PROJECT REPORT

Data-Driven Insights into Employee Attrition, Salaries, and Performance Ratings

Group 10

Gopi Boppana

Radhika Gedela

Krishna Vamsi Gottipati

Akhila Siri Godana

Elisha Yallamati

Department of Health Informatics, IUI

INFO B – 518 Applied Statistical Methods for Biomedical Informatics

Dr. Zeyana Hamid, PhD

December 14, 2023

Introduction

In the ever-evolving landscape of organizational management, the nexus of employee attrition, salaries, and performance ratings emerges as a critical determinant for the triumph and endurance of any enterprise (Zhenjing et al., 2022, pp. 1,2). It is imperative for organizations to comprehend the intricate interplay of factors influencing these aspects to foster a conducive and thriving work environment. Job satisfaction emerges as a cornerstone for employee retention and enhanced performance, with individuals who derive fulfillment and purpose from their roles being more inclined to stay within an organization (Zhenjing et al., 2022, pp. 1,2). Furthermore, the delicate equilibrium of work-life balance, encapsulating flexible schedules and supportive policies, assumes a pivotal role in ensuring employee well-being and long-term commitment (Alblihed & Alzghaibi, 2022, p. 2). Equally foundational are competitive and fair compensation structures, as employees perceiving their salaries as commensurate with skills and contributions are more likely to remain (Balushi et al., 2022, p. 2).

One of the pivotal contributors is the cumulative years of experience, where the intuitive assumption is that individuals with extensive professional backgrounds bring a wealth of knowledge and skills, thereby often warranting higher compensation (De Jong et al., 2019, p. 1). Beyond experience, educational qualifications serve as a key benchmark for evaluating an employee's expertise and qualifications, establishing a critical link between academic achievements and compensation levels (Zhenjing et al., 2022, pp. 1,2). The hierarchical position within an organization, demarcated by job levels, assumes a central role in salary determination, with executives and managers traditionally commanding higher salaries compared to their entry or mid-level counterparts. Additionally, the department or functional area within which an employee operates emerges as a significant influencer, as certain departments inherently require specialized skills or expertise, impacting compensation structures.

This study delves into these multifaceted dynamics, aiming to unravel the complexities of salary determination by examining the interplay of experience, education, job level, and departmental functions. Through this exploration, we seek to contribute insights that enable organizations to make informed decisions in shaping compensation structures that align with industry norms, organizational goals, and the unique contributions of their workforce.

Problem Statement

This study aims to investigate the intricate interplay between various workplace factors and employee outcomes within a corporate setting. By employing advanced statistical analyses, including correlation and regression methods, the research seeks to uncover essential insights that can inform human resource management strategies and enhance overall organizational performance. The focus will be on understanding the relationships between employee satisfaction, work-related variables, and critical outcomes such as performance ratings, salaries, and attrition.

Dataset

The dataset, sourced from Kaggle's Employee Attrition dataset by Prashant Patel, consists of 1470 observations and 28 variables.
Link for dataset: https://www.kaggle.com/datasets/patelprashant/employee-attrition

Types and Summary of the Variables:

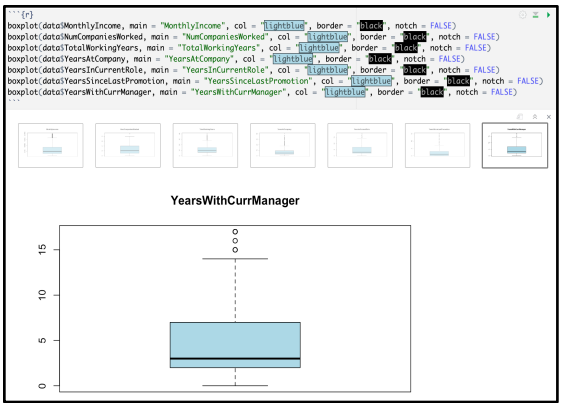
Categorical Variables		Numerical Variables	
Ordinal	**Nominal**	**Continuous**	**Discrete**
Education	Attrition	Age	DistanceFromHome
EnvironmentSatisfaction	BusinessTravel	MonthlyIncome	NumCompaniesWorked
JobInvolvement	Department		PercentSalaryHike
JobSatisfaction	EducationField		TotalWorkingYears
PerformanceRating	Gender		TrainingTimesLastYear
RelationshipSatisfaction	JobRole		YearsAtCompany
WorkLifeBalance	MaritalStatus		YearsInCurrentRole
	Over18		YearsSinceLastPromotion
	OverTime		YearsWithCurrManager

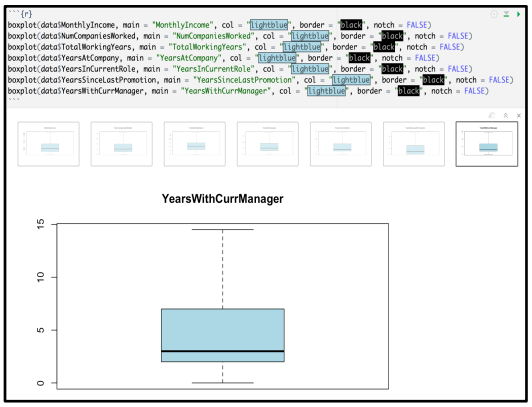
The numerical variables exhibit various characteristics. The average age of individuals in the dataset is 36.92, with a standard deviation of 9.14, ranging from 18 to 60. MonthlyIncome has a mean of 6502.93, a standard deviation of 4707.96, and ranges from 1009 to 19999. PercentSalaryHike shows a mean of 15.21, a standard deviation of 3.66, and a range from 11 to 25. PerformanceRating has a mean of 3.15, a standard deviation of 0.36, and ranges between 3 and 4. DistanceFromHome has a mean of 9.19, a standard deviation of 8.11, and a range from 1 to 29. Education exhibits a mean of 2.91, a standard deviation of 1.02, and ranges from 1 to 5.

Several work-related variables provide insights into the workforce. JobInvolvement has a mean of 2.73, a standard deviation of 0.71, and ranges from 1 to 4. JobLevel exhibits a mean of 2.06, a standard deviation of 1.11, and ranges from 1 to 5. JobSatisfaction has a mean of 2.73, a standard deviation of 1.10, and ranges from 1 to 4. NumCompaniesWorked shows a mean of 2.69, a standard deviation of 2.50, and ranges from 0 to 9. TotalWorkingYears has a mean of 11.28, a standard deviation of 7.78, and ranges from 0 to 40. TrainingTimesLastYear exhibits a mean of 2.80, a standard deviation of 1.29, and ranges from 0 to 6. WorkLifeBalance has a mean of 2.76, a standard deviation of 0.71, and ranges from 1 to 4. YearsAtCompany shows a mean of 7.01, a standard deviation of 6.13, and ranges from 0 to 40. YearsInCurrentRole has a mean of 4.23, a standard deviation of 3.62, and ranges from 0 to 18. YearsSinceLastPromotion exhibits a mean of 2.19, a standard deviation of 3.22, and ranges from 0 to 15. Finally, YearsWithCurrManager has a mean of 4.12, a standard deviation of 3.57, and ranges from 0 to 17.

Data Cleaning and Exploration:

The CSV file containing employee attrition data was initially imported into R using the "read.csv" command. To streamline the dataset according to our objectives, we removed unnecessary columns. The first few rows of the modified dataset were inspected using "head(data)," and a comprehensive summary was obtained with "summary(data)." Further exploration involved assessing the dataset's dimensions with "dim(data)," revealing the number of attributes and instances. The structure of the dataset was examined using "str(data)," providing valuable insights into variable types and attributes. A meticulous check for null values indicated an absence of missing entries. Duplicate values were scrutinized and determined to be absent. Moving on to outliers, we utilized the Interquartile Range (IQR) method, substantiated by box plots, to identify and address outliers in the dataset. Winsorization, employing the clipping method, was then applied to mitigate the impact of outliers, aiming to improve the distribution for enhanced model performance.

