Statistical Learning Project

Marlon Helbing, Nemanja Ilic, Daniele Virzì

2024-06-13

1. Introduction

This project is a collaborative effort between three students, **Marlon Helbing**, **Nemanja Ilic**, **Daniele Virzì**. It is an academic project that will be graded based on the quality and depth of our analysis. The project aims to apply the concepts and techniques from the **Statistical Learning Course** to a real-world dataset. Our project is based on the **HR Analytics Case Study** and we will be using the **R** programming language to perform the analysis. As previously stated, the scope of this project is to assess the knowledge we have gained from the course. Because of this, in our project work, we were only allowed to utilize the models and techniques covered in the lectures; we were not permitted to use any `Tidyverse` R-packages, like `ggplot` or `ggplot2`.

1.1 Dataset

- **HR Analytics Case Study**: This set of datasets, sourced from *Kaggle*, contains information about employees working in a company. These data are collected to understand why the employees are leaving the company and to predict the employees who are likely to leave the company. There are several datasets available for this case study but for our purposes we have chosen and merged just two of them, `general_data` and `employee_survey_data`. The final dataset contains 4410 observations and 27 variables. The variables are as follows:
  - **`Age`**: Age of the employee.
  - **`Attrition`**: Whether the employee has left the company or not.
  - **`BusinessTravel`**: Frequency of travel for the employee.
  - **`Department`**: Department of the employee.
  - **`DistanceFromHome`**: Distance of the employee's residence from the company.
  - **`Education`**: Education level of the employee.
  - **`EducationField`**: Field of education of the employee.
  - **`EmployeeCount`**: Employee count.
  - **`EmployeeID`**: Employee ID.
  - **`Gender`**: The gender of the employee.
  - **`JobLevel`**: Job level of the employee.
  - **`JobRole`**: Job role of the employee.
  - **`MaritalStatus`**: Marital status of the employee.
  - **`MonthlyIncome`**: Monthly income of the employee.
  - **`NumCompaniesWorked`**: Number of companies the employee has worked for.
  - **`Over18`**: Whether the employee is over 18 years old or not.
  - **`PercentSalaryHike`**: Percentage increase in salary.
  - **`StandardHours`**: Standard hours of work.
  - **`StockOptionLevel`**: Stock option level of the employee.
  - **`TotalWorkingYears`**: Total years the employee has worked.

  - `TrainingTimesLastYear`: Number of times the employee was trained last year.
  - `YearsAtCompany`: Number of years the employee has worked at the company.
  - `YearsSinceLastPromotion`: Number of years since the last promotion.
  - `YearsWithCurrManager`: Number of years the employee has worked with the current manager.
  - `EnvironmentSatisfaction`: Environment satisfaction level of the employee.
  - `JobSatisfaction`: Job satisfaction level of the employee.
  - `WorkLifeBalance`: Work-life balance level of the employee.

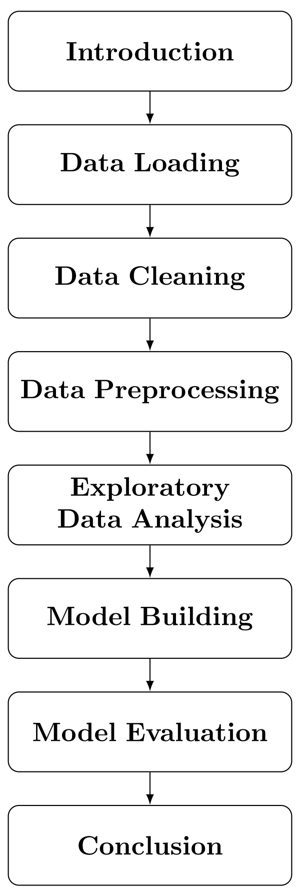
1.2 Project goals

The main objectives of this project are:

- **Regression Model**: To predict `YearsAtCompany` based on the available features, in order to understand the factors that influence the number of years an employee stays in the company. In this way, the company can take actions to retain employees for a longer period of time.

- **Classification Model**: To predict `Attrition` based on the available features, in order to understand the factors that influence the attrition of employees in the company. In this way, the company can take actions to reduce the attrition rate.

