7
##   YearsSinceLastPromotion YearsWithCurrManager
## 1                       0                    5
## 2                       1                    7
## 3                       0                    0
## 4                       3                    0
## 5                       2                    2
## 6                       3                    6
```

3. Checking for Duplicate Records

```r
sum(duplicated(data))
```

```
## [1] 0
```

Inference

- **No Duplicate Records**:
  - Since the output is **0**, it indicates that there are **no duplicate records** present in the dataset, ensuring data integrity for further analysis.

4. Checking for Missing Values and the Percentage of Missing Values

```r
# Calculate the number of missing values in each column
missing_values <- colSums(is.na(data))

# Create a data frame to store the results
missing_data <- data.frame(
  "Total No. of Missing Values" = missing_values,
  "% of Missing Values" = round((missing_values / nrow(data)) * 100, 2)
)

# Print the data frame
print(missing_data)
```

```
##                          Total.No..of.Missing.Values X..of.Missing.Values
## Age                                                0                    0
## Attrition                                          0                    0
## BusinessTravel                                     0                    0
## DailyRate                                          0                    0
## Department                                         0                    0
## DistanceFromHome                                   0                    0
## Education                                          0                    0
## EducationField                                     0                    0
## EmployeeCount                                      0                    0
## EmployeeNumber                                     0                    0
## EnvironmentSatisfaction                            0                    0
## Gender                                             0                    0
## HourlyRate                                         0                    0
## JobInvolvement                                     0                    0
## JobLevel                                           0                    0
## JobRole                                            0                    0
```

```
## JobSatisfaction                                        0                    0
## MaritalStatus                                          0                    0
## MonthlyIncome                                          0                    0
## MonthlyRate                                            0                    0
## NumCompaniesWorked                                     0                    0
## Over18                                                 0                    0
## OverTime                                               0                    0
## PercentSalaryHike                                      0                    0
## PerformanceRating                                      0                    0
## RelationshipSatisfaction                               0                    0
## StandardHours                                          0                    0
## StockOptionLevel                                       0                    0
## TotalWorkingYears                                      0                    0
## TrainingTimesLastYear                                  0                    0
## WorkLifeBalance                                        0                    0
## YearsAtCompany                                         0                    0
## YearsInCurrentRole                                     0                    0
## YearsSinceLastPromotion                                0                    0
## YearsWithCurrManager                                   0                    0
```

Inference

- **No Missing Values**:
  - None of the attributes have **missing values**. This ensures that the analysis will be **unbiased** and **consistent**, as no imputation or handling of missing data is necessary.

5. Descriptive Analysis of the Attributes

```r
summary(data)
```

```
##       Age         Attrition         BusinessTravel       DailyRate     
##  Min.   :18.00   Length:1470        Length:1470        Min.   : 102.0  
##  1st Qu.:30.00   Class :character   Class :character   1st Qu.: 465.0  
##  Median :36.00   Mode  :character   Mode  :character   Median : 802.0  
##  Mean   :36.92                                         Mean   : 802.5  
##  3rd Qu.:43.00                                         3rd Qu.:1157.0  
##  Max.   :60.00                                         Max.   :1499.0  
##   Department        DistanceFromHome          Education    EducationField    
##  Length:1470        Min.   : 1.000   Below College:170   Length:1470       
##  Class :character   1st Qu.: 2.000   College      :282   Class :character  
##  Mode  :character   Median : 7.000   Bachelor     :572   Mode  :character  
##                     Mean   : 9.193   Master       :398                     
##                     3rd Qu.:14.000   Doctor       : 48                     
##                     Max.   :29.000                                         
##  EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender         
##  Min.   :1     Min.   :   1.0   Low      :284            Length:1470       
##  1st Qu.:1     1st Qu.: 491.2   Medium   :287            Class :character  
##  Median :1     Median :1020.5   High     :453            Mode  :character  
##  Mean   :1     Mean   :1024.9   Very High:446                              
##  3rd Qu.:1     3rd Qu.:1555.8                                              
##  Max.   :1     Max.   :2068.0                                              
##    HourlyRate       JobInvolvement            JobLevel     JobRole         
##  Min.   : 30.00   Low      : 83    Entry Level    :543   Length:1470       
##  1st Qu.: 48.00   Medium   :375    Junior Level   :534   Class :character  
```

```
##  Median : 66.00   High     :868    Mid Level      :218   Mode  :character
##  Mean   : 65.89   Very High:144    Senior Level   :106
##  3rd Qu.: 83.75                    Executive Level: 69
##  Max.   :100.00
##   JobSatisfaction MaritalStatus      MonthlyIncome    MonthlyRate
##  Low      :289    Length:1470        Min.   : 1009   Min.   : 2094
##  Medium   :280    Class :character   1st Qu.: 2911   1st Qu.: 8047
##  High     :442    Mode  :character   Median : 4919   Median :14236
##  Very High:459                       Mean   : 6503   Mean   :14313
##                                      3rd Qu.: 8379   3rd Qu.:20462
##                                      Max.   :19999   Max.   :26999
##  NumCompaniesWorked    Over18            OverTime         PercentSalaryHike
##  Min.   :0.000      Length:1470        Length:1470        Min.   :11.00
##  1st Qu.:1.000      Class :character   Class :character   1st Qu.:12.00
##  Median :2.000      Mode  :character   Mode  :character   Median :14.00
##  Mean   :2.693                                            Mean   :15.21
##  3rd Qu.:4.000                                            3rd Qu.:18.00
##  Max.   :9.000                                            Max.   :25.00
##    PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Low        :   0    Low      :276            Min.   :80    Min.   :0.0000
##  Good       :   0    Medium   :303            1st Qu.:80    1st Qu.:0.0000
##  Excellent  :1244    High     :459            Median :80    Median :1.0000
##  Outstanding: 226    Very High:432            Mean   :80    Mean   :0.7939
##                                               3rd Qu.:80    3rd Qu.:1.0000
##                                               Max.   :80    Max.   :3.0000
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         Bad   : 80      Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         Good  :344      1st Qu.: 3.000
##  Median :10.00     Median :3.000         Better:893      Median : 5.000
##  Mean   :11.28     Mean   :2.799         Best  :153      Mean   : 7.008
##  3rd Qu.:15.00     3rd Qu.:3.000                         3rd Qu.: 9.000
##  Max.   :40.00     Max.   :6.000                         Max.   :40.000
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
##  Median : 3.000     Median : 1.000          Median : 3.000
##  Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :18.000     Max.   :15.000          Max.   :17.000
```

Inference

1. **Minimum Age**:
   - The **minimum age** is **18**, which implies that all employees are **adults**.
2. **EmployeeNumber**:
   - The **EmployeeNumber** attribute represents a **unique identifier** for each employee, ensuring that every employee in the dataset can be distinctly recognized.

6. Drop Irrelevant Attributes

```
cols <-  c("Over18","EmployeeCount","EmployeeNumber","StandardHours")
data <- data %>%
  select(-cols)
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(cols)
##
##   # Now:
##   data %>% select(all_of(cols))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
colnames(data)
```

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
##  [7] "Education"                "EducationField"
##  [9] "EnvironmentSatisfaction"  "Gender"
## [11] "HourlyRate"               "JobInvolvement"
## [13] "JobLevel"                 "JobRole"
## [15] "JobSatisfaction"          "MaritalStatus"
## [17] "MonthlyIncome"            "MonthlyRate"
## [19] "NumCompaniesWorked"       "OverTime"
## [21] "PercentSalaryHike"        "PerformanceRating"
## [23] "RelationshipSatisfaction" "StockOptionLevel"
## [25] "TotalWorkingYears"        "TrainingTimesLastYear"
## [27] "WorkLifeBalance"          "YearsAtCompany"
## [29] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [31] "YearsWithCurrManager"
```

```r
# copy of processed dataframe for statistical analysis
new_df <- data
```

Data Cleaning Update

- We have successfully **dropped** the following columns from the dataset:
  - **Over18**
  - **EmployeeCount**
  - **EmployeeNumber**
  - **StandardHours**

These columns were either irrelevant or redundant for the analysis and were removed to streamline the dataset.

7. Checking the Unique Values of the Categorical Attributes

```r
library(dplyr)

# Assuming 'data' is your data frame

# Identify categorical columns
categorical_cols <- sapply(data, is.character) | sapply(data, is.factor)
```

```r
# Iterate over categorical columns and print details
for (col_name in names(data)[categorical_cols]) {
  print(paste("Unique values of", col_name))

  if (is.factor(data[[col_name]])) {
    print(levels(data[[col_name]]))
  } else {
    print(unique(data[[col_name]]))
  }
  cat("\n")
}
```

```
## [1] "Unique values of Attrition"
## [1] "Yes" "No"
##
## [1] "Unique values of BusinessTravel"
## [1] "Travel_Rarely"     "Travel_Frequently" "Non-Travel"
##
## [1] "Unique values of Department"
## [1] "Sales"                  "Research & Development" "Human Resources"
##
## [1] "Unique values of Education"
## [1] "Below College" "College"       "Bachelor"      "Master"
## [5] "Doctor"
##
## [1] "Unique values of EducationField"
## [1] "Life Sciences"    "Other"            "Medical"          "Marketing"
## [5] "Technical Degree" "Human Resources"
##
## [1] "Unique values of EnvironmentSatisfaction"
## [1] "Low"       "Medium"    "High"      "Very High"
##
## [1] "Unique values of Gender"
## [1] "Female" "Male"
##
## [1] "Unique values of JobInvolvement"
## [1] "Low"       "Medium"    "High"      "Very High"
##
## [1] "Unique values of JobLevel"
## [1] "Entry Level"     "Junior Level"    "Mid Level"       "Senior Level"
## [5] "Executive Level"
##
## [1] "Unique values of JobRole"
## [1] "Sales Executive"           "Research Scientist"
## [3] "Laboratory Technician"     "Manufacturing Director"
## [5] "Healthcare Representative" "Manager"
## [7] "Sales Representative"      "Research Director"
## [9] "Human Resources"
##
## [1] "Unique values of JobSatisfaction"
## [1] "Low"       "Medium"    "High"      "Very High"
##
## [1] "Unique values of MaritalStatus"
## [1] "Single"   "Married"  "Divorced"
```

```
##
## [1] "Unique values of OverTime"
## [1] "Yes" "No"
##
## [1] "Unique values of PerformanceRating"
## [1] "Low"         "Good"        "Excellent"   "Outstanding"
##
## [1] "Unique values of RelationshipSatisfaction"
## [1] "Low"       "Medium"    "High"      "Very High"
##
## [1] "Unique values of WorkLifeBalance"
## [1] "Bad"    "Good"   "Better" "Best"
```

Inference

1. **Categorical Attributes**:
   - The **value set** of the **categorical attributes** is **complete** and **easy to understand**.
2. **Preprocessing Requirements**:
   - Since the categorical data is well-defined, there is no need to perform additional **preprocessing steps** for these attributes.

Exploratory data Analysis

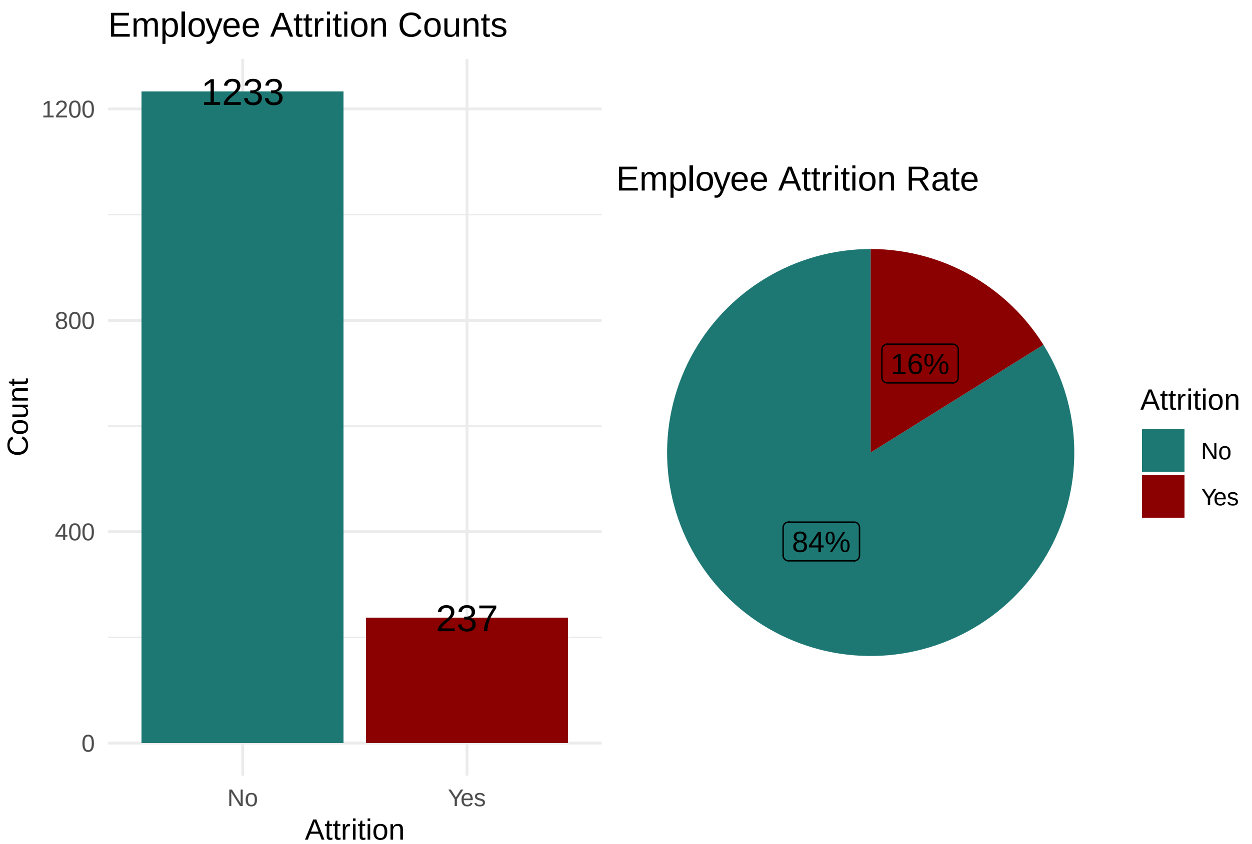
1. Visualizing the Employee Attrition Rate

```r
# Visualization to show Employee Attrition in Counts
# Get attrition counts
attrition_rate <- data %>% count(Attrition)

# Add value labels for bars (optional)
attrition_bar <- ggplot(attrition_rate, aes(x = Attrition, y = n)) +
  geom_bar(stat = "identity", fill = c("#1d7874", "#8B0000")) +
  labs(title = "Employee Attrition Counts", x = "Attrition", y = "Count",
       fontweight = "bold", fontsize = 16) +  # Adjust font weight and size
  theme_minimal() +
  annotate("text", x = seq_along(attrition_rate$Attrition),
           y = attrition_rate$n + 0.1,   # Adjust label position
           label = attrition_rate$n, vjust = 0.5, color = "black", size = 5)  # Adjust label properties

# Create pie chart for percentage
# Basic piechart
attrition_pie <- ggplot(attrition_rate, aes(x = "", y = n, fill = Attrition)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Employee Attrition Rate") +
  theme_void() +
  scale_fill_manual(values = c("#1d7874", "#8B0000")) +
  geom_label(aes(label = paste0(round(n / sum(n) * 100), "%")),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE)


# You can combine both plots using gridExtra package (optional)
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
grid.arrange(attrition_bar, attrition_pie, ncol = 2)
```



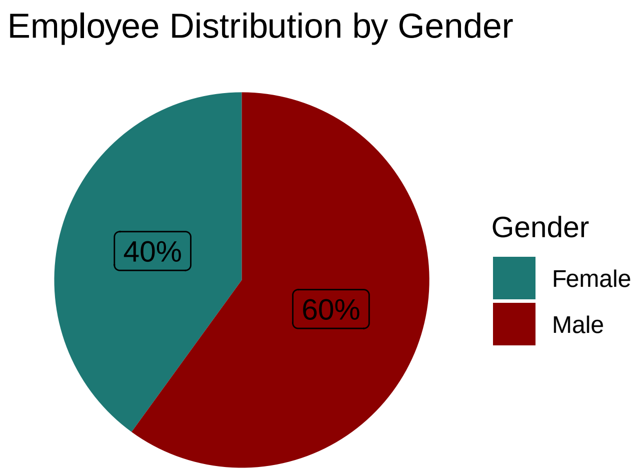
Inference

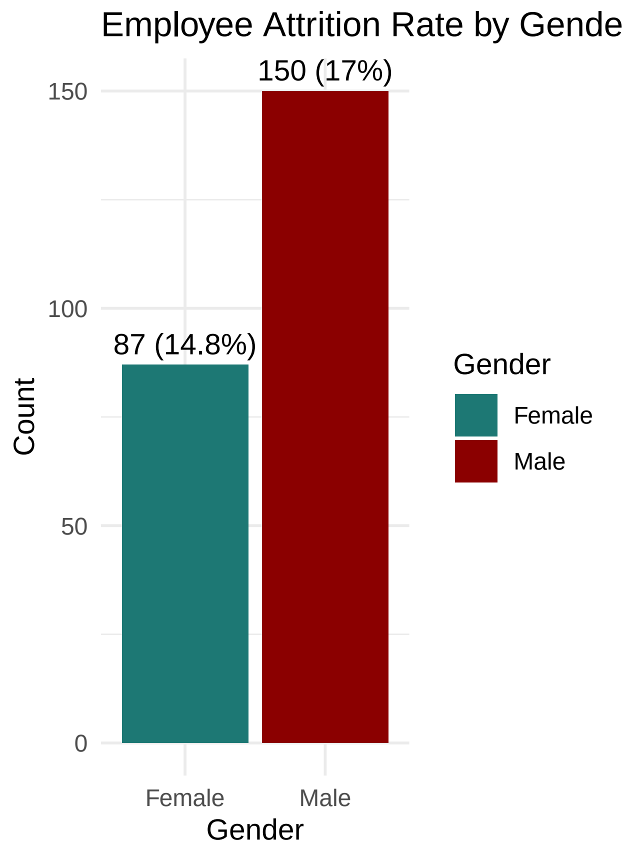
1. **Overall Attrition Rate**:
   - The employee **attrition rate** of this organization is **16.12%**.
2. **Healthy Attrition Rate Benchmark**:
   - According to **Rippling**, a cloud-based software platform, a **healthy attrition rate** typically ranges from **10% to 15%** annually.
3. **Comparison to Healthy Range**:
   - The organization's **attrition rate** exceeds the **healthy threshold**, indicating that it is at a **dangerous level**.
4. **Need for Intervention**:
   - Therefore, **measures must be taken** to reduce attrition rates and improve employee retention. .

2. Analyzing Employee Attrition by Gender

```r
#Visualization to show Total Employees by Gender
gender_dist <- data %>% count(Gender)
```

```
#plot the gender distribution in a pie chart
gender_dist_plot <- ggplot(gender_dist, aes(x="", y=n, fill=Gender)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title="Employee Distribution by Gender") +
  theme_void() +
  scale_fill_manual(values = c("#1d7874", "#8B0000")) +
  geom_label(aes(label = paste0(round(n / sum(n) * 100), "%")),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE)

# Calculate attrition counts
attrition_data <- data %>% filter(Attrition == "Yes")

# Calculate gender counts for all employees and those who left
gender_counts <- data %>% count(Gender)
attrition_counts <- attrition_data %>% count(Gender)

# Merge data frames
merged_data <- data.frame(
  Gender = gender_counts$Gender,
  Total = gender_counts$n,
  Left = attrition_counts$n,
  Attrition_Rate = round((attrition_counts$n / gender_counts$n) * 100, 1)
)

# Create the bar chart
attrition_gen <- ggplot(merged_data, aes(x = Gender, y = Left, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)")), position = position_dodge(width = 0.9
  labs(title = "Employee Attrition Rate by Gender", x = "Gender", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#1d7874", "#8B0000"))

grid.arrange(gender_dist_plot, attrition_gen, ncol = 2)
```





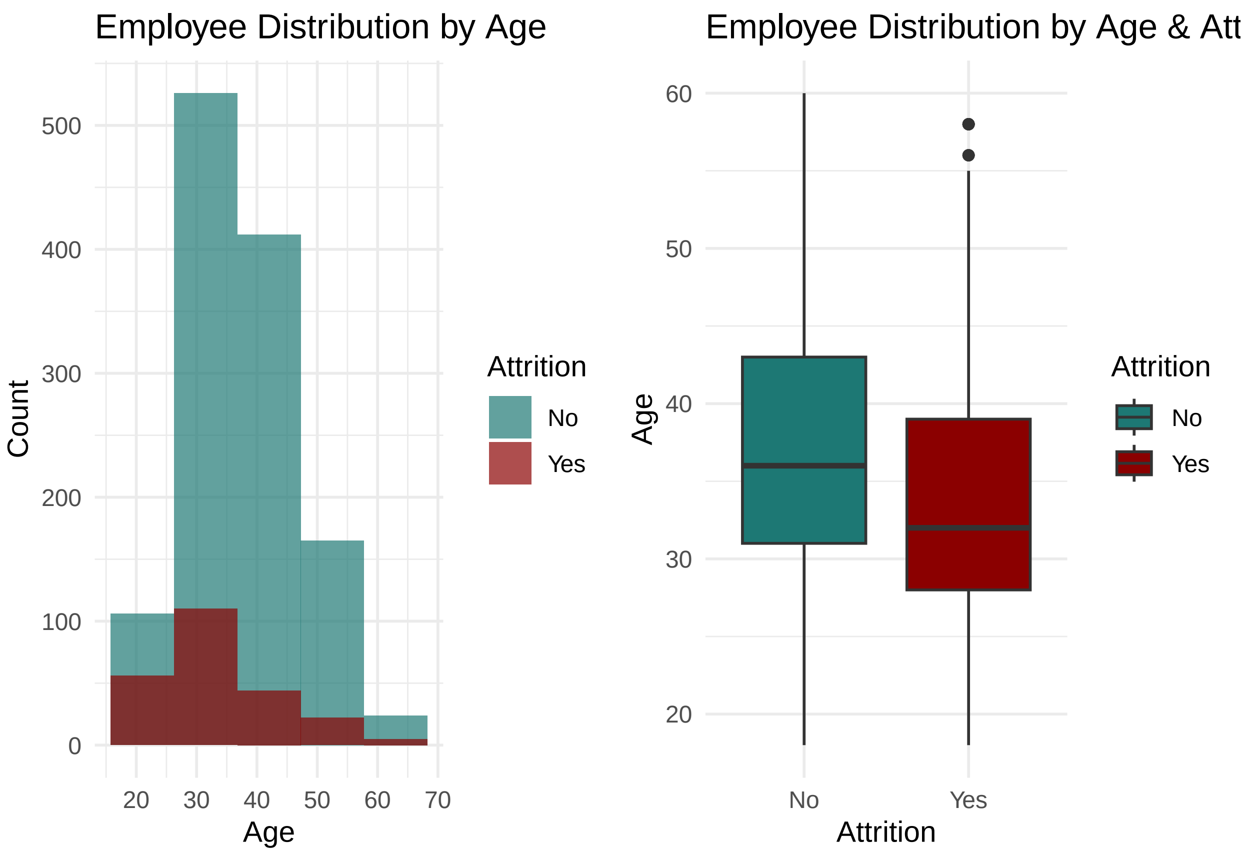
Inference

1. **Age Distribution**:
   - The majority of employees are between the ages of **30 to 40**.
2. **Attrition and Age Relationship**:
   - As **age increases**, **attrition decreases**, indicating that older employees are more likely to stay with the organization.
3. **Median Age Comparison**:
   - From the boxplot, it is evident that the **median age** of employees who **left** the organization is **lower** than that of employees who are still working.
4. **Attrition Among Younger Employees**:
   - **Younger employees** tend to leave the company more frequently compared to their older counterparts.