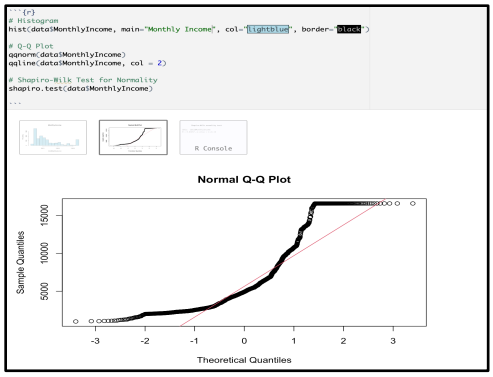




Exploratory Data Analysis

Conducting a thorough analysis of summary statistics, we explored key metrics such as mean, median, and other relevant measures. This comprehensive examination provided detailed insights into the central tendencies and distribution characteristics of the dataset.

Checking for Normality: To assess the normality of the data, a combination of a histogram, Q-Q plot, and the Shapiro-Wilk test was employed. This approach allowed us to evaluate the distribution characteristics and determine if the data adheres to a normal distribution. Notably, the analysis revealed that the monthly income variable deviates from a normal distribution.



Log Transformation: Given the observed non-normal distribution in certain dataset columns, our strategy involved implementing a log transformation on these variables. The goal of this transformation is to bring the variables closer to a normal distribution, facilitating more robust statistical analyses.

Histograms Before and After Log Transformation: The accompanying histograms illustrate the distribution of selected columns in the dataset both before and after the log transformation. This visual representation provides clarity on the impact of the log transformation in achieving a more normalized distribution for the analyzed variables.

Correlation Test*:* A comprehensive correlation test was conducted to evaluate the strength of association among the variables in our dataset. This analytical approach provided valuable insights into the relationships and dependencies existing between different factors, aiding in the identification of potential patterns and trends.

Fisher's Exact Test: In addition to the correlation test, we performed Fisher's Exact Test to rigorously assess the association between variables in the dataset. Fisher's Exact Test is particularly useful for contingency tables and scenarios with small sample sizes. This statistical test enhances our understanding of the

interdependencies and associations among categorical variables, contributing to a more nuanced interpretation of the dataset.

Statistical Analysis

Multiple linear regression
What factors collectively influence employee salaries in the organization, and how effectively can we estimate salaries using variables like job role, job level, performance ratings and department?

Null Hypothesis (H0): There is no significant relationship between employee salaries and variables such as job role, job level, performance ratings and department.
Alternative Hypothesis (H1): There is a significant relationship between employee salaries and variables such as job role, job level, performance ratings and department, allowing us to predict salaries effectively.

Rationale:
The purpose of employing multiple linear regression analysis in this study is to discern the significant factors that contribute to determining employee salaries within the company. This statistical method allows us to assess the relationship between the dependent variable, i.e., employee salaries, and multiple independent variables, including years of experience, education level, job level, and department.

Upon conducting the initial multiple linear regression, the statistical output provides p-values for each predictor variable. The significance of these p-values lies in their ability to indicate whether a particular variable significantly contributes to explaining the variability in monthly income. A p-value below the conventional threshold of 0.05 suggests statistical significance.

Considering our findings, where certain variables exhibit p-values below 0.05, we recognize these as potentially influential factors in determining employee salaries. Consequently, we plan to refine our model by selectively including these specific variables. This iterative process aims to enhance the model's accuracy and predictive power. Subsequent iterations of the model will not only assess the impact on the R-squared value but also consider the adjusted R-squared value, a metric that accounts for the number of predictors and aids in preventing overfitting.

This approach aligns with best practices in regression analysis, ensuring that our model is not only statistically robust but also practically insightful for informing decisions related to salary structures and human resource management within the organization.

Following the selection of variables demonstrating statistical significance, we proceeded to re-execute the model. The obtained F-statistic of 2045, coupled with a p-value below 0.005, provides compelling evidence to reject the null hypothesis. However, to fortify this rejection, we conducted an analysis of variance (ANOVA) and subsequent pairwise tests to scrutinize the relationships among variable groups.

ANOVA results revealed a p-value below 0.05, substantiating our rejection of the null hypothesis that posits no significant difference among the variable groups. Encouraged by these findings, we pursued pairwise t-tests to discern specific differences between the groups, aligning with the F-statistic outcomes. The consistency between ANOVA and pairwise tests enhance our confidence in the observed associations, affirming the robustness of our findings.

In addressing nuanced associations within the job role variable and their impact on monthly income, we implemented encoding for distinct job role categories. The subsequent model rerun yielded an adjusted R-squared value of 0.8596, indicating a robust fit. Introducing interactions with predictors contributed to a

slight improvement in adjusted R-squared, but further scrutiny revealed limited impact, suggesting negligible enhancement in explanatory power.
To ensure model reliability, we scrutinized assumptions, finding residuals adhering to normal density within the central range but deviating in the tails. Large residuals in the tails raised concerns, and observations indicated variable predictive accuracy for higher values. These insights call for refined model considerations to enhance accuracy.
In conclusion, our analysis rejects the null hypothesis, establishing a significant relationship between employee salaries and variables like job level and department. The final model, with an R-squared of 0.86, elucidates a substantial portion of income variability. Job level and specific roles (Human resources, Laboratory technician, Research scientist, Sales representative) exhibit significant associations, with higher job levels correlating with increased income. However, residual plots suggest potential inaccuracies, prompting consideration of higher-order models for nuanced understanding.

Multiple Logistic Regression
How do demographic and job-related variables (education, job level, percentage salary hike, job role) collectively affect employee attrition, and can we create an effective prediction model for it?

Null Hypothesis (H0): There is no significant relationship between demographics, job-related factors (job level, percentage salary hike, job role) and employee attrition within the organization.
Alternative Hypothesis (H1): There is a significant relationship between demographics, job-related factors (job level, percentage salary hike, job role) and employee attrition within the organization.

Rationale
The rationale behind employing logistic regression for the second research question was to examine how demographic and job-related factors collectively influence employee attrition, with a focus on variables such as age, gender, job level, department, and job satisfaction. Logistic regression, with attrition as the response variable, allowed us to assess the significance of these predictors. Notably, before conducting the regression analysis, we identified a class imbalance issue within the response variable. To address this, oversampling was employed, ensuring a more equitable representation of attrition classes in the binary classification context. The analysis revealed that most predictors exhibit a significant association with the response variable. Considering these findings, the null hypothesis is rejected, confirming a substantial relationship between job-related factors (job level, percentage salary hike, job role) and employee attrition within the organization.

In summary, the logistic regression analysis supports the rejection of the null hypothesis, indicating a significant relationship between demographic and job-related factors (job level, department, and job satisfaction) and employee attrition within the company. The study underscores the importance of variables such as job level, department, and job satisfaction in predicting and understanding employee attrition patterns.

Ordinary Least Squares regression
How do variables such as environment satisfaction, job satisfaction, relationship satisfaction, and work-life balance influence the chance of earning higher performance ratings, and can we develop an accurate prediction model for performance ratings?

Null Hypotheses (H0): There is no significant association between environment satisfaction, job satisfaction, relationship satisfaction, work-life balance, and performance ratings.
Alternative Hypotheses (H1): There is a significant association between predictors (environment satisfaction, job satisfaction, relationship satisfaction, work-life balance) and response variables (performance ratings).

Rationale

The rationale behind employing Ordinary Least Squares (OLS) regression for the third research question was to examine the impact of factors such as predictors (environment satisfaction, job satisfaction, relationship satisfaction, and work-life balance) on chance of obtaining better performance ratings. The initial analysis, involving specific predictors, did not yield any significant associations with the response variable (performance ratings). In response, we made the decision to broaden the model by including almost all predictors, yet none exhibited a significant relationship with performance ratings.

To explore potential associations, we turned our attention to the correlation heatmap, identifying a robust positive association ($r = 0.77$) between percent salary hike and performance rating. Despite this correlation, when we specifically examined the model for percent salary hike and performance rating, we encountered challenges due to quasi-complete separation. This phenomenon emerged from the observation that, for instances where PercentSalaryHike is less than or equal to 20, all corresponding PerformanceRating values were consistently equal to 4. This condition, referred to as complete separation, introduces complexities in logistic regression analysis and can result in infinite maximum likelihood estimates of coefficients. Consequently, these challenges prompted a reconsideration of the model's structure and raised questions about its suitability for predicting performance ratings based on the selected predictors.

