What Determines Employee Salaries: A Regression Analysis Study

Neeraj Namani & Srisailam Gitte

2024-04-26

```
# Read the data from Salary data csv file
data1 <- read.csv("~/Downloads/Salary Data.csv")
```

Exploratory Data Analysis:

```
# Display the first few rows of the data
head(data1)
```

```
##   Age Gender Education.Level         Job.Title Years.of.Experience Salary
## 1  32   Male      Bachelor's Software Engineer                  5  90000
## 2  28 Female        Master's      Data Analyst                  3  65000
## 3  45   Male             PhD    Senior Manager                 15 150000
## 4  36 Female      Bachelor's   Sales Associate                  7  60000
## 5  52   Male        Master's          Director                 20 200000
## 6  29   Male      Bachelor's Marketing Analyst                  2  55000
```

```
# Display the last few rows of the data
tail(data1)
```

```
##     Age Gender Education.Level                     Job.Title
## 370  33   Male      Bachelor's       Junior Business Analyst
## 371  35 Female      Bachelor's      Senior Marketing Analyst
## 372  43   Male        Master's        Director of Operations
## 373  29 Female      Bachelor's        Junior Project Manager
## 374  34   Male      Bachelor's Senior Operations Coordinator
## 375  44 Female             PhD       Senior Business Analyst
##     Years.of.Experience Salary
## 370                   4  60000
## 371                   8  85000
## 372                  19 170000
## 373                   2  40000
## 374                   7  90000
## 375                  15 150000
```

```
# Display the shape of the dataset
dim(data1)
```

```
## [1] 375   6
```

```
# Check for missing values
colSums(is.na(data1))
```

```
##                 Age              Gender     Education.Level           Job.Title
##                   2                   0                   0                   0
## Years.of.Experience              Salary
##                   2                   2
```

```r
cleaned_data <- na.omit(data1)
nrow(cleaned_data)
```

```
## [1] 373
```

```r
# Re check for missing values in the data
colSums(is.na(cleaned_data))
```

```
##                 Age              Gender     Education.Level           Job.Title
##                   0                   0                   0                   0
## Years.of.Experience              Salary
##                   0                   0
```

```r
# Data Frame
str(cleaned_data)
```

```
## 'data.frame':    373 obs. of  6 variables:
##  $ Age                : int  32 28 45 36 52 29 42 31 26 38 ...
##  $ Gender             : chr  "Male" "Female" "Male" "Female" ...
##  $ Education.Level    : chr  "Bachelor's" "Master's" "PhD" "Bachelor's" ...
##  $ Job.Title          : chr  "Software Engineer" "Data Analyst" "Senior Manager" "Sales Associate" .
##  $ Years.of.Experience: num  5 3 15 7 20 2 12 4 1 10 ...
##  $ Salary             : int  90000 65000 150000 60000 200000 55000 120000 80000 45000 110000 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2] 173 261
##   ..- attr(*, "names")= chr [1:2] "173" "261"
```

```r
# Load the necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Dataset Structure
glimpse(cleaned_data)
```

```
## Rows: 373
## Columns: 6
## $ Age                <int> 32, 28, 45, 36, 52, 29, 42, 31, 26, 38, 29, 48, 35~
## $ Gender             <chr> "Male", "Female", "Male", "Female", "Male", "Male"~
## $ Education.Level    <chr> "Bachelor's", "Master's", "PhD", "Bachelor's", "Ma~
## $ Job.Title          <chr> "Software Engineer", "Data Analyst", "Senior Manag~
## $ Years.of.Experience <dbl> 5, 3, 15, 7, 20, 2, 12, 4, 1, 10, 3, 18, 6, 14, 2,~
## $ Salary             <int> 90000, 65000, 150000, 60000, 200000, 55000, 120000~
```

```r
# Generate Summary statistics for the data
summary(cleaned_data)
```

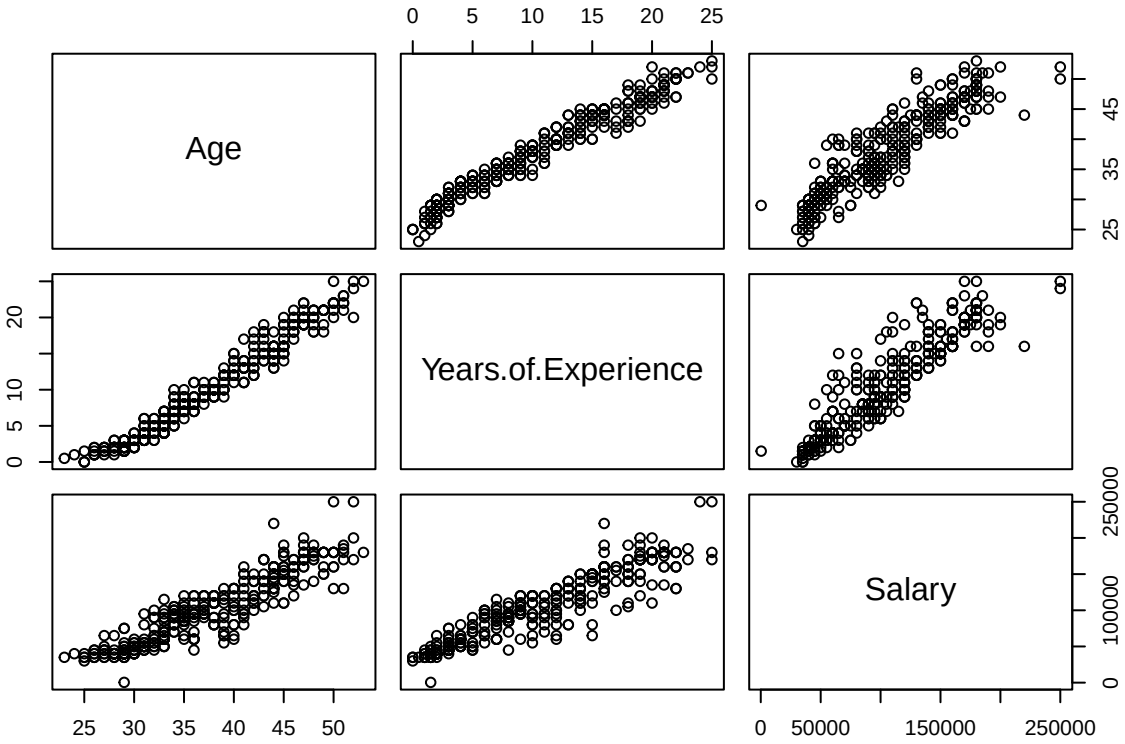
```
##       Age           Gender          Education.Level     Job.Title
##  Min.   :23.00   Length:373         Length:373         Length:373
##  1st Qu.:31.00   Class :character   Class :character   Class :character
##  Median :36.00   Mode  :character   Mode  :character   Mode  :character
```

```
##  Mean   :37.43
##  3rd Qu.:44.00
##  Max.   :53.00
##  Years.of.Experience     Salary
##  Min.   : 0.00       Min.   :   350
##  1st Qu.: 4.00       1st Qu.: 55000
##  Median : 9.00       Median : 95000
##  Mean   :10.03       Mean   :100577
##  3rd Qu.:15.00       3rd Qu.:140000
##  Max.   :25.00       Max.   :250000
```

```r
# Load the required libraries for EDA
library(dplyr) # For data manipulation
```

```r
library(ggplot2) # For data visualization
```

```r
# Implement Pairwise Scatterplot
pairs(cleaned_data[sapply(cleaned_data,is.numeric)])
```



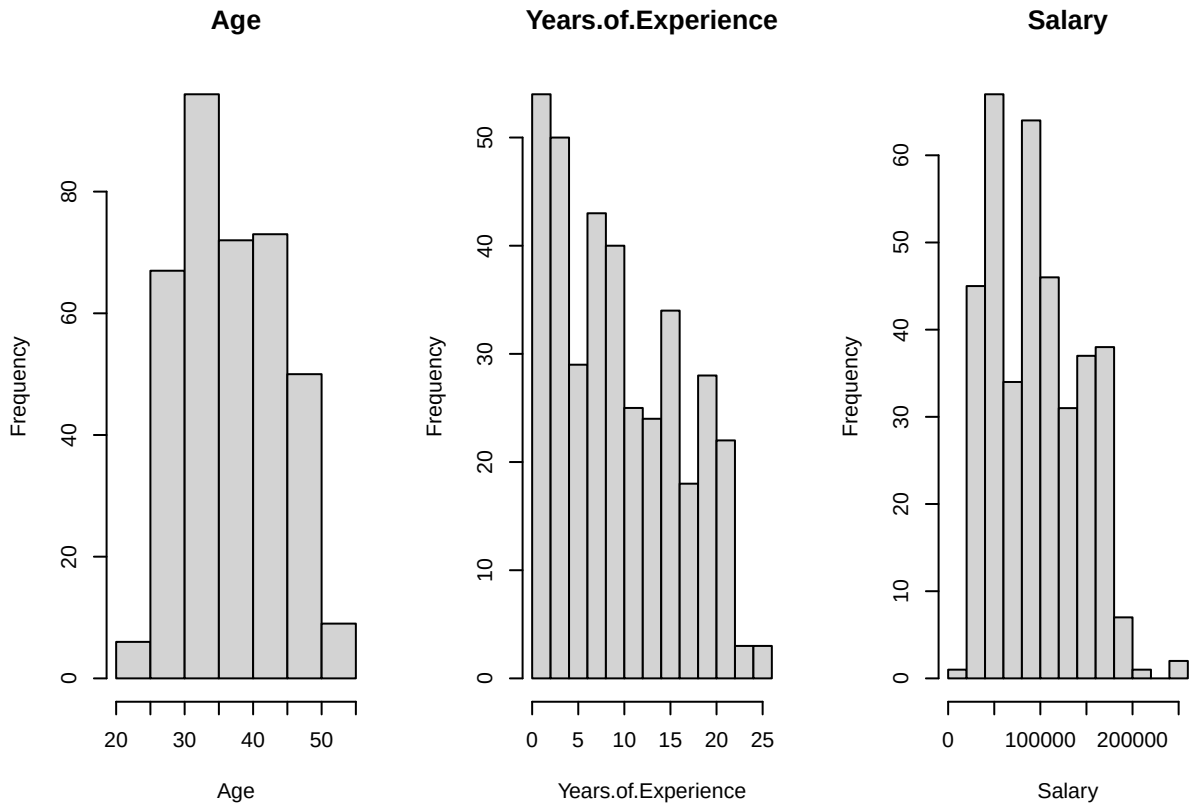

```r
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

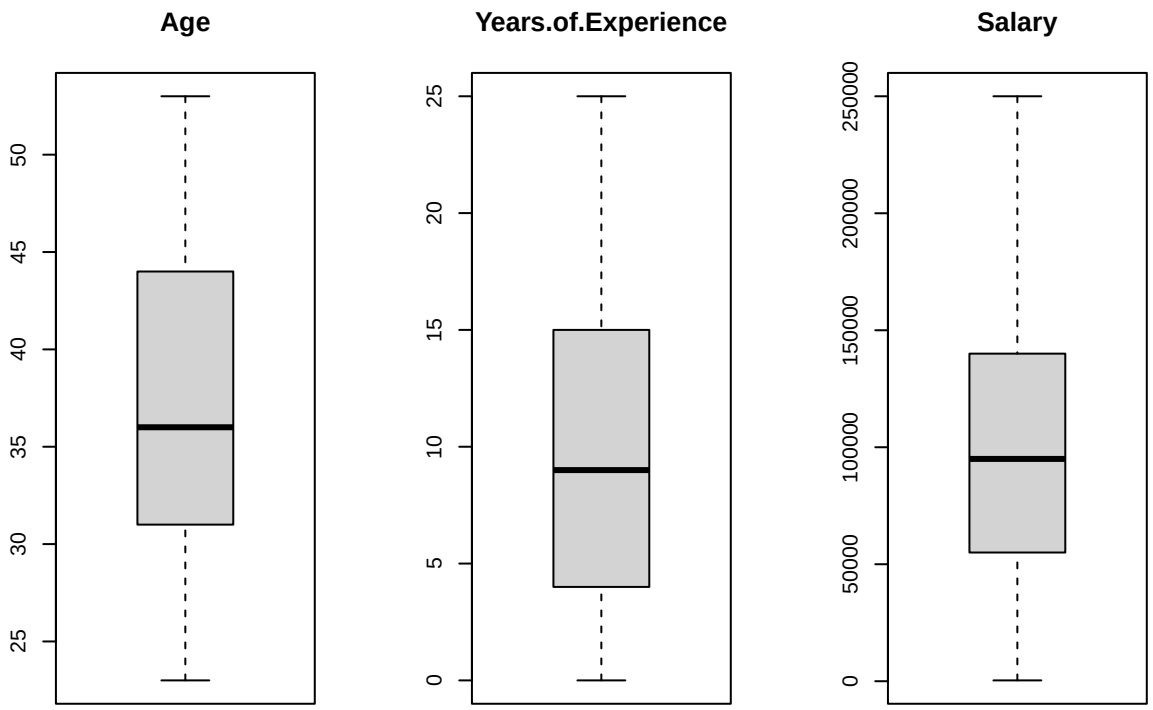
```r
library(ggplot2)
```

```r
# Use ggpairs to plot numerical variables
ggpairs(cleaned_data[sapply(cleaned_data,is.numeric)])
```