3. Analyzing Employee Attrition by Age

```
# Visualization to show Employee Distribution by Age
employee_dist_age <- ggplot(data, aes(x = Age, fill = Attrition)) +
  geom_histogram(alpha = 0.7, position = "identity", bins = 5) +
  labs(title = "Employee Distribution by Age", x = "Age", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#1d7874", "#8B0000"))

# Visualization to show Employee Distribution by Age & Attrition
employee_dist_attvsage <- ggplot(data = data, aes(x = Attrition, y = Age, fill = Attrition)) +
  geom_boxplot() +
  labs(title = "Employee Distribution by Age & Attrition", x = "Attrition", y = "Age") +
  theme_minimal() +
```

```r
  scale_fill_manual(values = c("#1d7874", "#8B0000")) +
  guides(color = guide_legend(title = NULL))

grid.arrange(employee_dist_age, employee_dist_attvsage, ncol = 2)
```



Inference

1. **Age Distribution**:
   - The majority of employees are between the ages of **30 to 40**.
2. **Attrition and Age Relationship**:
   - As **age increases**, **attrition decreases**, indicating that older employees are more likely to stay with the organization.
3. **Median Age Comparison**:
   - From the boxplot, it is evident that the **median age** of employees who **left** the organization is **lower** than that of employees who are still working.
4. **Attrition Among Younger Employees**:
   - **Younger employees** tend to leave the company more frequently compared to their older counterparts.

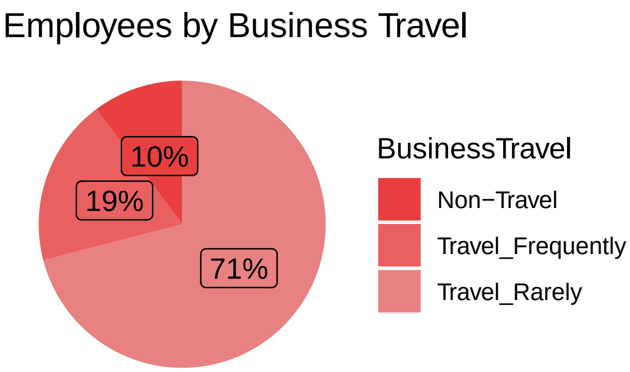
4. Analyzing Employee Attrition by Business Travel

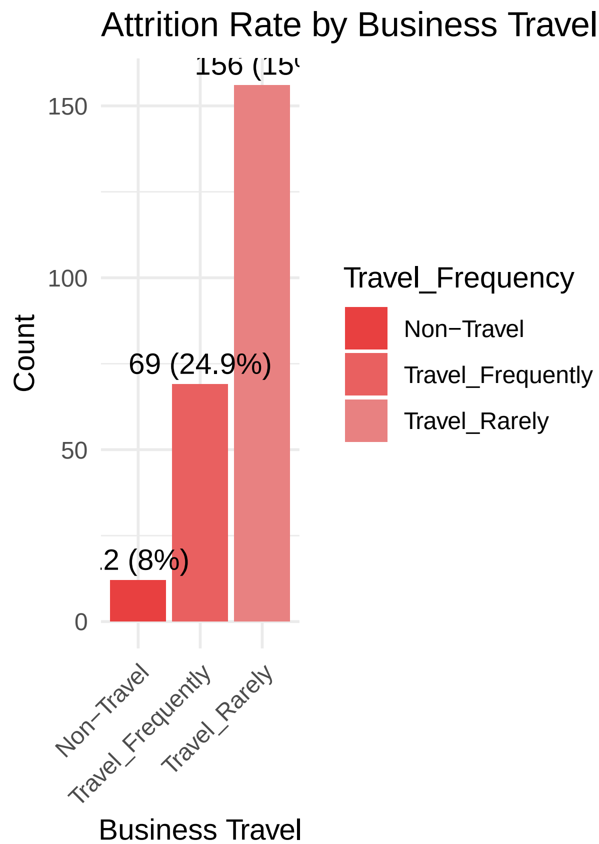
```r
# Visualization to show Total Employees by Business Travel
businesstravel_count <- data %>% count(BusinessTravel)
businesstravel_count
```

```
##      BusinessTravel   n
## 1        Non-Travel 150
## 2 Travel_Frequently 277
```

```
## 3     Travel_Rarely 1043
```

```
businesstravel_plot <- ggplot(businesstravel_count, aes(x="", y=n, fill=BusinessTravel)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title="Employees by Business Travel") +
  theme_void() +
  scale_fill_manual(values = c('#E84040', '#E96060', '#E88181')) +
  geom_label(aes(label = paste0(round(n / sum(n) * 100), "%")),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE)

# visualization to show Attrition by Business Travel
attrition_data <- data %>% filter(Attrition == "Yes")

# Calculate business travel counts for all employees and those who left
business_travel_counts <- data %>% count(BusinessTravel)
attrition_by_travel <- attrition_data %>% count(BusinessTravel)

# Merge data frames
merged_data <- data.frame(
  Travel_Frequency = business_travel_counts$BusinessTravel,
  Total_Employees = business_travel_counts$n,
  Left = attrition_by_travel$n,
  Attrition_Rate = round((attrition_by_travel$n / business_travel_counts$n) * 100, 1)
)

# Create the bar chart
attrition_business_travel <-  ggplot(merged_data, aes(x = Travel_Frequency, y = Left, fill = Travel_Fre
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Attrition Rate by Business Travel", x = "Business Travel", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c('#E84040', '#E96060', '#E88181')) +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)")), x = merged_data$Travel_Frequency,
            vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))


grid.arrange(businesstravel_plot, attrition_business_travel,ncol=2)
```





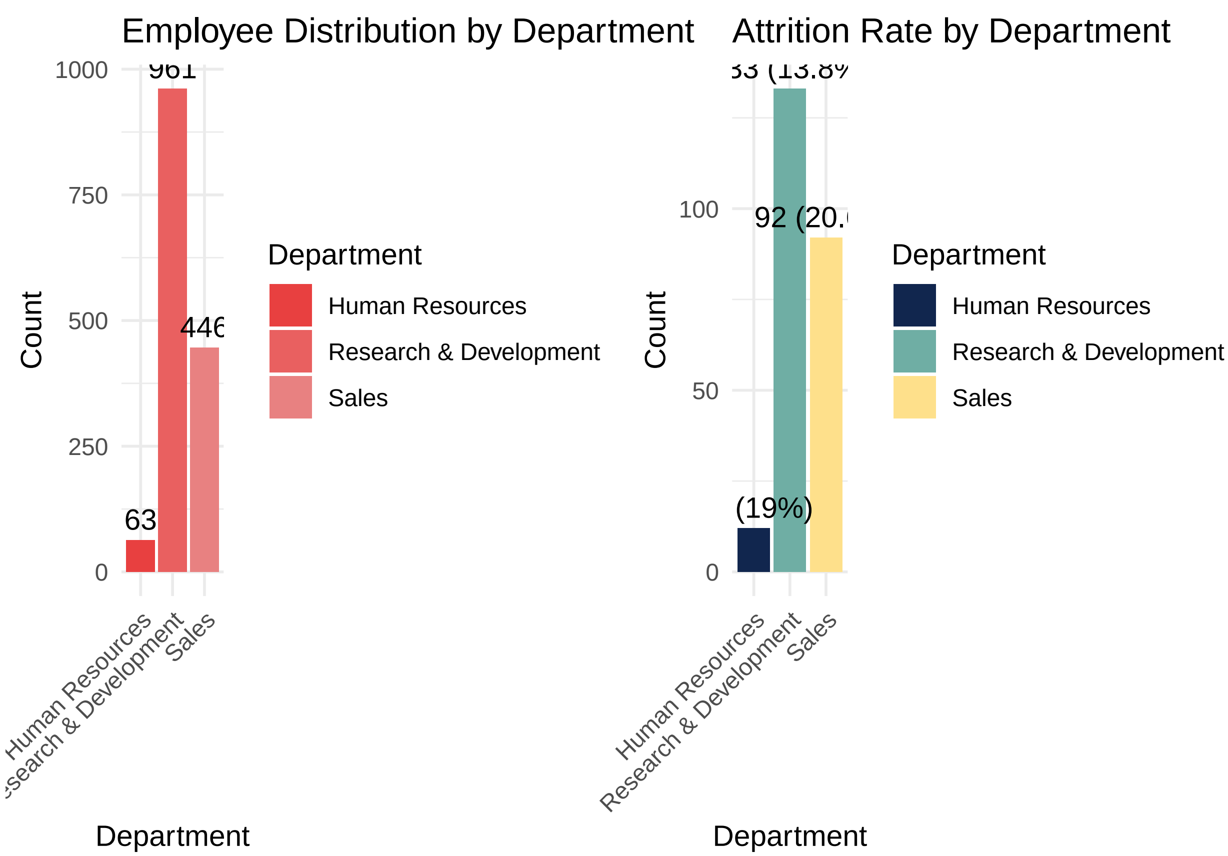
Inference

1. **Travel Frequency Distribution**:
   - **71%** of the employees in the organization **rarely** travel.
2. **Attrition Rates by Travel Frequency**:
   - The **highest attrition rate** is observed among employees who **travel frequently** (**24.9%**).
3. **Attrition Rate Among Non-Travellers**:
   - The **lowest attrition rate** is found among employees who **do not travel** (**8%**), suggesting that travel may contribute to higher turnover.

5. Analyzing Employee Attrition by Department

```
# Visualizing employees by department
department_counts <- data %>% count(Department)

department_dist <- ggplot(data = department_counts, aes(x=Department, y=n, fill=Department)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  scale_fill_manual(values = c('#E84040', '#E96060', '#E88181')) +
  labs(title="Employee Distribution by Department", x="Department", y = "Count") +
  geom_text(aes(label = paste0(n)), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

# Visualization to show Employee Attrition Rate by Department.
attrition_data <- data %>% filter(Attrition == "Yes")

# Calculate department-wise counts
department_counts <- data %>% count(Department)
```

```r
attrition_by_dept <- attrition_data %>% count(Department)

# Merge data frames
merged_data <- data.frame(
  Department = department_counts$Department,
  Total_Employees = department_counts$n,
  Left = attrition_by_dept$n,
  Attrition_Rate = round((attrition_by_dept$n / department_counts$n) * 100, 1)
)

# Create the bar chart
department_att_plot <-  ggplot(merged_data, aes(x = Department, y = Left, fill = Department)) +
  geom_bar(stat = "identity") +
  labs(title = "Attrition Rate by Department", x = "Department", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#11264e", "#6faea4", "#FEE08B")) +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)")), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

grid.arrange(department_dist, department_att_plot,ncol = 2)
```



Inference

1. **Department Distribution**:
   - The majority of employees are from the **Research and Development** department.
2. **Attrition Rates by Department**:
   - The **highest attrition rate** is observed in the **Sales** department.

- The **attrition rate** in the **Human Resources** department is also **very high**.

3. **Attrition in Research and Development**:
   - Although the **attrition rate** in the **Research and Development** department is **high**, it remains the **lowest** compared to other departments.

6. Analyzing Employee Attrition by Daily Rate

Note

1. **Daily Rate Definition**:
   - *DailyRate* represents the **daily wages** for the employees.
2. **Data Segmentation**:
   - To generate more **meaningful insights**, the daily rates can be **divided into three groups** (e.g., Low, Average, High) based on their values. This will help in better understanding the relationship between **daily wages** and **attrition rates**.

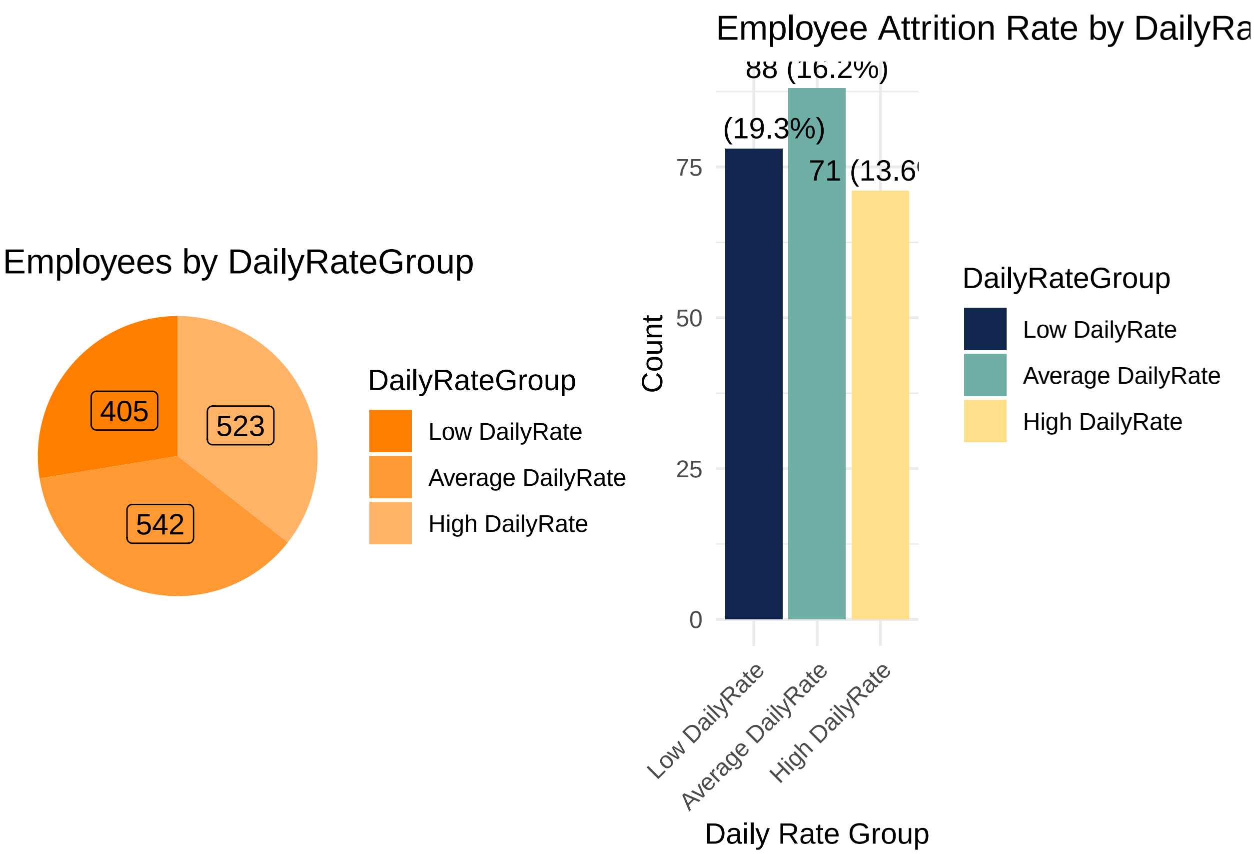
```r
# Creating salary groups
data$DailyRateGroup <- cut(data$DailyRate, breaks = c(0, 500, 1000, 1500),
                    labels = c("Low DailyRate", "Average DailyRate", "High DailyRate"))
head(data)
```

```
##   Age Attrition    BusinessTravel DailyRate              Department
## 1  41       Yes     Travel_Rarely      1102                   Sales
## 2  49        No Travel_Frequently       279 Research & Development
## 3  37       Yes     Travel_Rarely      1373 Research & Development
## 4  33        No Travel_Frequently      1392 Research & Development
## 5  27        No     Travel_Rarely       591 Research & Development
## 6  32        No Travel_Frequently      1005 Research & Development
##   DistanceFromHome     Education EducationField EnvironmentSatisfaction Gender
## 1                1       College  Life Sciences                  Medium Female
## 2                8 Below College  Life Sciences                    High   Male
## 3                2       College          Other               Very High   Male
## 4                3        Master  Life Sciences               Very High Female
## 5                2 Below College        Medical                     Low   Male
## 6                2       College  Life Sciences               Very High   Male
##   HourlyRate JobInvolvement     JobLevel               JobRole JobSatisfaction
## 1         94           High Junior Level       Sales Executive       Very High
## 2         61         Medium Junior Level    Research Scientist          Medium
## 3         92         Medium  Entry Level Laboratory Technician            High
## 4         56           High  Entry Level    Research Scientist            High
## 5         40           High  Entry Level Laboratory Technician          Medium
## 6         79           High  Entry Level Laboratory Technician       Very High
##   MaritalStatus MonthlyIncome MonthlyRate NumCompaniesWorked OverTime
## 1        Single          5993       19479                  8      Yes
## 2       Married          5130       24907                  1       No
## 3        Single          2090        2396                  6      Yes
## 4       Married          2909       23159                  1      Yes
## 5       Married          3468       16632                  9       No
## 6        Single          3068       11864                  0       No
##   PercentSalaryHike PerformanceRating RelationshipSatisfaction StockOptionLevel
## 1                11         Excellent                      Low                0
## 2                23       Outstanding                Very High                1
## 3                15         Excellent                   Medium                0
## 4                11         Excellent                     High                0
## 5                12         Excellent                Very High                1
```

```
## 6                13         Excellent                       High             0
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                 8                     0             Bad              6
## 2                10                     3          Better             10
## 3                 7                     3          Better              0
## 4                 8                     3          Better              8
## 5                 6                     3          Better              2
## 6                 8                     2            Good              7
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  4                       0                    5
## 2                  7                       1                    7
## 3                  0                       0                    0
## 4                  7                       3                    0
## 5                  2                       2                    2
## 6                  7                       3                    6
##      DailyRateGroup
## 1    High DailyRate
## 2     Low DailyRate
## 3    High DailyRate
## 4    High DailyRate
## 5 Average DailyRate
## 6    High DailyRate
```

```r
# Visualization to show Total Employees by DailyRate groups

daily_rate_counts <- data %>% count(DailyRateGroup)

# Create the pie chart
dailyrate_dist <-  ggplot(daily_rate_counts, aes(x = "", y = n, fill = DailyRateGroup)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Employees by DailyRateGroup") +
  theme_void() +
  scale_fill_manual(values = c("#FF8000", "#FF9933", "#FFB366")) +
  geom_label(aes(label = paste0(n)),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE)

# Visualization of Attrition by daily rate groups
attrition_data <- data %>% filter(Attrition == "Yes")

# Calculate daily rate group counts for all employees and those who left
daily_rate_counts <- data %>% count(DailyRateGroup)
attrition_by_rate <- attrition_data %>% count(DailyRateGroup)

# Merge data frames
merged_data <- data.frame(
  DailyRateGroup = daily_rate_counts$DailyRateGroup,
  Total_Employees = daily_rate_counts$n,
  Left = attrition_by_rate$n,
  Attrition_Rate = round((attrition_by_rate$n / daily_rate_counts$n) * 100, 1)
)

# Create the bar chart
```

```r
dailyrate_att_dist <-  ggplot(merged_data, aes(x = DailyRateGroup, y = Left, fill = DailyRateGroup)) +
  geom_bar(stat = "identity") +
  labs(title = "Employee Attrition Rate by DailyRateGroup", x = "Daily Rate Group", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#11264e", "#6faea4", "#FEE08B")) +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)")), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

grid.arrange(dailyrate_dist, dailyrate_att_dist, ncol=2)
```



Inference

1. **Daily Rate Distribution**:
   - The number of employees with **Average Daily Rate** and **High Daily Rate** is approximately equal.
2. **Attrition and Daily Rate**:
   - Employees with an **average daily rate** exhibit a **very high attrition rate** compared to those with a **high daily rate**.
3. **Attrition Among Low Daily Rate Employees**:
   - The attrition rate is also **very high** among employees with a **low daily rate**, indicating that salary may be a significant factor in employee retention.

