DDSAnalytics

2023-12-05

#DDSAnalytics Talent Management Report: Employee Attrition Analysis Introduction: DDSAnalytics, a leading analytics firm serving Fortune 100 companies, is embarking on a data science initiative to enhance talent management. Talent management encompasses workforce planning, employee development, and reducing attrition. Predicting employee turnover is the first focus area identified by the executive leadership.

This report, prepared by our data science team, analyzes existing employee data (CaseStudy2-data.csv) to identify the top factors contributing to attrition.Our evidence-based findings aim to inform strategies for mitigating attrition risks and improving workforce stability.

```
# Clean the global environment & load libraries
rm(list = ls())

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

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

```
library(e1071)
library(ggplot2)
library(ROSE)
```

```
## Warning: package 'ROSE' was built under R version 4.3.2
```

```
## Loaded ROSE 0.0-4
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.2

## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

#Employee Attrition Analysis: Identifying Top Factors and Model Development: Develop a model with one variable. Find the accuracy, specificity, sensitivity, using KNN, Naïve Bayes or Linear Regression.

```r
#Data Reading and Initial Exploration
data <- read.csv("CaseStudy2-data.csv")
head(data)
```

```
##   ID Age Attrition    BusinessTravel DailyRate             Department
## 1  1  32        No     Travel_Rarely       117                  Sales
## 2  2  40        No     Travel_Rarely      1308 Research & Development
## 3  3  35        No Travel_Frequently       200 Research & Development
## 4  4  32        No     Travel_Rarely       801                  Sales
## 5  5  24        No Travel_Frequently       567 Research & Development
## 6  6  27        No Travel_Frequently       294 Research & Development
##   DistanceFromHome Education   EducationField EmployeeCount EmployeeNumber
## 1               13         4    Life Sciences             1            859
## 2               14         3          Medical             1           1128
## 3               18         2    Life Sciences             1           1412
## 4                1         4        Marketing             1           2016
## 5                2         1 Technical Degree             1           1646
## 6               10         2    Life Sciences             1            733
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       2   Male         73              3        2
## 2                       3   Male         44              2        5
## 3                       3   Male         60              3        3
## 4                       3 Female         48              3        3
## 5                       1 Female         32              3        1
## 6                       4   Male         32              3        3
##                  JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1        Sales Executive               4      Divorced          4403
## 2      Research Director               3        Single         19626
## 3 Manufacturing Director               4        Single          9362
## 4        Sales Executive               4       Married         10422
## 5     Research Scientist               4        Single          3760
```

```
## 6 Manufacturing Director                1      Divorced          8793
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1        9250                  2      Y       No                11
## 2       17544                  1      Y       No                14
## 3       19944                  2      Y       No                11
## 4       24032                  1      Y       No                19
## 5       17218                  1      Y      Yes                13
## 6        4809                  1      Y       No                21
##   PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## 1                 3                        3            80                1
## 2                 3                        1            80                0
## 3                 3                        3            80                0
## 4                 3                        3            80                2
## 5                 3                        3            80                0
## 6                 4                        3            80                2
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                 8                     3               2              5
## 2                21                     2               4             20
## 3                10                     2               3              2
## 4                14                     3               3             14
## 5                 6                     2               3              6
## 6                 9                     4               2              9
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  2                       0                    3
## 2                  7                       4                    9
## 3                  2                       2                    2
## 4                 10                       5                    7
## 5                  3                       1                    3
## 6                  7                       1                    7
```

```r
#View(data)
str(data)
```

```
## 'data.frame':    870 obs. of  36 variables:
##  $ ID                      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age                     : int  32 40 35 32 24 27 41 37 34 34 ...
##  $ Attrition               : chr  "No" "No" "No" "No" ...
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Rarely" "Travel_Frequently" "Travel_Rarely
##  $ DailyRate               : int  117 1308 200 801 567 294 1283 309 1333 653 ...
##  $ Department              : chr  "Sales" "Research & Development" "Research & Development" "Sales"
##  $ DistanceFromHome        : int  13 14 18 1 2 10 5 10 10 10 ...
##  $ Education               : int  4 3 2 4 1 2 5 4 4 4 ...
##  $ EducationField          : chr  "Life Sciences" "Medical" "Life Sciences" "Marketing" ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  859 1128 1412 2016 1646 733 1448 1105 1055 1597 ...
##  $ EnvironmentSatisfaction : int  2 3 3 3 1 4 2 4 3 4 ...
##  $ Gender                  : chr  "Male" "Male" "Male" "Female" ...
##  $ HourlyRate              : int  73 44 60 48 32 32 90 88 87 92 ...
##  $ JobInvolvement          : int  3 2 3 3 3 3 4 2 3 2 ...
##  $ JobLevel                : int  2 5 3 3 1 3 1 2 1 2 ...
##  $ JobRole                 : chr  "Sales Executive" "Research Director" "Manufacturing Director" "Sa
##  $ JobSatisfaction         : int  4 3 4 4 4 1 3 4 3 3 ...
##  $ MaritalStatus           : chr  "Divorced" "Single" "Single" "Married" ...
##  $ MonthlyIncome           : int  4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...
```

```
## $ MonthlyRate             : int  9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...
## $ NumCompaniesWorked      : int  2 1 2 1 1 1 2 2 1 1 ...
## $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
## $ OverTime                : chr  "No" "No" "No" "No" ...
## $ PercentSalaryHike       : int  11 14 11 19 13 21 12 14 19 14 ...
## $ PerformanceRating       : int  3 3 3 3 3 4 3 3 3 3 ...
## $ RelationshipSatisfaction: int  3 1 3 3 3 3 1 3 4 2 ...
## $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
## $ StockOptionLevel        : int  1 0 0 2 0 2 0 3 1 1 ...
## $ TotalWorkingYears       : int  8 21 10 14 6 9 7 8 1 8 ...
## $ TrainingTimesLastYear   : int  3 2 2 3 2 4 5 5 2 3 ...
## $ WorkLifeBalance         : int  2 4 3 3 3 2 2 3 3 2 ...
## $ YearsAtCompany          : int  5 20 2 14 6 9 4 1 1 8 ...
## $ YearsInCurrentRole      : int  2 7 2 10 3 7 2 0 1 2 ...
## $ YearsSinceLastPromotion : int  0 4 2 5 1 1 0 0 0 7 ...
## $ YearsWithCurrManager    : int  3 9 2 7 3 7 3 0 0 7 ...
```

```r
summary(data)
```

```
##        ID             Age         Attrition         BusinessTravel    
##  Min.   :  1.0   Min.   :18.00   Length:870         Length:870        
##  1st Qu.:218.2   1st Qu.:30.00   Class :character   Class :character  
##  Median :435.5   Median :35.00   Mode  :character   Mode  :character  
##  Mean   :435.5   Mean   :36.83                                        
##  3rd Qu.:652.8   3rd Qu.:43.00                                        
##  Max.   :870.0   Max.   :60.00                                        
##    DailyRate       Department        DistanceFromHome   Education    
##  Min.   : 103.0   Length:870         Min.   : 1.000   Min.   :1.000  
##  1st Qu.: 472.5   Class :character   1st Qu.: 2.000   1st Qu.:2.000  
##  Median : 817.5   Mode  :character   Median : 7.000   Median :3.000  
##  Mean   : 815.2                      Mean   : 9.339   Mean   :2.901  
##  3rd Qu.:1165.8                      3rd Qu.:14.000   3rd Qu.:4.000  
##  Max.   :1499.0                      Max.   :29.000   Max.   :5.000  
##  EducationField     EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Length:870         Min.   :1     Min.   :   1.0   Min.   :1.000          
##  Class :character   1st Qu.:1     1st Qu.: 477.2   1st Qu.:2.000          
##  Mode  :character   Median :1     Median :1039.0   Median :3.000          
##                     Mean   :1     Mean   :1029.8   Mean   :2.701          
##                     3rd Qu.:1     3rd Qu.:1561.5   3rd Qu.:4.000          
##                     Max.   :1     Max.   :2064.0   Max.   :4.000          
##     Gender            HourlyRate     JobInvolvement     JobLevel    
##  Length:870         Min.   : 30.00   Min.   :1.000   Min.   :1.000  
##  Class :character   1st Qu.: 48.00   1st Qu.:2.000   1st Qu.:1.000  
##  Mode  :character   Median : 66.00   Median :3.000   Median :2.000  
##                     Mean   : 65.61   Mean   :2.723   Mean   :2.039  
##                     3rd Qu.: 83.00   3rd Qu.:3.000   3rd Qu.:3.000  
##                     Max.   :100.00   Max.   :4.000   Max.   :5.000  
##    JobRole          JobSatisfaction MaritalStatus      MonthlyIncome  
##  Length:870         Min.   :1.000   Length:870         Min.   : 1081  
##  Class :character   1st Qu.:2.000   Class :character   1st Qu.: 2840  
##  Mode  :character   Median :3.000   Mode  :character   Median : 4946  
##                     Mean   :2.709                      Mean   : 6390  
##                     3rd Qu.:4.000                      3rd Qu.: 8182  
##                     Max.   :4.000                      Max.   :19999  
```

```
##   MonthlyRate    NumCompaniesWorked    Over18             OverTime
##  Min.   : 2094   Min.   :0.000      Length:870         Length:870
##  1st Qu.: 8092   1st Qu.:1.000      Class :character   Class :character
##  Median :14074   Median :2.000      Mode  :character   Mode  :character
##  Mean   :14326   Mean   :2.728
##  3rd Qu.:20456   3rd Qu.:4.000
##  Max.   :26997   Max.   :9.000
##  PercentSalaryHike PerformanceRating RelationshipSatisfaction StandardHours
##  Min.   :11.0      Min.   :3.000     Min.   :1.000            Min.   :80
##  1st Qu.:12.0      1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80
##  Median :14.0      Median :3.000     Median :3.000            Median :80
##  Mean   :15.2      Mean   :3.152     Mean   :2.707            Mean   :80
##  3rd Qu.:18.0      3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80
##  Max.   :25.0      Max.   :4.000     Max.   :4.000            Max.   :80
##  StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
##  Min.   :0.0000   Min.   : 0.00     Min.   :0.000         Min.   :1.000
##  1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000
##  Median :1.0000   Median :10.00     Median :3.000         Median :3.000
##  Mean   :0.7839   Mean   :11.05     Mean   :2.832         Mean   :2.782
##  3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000
##  Max.   :3.0000   Max.   :40.00     Max.   :6.000         Max.   :4.000
##  YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion
##  Min.   : 0.000   Min.   : 0.000     Min.   : 0.000
##  1st Qu.: 3.000   1st Qu.: 2.000     1st Qu.: 0.000
##  Median : 5.000   Median : 3.000     Median : 1.000
##  Mean   : 6.962   Mean   : 4.205     Mean   : 2.169
##  3rd Qu.:10.000   3rd Qu.: 7.000     3rd Qu.: 3.000
##  Max.   :40.000   Max.   :18.000     Max.   :15.000
##  YearsWithCurrManager
##  Min.   : 0.00
##  1st Qu.: 2.00
##  Median : 3.00
##  Mean   : 4.14
##  3rd Qu.: 7.00
##  Max.   :17.00
```

```r
sapply(data, class)
```

```
##                       ID                      Age                Attrition
##                "integer"                "integer"              "character"
##           BusinessTravel                DailyRate               Department
##              "character"                "integer"              "character"
##         DistanceFromHome                Education           EducationField
##                "integer"                "integer"              "character"
##            EmployeeCount           EmployeeNumber  EnvironmentSatisfaction
##                "integer"                "integer"                "integer"
##                   Gender               HourlyRate           JobInvolvement
##              "character"                "integer"                "integer"
##                 JobLevel                  JobRole          JobSatisfaction
##                "integer"              "character"                "integer"
##            MaritalStatus            MonthlyIncome              MonthlyRate
##              "character"                "integer"                "integer"
##       NumCompaniesWorked                   Over18                 OverTime
##                "integer"              "character"              "character"
```

```
##       PercentSalaryHike        PerformanceRating RelationshipSatisfaction
##               "integer"                "integer"                "integer"
##           StandardHours         StockOptionLevel        TotalWorkingYears
##               "integer"                "integer"                "integer"
##    TrainingTimesLastYear          WorkLifeBalance           YearsAtCompany
##               "integer"                "integer"                "integer"
##      YearsInCurrentRole  YearsSinceLastPromotion     YearsWithCurrManager
##               "integer"                "integer"                "integer"
```

```r
colSums(is.na(data))
```

```
##                      ID                      Age                Attrition
##                       0                        0                        0
##          BusinessTravel                DailyRate               Department
##                       0                        0                        0
##        DistanceFromHome                Education           EducationField
##                       0                        0                        0
##           EmployeeCount           EmployeeNumber  EnvironmentSatisfaction
##                       0                        0                        0
##                  Gender               HourlyRate           JobInvolvement
##                       0                        0                        0
##                JobLevel                  JobRole          JobSatisfaction
##                       0                        0                        0
##           MaritalStatus            MonthlyIncome              MonthlyRate
##                       0                        0                        0
##      NumCompaniesWorked                   Over18                 OverTime
##                       0                        0                        0
##       PercentSalaryHike        PerformanceRating RelationshipSatisfaction
##                       0                        0                        0
##           StandardHours         StockOptionLevel        TotalWorkingYears
##                       0                        0                        0
##    TrainingTimesLastYear          WorkLifeBalance           YearsAtCompany
##                       0                        0                        0
##      YearsInCurrentRole  YearsSinceLastPromotion     YearsWithCurrManager
##                       0                        0                        0
```

```r
# Data Preprocessing
# Converting Attrition to a factor
data$Attrition <- factor(data$Attrition, levels = c("No", "Yes"))
str(data$Attrition)
```

```
##  Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Identify continuous and categorical variables
continuous_vars <- c("Age", "DailyRate", "DistanceFromHome", "Education", "HourlyRate", "MonthlyIncome"

# Convert categorical variables to factors
categorical_vars <- c("BusinessTravel", "Department", "EducationField", "Gender", "JobInvolvement", "Jo

data[categorical_vars] <- lapply(data[categorical_vars], factor)

str(data[categorical_vars])
```

```
## 'data.frame':	870 obs. of  12 variables:
##  $ BusinessTravel         : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 2 2 3 3
##  $ Department             : Factor w/ 3 levels "Human Resources",..: 3 2 2 3 2 2 2 3 3 2 ...
##  $ EducationField         : Factor w/ 6 levels "Human Resources",..: 2 4 2 3 6 2 4 2 2 6 ...
##  $ Gender                 : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 2 2 1 1 2 ...
##  $ JobInvolvement         : Factor w/ 4 levels "1","2","3","4": 3 2 3 3 3 3 4 2 3 2 ...
##  $ JobLevel               : Factor w/ 5 levels "1","2","3","4",..: 2 5 3 3 1 3 1 2 1 2 ...
##  $ JobRole                : Factor w/ 9 levels "Healthcare Representative",..: 8 6 5 8 7 5 7 8 9 1 .
##  $ JobSatisfaction        : Factor w/ 4 levels "1","2","3","4": 4 3 4 4 4 1 3 4 3 3 ...
##  $ MaritalStatus          : Factor w/ 3 levels "Divorced","Married",..: 1 3 3 2 3 1 2 1 2 2 ...
##  $ OverTime               : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 2 1 ...
##  $ WorkLifeBalance        : Factor w/ 4 levels "1","2","3","4": 2 4 3 3 3 2 2 3 3 2 ...
##  $ YearsSinceLastPromotion: Factor w/ 16 levels "0","1","2","3",..: 1 5 3 6 2 2 1 1 1 8 ...
```

```r
# Final structure check
str(data)
```

```
## 'data.frame':	870 obs. of  36 variables:
##  $ ID                      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age                     : int  32 40 35 32 24 27 41 37 34 34 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 2 2 3 3
##  $ DailyRate               : int  117 1308 200 801 567 294 1283 309 1333 653 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 3 2 2 2 3 3 2 ...
##  $ DistanceFromHome        : int  13 14 18 1 2 10 5 10 10 10 ...
##  $ Education               : int  4 3 2 4 1 2 5 4 4 4 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 4 2 3 6 2 4 2 2 6 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  859 1128 1412 2016 1646 733 1448 1105 1055 1597 ...
##  $ EnvironmentSatisfaction : int  2 3 3 3 1 4 2 4 3 4 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 2 2 1 1 2 ...
##  $ HourlyRate              : int  73 44 60 48 32 32 90 88 87 92 ...
##  $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 2 3 3 3 3 4 2 3 2 ...
##  $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 5 3 3 1 3 1 2 1 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 6 5 8 7 5 7 8 9 1
##  $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 3 4 4 4 1 3 4 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 1 3 3 2 3 1 2 1 2 2 ...
##  $ MonthlyIncome           : int  4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...
##  $ MonthlyRate             : int  9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...
##  $ NumCompaniesWorked      : int  2 1 2 1 1 1 2 2 1 1 ...
##  $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 2 1 ...
##  $ PercentSalaryHike       : int  11 14 11 19 13 21 12 14 19 14 ...
##  $ PerformanceRating       : int  3 3 3 3 3 4 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  3 1 3 3 3 3 1 3 4 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  1 0 0 2 0 2 0 3 1 1 ...
##  $ TotalWorkingYears       : int  8 21 10 14 6 9 7 8 1 8 ...
##  $ TrainingTimesLastYear   : int  3 2 2 3 2 4 5 5 2 3 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 2 4 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  5 20 2 14 6 9 4 1 1 8 ...
##  $ YearsInCurrentRole      : int  2 7 2 10 3 7 2 0 1 2 ...
##  $ YearsSinceLastPromotion : Factor w/ 16 levels "0","1","2","3",..: 1 5 3 6 2 2 1 1 1 8 ...
##  $ YearsWithCurrManager    : int  3 9 2 7 3 7 3 0 0 7 ...
```

```r
summary(data)
```

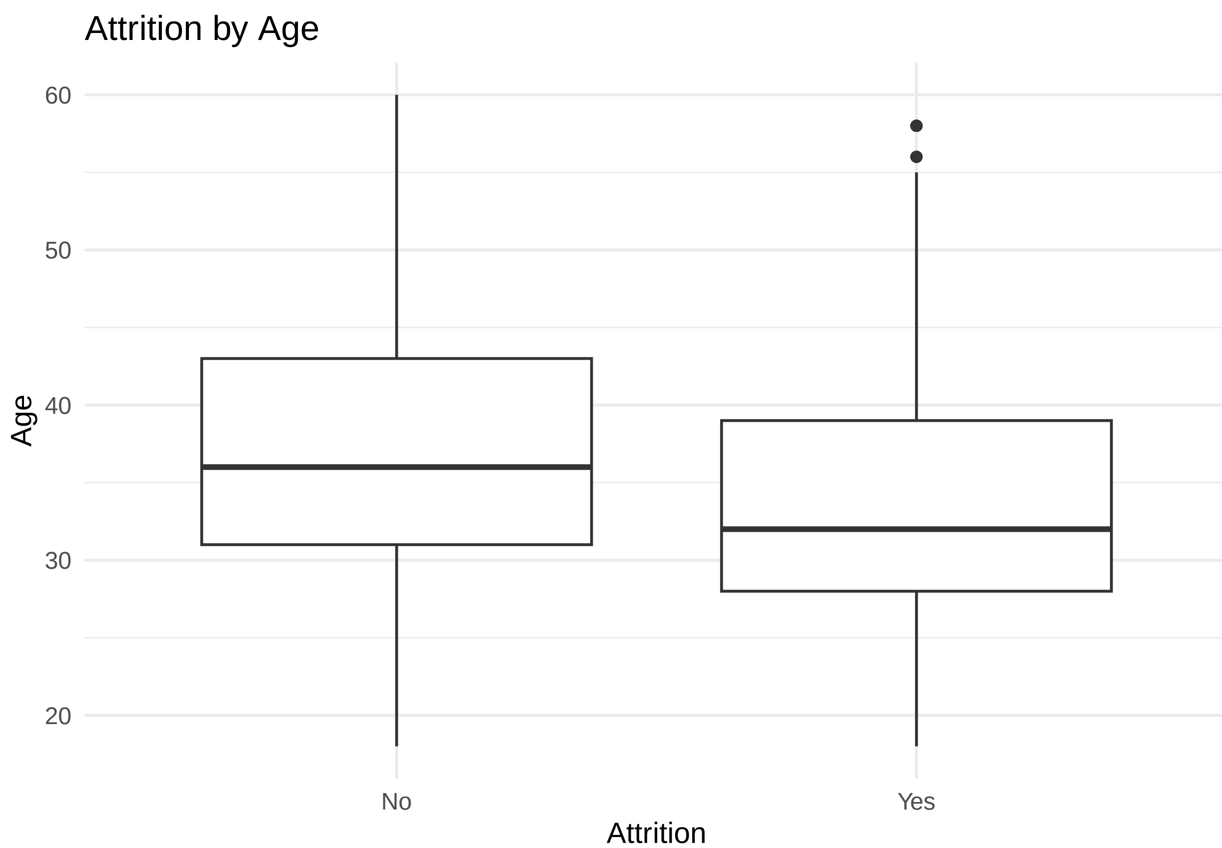
```
##        ID             Age        Attrition           BusinessTravel
##  Min.   :  1.0   Min.   :18.00   No :730   Non-Travel       : 94
##  1st Qu.:218.2   1st Qu.:30.00   Yes:140   Travel_Frequently:158
##  Median :435.5   Median :35.00             Travel_Rarely    :618
##  Mean   :435.5   Mean   :36.83
##  3rd Qu.:652.8   3rd Qu.:43.00
##  Max.   :870.0   Max.   :60.00
##
##    DailyRate                       Department  DistanceFromHome   Education
##  Min.   : 103.0   Human Resources       : 35   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 472.5   Research & Development:562   1st Qu.: 2.000   1st Qu.:2.000
##  Median : 817.5   Sales                 :273   Median : 7.000   Median :3.000
##  Mean   : 815.2                                Mean   : 9.339   Mean   :2.901
##  3rd Qu.:1165.8                                3rd Qu.:14.000   3rd Qu.:4.000
##  Max.   :1499.0                                Max.   :29.000   Max.   :5.000
##
##           EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfaction
##  Human Resources : 15    Min.   :1     Min.   :   1.0   Min.   :1.000
##  Life Sciences   :358    1st Qu.:1     1st Qu.: 477.2   1st Qu.:2.000
##  Marketing       :100    Median :1     Median :1039.0   Median :3.000
##  Medical         :270    Mean   :1     Mean   :1029.8   Mean   :2.701
##  Other           : 52    3rd Qu.:1     3rd Qu.:1561.5   3rd Qu.:4.000
##  Technical Degree: 75    Max.   :1     Max.   :2064.0   Max.   :4.000
##
##     Gender      HourlyRate     JobInvolvement JobLevel
##  Female:354   Min.   : 30.00   1: 47          1:329
##  Male  :516   1st Qu.: 48.00   2:228          2:312
##               Median : 66.00   3:514          3:132
##               Mean   : 65.61   4: 81          4: 60
##               3rd Qu.: 83.00                  5: 37
##               Max.   :100.00
##
##                       JobRole    JobSatisfaction  MaritalStatus MonthlyIncome
##  Sales Executive          :200   1:179           Divorced:191   Min.   : 1081
##  Research Scientist       :172   2:166           Married :410   1st Qu.: 2840
##  Laboratory Technician    :153   3:254           Single  :269   Median : 4946
##  Manufacturing Director   : 87   4:271                          Mean   : 6390
##  Healthcare Representative: 76                                  3rd Qu.: 8182
##  Sales Representative     : 53                                  Max.   :19999
##  (Other)                  :129
##   MonthlyRate    NumCompaniesWorked    Over18           OverTime
##  Min.   : 2094   Min.   :0.000      Length:870         No :618
##  1st Qu.: 8092   1st Qu.:1.000      Class :character   Yes:252
##  Median :14074   Median :2.000      Mode  :character
##  Mean   :14326   Mean   :2.728
##  3rd Qu.:20456   3rd Qu.:4.000
##  Max.   :26997   Max.   :9.000
##
##  PercentSalaryHike PerformanceRating RelationshipSatisfaction StandardHours
##  Min.   :11.0      Min.   :3.000     Min.   :1.000            Min.   :80
##  1st Qu.:12.0      1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80
```