In conclusion, the Ordinary Least Squares (OLS) regression analysis failed to reveal a significant association between various predictors and response variable (performance ratings). Due to challenges associated with quasi-complete separation, it is recommended to explore alternative modeling approaches, such as logistic regression with regularization techniques, for more robust results in predicting performance ratings based on the selected factors.

Limitations

1. Causation and Correlation Distinction:
    - Statistical methods excel in identifying correlations between variables but lack the capacity to establish causation.
    - Correlations, as seen between job role, job level do not inherently indicate a causal link resulting in increased/decreased monthly income.

2. Impact of Confounding Variables:
    - Confounding variables, correlated with both independent and dependent variables, can introduce distortions into statistical results.

3. Influence of Biased Data:
    - The accuracy of statistical analyses hinges on the quality of the data used.
    - Biased or incomplete data may yield results that do not faithfully represent the true relationship between variables.

4. Predictive Limitations at Extremes:
    - Developing forecasts for independent variables that are marginally bigger or lower than the minimum might be risky.
    - In many datasets, few observations around the lowest or maximum values of the explanatory variable may jeopardize the predictability of forecasts in these extreme ranges.

References

Alblihed, M. A., & Alzghaibi, H. A. (2022). The impact of job stress, role ambiguity and work–life imbalance on turnover intention during COVID-19: A case study of frontline health workers in Saudi Arabia. *International Journal of Environmental Research and Public Health*, *19*(20), 13132. https://doi.org/10.3390/ijerph192013132

Balushi, A. K. A., Thumiki, V. R. R., Nawaz, N., Jurčić, A., & Gajenderan, V. (2022). Role of organizational commitment in career growth and turnover intention in public sector of Oman. *PLOS ONE*, *17*(5), e0265535. https://doi.org/10.1371/journal.pone.0265535

De Jong, N., Wisse, B., Heesink, J., & Van Der Zee, K. I. (2019). Personality traits and career role enactment: Career role preferences as a mediator. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01720

Patel, P. (2018). *Employee Attrition* [Data set]. Kaggle. https://www.kaggle.com/datasets/patelprashant/employee- attrition/

Vu, J., & Harrington, D. (2020). Introductory statistics for the life and biomedical sciences. In Diez, D., Çetinkaya-Rundel, M., & Barr, C. (2019). OpenIntro Statistics. (1st ed., pp. 330-372). https://www.openintro.org/book/biostat/

Zhenjing, G., Chupradit, S., Ku, K. Y., Nassani, A. A., & Haffar, M. (2022). Impact of employees' workplace environment on employees' performance: A multi-mediation model. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.890400

Appendix

Data Importing

CSV file containing employee attrition data was uploaded to R using the "read.csv" code.

```r
data <- read.csv("HR-Employee-Attrition.csv")
```

```r
columns_to_remove <- c("DailyRate","EmployeeNumber", "HourlyRate", "MonthlyRate", "StandardHours", "StockOptionLevel", "EmployeeCount")

data <- data[, !(names(data) %in% columns_to_remove)]
```

Data Description

Examined the first few rows of the modified dataset with the command "head(data)" and obtained a summary of the dataset using "summary(data).”

```r
head(data)
summary(data)
```

	Age <int>	Attrition <chr>	BusinessTravel <chr>	Department <chr>	DistanceFromHome <int>	Education <int>	EducationField <chr>
1	41	Yes	Travel_Rarely	Sales	1	2	Life Sciences
2	49	No	Travel_Frequently	Research & Development	8	1	Life Sciences
3	37	Yes	Travel_Rarely	Research & Development	2	2	Other
4	33	No	Travel_Frequently	Research & Development	3	4	Life Sciences
5	27	No	Travel_Rarely	Research & Development	2	1	Medical
6	32	No	Travel_Frequently	Research & Development	2	2	Life Sciences

```
:character
 Mean   :2.729                      Mean   : 6503   Mean   :2.693
 3rd Qu.:4.000                      3rd Qu.: 8379   3rd Qu.:4.000
 Max.   :4.000                      Max.   :19999   Max.   :9.000
 PercentSalaryHike PerformanceRating RelationshipSatisfaction TotalWorkingYears
TrainingTimesLastYear WorkLifeBalance
 Min.   :11.00     Min.   :3.000     Min.   :1.000            Min.   : 0.00     Min.   :0.000
Min.   :1.000
 1st Qu.:12.00     1st Qu.:3.000     1st Qu.:2.000            1st Qu.: 6.00     1st Qu.:2.000
1st Qu.:2.000
 Median :14.00     Median :3.000     Median :3.000            Median :10.00     Median :3.000
Median :3.000
 Mean   :15.21     Mean   :3.154     Mean   :2.712            Mean   :11.28     Mean   :2.799
Mean   :2.761
 3rd Qu.:18.00     3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:15.00     3rd Qu.:3.000
3rd Qu.:3.000
 Max.   :25.00     Max.   :4.000     Max.   :4.000            Max.   :40.00     Max.   :6.000
Max.   :4.000
 YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.000   Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
 1st Qu.: 3.000   1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
 Median : 5.000   Median : 3.000     Median : 1.000          Median : 3.000
 Mean   : 7.008   Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
 3rd Qu.: 9.000   3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :40.000   Max.   :18.000     Max.   :15.000          Max.   :17.000
```

```r
dim(data)
str(data)
```

Explored the dimensions of the dataset using "dim(data)," revealing the number of rows and columns. Additionally, we examined the structure of the dataset with "str(data)," obtaining information about the variable types and their respective attributes.

```
'data.frame':   1470 obs. of  28 variables:
 $ Age                     : int  41 49 37 33 27 32 59 30 38 36 ...
 $ Attrition               : chr  "Yes" "No" "Yes" "No" ...
 $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel_Frequently" ...
 $ Department              : chr  "Sales" "Research & Development" "Research & Development" "Research & Development" ...
 $ DistanceFromHome        : int  1 8 2 3 2 2 3 24 23 27 ...
 $ Education               : int  2 1 2 4 1 2 3 1 3 3 ...
 $ EducationField          : chr  "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...
 $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
 $ Gender                  : chr  "Female" "Male" "Male" "Female" ...
 $ JobInvolvement          : int  3 2 2 3 3 3 4 3 2 3 ...
 $ JobLevel                : int  2 2 1 1 1 1 1 1 3 2 ...
 $ JobRole                 : chr  "Sales Executive" "Research Scientist" "Laboratory Technician" "Research Scientist" ...
 $ JobSatisfaction         : int  4 2 3 3 2 4 1 3 3 3 ...
 $ MaritalStatus           : chr  "Single" "Married" "Single" "Married" ...
 $ MonthlyIncome           : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
 $ NumCompaniesWorked      : int  8 1 6 1 9 0 4 1 0 6 ...
 $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
 $ OverTime                : chr  "Yes" "No" "Yes" "Yes" ...
 $ PercentSalaryHike       : int  11 23 15 11 12 13 20 22 21 13 ...
 $ PerformanceRating       : int  3 4 3 3 3 3 4 4 4 3 ...
 $ RelationshipSatisfaction: int  1 4 2 3 4 3 1 2 2 2 ...
 $ TotalWorkingYears       : int  8 10 7 8 6 8 12 1 10 17 ...
 $ TrainingTimesLastYear   : int  0 3 3 3 3 2 3 2 2 3 ...
 $ WorkLifeBalance         : int  1 3 3 3 3 2 2 3 3 2 ...
 $ YearsAtCompany          : int  6 10 0 8 2 7 1 1 9 7 ...
 $ YearsInCurrentRole      : int  4 7 0 7 2 7 0 0 7 7 ...
 $ YearsSinceLastPromotion : int  0 1 0 3 2 3 0 0 1 7 ...
 $ YearsWithCurrManager    : int  5 7 0 0 2 6 0 0 8 7 ...
```

Data Cleaning

Examined the dataset for null values, search for duplicate values in our dataset and determined that there were nonpresent.

```
```{r}
# Checking for NA values in the entire data frame
if (any(is.na(data))) {
  print("There are NA values in the data frame.")
} else {
  print("There are no NA values in the data frame.")
}

```
```

```
[1] "There are no NA values in the data frame."
```

```
```{r}
any(duplicated(data))
```
```