7. Analyzing Employee Attrition by Distance From Home

```r
# Unique values in DistanceFromHome attribute
unique_distances <- unique(data$DistanceFromHome)
length(unique_distances)
```

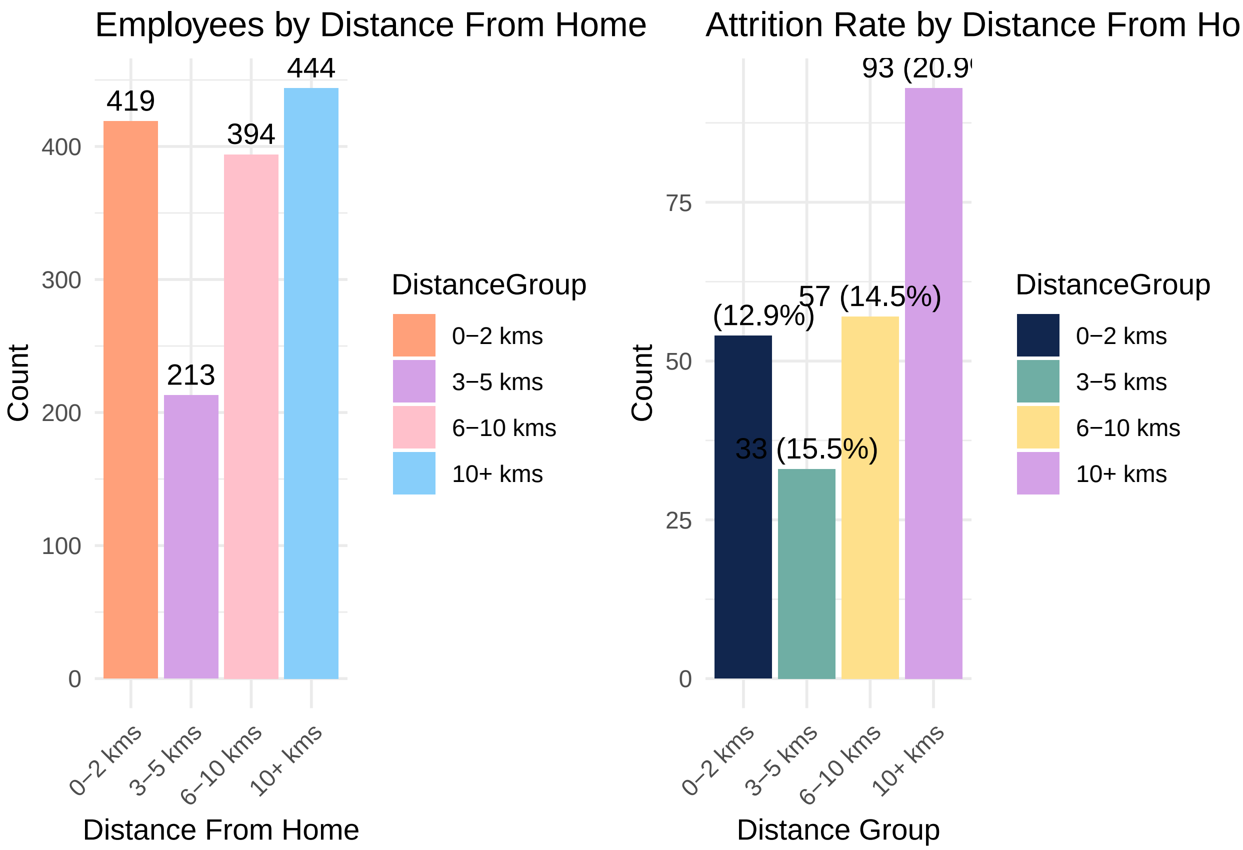
```
## [1] 29
```

```r
# description of distance from home attribute
distance_summary <- summary(data$DistanceFromHome)

distance_summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   7.000   9.193  14.000  29.000
```

```r
# Define the bins
data$DistanceGroup <- cut(data$DistanceFromHome, breaks = c(0, 2, 5, 10, 30), labels = c('0-2 kms', '3-5
```

```r
# Visualization to show total employees by distance from home
distance_group_count <- data %>% count(DistanceGroup)

distance_group_dist <- ggplot(data=distance_group_count, aes(x=DistanceGroup, y=n, fill=DistanceGroup))
  geom_bar(stat = "identity") +
  theme_minimal() +
  scale_fill_manual(values = c("#FFA07A", "#D4A1E7", "#FFC0CB","#87CEFA")) +
  labs(title="Employees by Distance From Home", x="Distance From Home", y="Count") +
  geom_text(aes(label = paste0(n)), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

#Visualization to show attrition rate by distance from home
attrition_data <- data %>% filter(Attrition == "Yes")

# Calculate distance group counts for all employees and those who left
distance_group_counts <- data %>% count(DistanceGroup)
attrition_by_distance <- attrition_data %>% count(DistanceGroup)

# Merge data frames
merged_data <- data.frame(
  DistanceGroup = distance_group_counts$DistanceGroup,
  Total_Employees = distance_group_counts$n,
  Left = attrition_by_distance$n,
  Attrition_Rate = round((attrition_by_distance$n / distance_group_counts$n) * 100, 1)
)

# Create the bar chart
att_by_distance_plot <-  ggplot(merged_data, aes(x = DistanceGroup, y = Left, fill = DistanceGroup)) +
  geom_bar(stat = "identity") +
  labs(title = "Attrition Rate by Distance From Home", x = "Distance Group", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#11264e", "#6faea4", "#FEE08B", "#D4A1E7", "#E7A1A1")) +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)")), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))



grid.arrange(distance_group_dist, att_by_distance_plot, ncol=2)
```



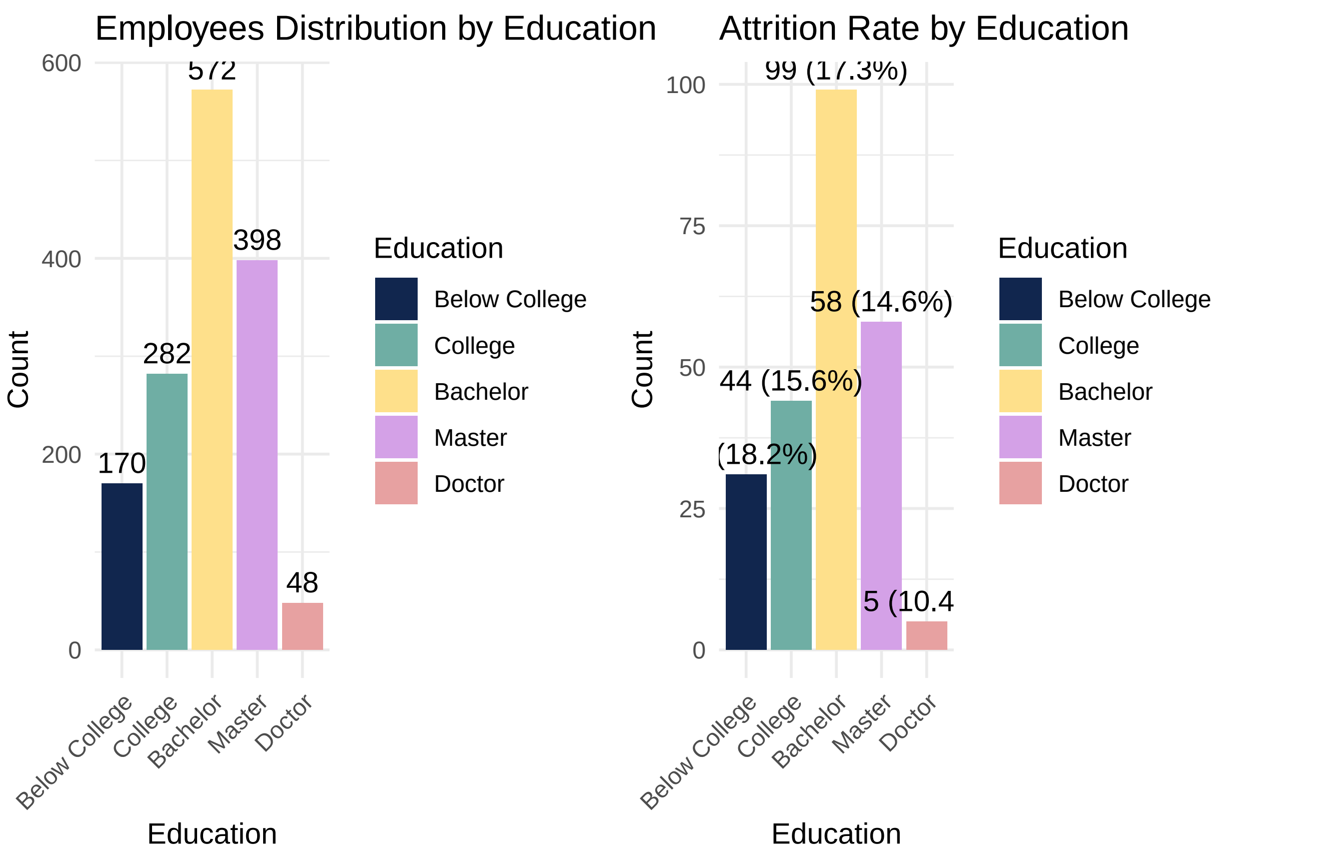
Inference

1. **Employee Commuting Distance Distribution**:
   - The organization has employees living in **close proximity** to the office, as well as others commuting from **farther distances**.
2. **Attrition Rates by Distance**:
   - There isn't a clear, observable trend in attrition rates based on commuting distance.
3. **Attrition Across Distance Groups**:
   - The attrition rate is **above 10%** for all **commuting distance groups**.
4. **Higher Attrition Among Long-Distance Commuters**:
   - Employees commuting **further than 10 km** from the company exhibit a **higher attrition rate** of **20.9%**, suggesting that long commuting distances may contribute to higher turnover.

8. Analyzing Employee Attrition by Education

```
# Visualization to show total employees by education
education_count <- data %>% count(Education)

education_dist <- ggplot(data=education_count, aes(x=Education, y=n, fill=Education)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title="Employees Distribution by Education", x="Education", y="Count") +
  scale_fill_manual(values=c("#11264e", "#6faea4", "#FEE08B", "#D4A1E7", "#E7A1A1")) +
  geom_text(aes(label = paste0(n)), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

# Visualization to show attrition by education
education_att_count <- attrition_data %>% count(Education)
```

```r
# Merge data frames
merged_data <- data.frame(
  Education = education_count$Education,
  Total_Employees = education_count$n,
  Left = education_att_count$n,
  Attrition_Rate = round((education_att_count$n / education_count$n) * 100, 1)
)

attrition_by_education <- ggplot(data=merged_data, aes(x=Education, y=Left, fill=Education)) +
  geom_bar(stat = "identity") +
  labs(title = "Attrition Rate by Education", x = "Education", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#11264e", "#6faea4", "#FEE08B", "#D4A1E7", "#E7A1A1")) +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)")), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

grid.arrange(education_dist, attrition_by_education, ncol=2)
```

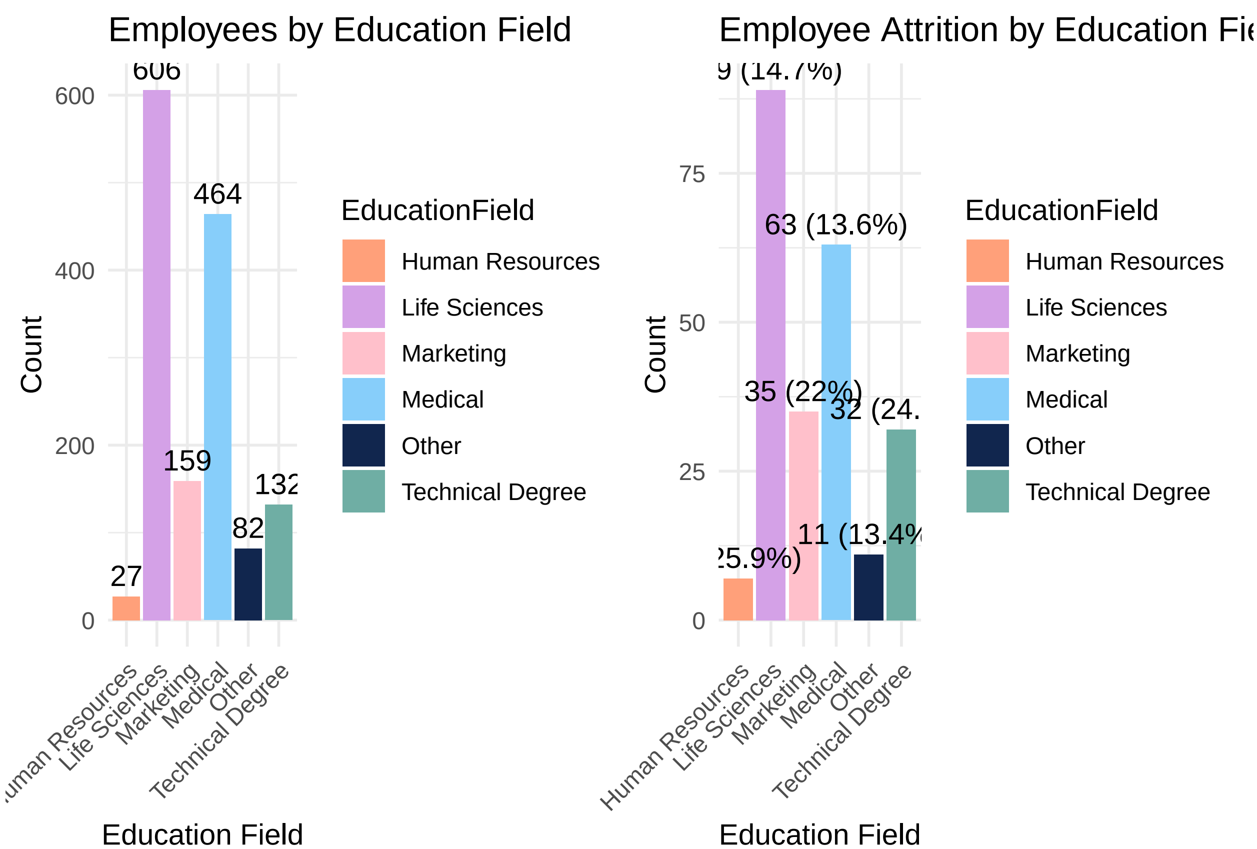


Inference

1. **Education Qualification Distribution**:
   - Most employees in the organization have completed **Bachelors** or **Masters** degrees.
2. **Employees with Doctorate Degrees**:
   - A **very small proportion** of employees in the organization have completed **Doctorate** degrees.
3. **Attrition Rates by Education Qualification**:
   - **Attrition rates decrease** as **education qualification increases**, suggesting that employees with higher qualifications may have more job stability or satisfaction.

9. Analyzing Employee Attrition by Education Field

```r
# Visualization showing total employees by education field
education_field_count <- data %>% count(EducationField)

education_field_dist <- ggplot(data=education_field_count, aes(x=EducationField, y=n, fill=EducationFiel
  geom_bar(stat="identity") +
  theme_minimal()+
  labs(title="Employees by Education Field", x="Education Field", y="Count") +
  scale_fill_manual(values = c("#FFA07A", "#D4A1E7", "#FFC0CB","#87CEFA","#11264e", "#6faea4")) +
  geom_text(aes(label=paste0(n)), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

# Visualization od attrition by education field
education_field_att_count <- attrition_data %>% count(EducationField)

merged_data <- data.frame(
  EducationField = education_field_count$EducationField,
  TotalEmployees = education_field_count$n,
  Left = education_field_att_count$n,
  Attrition_Rate = round((education_field_att_count$n / education_field_count$n) * 100, 1)
)

attrition_by_education_field <- ggplot(data=merged_data, aes(x=EducationField, y=Left, fill=EducationFi
  geom_bar(stat="identity") +
  theme_minimal() +
  scale_fill_manual(values = c("#FFA07A", "#D4A1E7", "#FFC0CB","#87CEFA","#11264e", "#6faea4")) +
  labs(title = "Employee Attrition by Education Field", x="Education Field", y="Count") +
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)")), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle =45, hjust =1))

grid.arrange(education_field_dist, attrition_by_education_field, ncol=2)
```



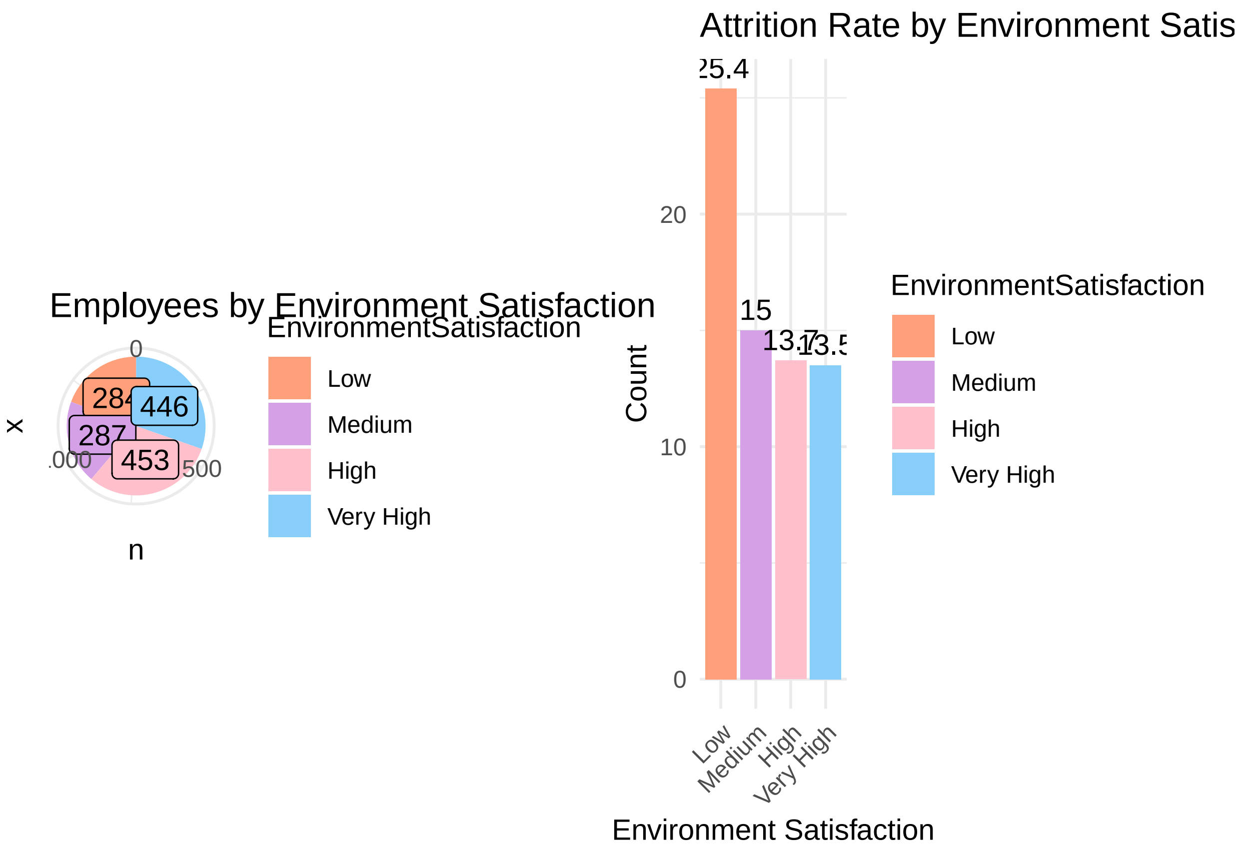
Inference

1. **Education Field Distribution**:
   - Most employees are from the **Life Science** (**606**) or **Medical** (**464**) education fields.
2. **Employees from Human Resources Education Field**:
   - Only a **small number** of employees come from the **Human Resources** (**27**) education field.
3. **Attrition Rates by Education Field**:
   - Education fields like **Human Resources** (**25.9%**), **Marketing** (**22%**), and **Technical** (**24.2%**) show **very high attrition rates**, indicating that employees in these fields may be more likely to leave the organization.

10. Analyzing employee Attrition by Environment Satisfaction

```
# Visualization to show total employees by job satisfaction
employee_satisfaction_count <- data %>% count(EnvironmentSatisfaction)

employee_satisfaction_dist <- ggplot(data=employee_satisfaction_count, aes(x="", y=n, fill=EnvironmentS
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  theme_minimal() +
  labs(title = "Employees by Environment Satisfaction") +
  scale_fill_manual(values = c("#FFA07A", "#D4A1E7", "#FFC0CB","#87CEFA")) +
  geom_label(aes(label = paste0(n)),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE)

# Visualization to show attrition by environment satisfaction
environment_satisfaction_att_count <- attrition_data %>% count(EnvironmentSatisfaction)
```

```
merged_data <- data.frame(
  EnvironmentSatisfaction = employee_satisfaction_count$EnvironmentSatisfaction,
  Total_Employees = employee_satisfaction_count$n,
  Left = environment_satisfaction_att_count$n,
  Attrition_Rate =round((environment_satisfaction_att_count$n/employee_satisfaction_count$n) *100, 1)
)

attrition_by_environment_sat <- ggplot(data=merged_data, aes(x=EnvironmentSatisfaction, y=Attrition_Rat
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title="Attrition Rate by Environment Satisfaction", x="Environment Satisfaction", y="Count") +
  geom_text(aes(label = paste0(Attrition_Rate)), vjust = -0.5, hjust = 0.5) +
  scale_fill_manual(values = c("#FFA07A", "#D4A1E7", "#FFC0CB","#87CEFA")) +
  theme(axis.text.x = element_text(angle =45, hjust=1))

grid.arrange(employee_satisfaction_dist, attrition_by_environment_sat, ncol=2)
```

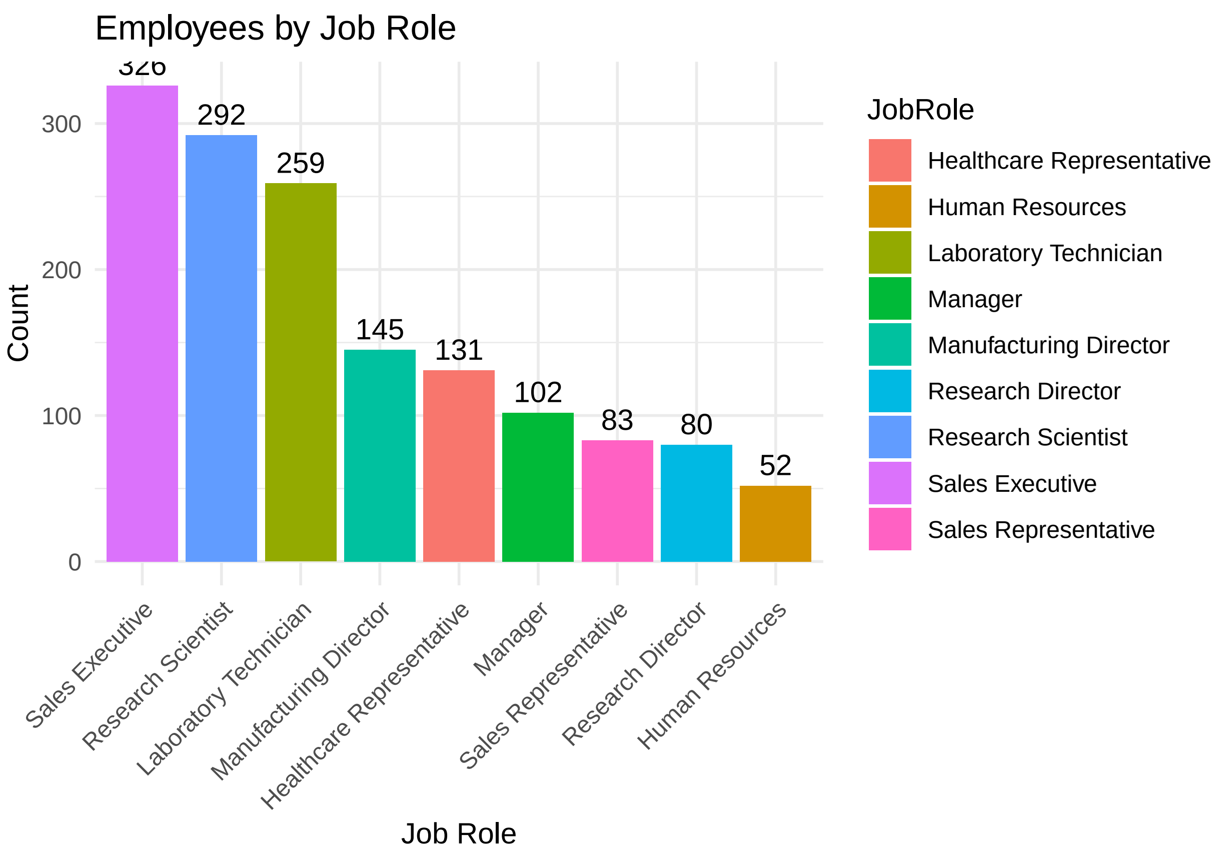


Inference

1. **Environment Satisfaction Distribution**:
   - Most employees have rated the organization's **environment satisfaction** as **High** or **Very High**.
2. **Attrition Despite High Environment Satisfaction**:
   - Despite high ratings for environment satisfaction, there is a **very high attrition rate** in this environment, suggesting other factors may be influencing retention.
3. **Attrition and Environment Satisfaction**:
   - **Attrition rates increase** as **environment satisfaction decreases**, indicating that a less satisfying work environment is a significant factor in employee turnover.