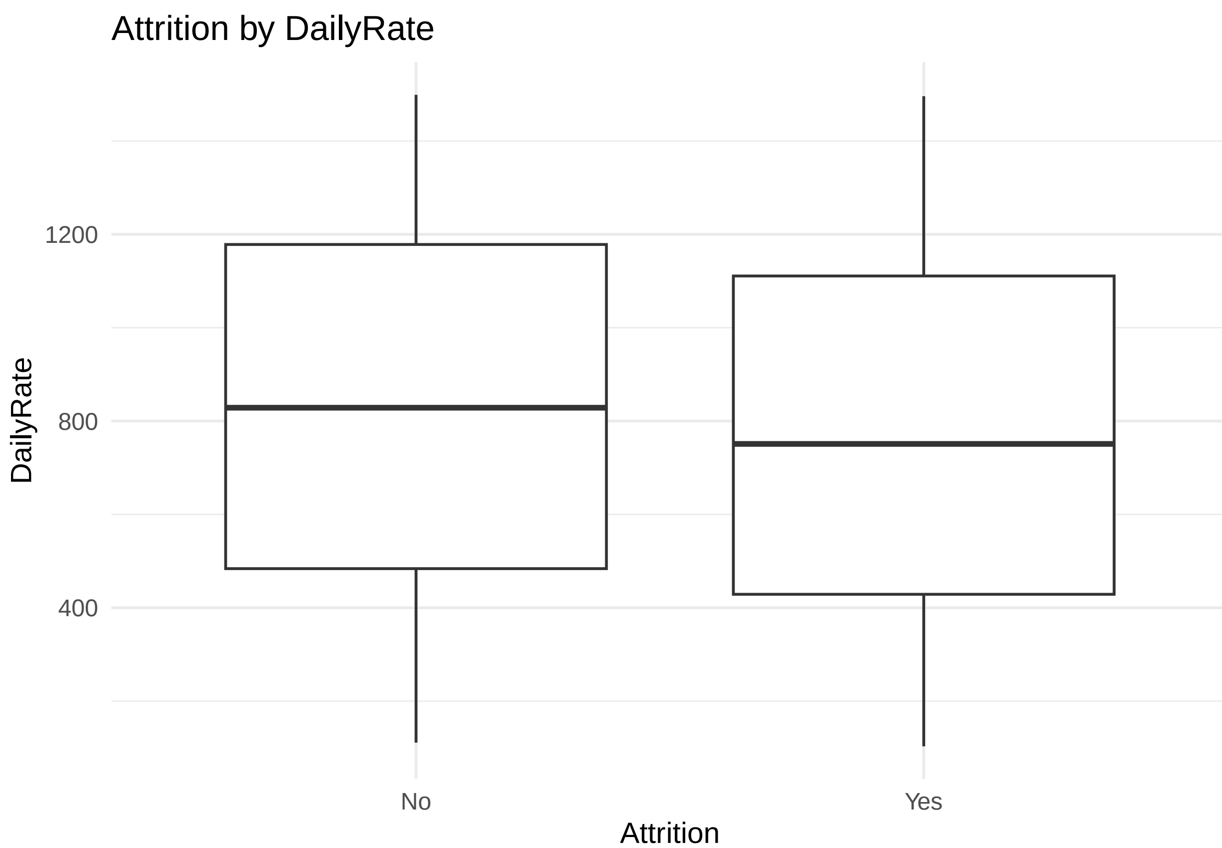
```
##  Median :14.0      Median :3.000     Median :3.000            Median :80
##  Mean   :15.2      Mean   :3.152     Mean   :2.707            Mean   :80
##  3rd Qu.:18.0      3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80
##  Max.   :25.0      Max.   :4.000     Max.   :4.000            Max.   :80
##
##  StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
##  Min.   :0.0000   Min.   : 0.00     Min.   :0.000         1: 48
##  1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000         2:192
##  Median :1.0000   Median :10.00     Median :3.000         3:532
##  Mean   :0.7839   Mean   :11.05     Mean   :2.832         4: 98
##  3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000
##  Max.   :3.0000   Max.   :40.00     Max.   :6.000
##
##  YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion
##  Min.   : 0.000   Min.   : 0.000     0      :342
##  1st Qu.: 3.000   1st Qu.: 2.000     1      :214
##  Median : 5.000   Median : 3.000     2      : 94
##  Mean   : 6.962   Mean   : 4.205     7      : 41
##  3rd Qu.:10.000   3rd Qu.: 7.000     3      : 32
##  Max.   :40.000   Max.   :18.000     4      : 32
##                                      (Other):115
##  YearsWithCurrManager
##  Min.   : 0.00
##  1st Qu.: 2.00
##  Median : 3.00
##  Mean   : 4.14
##  3rd Qu.: 7.00
##  Max.   :17.00
##
```

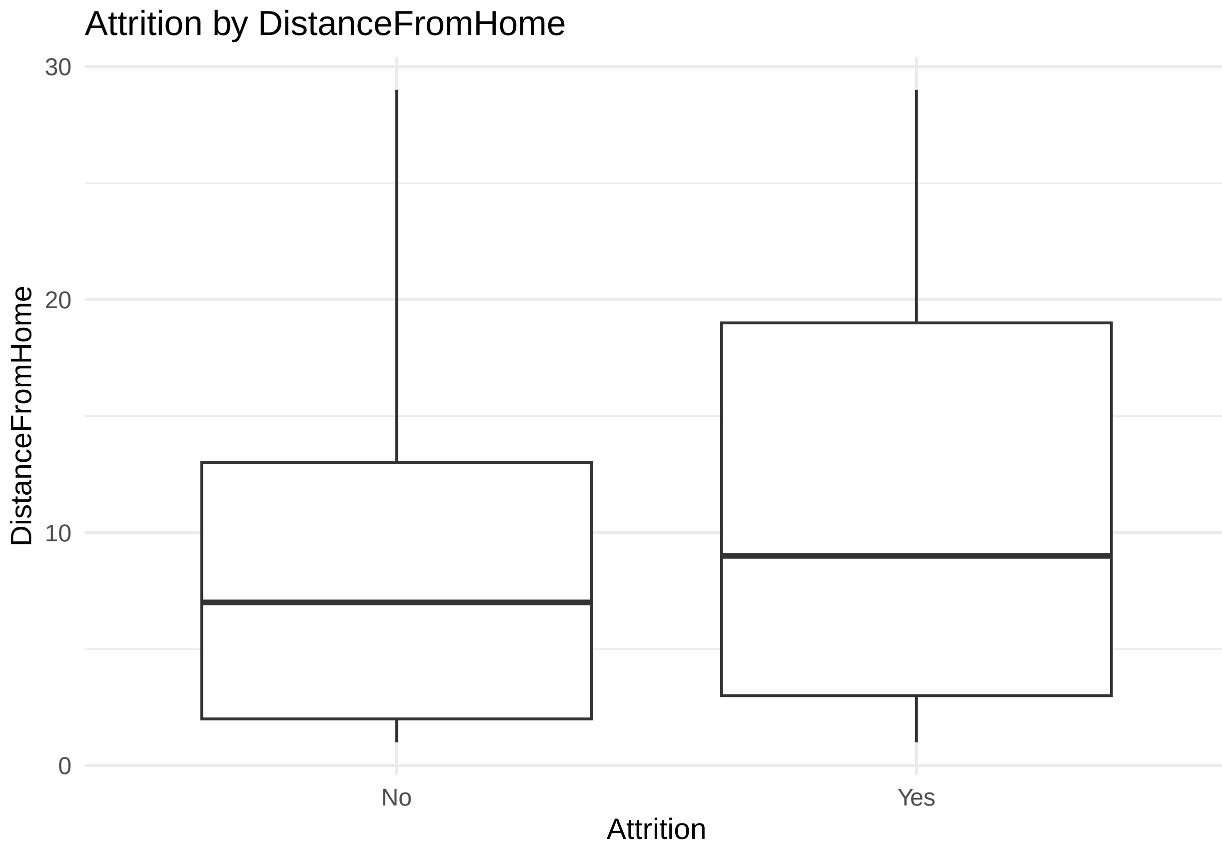
#DATA VIZ

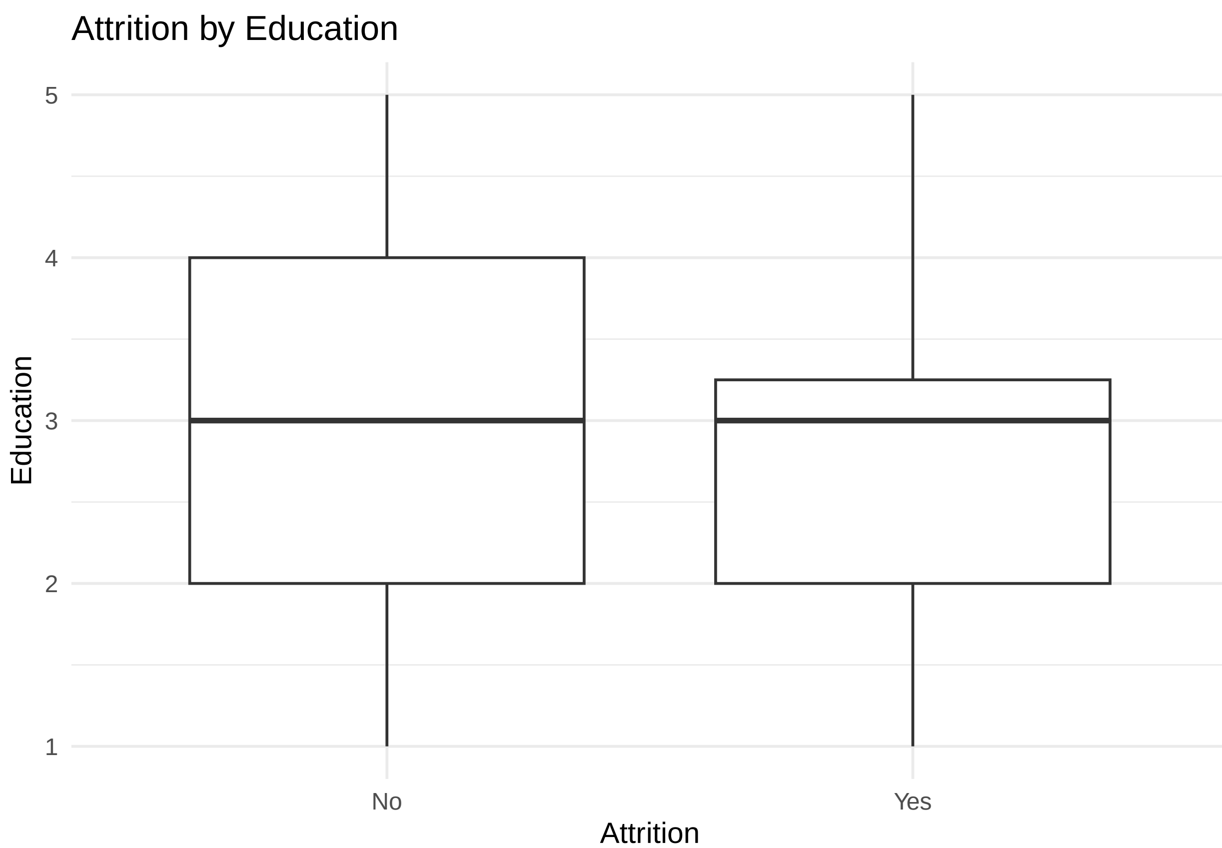
```r
# Create box plots for continuous variables
for (var in continuous_vars) {
  p <- ggplot(data, aes_string(x = "Attrition", y = var)) +
    geom_boxplot() +
    labs(title = paste("Attrition by", var), y = var, x = "Attrition") +
    theme_minimal()
  print(p)
}
```

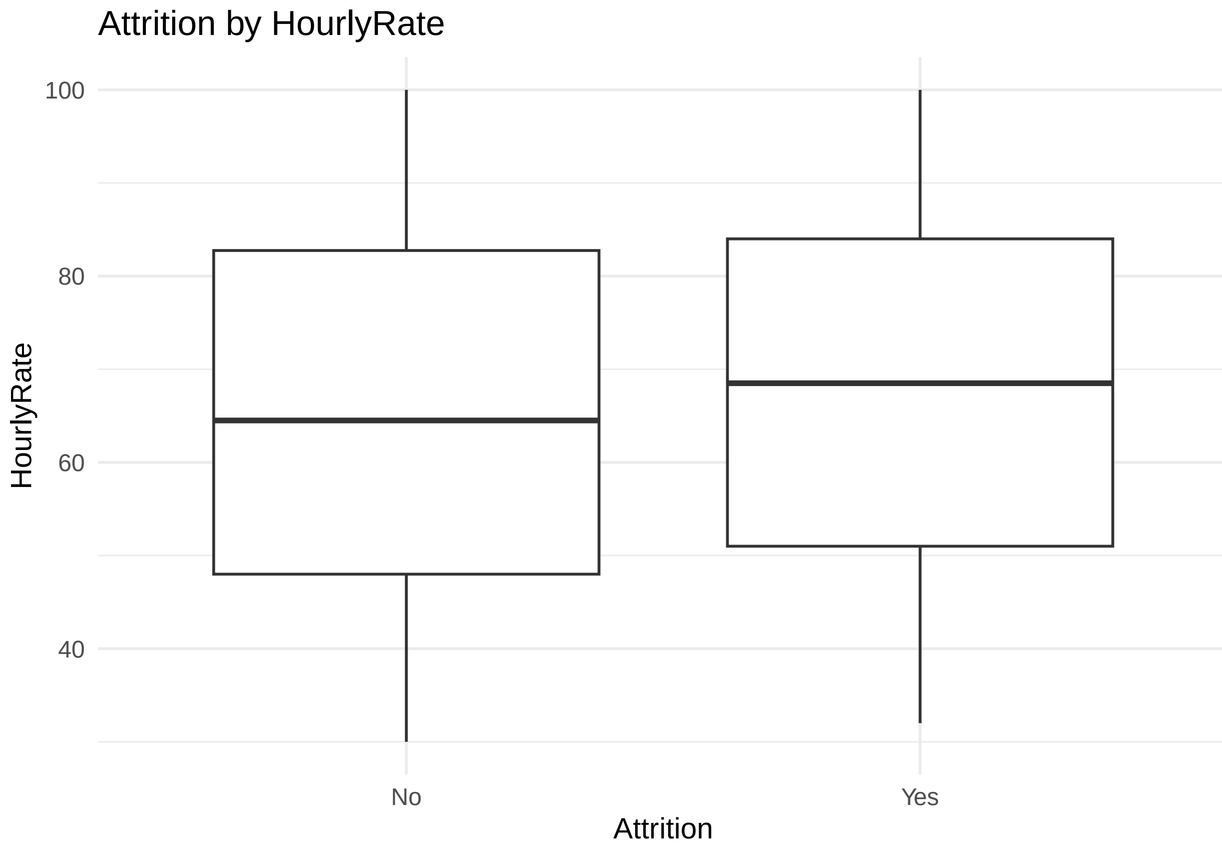
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