```r
# Define a function to plot histograms for numeric columns
plot_numeric_histograms <- function(cleaned_data) {
  par(mfrow = c(1,3)) # Set up a 1x3 grid for plotting

  # Get the names of numeric columns
  numeric_cols <- names(cleaned_data)[sapply(cleaned_data, is.numeric)]

  # Iterate through numeric columns and create histograms
  for (col_name in numeric_cols) {
    hist(cleaned_data[[col_name]], main = col_name, xlab = col_name)
  }
}
# Call the function with my dataset
plot_numeric_histograms(cleaned_data)
```

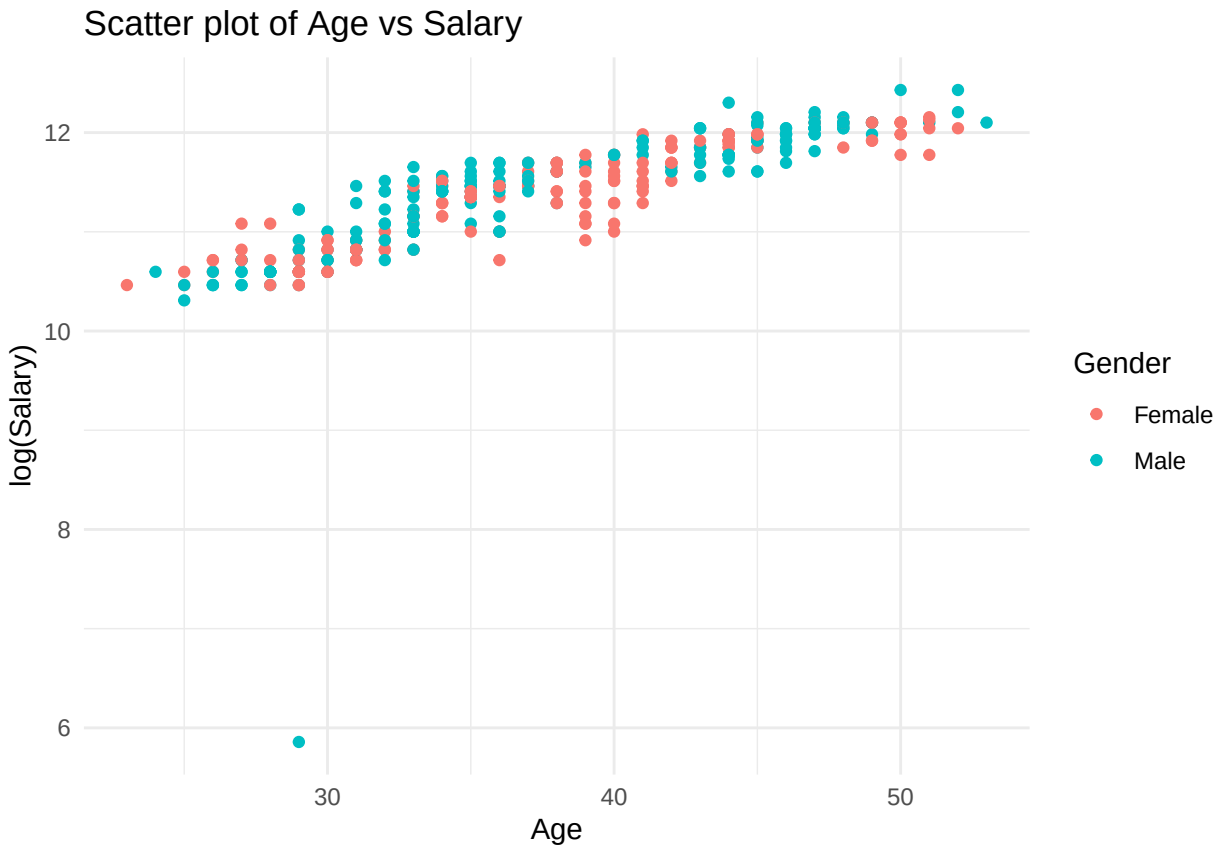



```r
# Set up the layout for the boxplots
par(mfrow = c(1,3)) # 1 row and 3 columns
# Get the names of numeric columns
numeric_cols <- names(cleaned_data)[sapply(cleaned_data, is.numeric)]
# Iterate through numeric columns and create boxplots
for (col_name in numeric_cols) {
  boxplot(cleaned_data[[col_name]], main = col_name, xlab = col_name)
}
```



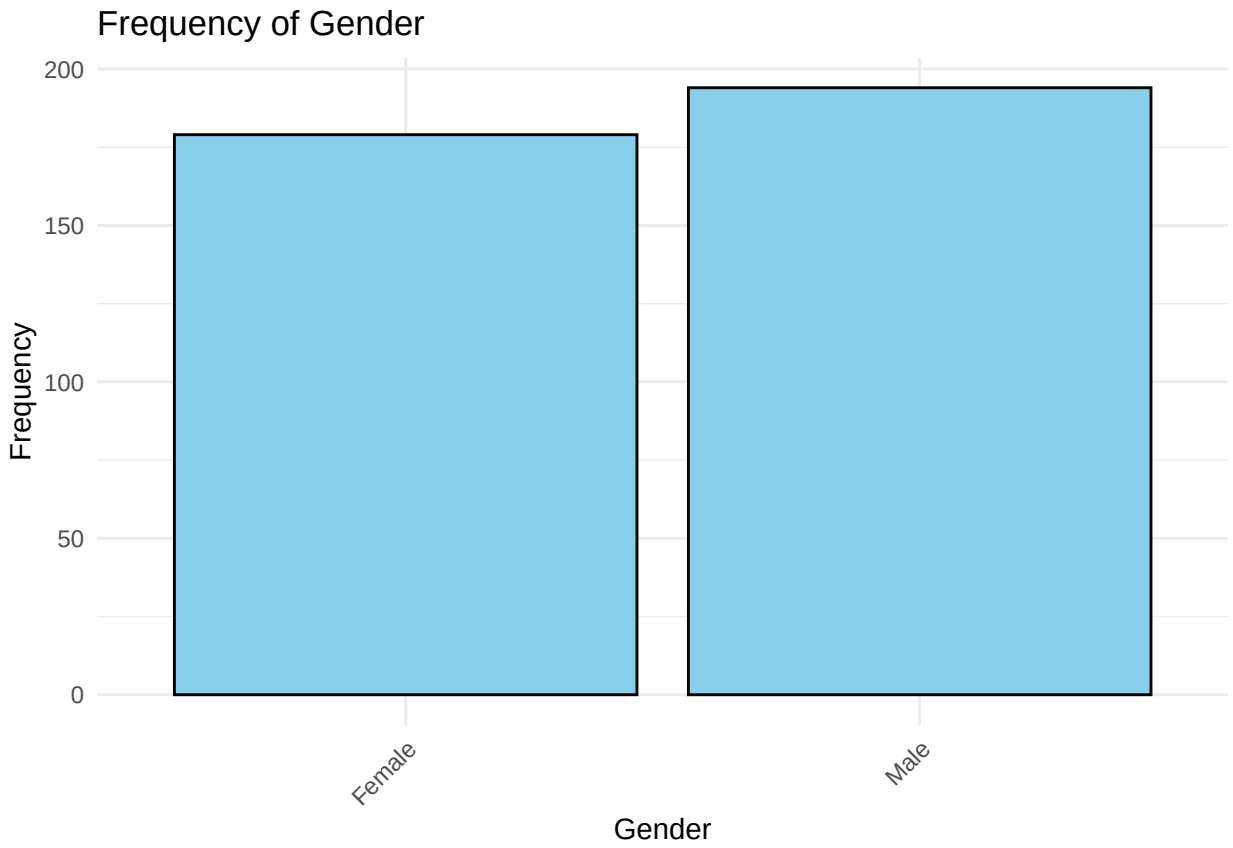

```
# Reset the layout to the default
```

```
library(ggplot2)

# Create a scatter plot with Age on the x-axis and Salary on the y-axis
ggplot(cleaned_data, aes(x = Age, y = log(Salary), color = Gender)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Scatter plot of Age vs Salary", x = "Age", y = "log(Salary)")
```

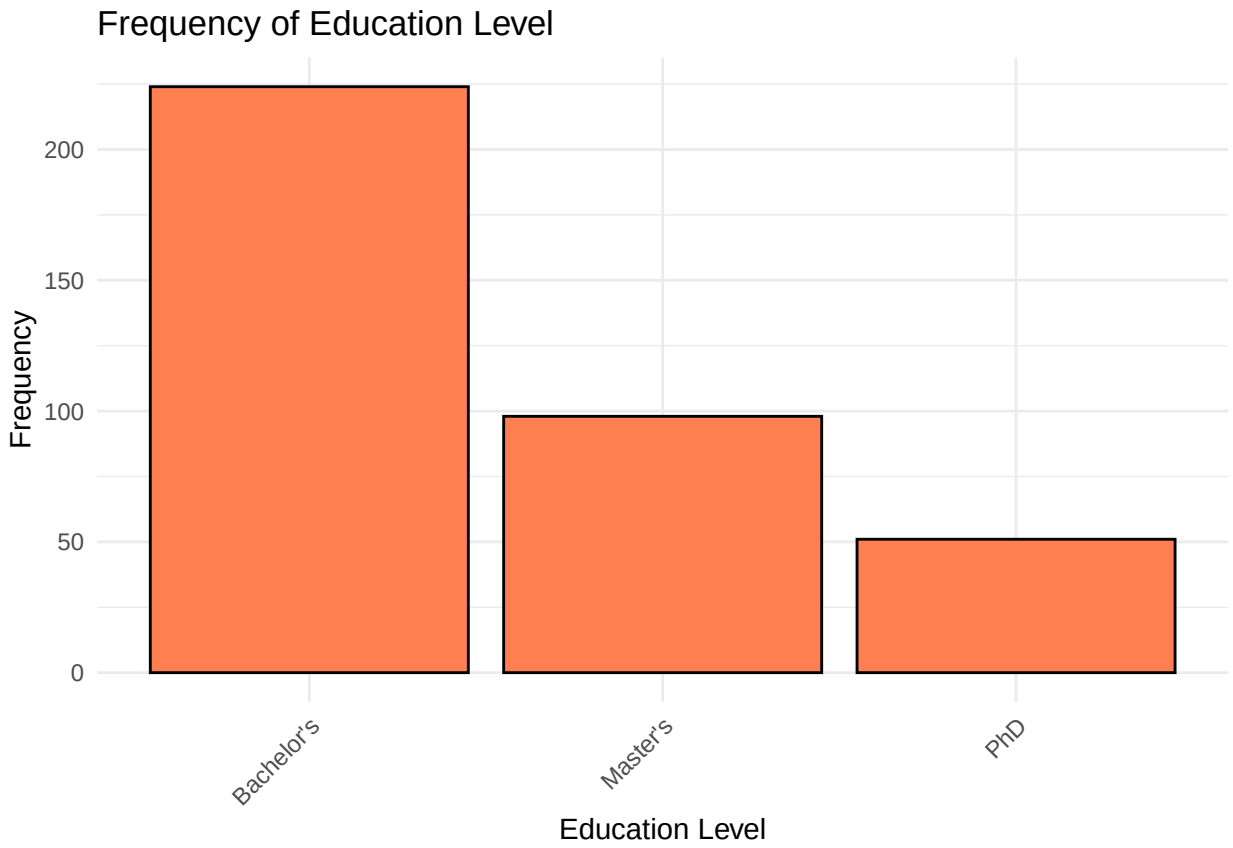