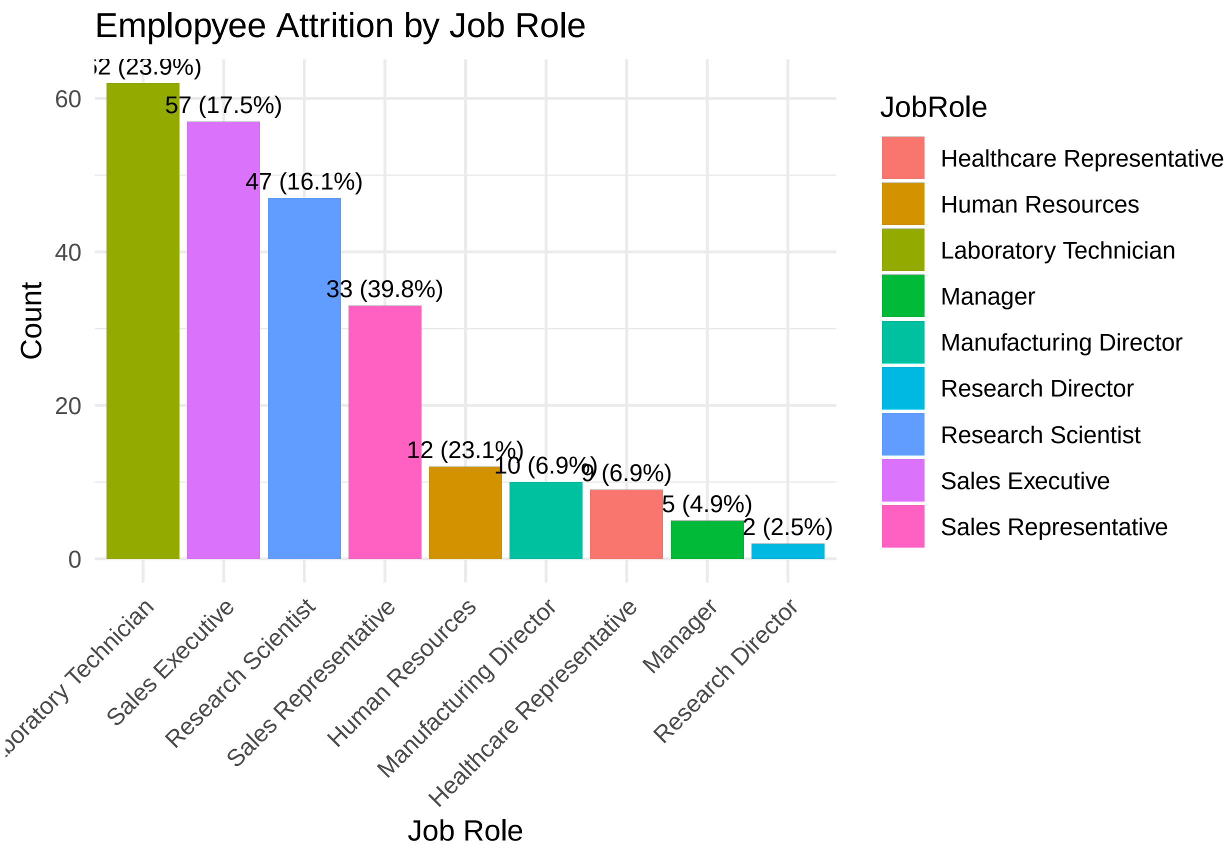
11. Analyzing Employee Attrition by Job Roles

```r
# Visualization to show total employees by Job Role
job_role_count <- data %>% count(JobRole)

job_role_dist <- ggplot(data=job_role_count, aes(x=reorder(JobRole, -n), y=n, fill=JobRole)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Employees by Job Role", x="Job Role", y="Count") +
  geom_text(aes(label=paste0(n)), vjust = -0.5, hjust = 0.5) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

grid.arrange(job_role_dist, ncol=1)
```




```r
# Visualization to show attrition by job role

job_role_att_count <- attrition_data %>% count(JobRole)

merged_data <- data.frame(
  JobRole = job_role_count$JobRole,
  TotalEmployees = job_role_count$n,
  Left = job_role_att_count$n,
  Attrition_Rate = round((job_role_att_count$n/job_role_count$n)*100, 1)
)

attrition_by_job_role <- ggplot(data=merged_data, aes(x=fct_reorder(JobRole, Left, .desc = TRUE), y=Lef
  geom_bar(stat="identity") +
```

```r
  theme_minimal() +
  labs(title="Employee Attrition by Job Role", x="Job Role", y="Count") +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)")), vjust=-0.5, hjust=0.5, size = 3) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

grid.arrange(attrition_by_job_role, ncol=1)
```

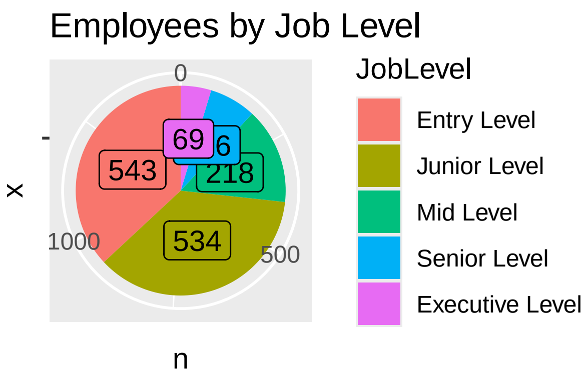


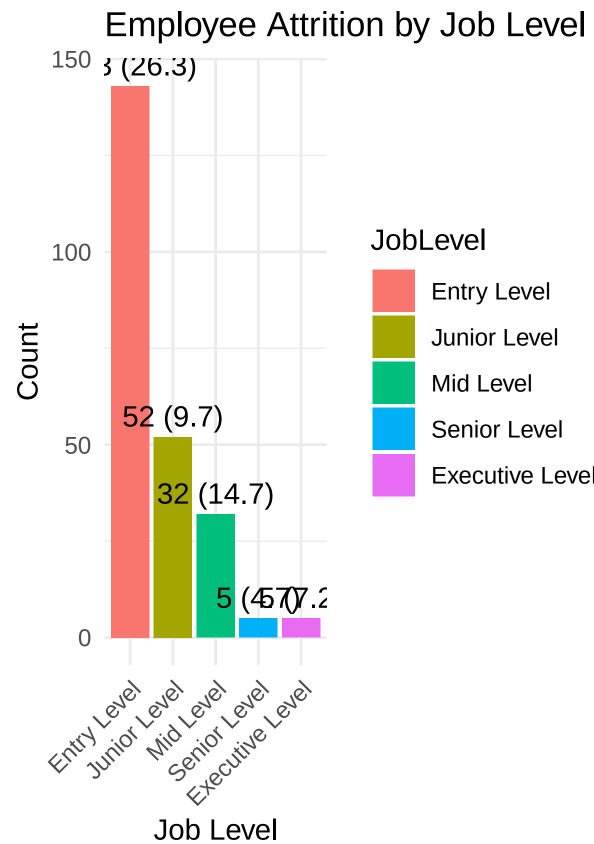
Inference

1. **Job Role Distribution**:
   - The majority of employees are working as **Sales Executives**, **Research Scientists**, and **Laboratory Technicians** in the organization.
2. **Attrition Rates by Job Role**:
   - The **highest attrition rates** are observed among employees in the following roles:
     - **Laboratory Technicians**
     - **Sales Representatives**
     - **Sales Executives**
     - **Research Scientists**
     - **Human Resources**
   - These roles have **attrition rates over 10%**, indicating potential areas for improvement in employee retention strategies.

12. Analyzing employee Attrition by Job Level

```r
# Visualization to show Total Employees by Job Level

job_level_count <- data %>% count(JobLevel)

job_level_dist <- ggplot(data=job_level_count, aes(x="", y=n, fill=JobLevel)) +
  geom_bar(stat="identity") +
  coord_polar("y", start = 0) +
  labs(title = "Employees by Job Level") +
  geom_label(aes(label = paste0(n)),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE)

# Visualization to show employee attrition by job level
job_level_att_count <- attrition_data %>% count(JobLevel)

merged_data <-  data.frame(
  JobLevel = job_level_count$JobLevel,
  TotalEmployees = job_level_count$n,
  Left = job_level_att_count$n,
  Attrition_Rate = round((job_level_att_count$n/job_level_count$n)*100, 1)
)

attrition_by_job_level <- ggplot(data=merged_data, aes(x=JobLevel, y=Left, fill=JobLevel)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title="Employee Attrition by Job Level", x="Job Level", y="Count") +
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, ")")), vjust=-0.5, hjust=0.5) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

grid.arrange(job_level_dist, attrition_by_job_level, ncol=2)
```





Inference

1. **Employee Job Level Distribution**:
   - The majority of employees in the organization are at an **Entry Level** or **Junior Level**.
2. **Attrition Among Entry-Level Employees**:
   - The **highest attrition rate** is observed among employees at the **Entry Level**, indicating potential challenges in retaining newer or less experienced employees.
3. **Attrition and Job Level Relationship**:
   - As the **job level increases**, the **attrition rate decreases**, suggesting that employees in higher positions may experience better job satisfaction or incentives to stay.

13. Analyzing Employee Attrition by Job Satisfaction

```
# Visualization to show total employees by job satisfaction
job_satisfaction_count <- data %>% count(JobSatisfaction)

job_satisfaction_dist <- ggplot(data=job_satisfaction_count, aes(x="", y=n, fill=JobSatisfaction)) +
  geom_bar(stat="identity") +
  coord_polar("y", start = 0) +
  theme_minimal() +
  labs(title="Employees by Job Satisfaction") +
  geom_label(aes(label=paste0(n)), position = position_stack(vjust = 0.5),
             show.legend = FALSE)

# Visualization to show attrition rate by Job Satisfaction
job_satisfaction_att_count <- attrition_data %>% count(JobSatisfaction)

merged_data <- data.frame(
```

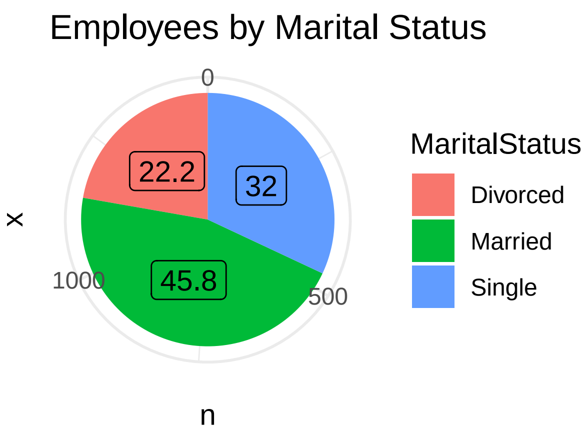
```
  JobSatisfaction = job_satisfaction_count$JobSatisfaction,
  TotalEmployees = job_satisfaction_count$n,
  Left = job_satisfaction_att_count$n,
  Attrition_Rate = round((job_satisfaction_att_count$n/job_satisfaction_count$n)*100, 1)
)

attrition_by_job_satisfaction <- ggplot(data=merged_data, aes(x=JobSatisfaction, y=Left, fill=JobSatisf
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title="Attrition by Employee Job Satisfaction", x="Job Satisfaction", y="Count") +
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)")), vjust=-0.5, hjust=0.5) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

grid.arrange(job_satisfaction_dist, attrition_by_job_satisfaction, ncol=2)
```

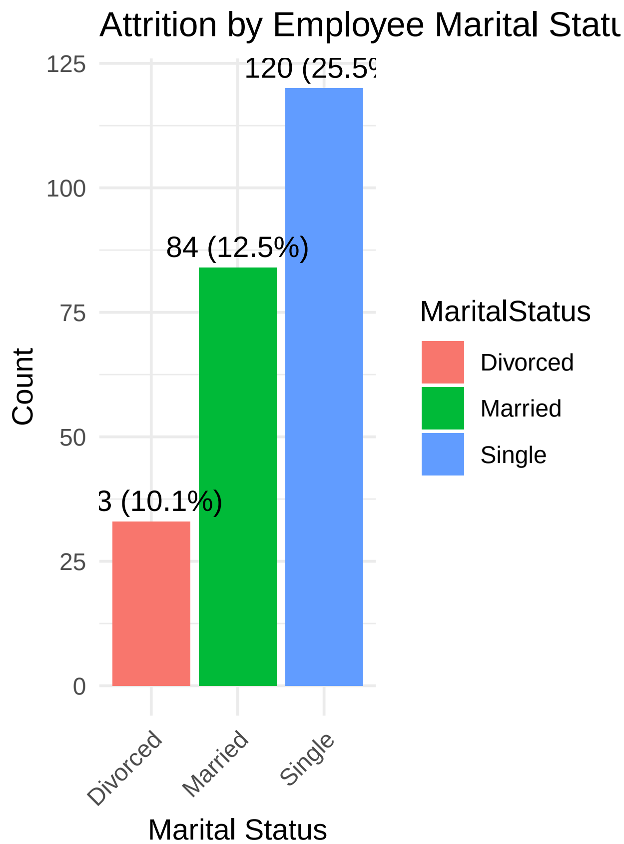
Inference

1. **Job Satisfaction Distribution**:
    - Most employees have rated their job satisfaction as **High** or **Very High**.
2. **Attrition Among Low Job Satisfaction**:
    - Employees who rated their job satisfaction as **Low** exhibit a **very high attrition rate** of **22.8%**, indicating dissatisfaction as a key driver of attrition.
3. **High Attrition Across All Categories**:
    - Despite differences in satisfaction levels, all job satisfaction categories exhibit **high attrition rates**, suggesting that additional factors may also influence employee retention.

14. Analyzing Employee Attrition by Marital Status

```
# Visualization of Total Employees by Marital Status
marital_status_count <- data %>% count(MaritalStatus)

marital_status_dist <- ggplot(data=marital_status_count, aes(x="", y=n, fill=MaritalStatus)) +
  geom_bar(stat="identity") +
  coord_polar("y", start=0) +
  theme_minimal() +
  labs(title="Employees by Marital Status") +
  geom_label(aes(label=paste0(round((n/sum(n))*100, 1))), position = position_stack(vjust=0.5), show.leg

# Visualization to show Attrition Rate by Marital Status
marital_status_att_count <- attrition_data %>% count(MaritalStatus)

merged_data <- data.frame(
  MaritalStatus = marital_status_count$MaritalStatus,
  Total_Employees = marital_status_count$n,
  Left = marital_status_att_count$n,
  Attrition_Rate = round((marital_status_att_count$n/marital_status_count$n)*100, 1)
)

attrition_by_marital_status <- ggplot(data=merged_data, aes(x=MaritalStatus, y=Left, fill=MaritalStatus)
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title="Attrition by Employee Marital Status", x="Marital Status", y="Count") +
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)")), vjust=-0.5, hjust=0.5) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))


grid.arrange(marital_status_dist, attrition_by_marital_status ,ncol=2)
```





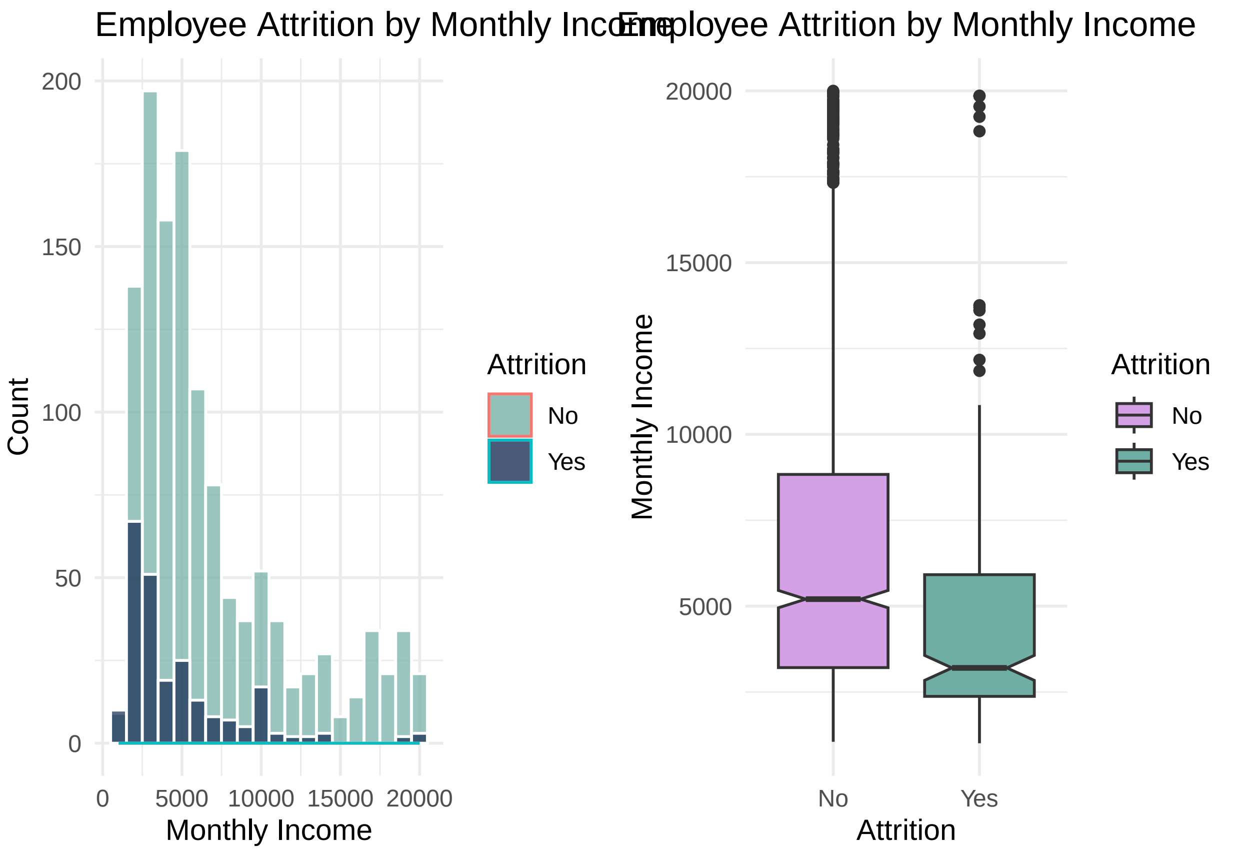
Inference

1. **Marital Status Distribution**:
   - The majority of employees in the organization are **married**.
2. **Attrition Among Singles**:
   - **Single employees** exhibit a **very high attrition rate**, indicating they may be more likely to leave the organization.
3. **Attrition Among Divorced Employees**:
   - The attrition rate is **lowest** among **divorced employees**, suggesting they might be more stable in their roles.

15. Analyzing Employee Attrition by Monthly Income

```r
# Visualization to show Employee Distribution by Monthly Income
monthly_income_plot <- ggplot(data, aes(x = MonthlyIncome, fill = Attrition)) +
  geom_histogram(alpha = 0.7, position = "identity", bins = 20, color = "white") +
  geom_density(alpha = 0.2, aes(color = Attrition)) +
  labs(title = "Employee Attrition by Monthly Income", x = "Monthly Income", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#6faea4", "#11264e"))

# Visualization to show Employee Attrition by Monthly Income
attrition_by_monthly_income <- ggplot(data, aes(x = Attrition, y = MonthlyIncome, fill = Attrition)) +
  geom_boxplot(notch = TRUE) +  # Add notches for better comparison between groups
  labs(title = "Employee Attrition by Monthly Income", x = "Attrition", y = "Monthly Income") +
  theme_minimal() +
  scale_fill_manual(values = c("#D4A1E7", "#6faea4")) +
  theme(plot.title = element_text(hjust = 0.5))  # Center title horizontally
```

```r
grid.arrange(monthly_income_plot, attrition_by_monthly_income, ncol=2)
```




```r
mean_monthly_income <- data %>%
  select(Attrition, MonthlyIncome) %>%
  group_by(Attrition) %>%
  summarize(mean_income = mean(MonthlyIncome))

mean_monthly_income
```

```
## # A tibble: 2 x 2
##   Attrition mean_income
##   <chr>           <dbl>
## 1 No              6833.
## 2 Yes             4787.
```

Inference

1. **Monthly Income Distribution**:
   - The majority of employees in the organization earn **less than 10,000**.
2. **Attrition and Income Comparison**:
   - The **average monthly income** of employees who have left is **comparatively lower** than that of employees who are still working.
3. **Income and Attrition Relationship**:
   - As **monthly income increases**, the **attrition rate decreases**, indicating that higher income may act as a retention factor.

16. Analyzing Employee Attrition by Number of Companies Worked

```
data %>%
  select(NumCompaniesWorked) %>%
  summary()
```

```
##  NumCompaniesWorked
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.693
##  3rd Qu.:4.000
##  Max.   :9.000
```

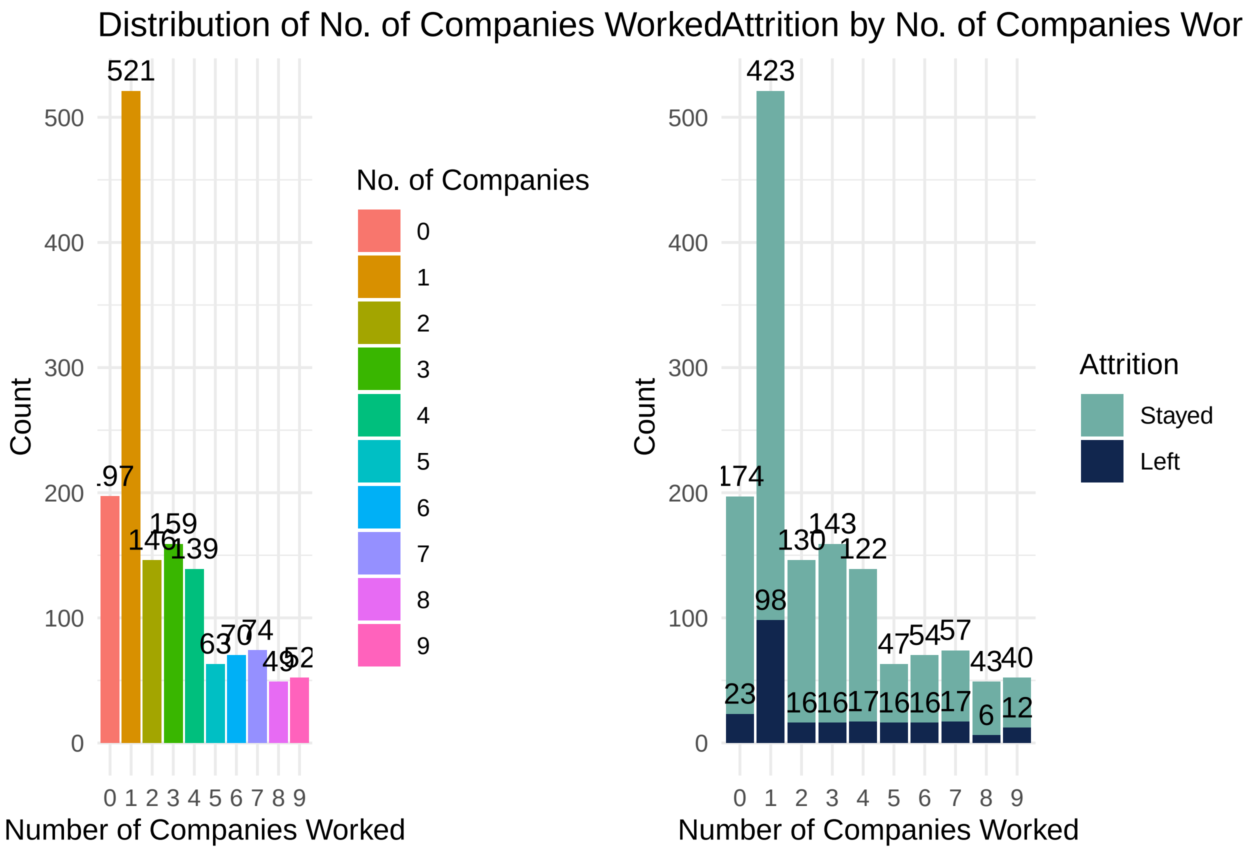
```
number_companies_worked_count <- data %>% count(NumCompaniesWorked)

number_companies_worked_count
```

```
##    NumCompaniesWorked   n
## 1                   0 197
## 2                   1 521
## 3                   2 146
## 4                   3 159
## 5                   4 139
## 6                   5  63
## 7                   6  70
## 8                   7  74
## 9                   8  49
## 10                  9  52
```