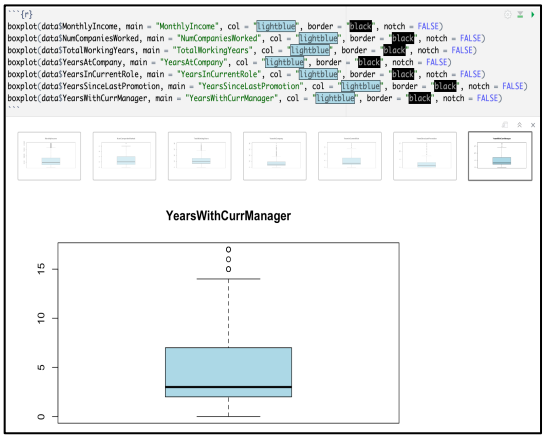
```
[1] FALSE
```

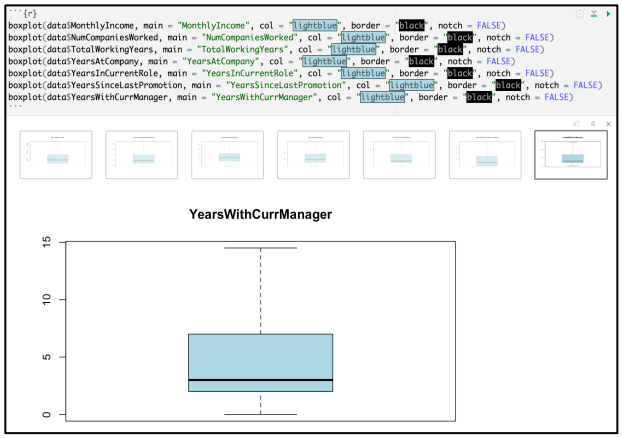
Opted to address outliers in the dataset through the application of the Winsorization technique, employing the clipping method.

```r
# Function to detect outliers using IQR method
detect_outliers <- function(variable) {
  Q1 <- quantile(variable, 0.25)
  Q3 <- quantile(variable, 0.75)
  IQR <- Q3 - Q1

  # Identify potential outliers
  potential_outliers <- variable < (Q1 - 1.5 * IQR) | variable > (Q3 + 1.5 * IQR)

  return(potential_outliers)
}

# Columns to check for outliers
columns_to_check <- c(
  "DistanceFromHome",
  "MonthlyIncome",
  "NumCompaniesWorked",
  "PercentSalaryHike",
  "TotalWorkingYears",
  "YearsAtCompany",
  "YearsInCurrentRole",
  "YearsSinceLastPromotion",
  "YearsWithCurrManager"
)

# Check for outliers in each column
for (col in columns_to_check) {
  variable <- data[[col]]
  outliers <- detect_outliers(variable)
  print(paste("Outliers in", col, ":", any(outliers)))
}
```

```
[1] "Outliers in DistanceFromHome : FALSE"
[1] "Outliers in MonthlyIncome : TRUE"
[1] "Outliers in NumCompaniesWorked : TRUE"
[1] "Outliers in PercentSalaryHike : FALSE"
[1] "Outliers in TotalWorkingYears : TRUE"
[1] "Outliers in YearsAtCompany : TRUE"
[1] "Outliers in YearsInCurrentRole : TRUE"
[1] "Outliers in YearsSinceLastPromotion : TRUE"
[1] "Outliers in YearsWithCurrManager : TRUE"
```

```r
# Function to clip (cap) outliers based on IQR method
clip_outliers <- function(variable) {
  Q1 <- quantile(variable, 0.25)
  Q3 <- quantile(variable, 0.75)
  IQR <- Q3 - Q1

  # Set the clipping threshold
  threshold <- 1.5

  # Clip (cap) values beyond the threshold
  variable[variable < (Q1 - threshold * IQR)] <- (Q1 - threshold * IQR)
  variable[variable > (Q3 + threshold * IQR)] <- (Q3 + threshold * IQR)

  return(variable)
}

# Columns to clip (cap) outliers
columns_to_clip <- c(
  "DistanceFromHome",
  "MonthlyIncome",
  "NumCompaniesWorked",
  "PercentSalaryHike",
  "TotalWorkingYears",
  "YearsAtCompany",
  "YearsInCurrentRole",
  "YearsSinceLastPromotion",
  "YearsWithCurrManager"
)

# Clip (cap) outliers in each column
for (col in columns_to_clip) {
  variable <- data[[col]]
  data[[col]] <- clip_outliers(variable)
}
```





Exploratory Data Analysis

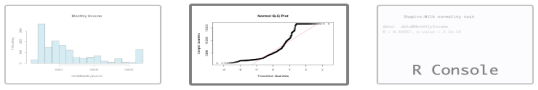
Conducted a comprehensive analysis of summary statistics, delving into various metrics such as mean, median, and other relevant measures.

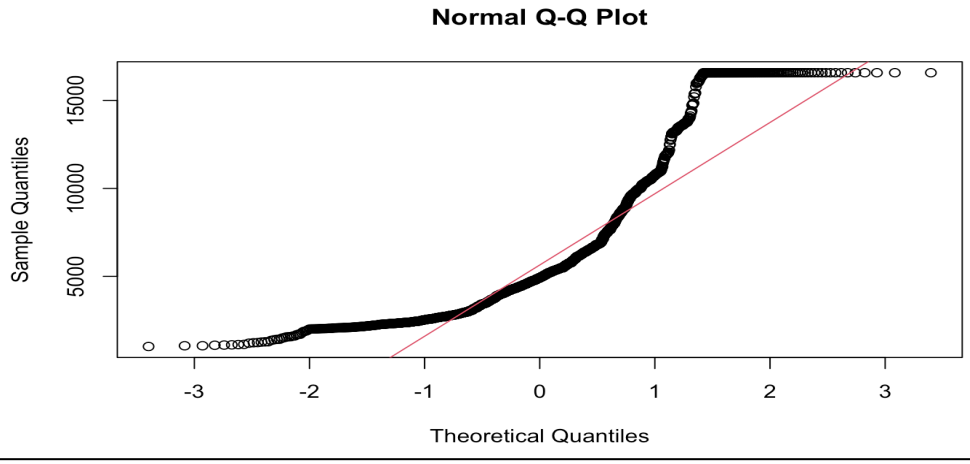
```r
summary(data)
```

```
      Age          Attrition         BusinessTravel      Department        DistanceFromHome   Education
 Min.   :18.00   Length:1470        Length:1470        Length:1470        Min.   : 1.000   Min.   :1.000
 1st Qu.:30.00   Class :character   Class :character   Class :character   1st Qu.: 2.000   1st Qu.:2.000
 Median :36.00   Mode  :character   Mode  :character   Mode  :character   Median : 7.000   Median :3.000
 Mean   :36.92                                                           Mean   : 9.193   Mean   :2.913
 3rd Qu.:43.00                                                           3rd Qu.:14.000   3rd Qu.:4.000
 Max.   :60.00                                                           Max.   :29.000   Max.   :5.000
 EducationField     EnvironmentSatisfaction    Gender          JobInvolvement    JobLevel       JobRole
 Length:1470        Min.   :1.000           Length:1470        Min.   :1.00   Min.   :1.000   Length:1470
 Class :character   1st Qu.:2.000           Class :character   1st Qu.:2.00   1st Qu.:1.000   Class :character
 Mode  :character   Median :3.000           Mode  :character   Median :3.00   Median :2.000   Mode  :character
                    Mean   :2.722                              Mean   :2.73   Mean   :2.064
                    3rd Qu.:4.000                              3rd Qu.:3.00   3rd Qu.:3.000
                    Max.   :4.000                              Max.   :4.00   Max.   :5.000
 JobSatisfaction MaritalStatus      MonthlyIncome   NumCompaniesWorked    Over18            OverTime
 Min.   :1.000   Length:1470        Min.   : 1009   Min.   :0.000      Length:1470        Length:1470
 1st Qu.:2.000   Class :character   1st Qu.: 2911   1st Qu.:1.000      Class :character   Class :character
 Median :3.000   Mode  :character   Median : 4919   Median :2.000      Mode  :character   Mode  :character
 Mean   :2.729                      Mean   : 6362   Mean   :2.676
 3rd Qu.:4.000                      3rd Qu.: 8379   3rd Qu.:4.000
 Max.   :4.000                      Max.   :16581   Max.   :8.500
 PercentSalaryHike PerformanceRating RelationshipSatisfaction TotalWorkingYears TrainingTimesLastYear WorkLifeBalance
 Min.   :11.00     Min.   :3.000     Min.   :1.000            Min.   : 0.0      Min.   :0.000         Min.   :1.000
 1st Qu.:12.00     1st Qu.:3.000     1st Qu.:2.000            1st Qu.: 6.0      1st Qu.:2.000         1st Qu.:2.000
 Median :14.00     Median :3.000     Median :3.000            Median :10.0      Median :3.000         Median :3.000
 Mean   :15.21     Mean   :3.154     Mean   :2.712            Mean   :11.1      Mean   :2.799         Mean   :2.761
 3rd Qu.:18.00     3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:15.0      3rd Qu.:3.000         3rd Qu.:3.000
 Max.   :25.00     Max.   :4.000     Max.   :4.000            Max.   :28.5      Max.   :6.000         Max.   :4.000
 YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.000   Min.   : 0.000     Min.   :0.000           Min.   : 0.000
 1st Qu.: 3.000   1st Qu.: 2.000     1st Qu.:0.000           1st Qu.: 2.000
 Median : 5.000   Median : 3.000     Median :1.000           Median : 3.000
 Mean   : 6.618   Mean   : 4.208     Mean   :1.923           Mean   : 4.107
 3rd Qu.: 9.000   3rd Qu.: 7.000     3rd Qu.:3.000           3rd Qu.: 7.000
 Max.   :18.000   Max.   :14.500     Max.   :7.500           Max.   :14.500
```

