What Determines Employee Salaries: A Regression Analysis

Neeraj Namani & Srisailam Gitte

Introduction:

In today's data-driven world, understanding the factors that influencing salary can provide valuable insights for both employees and employers. The data "Salary Data" offers a comprehensive overview of salary distributions across different demographics and job specifications. It can be utilized to analyze trends, benchmark salaries, and guide career decisions based on education, experience, and other relevant factors.

This study utilizes multiple regression analysis conducted in R to explore the determinants of employee salaries, focusing on variables such as experience, education, job role and industry. Through meticulous data preprocessing and model selection, we aim to develop a predictive model. The objective is to offer actionable insights into a salary structuring and strategies for enhancing employee retention.

Research Question:

How do factors such as Years of Experience, Education Levels, Job roles and Industry influence employee salaries and how can this information be utilized to inform salary structuring and retention strategies in organizations?

Dataset:

The dataset comprises 375 entries, representing individual records of employment and salary details. Each entry includes the following attributes:

- Age: The age of the employee.
- Gender: The gender of the employee listed as Male or Female.
- Education.Level: The highest level of education attained by the employee, categorized into Bachelor's, Master's and PhD.
- Job Title: The professional designation of the employee, such as Software Engineer, Data Analyst, Senior Manager, Sales Associate, and Director.
- Years of Experience: The total number of years the employee has spent in their professional field.
- Salary: The annual salary of the employee in USD.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) on the salary dataset provides insights into various aspects that influence salary differences, such as age, gender, education level, job title and years of experience.

<u>Data Cleaning:</u>

Before conducting EDA, initial steps involve handing missing data. The dataset has two missing values in all columns. These are dealt with by omitting rows with missing values, ensuring that subsequent analyses on clean, robust data are not misleading or biased.

<u>Pairwise Relationships:</u>

We have used "ggplot2" library in R studio to explore the pairwise relationships. The below shown visualization is matrix of scatter plots between three key variables: Age, Years of Experience, and Salary. This visualization is commonly used in the part of EDA to investigate the potential relationships between different variables.

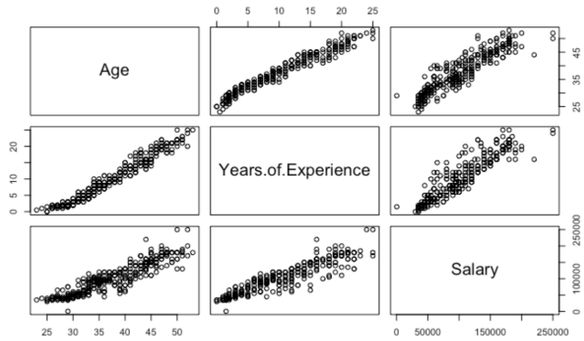


Figure 1: Matrix plot for Pairwise Relationships

From the above plot, we can infer:

- Age vs Years of Experience: This plot is in the lower left corner of the matrix. As expected, it shows a general trend where the number of years of experience tends to increase with age. This positive correlation suggests that as employees get older, they also accumulate more experience.
- Age vs Salary: In the middle of the bottom row, the scatter plot indicates a positive relationship between age and salary. Generally, the salary tends to increase as the age of the employees increases. This could reflect that older employees, who are often more experienced, may hold higher-paying positions.
- Years of Experience vs Salary: It is shown on the bottom right of the matrix, this scatter plot exhibits a strong positive correlation between years of experience and salary. This is a typical finding in salary data, where employees with more years of experience tend to earn higher salaries, likely due to advancing to more senior roles or accumulating merit-based raises over time.

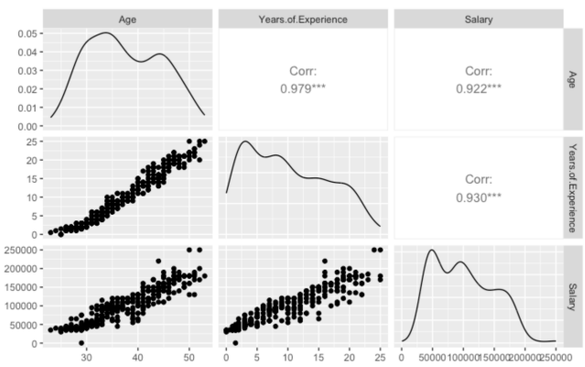


Figure 2: Matrix plot for numerical variables

This visualization is a comprehensive scatterplot matrix, also known as a pairs plot, showing the relationships between Age, Years of Experience, and Salary. Along the diagonal, instead of comparing a variable to itself, we have density plots for each variable, showing the distribution of values.

- Age Distribution: The density plot for age shows a unimodal distribution, possibly indicating a concentration of employees within a certain age range.
- Years of Experience Distribution: The density plot for Years of Experience appears right-skewed, indicating a larger number of employees with fewer years of experience.
- Salary Distribution: The density plot for salary also appears right-skewed, suggesting that a majority of employees earn on the lower end of the salary range.
- Age vs Years of Experience Distribution: The scatterplot shows a strong positive correlation, as indicated by the correlation coefficient (Corr: 0.979***), which is consistent with the idea that as people get older, they typically accumulate more work experience.