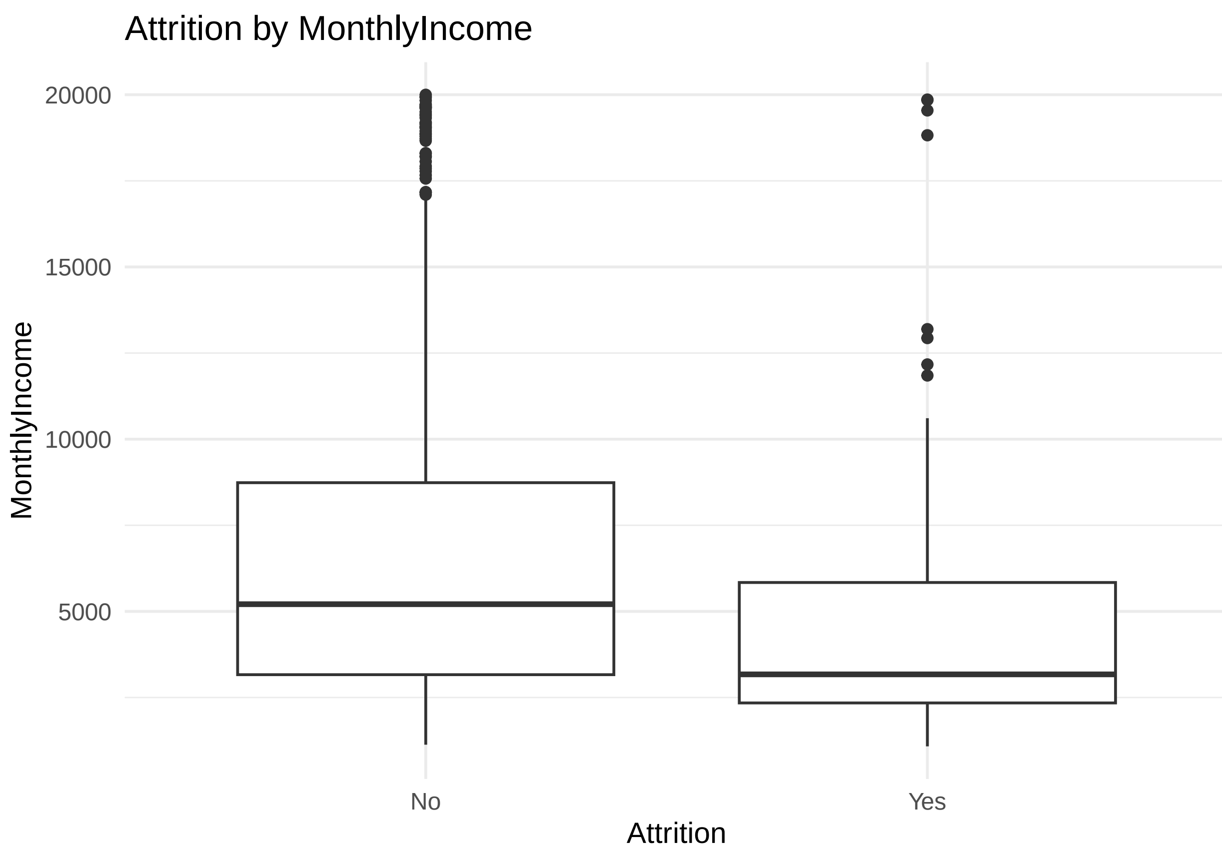

Attrition by Age
60
50
40
30
20
Age
No
Yes
Attrition

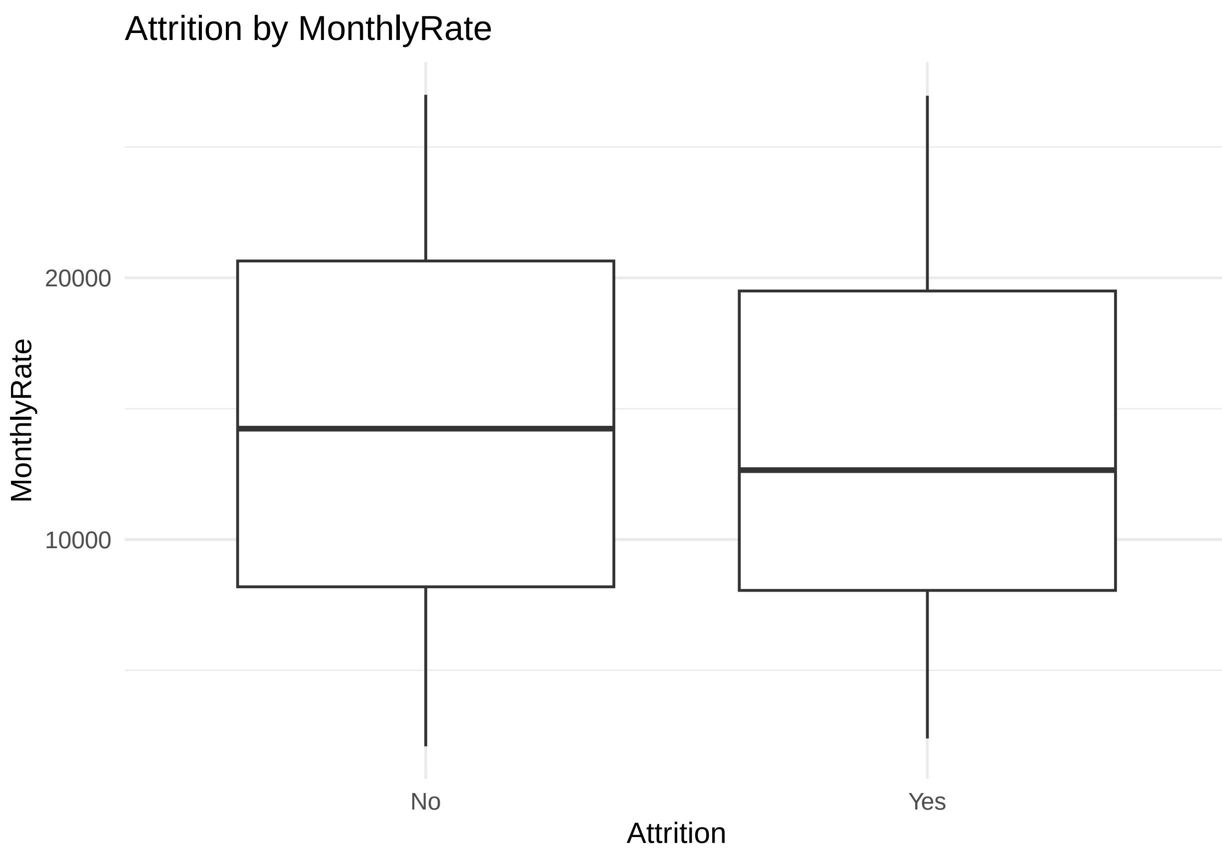

Attrition by DailyRate
DailyRate
1200
800
400
No
Yes
Attrition

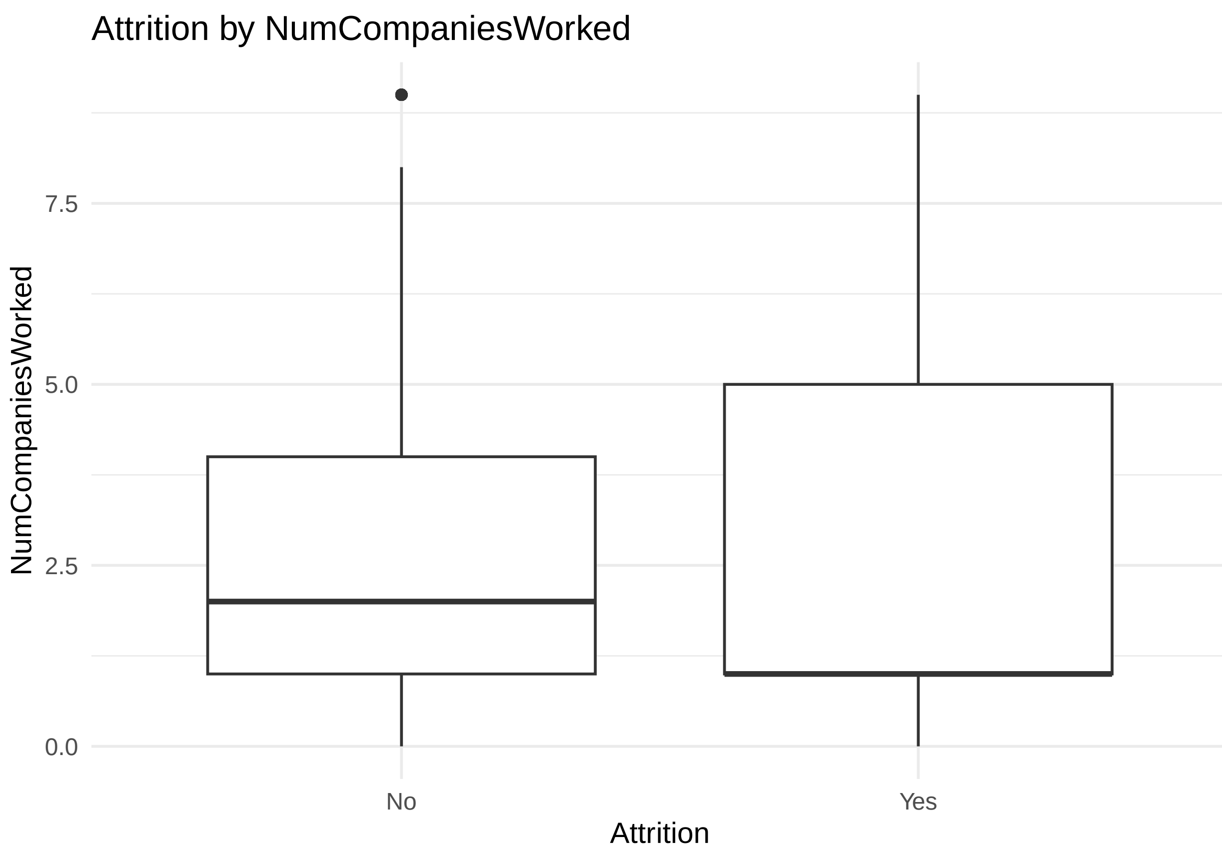

Attrition by DistanceFromHome
30
20
10
0
DistanceFromHome
No
Yes
Attrition


Attrition by Education
Education
5
4
3
2
1
No
Yes
Attrition

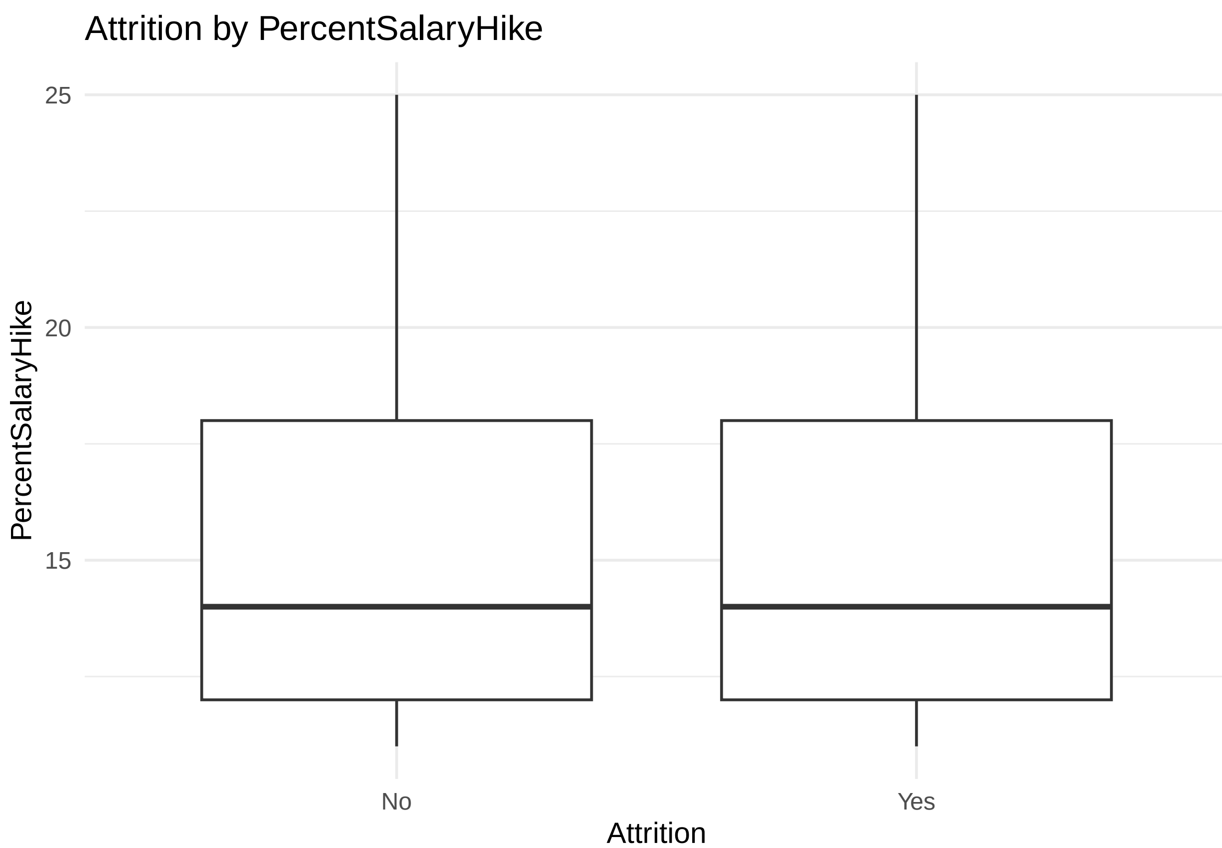

Attrition by HourlyRate
100
80
60
40
HourlyRate
No
Yes
Attrition

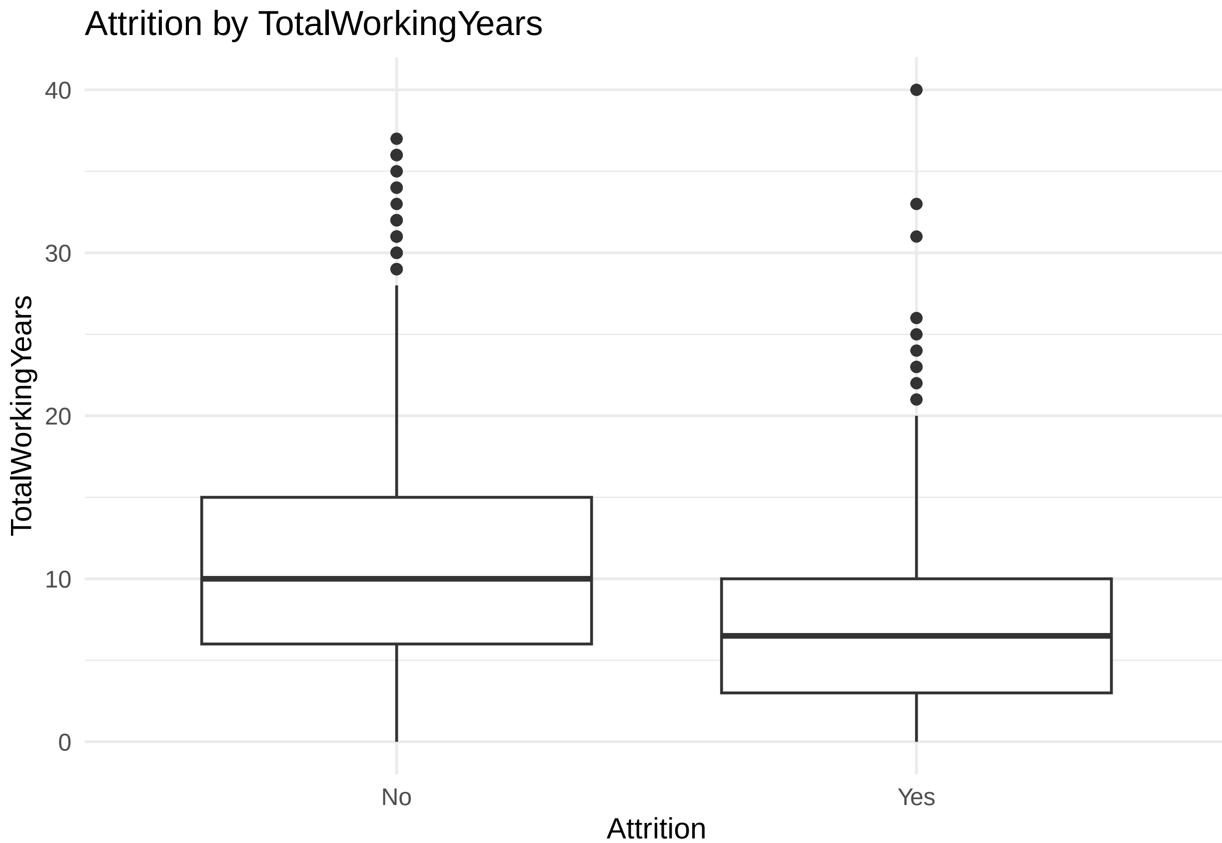

Attrition by MonthlyIncome
20000
15000
10000
5000
MonthlyIncome
No
Yes
Attrition

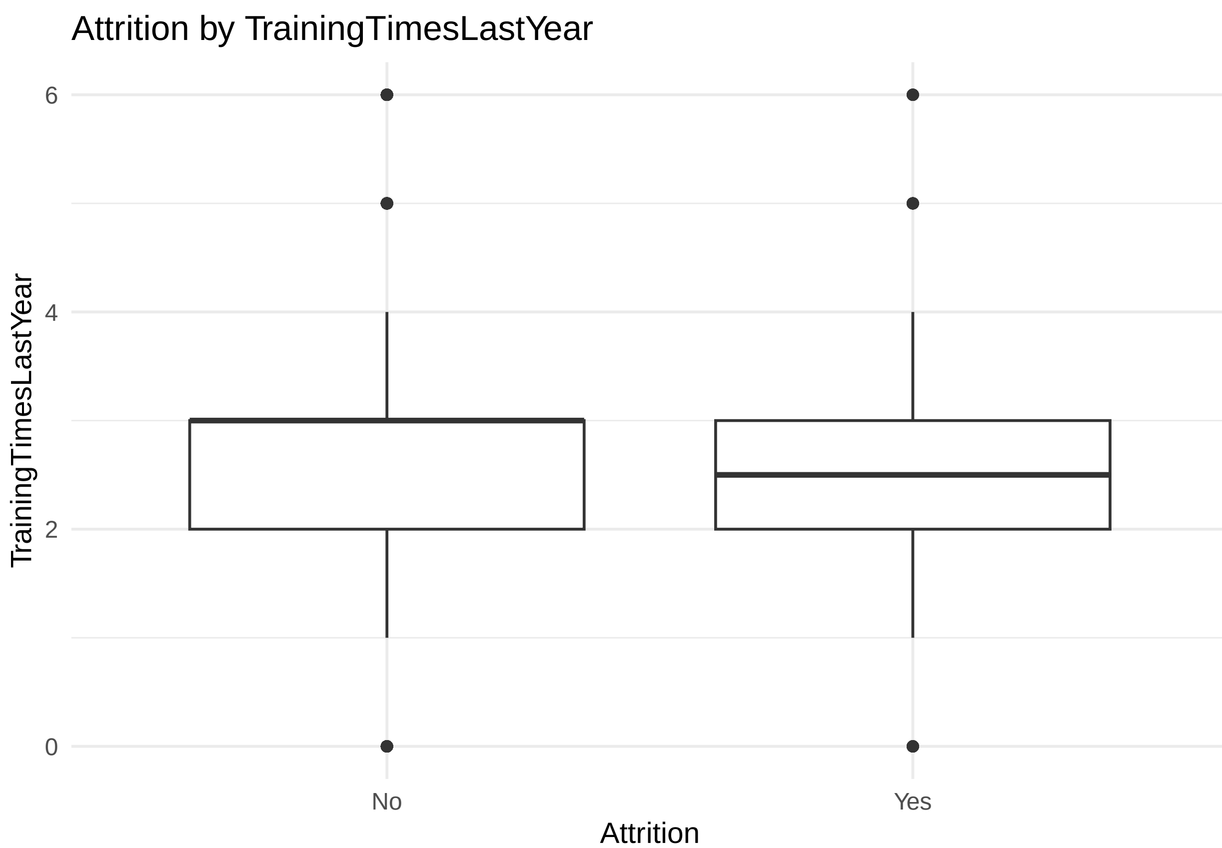

Attrition by MonthlyRate
20000
10000
MonthlyRate
No
Yes
Attrition

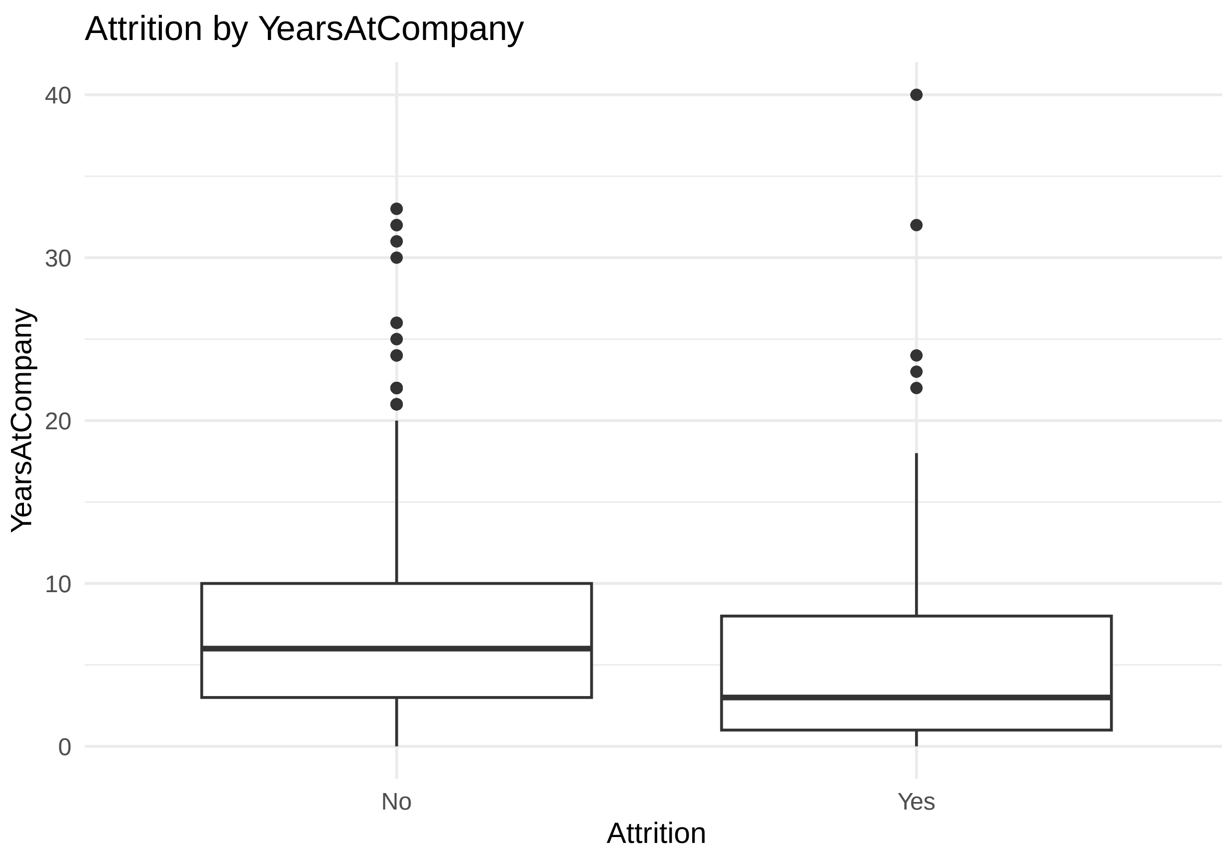


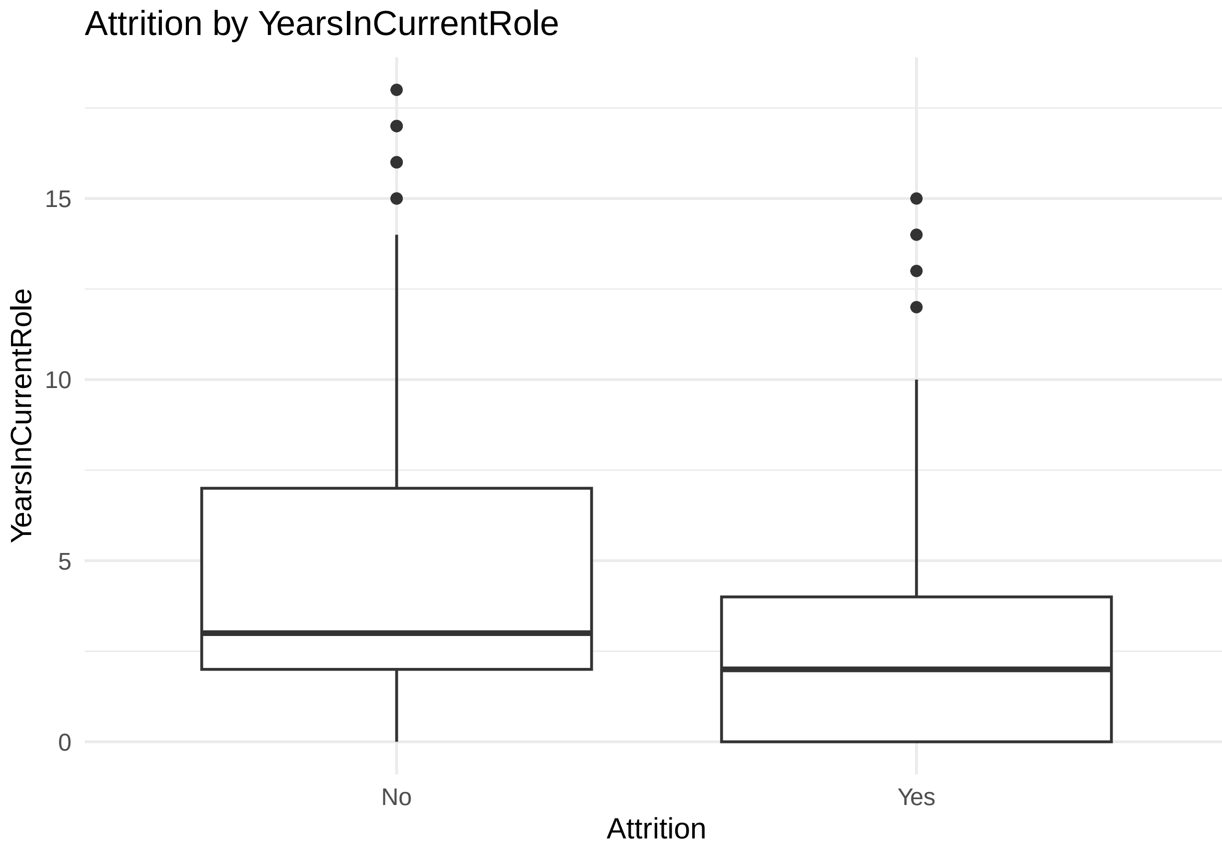
Attrition by NumCompaniesWorked
NumCompaniesWorked
7.5
5.0
2.5
0.0
No
Yes
Attrition


Attrition by PercentSalaryHike
25
20
15
PercentSalaryHike
No
Yes
Attrition

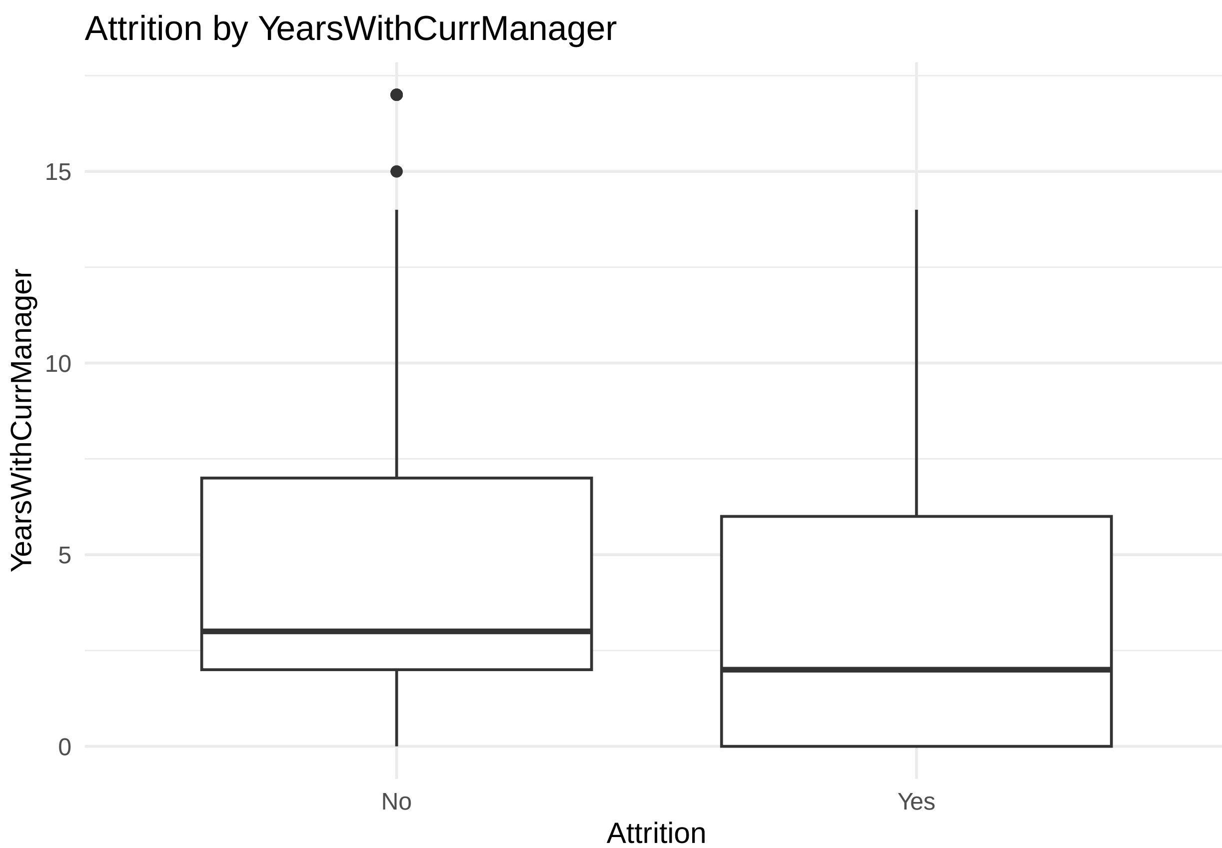

Attrition by TotalWorkingYears
40
30
20
10
0
TotalWorkingYears
No
Yes
Attrition


Attrition by TrainingTimesLastYear
6
4
2
0
TrainingTimesLastYear
No
Yes
Attrition