```
# Create a scatterplot with years of experience on the x-axis, Salary on the y-axis and Color by Educat
ggplot(cleaned_data, aes(x = `Years.of.Experience`,y = Salary, color = `Education.Level`)) +
  geom_point()+
  theme_minimal() +
  labs(title = "Scatter plot of Experience Level vs Salary", x = "Experience", y = "Salary")
```

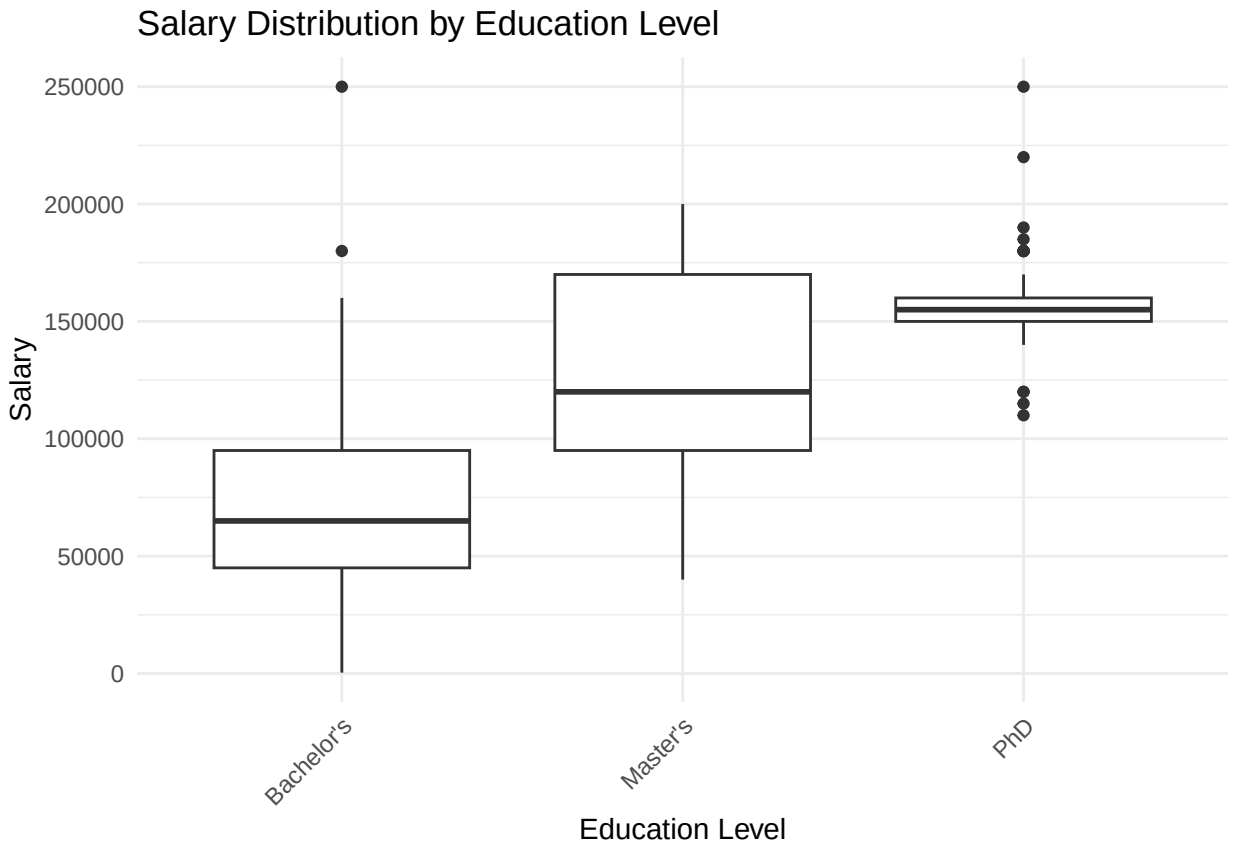

```r
# Bar plot for Gender Frequency
ggplot(cleaned_data, aes(x = Gender)) +
  geom_bar(fill = "skyblue", color = "black") +
  theme_minimal() +
  ggtitle("Frequency of Gender") +
  xlab("Gender") +
  ylab("Frequency") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Bar plot for Education Frequency
ggplot(cleaned_data, aes(x = Education.Level)) +
  geom_bar(fill = "coral", color = "black") +
  theme_minimal() +
  ggtitle("Frequency of Education Level") +
  xlab("Education Level") +
  ylab("Frequency") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Box plot of Salary by Education Level
ggplot(cleaned_data, aes(x = Education.Level, y = Salary)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Salary Distribution by Education Level") +
  xlab("Education Level") +
  ylab("Salary") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Let us calculate the average salary for each gender
average_salary_by_gender <- cleaned_data %>%
  group_by(Gender) %>%
  summarise(Average_Salary = mean(Salary, na.rm = TRUE))

# Print the results
print(average_salary_by_gender)
```

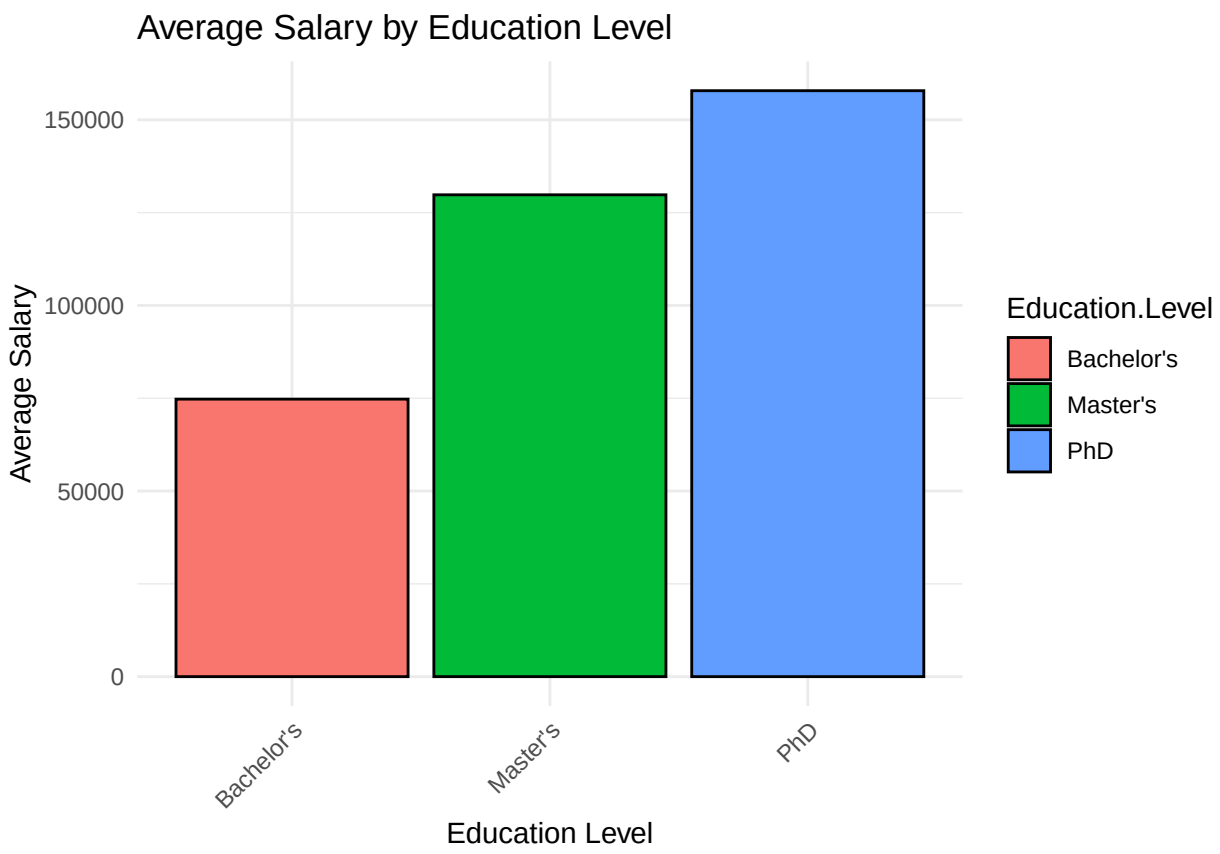
```
## # A tibble: 2 x 2
##   Gender Average_Salary
##   <chr>           <dbl>
## 1 Female         97011.
## 2 Male          103868.
```

```
# Bar plot for Average Salary by Gender
ggplot(average_salary_by_gender, aes(x = Gender, y = Average_Salary, fill = Gender)) +
  geom_bar(stat = "identity", color = "black") +
  theme_minimal() +
  ggtitle("Average Salary by Gender") +
  xlab("Gender") +
  ylab("Average Salary")
```


```
# Calculate the average salary for different education levels
average_salary_by_education <- cleaned_data %>%
  group_by(Education.Level) %>%
  summarise(Average_Salary = mean(Salary, na.rm = TRUE))

print(average_salary_by_education)
```

```
## # A tibble: 3 x 2
##   Education.Level Average_Salary
##   <chr>                    <dbl>
## 1 Bachelor's               74756.
## 2 Master's                129796.
## 3 PhD                     157843.
```

```
ggplot(average_salary_by_education, aes(x = Education.Level, y = Average_Salary, fill = Education.Level
  geom_bar(stat = "identity", color = "black") +
  theme_minimal() +
  ggtitle("Average Salary by Education Level") +
  xlab("Education Level") +
  ylab("Average Salary") +
  theme(axis.text.x = element_text(angle = 45,hjust = 1))
```

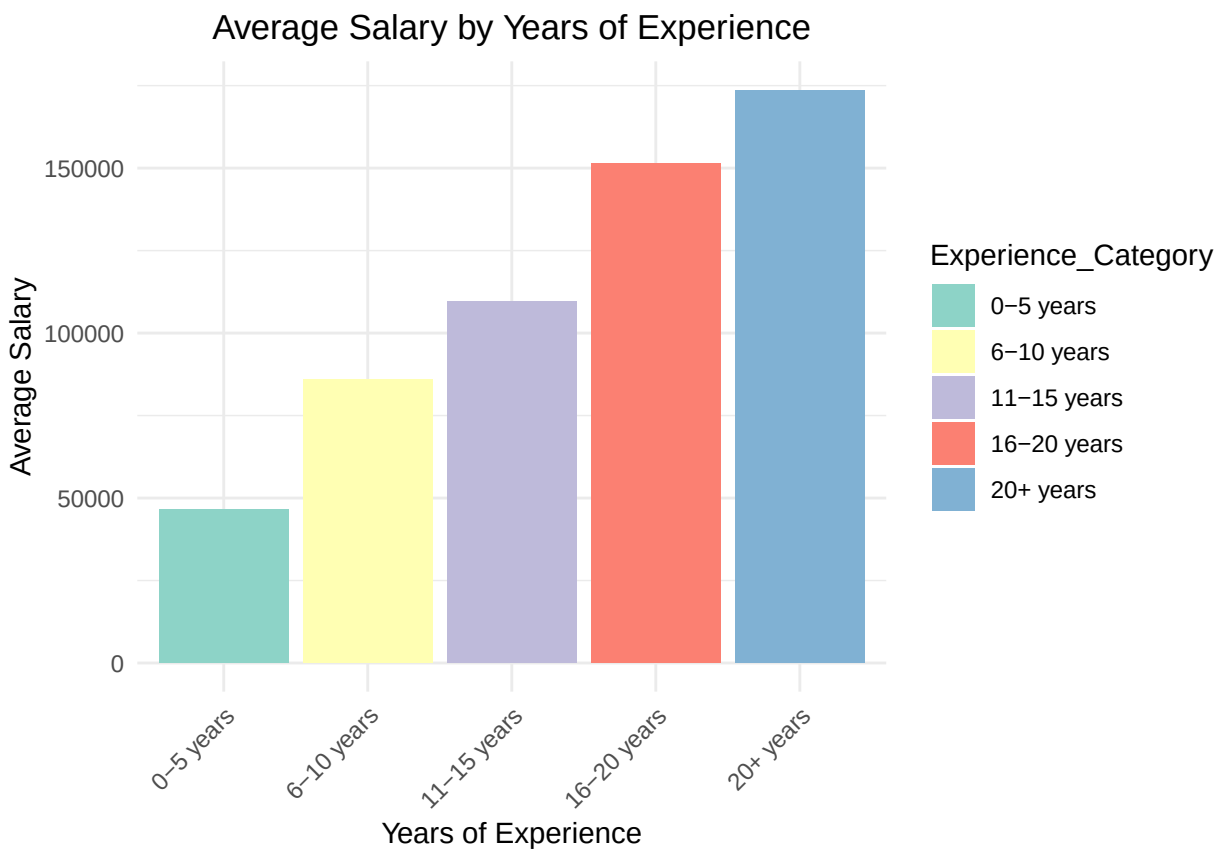



```
# Load the necessary library
library(dplyr)

# Assuming your data is already loaded into a data frame called `data`

# Define the breaks and labels for the experience categories
exp_breaks <- c(-Inf, 5, 10, 15, 20, Inf)
exp_labels <- c("0-5 years", "6-10 years", "11-15 years", "16-20 years", "20+ years")

# Group years of experience into categories and calculate the average salary
salary_by_exp <- cleaned_data %>%
  mutate(Experience_Category = cut(Years.of.Experience, breaks = exp_breaks, labels = exp_labels, right
  group_by(Experience_Category) %>%
  summarise(Average_Salary = mean(Salary, na.rm = TRUE)) %>%
  arrange(desc(Average_Salary))

# Format the average salary as currency
salary_by_exp$Average_Salary <- scales::dollar(salary_by_exp$Average_Salary)

# Print the result
print(salary_by_exp)
```

```
## # A tibble: 5 x 2
##   Experience_Category Average_Salary
##   <fct>               <chr>
## 1 20+ years           $173,659
## 2 16-20 years         $151,343
```

```
## 3 11-15 years         $109,627
## 4 6-10 years          $85,904
## 5 0-5 years           $46,494
```

```r
# Load the necessary library
library(ggplot2)
library(dplyr)

# Assuming your data is already loaded into a variable called `data`

# Categorize 'Years of Experience' and compute the average salary for each category
salary_by_exp <- cleaned_data %>%
  mutate(Experience_Category = cut(Years.of.Experience,
                                   breaks = c(-Inf, 5, 10, 15, 20, Inf),
                                   labels = c("0-5 years", "6-10 years", "11-15 years", "16-20 years",
                                   right = FALSE)) %>%
  group_by(Experience_Category) %>%
  summarise(Average_Salary = mean(Salary, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(desc(Average_Salary))

# Create a simple bar plot
ggplot(salary_by_exp, aes(x = Experience_Category, y = Average_Salary, fill = Experience_Category)) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal() +
  labs(title = "Average Salary by Years of Experience",
       x = "Years of Experience",
       y = "Average Salary") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), # Rotate x labels for readability
        plot.title = element_text(hjust = 0.5)) # Center the plot title
```




```r
library(reshape2)

# Create a correlation matrix for your data
corr_matrix <- cor(cleaned_data[sapply(cleaned_data,is.numeric)])

# Reduce the size of the correlation matrix
melted_corr_matrix <- melt(corr_matrix)

# Create a correlation heatmap using ggplot2
ggplot(data = melted_corr_matrix, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +
  scale_fill_gradient(low = "grey40", high = "grey80")+
  labs(title = "Correlation Heatmap")
```


```
# Install and load the fastDummies library if not already installed
# install.packages("fastDummies")
library(fastDummies)
```

```
## Thank you for using fastDummies!