1.3 Methodology



2. Data Loading

We start by loading the necessary libraries and the data into the R environment. The libraries that we will be using in this project are:

```r
library(MASS) # For step, glm, lda, qda
library(car) # For vif
library(corrplot) # For plotting correlation matrix
library(pROC) # For ROC curve
```

The data is loaded from the `general_data.csv` and `employee_survey_data.csv` files. We then merge the two datasets based on the `EmployeeID` variable.

```r
general_data <- read.csv("./data/general_data.csv")
employee_survey_data <- read.csv("./data/employee_survey_data.csv")
data <- merge(general_data, employee_survey_data, by = "EmployeeID")
```

3. Data Cleaning

3.1 Handling missing values

We check for missing values in the dataset and find that there are 111 misssing values.

```r
missing_values <- sum(is.na(data))
missing_values
```

```
## [1] 111
```

```r
data <- na.omit(data)
```

3.2 Handling duplicate rows

We check for duplicate rows in the dataset and find that there are no duplicate rows.

```r
duplicates <- sum(duplicated(data))
duplicates
```

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

3.3 Removing unnecessary columns

We remove the `EmployeeID`, because it is a unique identifier and does not provide any useful information for the analysis. We also remove the `Over18`, `StandardHours`, and `EmployeeCount` columns because they have the same value for all employees and so the variance is zero.

```r
data <- data[, !(names(data) %in% c("EmployeeID", "Over18", "StandardHours", "EmployeeCount"))]
```

4. Data Preprocessing

4.1 Encoding categorical variables

We encode the categorical variables as factors in order to use them in the regression and classification models.

```r
data$Attrition <- factor(data$Attrition)
data$Gender <- factor(data$Gender)
data$BusinessTravel <- factor(data$BusinessTravel)
data$JobRole <- factor(data$JobRole)
data$Department <- factor(data$Department)
data$EducationField <- factor(data$EducationField)
data$MaritalStatus <- factor(data$MaritalStatus)
data$StockOptionLevel <- factor(data$StockOptionLevel)
data$Education <- factor(data$Education)
data$JobLevel <- factor(data$JobLevel)
data$EnvironmentSatisfaction <- factor(data$EnvironmentSatisfaction)
data$JobSatisfaction <- factor(data$JobSatisfaction)
data$WorkLifeBalance <- factor(data$WorkLifeBalance)
```

4.2 Log transformation

We perform log transformation on the `MonthlyIncome` variable to make it more normally distributed.

```r
data$MonthlyIncome <- log(data$MonthlyIncome)
```

4.3 Check the structure of the dataset

We check the structure of the dataset to ensure that the data preprocessing steps have been applied correctly.

```r
str(data)
```