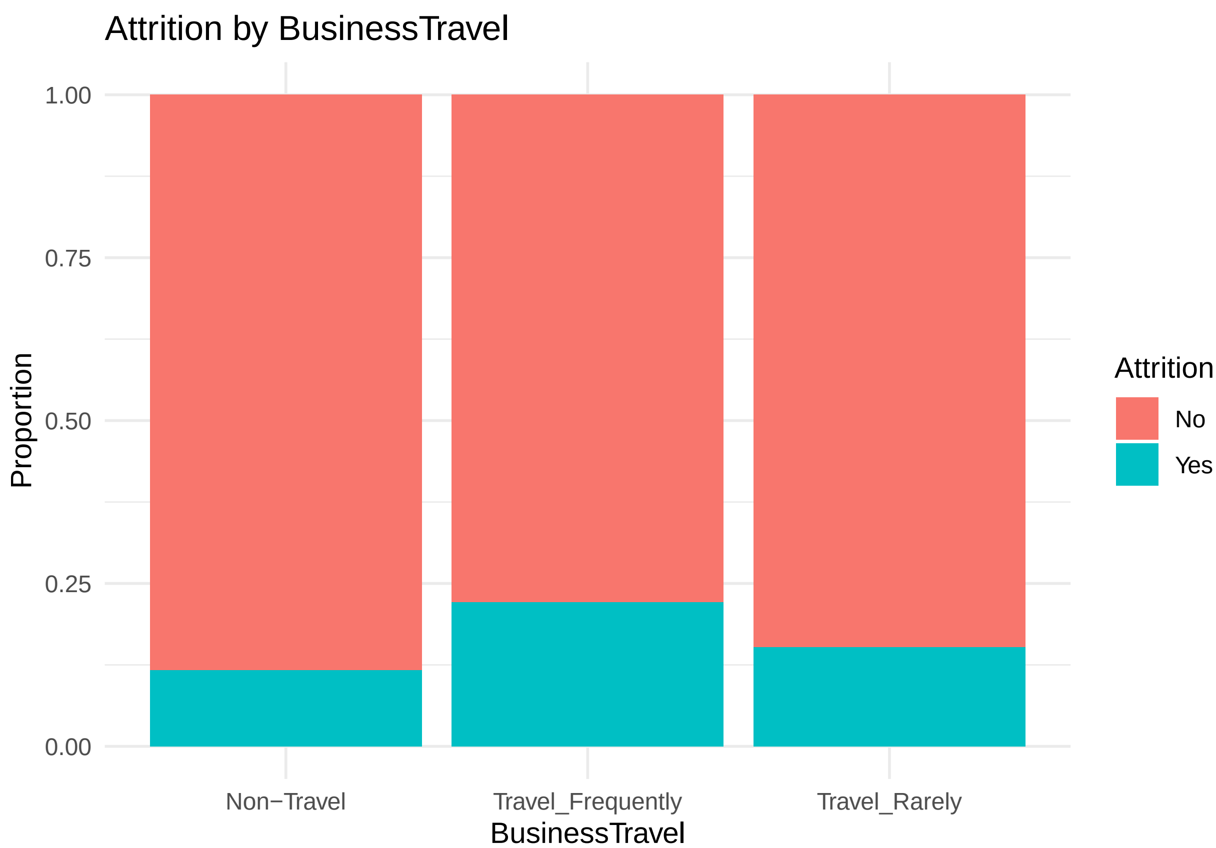

Attrition by YearsAtCompany
40
30
20
10
0
YearsAtCompany
No
Yes
Attrition

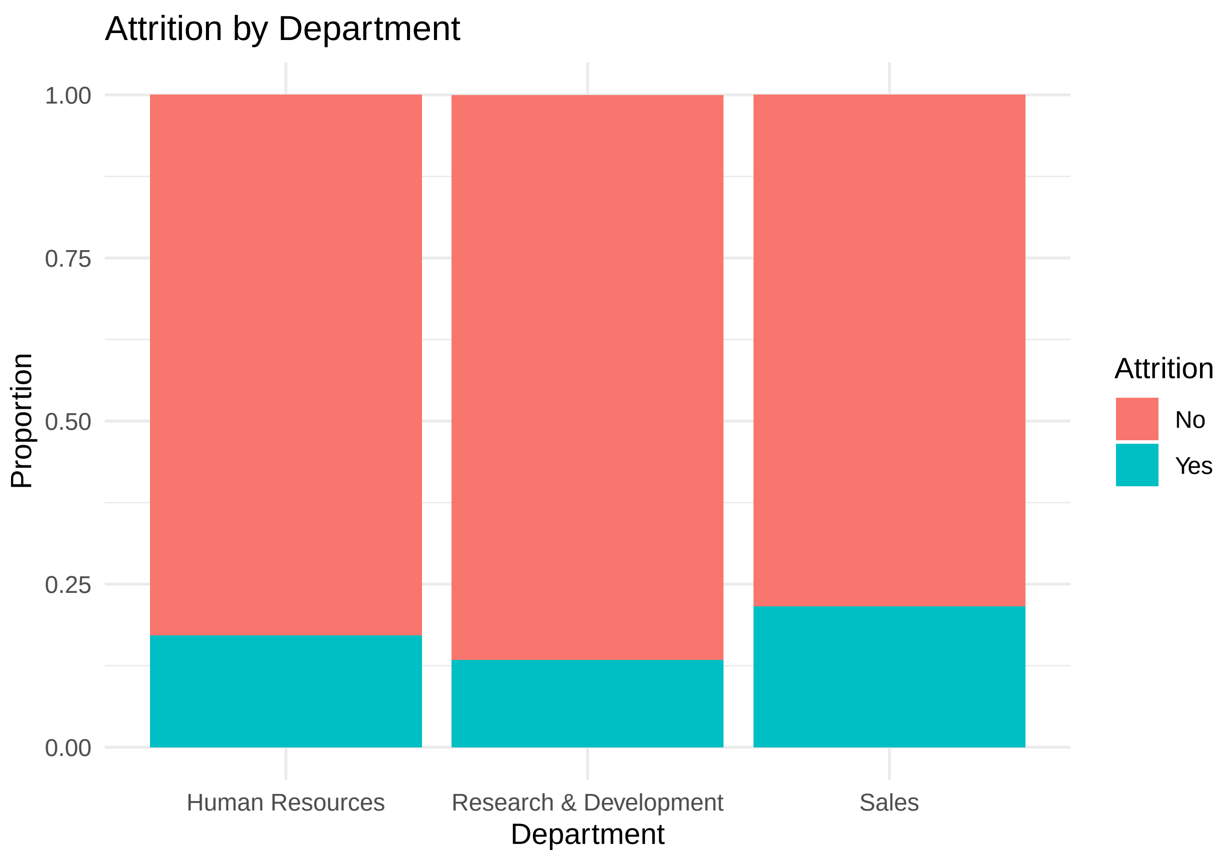


Attrition by YearsInCurrentRole
YearsInCurrentRole
15
10
5
0
No
Yes
Attrition

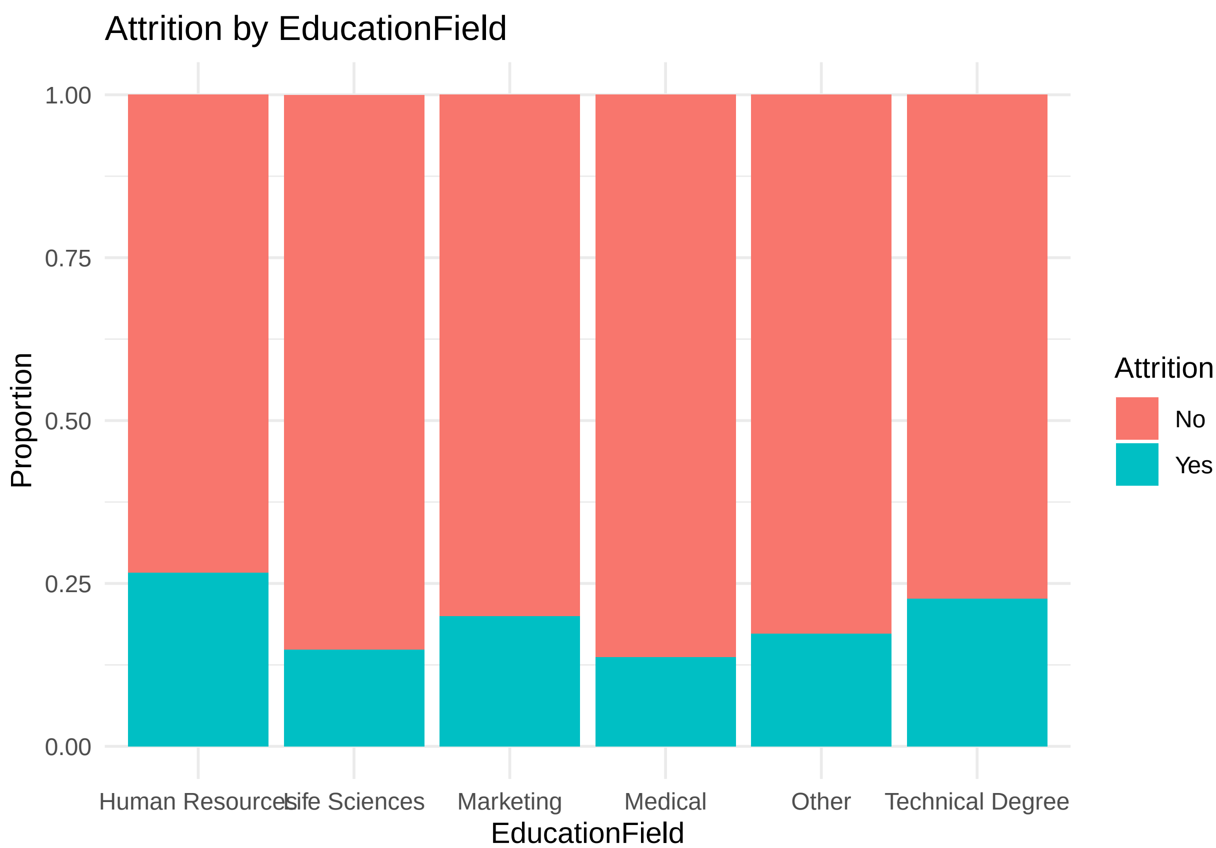



```
# Create bar plots for categorical variables
for (var in categorical_vars) {
  p <- ggplot(data, aes_string(x = var, fill = "Attrition")) +
    geom_bar(position = "fill") +
    labs(title = paste("Attrition by", var), y = "Proportion", x = var) +
    theme_minimal()
  print(p)
}
```

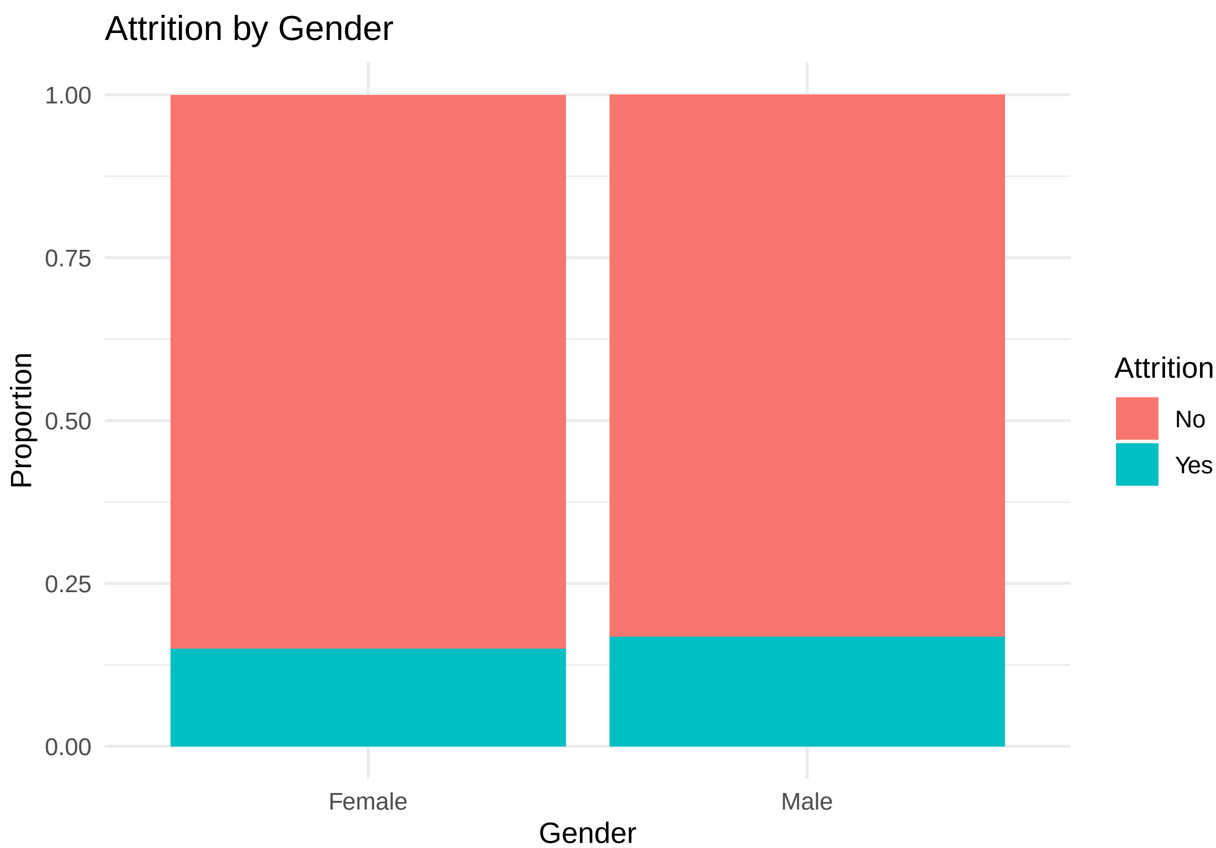

Attrition by BusinessTravel
1.00
0.75
0.50
0.25
0.00
Proportion
Non−Travel
Travel_Frequently
Travel_Rarely
BusinessTravel
Attrition
No
Yes

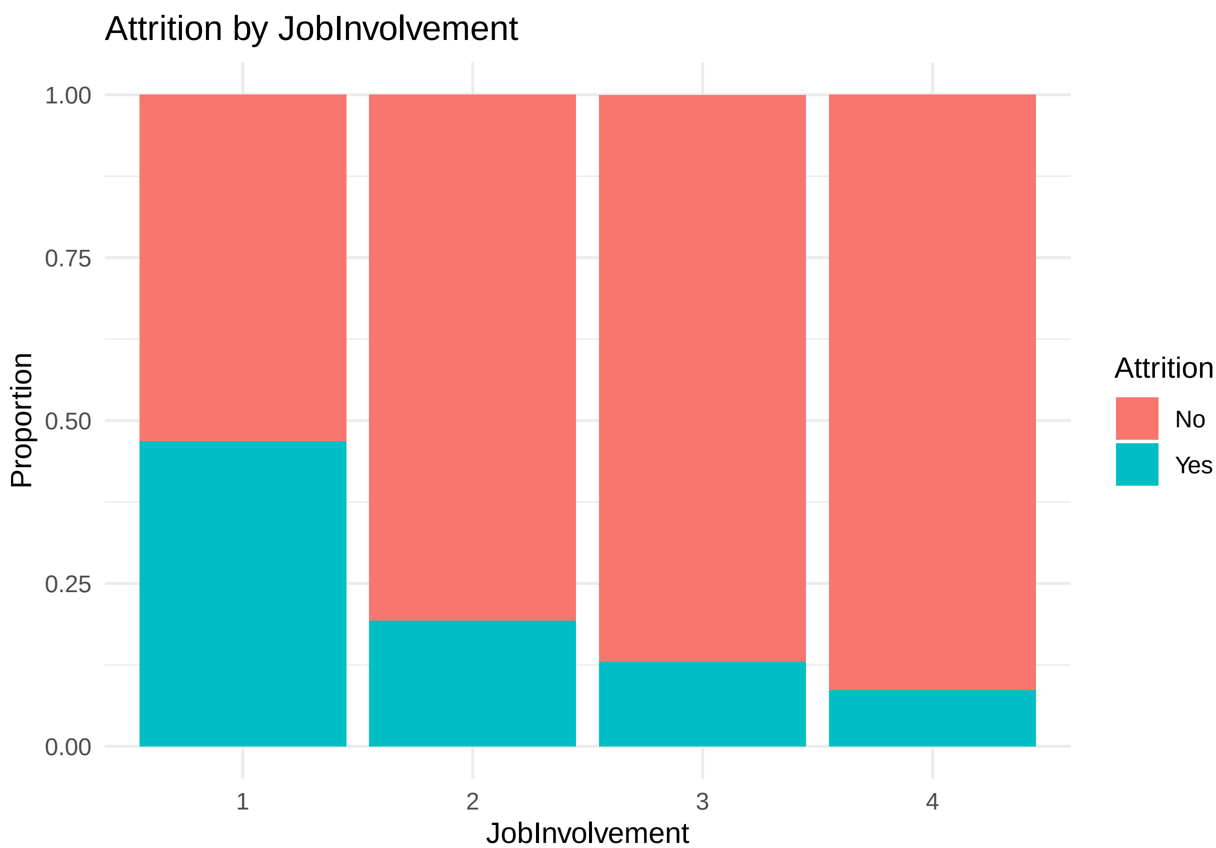


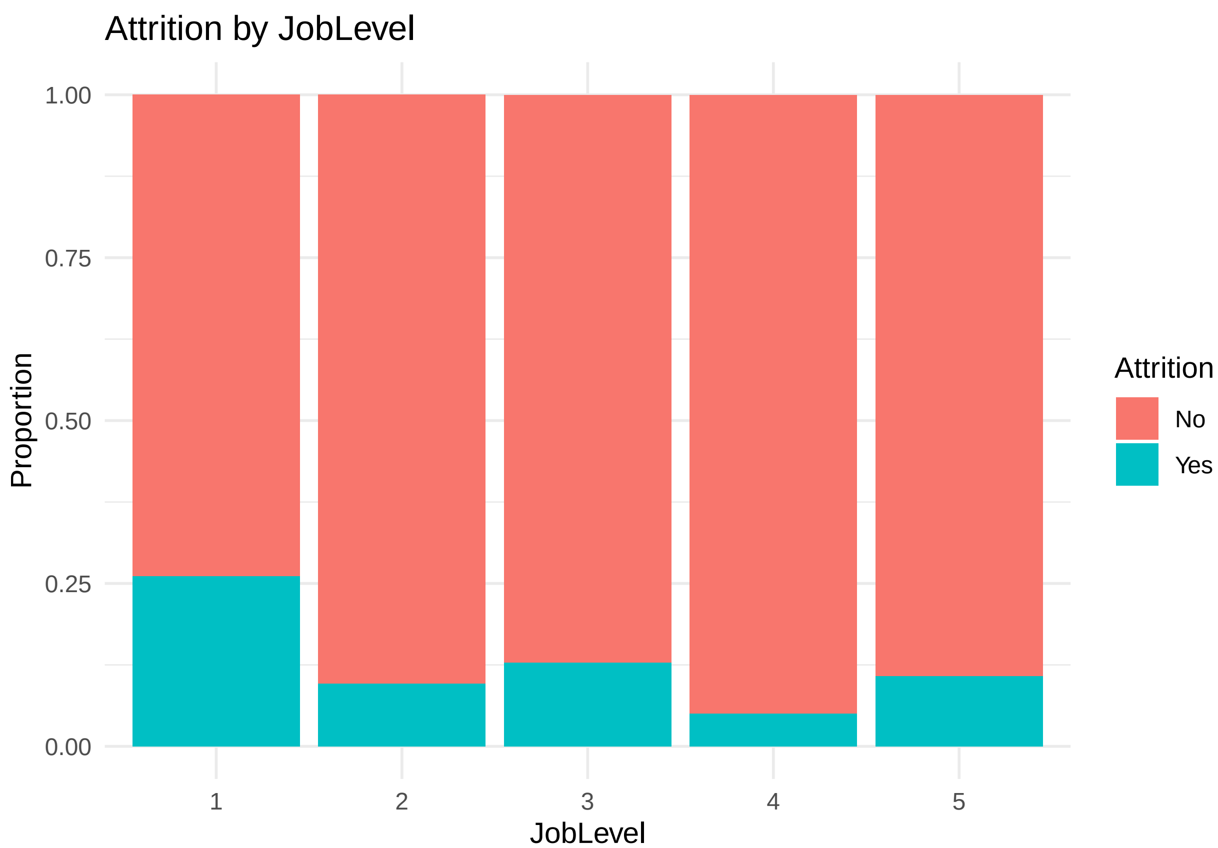
Attrition by Department
1.00
0.75
0.50
0.25
0.00
Proportion
Human Resources
Research & Development
Sales
Department
Attrition
No
Yes



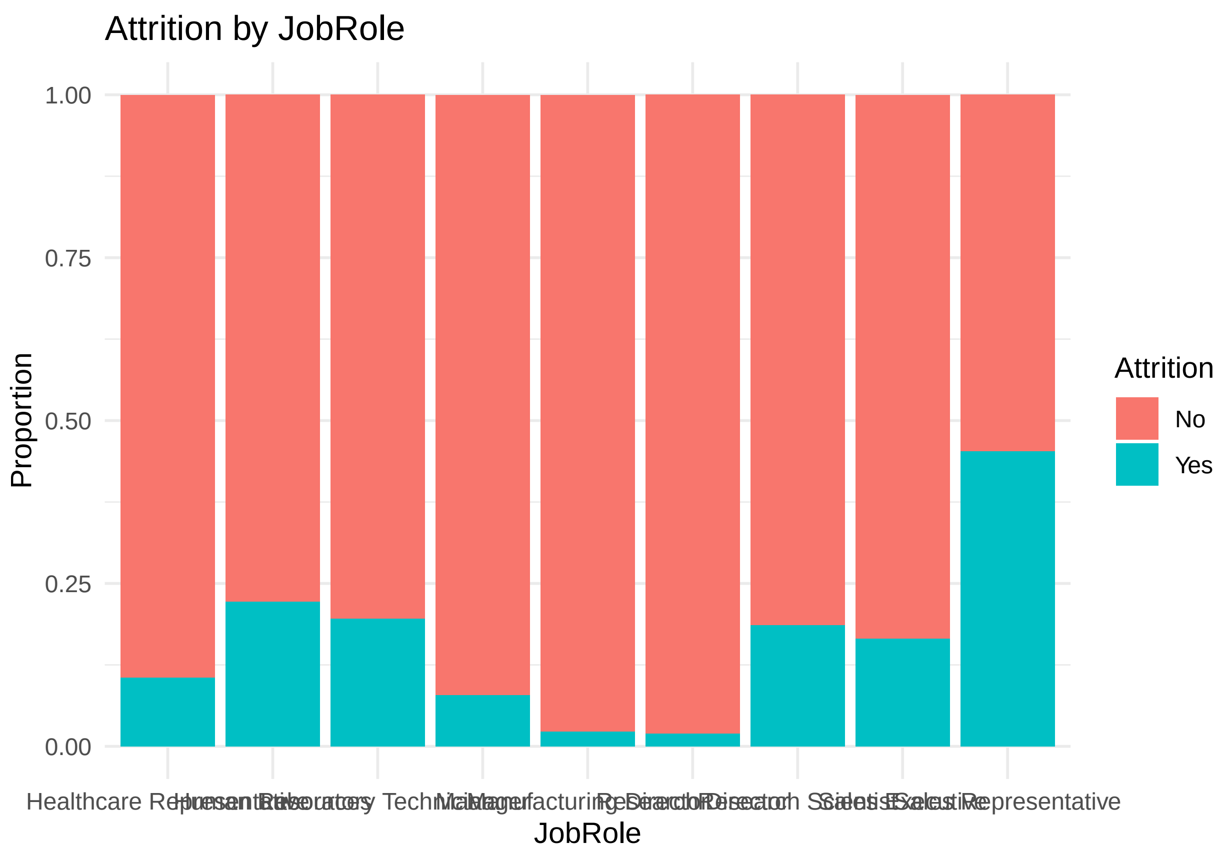
Attrition by EducationField
1.00
0.75
0.50
0.25
0.00
Proportion
Human Resources
Life Sciences
Marketing
Medical
Other
Technical Degree
EducationField
Attrition
No
Yes

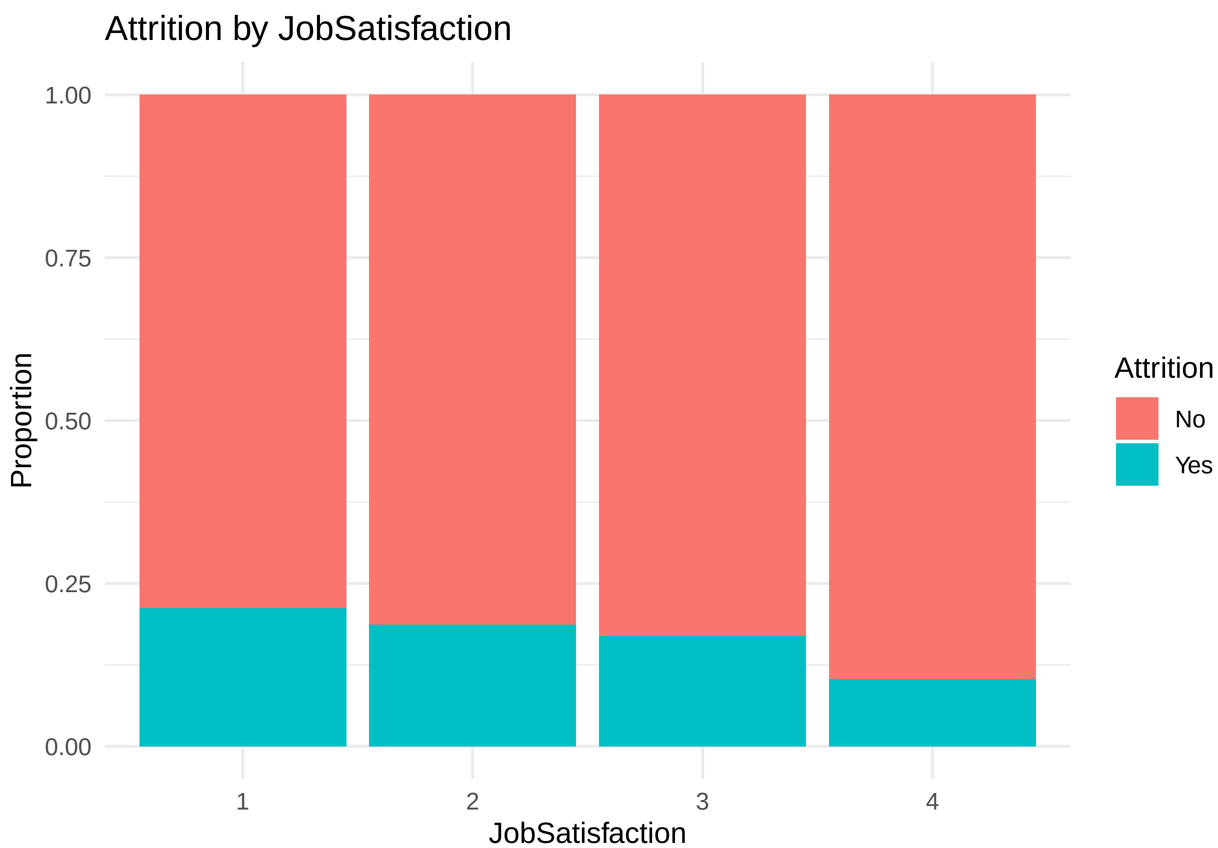

Attrition by Gender
Proportion
1.00
0.75
0.50
0.25
0.00
Female
Male
Gender
Attrition
No
Yes

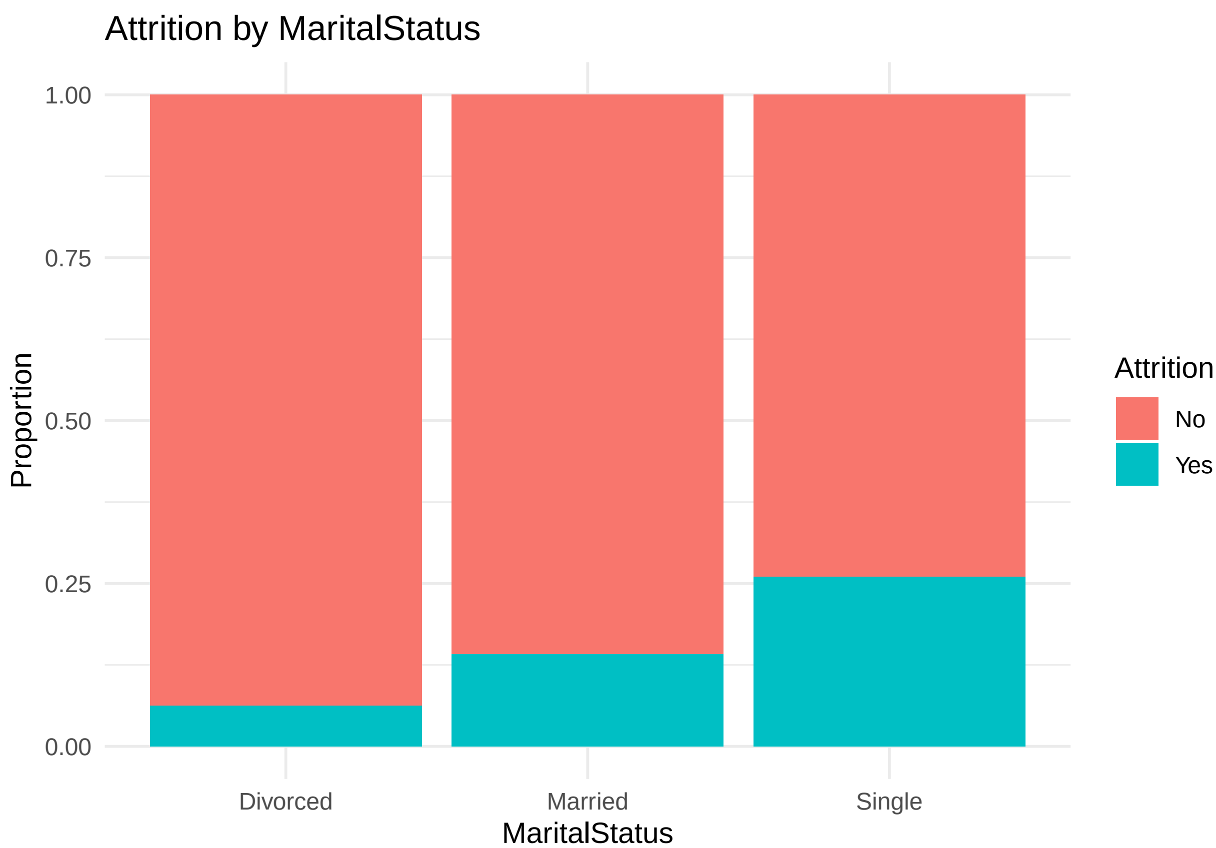

Attrition by JobInvolvement
Proportion
1.00
0.75
0.50
0.25
0.00
1
2
3
4
JobInvolvement
Attrition
No
Yes