Checking for Normality

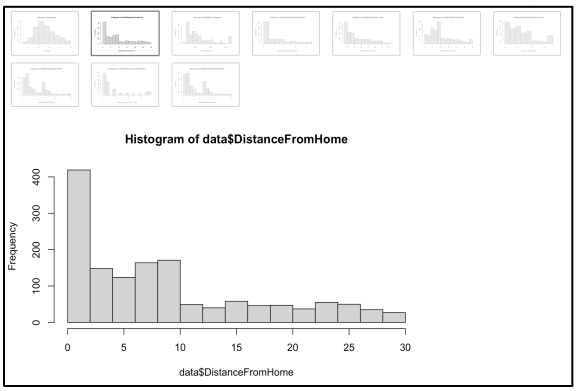
```r
# Histogram
hist(data$MonthlyIncome, main="Monthly Income", col="lightblue", border="black")

# Q-Q Plot
qqnorm(data$MonthlyIncome)
qqline(data$MonthlyIncome, col = 2)

# Shapiro-Wilk Test for Normality
shapiro.test(data$MonthlyIncome)

```





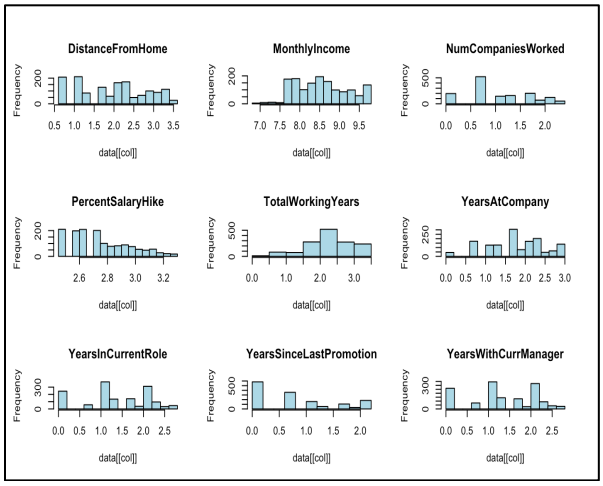
Log Transformation

```r
hist(data$Age)
hist(data$DistanceFromHome)
hist(data$MonthlyIncome)
hist(data$NumCompaniesWorked)
hist(data$PercentSalaryHike)
hist(data$TotalWorkingYears)
hist(data$YearsAtCompany)
hist(data$YearsInCurrentRole)
hist(data$YearsSinceLastPromotion)
hist(data$YearsWithCurrManager)
```



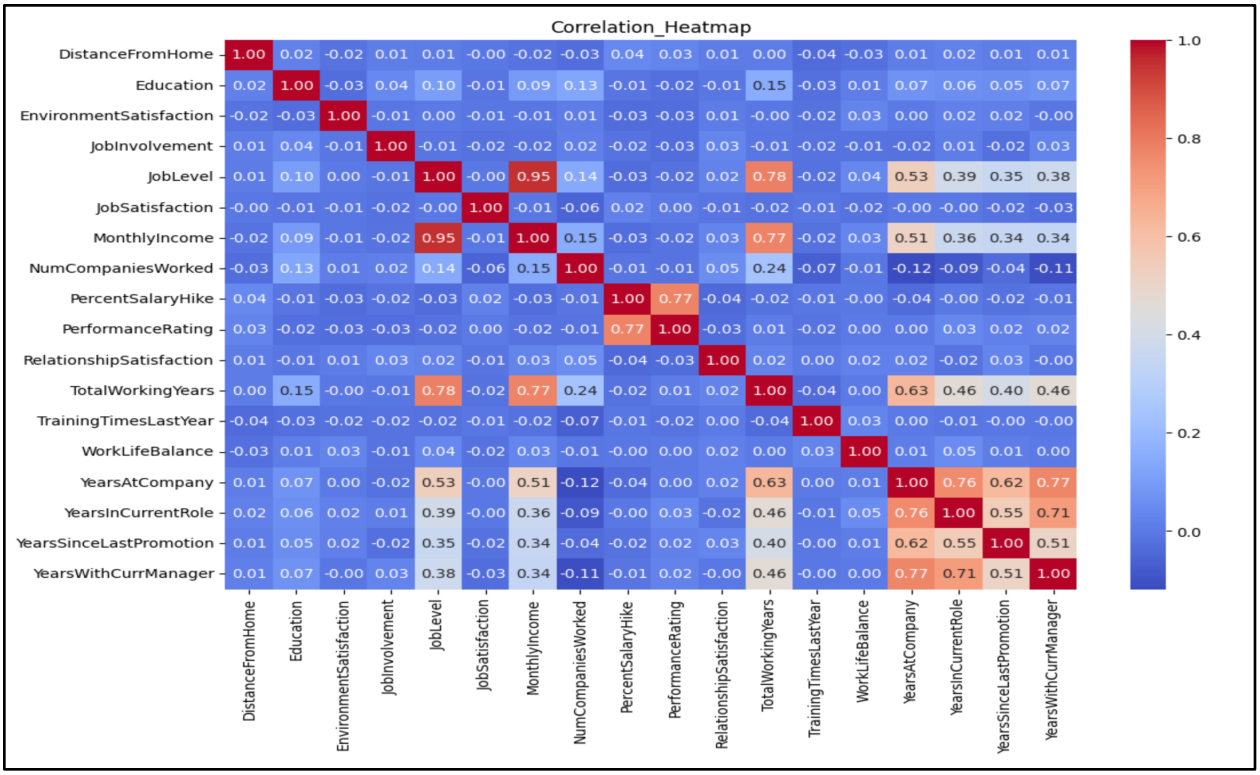
After performing Log Transformation

```r
columns_to_transform <- c(
  "DistanceFromHome",
  "MonthlyIncome",
  "NumCompaniesWorked",
  "PercentSalaryHike",
  "TotalWorkingYears",
  "YearsAtCompany",
  "YearsInCurrentRole",
  "YearsSinceLastPromotion",
  "YearsWithCurrManager"
)

for (col in columns_to_transform) {
  data[[col]] <- ifelse(data[[col]] > 0, log(data[[col]] + 1), 0)
}

par(mfrow = c(3, 3))
for (col in columns_to_transform) {
  hist(data[[col]], main = col, col = "lightblue", border = "black")
}
```



Correlation coefficient

```python
corr_matrix = df.corr(numeric_only=True)
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation_Heatmap')
plt.show()
```



Fisher's exact test

```r
```{r}

your_data <- lapply(data, as.factor)
your_data <- as.data.frame(your_data)


variable_pairs <- combn(names(your_data), 2, simplify = TRUE)

associations <- list()