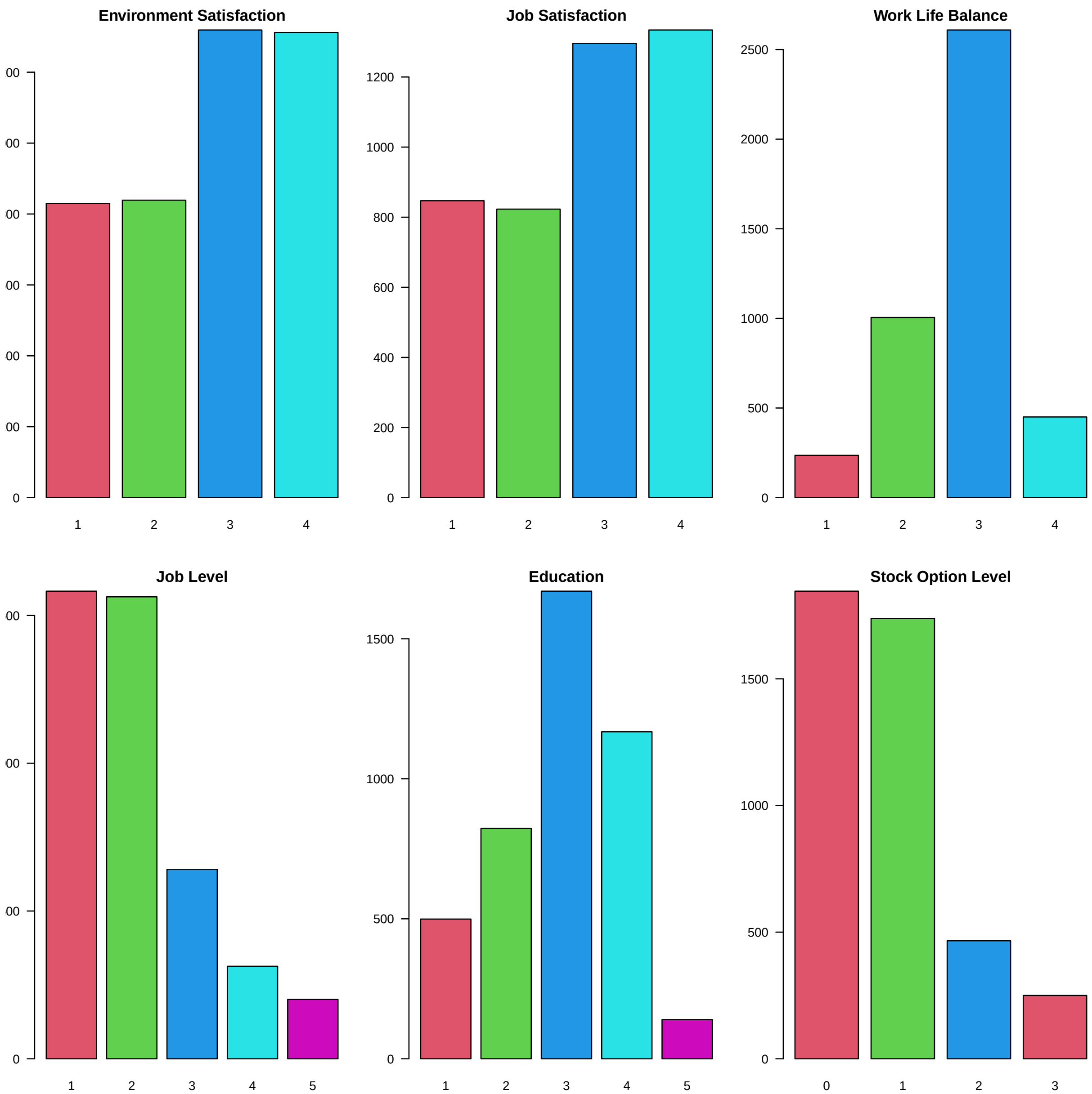
```
## 'data.frame':    4300 obs. of  23 variables:
##  $ Age                    : int  51 31 32 38 32 46 28 29 31 25 ...
##  $ Attrition              : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 1 1 ...
##  $ BusinessTravel         : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 2 2 1 3 3 3 3 3
##  $ Department             : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...
##  $ DistanceFromHome       : int  6 10 17 2 10 8 11 18 1 7 ...
##  $ Education              : Factor w/ 5 levels "1","2","3","4",..: 2 1 4 5 1 3 2 3 3 4 ...
##  $ EducationField         : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...
##  $ Gender                 : Factor w/ 2 levels "Female","Male": 1 1 2 2 2 1 2 2 2 1 ...
##  $ JobLevel               : Factor w/ 5 levels "1","2","3","4",..: 1 1 4 3 1 4 2 2 3 4 ...
##  $ JobRole                : Factor w/ 9 levels "Healthcare Representative",..: 1 7 8 2 8 6 8 8 3 3 .
##  $ MaritalStatus          : Factor w/ 3 levels "Divorced","Married",..: 2 3 2 2 3 2 3 2 2 1 ...
##  $ MonthlyIncome          : num  11.8 10.6 12.2 11.3 10.1 ...
##  $ NumCompaniesWorked     : int  1 0 1 3 4 3 2 2 0 1 ...
##  $ PercentSalaryHike      : int  11 23 15 11 12 13 20 22 21 13 ...
##  $ StockOptionLevel       : Factor w/ 4 levels "0","1","2","3": 1 2 4 4 3 1 2 4 1 2 ...
##  $ TotalWorkingYears      : int  1 6 5 13 9 28 5 10 10 6 ...
##  $ TrainingTimesLastYear  : int  6 3 2 5 2 5 2 2 2 2 ...
##  $ YearsAtCompany         : int  1 5 5 8 6 7 0 0 9 6 ...
```

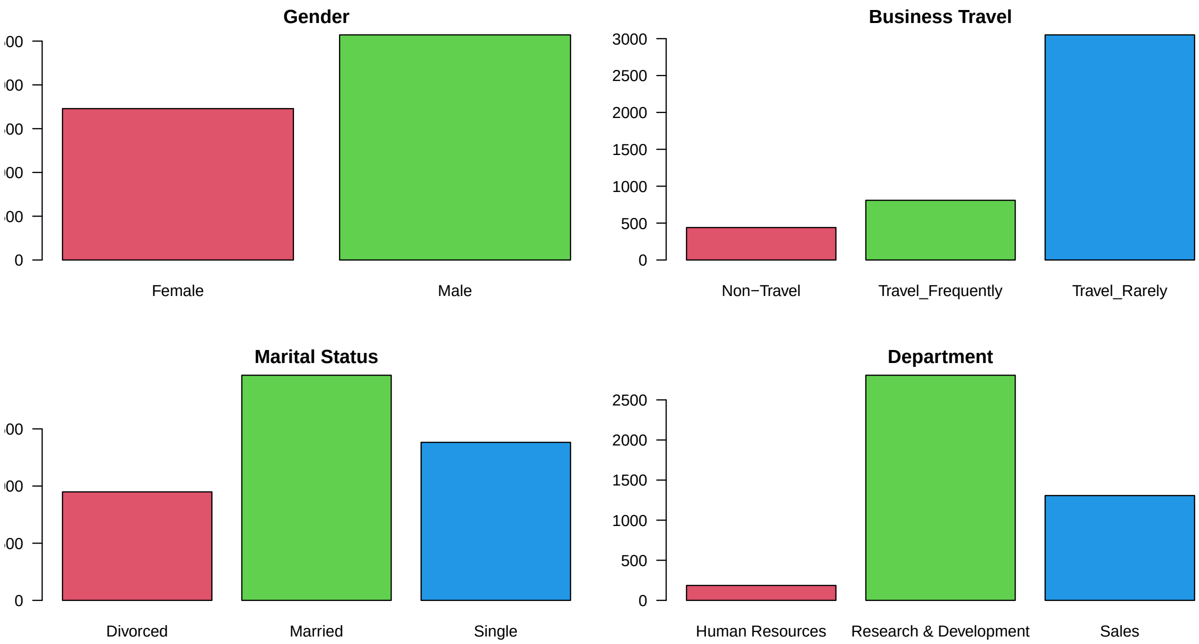
```
##  $ YearsSinceLastPromotion: int  0 1 0 7 0 7 0 0 7 1 ...
##  $ YearsWithCurrManager   : int  0 4 3 5 4 7 0 0 8 5 ...
##  $ EnvironmentSatisfaction: Factor w/ 4 levels "1","2","3","4": 3 3 2 4 4 3 1 1 2 2 ...
##  $ JobSatisfaction        : Factor w/ 4 levels "1","2","3","4": 4 2 2 4 1 2 3 2 4 1 ...
##  $ WorkLifeBalance        : Factor w/ 4 levels "1","2","3","4": 2 4 1 3 3 2 1 3 3 3 ...
```