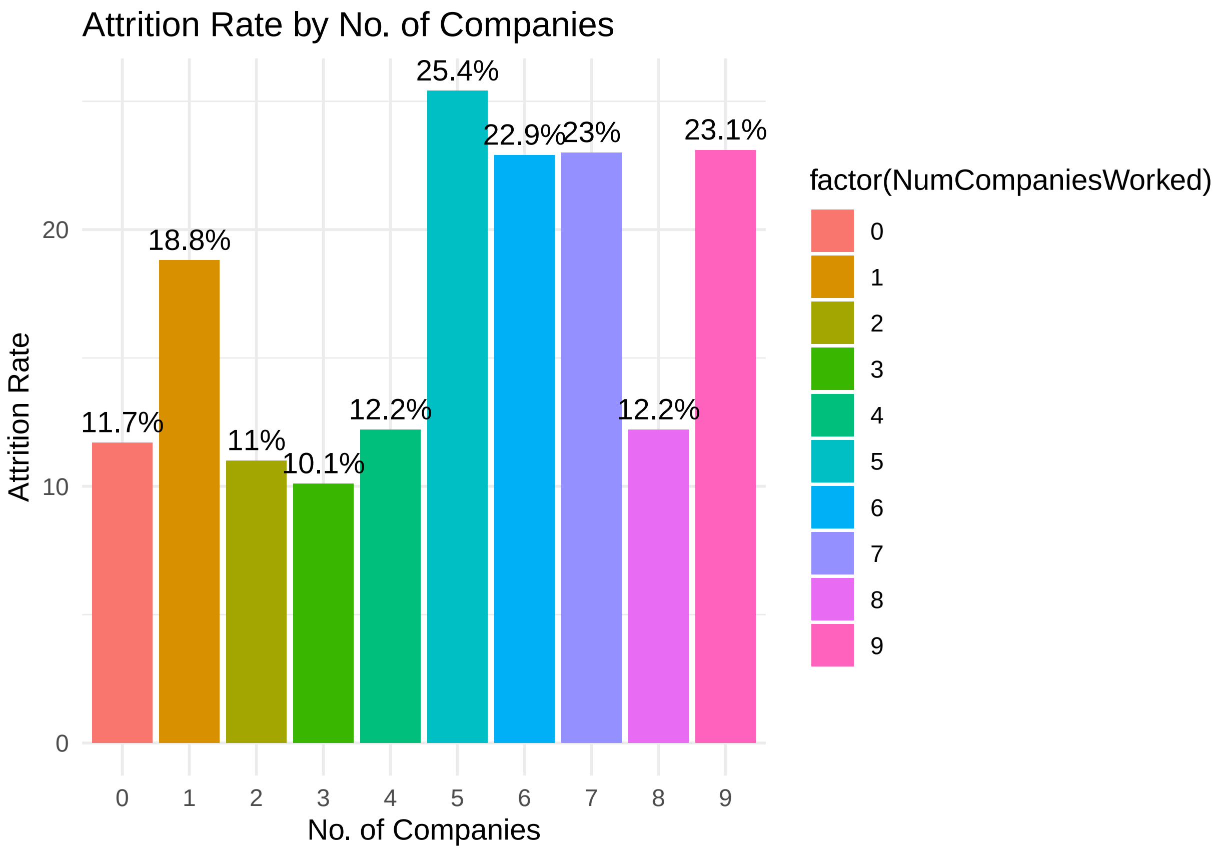
```
# Visualization to show employee number of companies worked distribution
number_of_companies_dist <-  ggplot(number_companies_worked_count, aes(x = factor(NumCompaniesWorked), y
  geom_bar(stat = "identity") +
  labs(title = "Distribution of No. of Companies Worked", x = "Number of Companies Worked", y = "Count"
  theme_minimal() +
  geom_text(aes(label = paste0(n), vjust = -.5, hjust=0.5))

# Calculate the number of employees who left and stayed for each number of companies worked
attrition_by_num_companies <- data %>%
  group_by(NumCompaniesWorked, Attrition) %>%
  summarize(count = n())
```

```
## `summarise()` has grouped output by 'NumCompaniesWorked'. You can override
## using the `.groups` argument.
```

```
# Create the stacked bar chart
attrition_by_number_of_companies <- ggplot(attrition_by_num_companies, aes(x = factor(NumCompaniesWorke
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(count), vjust=-0.5, hjust=0.5), position = position_stack(vjust = 1.0, re
  scale_x_discrete(labels = unique(attrition_by_num_companies$NumCompaniesWorked)) +
  labs(title = "Attrition by No. of Companies Worked", x = "Number of Companies Worked", y = "Count") +
  theme_minimal() +
  scale_fill_manual(values = c("#6faea4", "#11264e"), labels = c("Stayed", "Left"))

grid.arrange(number_of_companies_dist, attrition_by_number_of_companies, ncol=2)
```




```r
# Visualization to show just the attrition rates by NUmber of Companies Worked
number_companies_worked_att_count <- attrition_data %>% count(NumCompaniesWorked)

merged_data <- data.frame(
  NumCompaniesWorked = number_companies_worked_count$NumCompaniesWorked,
  TotalEmployees = number_companies_worked_count$n,
  Left = number_companies_worked_att_count$n,
  Attrition_Rate = round((number_companies_worked_att_count$n/number_companies_worked_count$n)*100, 1)
)

attrition_rate_by_no_companies <- ggplot(data=merged_data, aes(x=factor(NumCompaniesWorked), y=Attritio
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Attrition Rate by No. of Companies", x="No. of Companies", y="Attrition Rate") +
  geom_text(aes(label=paste0(Attrition_Rate, "%"), vjust=-0.5, hjust=0.5))

grid.arrange(attrition_rate_by_no_companies, ncol=1)
```



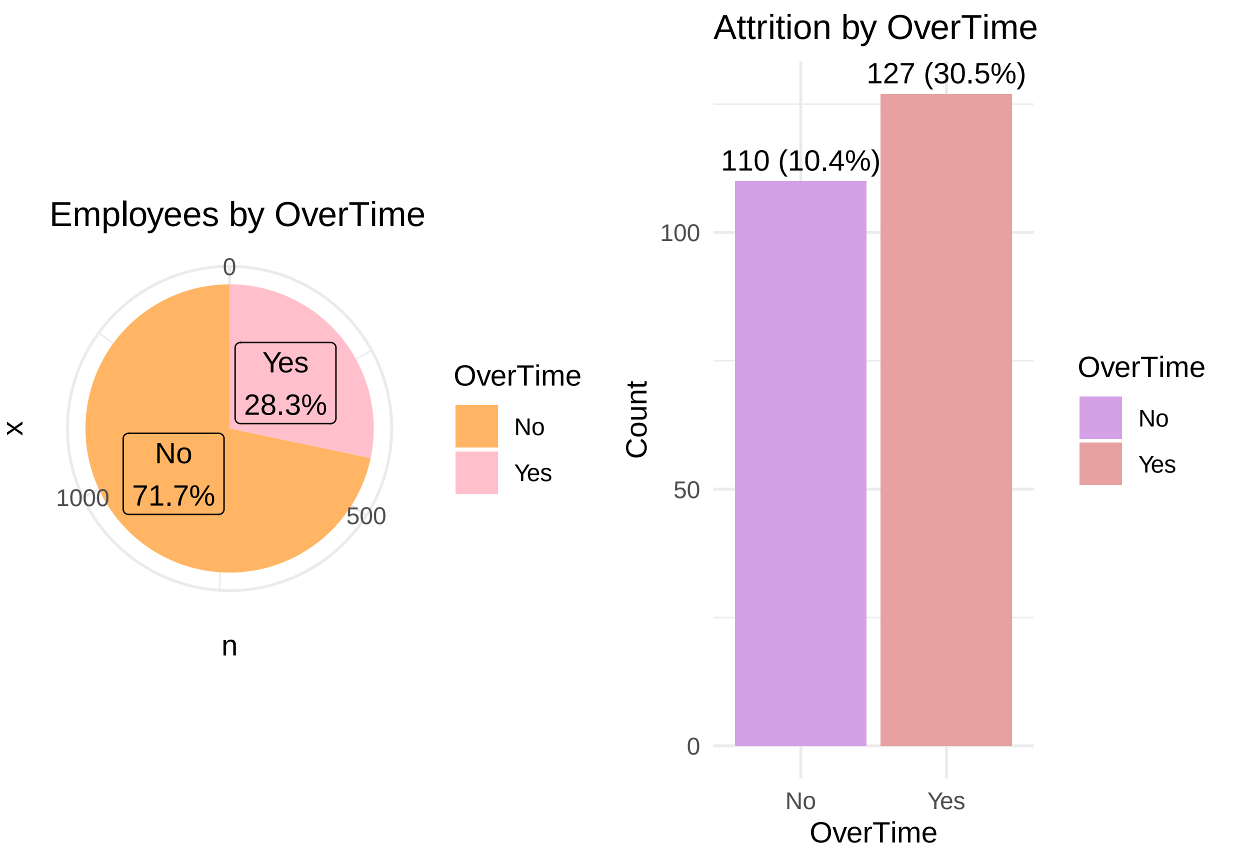
Inference

1. **Number of Companies Worked Distribution**:
   - The majority of employees have worked for **fewer than 2 companies**.
2. **Attrition Among Employees with Limited Experience**:
   - Employees who have worked for **1 company** exhibit a **high attrition rate**.
3. **Attrition Trends Across Multiple Companies**:
   - There is a noticeable **increase in attrition rates** among employees who have worked for **5-7 companies**, suggesting a possible trend of job-hopping behavior.

17. Analyzing Employee Attrition by Over Time

```
# Visualization to show Total Employees by Overtime.
overtime_count <- data %>% count(OverTime)

overtime_dist <- ggplot(data=overtime_count, aes(x="", y=n, fill=OverTime)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  coord_polar("y", start=0) +
  labs(title="Employees by OverTime") +
  scale_fill_manual(values=c("#ffb563","#FFC0CB")) +
  geom_label(aes(label=paste0(OverTime,"\n",round(n/sum(n)*100, 1), "%")), position = position_stack(vj

# Visualization to show Attrition Rate by Overtime
overtime_att_count <- attrition_data %>% count(OverTime)
```

```r
merged_data <- data.frame(
  OverTime = overtime_count$OverTime,
  TotalEmployees = overtime_count$n,
  Left = overtime_att_count$n,
  Attrition_Rate = round((overtime_att_count$n/overtime_count$n)*100, 1)
)

attrition_by_overtime <- ggplot(data=merged_data, aes(x=OverTime, y=Left, fill=OverTime)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title="Attrition by OverTime", x="OverTime", y="Count") +
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)"), vjust=-0.5, hjust=0.5)) +
  scale_fill_manual(values = c("#D4A1E7","#E7A1A1"))

grid.arrange(overtime_dist, attrition_by_overtime, ncol=2)
```



Inference

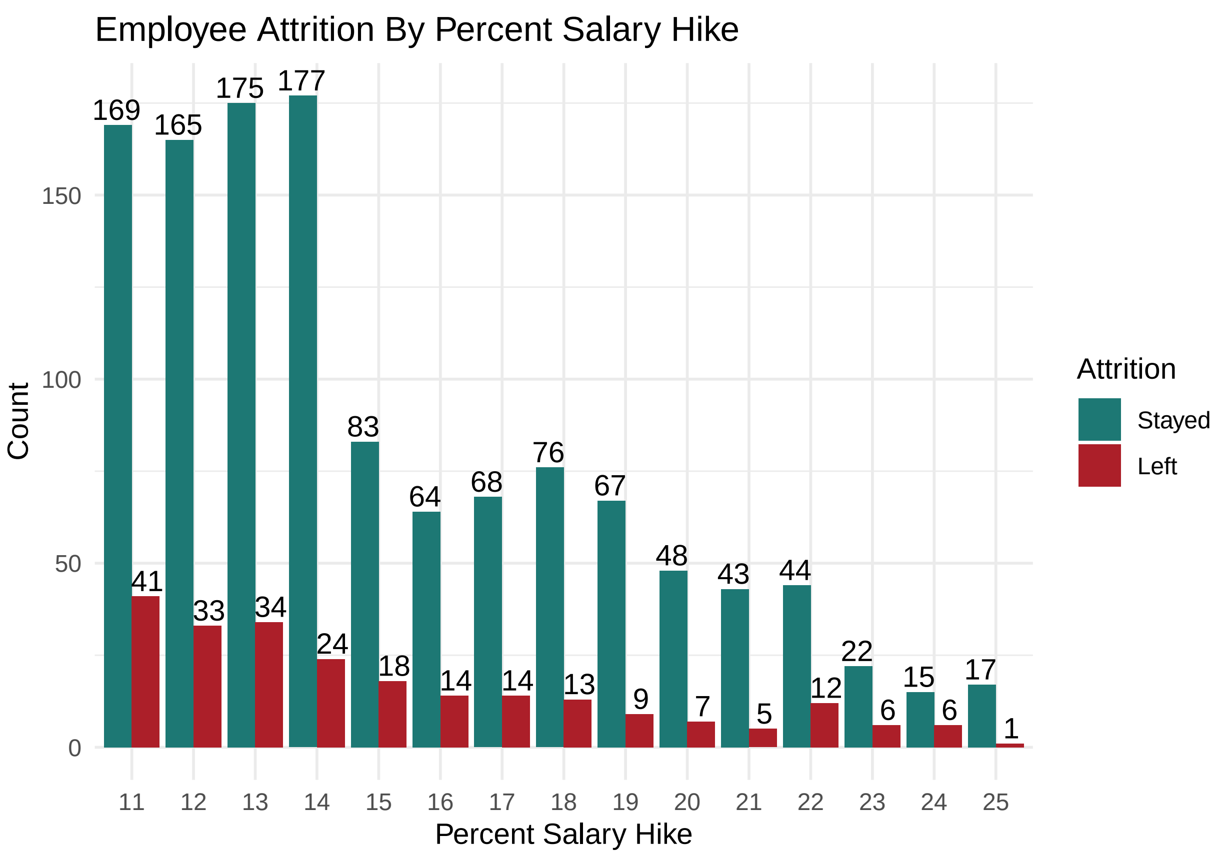
1. **OverTime Distribution**:
    - Approximately **72%** of employees in the organization do not work overtime.
2. **Attrition and OverTime**:
    - The attrition rate is **higher** among employees who work overtime.

    - However, the **OverTime** feature exhibits a significant class imbalance, limiting the ability to draw **meaningful insights** from this attribute.

18. Analyzing Employee Attrition by Percentage Salary Hike

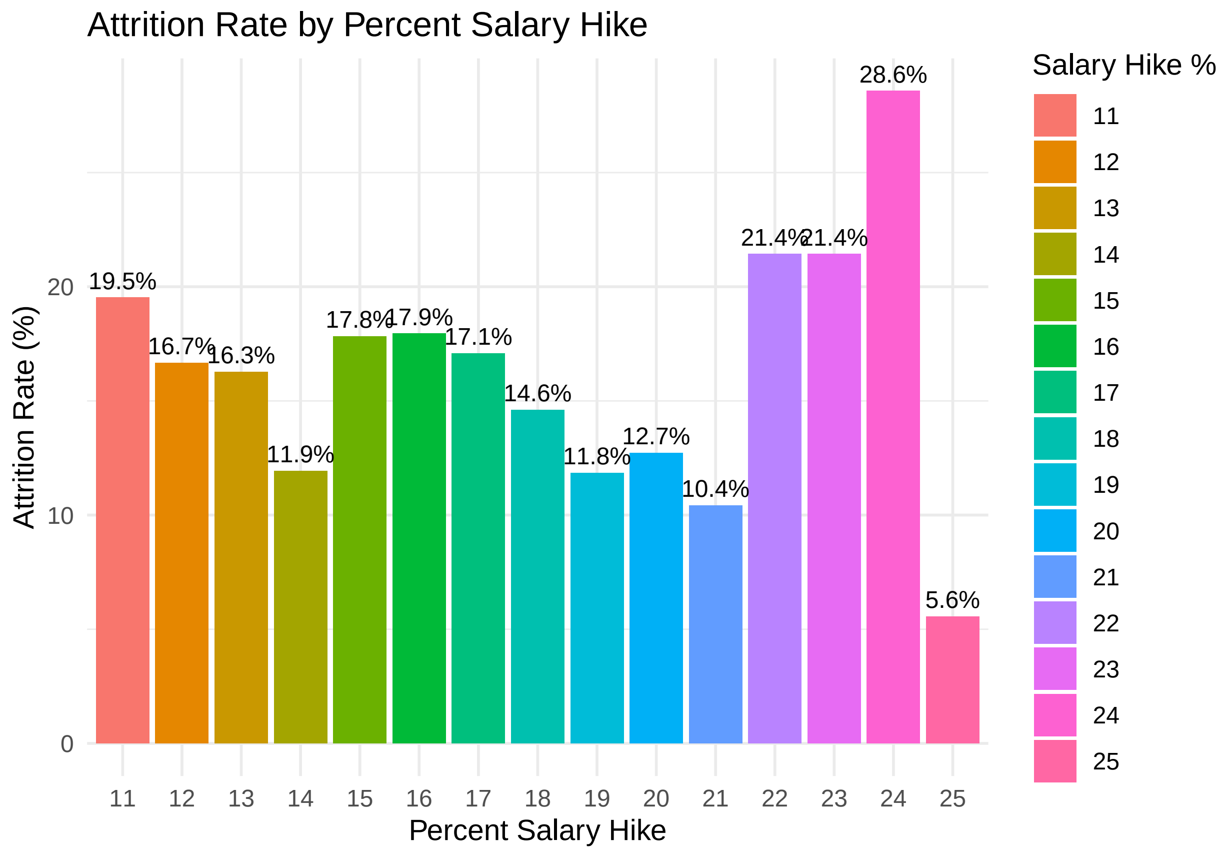
```r
salary_hike_dist <- ggplot(data, aes(x = factor(PercentSalaryHike), fill = Attrition)) +
  geom_bar(stat = "count", position = "dodge") +
  labs(title = "Employee Attrition By Percent Salary Hike",
       x = "Percent Salary Hike",
       y = "Count",
       fontweight = "bold",
       title.size = 20,
       title.x = 0.5,
       title.y = 1.0) +
  theme_minimal() +
  scale_fill_manual(values = c("#1d7874", "#AC1F29"), labels = c("Stayed", "Left")) +
  geom_text(aes(label = ..count..), stat = "count", position = position_dodge(width = 1), vjust = -0.25)

grid.arrange(salary_hike_dist, ncol=1)
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```r
# Attrition Rate by percent salary hike
# Calculate attrition rate for each salary hike
attrition_rate <- data %>%
  group_by(PercentSalaryHike) %>%
```

```
  summarize(
    Total = n(),
    Left = sum(Attrition == "Yes"),
    Attrition_Rate = (Left / Total) * 100
  )

# Create a bar plot
attrition_rate_by_salaryhike <-  ggplot(attrition_rate, aes(x =factor( PercentSalaryHike), y = Attritio
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(Attrition_Rate, 1), "%"), vjust = -0.5), size = 3) +
  labs(title = "Attrition Rate by Percent Salary Hike", x = "Percent Salary Hike", y = "Attrition Rate
  theme_minimal()

grid.arrange(attrition_rate_by_salaryhike, ncol=1)
```



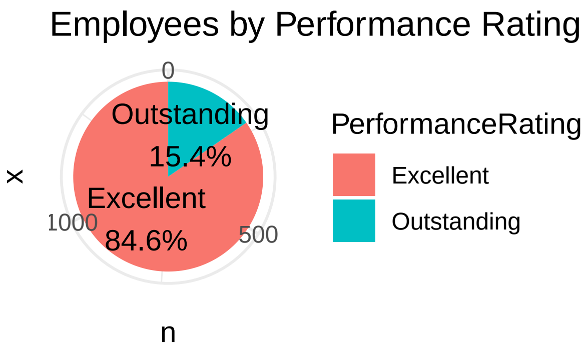
Inference

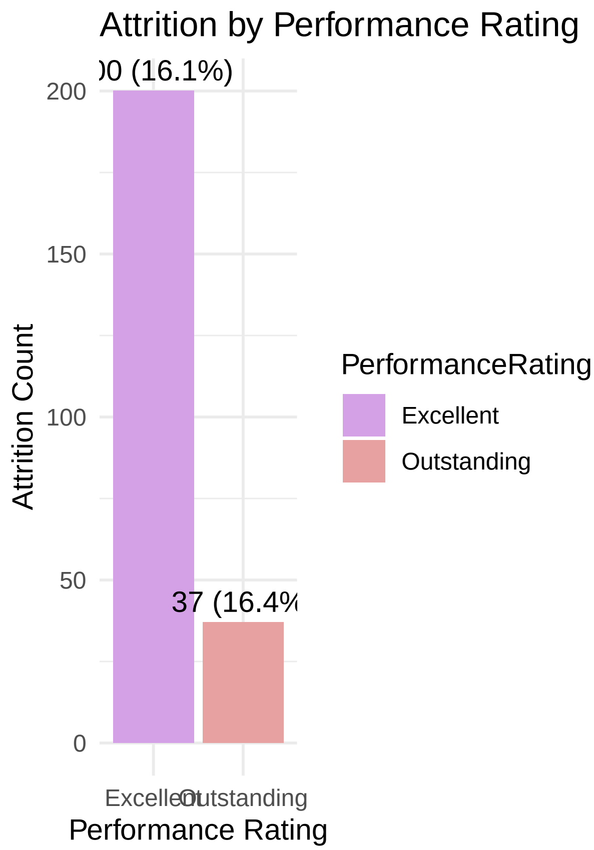
1. **Salary Hike Distribution**:
    - Only a **small proportion of employees** receive a **high percentage salary hike**.
2. **Attrition and Salary Hikes**:
    - Higher salary hikes are generally associated with **lower attrition rates**.

    - However, the data for salary hikes above **20%** is limited, making it challenging to draw **definitive conclusions** about their impact on attrition.