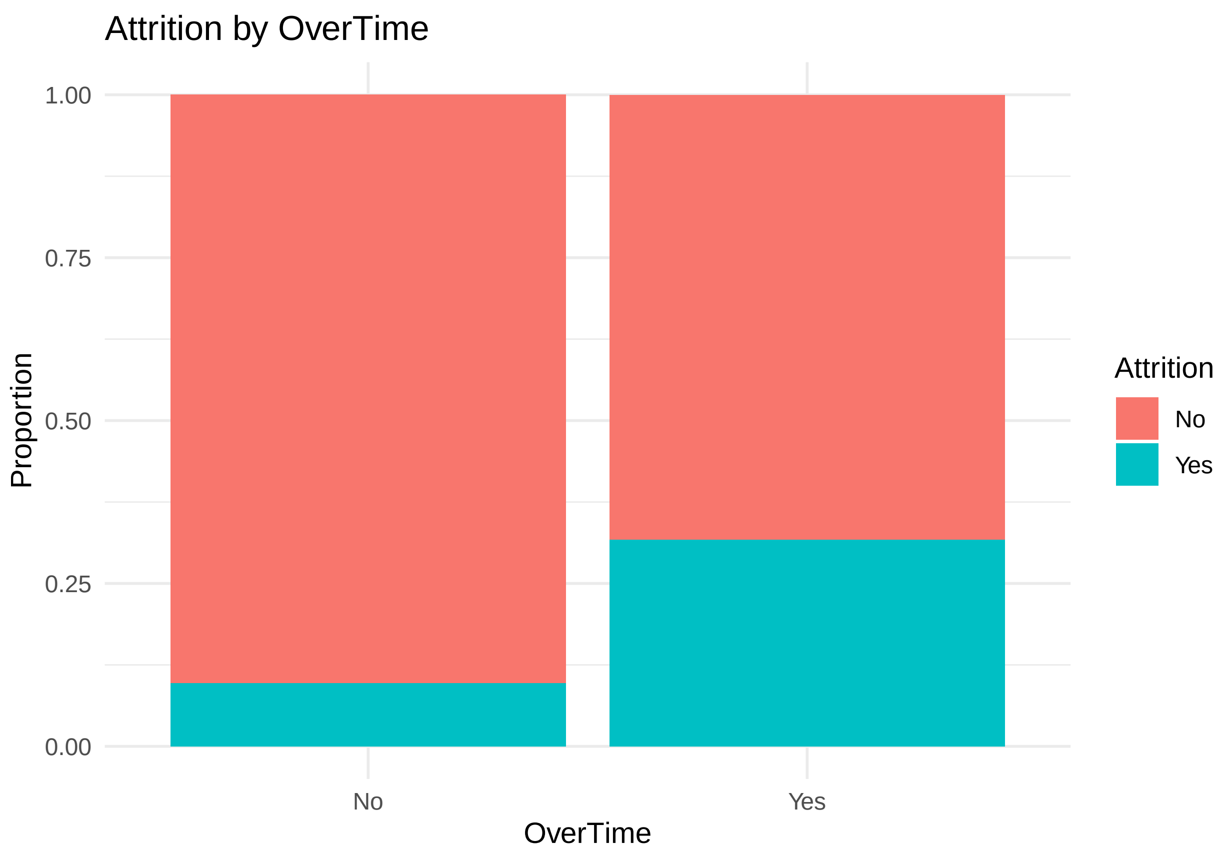


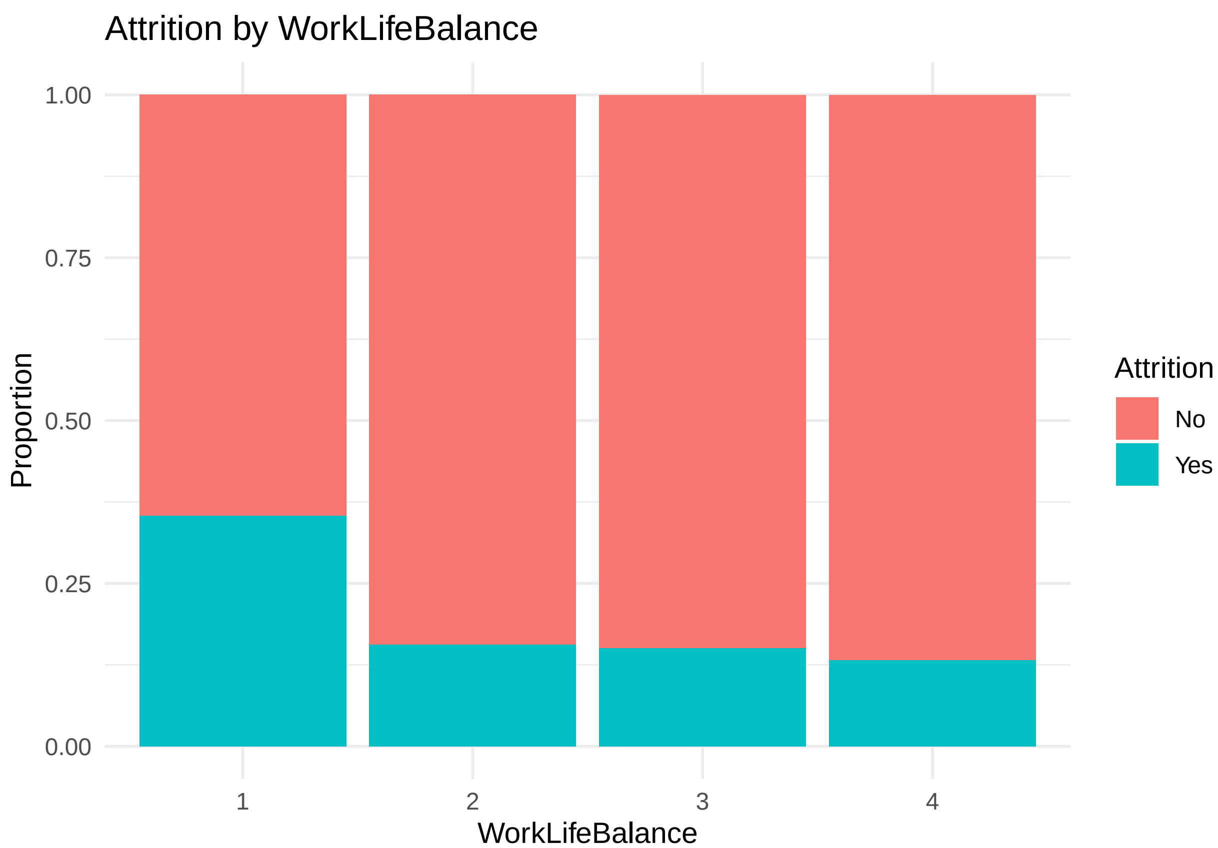
Attrition by JobLevel
Proportion
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
JobLevel
Attrition
No
Yes

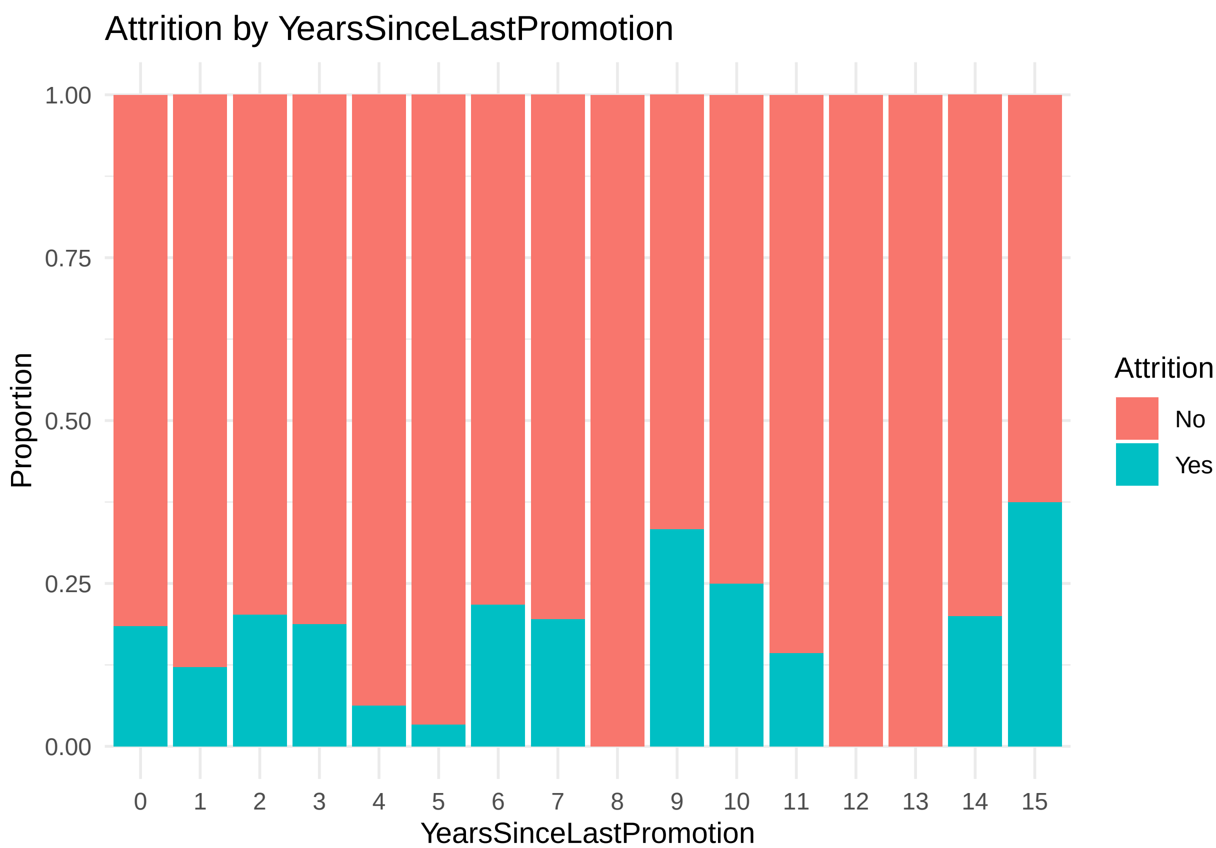

Attrition by JobRole
1.00
0.75
0.50
0.25
0.00
Proportion
JobRole
Attrition
No
Yes


Attrition by JobSatisfaction
1.00
0.75
0.50
0.25
0.00
Proportion
1
2
3
4
JobSatisfaction
Attrition
No
Yes

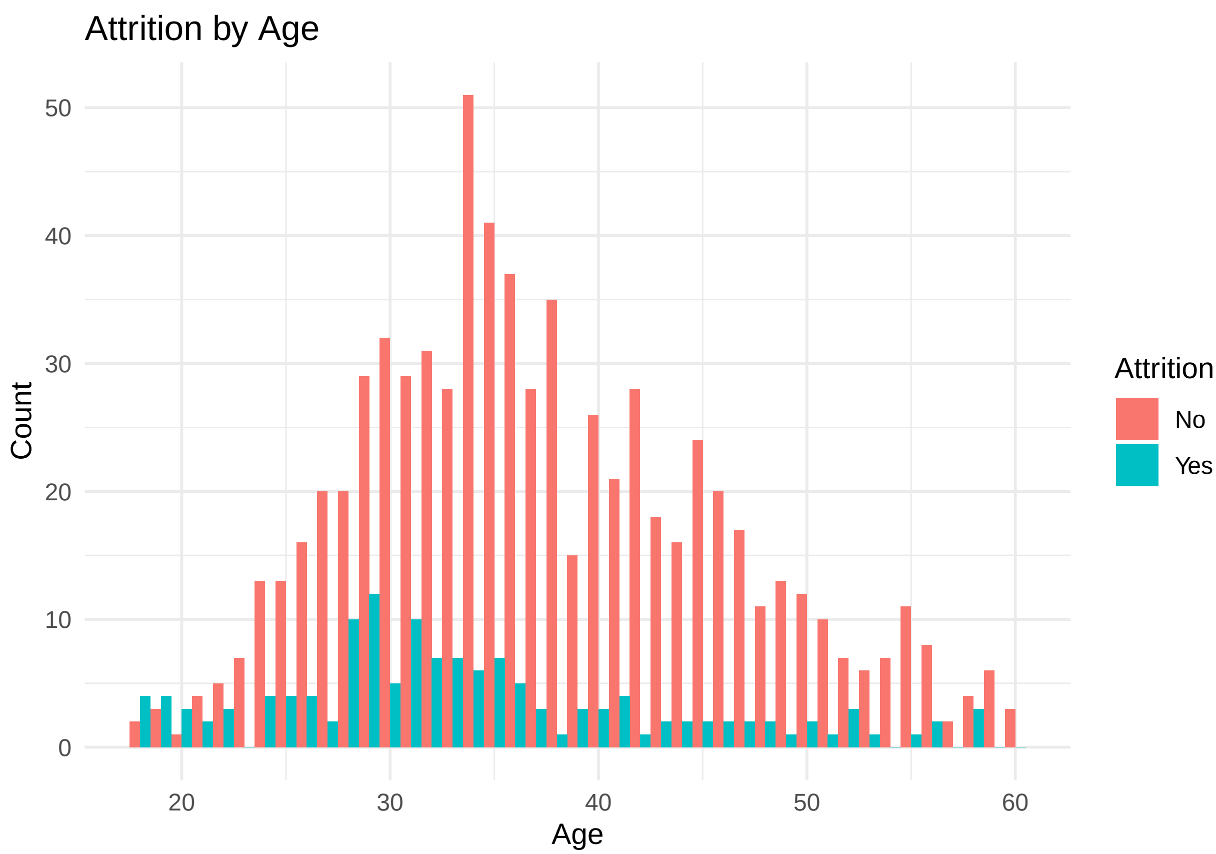

Attrition by MaritalStatus
1.00
0.75
0.50
0.25
0.00
Proportion
Divorced
Married
Single
MaritalStatus
Attrition
No
Yes


Attrition by OverTime
Proportion
1.00
0.75
0.50
0.25
0.00
No
Yes
OverTime
Attrition
No
Yes

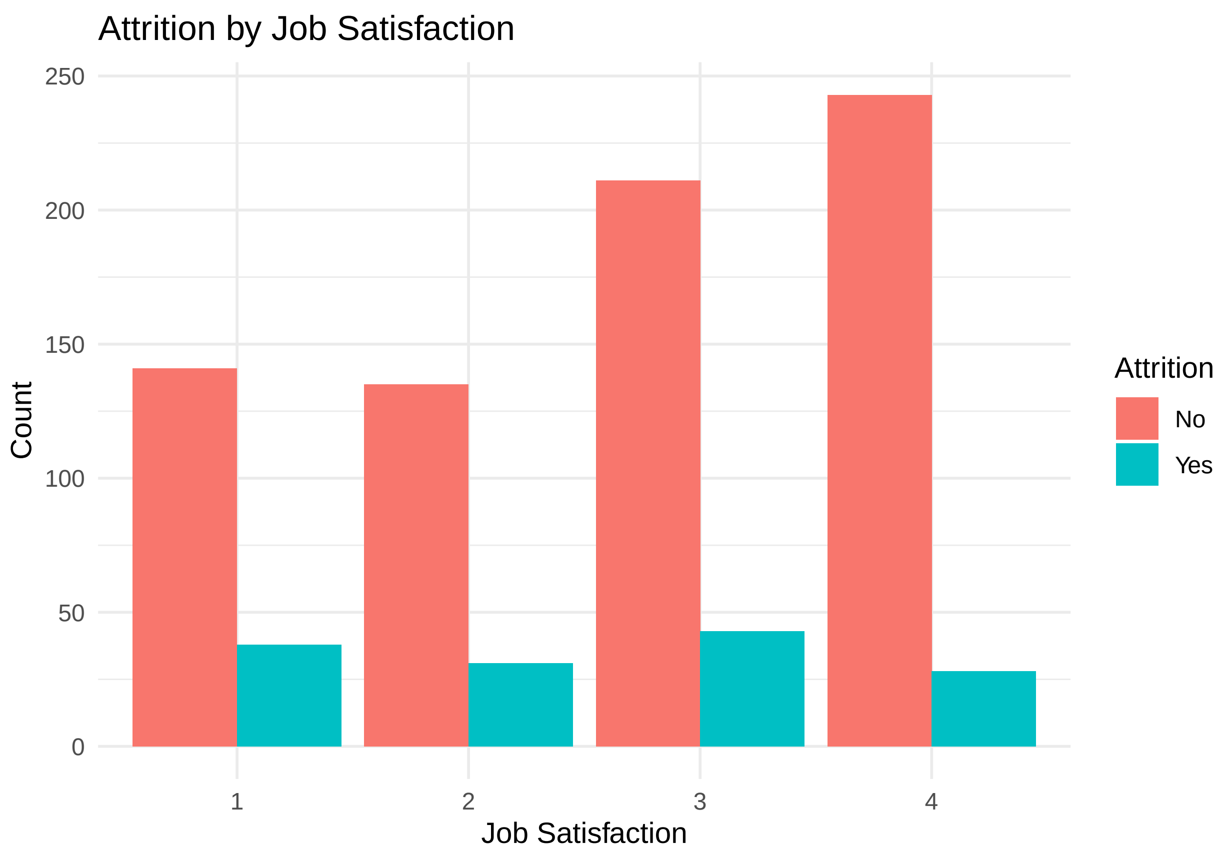

Attrition by WorkLifeBalance
1.00
0.75
0.50
0.25
0.00
Proportion
1
2
3
4
WorkLifeBalance
Attrition
No
Yes



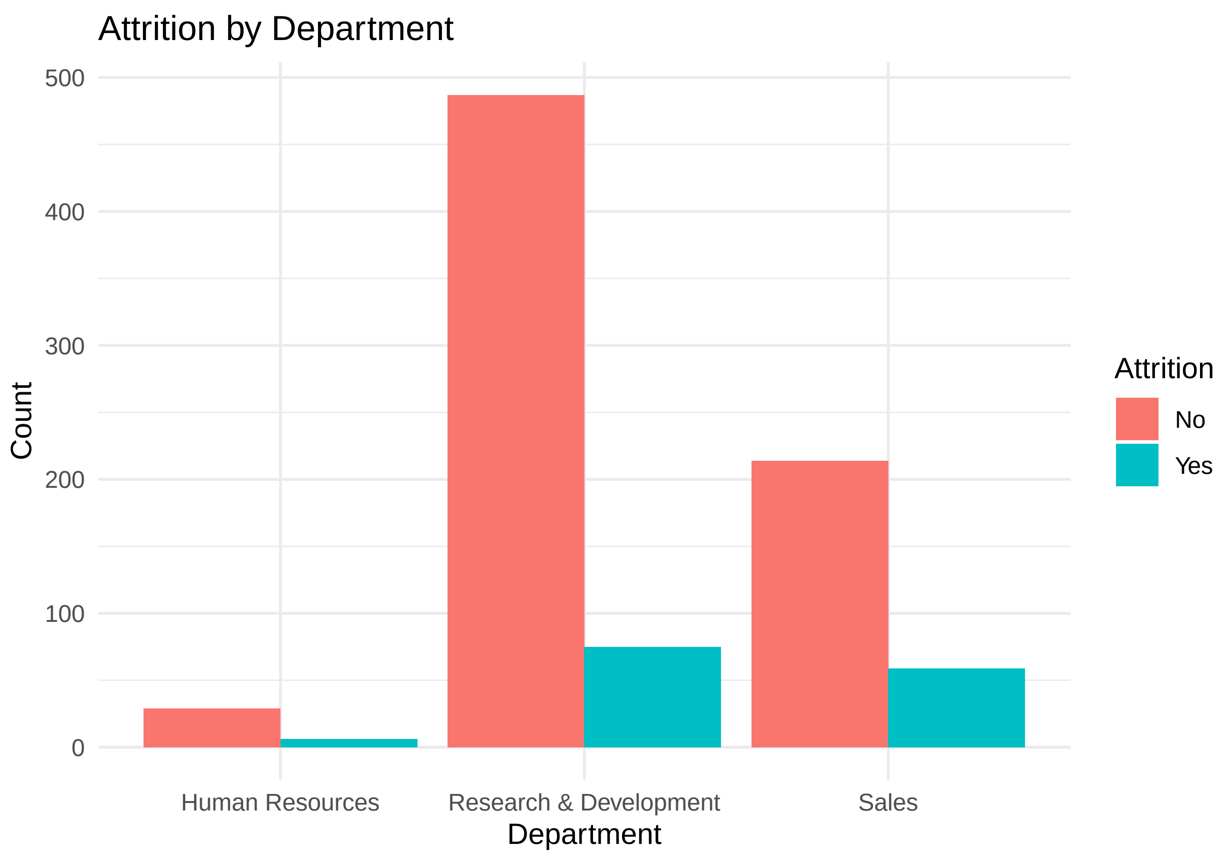

```
# Example: Visualize the relationship between Age and Attrition
ggplot(data, aes(x = Age, fill = Attrition)) +
  geom_histogram(binwidth = 1, position = "dodge") +
  labs(title = "Attrition by Age", x = "Age", y = "Count") +
  theme_minimal()
```



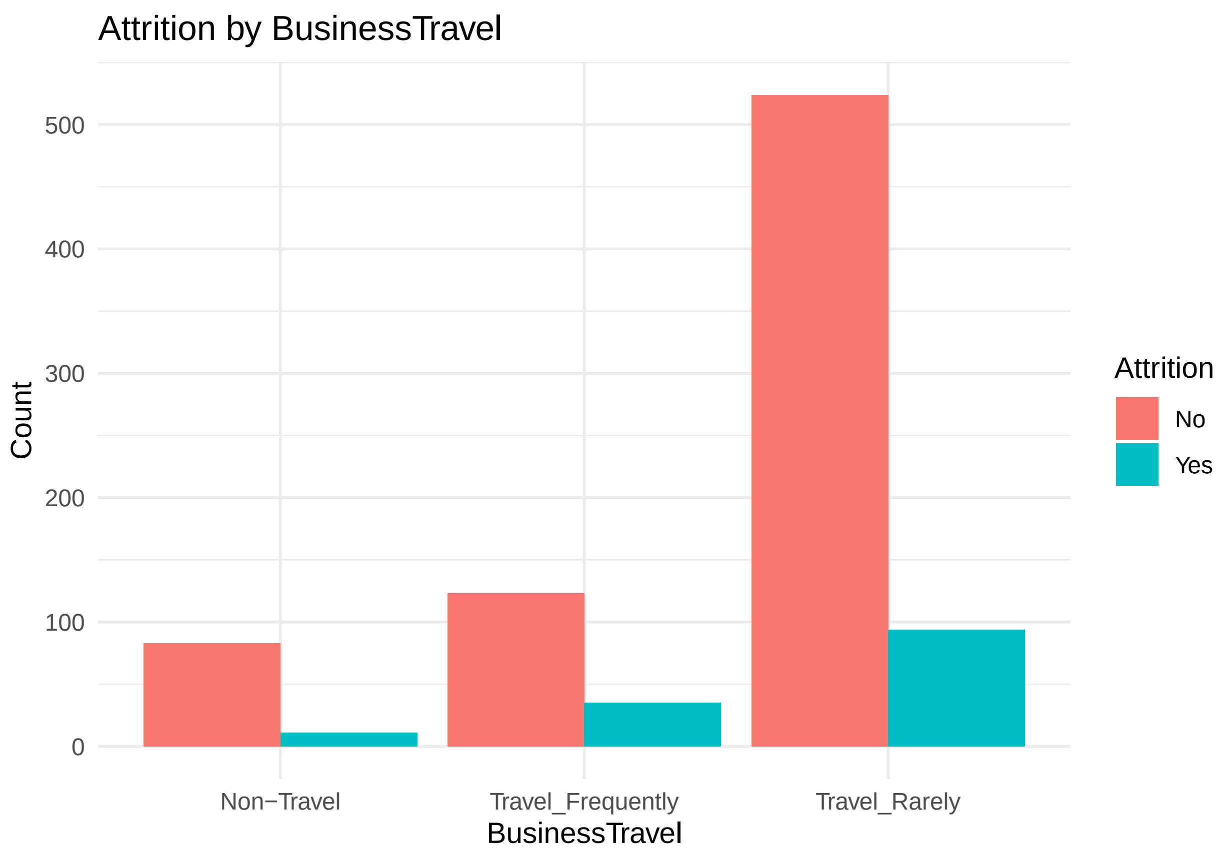

```
ggplot(data, aes(x = JobSatisfaction, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by Job Satisfaction", x = "Job Satisfaction", y = "Count") +
  theme_minimal()
```