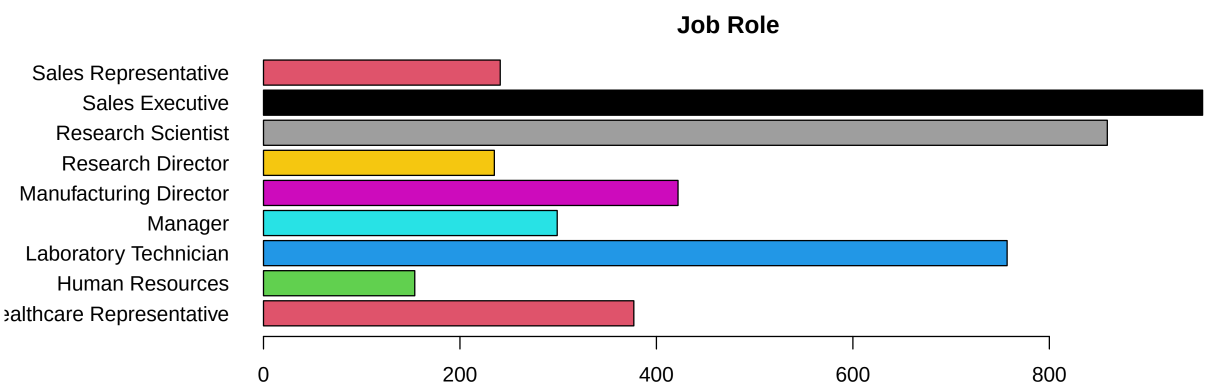
5. Exploratory Data Analysis

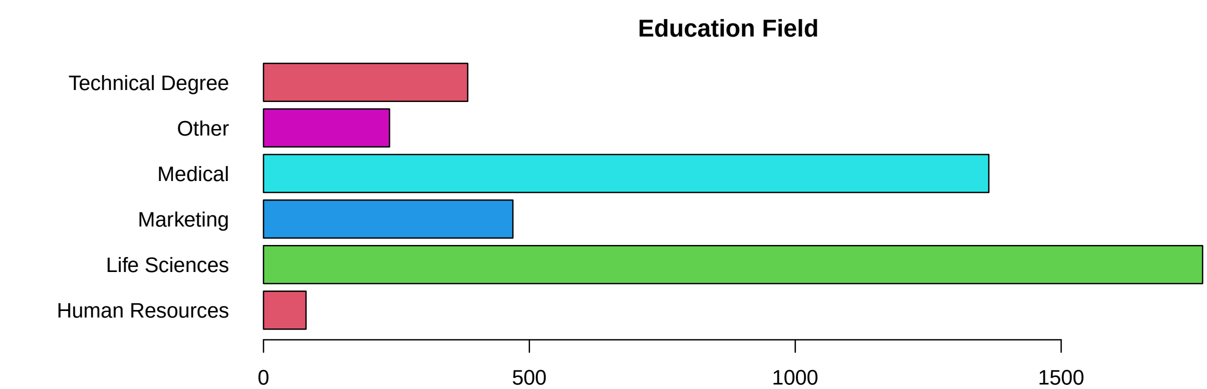
5.1 Categorical variables

We plot the distribution of the categorical variables in the dataset.




Gender
0
Female
Male
Business Travel
3000
2500
2000
1500
1000
500
0
Non−Travel
Travel_Frequently
Travel_Rarely
Marital Status
0
Divorced
Married
Single
Department
2500
2000
1500
1000
500
0
Human Resources
Research & Development
Sales


Job Role
Sales Representative
Sales Executive
Research Scientist
Research Director
Manufacturing Director
Manager
Laboratory Technician
Human Resources
0
200
400
600
800


Education Field
Technical Degree
Other
Medical
Marketing
Life Sciences
Human Resources
0
500
1000
1500

We notice that `JobSatisfaction` and `EnvironmentSatisfaction` are extremely similar. To check this we compute the chi squared statistic between these two.