## To acknowledge our work, please cite the package:

## Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from
```

```
# Encode the 'Education Level' categorical data
df_encoded <- dummy_cols(cleaned_data, select_columns = "Education.Level", remove_first_dummy = TRUE)

# View the first few rows of the new data frame
head(df_encoded)
```

```
##   Age Gender Education.Level         Job.Title Years.of.Experience Salary
## 1  32   Male      Bachelor's Software Engineer                  5  90000
## 2  28 Female        Master's      Data Analyst                  3  65000
## 3  45   Male             PhD    Senior Manager                 15 150000
## 4  36 Female      Bachelor's   Sales Associate                  7  60000
## 5  52   Male        Master's          Director                 20 200000
## 6  29   Male      Bachelor's Marketing Analyst                  2  55000
##   Education.Level_Master's Education.Level_PhD
## 1                        0                   0
## 2                        1                   0
## 3                        0                   1
## 4                        0                   0
## 5                        1                   0
```

```
## 6                         0                    0
```

```r
# Perform simple linear Regression
full_model <- lm(Salary ~ ., data = cleaned_data)
summary(full_model)
```

```
##
## Call:
## lm(formula = Salary ~ ., data = cleaned_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20404  -2385      0   2533  21938
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 -4437.40   15520.45  -0.286
## Age                                          2221.39     474.43   4.682
## GenderMale                                    830.25    1448.30   0.573
## Education.LevelMaster's                      4184.73    2349.57   1.781
## Education.LevelPhD                          13251.93    3679.17   3.602
## Job.TitleAccountant                        -15443.83   12096.40  -1.277
## Job.TitleAdministrative Assistant          -42405.50   10522.48  -4.030
## Job.TitleBusiness Analyst                   -4013.85   10482.14  -0.383
## Job.TitleBusiness Development Manager        1858.91   12209.58   0.152
## Job.TitleBusiness Intelligence Analyst        584.83   12205.72   0.048
## Job.TitleCEO                               104924.45   13089.83   8.016
## Job.TitleChief Data Officer                 88541.60   12631.16   7.010
## Job.TitleChief Technology Officer           88734.27   13015.41   6.818
## Job.TitleContent Marketing Manager         -11717.92   12105.82  -0.968
## Job.TitleCopywriter                        -20770.61   12135.18  -1.712
## Job.TitleCreative Director                   5890.88   12386.03   0.476
## Job.TitleCustomer Service Manager          -14654.31   10762.78  -1.362
## Job.TitleCustomer Service Rep              -19379.47   12052.00  -1.608
## Job.TitleCustomer Service Representative   -26948.36   12113.46  -2.225
## Job.TitleCustomer Success Manager          -21600.86   12030.47  -1.796
## Job.TitleCustomer Success Rep              -14823.30   12183.65  -1.217
## Job.TitleData Analyst                      -15921.53   10706.13  -1.487
## Job.TitleData Entry Clerk                  -16097.39   12203.61  -1.319
## Job.TitleData Scientist                      5488.50   12239.86   0.448
## Job.TitleDigital Content Producer          -16717.92   12105.82  -1.381
## Job.TitleDigital Marketing Manager          -9205.45   12306.83  -0.748
## Job.TitleDirector                           53819.57   12830.13   4.195
## Job.TitleDirector of Business Development   36431.05   12572.35   2.898
## Job.TitleDirector of Engineering            43366.58   11046.99   3.926
## Job.TitleDirector of Finance                37009.56   11106.82   3.332
## Job.TitleDirector of HR                     43535.38   12648.41   3.442
## Job.TitleDirector of Human Capital          39809.47   12708.31   3.133
## Job.TitleDirector of Human Resources        36828.56   11228.69   3.280
## Job.TitleDirector of Marketing              40732.02    9633.72   4.228
## Job.TitleDirector of Operations             32177.55    9668.33   3.328
## Job.TitleDirector of Product Management     36200.61   12687.56   2.853
## Job.TitleDirector of Sales                  45873.83   12574.11   3.648
## Job.TitleDirector of Sales and Marketing    27016.35   12878.73   2.098
## Job.TitleEvent Coordinator                 -34710.46   10524.62  -3.298
```

```
## Job.TitleFinancial Advisor                        4680.11   12354.91   0.379
## Job.TitleFinancial Analyst                      -18168.70   12040.56  -1.509
## Job.TitleFinancial Manager                       28786.55   12326.64   2.335
## Job.TitleGraphic Designer                       -14496.53   12122.95  -1.196
## Job.TitleHelp Desk Analyst                      -16927.63   12083.50  -1.401
## Job.TitleHR Generalist                          -25414.55   10604.32  -2.397
## Job.TitleHR Manager                              -5871.70   10726.13  -0.547
## Job.TitleHuman Resources Director                43535.38   12648.41   3.442
## Job.TitleIT Manager                              -7162.83   12489.72  -0.573
## Job.TitleIT Support                             -16043.64   12043.50  -1.332
## Job.TitleIT Support Specialist                  -26894.62   12043.77  -2.233
## Job.TitleJunior Account Manager                 -16774.61   10444.06  -1.606
## Job.TitleJunior Accountant                      -19695.24    9843.93  -2.001
## Job.TitleJunior Advertising Coordinator         -22548.17   12015.04  -1.877
## Job.TitleJunior Business Analyst                -19397.85    9034.57  -2.147
## Job.TitleJunior Business Development Associate -22741.00    9145.35  -2.487
## Job.TitleJunior Business Operations Analyst     -45394.99   10439.30  -4.348
## Job.TitleJunior Copywriter                      -21717.92   12105.82  -1.794
## Job.TitleJunior Customer Support Specialist     -22158.08   12092.12  -1.832
## Job.TitleJunior Data Analyst                    -26894.62   12043.77  -2.233
## Job.TitleJunior Data Scientist                  -14760.29   12362.35  -1.194
## Job.TitleJunior Designer                        -22992.00   12120.22  -1.897
## Job.TitleJunior Developer                       -11210.77   12159.48  -0.922
## Job.TitleJunior Financial Advisor                -8495.47   12019.01  -0.707
## Job.TitleJunior Financial Analyst               -18249.74    9126.05  -2.000
## Job.TitleJunior HR Coordinator                  -22828.61   10511.06  -2.172
## Job.TitleJunior HR Generalist                   -20385.31   10434.32  -1.954
## Job.TitleJunior Marketing Analyst               -15171.98    9953.84  -1.524
## Job.TitleJunior Marketing Coordinator           -20824.81    9232.46  -2.256
## Job.TitleJunior Marketing Manager               -12426.27    9908.22  -1.254
## Job.TitleJunior Marketing Specialist            -15504.37    9341.21  -1.660
## Job.TitleJunior Operations Analyst              -19634.00    9330.25  -2.104
## Job.TitleJunior Operations Coordinator          -25213.39   12123.82  -2.080
## Job.TitleJunior Operations Manager              -15458.03    9866.44  -1.567
## Job.TitleJunior Product Manager                 -14827.17    9507.68  -1.559
## Job.TitleJunior Project Manager                 -19003.93    9361.65  -2.030
## Job.TitleJunior Recruiter                       -19496.53   12122.95  -1.608
## Job.TitleJunior Research Scientist              -11981.68   12350.12  -0.970
## Job.TitleJunior Sales Representative            -23774.41    9555.10  -2.488
## Job.TitleJunior Social Media Manager            -21717.92   12105.82  -1.794
## Job.TitleJunior Social Media Specialist         -23939.31   12107.26  -1.977
## Job.TitleJunior Software Developer              -23937.47   10408.43  -2.300
## Job.TitleJunior Software Engineer               -21600.86   12030.47  -1.796
## Job.TitleJunior UX Designer                     -16981.68   12350.12  -1.375
## Job.TitleJunior Web Designer                    -20326.78   12020.05  -1.691
## Job.TitleJunior Web Developer                   -27221.39   12014.32  -2.266
## Job.TitleMarketing Analyst                      -12074.51   10414.35  -1.159
## Job.TitleMarketing Coordinator                  -16980.47    9939.61  -1.708
## Job.TitleMarketing Manager                        6415.08   12273.65   0.523
## Job.TitleMarketing Specialist                   -32242.12   12136.26  -2.657
## Job.TitleNetwork Engineer                        -7548.17   12015.04  -0.628
## Job.TitleOffice Manager                         -41986.13   12423.73  -3.380
## Job.TitleOperations Analyst                     -16445.96   12657.75  -1.299
## Job.TitleOperations Director                     55714.18   12593.75   4.424
```

```
## Job.TitleOperations Manager                     13530.90   10932.12   1.238
## Job.TitlePrincipal Engineer                     23637.94   12786.88   1.849
## Job.TitlePrincipal Scientist                    13835.36   12514.29   1.106
## Job.TitleProduct Designer                        7089.36   12224.52   0.580
## Job.TitleProduct Manager                        12520.99   10738.67   1.166
## Job.TitleProduct Marketing Manager               4637.52   12223.00   0.379
## Job.TitleProject Engineer                          60.63   12355.68   0.005
## Job.TitleProject Manager                        10852.04   10576.57   1.026
## Job.TitlePublic Relations Manager               -9205.45   12306.83  -0.748
## Job.TitleRecruiter                             -28910.03   10547.47  -2.741
## Job.TitleResearch Director                      56320.21   12665.88   4.447
## Job.TitleResearch Scientist                     12850.28   12931.27   0.994
## Job.TitleSales Associate                       -25280.45   10492.64  -2.409
## Job.TitleSales Director                         37216.64   12678.74   2.935
## Job.TitleSales Executive                        23163.95   12610.19   1.837
## Job.TitleSales Manager                           -205.24    9904.03  -0.021
## Job.TitleSales Operations Manager              -13436.91   12424.07  -1.082
## Job.TitleSales Representative                  -21927.63   12083.50  -1.815
## Job.TitleSenior Account Executive               -5251.17   12200.92  -0.430
## Job.TitleSenior Account Manager                  4518.39   12316.62   0.367
## Job.TitleSenior Accountant                      -5957.67   10453.99  -0.570
## Job.TitleSenior Business Analyst                13821.24    9072.19   1.523
## Job.TitleSenior Business Development Manager    19050.15    9546.08   1.996
## Job.TitleSenior Consultant                      13097.77   12602.91   1.039
## Job.TitleSenior Data Analyst                    17331.74   10626.39   1.631
## Job.TitleSenior Data Engineer                   22846.03   10259.16   2.227
## Job.TitleSenior Data Scientist                  25908.60    9897.14   2.618
## Job.TitleSenior Engineer                         4962.56   11042.70   0.449
## Job.TitleSenior Financial Advisor                9031.55   10168.65   0.888
## Job.TitleSenior Financial Analyst               15036.26    9203.03   1.634
## Job.TitleSenior Financial Manager               12228.22    9390.55   1.302
## Job.TitleSenior Graphic Designer                -5613.64   12413.23  -0.452
## Job.TitleSenior HR Generalist                    -479.54   12299.27  -0.039
## Job.TitleSenior HR Manager                      19469.34   10374.37   1.877
## Job.TitleSenior HR Specialist                   19943.57   12440.79   1.603
## Job.TitleSenior Human Resources Coordinator    -15020.72   12131.35  -1.238
## Job.TitleSenior Human Resources Manager         18506.10   10816.34   1.711
## Job.TitleSenior Human Resources Specialist      11838.19   12349.94   0.959
## Job.TitleSenior IT Consultant                   11468.82   10744.16   1.067
## Job.TitleSenior IT Project Manager              17023.97   12205.96   1.395
## Job.TitleSenior IT Support Specialist           -5197.42   12232.76  -0.425
## Job.TitleSenior Manager                         22952.73   11170.21   2.055
## Job.TitleSenior Marketing Analyst                -730.81    9074.38  -0.081
## Job.TitleSenior Marketing Coordinator           -1200.08    9906.42  -0.121
## Job.TitleSenior Marketing Director              34531.50   12751.95   2.708
## Job.TitleSenior Marketing Manager               14279.67    9267.76   1.541
## Job.TitleSenior Marketing Specialist             8140.12    9536.15   0.854
## Job.TitleSenior Operations Analyst               -425.49   10414.35  -0.041
## Job.TitleSenior Operations Coordinator          12282.08    9513.64   1.291
## Job.TitleSenior Operations Manager              14458.86    9362.15   1.544
## Job.TitleSenior Product Designer                17529.38   10092.79   1.737
## Job.TitleSenior Product Development Manager     19769.56   12028.74   1.644
## Job.TitleSenior Product Manager                 25615.05    9263.61   2.765
## Job.TitleSenior Product Marketing Manager        3669.49   12411.03   0.296
```