19. Analyzing Employee Attrition by Performance Rating

```
# Visualization to show total employees by performance rating
performance_rating_count <- data %>% count(PerformanceRating)

performance_rating_dist <- ggplot(data=performance_rating_count, aes(x="", y=n, fill=PerformanceRating))
  geom_bar(stat="identity") +
  coord_polar("y", start=0) +
  theme_minimal() +
  labs(title="Employees by Performance Rating") +
  geom_text(aes(label=paste0(PerformanceRating, "\n", round((n/sum(n))*100, 1), "%")), position = positi

# Visualization to show attrition rate by Performance Rating
performance_rating_att_count <- attrition_data %>% count(PerformanceRating)

merged_data <- data.frame(
  PerformanceRating = performance_rating_count$PerformanceRating,
  TotalEmployees = performance_rating_count$n,
  Left = performance_rating_att_count$n,
  Attrition_Rate = round((performance_rating_att_count$n/performance_rating_count$n)*100, 1)
)

attrition_rate_by_performance <- ggplot(data=merged_data, aes(x=PerformanceRating, y=Left, fill=Performa
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title="Attrition by Performance Rating", x="Performance Rating", y="Attrition Count") +
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)"), vjust=-0.5, hjust=0.5)) +
  scale_fill_manual(values = c("#D4A1E7","#E7A1A1"))

grid.arrange(performance_rating_dist, attrition_rate_by_performance, ncol=2)
```





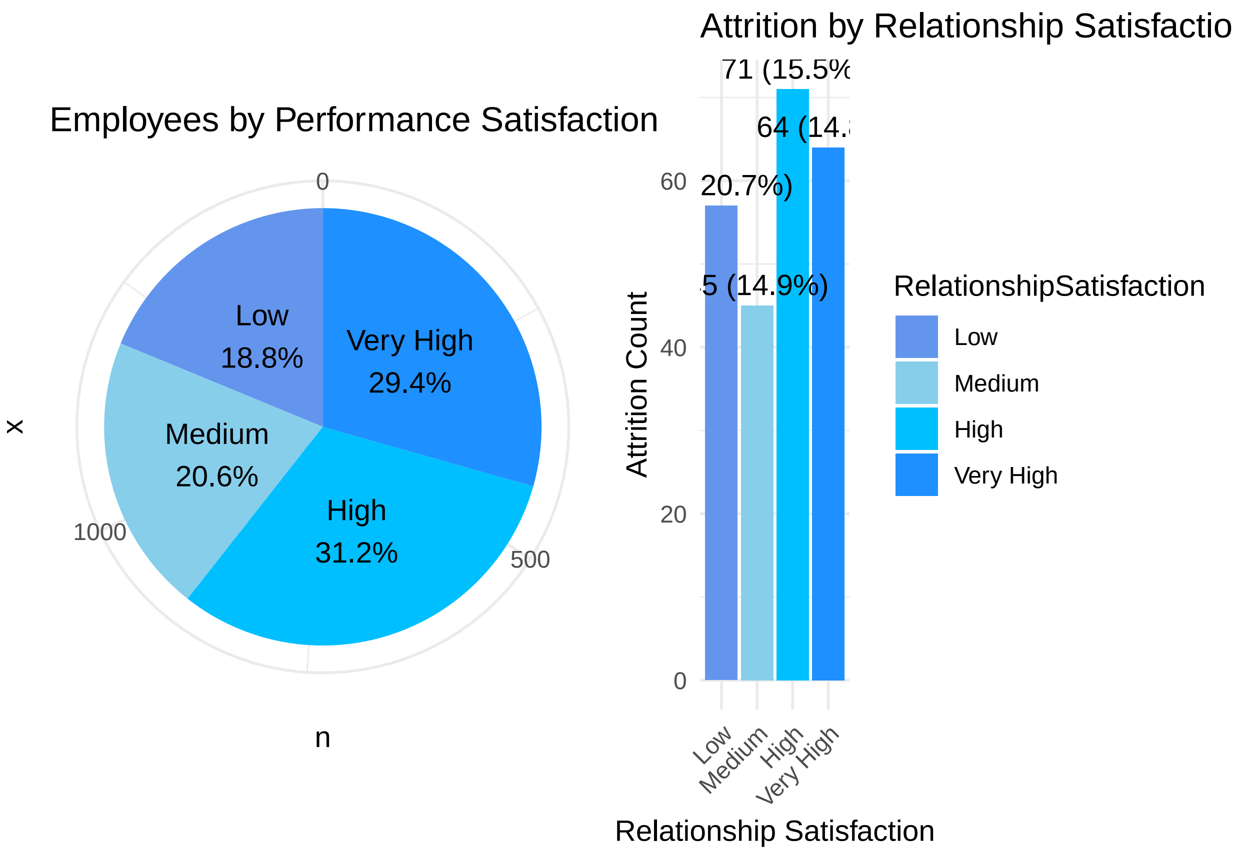
Inference

1. **Performance Ratings Distribution**:
   - The majority of employees have received **excellent performance ratings**.
2. **Attrition Rates by Performance Ratings**:
   - All performance rating categories exhibit **similar attrition rates**.
3. **Conclusion**:
   - The **Performance Ratings** attribute does not provide meaningful insights for understanding or predicting employee attrition.

20. Analyzing Employee Attrition by Relationship Satisfaction

```
# Visualization to show total employees by relationship satisfaction
relationship_satisfaction_count <- data %>% count(RelationshipSatisfaction)

relationship_satisfaction_dist <- ggplot(data=relationship_satisfaction_count, aes(x="", y=n, fill=Rela
  geom_bar(stat="identity") +
  coord_polar("y", start=0) +
  theme_minimal() +
  labs(title="Employees by Performance Satisfaction") +
  geom_text(aes(label=paste0( RelationshipSatisfaction, "\n", round((n/sum(n))*100, 1), "%")), position
  scale_fill_manual(values=c('#6495ED', '#87CEEB', '#00BFFF', '#1E90FF')) +
  theme(legend.position = "none")


# Visualization to show Attrition Rate by relationship satisfaction
relationship_satisfaction_att_count <- attrition_data %>% count(RelationshipSatisfaction)
```

```
merged_data = data.frame(
  RelationshipSatisfaction = relationship_satisfaction_count$RelationshipSatisfaction,
  TotalEmployees = relationship_satisfaction_count$n,
  Left = relationship_satisfaction_att_count$n,
  Attrition_Rate = round((relationship_satisfaction_att_count$n / relationship_satisfaction_count$n)*10
)

attrition_by_relationship_satisfaction <- ggplot(data=merged_data, aes(x=RelationshipSatisfaction, y=Le
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Attrition by Relationship Satisfaction", x="Relationship Satisfaction", y="Attrition Co
  scale_fill_manual(values = c('#6495ED', '#87CEEB', '#00BFFF', '#1E90FF')) +
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)"), vjust=-0.5, hjust=0.5)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

grid.arrange(relationship_satisfaction_dist, attrition_by_relationship_satisfaction, ncol=2)
```

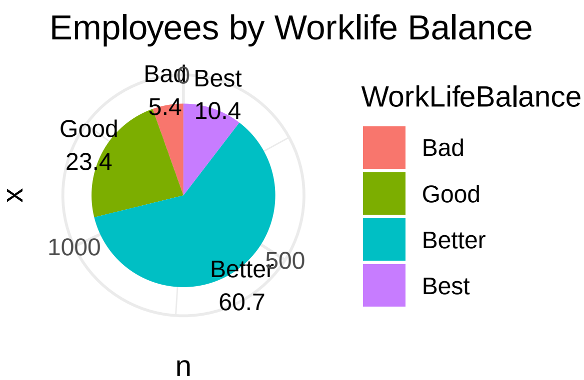


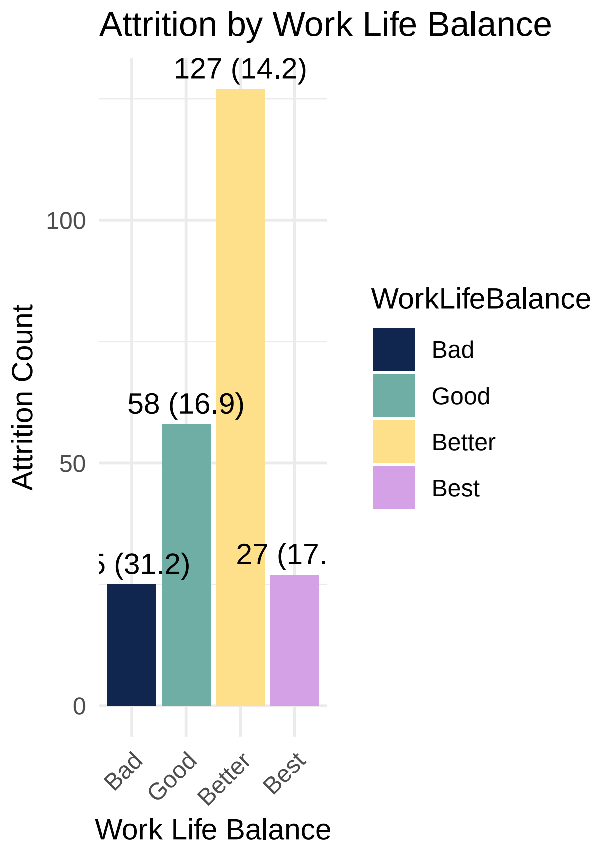
Inference

1. **Relationship Satisfaction Distribution**:
   - Most employees report having **high** or **very high relationship satisfaction**.
2. **Attrition Despite High Satisfaction**:
   - Despite the high relationship satisfaction levels, there is still a **high attrition rate** observed across the workforce.
3. **Attrition Across All Levels**:
   - All levels of relationship satisfaction exhibit **high attrition rates**, suggesting that this attribute alone may not be a strong predictor of employee retention.

21. Analyzing Employee Attrition by Work Life balance

```r
# visualization to show Total Employees by Work Life Balance
worklife_balance_count <- data %>% count(WorkLifeBalance)

worklife_balance_dist <- ggplot(data=worklife_balance_count, aes(x="", y=n, fill=WorkLifeBalance)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  coord_polar("y", start=0) +
  labs(title="Employees by Worklife Balance") +
  geom_text(aes(label=paste0(WorkLifeBalance, "\n",round((n/sum(n))*100, 1)), x=1.6), position = positi

# Visualization to show attrition by worklife balance
worklife_balance_att_count <- attrition_data %>% count(WorkLifeBalance)

merged_data <- data.frame(
  WorkLifeBalance = worklife_balance_count$WorkLifeBalance,
  TotalEmployees = worklife_balance_count$n,
  Left = worklife_balance_att_count$n,
  Attrition_Rate = round((worklife_balance_att_count$n / worklife_balance_count$n)*100, 1)
)

attrition_by_worklife_balance <- ggplot(data=merged_data, aes(x=WorkLifeBalance, y=Left, fill=WorkLifeBa
  geom_bar(stat="identity") +
  theme_minimal()+
  labs(title="Attrition by Work Life Balance", x="Work Life Balance", y="Attrition Count") +
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, ")"), vjust=-0.5, hjust=0.5)) +
  scale_fill_manual(values=c("#11264e","#6faea4","#FEE08B","#D4A1E7","#E7A1A1")) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

grid.arrange(worklife_balance_dist, attrition_by_worklife_balance, ncol=2)
```





Inference

1. **Work-Life Balance Distribution**:
   - Over **60%** of employees report having a **"Better" work-life balance**.
2. **Attrition Rates by Work-Life Balance**:
   - Employees with a **"Bad" work-life balance** exhibit a **very high attrition rate**.

   - Despite this, the other categories also show **high attrition rates**, indicating that work-life balance significantly impacts employee retention across all levels.

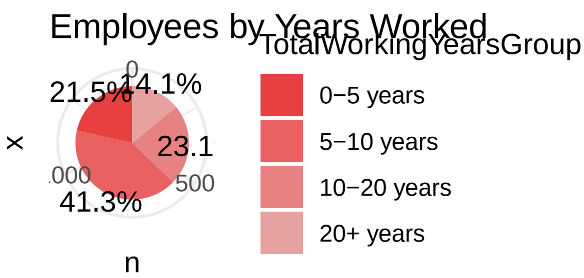
22. Analyzing Employee Attrition by Total Working Years

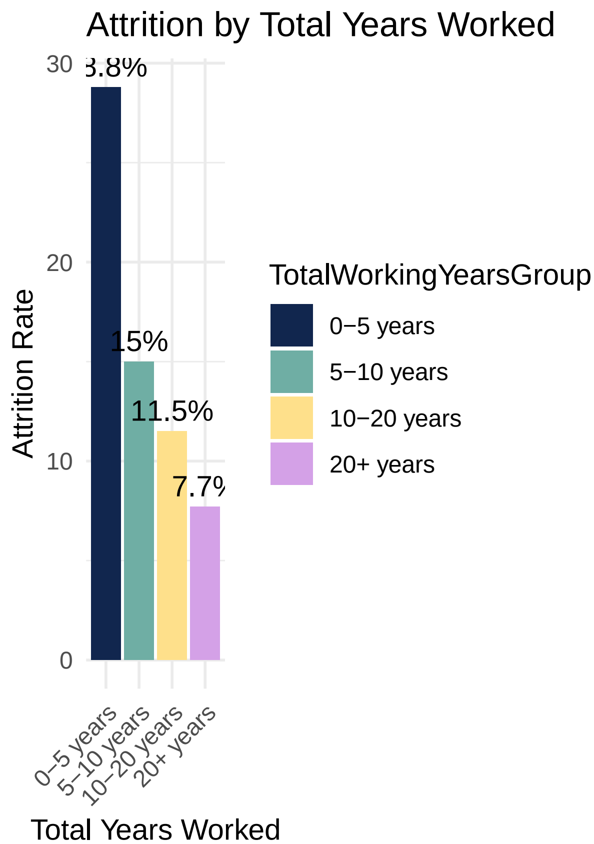
```
# Cut the TotalWorkingYears into groups
data$TotalWorkingYearsGroup <- cut(data$TotalWorkingYears,
                                   breaks = c(-Inf, 5, 10, 20, Inf),
                                   labels = c("0-5 years", "5-10 years", "10-20 years", "20+ years"))

data %>% count(TotalWorkingYearsGroup)
```

```
##   TotalWorkingYearsGroup   n
## 1              0-5 years 316
## 2             5-10 years 607
## 3            10-20 years 340
## 4              20+ years 207
```

```
# Visualization to show total employees by total working years group
workyears_count <- data %>% count(TotalWorkingYearsGroup)
```

```r
workyears_dist <- ggplot(data=workyears_count, aes(x="", y=n, fill=TotalWorkingYearsGroup)) +
  geom_bar(stat="identity") +
  coord_polar("y", start=0) +
  labs(title="Employees by Years Worked") +
  geom_text(aes(label=paste0(round((n/sum(n))*100, 1), "%"), x=1.6), position = position_stack(vjust=0.5
  scale_fill_manual(values=c('#E84040', '#E96060', '#E88181', '#E7A1A1')) +
  theme_minimal()

# Visualization to show Attrition Rate by TotalWorkingYears Groups
attrition_data <- data %>%
  filter(Attrition == "Yes")

workingyears_att_count <- attrition_data %>% count(TotalWorkingYearsGroup)

merged_data <- data.frame(
  TotalWorkingYearsGroup = workyears_count$TotalWorkingYearsGroup,
  TotalEmployees = workyears_count$n,
  Left = workingyears_att_count$n,
  Attrition_Rate = round((workingyears_att_count$n / workyears_count$n)*100, 1)
)

attrition_by_workyears <- ggplot(data=merged_data, aes(x=TotalWorkingYearsGroup, y=Attrition_Rate, fill=
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title="Attrition by Total Years Worked", x="Total Years Worked", y="Attrition Rate") +
  geom_text(aes(label=paste0(Attrition_Rate, "%"), vjust=-0.5, hjust=0.5)) +
  scale_fill_manual(values = c("#11264e","#6faea4","#FEE08B","#D4A1E7","#E7A1A1")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

grid.arrange(workyears_dist, attrition_by_workyears, ncol=2)
```



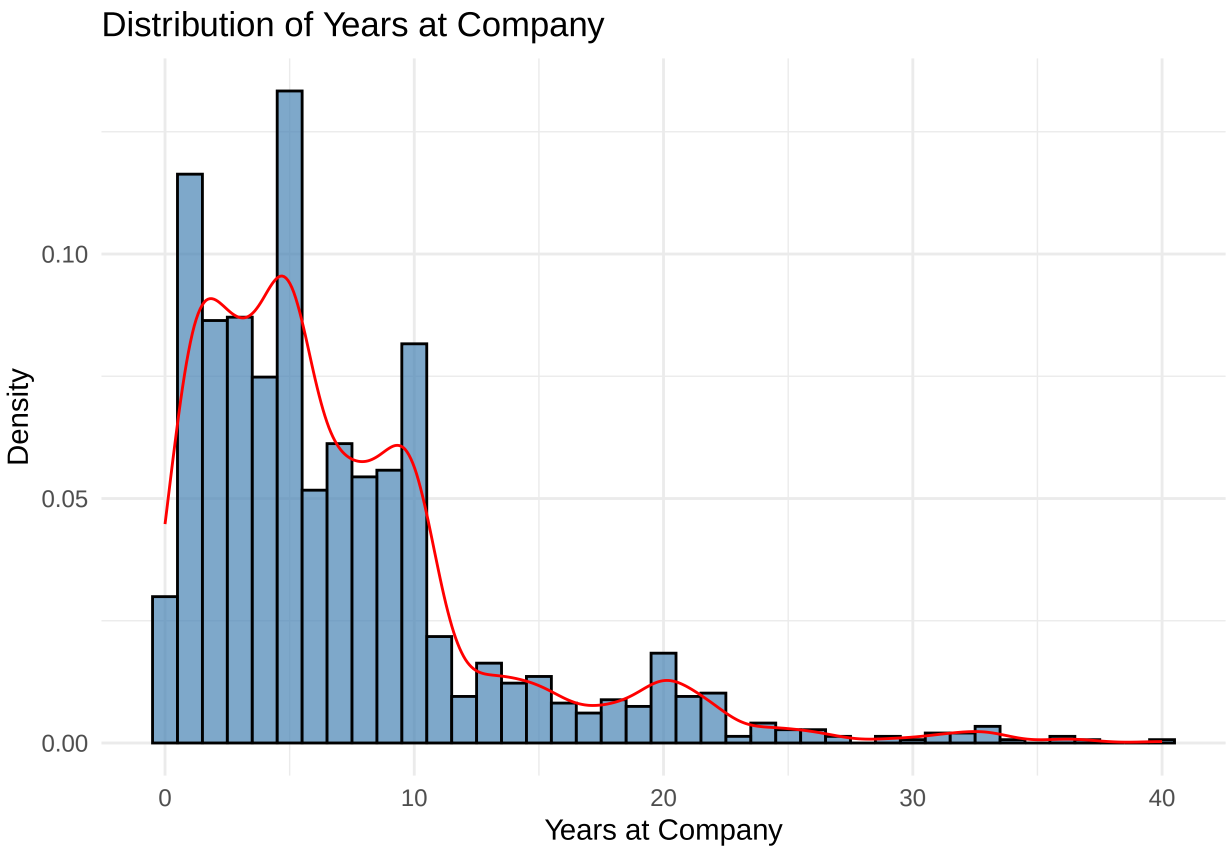


Inference

1. **Experience Distribution**:
   - The majority of employees have a total of **5-10 years of work experience**, but this group also exhibits a **very high attrition rate**.
2. **Attrition Rates by Experience Level**:
   - Employees with **less than 10 years of total work experience** tend to have a **high attrition rate**.

23. Analyzing Employee Attrition by Years at Company

```
# Visualization to show total employees by Years At Company
ggplot(data, aes(x = YearsAtCompany)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "steelblue", color = "black", alpha
  geom_density(aes(y = after_stat(density)), color = "red") +
  labs(title = "Distribution of Years at Company", x = "Years at Company", y = "Density") +
  theme_minimal()
```

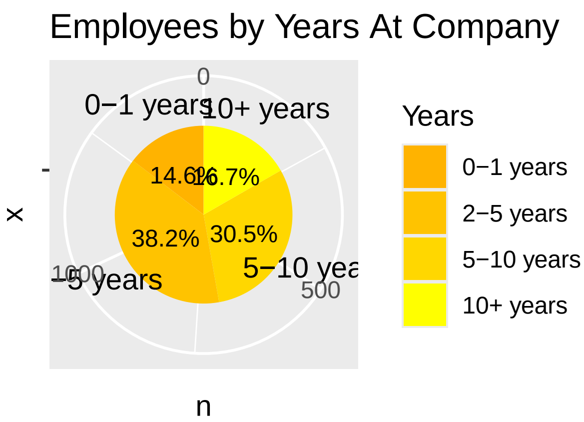


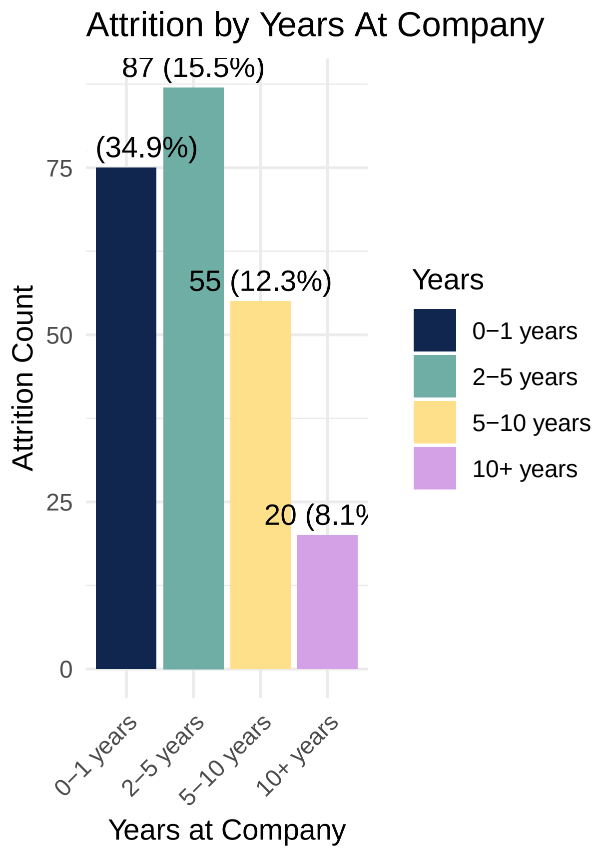

```r
data %>% count(YearsAtCompany)
```

```
##    YearsAtCompany   n
## 1               0  44
## 2               1 171
## 3               2 127
## 4               3 128
## 5               4 110
## 6               5 196
## 7               6  76
## 8               7  90
## 9               8  80
## 10              9  82
## 11             10 120
## 12             11  32
## 13             12  14
## 14             13  24
## 15             14  18
## 16             15  20
## 17             16  12
## 18             17   9
## 19             18  13
## 20             19  11
## 21             20  27
## 22             21  14
## 23             22  15
## 24             23   2
```

```
## 25              24   6
## 26              25   4
## 27              26   4
## 28              27   2
## 29              29   2
## 30              30   1
## 31              31   3
## 32              32   3
## 33              33   5
## 34              34   1
## 35              36   2
## 36              37   1
## 37              40   1
```

```r
# employee years at company by groups
data$YearsAtCompanyGroups <- cut(data$YearsAtCompany, breaks = c(-Inf, 1, 5, 10, Inf), labels = c('0-1 y


# data %>% count(YearsAtCompanyGroups)

years_at_company_count <- data %>% count(YearsAtCompanyGroups)

years_at_company_dist <- ggplot(data=years_at_company_count, aes(x="", y=n, fill=YearsAtCompanyGroups))
  geom_bar(stat="identity") +
  coord_polar(start=0, "y") +
  labs(title="Employees by Years At Company", fill="Years") +
  geom_text(aes(label=paste0(round((n/sum(n))*100, 1), "%")), position = position_stack(vjust = 0.5), si
  geom_text(aes(label=paste0(YearsAtCompanyGroups),x=1.8), position = position_stack(vjust = 0.5)) +
  scale_fill_manual(values=c('#FFB300', '#FFC300', '#FFD700', '#FFFF00'))

# Visualization to show attrition by YearsAtCompanyGroups
attrition_data <- data %>% filter(Attrition == "Yes")

years_at_company_att_count <- attrition_data %>% count(YearsAtCompanyGroups)

merged_data <- data.frame(
  YearsAtCompanyGroups = years_at_company_count$YearsAtCompanyGroups,
  TotalEmployees = years_at_company_count$n,
  Left = years_at_company_att_count$n,
  Attrition_Rate = round((years_at_company_att_count$n / years_at_company_count$n)*100, 1)
)

attrition_by_years_at_company <- ggplot(data=merged_data, aes(x=YearsAtCompanyGroups, y=Left, fill=Years
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Attrition by Years At Company", x="Years at Company", y="Attrition Count", fill="Years")
  geom_text(aes(label = paste0(Left, " (", Attrition_Rate, "%)"), vjust=-0.5, hjust=0.5)) +
  scale_fill_manual(values = c("#11264e","#6faea4","#FEE08B","#D4A1E7","#E7A1A1")) +
  theme(axis.text.x = element_text(angle=45, hjust=1))

grid.arrange(years_at_company_dist, attrition_by_years_at_company, ncol=2  )
```





Inference

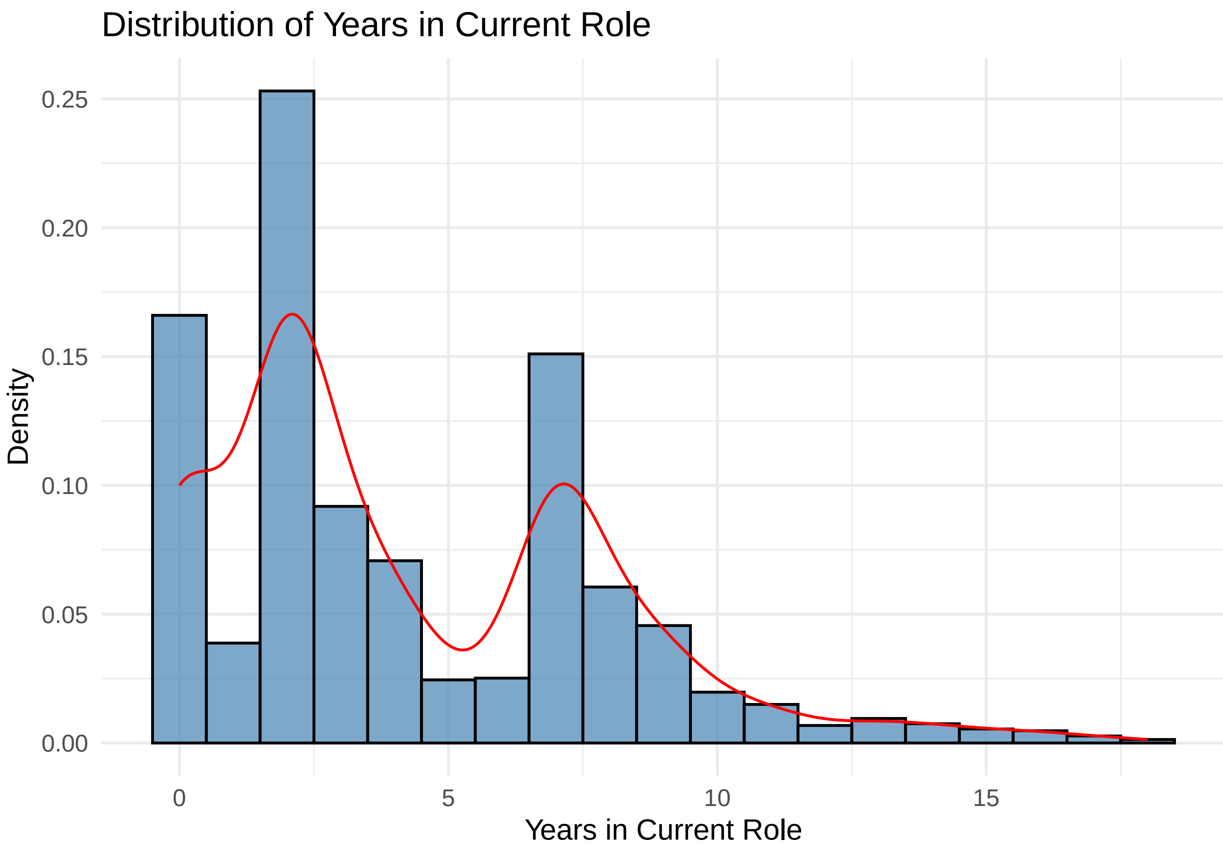
1. **Company Tenure Distribution**:
   - The majority of employees have worked at the company for **2 to 5 years** (**38.2%**) and **5 to 10 years** (**30.5**).