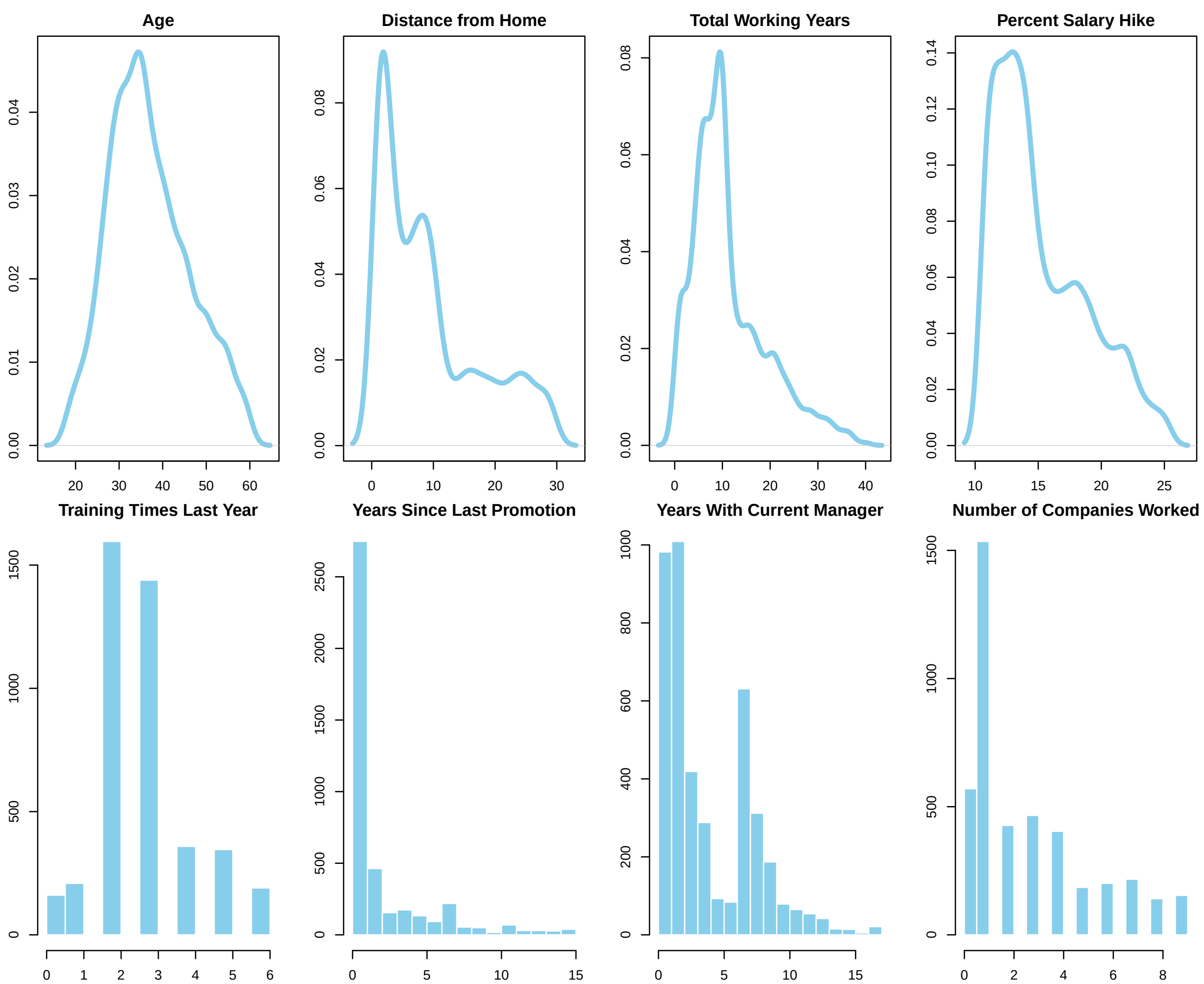
```
contingency_table <- table(data$EnvironmentSatisfaction, data$JobSatisfaction)
chi_squared <- chisq.test(contingency_table)
chi_squared #
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 15.327, df = 9, p-value = 0.08235
```

```
data$EnvironmentSatisfaction <- NULL
```