# Perform Fisher's Exact Test for each pair
for (i in seq(ncol(variable_pairs))) {
  # Create contingency table
  contingency_table <- table(your_data[, variable_pairs[1, i]], your_data[, variable_pairs[2, i]])


  if (all(dim(contingency_table) >= 2)) {
    test_result <- fisher.test(contingency_table, simulate.p.value = TRUE)

    # Check if the p-value is below a significance threshold (e.g., 0.05)
    if (test_result$p.value < 0.05) {
      associations[[paste(variable_pairs[1, i], variable_pairs[2, i], sep = "_")]] <- test_result
    }
  } else {
    cat("Insufficient data for Fisher's Exact Test for", variable_pairs[1, i], "and", variable_pairs[2, i], "\n")
  }
}

# Print the list of associations
cat("List of associations:\n")
print(associations)
```
```

```
List of associations:
$Age_Attrition

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided


$Age_Education

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided


$Age_JobLevel

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided


$Age_JobRole

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided
```

```
$Age_MaritalStatus

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided


$Age_MonthlyIncome

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided


$Age_NumCompaniesWorked

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided


$Age_TotalWorkingYears

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  contingency_table
p-value = 0.0004998
alternative hypothesis: two.sided
```

Multiple Linear Regression

```r
model = lm(MonthlyIncome ~ Age + Attrition + BusinessTravel + Department + DistanceFromHome + Education + EducationField + EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction + TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager, data = data)

summary(model)
```

```
Call:
lm(formula = MonthlyIncome ~ Age + Attrition + BusinessTravel +
    Department + DistanceFromHome + Education + EducationField +
    EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel +
    JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked +
    OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +
    TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3874.3  -807.9   -76.5   764.1  4263.1

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       853.4386   619.2415   1.378 0.168357
Age                                -4.9529     4.7598  -1.041 0.298253
AttritionYes                      -39.4037    94.9364  -0.415 0.678165
BusinessTravelTravel_Frequently   153.6341   119.4688   1.286 0.198660
BusinessTravelTravel_Rarely       143.1395   102.4193   1.398 0.162455
DepartmentResearch & Development  494.9235   418.9965   1.181 0.237715
DepartmentSales                   558.8567   434.6649   1.286 0.198750
DistanceFromHome                   -0.6997     3.7680  -0.186 0.852701
Education                           7.5249    30.6090   0.246 0.805841
EducationFieldLife Sciences        30.1462   300.3375   0.100 0.920061
EducationFieldMarketing            39.4030   319.8898   0.123 0.901984
EducationFieldMedical               5.4063   301.4948   0.018 0.985696
EducationFieldOther               -69.1657   322.5355  -0.214 0.830232
EducationFieldTechnical Degree    140.9911   313.4248   0.450 0.652894
EnvironmentSatisfaction           -14.9401    28.2282  -0.529 0.596707
GenderMale                         53.3323    62.5225   0.853 0.393797
JobInvolvement                    -73.6859    43.2581  -1.703 0.088711 .
JobLevel                         2354.3484    70.1638  33.555  < 2e-16 ***
JobRoleHuman Resources           -110.9667   438.7394  -0.253 0.800365
JobRoleLaboratory Technician     -978.9379   143.0894  -6.841 1.16e-11 ***
JobRoleManager                   3543.8164   214.8507  16.494  < 2e-16 ***
JobRoleManufacturing Director     -55.3790   140.6398  -0.394 0.693813
JobRoleResearch Director         3636.0091   186.7315  19.472  < 2e-16 ***
JobRoleResearch Scientist        -928.5219   141.2778  -6.572 6.92e-11 ***
JobRoleSales Executive           -139.1149   277.5854  -0.501 0.616335
JobRoleSales Representative     -1172.6704   309.3222  -3.791 0.000156 ***
JobSatisfaction                   -16.6929    27.8134  -0.600 0.548485
MaritalStatusMarried               64.8594    78.6620   0.825 0.409775
MaritalStatusSingle                52.7054    85.4619   0.617 0.537522
NumCompaniesWorked                 15.5706    14.0264   1.110 0.267146
OverTimeYes                        93.8492    70.9577   1.323 0.186177
PercentSalaryHike                  20.6346    13.1821   1.565 0.117724
PerformanceRating                -173.8060   133.3113  -1.304 0.192525
RelationshipSatisfaction           12.1528    28.3088   0.429 0.667775
TotalWorkingYears                  58.7051     9.1827   6.393 2.20e-10 ***
TrainingTimesLastYear             -29.2607    23.8241  -1.228 0.219576
WorkLifeBalance                   -21.1924    43.3522  -0.489 0.625029
YearsAtCompany                     28.6468    15.6904   1.826 0.068095 .
YearsInCurrentRole                  0.9785    15.7646   0.062 0.950516
YearsSinceLastPromotion            24.4625    15.5024   1.578 0.114791
YearsWithCurrManager              -53.1801    15.9628  -3.332 0.000886 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1156 on 1429 degrees of freedom
Multiple R-squared:  0.9314,    Adjusted R-squared:  0.9295
F-statistic: 485.4 on 40 and 1429 DF,  p-value: < 2.2e-16
```

```r
model1 = lm(MonthlyIncome ~ JobLevel + JobRole, data = data)

summary(model1)
```

```
Call:
lm(formula = MonthlyIncome ~ JobLevel + JobRole, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3280.8  -851.2   -81.4   762.5  3967.9

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                      790.73     173.77   4.551 5.80e-06 ***
JobLevel                        2724.33      56.46  48.253  < 2e-16 ***
JobRoleHuman Resources          -589.08     201.94  -2.917  0.00359 **
JobRoleLaboratory Technician    -930.04     144.77  -6.424 1.79e-10 ***
JobRoleManager                  3392.06     187.39  18.102  < 2e-16 ***
JobRoleManufacturing Director   -165.50     142.69  -1.160  0.24630
JobRoleResearch Director        3445.84     188.15  18.315  < 2e-16 ***
JobRoleResearch Scientist       -825.55     143.69  -5.746 1.11e-08 ***
JobRoleSales Executive          -209.29     122.72  -1.705  0.08834 .
JobRoleSales Representative    -1118.82     183.65  -6.092 1.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1184 on 1460 degrees of freedom
Multiple R-squared:  0.9265,    Adjusted R-squared:  0.9261
F-statistic:  2045 on 9 and 1460 DF,  p-value: < 2.2e-16
```

ANOVA

```r
#ANOVA
summary(aov(MonthlyIncome ~ JobRole , data = data))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
JobRole       8  464.5   58.06   547.5 <2e-16 ***
Residuals  1461  155.0    0.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```r
pairwise.t.test(data$MonthlyIncome, data$JobRole, p.adj = "none")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  data$MonthlyIncome and data$JobRole

                      Healthcare Representative Human Resources Laboratory Technician Manager Manufacturing Director
Human Resources       < 2e-16                   -               -                     -       -
Laboratory Technician < 2e-16                   0.00013         -                     -       -
Manager               < 2e-16                   < 2e-16         < 2e-16               -       -
Manufacturing Director 0.31740                  < 2e-16         < 2e-16               < 2e-16 -
Research Director     < 2e-16                   < 2e-16         < 2e-16               0.23440 < 2e-16
Research Scientist    < 2e-16                   9.9e-05         0.95828               < 2e-16 < 2e-16
Sales Executive       0.01511                   < 2e-16         < 2e-16               < 2e-16 0.18931
Sales Representative  < 2e-16                   2.1e-11         1.4e-06               < 2e-16 < 2e-16
                      Research Director Research Scientist Sales Executive
Human Resources       -                 -                  -
Laboratory Technician -                 -                  -
Manager               -                 -                  -
Manufacturing Director -                -                  -
Research Director     -                 -                  -
Research Scientist    < 2e-16           -                  -
Sales Executive       < 2e-16           < 2e-16            -
Sales Representative  < 2e-16           1.2e-06            < 2e-16

P value adjustment method: none
```

```r
encoded_data <- cbind(data, model.matrix(~ JobRole - 1, data = data))

encoded_data <- encoded_data[, -which(names(encoded_data) %in% c("JobRole"))]

colnames(encoded_data)[colnames(encoded_data) == "JobRoleHuman Resources"] <- "JobRoleHumanResources"
colnames(encoded_data)[colnames(encoded_data) == "JobRoleLaboratory Technician"] <- "JobRoleLaboratoryTechnician"
colnames(encoded_data)[colnames(encoded_data) == "JobRoleManufacturing Director"] <- "JobRoleManufacturingDirector"
colnames(encoded_data)[colnames(encoded_data) == "JobRoleResearch Scientist"] <- "JobRoleResearchScientist"
colnames(encoded_data)[colnames(encoded_data) == "JobRoleSales Executive"] <- "JobRoleSalesExecutive"
colnames(encoded_data)[colnames(encoded_data) == "JobRoleSales Representative"] <- "JobRoleSalesRepresentative"
colnames(encoded_data)[colnames(encoded_data) == "JobRoleResearch Director"] <- "JobRoleResearchDirector"
```