```
## Job.TitleSenior Project Coordinator            -5536.76    9418.50  -0.588
## Job.TitleSenior Project Manager                15449.64    9375.04   1.648
## Job.TitleSenior Quality Assurance Analyst      14539.12   12099.86   1.202
## Job.TitleSenior Research Scientist             12850.28   12931.27   0.994
## Job.TitleSenior Researcher                     11089.77   12699.72   0.873
## Job.TitleSenior Sales Manager                   7952.39   10864.82   0.732
## Job.TitleSenior Sales Representative           -6361.86   10609.37  -0.600
## Job.TitleSenior Scientist                       6553.84   10436.66   0.628
## Job.TitleSenior Software Architect             23690.22   12259.31   1.932
## Job.TitleSenior Software Developer             16793.11   10110.16   1.661
## Job.TitleSenior Software Engineer              16147.42    9340.47   1.729
## Job.TitleSenior Training Specialist            -6657.29   12336.68  -0.540
## Job.TitleSenior UX Designer                    23392.43   10573.95   2.212
## Job.TitleSocial Media Manager                 -13222.44   12115.71  -1.091
## Job.TitleSocial Media Specialist              -17275.14   12158.59  -1.421
## Job.TitleSoftware Developer                     5488.50   12239.86   0.448
## Job.TitleSoftware Engineer                     15000.00   12004.95   1.249
## Job.TitleSoftware Manager                      11511.41   12311.06   0.935
## Job.TitleSoftware Project Manager               2874.94   12045.91   0.239
## Job.TitleStrategy Consultant                   18786.55   12326.64   1.524
## Job.TitleSupply Chain Analyst                  -9731.72   12727.28  -0.765
## Job.TitleSupply Chain Manager                 -13994.13   12514.67  -1.118
## Job.TitleTechnical Recruiter                  -18399.14   12030.47  -1.529
## Job.TitleTechnical Support Specialist         -21600.86   12030.47  -1.796
## Job.TitleTechnical Writer                     -14379.47   12052.00  -1.193
## Job.TitleTraining Specialist                  -35251.17   12200.92  -2.889
## Job.TitleUX Designer                           -2797.26   12362.45  -0.226
## Job.TitleUX Researcher                          2266.05   12341.89   0.184
## Job.TitleVP of Finance                         66431.05   12572.35   5.284
## Job.TitleVP of Operations                      56431.05   12572.35   4.489
## Job.TitleWeb Developer                        -13725.92   12007.47  -1.143
## Years.of.Experience                            1504.53     581.33   2.588
##                                              Pr(>|t|)
## (Intercept)                                  0.775255
## Age                                          5.32e-06 ***
## GenderMale                                   0.567134
## Education.LevelMaster's                      0.076468 .
## Education.LevelPhD                           0.000401 ***
## Job.TitleAccountant                          0.203224
## Job.TitleAdministrative Assistant            8.01e-05 ***
## Job.TitleBusiness Analyst                    0.702196
## Job.TitleBusiness Development Manager        0.879148
## Job.TitleBusiness Intelligence Analyst       0.961834
## Job.TitleCEO                                 1.00e-13 ***
## Job.TitleChief Data Officer                  3.85e-11 ***
## Job.TitleChief Technology Officer            1.14e-10 ***
## Job.TitleContent Marketing Manager           0.334271
## Job.TitleCopywriter                          0.088568 .
## Job.TitleCreative Director                   0.634890
## Job.TitleCustomer Service Manager            0.174912
## Job.TitleCustomer Service Rep                0.109464
## Job.TitleCustomer Service Representative     0.027256 *
## Job.TitleCustomer Success Manager            0.074128 .
## Job.TitleCustomer Success Rep                0.225213
```

```
## Job.TitleData Analyst                              0.138601
## Job.TitleData Entry Clerk                          0.188701
## Job.TitleData Scientist                            0.654356
## Job.TitleDigital Content Producer                  0.168873
## Job.TitleDigital Marketing Manager                 0.455369
## Job.TitleDirector                                  4.15e-05 ***
## Job.TitleDirector of Business Development          0.004191 **
## Job.TitleDirector of Engineering                   0.000120 ***
## Job.TitleDirector of Finance                       0.001032 **
## Job.TitleDirector of HR                            0.000707 ***
## Job.TitleDirector of Human Capital                 0.002001 **
## Job.TitleDirector of Human Resources               0.001231 **
## Job.TitleDirector of Marketing                     3.63e-05 ***
## Job.TitleDirector of Operations                    0.001046 **
## Job.TitleDirector of Product Management            0.004797 **
## Job.TitleDirector of Sales                         0.000339 ***
## Job.TitleDirector of Sales and Marketing           0.037222 *
## Job.TitleEvent Coordinator                         0.001158 **
## Job.TitleFinancial Advisor                         0.705246
## Job.TitleFinancial Analyst                         0.132936
## Job.TitleFinancial Manager                         0.020549 *
## Job.TitleGraphic Designer                          0.233237
## Job.TitleHelp Desk Analyst                         0.162845
## Job.TitleHR Generalist                             0.017496 *
## Job.TitleHR Manager                                0.584719
## Job.TitleHuman Resources Director                  0.000707 ***
## Job.TitleIT Manager                                0.566972
## Job.TitleIT Support                                0.184377
## Job.TitleIT Support Specialist                     0.026687 *
## Job.TitleJunior Account Manager                    0.109870
## Job.TitleJunior Accountant                         0.046813 *
## Job.TitleJunior Advertising Coordinator            0.062066 .
## Job.TitleJunior Business Analyst                   0.033029 *
## Job.TitleJunior Business Development Associate     0.013741 *
## Job.TitleJunior Business Operations Analyst        2.21e-05 ***
## Job.TitleJunior Copywriter                         0.074369 .
## Job.TitleJunior Customer Support Specialist        0.068419 .
## Job.TitleJunior Data Analyst                       0.026687 *
## Job.TitleJunior Data Scientist                     0.233947
## Job.TitleJunior Designer                           0.059314 .
## Job.TitleJunior Developer                          0.357685
## Job.TitleJunior Financial Advisor                  0.480516
## Job.TitleJunior Financial Analyst                  0.046923 *
## Job.TitleJunior HR Coordinator                     0.031079 *
## Job.TitleJunior HR Generalist                      0.052177 .
## Job.TitleJunior Marketing Analyst                  0.129079
## Job.TitleJunior Marketing Coordinator              0.025212 *
## Job.TitleJunior Marketing Manager                  0.211301
## Job.TitleJunior Marketing Specialist               0.098574 .
## Job.TitleJunior Operations Analyst                 0.036637 *
## Job.TitleJunior Operations Coordinator             0.038871 *
## Job.TitleJunior Operations Manager                 0.118808
## Job.TitleJunior Product Manager                    0.120509
## Job.TitleJunior Project Manager                    0.043724 *
```

```
## Job.TitleJunior Recruiter                       0.109410
## Job.TitleJunior Research Scientist              0.333171
## Job.TitleJunior Sales Representative            0.013684 *
## Job.TitleJunior Social Media Manager            0.074369 .
## Job.TitleJunior Social Media Specialist         0.049428 *
## Job.TitleJunior Software Developer              0.022523 *
## Job.TitleJunior Software Engineer               0.074128 .
## Job.TitleJunior UX Designer                     0.170711
## Job.TitleJunior Web Designer                    0.092429 .
## Job.TitleJunior Web Developer                   0.024571 *
## Job.TitleMarketing Analyst                      0.247714
## Job.TitleMarketing Coordinator                  0.089168 .
## Job.TitleMarketing Manager                      0.601800
## Job.TitleMarketing Specialist                   0.008549 **
## Job.TitleNetwork Engineer                       0.530594
## Job.TitleOffice Manager                         0.000878 ***
## Job.TitleOperations Analyst                     0.195390
## Job.TitleOperations Director                    1.61e-05 ***
## Job.TitleOperations Manager                     0.217316
## Job.TitlePrincipal Engineer                     0.066036 .
## Job.TitlePrincipal Scientist                    0.270284
## Job.TitleProduct Designer                       0.562635
## Job.TitleProduct Manager                        0.245057
## Job.TitleProduct Marketing Manager              0.704799
## Job.TitleProject Engineer                       0.996090
## Job.TitleProject Manager                        0.306148
## Job.TitlePublic Relations Manager               0.455369
## Job.TitleRecruiter                              0.006699 **
## Job.TitleResearch Director                      1.47e-05 ***
## Job.TitleResearch Scientist                     0.321589
## Job.TitleSales Associate                        0.016915 *
## Job.TitleSales Director                         0.003734 **
## Job.TitleSales Executive                        0.067751 .
## Job.TitleSales Manager                          0.983488
## Job.TitleSales Operations Manager               0.280807
## Job.TitleSales Representative                   0.071118 .
## Job.TitleSenior Account Executive               0.667389
## Job.TitleSenior Account Manager                 0.714128
## Job.TitleSenior Accountant                      0.569409
## Job.TitleSenior Business Analyst                0.129269
## Job.TitleSenior Business Development Manager    0.047377 *
## Job.TitleSenior Consultant                      0.299975
## Job.TitleSenior Data Analyst                    0.104511
## Job.TitleSenior Data Engineer                   0.027104 *
## Job.TitleSenior Data Scientist                  0.009548 **
## Job.TitleSenior Engineer                        0.653647
## Job.TitleSenior Financial Advisor               0.375546
## Job.TitleSenior Financial Analyst               0.103914
## Job.TitleSenior Financial Manager               0.194397
## Job.TitleSenior Graphic Designer                0.651608
## Job.TitleSenior HR Generalist                   0.968939
## Job.TitleSenior HR Manager                      0.062064 .
## Job.TitleSenior HR Specialist                   0.110544
## Job.TitleSenior Human Resources Coordinator     0.217148
```

```
## Job.TitleSenior Human Resources Manager        0.088691 .
## Job.TitleSenior Human Resources Specialist     0.338972
## Job.TitleSenior IT Consultant                  0.287096
## Job.TitleSenior IT Project Manager             0.164693
## Job.TitleSenior IT Support Specialist          0.671397
## Job.TitleSenior Manager                        0.041237 *
## Job.TitleSenior Marketing Analyst              0.935894
## Job.TitleSenior Marketing Coordinator          0.903704
## Job.TitleSenior Marketing Director             0.007375 **
## Job.TitleSenior Marketing Manager              0.124997
## Job.TitleSenior Marketing Specialist           0.394376
## Job.TitleSenior Operations Analyst             0.967453
## Job.TitleSenior Operations Coordinator         0.198241
## Job.TitleSenior Operations Manager             0.124122
## Job.TitleSenior Product Designer               0.084006 .
## Job.TitleSenior Product Development Manager    0.101894
## Job.TitleSenior Product Manager                0.006239 **
## Job.TitleSenior Product Marketing Manager      0.767803
## Job.TitleSenior Project Coordinator            0.557310
## Job.TitleSenior Project Manager                0.100981
## Job.TitleSenior Quality Assurance Analyst      0.230986
## Job.TitleSenior Research Scientist             0.321589
## Job.TitleSenior Researcher                     0.383618
## Job.TitleSenior Sales Manager                  0.465089
## Job.TitleSenior Sales Representative           0.549442
## Job.TitleSenior Scientist                      0.530766
## Job.TitleSenior Software Architect             0.054763 .
## Job.TitleSenior Software Developer             0.098326 .
## Job.TitleSenior Software Engineer              0.085443 .
## Job.TitleSenior Training Specialist            0.590069
## Job.TitleSenior UX Designer                    0.028115 *
## Job.TitleSocial Media Manager                  0.276474
## Job.TitleSocial Media Specialist               0.156975
## Job.TitleSoftware Developer                    0.654356
## Job.TitleSoftware Engineer                     0.212993
## Job.TitleSoftware Manager                      0.350927
## Job.TitleSoftware Project Manager              0.811617
## Job.TitleStrategy Consultant                   0.129122
## Job.TitleSupply Chain Analyst                  0.445418
## Job.TitleSupply Chain Manager                  0.264856
## Job.TitleTechnical Recruiter                   0.127800
## Job.TitleTechnical Support Specialist          0.074128 .
## Job.TitleTechnical Writer                      0.234280
## Job.TitleTraining Specialist                   0.004301 **
## Job.TitleUX Designer                           0.821229
## Job.TitleUX Researcher                         0.854514
## Job.TitleVP of Finance                         3.38e-07 ***
## Job.TitleVP of Operations                      1.23e-05 ***
## Job.TitleWeb Developer                         0.254400
## Years.of.Experience                            0.010381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8489 on 194 degrees of freedom
```

```
## Multiple R-squared:  0.9839, Adjusted R-squared:  0.969
## F-statistic:  66.4 on 178 and 194 DF,  p-value: < 2.2e-16
```

```
# Sample Data Creation
# Let's assume you have a column 'Job.Title' and 'Actual_Salary' in your dataset
data <- data.frame(
  Job.Title = c("Senior", "CEO", "Junior", "HR", "Manager"),
  Actual_Salary = c(50000, 150000, 45000, 70000, 60000)
)

# Categorizing job titles
data$Job.Category <- ifelse(grepl("Senior", data$Job.Title), "Senior",
                        ifelse(grepl("CEO|Executive", data$Job.Title), "CEO",
                        ifelse(grepl("Junior", data$Job.Title), "Junior",
                        ifelse(grepl("HR", data$Job.Title), "HR",
                        ifelse(grepl("Manager", data$Job.Title), "Manager", "Other")))))

# Fit the model (using a hypothetical cleaned data)
# model <- lm(Actual_Salary ~ Job.Category, data = data)

# For demonstration, let's create some predicted salaries based on arbitrary coefficients
data$Predicted_Salary <- with(data, ifelse(Job.Category == "Senior", 55000,
                                    ifelse(Job.Category == "CEO", 155000,
                                    ifelse(Job.Category == "Junior", 40000,
                                    ifelse(Job.Category == "HR", 65000,
                                    ifelse(Job.Category == "Manager", 65000, 50000))))))