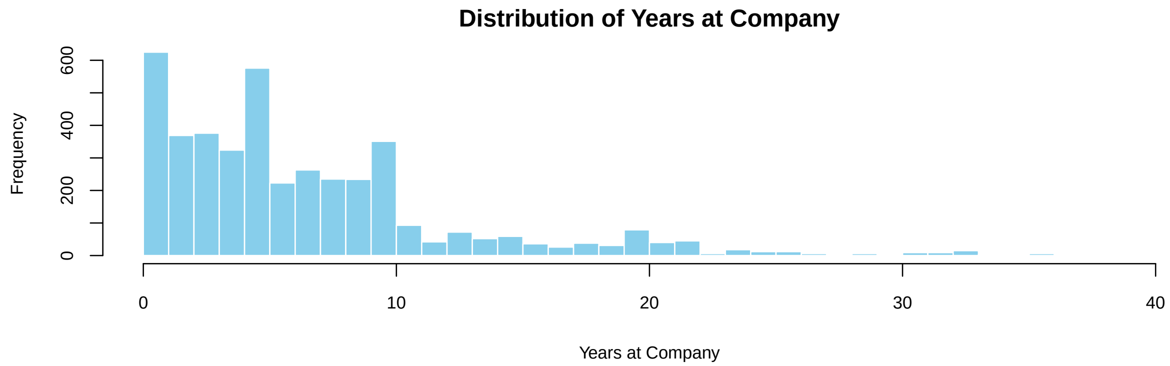
5.2 Numerical variables

We plot the distribution of the numerical variables in the dataset.

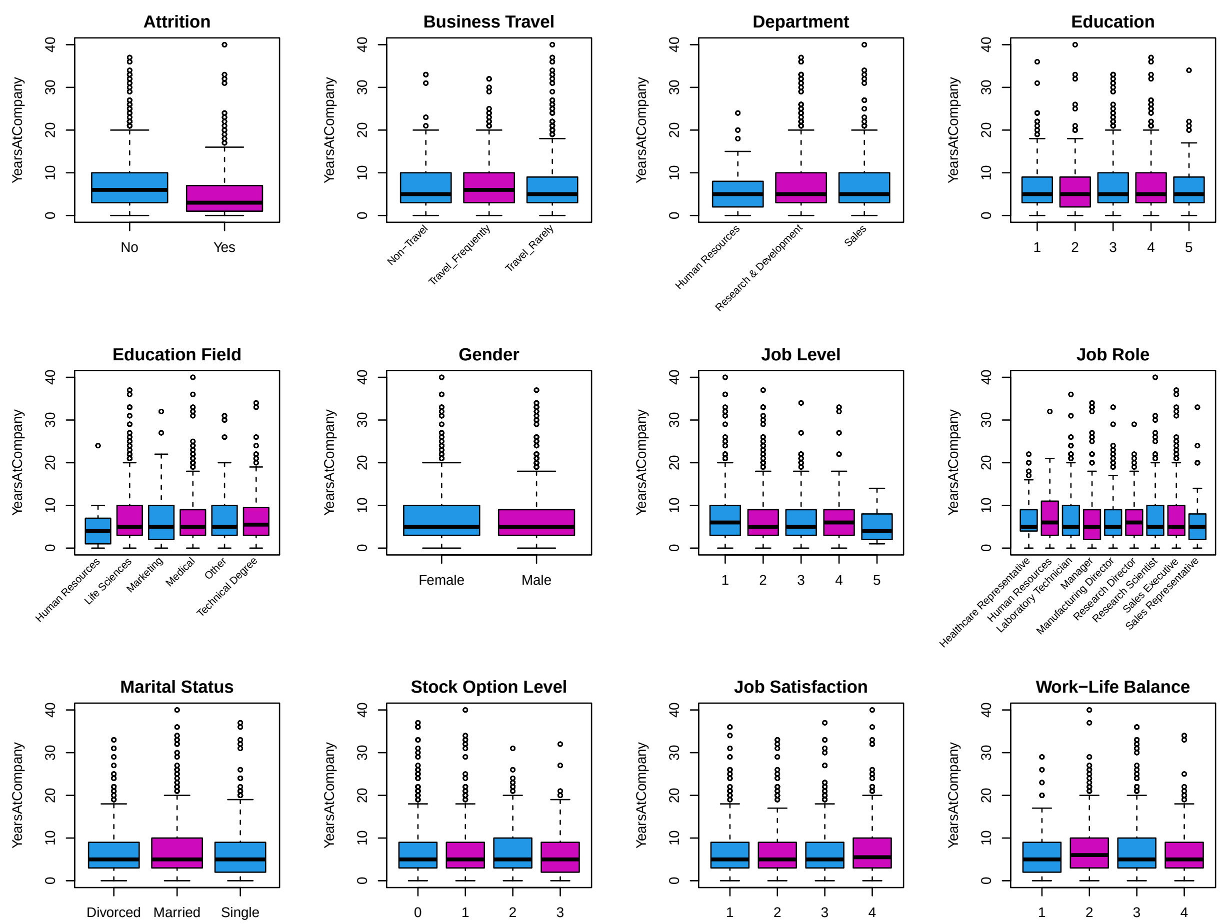


5.3 Years At Company Analysis

We plot the distribution of the `YearsAtCompany` variable, our target variable for the regression model.