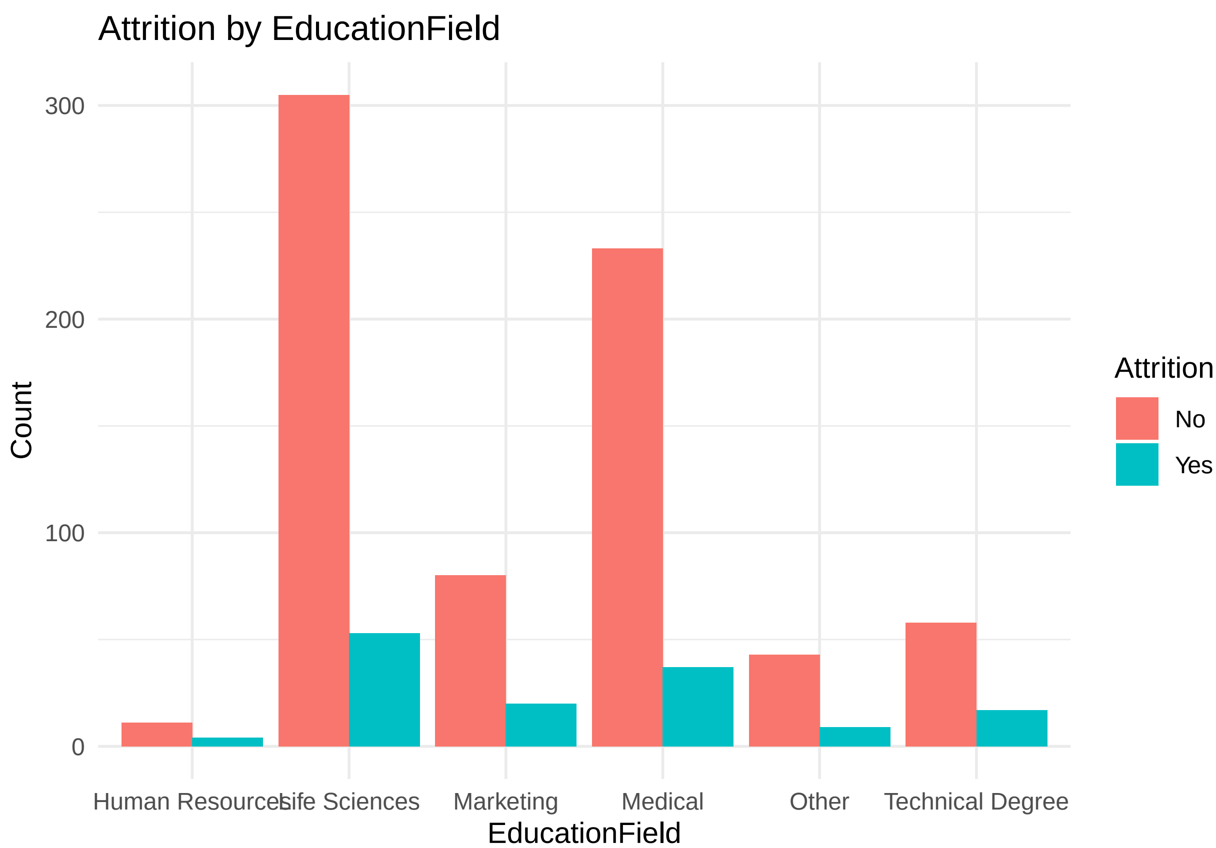

```r
ggplot(data, aes(x = Department, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by Department", x = "Department", y = "Count") +
  theme_minimal()
```



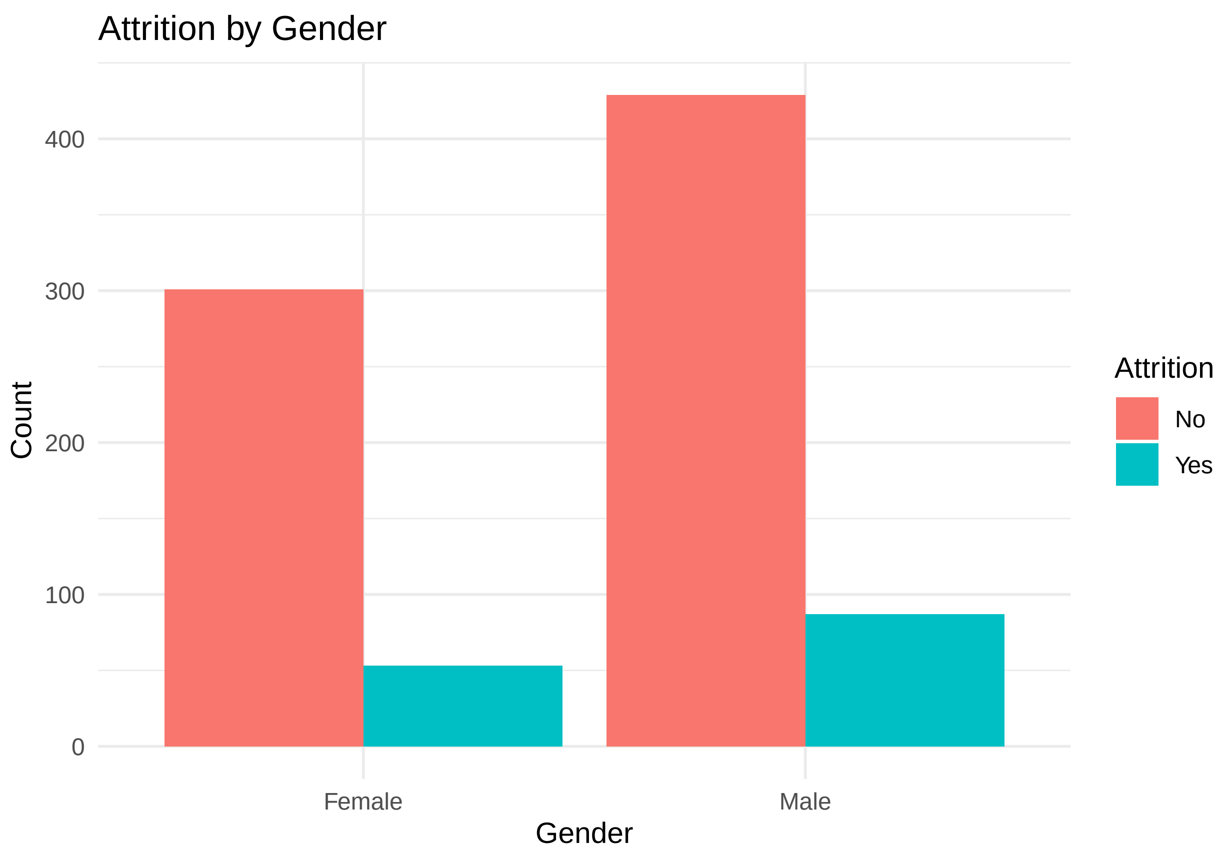

```
ggplot(data, aes(x = BusinessTravel, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by BusinessTravel", x = "BusinessTravel", y = "Count") +
  theme_minimal()
```



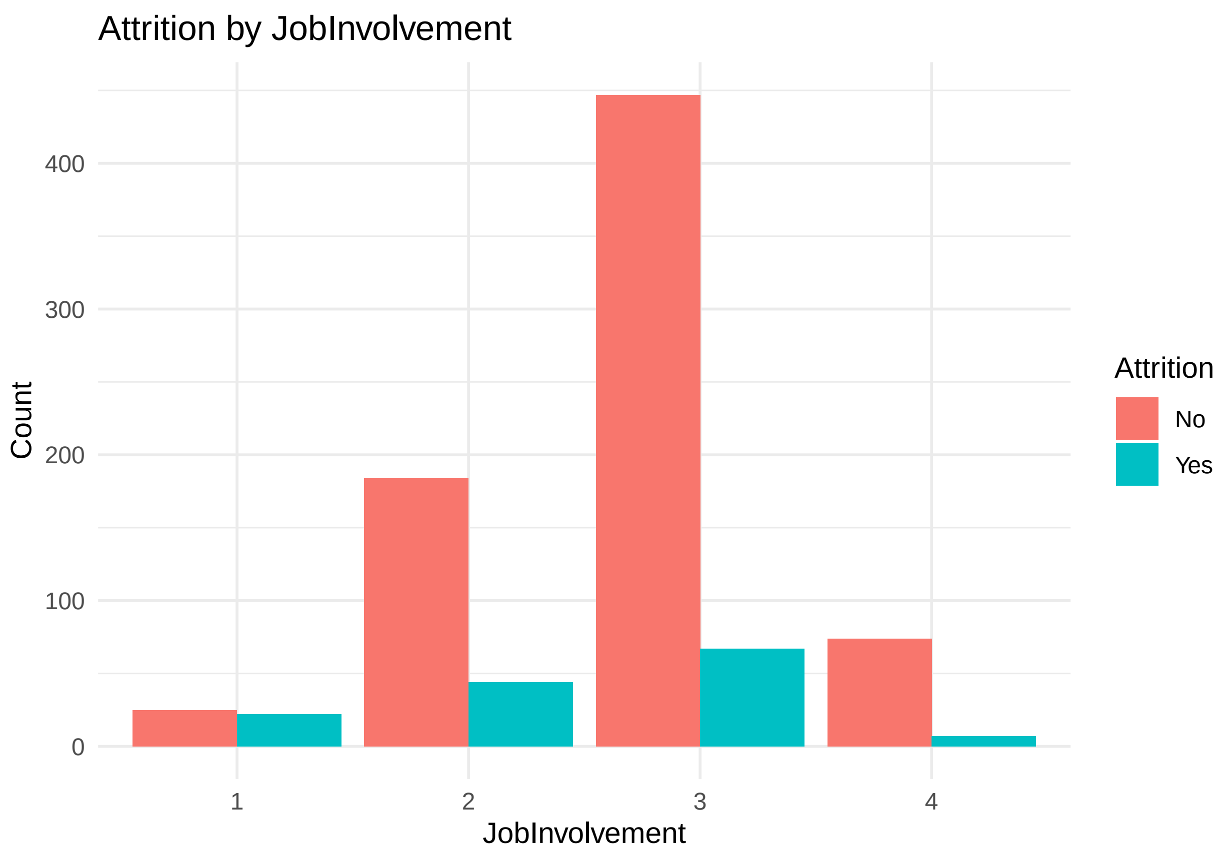

```r
ggplot(data, aes(x = EducationField, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by EducationField", x = "EducationField", y = "Count") +
  theme_minimal()
```



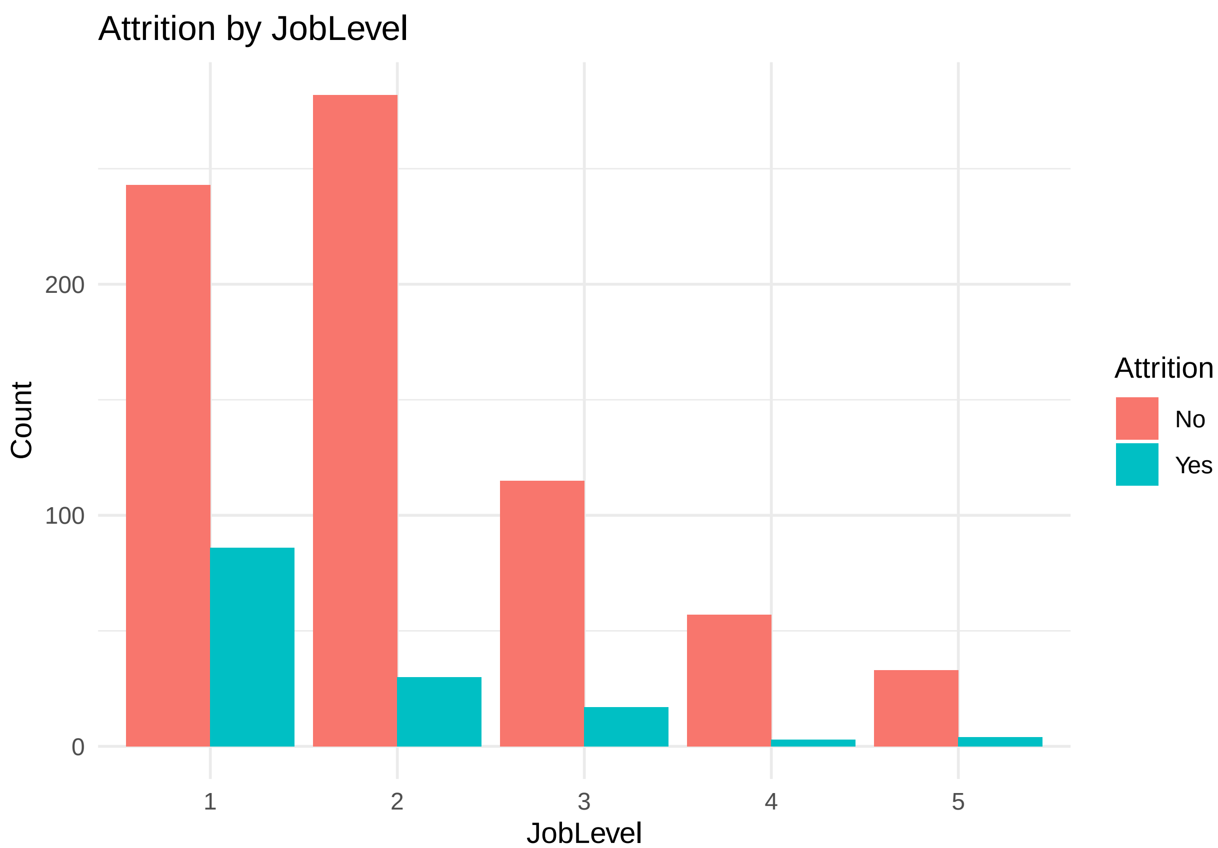

```r
ggplot(data, aes(x = Gender, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by Gender", x = "Gender", y = "Count") +
  theme_minimal()
```



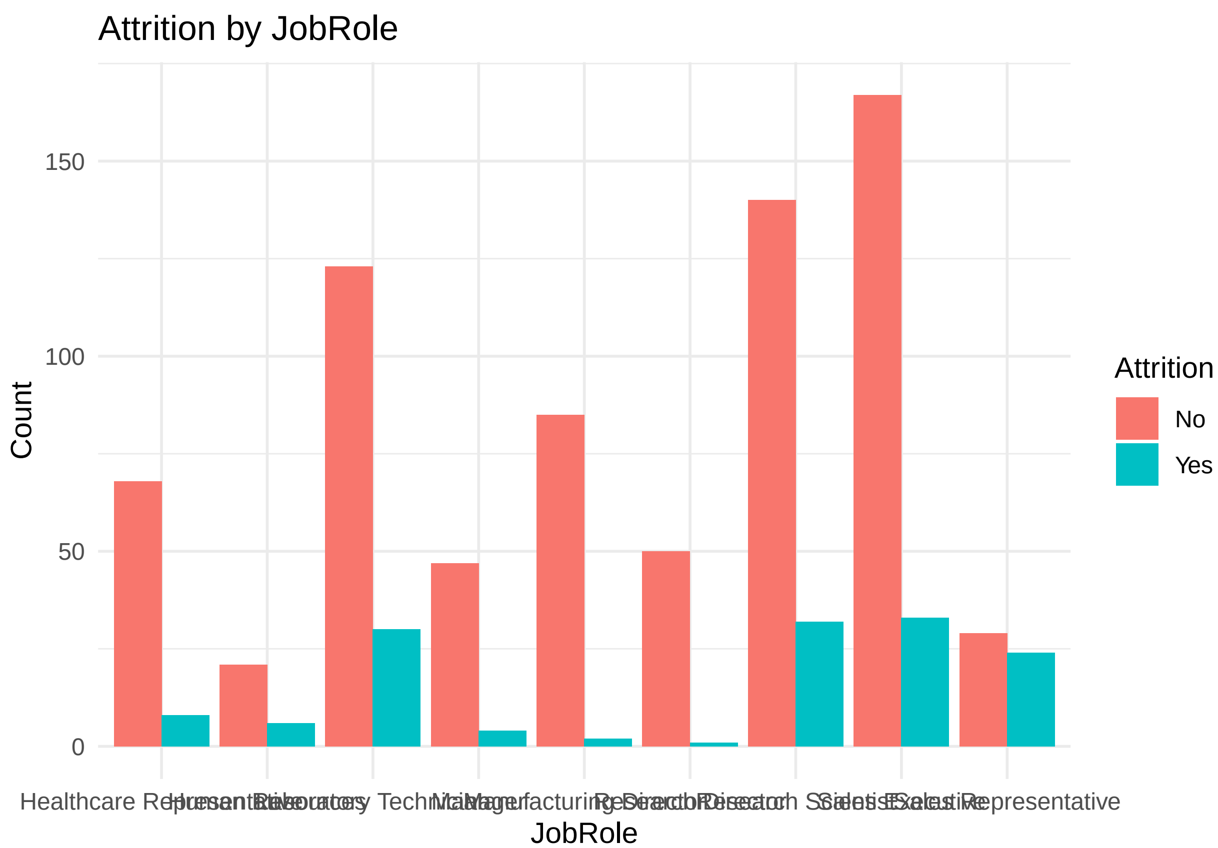

```r
ggplot(data, aes(x = JobInvolvement, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by JobInvolvement", x = "JobInvolvement", y = "Count") +
  theme_minimal()
```



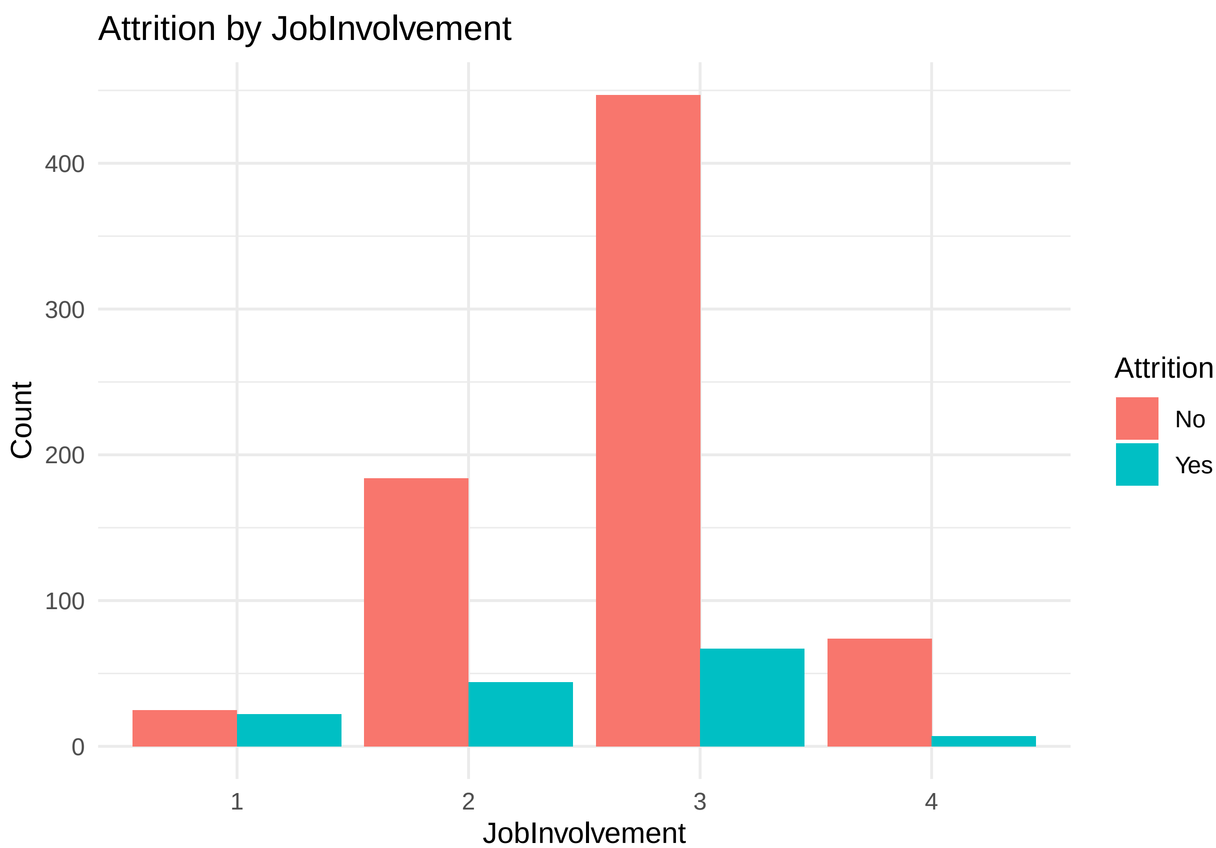

```r
ggplot(data, aes(x = JobLevel, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by JobLevel", x = "JobLevel", y = "Count") +
  theme_minimal()
```



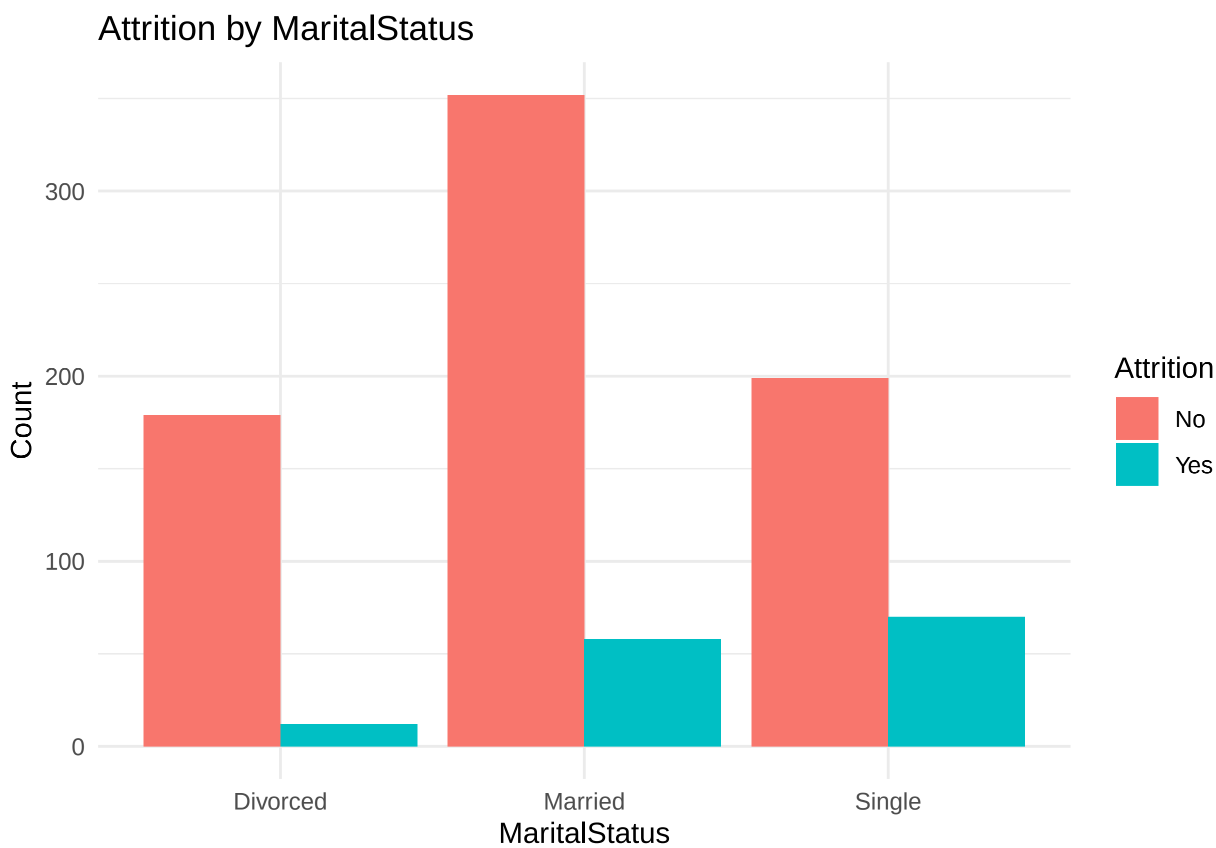

```r
ggplot(data, aes(x = JobRole, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by JobRole", x = "JobRole", y = "Count") +
  theme_minimal()
```