Equation:
Monthly income = 7.802 + 0.431 (Job Level) - 0.227 (JobRoleHumanResources) - 0.313(JobRoleLaboratoryTechnician)- 0.298(JobRoleResearchScientist) - 0.445 (JobRoleSalesRepresentative).

```{r}
model1 = lm(MonthlyIncome ~ JobLevel + JobRoleHumanResources + JobRoleLaboratoryTechnician +  JobRoleResearchScientist  +
JobRoleSalesRepresentative , data = encoded_data)

summary(model1)
```

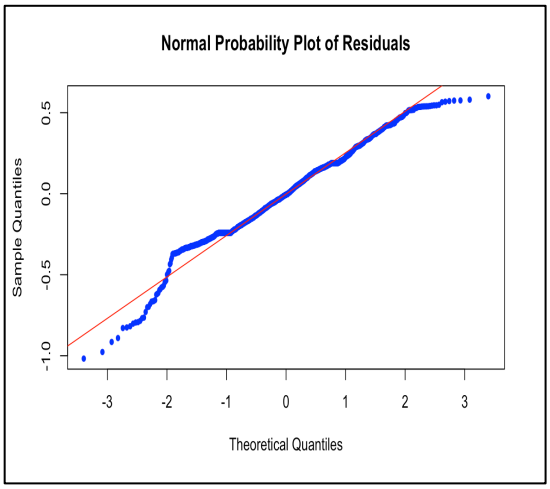
```
Call:
lm(formula = MonthlyIncome ~ JobLevel + JobRoleHumanResources +
    JobRoleLaboratoryTechnician + JobRoleResearchScientist +
    JobRoleSalesRepresentative, data = encoded_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0175 -0.1747 -0.0049  0.1699  0.6013

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   7.802519   0.024490 318.599  < 2e-16 ***
JobLevel                      0.431186   0.008178  52.725  < 2e-16 ***
JobRoleHumanResources        -0.227332   0.036474  -6.233 5.98e-10 ***
JobRoleLaboratoryTechnician -0.313102   0.021609 -14.490  < 2e-16 ***
JobRoleResearchScientist    -0.298462   0.021189 -14.086  < 2e-16 ***
JobRoleSalesRepresentative  -0.445451   0.031394 -14.189  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2433 on 1464 degrees of freedom
Multiple R-squared:  0.8601,    Adjusted R-squared:  0.8596
F-statistic:  1800 on 5 and 1464 DF,  p-value: < 2.2e-16
```

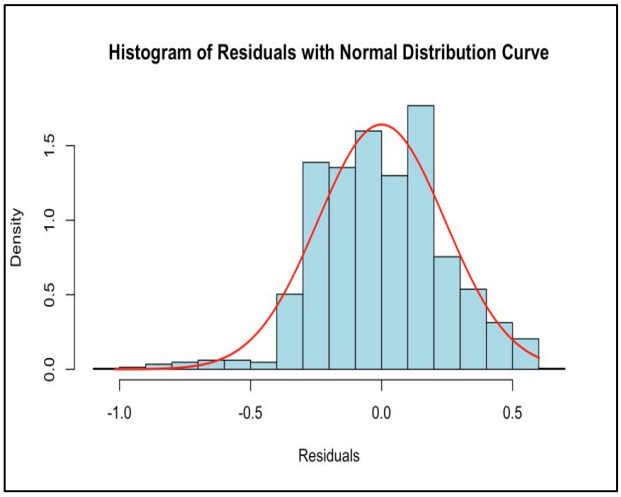
```{r}
model2 = lm(MonthlyIncome ~ JobLevel + JobRole + JobLevel*JobRole, data = data)

summary(model2)
```

```
Call:
lm(formula = MonthlyIncome ~ JobLevel + JobRole + JobLevel *
    JobRole, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99371 -0.13786 -0.01523  0.13747  0.61013

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               7.783578   0.080978  96.119  < 2e-16 ***
JobLevel                                  0.439946   0.031753  13.855  < 2e-16 ***
JobRoleHuman Resources                   -0.553317   0.109648  -5.046 5.07e-07 ***
JobRoleLaboratory Technician             -0.288425   0.090722  -3.179  0.00151 **
JobRoleManager                            1.404181   0.166781   8.419  < 2e-16 ***
JobRoleManufacturing Director            -0.122751   0.111311  -1.103  0.27031
JobRoleResearch Director                  1.291956   0.148844   8.680  < 2e-16 ***
JobRoleResearch Scientist                -0.421273   0.090753  -4.642 3.76e-06 ***
JobRoleSales Executive                   -0.038403   0.097396  -0.394  0.69342
JobRoleSales Representative              -0.397669   0.128616  -3.092  0.00203 **
JobLevel:JobRoleHuman Resources           0.224176   0.055252   4.057 5.23e-05 ***
JobLevel:JobRoleLaboratory Technician    -0.013389   0.044374  -0.302  0.76290
JobLevel:JobRoleManager                  -0.327938   0.046139  -7.108 1.85e-12 ***
JobLevel:JobRoleManufacturing Director    0.038594   0.043847   0.880  0.37889
JobLevel:JobRoleResearch Director        -0.304993   0.044215  -6.898 7.85e-12 ***
JobLevel:JobRoleResearch Scientist        0.109165   0.045265   2.412  0.01600 *
JobLevel:JobRoleSales Executive           0.008705   0.038982   0.223  0.82332
JobLevel:JobRoleSales Representative     -0.035358   0.094745  -0.373  0.70906
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.226 on 1452 degrees of freedom
Multiple R-squared:  0.8803,    Adjusted R-squared:  0.8789
F-statistic:   628 on 17 and 1452 DF,  p-value: < 2.2e-16
```

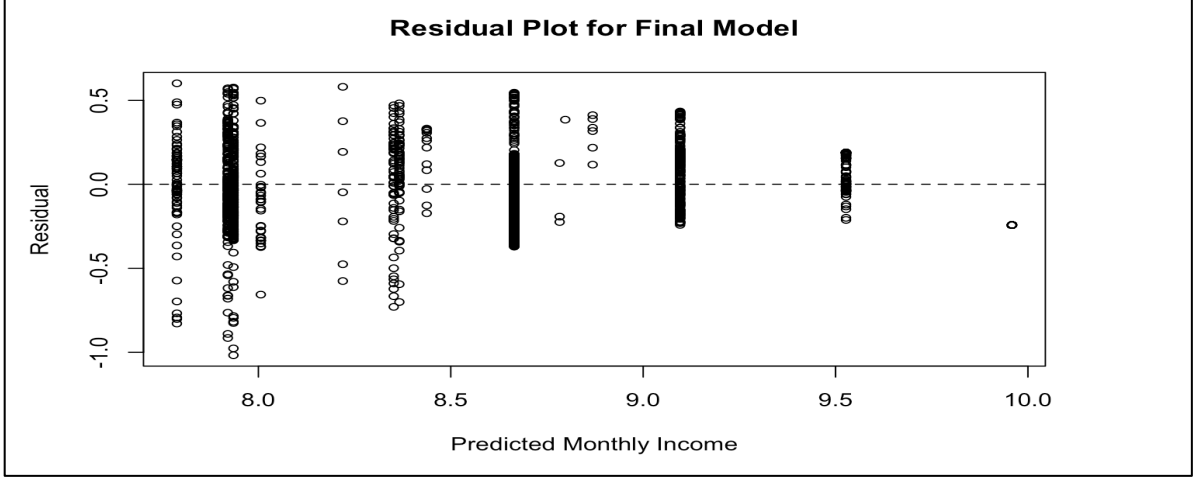
Normality of residuals

```r
residuals <- residuals(model1)

hist(residuals, main = "Histogram of Residuals with Normal Distribution Curve", col = "lightblue", border = "black", xlab =
"Residuals", prob = TRUE)


mu <- mean(residuals)
sigma <- sd(residuals)
x <- seq(min(residuals), max(residuals), length = 100)
lines(x, dnorm(x, mean = mu, sd = sigma), col = "red", lwd = 2)

qqnorm(residuals, main = "Normal Probability Plot of Residuals", col = "blue", pch = 20)
qqline(residuals, col = "red")
```






```r
#residual plot
plot(resid(model1) ~ fitted(model1),
     xlab = "Predicted Monthly Income",  ylab = "Residual",
     main = "Residual Plot for Final Model",
     pch = 21, cex = 0.8)
abline(h = 0, lty = 2)
```