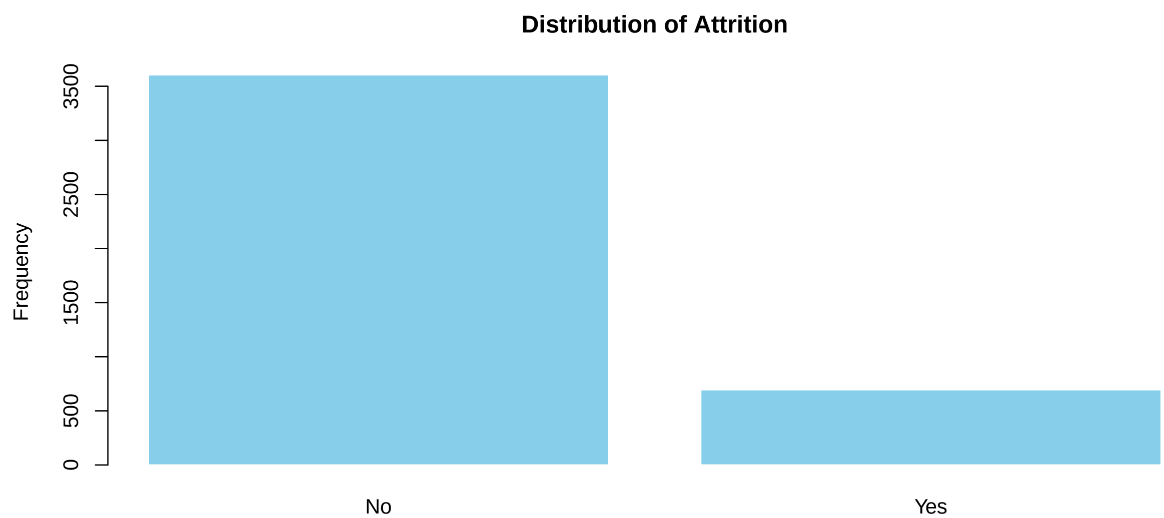


Then we plot the boxplot of the `YearsAtCompany` variable against all the categorical variables in the dataset.

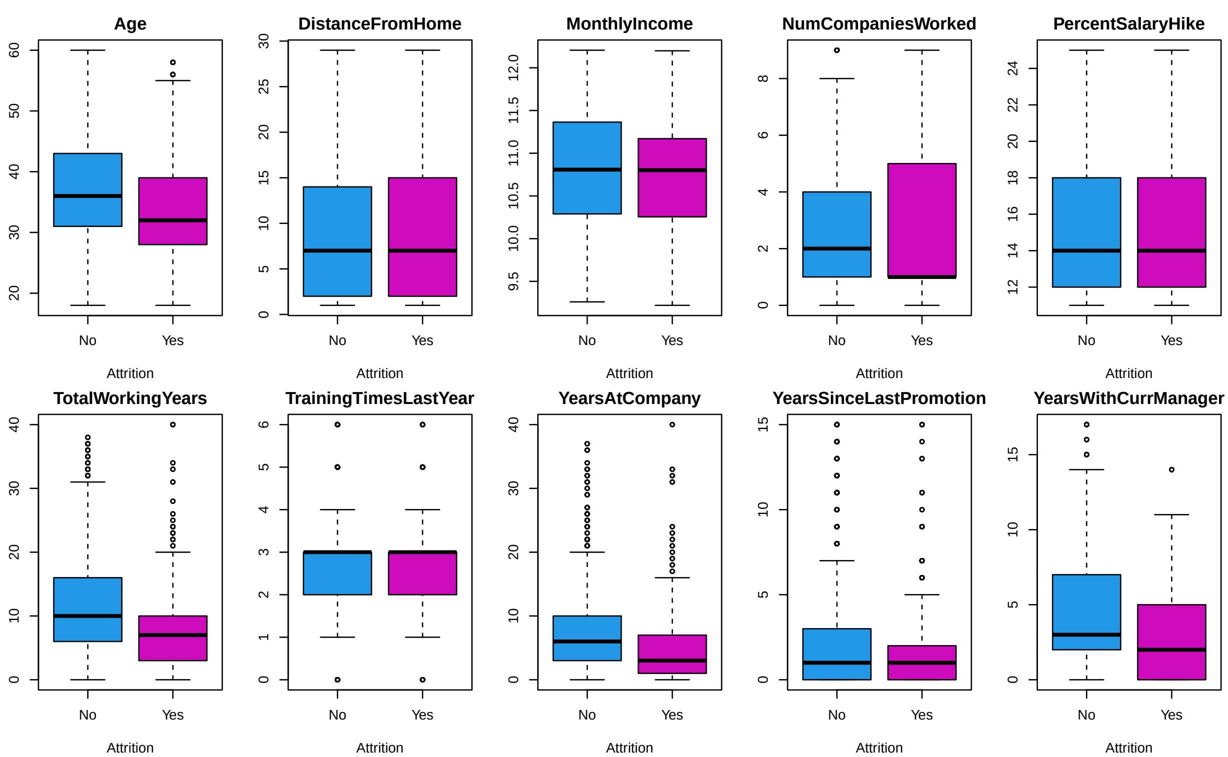


5.4 Attrition Analysis

We plot the distribution of the `Attrition` variable, our target variable for the classification model.