   - Only a small percentage of employees have worked for **less than a year** (**14.6%**) or **more than 10 years** (**16.7%**).
2. **Attrition Rates by Company Tenure**:
   - Employees with **2-5 years** of tenure exhibit a **very high attrition rate**.

   - Employees who have worked for **over 10 years** have a **very low attrition rate**.

24. Analyzing Employee Attrition by Years in Current Role

```
# Visualization to show Years in Current Role

ggplot(data, aes(x = YearsInCurrentRole)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "steelblue", color = "black", alpha
  geom_density(aes(y = after_stat(density)), color = "red") +
  labs(title = "Distribution of Years in Current Role", x = "Years in Current Role", y = "Density") +
  theme_minimal()
```

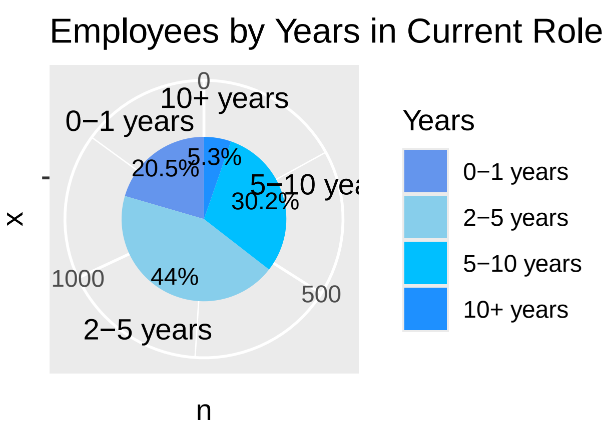


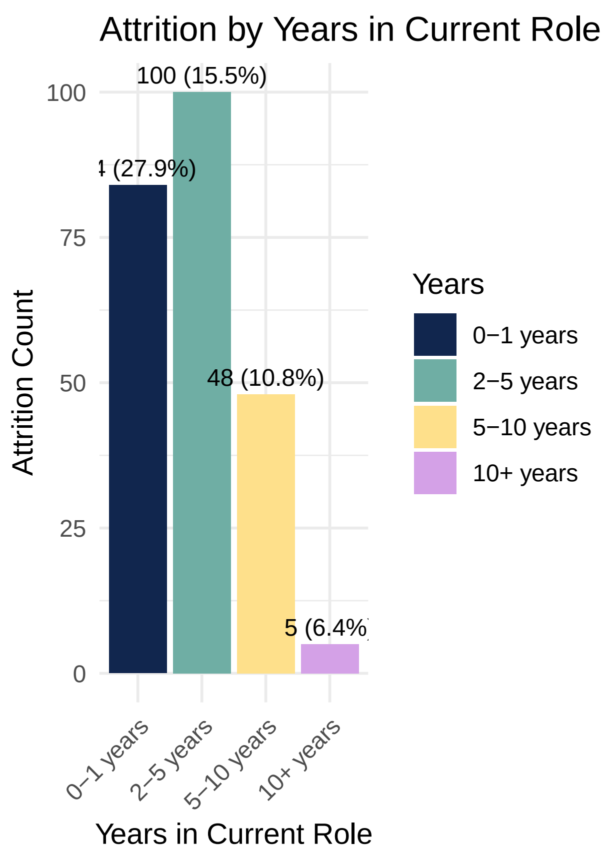

```r
# Creating bins
data$YearsInCurrentRoleGroups <- cut(data$YearsInCurrentRole, breaks = c(-Inf, 1, 5, 10, 29), labels=c(

data %>% count(YearsInCurrentRoleGroups)
```

```
##   YearsInCurrentRoleGroups   n
## 1                0-1 years 301
## 2                2-5 years 647
## 3               5-10 years 444
## 4                10+ years  78
```

```r
years_in_role_count <- data %>% count(YearsInCurrentRoleGroups)

years_in_role_dist <- ggplot(data=years_in_role_count, aes(x="", y=n, fill=YearsInCurrentRoleGroups)) +
  geom_bar(stat="identity") +
  coord_polar(start = 0, "y") +
  labs(title="Employees by Years in Current Role", fill="Years") +
  geom_text(aes(label=paste0(round((n/sum(n))*100, 1), "%"), x=1.25), position=position_stack(vjust=0.5
  scale_fill_manual(values = c('#6495ED', '#87CEEB', '#00BFFF', '#1E90FF')) +
  geom_text(aes(label=paste0(YearsInCurrentRoleGroups), x=1.9), position=position_stack(vjust=0.5))

# Visualization to show attrition rate by year in current role
attrition_data <- data %>% filter(Attrition == "Yes")

years_in_role_att_count <- attrition_data %>% count(YearsInCurrentRoleGroups)

merged_data = data.frame(
  YearsInCurrentRoleGroups = years_in_role_count$YearsInCurrentRoleGroups,
```

```r
  TotalEmployees = years_in_role_count$n,
  Left = years_in_role_att_count$n,
  Attrition_Rate = round((years_in_role_att_count$n / years_in_role_count$n)*100, 1)
)

attrition_by_years_in_role <- ggplot(data=merged_data, aes(x=YearsInCurrentRoleGroups, y=Left, fill=Yea
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(title = "Attrition by Years in Current Role", x="Years in Current Role", y="Attrition Count", fil
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)"), vjust=-0.5, hjust=0.5), size = 3) +
  scale_fill_manual(values=c("#11264e","#6faea4","#FEE08B","#D4A1E7","#E7A1A1")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

grid.arrange(years_in_role_dist, attrition_by_years_in_role, ncol=2)
```





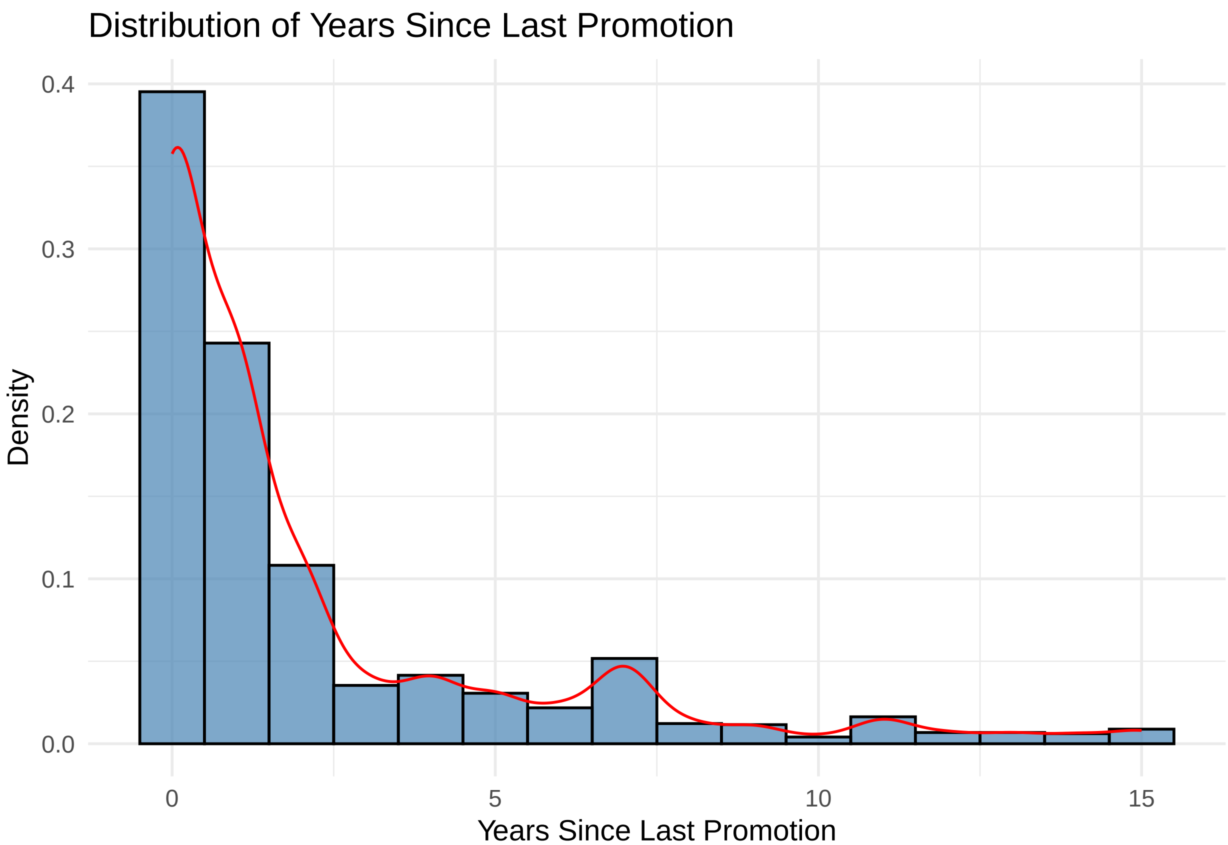
Inference

1. **Role Tenure Distribution**:
   - The majority of employees have worked in the same role for **2 to 5 years** and **5 to 10 years**.
   - Around **10%** of employees have worked in the same role for **less than 1 year**.
   - A small proportion of employees have worked in the same role for **more than 10 years**.
2. **Attrition Rates by Role Tenure**:
   - Employees who have worked in the same role for **0-1 years** exhibit a **very high attrition rate** of **27.9%**.

   - Employees with **2-5 years** of role tenure have an attrition rate of **15.5%**.

25. Analyzing Employee Attrition by Years Since Last Promotion.

```
# visualization to show employees by Years Since Last Promotion
ggplot(data, aes(x = YearsSinceLastPromotion)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "steelblue", color = "black", alpha
  geom_density(aes(y = after_stat(density)), color = "red") +
  labs(title = "Distribution of Years Since Last Promotion", x = "Years Since Last Promotion", y = "Dens
  theme_minimal()
```

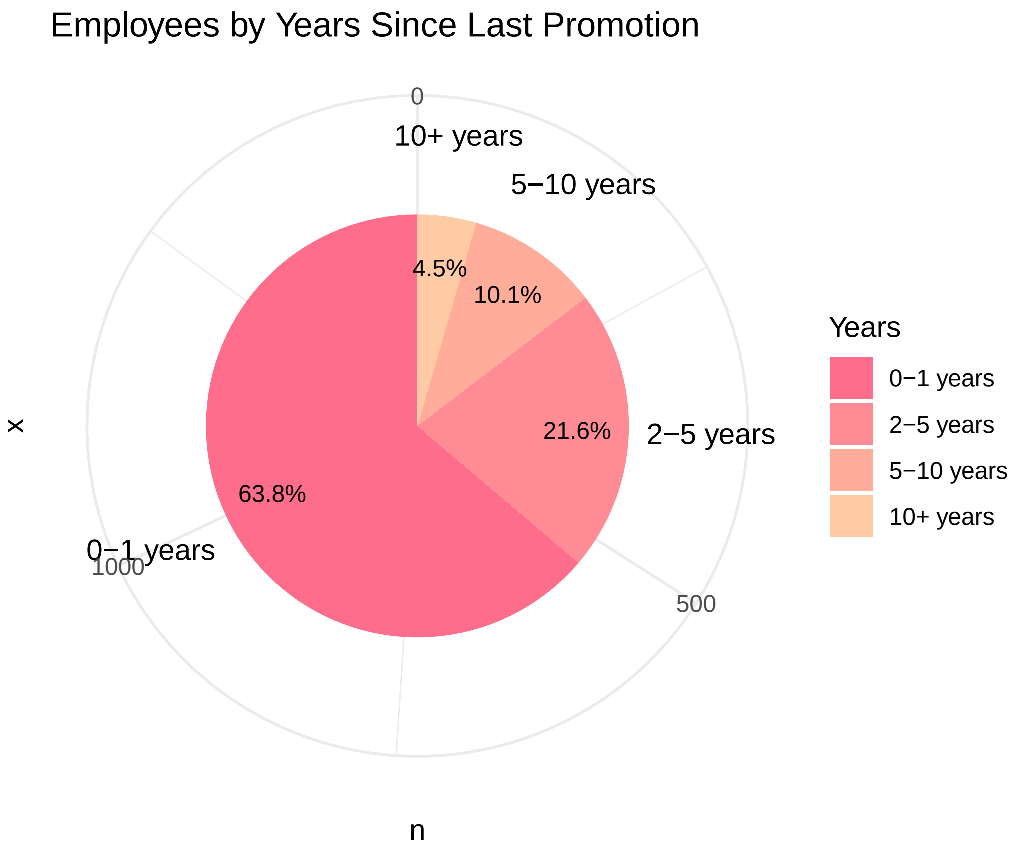



```
# creating bins
data$YearsSinceLastPromotionGroups <- cut(data$YearsSinceLastPromotion, breaks=c(-Inf, 1, 5, 10, Inf), l

data %>% count(YearsSinceLastPromotionGroups)
```

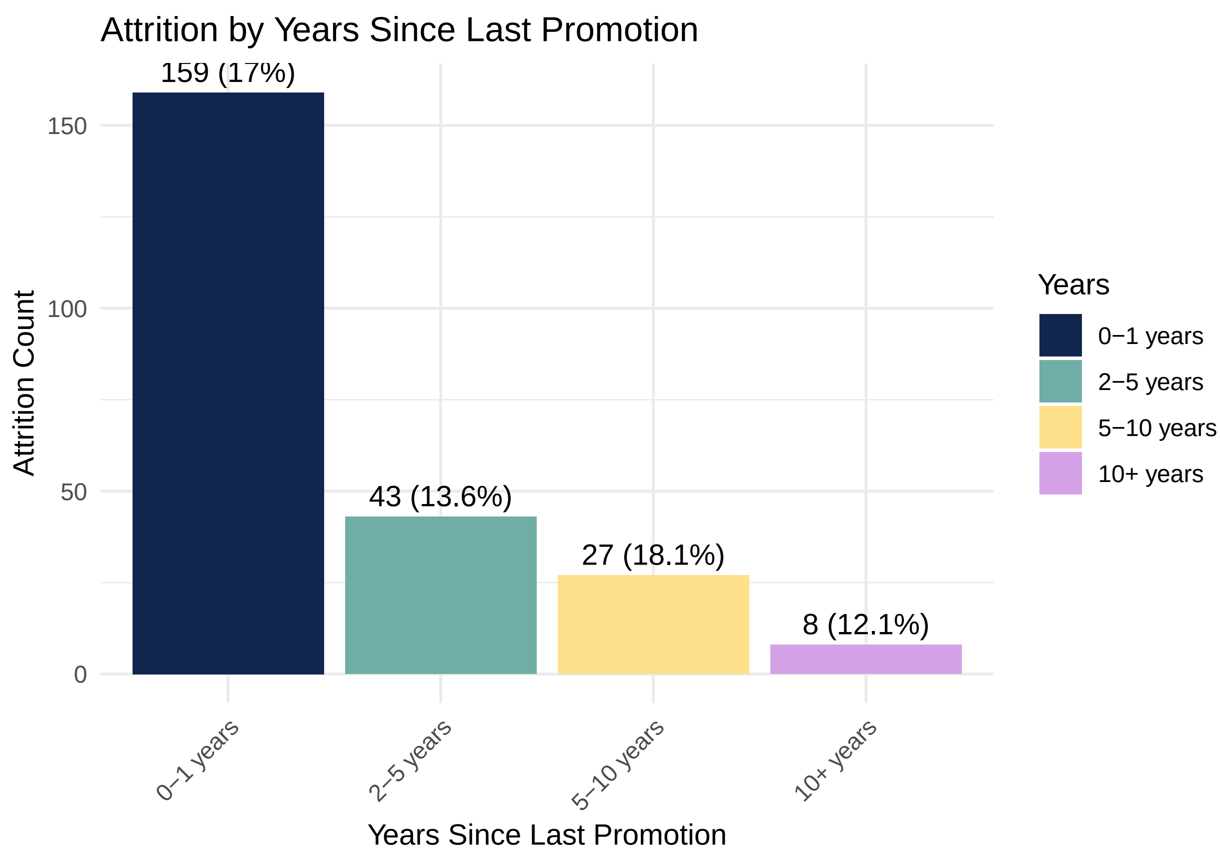
```
##   YearsSinceLastPromotionGroups   n
## 1                     0-1 years 938
## 2                     2-5 years 317
## 3                    5-10 years 149
## 4                      10+ years  66
```

```
# visualization to show employees by Years Since Last Promotion
years_since_promotion_count <- data %>% count(YearsSinceLastPromotionGroups)

years_since_promotion_dist <- ggplot(data=years_since_promotion_count, aes(x="", y=n, fill=YearsSinceLas
  geom_bar(stat="identity") +
  coord_polar(start=0, "y") +
  theme_minimal() +
  labs(title="Employees by Years Since Last Promotion", fill="Years") +
```

```
  geom_text(aes(label=paste0(round((n/sum(n))*100, 1), "%"), x=1.23), position = position_stack(vjust=0
  geom_text(aes(label=paste0(YearsSinceLastPromotionGroups), x=1.8), position = position_stack(vjust=0.5
  scale_fill_manual(values=c('#FF6D8C', '#FF8C94', '#FFAC9B', '#FFCBA4'))

grid.arrange(years_since_promotion_dist, ncol=1)
```




```
# Visualization to show Attrition by Years Since Last Promotion
attrition_data <- data %>% filter(Attrition == "Yes")

years_since_promotion_att_count <- attrition_data %>% count(YearsSinceLastPromotionGroups)

merged_data <- data.frame(
  YearsSinceLastPromotionGroups = years_since_promotion_count$YearsSinceLastPromotionGroups,
  TotalEmployees = years_since_promotion_count$n,
  Left = years_since_promotion_att_count$n,
  Attrition_Rate = round((years_since_promotion_att_count$n / years_since_promotion_count$n)*100, 1)
)

attrition_by_last_promotion <- ggplot(data=merged_data, aes(x=YearsSinceLastPromotionGroups, y=Left, fil
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title="Attrition by Years Since Last Promotion", x="Years Since Last Promotion", y="Attrition Co
  geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)"), vjust=-0.5, hjust=0.5)) +
  scale_fill_manual(values=c("#11264e","#6faea4","#FEE08B","#D4A1E7","#E7A1A1")) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

grid.arrange(attrition_by_last_promotion, ncol=1)
```



Inference

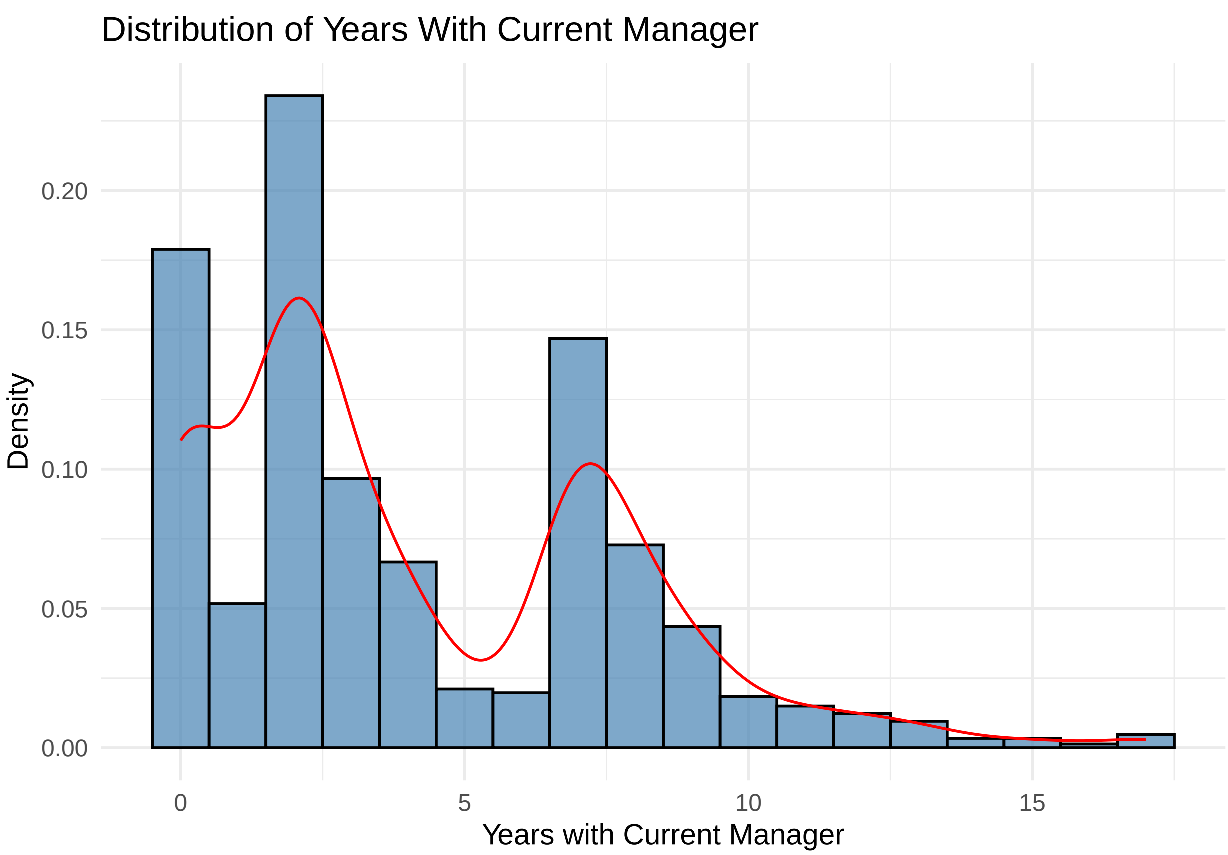
1. **Promotion Trends**:
   - Approximately **22%** of employees have not been promoted for the past **2-5 years**.