Multiple Logistic Regression

```r
model <- glm(Attrition ~ Age + MonthlyIncome + BusinessTravel + Department +
               DistanceFromHome + Education + EducationField + EnvironmentSatisfaction +
               Gender + JobInvolvement + JobLevel + JobRole + JobSatisfaction +
               MaritalStatus + NumCompaniesWorked + OverTime + PercentSalaryHike +
               PerformanceRating + RelationshipSatisfaction + TotalWorkingYears +
               TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany +
               YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager,
               data = data_oversampled, family = binomial(link = "logit"))

summary(model)
```

```
Call:
glm(formula = Attrition ~ Age + MonthlyIncome + BusinessTravel +
    Department + DistanceFromHome + Education + EducationField +
    EnvironmentSatisfaction + Gender + JobInvolvement + JobLevel +
    JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked +
    OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +
    TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager, family = binomial(link = "logit"),
    data = data_oversampled)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -5.536452 372.835812  -0.015 0.988152
Age                           -0.003997   0.007406  -0.540 0.589376
MonthlyIncome                 -0.616045   0.222382  -2.770 0.005602 **
BusinessTravelTravel_Frequently  1.785153   0.226491   7.882 3.23e-15 ***
BusinessTravelTravel_Rarely    0.993786   0.204030   4.871 1.11e-06 ***
DepartmentResearch & Development 15.025332 372.831100   0.040 0.967853
DepartmentSales               14.788808 372.831234   0.040 0.968359
DistanceFromHome               0.368945   0.063330   5.826 5.69e-09 ***
Education                      0.020877   0.052578   0.397 0.691324
EducationFieldLife Sciences   -1.094122   0.493709  -2.216 0.026683 *
EducationFieldMarketing       -0.680583   0.525747  -1.295 0.195490
EducationFieldMedical         -0.815641   0.489306  -1.667 0.095527 .
EducationFieldOther           -0.533487   0.524854  -1.016 0.309416
EducationFieldTechnical Degree  0.270395   0.507344   0.533 0.594060
EnvironmentSatisfaction       -0.503246   0.049237 -10.221  < 2e-16 ***
GenderMale                     0.476691   0.110623   4.309 1.64e-05 ***
JobInvolvement                -0.489756   0.074325  -6.589 4.42e-11 ***
JobLevel                       0.442615   0.141313   3.132 0.001735 **
JobRoleHuman Resources        16.342960 372.831176   0.044 0.965036
JobRoleLaboratory Technician   1.512235   0.286010   5.287 1.24e-07 ***
JobRoleManager                 0.438690   0.403344   1.088 0.276758
JobRoleManufacturing Director  0.843003   0.281340   2.996 0.002732 **
JobRoleResearch Director      -1.414074   0.516899  -2.736 0.006225 **
JobRoleResearch Scientist      0.670020   0.287902   2.327 0.019952 *
JobRoleSales Executive         1.443128   0.579892   2.489 0.012824 *
JobRoleSales Representative    2.311752   0.629742   3.671 0.000242 ***
JobSatisfaction               -0.370285   0.048075  -7.702 1.34e-14 ***
MaritalStatusMarried           0.691127   0.148825   4.644 3.42e-06 ***
MaritalStatusSingle            1.620060   0.156289  10.366  < 2e-16 ***
NumCompaniesWorked             0.696545   0.093807   7.425 1.13e-13 ***
OverTimeYes                    1.857652   0.113279  16.399  < 2e-16 ***
PercentSalaryHike             -1.125843   0.358913  -3.137 0.001708 **
PerformanceRating              0.188404   0.218592   0.862 0.388744
RelationshipSatisfaction      -0.265524   0.048286  -5.499 3.82e-08 ***
TotalWorkingYears             -0.787626   0.150225  -5.243 1.58e-07 ***
TrainingTimesLastYear         -0.190298   0.041367  -4.600 4.22e-06 ***
WorkLifeBalance               -0.337074   0.069592  -4.844 1.28e-06 ***
YearsAtCompany                 0.599085   0.191969   3.121 0.001804 **
YearsInCurrentRole            -0.551864   0.132632  -4.161 3.17e-05 ***
YearsSinceLastPromotion        0.677830   0.090991   7.449 9.38e-14 ***
YearsWithCurrManager          -0.529121   0.120542  -4.390 1.14e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3726.3  on 2702  degrees of freedom
Residual deviance: 2354.3  on 2662  degrees of freedom
AIC: 2436.3

Number of Fisher Scoring iterations: 14
```

Ordinary Least Squares regression

```r
install.packages("ordinal")
library(ordinal)

data$PerformanceRating <- ordered(data$PerformanceRating)

model3 <- clm(PerformanceRating ~ JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction +
                WorkLifeBalance, data = data)

summary(model3)
```

```
The downloaded binary packages are in
        /var/folders/l7/k8n1pmmx5fb2wnz3twy159300000gn/T//Rtmp9BqqU2/downloaded_packages
formula: PerformanceRating ~ JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction
+ WorkLifeBalance
data:    data


Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
JobSatisfaction           0.004497   0.065761   0.068    0.945
EnvironmentSatisfaction  -0.074077   0.065780  -1.126    0.260
RelationshipSatisfaction -0.079327   0.066427  -1.194    0.232
WorkLifeBalance           0.015907   0.102624   0.155    0.877

Threshold coefficients:
    Estimate Std. Error z value
3|4   1.3499     0.4212   3.205
```

```r
model4 <- clm(PerformanceRating ~ JobSatisfaction + EnvironmentSatisfaction + RelationshipSatisfaction + WorkLifeBalance +
JobInvolvement + DistanceFromHome + TotalWorkingYears + Education + JobLevel + NumCompaniesWorked +  TotalWorkingYears +
TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager
, data = data)

summary(model4)
```

```
+ YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager
data:    data


Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
JobSatisfaction            0.006125   0.066116   0.093    0.926
EnvironmentSatisfaction   -0.078050   0.066210  -1.179    0.238
RelationshipSatisfaction -0.070266   0.066814  -1.052    0.293
WorkLifeBalance            0.019066   0.103878   0.184    0.854
JobInvolvement            -0.117780   0.100651  -1.170    0.242
DistanceFromHome           0.008557   0.008733   0.980    0.327
TotalWorkingYears          0.026002   0.017006   1.529    0.126
Education                 -0.076855   0.071914  -1.069    0.285
JobLevel                  -0.174667   0.106976  -1.633    0.103
NumCompaniesWorked        -0.019232   0.033118  -0.581    0.561
TrainingTimesLastYear     -0.033477   0.056963  -0.588    0.557
YearsAtCompany            -0.042311   0.027130  -1.560    0.119
YearsInCurrentRole         0.051974   0.033983   1.529    0.126
YearsSinceLastPromotion    0.016416   0.028931   0.567    0.570
YearsWithCurrManager       0.020660   0.033970   0.608    0.543

Threshold coefficients:
    Estimate Std. Error z value
3|4    0.762      0.585   1.303
```

```{r}
table(data$PercentSalaryHike, data$PerformanceRating)

```

```
      3   4
  11 210   0
  12 198   0
  13 209   0
  14 201   0
  15 101   0
  16  78   0
  17  82   0
  18  89   0
  19  76   0
  20   0  55
  21   0  48
  22   0  56
  23   0  28
  24   0  21
  25   0  18
```