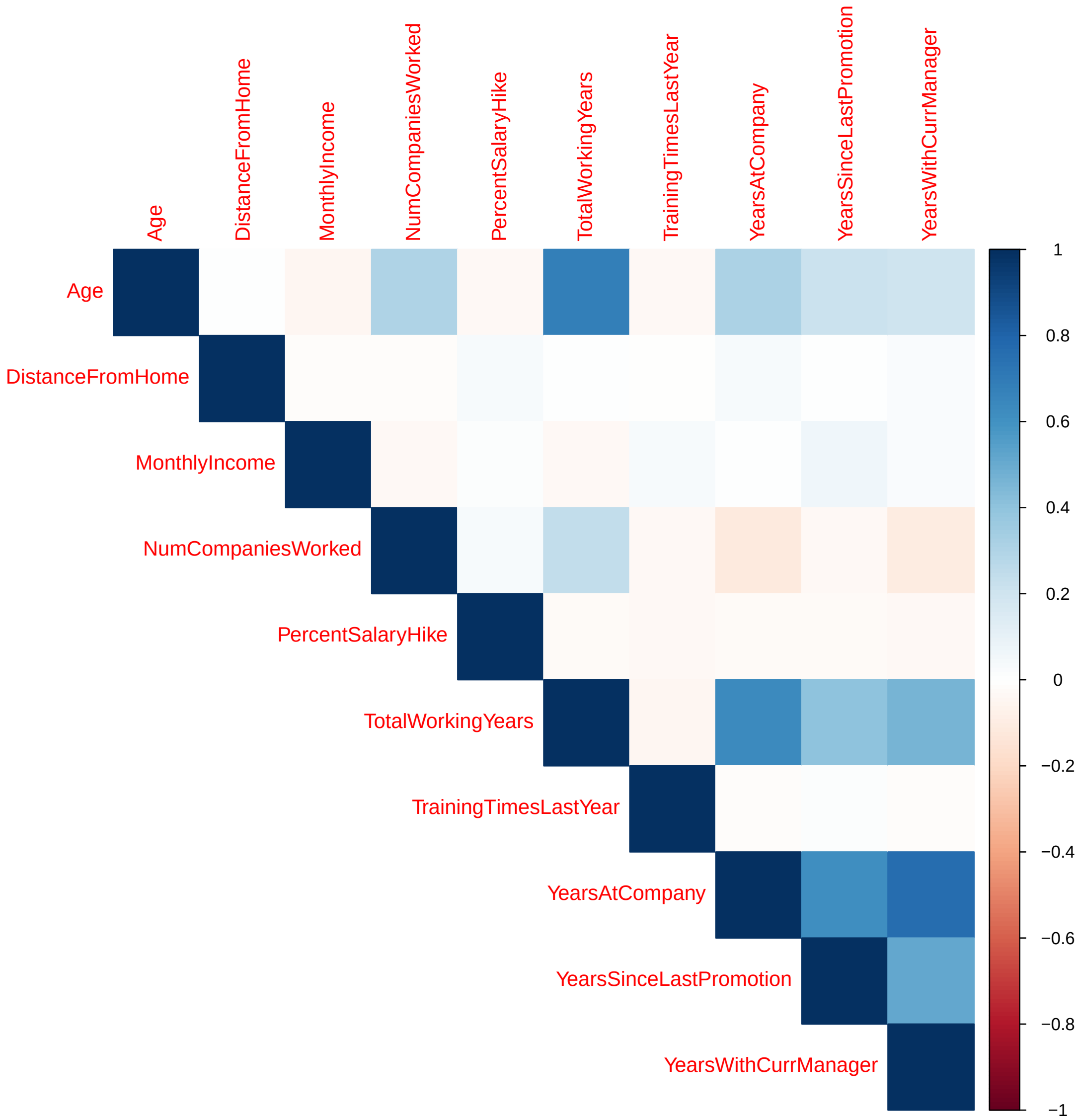


Then we plot the boxplot of the `Attrition` variable against all the numerical variables in the dataset.



5.5 Correlation Analysis

We calculate the correlation matrix of the numerical variables in the dataset.



We notice that `YearsAtCompany` is highly correlated with `YearsWithCurrManager`. We will start from this variable to build the regression model.

6. Regression Model