# Calculate the error
data$Error <- with(data, Actual_Salary - Predicted_Salary)

# Print the data frame
print(data[, c("Job.Title", "Actual_Salary", "Predicted_Salary", "Error")])
```

```
##   Job.Title Actual_Salary Predicted_Salary Error
## 1    Senior         50000            55000 -5000
## 2       CEO        150000           155000 -5000
## 3    Junior         45000            40000  5000
## 4        HR         70000            65000  5000
## 5   Manager         60000            65000 -5000
```

Now, let us remove the `Job.Title` from the full model as it contains many categories and let us see how much variance is captured without the `Job.Title` in the next model (simple_model)

```
simple_model <- lm(Salary ~ Age + Years.of.Experience + Education.Level + Gender, data = cleaned_data)
summary(simple_model)
```

```
##
## Call:
## lm(formula = Salary ~ Age + Years.of.Experience + Education.Level +
##     Gender, data = cleaned_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47335  -7050   -481   8495  74302
##
## Coefficients:
```

```
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -48720.0    15063.0  -3.234  0.00133 **
## Age                        2880.3      554.2   5.197 3.36e-07 ***
## Years.of.Experience        2873.5      613.9   4.681 4.03e-06 ***
## Education.LevelMaster's   18404.9     2088.1   8.814  < 2e-16 ***
## Education.LevelPhD        24635.4     2797.7   8.806  < 2e-16 ***
## GenderMale                 8566.1     1582.9   5.412 1.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15130 on 367 degrees of freedom
## Multiple R-squared:  0.903,  Adjusted R-squared:  0.9017
## F-statistic: 683.1 on 5 and 367 DF,  p-value: < 2.2e-16
```

Now, this model will be our first model since it captures maximum variance without the Job.title column and can be helpful in salary determining.

Intepretation:

1. Model Formula: The linear model predicts `Salary` using `Age`, `Years of Experience`, `Education Level`, and `Gender`.
2. Coefficients: Significant predictors include `Age`, `Years of Experience`, and higher levels of `Education`, with higher education and age associated with increased salary. Being male also predicts a higher salary.
3. Model Fit: The model explains approximately 90.3% of the variance in salaries, indicating a strong fit.
4. Statistical Significance: The model is statistically significant (F-statistic: 683.1, p < 2.2e-16), affirming the reliability of the predictors.

```r
# Log transformations
log_model <- lm(log(Salary) ~ Age + Years.of.Experience + Education.Level + Gender, data = cleaned_data)
summary(log_model)
```

```
##
## Call:
## lm(formula = log(Salary) ~ Age + Years.of.Experience + Education.Level +
##     Gender, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9327 -0.1280 -0.0031  0.1393  0.4445
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               9.47930    0.31552  30.043  < 2e-16 ***
## Age                       0.04198    0.01161   3.616 0.000341 ***
## Years.of.Experience       0.02182    0.01286   1.697 0.090512 .
## Education.LevelMaster's   0.20142    0.04374   4.605  5.7e-06 ***
## Education.LevelPhD        0.20106    0.05860   3.431 0.000670 ***
## GenderMale                0.06118    0.03316   1.845 0.065838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3169 on 367 degrees of freedom
## Multiple R-squared:  0.7159, Adjusted R-squared:  0.7121
## F-statistic:   185 on 5 and 367 DF,  p-value: < 2.2e-16
```

Interpretation:

1. Model Formula: The model uses the natural logarithm of `Salary` as a dependent variable, regressed on `Age`, `Years of Experience`, `Education Level`, and `Gender`.
2. Coefficients and Significance: Positive coefficients for `Age`, `Years of Experience`, both higher education levels, and `GenderMale` indicate their respective contributions to higher salary logarithm values, with all but `Years of Experience` and `GenderMale` showing strong statistical significance.
3. Model Fit: The model accounts for approximately 71.59% of the variability in the logarithmic salary (Adjusted R-squared: 0.7121), suggesting a good fit.
4. Statistical Significance of Model: With an F-statistic of 185 and a highly significant p-value ($p < 2.2e\text{-}16$), the overall model fit is statistically significant, validating the effectiveness of these predictors in explaining salary variations.

```
poly_model <- lm(poly(Salary,2) ~ Age + Years.of.Experience + Education.Level + Gender, data = cleaned_
summary(poly_model)
```

```
## Response 1 :
##
## Call:
## lm(formula = `1` ~ Age + Years.of.Experience + Education.Level +
##     Gender, data = cleaned_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.050875 -0.007577 -0.000517  0.009130  0.079858
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.1604624  0.0161895  -9.912  < 2e-16 ***
## Age                       0.0030957  0.0005956   5.197 3.36e-07 ***
## Years.of.Experience       0.0030883  0.0006598   4.681 4.03e-06 ***
## Education.LevelMaster's   0.0197813  0.0022443   8.814  < 2e-16 ***
## Education.LevelPhD        0.0264777  0.0030069   8.806  < 2e-16 ***
## GenderMale                0.0092067  0.0017013   5.412 1.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01626 on 367 degrees of freedom
## Multiple R-squared:  0.903,  Adjusted R-squared:  0.9017
## F-statistic: 683.1 on 5 and 367 DF,  p-value: < 2.2e-16
##
##
## Response 2 :
##
## Call:
## lm(formula = `2` ~ Age + Years.of.Experience + Education.Level +
##     Gender, data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07595 -0.03851 -0.01233  0.03251  0.35760
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.152395   0.050981   2.989  0.00299 **
## Age                       -0.005574   0.001876  -2.972  0.00315 **
## Years.of.Experience        0.005409   0.002078   2.603  0.00962 **
```

```
## Education.LevelMaster's -0.006224   0.007067  -0.881  0.37909
## Education.LevelPhD       0.011298   0.009469   1.193  0.23358
## GenderMale               0.004046   0.005357   0.755  0.45060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0512 on 367 degrees of freedom
## Multiple R-squared:  0.03784,    Adjusted R-squared:  0.02473
## F-statistic: 2.887 on 5 and 367 DF,  p-value: 0.01431
```

Interpretation:

Response 1 Interpretation:
1. The model uses `Age`, `Years of Experience`, `Education Level`, and `Gender` to predict an unnamed variable (`1`), explaining 90.3% of its variance.
2. All predictors are statistically significant, with positive effects on the response variable, suggesting increasing age, experience, higher education, and being male are associated with higher values of `1`.
3. The model is highly statistically significant ($p < 2.2e\text{-}16$), indicating reliable predictors.

Response 2 Interpretation:
1. A different model uses the same predictors for another unnamed variable (`2`), but only explains 3.78% of its variance, indicating a weak model fit.
2. `Age` and `Years of Experience` significantly influence `2`, with age negatively affecting it and experience positively affecting it.
3. `Education Level` and `Gender` are not statistically significant, suggesting they do not impact the response variable in this model context.
4. Despite low explanatory power, the model overall is statistically significant ($p = 0.01431$), suggesting some relationship between the variables and response `2`.

```
interaction_model <- lm(Salary ~ Age * Years.of.Experience + Education.Level + Gender, data = cleaned_da
summary(interaction_model)
```

```
##
## Call:
## lm(formula = Salary ~ Age * Years.of.Experience + Education.Level +
##     Gender, data = cleaned_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46598  -7205     63   8471  75627
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -51071.86   15361.70  -3.325 0.000975 ***
## Age                        2920.73     556.81   5.245 2.64e-07 ***
## Years.of.Experience        3597.23    1102.68   3.262 0.001209 **
## Education.LevelMaster's   18413.18    2089.24   8.813  < 2e-16 ***
## Education.LevelPhD        24619.06    2799.22   8.795  < 2e-16 ***
## GenderMale                 8587.43    1583.94   5.422 1.07e-07 ***
## Age:Years.of.Experience     -15.29      19.34  -0.790 0.429839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15140 on 366 degrees of freedom
## Multiple R-squared:  0.9031, Adjusted R-squared:  0.9015
## F-statistic: 568.8 on 6 and 366 DF,  p-value: < 2.2e-16
```

Interpretation:

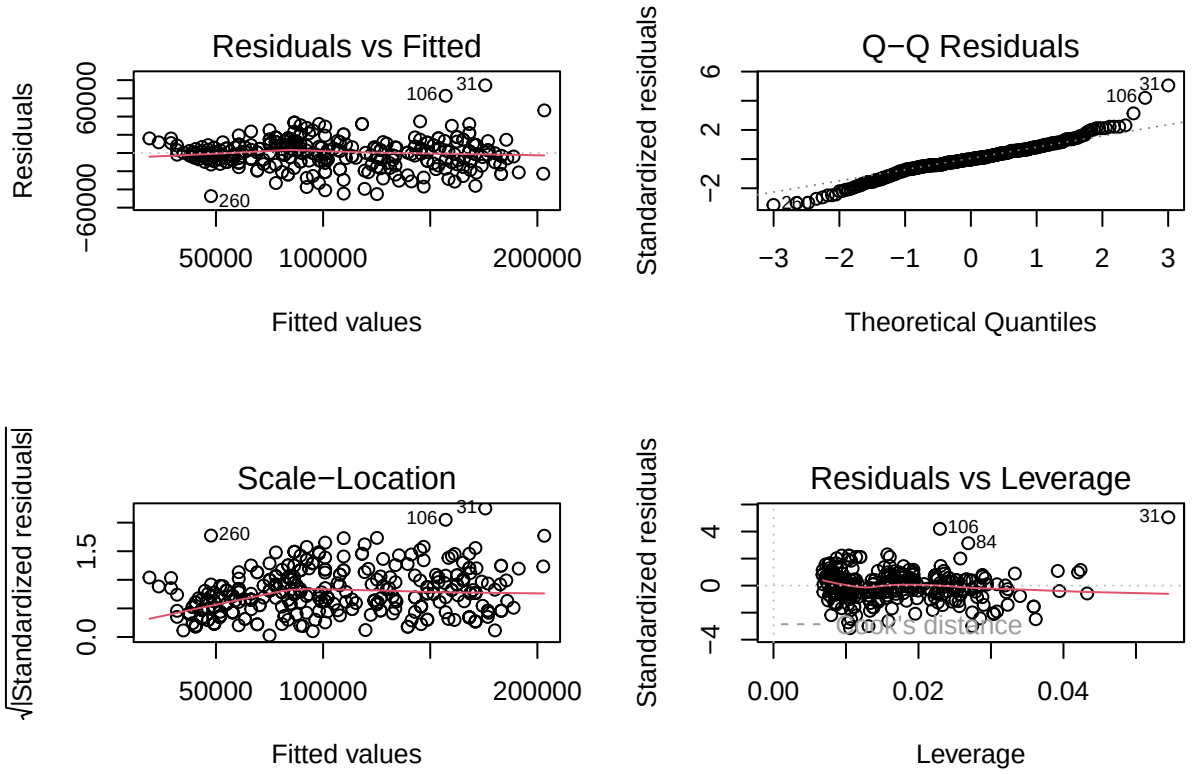
1. Model Details: The linear regression model predicts `Salary` using `Age`, `Years of Experience`, their interaction (`Age * Years of Experience`), `Education Level`, and `Gender`.

2. Significance and Coefficients: `Age`, `Years of Experience`, `Education Level` (both Master's and PhD), and `GenderMale` are significant predictors with positive coefficients, indicating their positive impact on `Salary`.

3. Interaction Term: The interaction term (`Age:Years.of.Experience`) is not significant ($p = 0.429839$), suggesting that the combined effect of age and years of experience does not significantly differ from their individual effects on salary.

4. Model Fit: The model explains 90.31% of the variability in `Salary` (Adjusted R-squared: 0.9015) and is statistically significant ($p < 2.2e\text{-}16$), indicating a strong fit to the data.

```
par(mfrow = c(2,2))
plot(simple_model)
```




```
# Assuming your encoded dataframe is named df_encoded

# Selecting features by dropping specific columns
X <- df_encoded[, !(names(df_encoded) %in% c("Job Title", "Salary", "Gender"))]

# Selecting the target variable
y <- df_encoded[["Salary"]]
```

```
head(X)
```

```
##   Age Education.Level         Job.Title Years.of.Experience
## 1  32      Bachelor's Software Engineer                    5
```

```
## 2  28        Master's      Data Analyst                   3
## 3  45             PhD    Senior Manager                  15
## 4  36      Bachelor's   Sales Associate                   7
## 5  52        Master's          Director                  20
## 6  29      Bachelor's Marketing Analyst                   2
##   Education.Level_Master's Education.Level_PhD
## 1                        0                   0
## 2                        1                   0
## 3                        0                   1
## 4                        0                   0
## 5                        1                   0
## 6                        0                   0
```

```r
# Load the caret package
library(caret)
```

```
## Loading required package: lattice
```

```r
library(lattice)
# Assuming your features and target variable are stored in X and y respectively

# Set seed for reproducibility
set.seed(90)

# Split the data
trainIndex <- createDataPartition(y, p = .8,
                                  list = FALSE,
                                  times = 1)

X_train <- X[trainIndex, ]
X_test <- X[-trainIndex, ]
y_train <- y[trainIndex]
y_test <- y[-trainIndex]
```

```r
# Assuming df_encoded is your dataframe from which you want to exclude 'Job.Title', 'Salary', and 'Gend
# and use 'Salary' as the target variable (y)

# Selecting features excluding 'Job Title', 'Salary', and 'Gender'
X <- subset(df_encoded, select = -c(Job.Title, Salary, Gender))

# Selecting the target variable 'Salary'
y <- df_encoded$Salary

# Set seed for reproducibility
set.seed(30)

# Define training control for 10-fold cross-validation
train_control1 <- trainControl(method = "cv", number = 10)

# Train the model using linear regression with 10-fold cross-validation
model1 <- train(x = X, y = y, method = "lm", trControl = train_control1)

# Print the results
print(model1)
```

```
## Linear Regression
```

```
##
## 373 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 335, 337, 336, 335, 336, 336, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   15658.61  0.8973575  11291.03
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Interpretation:

1. Dataset and Method: The model used 373 samples and 5 predictors, evaluated using 10-fold cross-validation, ensuring robust estimates by training on various subsets of the data.