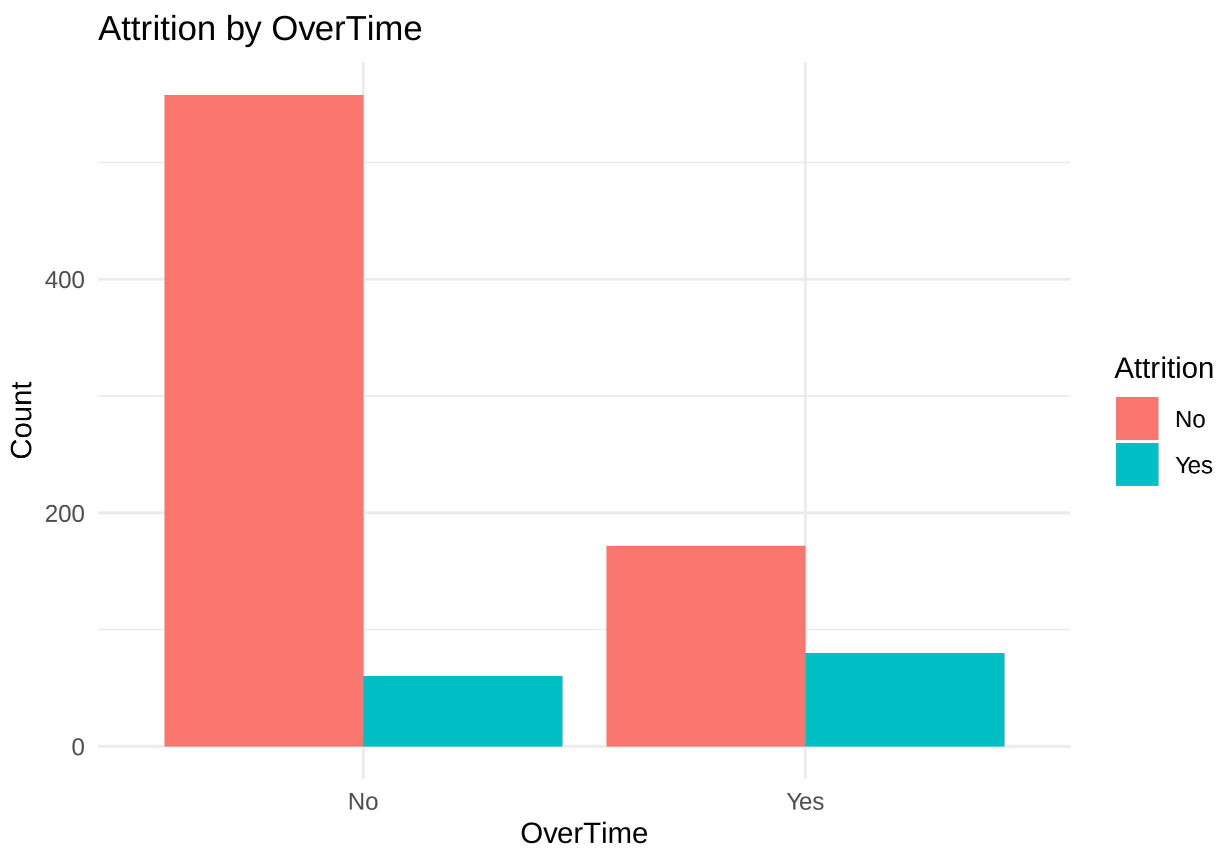

```r
ggplot(data, aes(x = JobInvolvement, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by JobInvolvement", x = "JobInvolvement", y = "Count") +
  theme_minimal()
```




```r
ggplot(data, aes(x = MaritalStatus, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by MaritalStatus", x = "MaritalStatus", y = "Count") +
  theme_minimal()
```




```r
ggplot(data, aes(x = OverTime, fill = Attrition)) +
  geom_bar(position = "dodge") +
  labs(title = "Attrition by OverTime", x = "OverTime", y = "Count") +
  theme_minimal()
```



Explore Job Role-Specific Trends: Examine trends related to specific job roles, such as variations in job satisfaction.

```r
# Check data structure
str(data)
```

```
## 'data.frame':    870 obs. of  36 variables:
##  $ ID                      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age                     : int  32 40 35 32 24 27 41 37 34 34 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 2 2 3 3
##  $ DailyRate               : int  117 1308 200 801 567 294 1283 309 1333 653 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 2 2 3 2 2 2 3 3 2 ...
##  $ DistanceFromHome        : int  13 14 18 1 2 10 5 10 10 10 ...
##  $ Education               : int  4 3 2 4 1 2 5 4 4 4 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 4 2 3 6 2 4 2 2 6 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  859 1128 1412 2016 1646 733 1448 1105 1055 1597 ...
##  $ EnvironmentSatisfaction : int  2 3 3 3 1 4 2 4 3 4 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 2 2 1 1 2 ...
##  $ HourlyRate              : int  73 44 60 48 32 32 90 88 87 92 ...
##  $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 2 3 3 3 3 4 2 3 2 ...
```

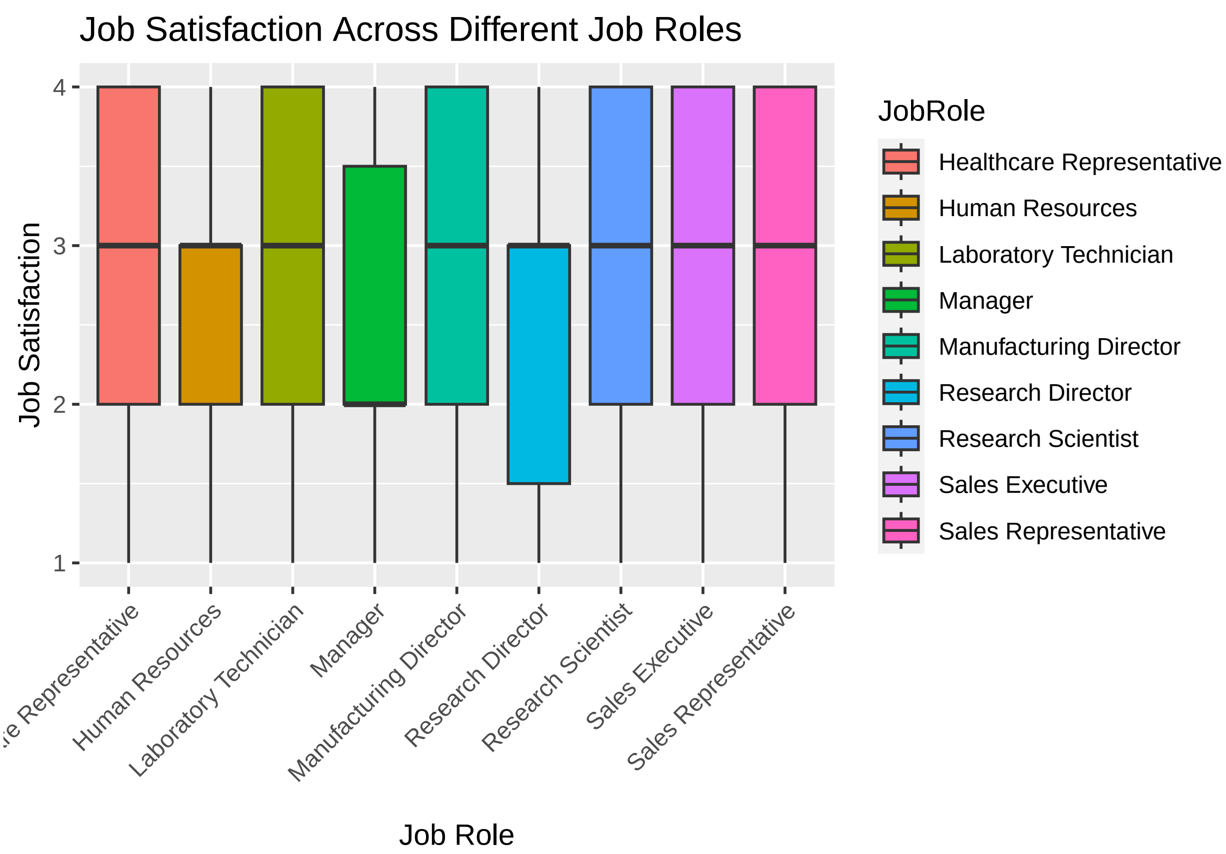
```
##  $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 5 3 3 1 3 1 2 1 2 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 6 5 8 7 5 7 8 9 1
##  $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 3 4 4 4 1 3 4 3 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 1 3 3 2 3 1 2 1 2 2 ...
##  $ MonthlyIncome           : int  4403 19626 9362 10422 3760 8793 2127 6694 2220 5063 ...
##  $ MonthlyRate             : int  9250 17544 19944 24032 17218 4809 5561 24223 18410 15332 ...
##  $ NumCompaniesWorked      : int  2 1 2 1 1 1 2 2 1 1 ...
##  $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 2 1 ...
##  $ PercentSalaryHike       : int  11 14 11 19 13 21 12 14 19 14 ...
##  $ PerformanceRating       : int  3 3 3 3 3 4 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  3 1 3 3 3 3 1 3 4 2 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  1 0 0 2 0 2 0 3 1 1 ...
##  $ TotalWorkingYears       : int  8 21 10 14 6 9 7 8 1 8 ...
##  $ TrainingTimesLastYear   : int  3 2 2 3 2 4 5 5 2 3 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 2 4 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : int  5 20 2 14 6 9 4 1 1 8 ...
##  $ YearsInCurrentRole      : int  2 7 2 10 3 7 2 0 1 2 ...
##  $ YearsSinceLastPromotion : Factor w/ 16 levels "0","1","2","3",..: 1 5 3 6 2 2 1 1 1 8 ...
##  $ YearsWithCurrManager    : int  3 9 2 7 3 7 3 0 0 7 ...
```

```r
# Convert factors to numeric
data$JobSatisfaction <- as.numeric(as.character(data$JobSatisfaction))

# Descriptive statistics
jobRoleTable <- table(data$JobRole)
jobSatisfactionSummary <- summary(data$JobSatisfaction)

# Descriptive Statistics by Job Role
job_satisfaction_by_role <- data %>%
  group_by(JobRole) %>%
  summarise(
    Count = n(),
    Mean = mean(JobSatisfaction, na.rm = TRUE),
    SD = sd(JobSatisfaction, na.rm = TRUE),
    Min = min(JobSatisfaction, na.rm = TRUE),
    Max = max(JobSatisfaction, na.rm = TRUE),
    Median = median(JobSatisfaction, na.rm = TRUE),
    IQR = IQR(JobSatisfaction, na.rm = TRUE)
  )
job_satisfaction_by_role
```

```
## # A tibble: 9 x 8
##   JobRole                   Count  Mean    SD   Min   Max Median   IQR
##   <fct>                     <int> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 Healthcare Representative    76  2.83  1.15     1     4      3   2
## 2 Human Resources              27  2.56  1.05     1     4      3   1
## 3 Laboratory Technician       153  2.69  1.12     1     4      3   2
## 4 Manager                      51  2.51  1.12     1     4      2   1.5
## 5 Manufacturing Director       87  2.72  1.01     1     4      3   2
## 6 Research Director            51  2.49  1.10     1     4      3   1.5
## 7 Research Scientist          172  2.80  1.12     1     4      3   2
## 8 Sales Executive             200  2.72  1.16     1     4      3   2
```

```
## 9 Sales Representative          53  2.70  1.08     1     4      3   2
```

```r
#View(job_satisfaction_by_role)

# Visualization
ggplot(data, aes(x = JobRole, y = JobSatisfaction, fill = JobRole)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Job Satisfaction Across Different Job Roles", x = "Job Role", y = "Job Satisfaction")
```




```r
# ANOVA Test
anova_result <- aov(JobSatisfaction ~ JobRole, data = data)
anova_summary <- summary(anova_result)

# LM
lm_model <- lm(JobSatisfaction ~ JobRole + WorkLifeBalance + YearsAtCompany + DistanceFromHome + Age + D
summary(lm_model)
```

```
##
## Call:
## lm(formula = JobSatisfaction ~ JobRole + WorkLifeBalance + YearsAtCompany +
##     DistanceFromHome + Age + DailyRate + Gender + JobLevel, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.0650 -0.7762  0.2255  1.1338  1.7337
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.740e+00  3.129e-01   8.757   <2e-16 ***
## JobRoleHuman Resources       -2.958e-01  2.814e-01  -1.051    0.293
## JobRoleLaboratory Technician -1.912e-01  1.950e-01  -0.980    0.327
## JobRoleManager               -3.232e-01  2.649e-01  -1.220    0.223
## JobRoleManufacturing Director -9.529e-02  1.759e-01  -0.542    0.588
## JobRoleResearch Director     -2.523e-01  2.365e-01  -1.067    0.286
## JobRoleResearch Scientist    -9.843e-02  1.986e-01  -0.496    0.620
## JobRoleSales Executive       -1.014e-01  1.516e-01  -0.669    0.504
## JobRoleSales Representative  -1.792e-01  2.464e-01  -0.727    0.467
## WorkLifeBalance2              2.799e-01  1.818e-01   1.540    0.124
## WorkLifeBalance3              6.822e-02  1.705e-01   0.400    0.689
## WorkLifeBalance4              1.431e-01  1.983e-01   0.721    0.471
## YearsAtCompany                1.309e-02  7.589e-03   1.725    0.085 .
## DistanceFromHome             -2.749e-03  4.695e-03  -0.585    0.558
## Age                          -3.997e-04  4.920e-03  -0.081    0.935
## DailyRate                    -1.666e-05  9.515e-05  -0.175    0.861
## GenderMale                    4.044e-02  7.801e-02   0.518    0.604
## JobLevel2                    -4.380e-02  1.508e-01  -0.290    0.772
## JobLevel3                    -2.531e-01  1.993e-01  -1.270    0.204
## JobLevel4                    -1.244e-01  2.794e-01  -0.445    0.656
## JobLevel5                    -2.049e-01  3.375e-01  -0.607    0.544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.116 on 849 degrees of freedom
## Multiple R-squared:  0.02044,    Adjusted R-squared:  -0.002635
## F-statistic: 0.8858 on 20 and 849 DF,  p-value: 0.6059
```

#Visual analysis indicates that the job roles of Human Resources, Manager, and Research Director have lower than average levels of job satisfaction. However, the output from the linear model reveals that none of these job roles have a statistically significant impact on job satisfaction, as evidenced by p-values all exceeding the typical alpha level of 0.05.

#The residuals of the model, which measure the differences between observed and predicted values of job satisfaction, range from -2.0650 to 1.7337, with a median close to zero. This suggests that the model's predictions are not biased towards overestimating or underestimating job satisfaction.

#Regarding outliers, the range of residuals indicates individual cases where actual job satisfaction is much higher or lower than predicted by the model. Additionally, the model's low multiple R-squared value of 0.02044, indicating that only about 2% of the variability in job satisfaction is explained by all the combined predictors, suggests that job satisfaction is influenced by factors not included in this model.

#The overall F-statistic p-value of 0.6059 confirms that the model does not provide a statistically significant fit to the data, implying that the included variables do not have strong predictive power for job satisfaction.

#Additional study is recommended to explore other influencing factors.

#Build a model to predict employee attrition. The model should achieve at least 60% sensitivity and specificity (60 each = 120 total) for both the training and validation sets.

```r
#LM model for predict employee attrition

#variables
continuous_vars <- c("Age", "DailyRate", "DistanceFromHome", "Education", "HourlyRate", "MonthlyIncome"

categorical_vars <- c("BusinessTravel", "Department", "EducationField", "Gender", "JobInvolvement", "Jo


#LM model with multiple variables
glmlog_model <- glm(Attrition ~ Age + DailyRate + DistanceFromHome + Education + HourlyRate + MonthlyIn
summary(glmlog_model)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + DailyRate + DistanceFromHome +
##     Education + HourlyRate + MonthlyIncome + MonthlyRate + NumCompaniesWorked +
##     PercentSalaryHike + TotalWorkingYears + TrainingTimesLastYear +
##     YearsAtCompany + YearsInCurrentRole + YearsWithCurrManager +
##     BusinessTravel + Department + EducationField + Gender + JobInvolvement +
##     JobLevel + JobRole + JobSatisfaction + MaritalStatus + OverTime +
##     WorkLifeBalance + YearsSinceLastPromotion, family = "binomial",
##     data = data)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.442e+01  1.047e+03  -0.014 0.989005
## Age                              -3.117e-02  1.963e-02  -1.588 0.112254
## DailyRate                        -2.477e-04  3.276e-04  -0.756 0.449574
## DistanceFromHome                  6.121e-02  1.657e-02   3.695 0.000220 ***
## Education                         1.644e-02  1.326e-01   0.124 0.901298
## HourlyRate                        1.546e-02  6.878e-03   2.248 0.024558 *
## MonthlyIncome                    -7.851e-06  1.376e-04  -0.057 0.954499
## MonthlyRate                      -1.721e-05  1.901e-05  -0.905 0.365374
## NumCompaniesWorked                2.317e-01  5.915e-02   3.917 8.97e-05 ***
## PercentSalaryHike                 5.243e-03  3.647e-02   0.144 0.885695
## TotalWorkingYears                -8.041e-02  4.415e-02  -1.821 0.068588 .
## TrainingTimesLastYear            -2.953e-01  1.113e-01  -2.653 0.007980 **
## YearsAtCompany                    1.077e-01  6.073e-02   1.773 0.076251 .
## YearsInCurrentRole               -1.878e-01  7.922e-02  -2.370 0.017785 *
## YearsWithCurrManager             -1.358e-01  6.885e-02  -1.972 0.048630 *
## BusinessTravelTravel_Frequently   1.802e+00  5.588e-01   3.225 0.001259 **
## BusinessTravelTravel_Rarely       9.453e-01  4.891e-01   1.933 0.053248 .
## DepartmentResearch & Development  1.672e+01  1.047e+03   0.016 0.987259
## DepartmentSales                   1.749e+01  1.047e+03   0.017 0.986672
## EducationFieldLife Sciences      -9.661e-01  1.240e+00  -0.779 0.435875
## EducationFieldMarketing          -1.072e+00  1.307e+00  -0.820 0.412022
## EducationFieldMedical            -9.974e-01  1.229e+00  -0.811 0.417202
## EducationFieldOther              -1.167e+00  1.316e+00  -0.887 0.375072
## EducationFieldTechnical Degree   -2.748e-01  1.287e+00  -0.214 0.830919
## GenderMale                        4.226e-01  2.712e-01   1.558 0.119145
## JobInvolvement2                  -1.945e+00  5.127e-01  -3.794 0.000148 ***
## JobInvolvement3                  -2.722e+00  4.975e-01  -5.471 4.48e-08 ***
## JobInvolvement4                  -3.156e+00  6.932e-01  -4.553 5.29e-06 ***
```

```
## JobLevel2                          -1.731e+00  6.788e-01  -2.550 0.010778 *
## JobLevel3                          -3.348e-01  1.054e+00  -0.318 0.750703
## JobLevel4                          -1.671e+00  1.756e+00  -0.951 0.341418
## JobLevel5                           2.642e+00  2.258e+00   1.170 0.241990
## JobRoleHuman Resources              1.685e+01  1.047e+03   0.016 0.987156
## JobRoleLaboratory Technician       -7.633e-02  7.767e-01  -0.098 0.921709
## JobRoleManager                     -1.847e+00  1.628e+00  -1.135 0.256423
## JobRoleManufacturing Director      -1.377e+00  9.253e-01  -1.488 0.136827
## JobRoleResearch Director           -2.872e+00  1.847e+00  -1.555 0.119931
## JobRoleResearch Scientist          -6.500e-01  7.952e-01  -0.817 0.413657
## JobRoleSales Executive              2.609e-01  1.558e+00   0.167 0.867026
## JobRoleSales Representative         3.320e-01  1.681e+00   0.197 0.843482
## JobSatisfaction                    -4.817e-01  1.217e-01  -3.958 7.56e-05 ***
## MaritalStatusMarried                1.219e+00  4.223e-01   2.887 0.003893 **
## MaritalStatusSingle                 2.158e+00  4.282e-01   5.041 4.64e-07 ***
## OverTimeYes                         2.247e+00  2.859e-01   7.857 3.95e-15 ***
## WorkLifeBalance2                   -1.441e+00  5.084e-01  -2.834 0.004592 **
## WorkLifeBalance3                   -1.898e+00  4.752e-01  -3.994 6.50e-05 ***
## WorkLifeBalance4                   -2.085e+00  6.350e-01  -3.284 0.001025 **
## YearsSinceLastPromotion1           -3.418e-01  3.508e-01  -0.975 0.329749
## YearsSinceLastPromotion2            2.427e-01  4.221e-01   0.575 0.565332
## YearsSinceLastPromotion3            9.876e-01  7.556e-01   1.307 0.191234
## YearsSinceLastPromotion4            3.722e-01  1.072e+00   0.347 0.728524
## YearsSinceLastPromotion5            2.579e-01  1.323e+00   0.195 0.845378
## YearsSinceLastPromotion6            2.961e+00  9.324e-01   3.176 0.001493 **
## YearsSinceLastPromotion7            1.444e+00  6.572e-01   2.198 0.027978 *
## YearsSinceLastPromotion8           -1.376e+01  9.098e+02  -0.015 0.987936
## YearsSinceLastPromotion9            2.899e+00  1.117e+00   2.596 0.009436 **
## YearsSinceLastPromotion10           3.445e+00  2.293e+00   1.503 0.132938
## YearsSinceLastPromotion11           1.559e+00  1.333e+00   1.170 0.242001
## YearsSinceLastPromotion12          -1.465e+01  1.484e+03  -0.010 0.992119
## YearsSinceLastPromotion13          -1.467e+01  1.441e+03  -0.010 0.991881
## YearsSinceLastPromotion14           1.810e+00  3.610e+00   0.501 0.616148
## YearsSinceLastPromotion15           4.925e+00  1.328e+00   3.710 0.000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 767.67  on 869  degrees of freedom
## Residual deviance: 427.77  on 808  degrees of freedom
## AIC: 551.77
##
## Number of Fisher Scoring iterations: 16
```

```
# stepwise to narrow down variables
stepwise_fit <- step(glmlog_model, direction = "both", trace = FALSE)
summary(stepwise_fit)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + DistanceFromHome + HourlyRate +
##     NumCompaniesWorked + TotalWorkingYears + TrainingTimesLastYear +
##     YearsAtCompany + YearsInCurrentRole + YearsWithCurrManager +
```

```
##     BusinessTravel + Department + Gender + JobInvolvement + JobLevel +
##     JobSatisfaction + MaritalStatus + OverTime + WorkLifeBalance +
##     YearsSinceLastPromotion, family = "binomial", data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      1.500e+00  1.275e+00   1.176 0.239441
## Age                             -3.381e-02  1.896e-02  -1.783 0.074573 .
## DistanceFromHome                 6.097e-02  1.586e-02   3.845 0.000121 ***
## HourlyRate                       1.470e-02  6.389e-03   2.302 0.021359 *
## NumCompaniesWorked               2.293e-01  5.571e-02   4.116 3.86e-05 ***
## TotalWorkingYears               -8.256e-02  4.077e-02  -2.025 0.042883 *
## TrainingTimesLastYear           -2.657e-01  1.072e-01  -2.478 0.013208 *
## YearsAtCompany                   1.045e-01  5.267e-02   1.985 0.047190 *
## YearsInCurrentRole              -1.863e-01  7.263e-02  -2.565 0.010311 *
## YearsWithCurrManager            -1.340e-01  6.453e-02  -2.077 0.037831 *
## BusinessTravelTravel_Frequently  1.783e+00  5.358e-01   3.328 0.000876 ***
## BusinessTravelTravel_Rarely      9.299e-01  4.734e-01   1.964 0.049505 *
## DepartmentResearch & Development -6.522e-01  6.120e-01  -1.066 0.286569
## DepartmentSales                  8.280e-01  6.396e-01   1.295 0.195479
## GenderMale                       4.264e-01  2.604e-01   1.637 0.101542
## JobInvolvement2                 -1.915e+00  4.919e-01  -3.893 9.90e-05 ***
## JobInvolvement3                 -2.726e+00  4.798e-01  -5.681 1.34e-08 ***
## JobInvolvement4                 -3.207e+00  6.680e-01  -4.801 1.58e-06 ***
## JobLevel2                       -1.878e+00  3.735e-01  -5.027 4.97e-07 ***
## JobLevel3                       -7.883e-01  5.053e-01  -1.560 0.118741
## JobLevel4                       -2.563e+00  1.034e+00  -2.478 0.013217 *
## JobLevel5                        8.200e-02  9.146e-01   0.090 0.928559
## JobSatisfaction                 -4.888e-01  1.186e-01  -4.122 3.75e-05 ***
## MaritalStatusMarried             1.176e+00  4.046e-01   2.906 0.003664 **
## MaritalStatusSingle              2.124e+00  4.164e-01   5.100 3.40e-07 ***
## OverTimeYes                      2.135e+00  2.720e-01   7.847 4.25e-15 ***
## WorkLifeBalance2                -1.506e+00  4.868e-01  -3.093 0.001982 **
## WorkLifeBalance3                -1.848e+00  4.493e-01  -4.113 3.91e-05 ***
## WorkLifeBalance4                -2.015e+00  6.033e-01  -3.340 0.000837 ***
## YearsSinceLastPromotion1        -3.187e-01  3.386e-01  -0.941 0.346686
## YearsSinceLastPromotion2         2.102e-01  4.039e-01   0.520 0.602780
## YearsSinceLastPromotion3         1.180e+00  7.347e-01   1.606 0.108340
## YearsSinceLastPromotion4         5.809e-01  9.802e-01   0.593 0.553421
## YearsSinceLastPromotion5         3.218e-01  1.228e+00   0.262 0.793245
## YearsSinceLastPromotion6         2.999e+00  8.984e-01   3.339 0.000842 ***
## YearsSinceLastPromotion7         1.496e+00  6.314e-01   2.369 0.017824 *
## YearsSinceLastPromotion8        -1.325e+01  9.818e+02  -0.013 0.989233
## YearsSinceLastPromotion9         3.153e+00  1.120e+00   2.816 0.004860 **
## YearsSinceLastPromotion10        1.812e+00  1.617e+00   1.121 0.262369
## YearsSinceLastPromotion11        2.056e+00  1.278e+00   1.609 0.107705
## YearsSinceLastPromotion12       -1.420e+01  1.485e+03  -0.010 0.992373
## YearsSinceLastPromotion13       -1.558e+01  1.411e+03  -0.011 0.991192
## YearsSinceLastPromotion14        1.697e+00  2.735e+00   0.620 0.535044
## YearsSinceLastPromotion15        5.436e+00  1.231e+00   4.418 9.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 767.67  on 869  degrees of freedom
## Residual deviance: 443.09  on 826  degrees of freedom
## AIC: 531.09
##
## Number of Fisher Scoring iterations: 16
```

```
stepwise_fit
```

```
##
## Call:  glm(formula = Attrition ~ Age + DistanceFromHome + HourlyRate +
##     NumCompaniesWorked + TotalWorkingYears + TrainingTimesLastYear +
##     YearsAtCompany + YearsInCurrentRole + YearsWithCurrManager +
##     BusinessTravel + Department + Gender + JobInvolvement + JobLevel +
##     JobSatisfaction + MaritalStatus + OverTime + WorkLifeBalance +
##     YearsSinceLastPromotion, family = "binomial", data = data)
##
## Coefficients:
##                      (Intercept)                               Age
##                          1.50000                          -0.03381
##                 DistanceFromHome                        HourlyRate
##                          0.06097                           0.01470
##               NumCompaniesWorked                 TotalWorkingYears
##                          0.22929                          -0.08256
##            TrainingTimesLastYear                    YearsAtCompany
##                         -0.26566                           0.10452
##               YearsInCurrentRole              YearsWithCurrManager
##                         -0.18630                          -0.13401
##  BusinessTravelTravel_Frequently       BusinessTravelTravel_Rarely
##                          1.78286                           0.92993
## DepartmentResearch & Development                   DepartmentSales
##                         -0.65223                           0.82797
##                       GenderMale                   JobInvolvement2
##                          0.42636                          -1.91498
##                  JobInvolvement3                   JobInvolvement4
##                         -2.72585                          -3.20727
##                        JobLevel2                         JobLevel3
##                         -1.87759                          -0.78828
##                        JobLevel4                         JobLevel5
##                         -2.56327                           0.08200
##                  JobSatisfaction              MaritalStatusMarried
##                         -0.48877                           1.17564
##              MaritalStatusSingle                       OverTimeYes
##                          2.12355                           2.13452
##                 WorkLifeBalance2                  WorkLifeBalance3
##                         -1.50556                          -1.84767
##                 WorkLifeBalance4          YearsSinceLastPromotion1
##                         -2.01505                          -0.31868
##         YearsSinceLastPromotion2          YearsSinceLastPromotion3
##                          0.21020                           1.17971
##         YearsSinceLastPromotion4          YearsSinceLastPromotion5
##                          0.58092                           0.32179
##         YearsSinceLastPromotion6          YearsSinceLastPromotion7
##                          2.99929                           1.49595
```

```
##        YearsSinceLastPromotion8        YearsSinceLastPromotion9
##                       -13.24921                         3.15270
##       YearsSinceLastPromotion10       YearsSinceLastPromotion11
##                         1.81215                         2.05640
##       YearsSinceLastPromotion12       YearsSinceLastPromotion13
##                       -14.19661                       -15.58036
##       YearsSinceLastPromotion14       YearsSinceLastPromotion15
##                         1.69659                         5.43586
##
## Degrees of Freedom: 869 Total (i.e. Null);  826 Residual
## Null Deviance:       767.7
## Residual Deviance: 443.1     AIC: 531.1
```

```
#choose variables: OverTime, YearsSinceLastPromotion, JobInvolvement, JobLevel
final_model <- glm(Attrition ~ OverTime + YearsSinceLastPromotion + JobInvolvement + JobLevel, data = da
summary(final_model)
```

```
##
## Call:
## glm(formula = Attrition ~ OverTime + YearsSinceLastPromotion +
##     JobInvolvement + JobLevel, family = "binomial", data = data)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.3236     0.4010   0.807 0.419583
## OverTimeYes                 1.6737     0.2172   7.705 1.31e-14 ***
## YearsSinceLastPromotion1   -0.6352     0.2828  -2.246 0.024707 *
## YearsSinceLastPromotion2   -0.1959     0.3338  -0.587 0.557252
## YearsSinceLastPromotion3    0.3042     0.5734   0.531 0.595694
## YearsSinceLastPromotion4   -0.6829     0.7889  -0.866 0.386701
## YearsSinceLastPromotion5   -1.4109     1.0473  -1.347 0.177944
## YearsSinceLastPromotion6    0.7368     0.6053   1.217 0.223504
## YearsSinceLastPromotion7    0.4928     0.4757   1.036 0.300230
## YearsSinceLastPromotion8  -14.6954   635.5961  -0.023 0.981554
## YearsSinceLastPromotion9    1.1161     0.8741   1.277 0.201617
## YearsSinceLastPromotion10   0.6220     1.2132   0.513 0.608150
## YearsSinceLastPromotion11   0.8841     0.8642   1.023 0.306279
## YearsSinceLastPromotion12 -13.9264  1052.4651  -0.013 0.989443
## YearsSinceLastPromotion13 -15.0032   957.6984  -0.016 0.987501
## YearsSinceLastPromotion14   0.7927     1.3515   0.587 0.557532
## YearsSinceLastPromotion15   2.2151     0.8274   2.677 0.007428 **
## JobInvolvement2            -1.4706     0.4009  -3.668 0.000244 ***
## JobInvolvement3            -2.0301     0.3845  -5.280 1.29e-07 ***
## JobInvolvement4            -2.6456     0.5568  -4.752 2.02e-06 ***
## JobLevel2                  -1.5685     0.2665  -5.886 3.95e-09 ***
## JobLevel3                  -0.9899     0.3388  -2.922 0.003480 **
## JobLevel4                  -2.7596     0.7426  -3.716 0.000202 ***
## JobLevel5                  -1.3899     0.6175  -2.251 0.024384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 767.67  on 869  degrees of freedom
```

```
## Residual deviance: 599.39  on 846  degrees of freedom
## AIC: 647.39
##
## Number of Fisher Scoring iterations: 15
```