   - Around **5%** of employees have not received a promotion for the past **10 years**.
2. **Attrition Rates by Promotion Category**:
   - All employee categories exhibit attrition rates exceeding **10%**.

   - The **highest attrition rates** are observed among employees who have not been promoted for **5-10 years**.

##26. Analyzing employee Attrition by Years with Current Manager

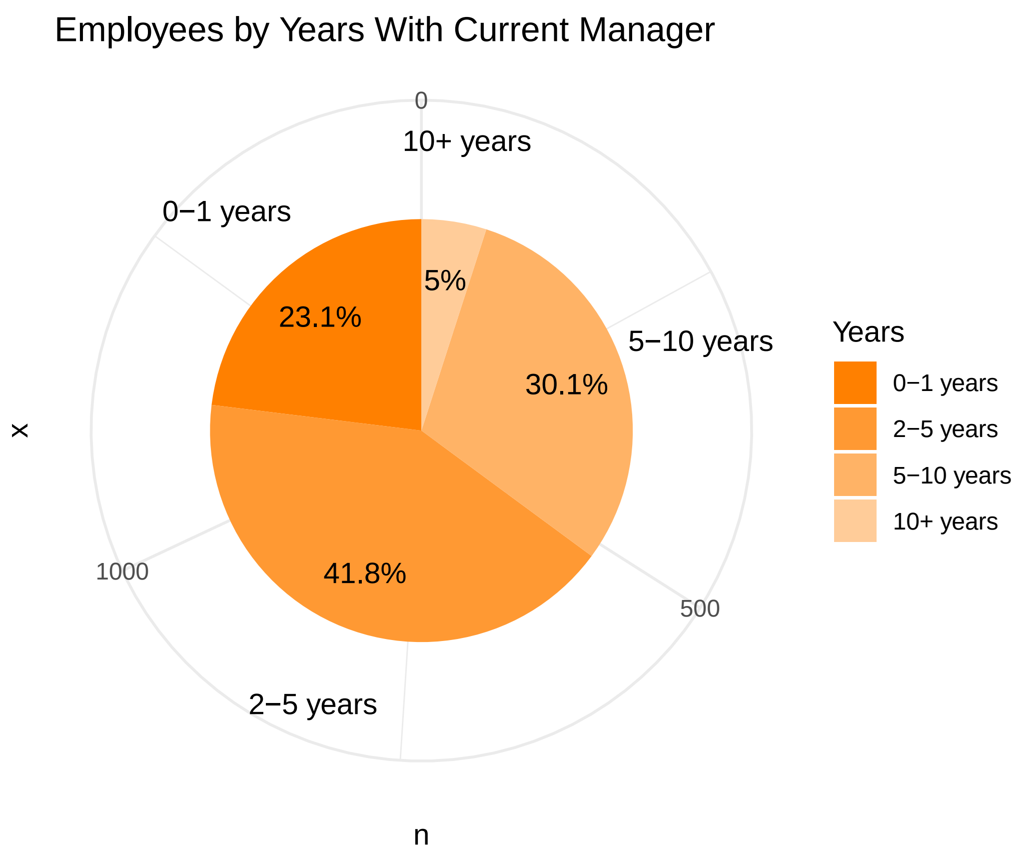
```
ggplot(data, aes(x = YearsWithCurrManager)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "steelblue", color = "black", alpha
  geom_density(aes(y = after_stat(density)), color = "red") +
  labs(title = "Distribution of Years With Current Manager", x = "Years with Current Manager", y = "Dens
  theme_minimal()
```



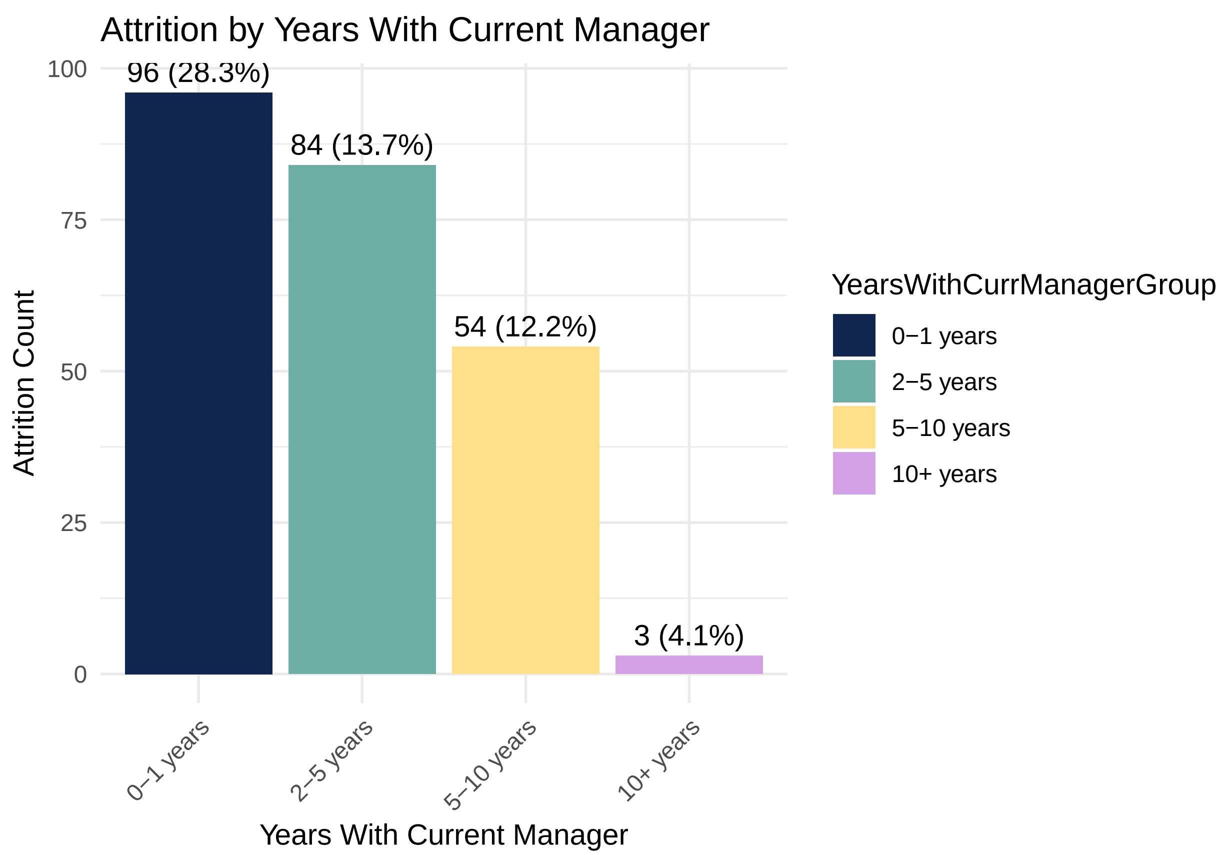

```
# Creating bins
data$YearsWithCurrManagerGroup <- cut(data$YearsWithCurrManager, breaks = c(-Inf, 1, 5, 10, Inf), label

data %>% count(YearsWithCurrManagerGroup)
```

```
##   YearsWithCurrManagerGroup   n
## 1                 0-1 years 339
## 2                 2-5 years 615
## 3                5-10 years 443
## 4                 10+ years  73
```

```
# Visualization to show employees by Years With Current Manager
years_current_manager_count <- data %>% count(YearsWithCurrManagerGroup)

years_current_manager_dist <- ggplot(data=years_current_manager_count, aes(x="", y=n, fill=YearsWithCur
  geom_bar(stat="identity") +
  coord_polar(start=0, "y") +
  theme_minimal() +
  labs(title="Employees by Years With Current Manager", fill="Years") +
  geom_text(aes(label=paste0(round((n/sum(n))*100, 1),"%"), x=1.2),
                                       position=position_stack(vjust=0.5)) +
  scale_fill_manual(values = c('#FF8000', '#FF9933', '#FFB366', '#FFCC99')) +
  geom_text(aes(label=paste0(YearsWithCurrManagerGroup), x=1.8),
                                     position=position_stack(vjust=0.5))

grid.arrange(years_current_manager_dist, ncol=)
```




```r
# Visualization to show Attrition Rate by YearsWithCurrentManager
attrition_data <- data %>% filter(Attrition == "Yes")

years_current_manager_att_count <- attrition_data %>% count(YearsWithCurrManagerGroup)

merged_data = data.frame(
  YearsWithCurrManagerGroup = years_current_manager_count$YearsWithCurrManagerGroup,
  TotalEmployees = years_current_manager_count$n,
  Left = years_current_manager_att_count$n,
  Attrition_Rate = round((years_current_manager_att_count$n / years_current_manager_count$n)*100, 1)
)
attrition_by_years_current_manager <- ggplot(data=merged_data, aes(x=YearsWithCurrManagerGroup, y=Left,
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Attrition by Years With Current Manager", x="Years With Current Manager", y="Attrition (
geom_text(aes(label=paste0(Left, " (", Attrition_Rate, "%)")), vjust=-0.5, hjust=0.5)) +
  scale_fill_manual(values=c("#11264e","#6faea4","#FEE08B","#D4A1E7","#E7A1A1")) +
  theme(axis.text.x = element_text(angle=45, hjust=1))

grid.arrange(attrition_by_years_current_manager, ncol=1)
```



Inference

1. **Years with Current Manager**:
    - Approximately **42%** of employees have worked with the same manager for **2-5 years**.

    - Around **31%** of employees have worked with the same manager for **5-10 years**.
2. **Attrition Rate by Tenure with Manager**:
    - Employees who have worked for **10+ years** with the same manager exhibit a **very low attrition rate**.

    - Other categories, especially those with shorter tenures, demonstrate a **higher attrition rate**.

Statistical Analysis

##1. CHi-Squared Test: Test to Analyze the Significance of Categorical Features on the Employee Attrition.

```
# Extracting the categorical columns in the dataset
copy_new_df <- new_df %>% select(-c(Attrition))

categorical_columns <- names(copy_new_df)[sapply(copy_new_df, is.factor) | sapply(copy_new_df, is.chara
```

```
categorical_columns
```

```
## [1] "BusinessTravel"          "Department"
## [3] "Education"               "EducationField"
## [5] "EnvironmentSatisfaction" "Gender"
## [7] "JobInvolvement"          "JobLevel"
## [9] "JobRole"                 "JobSatisfaction"
```

```
## [11] "MaritalStatus"            "OverTime"
## [13] "PerformanceRating"        "RelationshipSatisfaction"
## [15] "WorkLifeBalance"
```

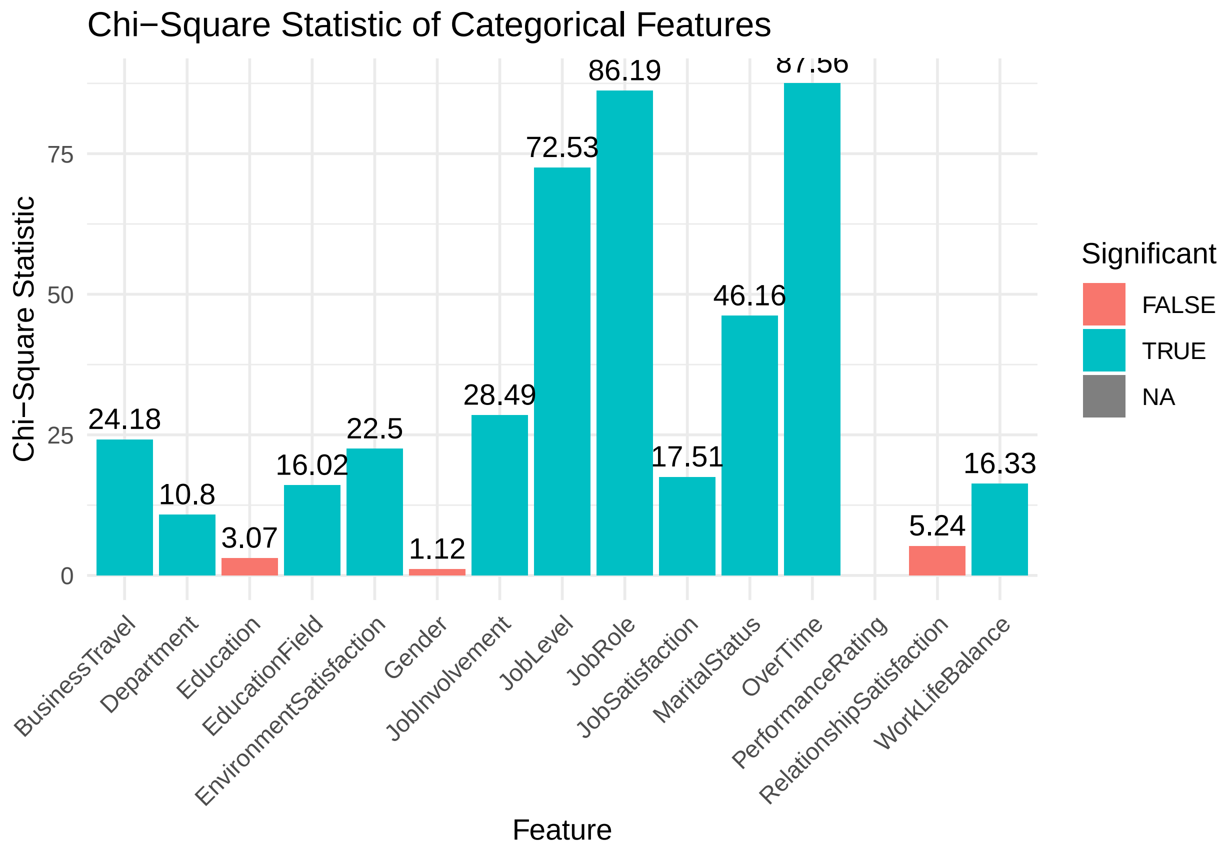
```r
chi2_statistic <- c()
p_values <- c()

for(col in categorical_columns){
  contingency_table <- table(new_df[[col]], new_df$Attrition)

  test <- chisq.test(contingency_table)

  chi2_statistic <- c(chi2_statistic, test$statistic)
  p_values <- c(p_values, test$p.value)
}
```

```
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect
```

Visualization of the Chi-Square Statistic Value of Each Categorical Column.

```r
results <- data.frame(
  Column = categorical_columns,
  Chi2 = chi2_statistic,
  P_Value = p_values
)

# Sort results by Chi-squared value
results <- results %>% arrange(desc(Chi2))

# Plot Chi-squared statistics
ggplot(results, aes(x = Column, y = Chi2, fill = P_Value < 0.05)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(Chi2, 2)), vjust = -0.5) +
  labs(title = "Chi-Square Statistic of Categorical Features", x = "Feature", y = "Chi-Square Statistic
  theme_minimal() +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_text()`).
```




```r
results[["Column"]]
```

```
##  [1] "OverTime"                 "JobRole"
##  [3] "JobLevel"                 "MaritalStatus"
##  [5] "JobInvolvement"           "BusinessTravel"
##  [7] "EnvironmentSatisfaction"  "JobSatisfaction"
##  [9] "WorkLifeBalance"          "EducationField"
## [11] "Department"               "RelationshipSatisfaction"
## [13] "Education"                "Gender"
## [15] "PerformanceRating"
```

Inference

Features Showing Statistically Significant Association with Employee Attrition

The following features demonstrate statistically significant associations with employee attrition:
1. **OverTime**
2. **JobRole**
3. **JobLevel**
4. **MaritalStatus**
5. **JobInvolvement**
6. **BusinessTravel**
7. **EnvironmentSatisfaction**
8. **JobSatisfaction**
9. **WorkLifeBalance**
10. **EducationField**
11. **Department**

Features Not Showing Statistically Significant Association with Employee Attrition

The following features do not exhibit statistically significant associations with employee attrition:
1. **RelationshipSatisfaction**
2. **Education**
3. **Gender**
4. **PerformanceRating**

##2. ANOVA Test: Test to Analyze the Significance of Numerical Features on the Employee Attrition.

```
all_columns <- colnames(copy_new_df)

numerical_columns <- setdiff(all_columns, categorical_columns)

numerical_columns
```

```
##  [1] "Age"                     "DailyRate"
##  [3] "DistanceFromHome"        "HourlyRate"
##  [5] "MonthlyIncome"           "MonthlyRate"
##  [7] "NumCompaniesWorked"      "PercentSalaryHike"
##  [9] "StockOptionLevel"        "TotalWorkingYears"
## [11] "TrainingTimesLastYear"   "YearsAtCompany"
## [13] "YearsInCurrentRole"      "YearsSinceLastPromotion"
## [15] "YearsWithCurrManager"
```

```
anova_results <- data.frame(
  Variable = character(),
  F_statistic = numeric(),
  P_value = numeric(),
  stringsAsFactors = FALSE
)

for (num_var in numerical_columns) {
  anova_model <- aov(new_df[[num_var]] ~ new_df$Attrition, data = new_df)
  anova_summary <- summary(anova_model)

  F_statistic <- anova_summary[[1]][["F value"]][1]
  P_value <- anova_summary[[1]][["Pr(>F)"]][1]

  anova_results <- rbind(anova_results, data.frame(
    Variable = num_var,
    F_statistic = F_statistic,
    P_value = P_value,
    stringsAsFactors = FALSE
  ))
}
```

Visualization of the Chi-Square Statistic Value of Each Numerical Column.

```
ggplot(data=anova_results, aes(x=Variable, y=F_statistic, fill=P_value<0.05)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=paste0(round(F_statistic, 2), "\n", "(p = ", round(P_value, 2), ")"), vjust=-0.5,
  theme_minimal() +
  labs(title="ANOVA Test F Scores Comparison", x="Feature", y="F Scores", fill="Significant") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```r
anova_results %>% arrange(desc(F_statistic))
```

```
##                   Variable F_statistic      P_value
## 1        TotalWorkingYears 44.25249144 4.061878e-11
## 2       YearsInCurrentRole 38.83830278 6.003186e-10
## 3            MonthlyIncome 38.48881898 7.147364e-10
## 4                      Age 38.17588679 8.356308e-10
## 5     YearsWithCurrManager 36.71231147 1.736987e-09
## 6         StockOptionLevel 28.14050091 1.301015e-07
## 7           YearsAtCompany 27.00162376 2.318872e-07
## 8         DistanceFromHome  8.96827659 2.793060e-03
## 9    TrainingTimesLastYear  5.21164607 2.257850e-02
## 10               DailyRate  4.72663984 2.985816e-02
## 11      NumCompaniesWorked  2.78228670 9.552526e-02
## 12 YearsSinceLastPromotion  1.60221841 2.057900e-01
## 13             MonthlyRate  0.33791646 5.611236e-01
## 14       PercentSalaryHike  0.26672817 6.056128e-01
## 15              HourlyRate  0.06879598 7.931348e-01
```

Inference

Features Showing Strong Association with Employee Attrition

The following features exhibit strong associations with employee attrition, as indicated by their high F-scores and very low p-values:

1. **Age**
2. **DailyRate**
3. **DistanceFromHome**

4. **MonthlyIncome**
5. **StockOptionLevel**
6. **TotalWorkYears**
7. **TrainingTimesLastYear**
8. **YearsAtCompany**
9. **YearsInCurrentRole**
10. **YearsWithCurrentManager**

Features with No Significant Relationship to Employee Attrition

The following features do not show a significant relationship with employee attrition, as evidenced by their moderate F-scores and extremely high p-values:
1. **HourlyRate** 2. **MonthRate** 3. **NumCompaniesWorked** 4. **PercentSalaryHike** 5. **YearsSinceLastPromotion**

CONCLUSION

1. **Key Findings**
   Several variables demonstrated strong relationships with employee attrition, highlighting their importance in predicting attrition risk.
2. **Key Numerical Variables**
   - **Demographic Factors**: Age.
   - **Compensation-Related Factors**: Monthly Income, Percent Salary Hike.
   - **Job Experience**: Total Working Years, Years at Company.
   - **Role-Specific Attributes**: Job Role, Years in Current Role.
3. **Key Categorical Variables**
   - **Job-Related Factors**: Department, Education Field, Job Role, Marital Status.
   - **Work-Related Factors**: Environment Satisfaction, Job Involvement, Job Satisfaction, OverTime, Work-Life Balance.

Limitations

- The analysis is confined to the available dataset and may not encompass all factors influencing employee attrition.
- There may be additional unmeasured variables that significantly contribute to attrition but are not included in this study.

RECOMMENDATIONS

Based on the findings, the following proposals aim to reduce attrition rates:

1. Age

- Establish strategies to address the unique requirements and career goals of employees across various age groups.

- Offer targeted growth opportunities, mentorship programs, and flexible work arrangements to support work-life balance.

2. **Income**

   - Regularly review and benchmark salary packages to remain competitive in the market.
   - Implement performance-based incentives and rewards to motivate and recognize employees' achievements.

3. **Job Experience**

   - Provide opportunities for career growth, skill development, and cross-functional training.
   - Develop clear career paths and conduct regular feedback sessions and performance reviews to promote employee growth and engagement.

4. **Specific Job-Related Factors**

   - Customize retention strategies for different roles and responsibilities.
   - Enhance job satisfaction by assigning challenging projects, fostering a positive work atmosphere, and ensuring recognition for contributions.

5. **Job-Related Issues**

   - Foster employee engagement and job satisfaction by cultivating a supportive work environment.
   - Provide growth opportunities, nurture a culture of continuous learning, and ensure fair and transparent promotion and career advancement processes.

6. **Work-Related Factors**

   - Focus on improving aspects such as environmental satisfaction, workplace involvement, job satisfaction, work-life balance, and effective overtime management.
   - Conduct regular employee surveys to understand concerns and feedback, and take proactive steps to address identified areas for improvement.

General Recommendations

- Develop a positive workplace culture that prioritizes employee well-being, work-life balance, and professional growth.
- Encourage open communication, seek input and suggestions, and regularly adapt retention strategies based on employee feedback and evolving needs.