2. Performance Metrics: The model achieved an RMSE (Root Mean Squared Error) of 15,658.61, an R-squared of 0.897, and an MAE (Mean Absolute Error) of 11,291.03, indicating strong predictive accuracy and consistency.

3. Model Consistency: High R-squared value suggests that about 89.7% of the variance in the dependent variable is predictable from the independent variables.

4. Tuning and Stability: The intercept was held constant, ensuring the model includes a baseline level from which the effects of predictors are measured, stabilizing comparisons and interpretations across different models.

```r
# Define training control for 10-fold cross-validation
train_control2 <- trainControl(method = "cv", number = 10)

# Train the model using linear regression with 10-fold cross-validation
model2 <- train(x = X, y = y, method = "rf", trControl = train_control2)

# Print the results
print(model2)
```

```
## Random Forest
##
## 373 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 336, 336, 336, 336, 335, 335, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     15166.94  0.9065717  10777.03
##   3     15065.59  0.9071382  10307.93
##   5     15484.23  0.9025122  10518.72
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 3.
```

Interpretation:

1. Model: The random forest model used 373 samples and 5 predictors, validated through 10-fold cross-validation to ensure robustness across different subsets of the data.
2. Tuning Parameter (mtry) Selection: The model was tested with different values of `mtry` (number of variables randomly sampled as candidates at each split): 2, 3, and 5. This parameter affects the complexity and potential overfitting of the model.
3. Performance Metrics: Among the tested mtry values, `mtry = 3` yielded the best results with an RMSE of 15,065.59, an R-squared of 0.9071, and an MAE of 10,307.93, indicating the highest model accuracy and prediction consistency compared to the other values.
4. Optimal Model Choice: The model with `mtry = 3` was selected as the optimal configuration based on having the lowest RMSE, signifying the best balance between model complexity and predictive accuracy. This model provided a good fit and explained approximately 90.71% of the variability in the target variable.

```r
# Create a data frame for Linear Regression results
linear_regression_results <- data.frame(
  Model = "Linear Regression",
  RMSE = 15658.61,
  Rsquared = 0.8973575,
  MAE = 11291.03
)

# Create a data frame for Random Forest results
random_forest_results <- data.frame(
  Model = "Random Forest",
  RMSE = 15065.59,  # Using the best RMSE corresponding to mtry = 3
  Rsquared = 0.9071382,  # Rsquared for mtry = 3
  MAE = 10307.93  # MAE for mtry = 3
)

# Combine the results into a single data frame
comparison_results <- rbind(linear_regression_results, random_forest_results)

# Print the results to compare the models
print(comparison_results)
```

```
##               Model     RMSE  Rsquared      MAE
## 1 Linear Regression 15658.61 0.8973575 11291.03
## 2     Random Forest 15065.59 0.9071382 10307.93
```

```r
# Make predictions on the test set
predicted_salary1 <- round(predict(model1, newdata = X_test))

# Create a data frame with Actual Salary, Predicted Salary, and Error
predicted_df1 <- data.frame(
  Actual_Salary = y_test,
  Predicted_Salary = predicted_salary1,
  Error = predicted_salary1 - y_test
)

# View the first few rows of the data frame
head(predicted_df1)
```

```
##    Actual_Salary Predicted_Salary Error
```

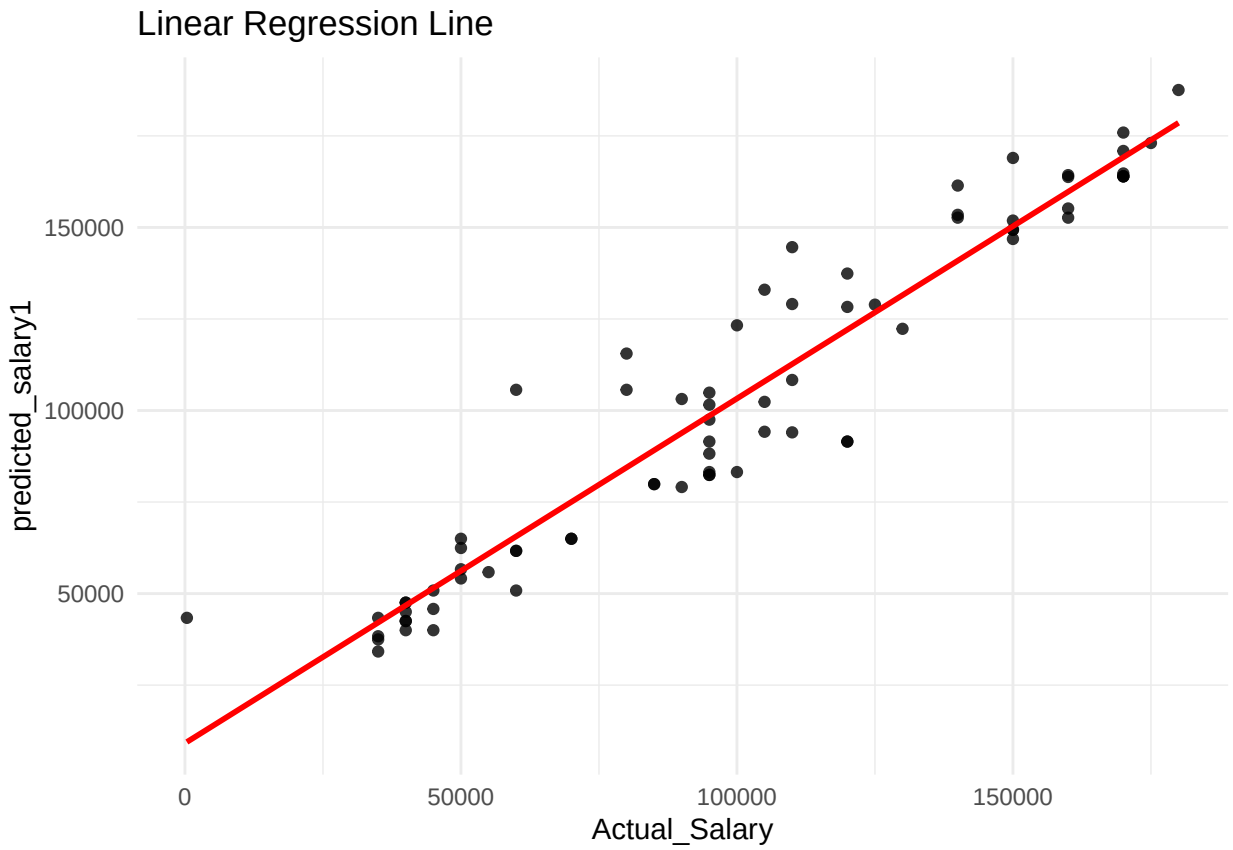
```
## 7          120000           128287  8287
## 16         125000           128905  3905
## 26          45000            45806   806
## 35         170000           170879   879
## 36          45000            39993 -5007
## 43          60000            50832 -9168
```

```r
# Make predictions on the test set
predicted_salary2 <- round(predict(model2, newdata = X_test))

# Create a data frame with Actual Salary, Predicted Salary, and Error
predicted_df2 <- data.frame(
  Actual_Salary = y_test,
  Predicted_Salary = predicted_salary2,
  Error = predicted_salary2 - y_test
)

# View the first few rows of the data frame
head(predicted_df2)
```

```
##    Actual_Salary Predicted_Salary  Error
## 7         120000           118748  -1252
## 16        125000           126267   1267
## 26         45000            46730   1730
## 35        170000           163362  -6638
## 36         45000            42036  -2964
## 43         60000            49672 -10328
```

```r
ggplot(predicted_df1, aes(x = Actual_Salary, y = predicted_salary1)) +
  geom_point(alpha = 0.8, color = "black") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  ggtitle("Linear Regression Line") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Load necessary libraries
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

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

```
library(caret) # For creating dummy variables and train/test splits

suppressPackageStartupMessages(library(randomForest))

# Assuming df_encoded is your dataframe and it contains 'Job.Title', 'Salary', 'Gender' among other fea

# Convert 'Job.Title' to a factor if it's not already
df_encoded$Job.Title <- as.factor(df_encoded$Job.Title)

# Create dummy variables for all categorical features, excluding 'Gender' and 'Salary'
df_dummies <- dummyVars(Salary ~ ., data = df_encoded, fullRank = TRUE)
df_transformed <- data.frame(predict(df_dummies, newdata = df_encoded))
```

```r
# Ensure 'Salary' is not included in the transformed data frame
df_transformed <- df_transformed[, !(names(df_transformed) %in% c("Salary", "X.Education.Level_Master.s
#print(names(df_transformed))

# Selecting the target variable 'Salary'
y <- df_encoded$Salary

# Set seed for reproducibility
set.seed(30)

# Split the data into training and testing sets
trainIndex <- createDataPartition(y, p = .8, list = FALSE, times = 1)
X_train <- df_transformed[trainIndex, ]
X_test <- df_transformed[-trainIndex, ]
y_train <- y[trainIndex]
y_test <- y[-trainIndex]

# Train a Random Forest model
rf_model <- randomForest(x = X_train, y = y_train)

# Print the model summary
print(rf_model)
```

```
##
## Call:
##  randomForest(x = X_train, y = y_train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 59
##
##           Mean of squared residuals: 221682732
##                     % Var explained: 90.45
```

```r
# Predict on the test set
predictions <- predict(rf_model, X_test)
```

```r
# metrics for the test data
test_results <- postResample(pred = predictions, obs = y_test)
print(round(test_results,3))
```

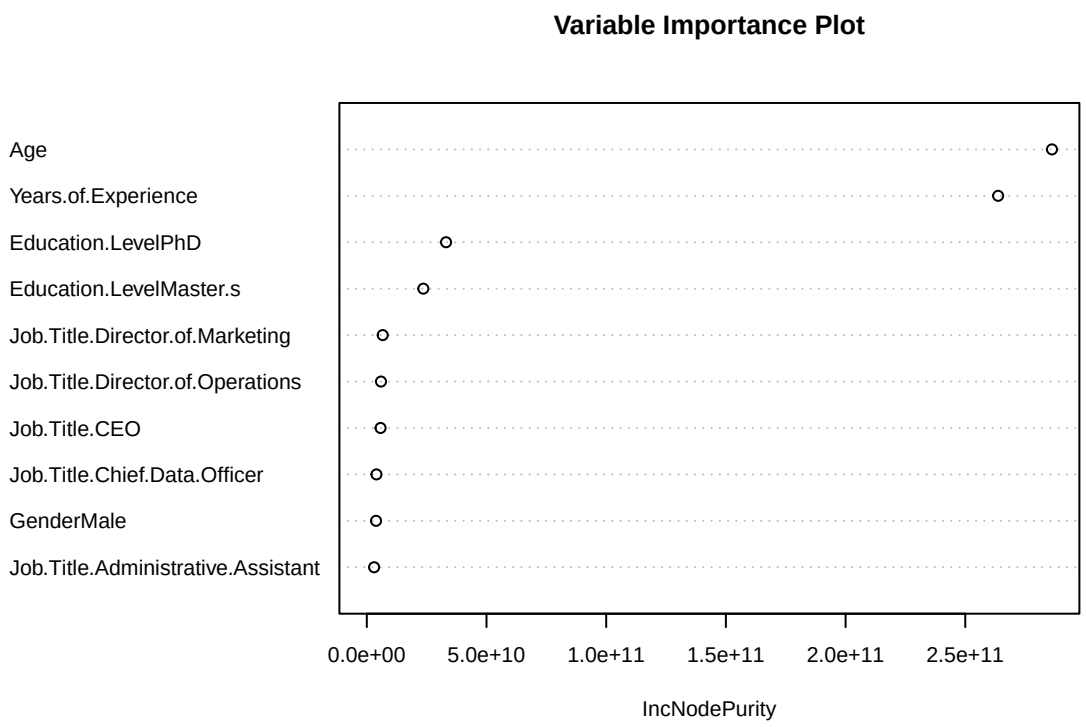
```
##      RMSE  Rsquared       MAE
## 14197.533     0.916  8840.320
```

```r
# Create a data frame for Linear Regression results
linear_regression_testing <- data.frame(
  Model = "Linear Regression",
  RMSE = 14352.312,
  Rsquared = 0.8825102,
  MAE = 9470.211
)

# Create a data frame for Random Forest results
random_forest_testing <- data.frame(
  Model = "Random Forest",
  RMSE = 14197.533,  # Using the best RMSE corresponding to mtry = 3
```