#logistic regression model (`glm`) using the binomial family was developed to predict the probability of 'Attrition', utilizing various explanatory variables. #Several predictors have been identified as statistically significant ($p < 0.05$), suggesting they meaningfully contribute to the model in this dataset's context. Statistically significant coefficients were found for 'DistanceFromHome', 'HourlyRate', 'NumCompaniesWorked', 'TrainingTimesLastYear', 'YearsInCurrentRole', 'YearsWithCurrManager', 'BusinessTravel', 'JobInvolvement', 'JobLevel2', 'JobSatisfaction', 'MaritalStatus', 'OverTime', 'WorkLifeBalance', and 'YearsSinceLastPromotion'. These factors are predictive of attrition when controlling for other variables. Further research on these variables is recommended.

#Specifically, 'DistanceFromHome', 'NumCompaniesWorked', and 'OverTimeYes' exhibit positive coefficients, indicating that higher values of these predictors are associated with increased log odds of attrition. Conversely, 'JobSatisfaction' has a negative coefficient, suggesting that higher job satisfaction correlates with lower log odds of attrition. Similarly, higher levels of JobInvolvement (levels 2, 3, and 4) are associated with lower log odds of attrition compared to the baseline level. If focusing on retention, further study of this variable is recommended.

#The model's overall fit is reflected in the AIC value of 551.77. Generally, lower AIC values indicate a better-fitting model, suggesting that this model fits the data better than a model with no predictors.

#The stepwise logistic regression identifies several predictors as significant for the likelihood of attrition. 'DistanceFromHome', 'NumCompaniesWorked', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsInCurrentRole', and 'YearsWithCurrManager' show a significant relationship with attrition. Higher values of 'DistanceFromHome' and 'NumCompaniesWorked', frequent business travel ('BusinessTravelTravel_Frequently'), and 'OverTimeYes' are linked to increasing the odds of attrition. Marital status plays a role, with 'MaritalStatusSingle' increasing attrition odds compared to the baseline. Gender is also significant, with 'GenderMale' showing a relationship with attrition. Various levels of job involvement ('JobInvolvement2', 'JobInvolvement3', 'JobInvolvement4') and work-life balance ('WorkLifeBalance2', 'WorkLifeBalance3', 'WorkLifeBalance4'), along with years since the last promotion at certain levels ('YearsSinceLastPromotion6', 'YearsSinceLastPromotion7', 'YearsSinceLastPromotion9', 'YearsSinceLastPromotion15'), are identified as significant predictors, all influencing the likelihood of an employee leaving the organization.

#Moreover, higher 'TrainingTimesLastYear' and greater job satisfaction ('JobSatisfaction') lower the odds of attrition.

```r
set.seed(123)
splitIndex <- createDataPartition(data$Attrition, p = 0.8, list = FALSE)
train_data <- data[splitIndex, ]
test_data <- data[-splitIndex, ]

#ran glm, Knn & NB without correcting for imbalance, none were predictive at required level. added code

# Calculate the number of 'Yes' and 'No' instances in the training data
yes_count <- nrow(train_data[train_data$Attrition == "Yes", ])
no_count <- nrow(train_data[train_data$Attrition == "No", ])

# Determine the desired number of 'Yes' instances after oversampling
oversampled_yes_count <- yes_count * 5

# Calculate the desired total size after oversampling
desired_size <- no_count + oversampled_yes_count
```

```r
# if ..Apply Oversampling on the Training Set
if (desired_size > nrow(train_data)) {
    train_data_balanced <- ovun.sample(Attrition ~ ., data = train_data, method = "over", N = desired_s
    table(train_data_balanced$Attrition)
} else {
    train_data_balanced <- train_data
}
```

```
##
##  No Yes
## 584 560
```

```r
# Inspect the first few rows of the balanced dataset
head(train_data_balanced)
```

```
##   ID Age Attrition    BusinessTravel DailyRate             Department
## 1  2  40        No     Travel_Rarely      1308 Research & Development
## 2  4  32        No     Travel_Rarely       801                  Sales
## 3  5  24        No Travel_Frequently       567 Research & Development
## 4  6  27        No Travel_Frequently       294 Research & Development
## 5  7  41        No     Travel_Rarely      1283 Research & Development
## 6  8  37        No     Travel_Rarely       309                  Sales
##   DistanceFromHome Education   EducationField EmployeeCount EmployeeNumber
## 1               14         3          Medical             1           1128
## 2                1         4        Marketing             1           2016
## 3                2         1 Technical Degree             1           1646
## 4               10         2    Life Sciences             1            733
## 5                5         5          Medical             1           1448
## 6               10         4    Life Sciences             1           1105
##   EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 1                       3   Male         44              2        5
## 2                       3 Female         48              3        3
## 3                       1 Female         32              3        1
## 4                       4   Male         32              3        3
## 5                       2   Male         90              4        1
## 6                       4 Female         88              2        2
##                  JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 1      Research Director               3        Single         19626
## 2        Sales Executive               4       Married         10422
## 3     Research Scientist               4        Single          3760
## 4 Manufacturing Director               1      Divorced          8793
## 5     Research Scientist               3       Married          2127
## 6        Sales Executive               4      Divorced          6694
##   MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 1       17544                  1      Y       No                14
## 2       24032                  1      Y       No                19
## 3       17218                  1      Y      Yes                13
## 4        4809                  1      Y       No                21
## 5        5561                  2      Y      Yes                12
## 6       24223                  2      Y      Yes                14
##   PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
## 1                 3                        1            80                0
## 2                 3                        3            80                2
```

```
## 3                 3                        3            80                0
## 4                 4                        3            80                2
## 5                 3                        1            80                0
## 6                 3                        3            80                3
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                21                     2               4             20
## 2                14                     3               3             14
## 3                 6                     2               3              6
## 4                 9                     4               2              9
## 5                 7                     5               2              4
## 6                 8                     5               3              1
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  7                       4                    9
## 2                 10                       5                    7
## 3                  3                       1                    3
## 4                  7                       1                    7
## 5                  2                       0                    3
## 6                  0                       0                    0
```

```r
# Convert Attrition to a factor if it's not already
data$Attrition <- as.factor(data$Attrition)

# Count the number of 'Yes' and 'No' in the Attrition column
attrition_counts_train <- table(train_data$Attrition)
# Output the counts
print(attrition_counts_train)
```

```
##
##  No Yes
## 584 112
```

```r
# Count the number of 'Yes' and 'No' in the Attrition column after oversample
attrition_counts_balance <- table(train_data_balanced$Attrition)

# Output the counts
print(attrition_counts_balance)
```

```
##
##  No Yes
## 584 560
```

#LM model using stepwise selected variables

"'{r{}} # Build the logistic regression model using stepwise selected variables #names(train_data) - go through and change from train_data to train_data_balanced for train but not predict names(train_data_balanced)

final_model <- glm(Attrition ~ DistanceFromHome + NumCompaniesWorked + BusinessTravel + OverTime, data = train_data_balanced, family = "binomial")

summary(final_model)

Predict and Evaluate on the test data

predictions <- predict(final_model, newdata = test_data, type = "response") predicted_classes <- ifelse(predictions > 0.5, "Yes", "No") predicted_classes <- factor(predicted_classes, levels = c("No", "Yes"))

Evaluate the model

conf_matrix <- confusionMatrix(predicted_classes, test_data$Attrition) conf_matrix

#CHOOSE THIS MODEL #Sensitivity : 0.6781
#Specificity : 0.6429

Save the logistic regression model to a file

saveRDS(final_model, "best_model.rds") # Load the saved logistic regression model #loaded_model <- readRDS("best_model.rds")

```
#before correct for imbalance
Sensitivity : 1.00000
Specificity : 0.03571
#After correct for imbalance
Sensitivity : 0.6781
Specificity : 0.6429
```

```
```r
# KNN Model
set.seed(123)
train_control <- trainControl(method = "cv", number = 10)
knn_model <- train(Attrition ~ OverTime + YearsSinceLastPromotion + JobInvolvement + JobLevel, data = t
```

```
# Model Evaluation
predictions_knn <- predict(knn_model, newdata = test_data)
conf_matrix_knn <- confusionMatrix(predictions_knn, test_data$Attrition)
conf_matrix_knn
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  106  14
##        Yes  40  14
##
##                Accuracy : 0.6897
##                  95% CI : (0.6152, 0.7575)
##     No Information Rate : 0.8391
##     P-Value [Acc > NIR] : 0.9999997
##
```
```

```
##                  Kappa : 0.1644
##
##  Mcnemar's Test P-Value : 0.0006688
##
##            Sensitivity : 0.7260
##            Specificity : 0.5000
##         Pos Pred Value : 0.8833
##         Neg Pred Value : 0.2593
##             Prevalence : 0.8391
##         Detection Rate : 0.6092
##   Detection Prevalence : 0.6897
##      Balanced Accuracy : 0.6130
##
##       'Positive' Class : No
##
```

#BEFORE CORRECTING imbalance Sensitivity: 0.9726 Specificity: 0.3571

#AFTER CORRECT IMBALANCE WITH OVERSAMPLING Sensitivity: 0.7808 Specificity: 0.5000

```r
# Naive Bayes Model
set.seed(123)
nb_model <- train(Attrition ~ OverTime + YearsSinceLastPromotion + JobInvolvement + JobLevel, data = tra

# Model Evaluation
predictions_nb <- predict(nb_model, newdata = test_data)
conf_matrix_nb <- confusionMatrix(predictions_nb, test_data$Attrition)
conf_matrix_nb
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No   26   2
##        Yes 120  26
##
##               Accuracy : 0.2989
##                 95% CI : (0.2319, 0.3728)
##    No Information Rate : 0.8391
##    P-Value [Acc > NIR] : 1
##
##                  Kappa : 0.0395
##
##  Mcnemar's Test P-Value : <2e-16
##
##            Sensitivity : 0.1781
##            Specificity : 0.9286
##         Pos Pred Value : 0.9286
##         Neg Pred Value : 0.1781
##             Prevalence : 0.8391
##         Detection Rate : 0.1494
##   Detection Prevalence : 0.1609
##      Balanced Accuracy : 0.5533
##
```

```
##        'Positive' Class : No
##
```

#BEFORE CORRECTING imbalance Sensitivity: 1.0000 (the model did not predict any 'Yes' cases) Specificity: 0.0000 (the model failed to correctly identify any of the 'Yes' cases)

#correct for imbalance Sensitivity : 0.1781
Specificity : 0.9286

```r
#Load the best model (lm)
loaded_model <- readRDS("best_model.rds")
```

```r
# Data Preprocessing test data -load and preprocess
# Load the saved logistic regression model
loaded_model <- readRDS("best_model.rds")

# Data Preprocessing test data -load and preprocess
# Load test data
comp_data <- read.csv("CaseStudy2CompSet No Attrition.csv")

# List of categorical variables
categorical_vars <- c("BusinessTravel", "Department", "EducationField", "Gender", "JobInvolvement", "Jol

# Apply factor levels to existing variables in test_data
comp_data[categorical_vars] <- lapply(comp_data[categorical_vars], factor)

str(comp_data[categorical_vars])
```

```
## 'data.frame':    300 obs. of  12 variables:
##  $ BusinessTravel         : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 3 3 3 2 1 3 3
##  $ Department             : Factor w/ 3 levels "Human Resources",..: 2 1 2 2 3 2 3 2 2 2 ...
##  $ EducationField         : Factor w/ 6 levels "Human Resources",..: 2 1 4 2 2 4 2 4 4 2 ...
##  $ Gender                 : Factor w/ 2 levels "Female","Male": 2 2 2 1 2 1 2 2 1 2 ...
##  $ JobInvolvement         : Factor w/ 4 levels "1","2","3","4": 4 3 3 3 2 2 3 3 2 3 ...
##  $ JobLevel               : Factor w/ 5 levels "1","2","3","4",..: 2 1 1 4 2 3 1 2 2 3 ...
##  $ JobRole                : Factor w/ 9 levels "Healthcare Representative",..: 3 2 3 4 8 1 9 5 1 1 .
##  $ JobSatisfaction        : Factor w/ 4 levels "1","2","3","4": 3 3 3 4 3 1 4 1 1 3 ...
##  $ MaritalStatus          : Factor w/ 3 levels "Divorced","Married",..: 2 2 1 3 1 3 1 2 3 2 ...
##  $ OverTime               : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ WorkLifeBalance        : Factor w/ 4 levels "1","2","3","4": 2 3 2 3 3 2 2 3 3 3 ...
##  $ YearsSinceLastPromotion: Factor w/ 16 levels "0","1","2","3",..: 7 2 1 2 1 13 1 8 2 1 ...
```

```r
# Final structure and summary check for test data
str(comp_data)
```

```
## 'data.frame':    300 obs. of  35 variables:
##  $ ID                      : int  1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 ...
##  $ Age                     : int  35 33 26 55 29 51 52 39 31 31 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 3 3 3 2 1 3
##  $ DailyRate               : int  750 147 1330 1311 1246 1456 585 1387 1062 534 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 2 1 2 2 3 2 3 2 2 2 ...
##  $ DistanceFromHome        : int  28 2 21 2 19 1 29 10 24 20 ...
##  $ Education               : int  3 3 3 3 3 4 4 5 3 3 ...
```

```
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 1 4 2 2 4 2 4 4 2 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : int  1596 1207 1107 505 1497 145 2019 1618 1252 587 ...
##  $ EnvironmentSatisfaction : int  2 2 1 3 3 1 1 2 3 1 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 2 2 2 1 2 1 2 2 1 2 ...
##  $ HourlyRate              : int  46 99 37 97 77 30 40 76 96 66 ...
##  $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 4 3 3 3 2 2 3 3 2 3 ...
##  $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 1 1 4 2 3 1 2 2 3 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 3 2 3 4 8 1 9 5 1 1
##  $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 3 3 3 4 3 1 4 1 1 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 2 2 1 3 1 3 1 2 3 2 ...
##  $ MonthlyIncome           : int  3407 3600 2377 16659 8620 7484 3482 5377 6812 9824 ...
##  $ MonthlyRate             : int  25348 8429 19373 23258 23757 25796 19788 3835 17198 22908 ...
##  $ NumCompaniesWorked      : int  1 1 1 2 1 3 2 2 1 3 ...
##  $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ PercentSalaryHike       : int  17 13 20 13 14 20 15 13 19 12 ...
##  $ PerformanceRating       : int  3 3 4 3 3 4 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  4 4 3 3 3 3 2 4 2 1 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  2 1 1 0 2 0 2 3 0 0 ...
##  $ TotalWorkingYears       : int  10 5 1 30 10 23 16 10 10 12 ...
##  $ TrainingTimesLastYear   : int  3 2 0 2 3 1 3 3 2 2 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 2 3 2 3 3 2 2 3 3 3 ...
##  $ YearsAtCompany          : int  10 5 1 5 10 13 9 7 10 1 ...
##  $ YearsInCurrentRole      : int  9 4 1 4 7 12 8 7 9 0 ...
##  $ YearsSinceLastPromotion : Factor w/ 16 levels "0","1","2","3",..: 7 2 1 2 1 13 1 8 2 1 ...
##  $ YearsWithCurrManager    : int  8 4 0 2 4 8 0 7 8 0 ...
```

```r
# Predict using the loaded model
comp_predictions <- predict(loaded_model, newdata = comp_data, type = "response")

# Convert probabilities to class labels (assuming a threshold of 0.5)
comp_predictions_class <- ifelse(comp_predictions > 0.5, "Yes", "No")

# Create a data frame to save the predictions
result_df <- data.frame(ID = comp_data$ID, Attrition = comp_predictions_class)

# Write the predictions to a CSV file
write.csv(result_df, "Case2PredictionsMirzaAttrition.csv", row.names = FALSE)
```

#Develop a regression model to predict missing monthly incomes in another dataset. The model should achieve a Root Mean Square Error (RMSE) of less than $3000 for both training and validation sets. Validation Requirement for Salary(RMSE < $4000)

```r
# Read Training Data
train_data <- read.csv("CaseStudy2-data.csv")

# Data Preprocessing
# Convert categorical variables in the training data to factors
categorical_vars <- c("BusinessTravel", "Department", "EducationField", "Gender", "JobInvolvement",
                      "JobLevel", "JobRole", "JobSatisfaction", "MaritalStatus", "OverTime",
                      "WorkLifeBalance", "YearsSinceLastPromotion")
train_data[categorical_vars] <- lapply(train_data[categorical_vars], factor)
```

```r
# Log-transform the 'MonthlyIncome' variable
train_data$MonthlyIncome <- log(train_data$MonthlyIncome)

# Check for and remove categorical variables with only one level
single_level_vars <- sapply(train_data, function(x) length(unique(x)) == 1)
train_data <- train_data[, !single_level_vars]

# Split the data into training (70%) and validation (30%) sets
set.seed(123) # For reproducibility
train_index <- createDataPartition(train_data$MonthlyIncome, p = 0.7, list = FALSE)
train_set <- train_data[train_index, ]
validation_set <- train_data[-train_index, ]

# Building a regression model on the training set
model <- train(MonthlyIncome ~ ., data = train_set, method = "lm", trControl = trainControl(method = "c

# Evaluate model performance on the validation set
validation_predictions <- predict(model, newdata = validation_set)

# Reverse the log transformation for predictions and actual values
predicted_values_validation <- exp(validation_predictions)
actual_values_validation <- exp(validation_set$MonthlyIncome)

# Calculate RMSE on the original scale
RMSE_train_original_scale <- sqrt(mean((exp(train_set$MonthlyIncome) - exp(predict(model, newdata = trai
RMSE_validation_original_scale <- sqrt(mean((actual_values_validation - predicted_values_validation)^2)

# Print RMSE on training and validation sets on the original scale
cat("RMSE on training data (original scale):", RMSE_train_original_scale, "\n")
```

```
## RMSE on training data (original scale): 1029.428
```

```r
cat("RMSE on validation data (original scale):", RMSE_validation_original_scale, "\n")
```

```
## RMSE on validation data (original scale): 1118.905
```

```r
# Read the Dataset with Missing Monthly Incomes
comp_salary_data <- read.csv("CaseStudy2CompSet No Salary.csv")

# Convert categorical variables in this dataset to factors
comp_salary_data[categorical_vars] <- lapply(comp_salary_data[categorical_vars], factor)

# Apply the model to the competition data
comp_salary_predictions <- predict(model, newdata = comp_salary_data, type = "raw")
comp_salary_predictions <- exp(comp_salary_predictions)  # Reverse log transformation

# Create a data frame to save the predictions
result_df <- data.frame(ID = comp_salary_data$EmployeeNumber, PredictedSalary = comp_salary_predictions

# Write the predictions to a CSV file
write.csv(result_df, "Case2PredictionsMirzaSalary.csv", row.names = FALSE)
```