```
  Rsquared = 0.9162342,  # Rsquared for mtry = 3
  MAE = 8840.320  # MAE for mtry = 3
)

# Combine the results into a single data frame
testing_results <- rbind(linear_regression_testing, random_forest_testing)

# Print the results to compare the models
print(testing_results)
```

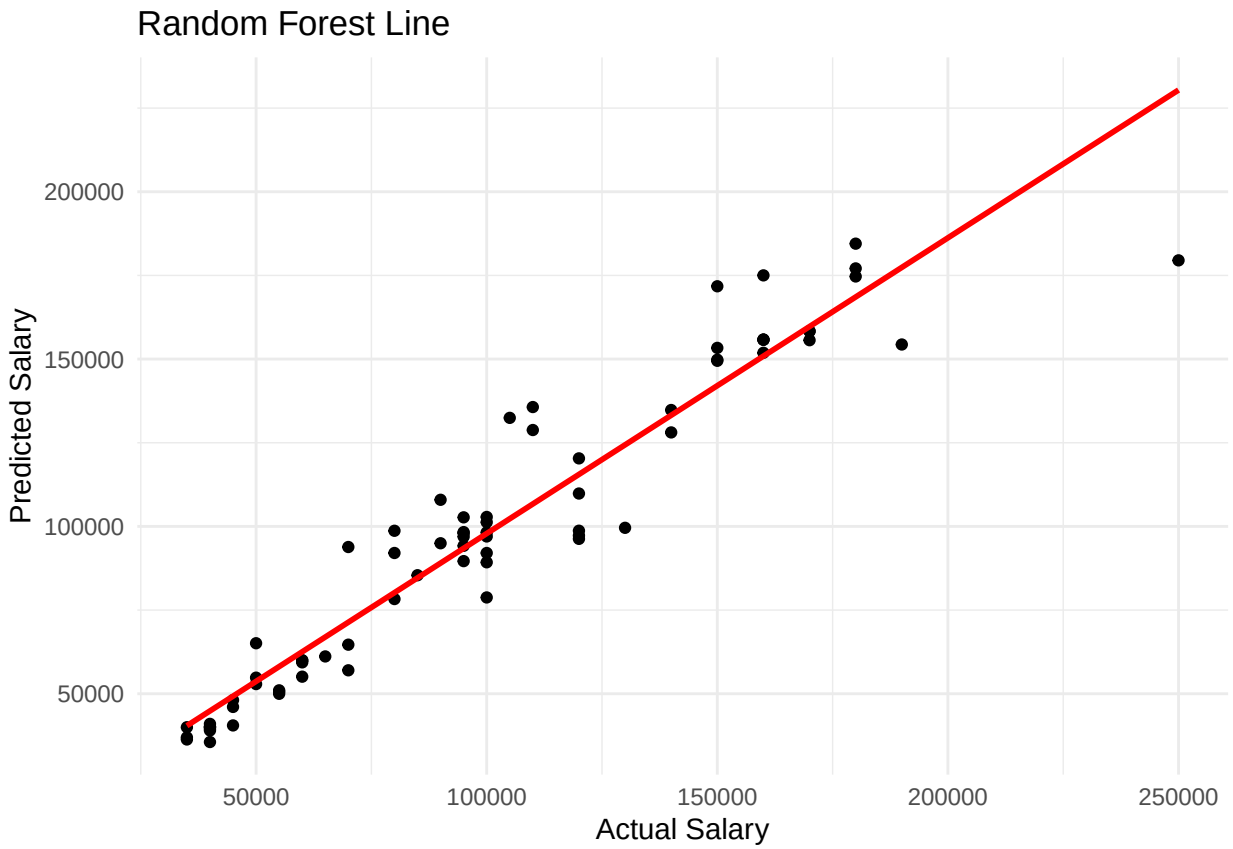
```
##               Model     RMSE  Rsquared      MAE
## 1 Linear Regression 14352.31 0.8825102 9470.211
## 2     Random Forest 14197.53 0.9162342 8840.320
```

```
# Plot variable importance
varImpPlot(rf_model, main = "Variable Importance Plot", cex = 0.7, n.var = 10)
```




```
p <- predict(rf_model, newdata = X_test)

p_df <- data.frame(Actual_Salary = y_test, Predicted_Salary = p)

library(ggplot2)

ggplot(p_df, aes(x = Actual_Salary, y = Predicted_Salary)) +
  geom_point() +
  geom_smooth(method = "lm",se = FALSE, color = "red") +
  ggtitle("Random Forest Line") +
  theme_minimal() +
  xlab("Actual Salary") +
  ylab("Predicted Salary")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

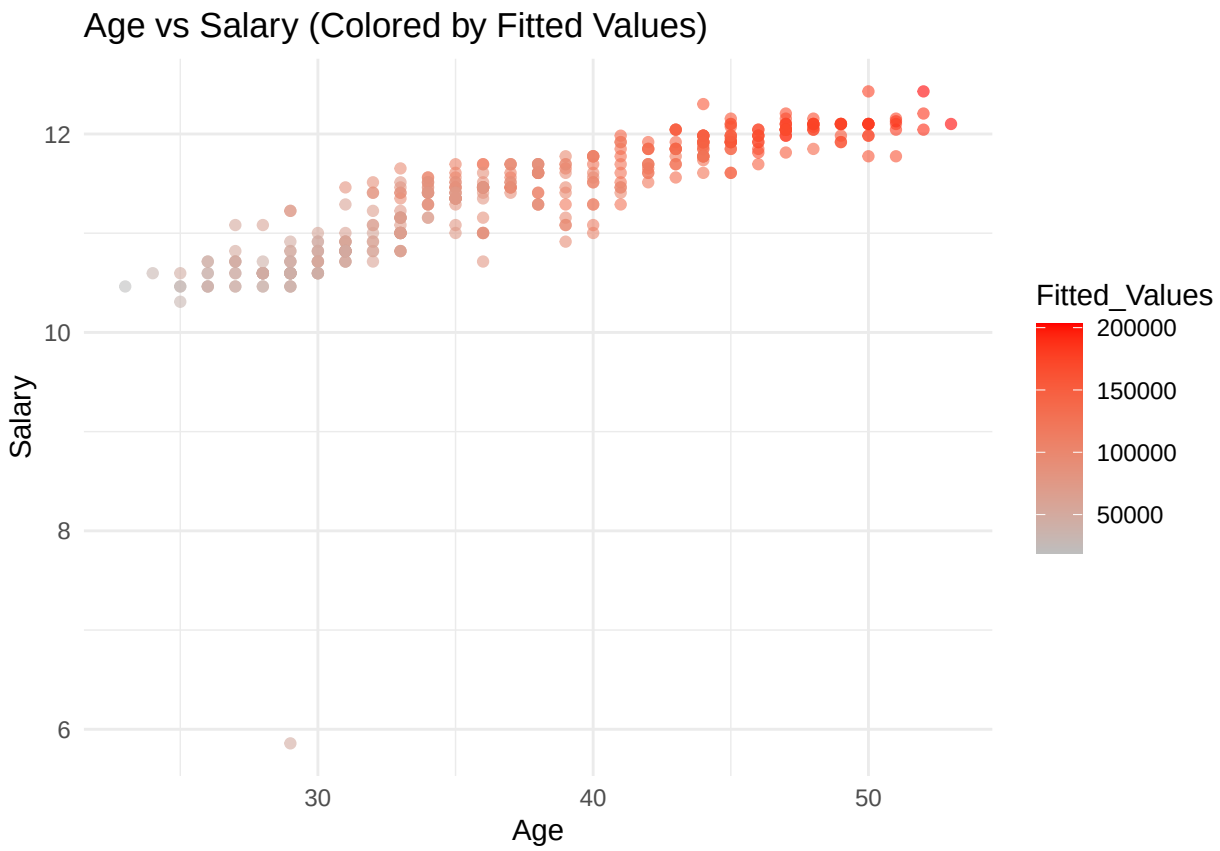



```r
# Create a data frame for the metrics
metrics_df <- data.frame(
  Model = c("Linear Regression", "Random Forest", "Linear Regression", "Random Forest"),
  Dataset = c("Training", "Training", "Testing", "Testing"),
  RMSE = c(15658.61, 15065.59, 14352.31, 14197.53),
  Rsquared = c(0.8973575, 0.9071382, 0.8825102, 0.9162342),
  MAE = c(11291.03, 10307.93, 9470.211, 8840.320)
)

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

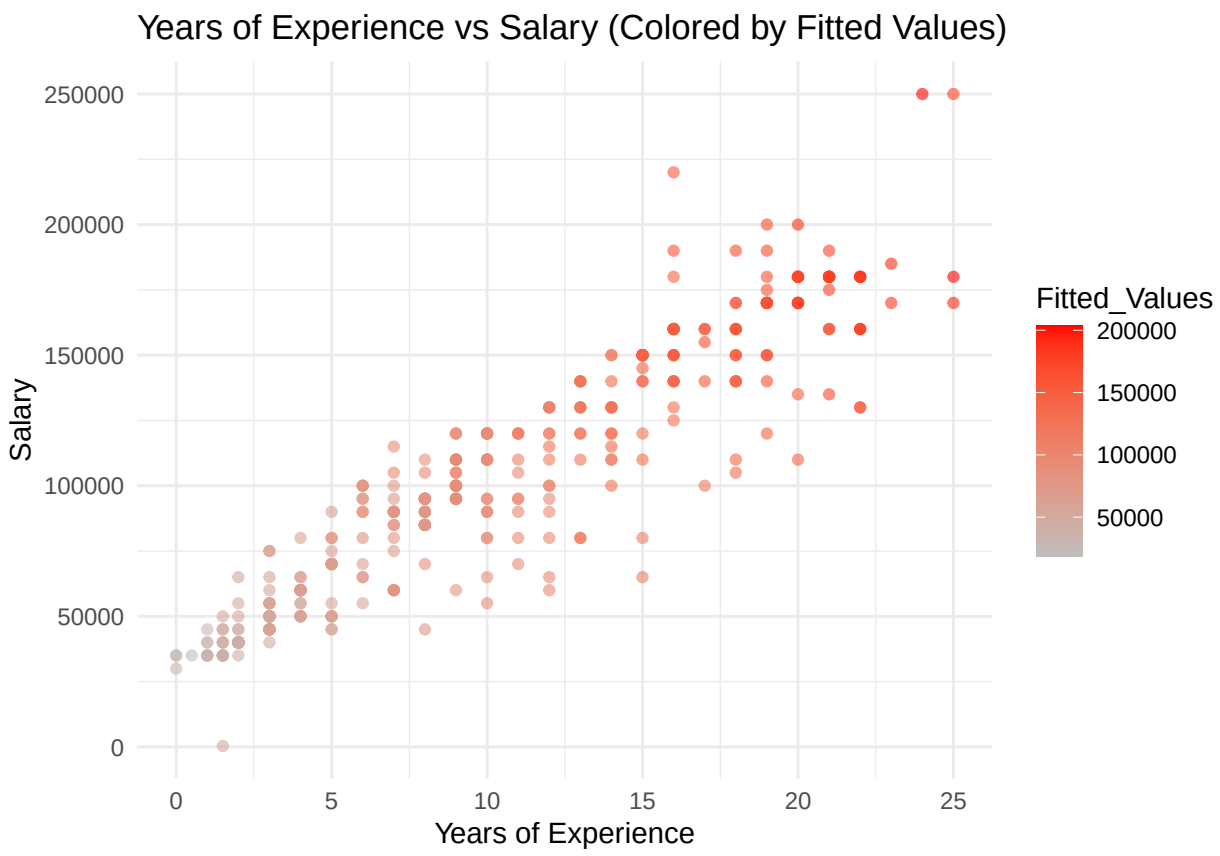
```
##               Model  Dataset     RMSE  Rsquared       MAE
## 1 Linear Regression Training 15658.61 0.8973575 11291.030
## 2     Random Forest Training 15065.59 0.9071382 10307.930
## 3 Linear Regression  Testing 14352.31 0.8825102  9470.211
## 4     Random Forest  Testing 14197.53 0.9162342  8840.320
```

```r
# Adding fitted values to the original data frame
cleaned_data$Fitted_Values <- fitted(simple_model)

# Load necessary library
library(ggplot2)

# Plotting Age vs Fitted Salary using color for fitted values
ggplot(cleaned_data, aes(x = Age, y = log(Salary), color = Fitted_Values)) +
  geom_point(alpha = 0.6) +  # Using semi-transparent points
  scale_color_gradient(low = "gray", high = "red") +  # Gradient color from blue to red
```

```
  labs(title = "Age vs Salary (Colored by Fitted Values)", x = "Age", y = "Salary") +
  theme_minimal()  # Using a minimal theme for better visibility
```




```
# Plotting Years of Experience vs Salary using color for fitted values
ggplot(cleaned_data, aes(x = Years.of.Experience, y = Salary, color = Fitted_Values)) +
  geom_point(alpha = 0.6) +  # Using semi-transparent points
  scale_color_gradient(low = "gray", high = "red") +  # Gradient color from blue to red
  labs(title = "Years of Experience vs Salary (Colored by Fitted Values)", x = "Years of Experience", y
  theme_minimal()  # Using a minimal theme for better visibility
```


Years of Experience vs Salary (Colored by Fitted Values)
Salary
250000
200000
150000
100000
50000
0
0
5
10
15
20
25
Years of Experience
Fitted_Values
200000
150000
100000
50000