- Age vs Salary: There is a positive correlation between Age and Salary (Corr: 0.922***), suggesting that older employees tend to have higher salaries, potentially reflecting career progression.
- Years of Experience vs Salary: The positive correlation here (Corr: 0.930***) is one of the strongest observed, indicating that employees with more experience tend to have higher salaries.

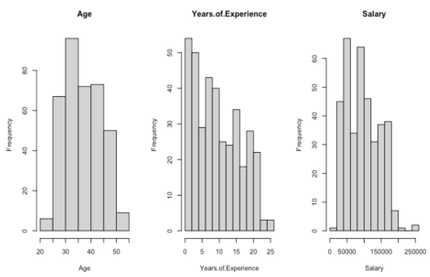


Figure 3: Histogram Plot

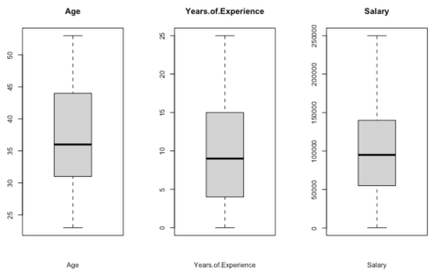


Figure 4: Box Plot

- Age Histogram: The distribution of age seems to be slightly right-skewed, suggesting a young workforce with fewer older employees. The highest frequency occurs in the 30-35 age bracket.
- Age Box plot: The median age is around the mid-30s and the interquartile range (IQR), which spans from approximately the 25th to the 75th percentile, suggests a middle age group clustered between the early 30s and early 40s. The "whiskers" extend to show the full range of the data, which appears to be from the mid-20s to just over 50. There do not seem to be any outliers, indicating a relatively even distribution of ages.
- Years of Experience Histogram: This distribution is right-skewed, showing that the majority of employees have between 0-10 years of experience, with numbers steadily declining as experience increases.
- Years of Experience Box plot: The median years of experience id approximately 10 years, The IQR is broad, extending from around 5 to over 15 years, which indicates a wide middle range of employee experience levels. There are no visible outliers, and the distribution appears slightly right-skewed, with a longer whisker extending toward the maximum value of around 25 years, suggesting that fewer employees have very high levels of experience.
- Salary Histogram: The Salary histogram is also right-skewed, indicating that most employees earn lower salaries, with fewer individuals earning higher salaries. The most common salary range seems to be around $50,000 to $100,000.
- Salary Box plot: The median salary is around $100,000. The IQR for salary is wide, indicating substantial variability in pay between the 25th and 75th percentiles. The whiskers extend from the lower to the upper extremes of the salary range, and there don't appear to be outliers, suggesting that very high or very low salaries are within the expected range for this dataset.

<u>Scatter Plots:</u>

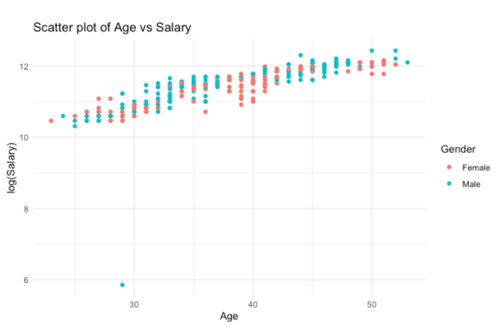


Figure 5: Scatterplot

Figure 6: Scatterplot

The first scatter plot visualizes the relationship between Age and Salary, with Salary values being log-transformed. The use of a log-scale for salary often helps in dealing with right-skewed distributions, making patterns more evident when there is a wide range of income. In this plot,

points are color-coded to represent Gender, providing a visual distinction between salaries of females (in red) and males (in skyblue). From this plot, it seems that there is a general upward trend, indicating that as age increases, so does salary, regardless of gender. The distribution of male and female points suggests that there might be a gender disparity in salaries, with males potentially occupying a higher salary range.

The second scatter plot shows the relationship between Years of Experience and Salary. This plot does not use a log scale for salary, allowing direct interpretation of salary values. Points are color – coded according to Education level, indicating that individuals with Bachelor's, Master's or PhD degrees can be distinguished. We see a positive trend, where more experience tends to correspond to higher salaries. Additionally, the color coding suggests that higher education levels might correlate with higher salaries, particularly noticeable with individuals who have a PhD. However, it is also clear that education alone does not dictate salary – experience plays a significant role, and there is a spread of salaries at each education level.

Bar Plots:

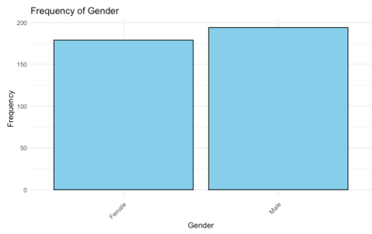


Figure 7: Frequency of Gender

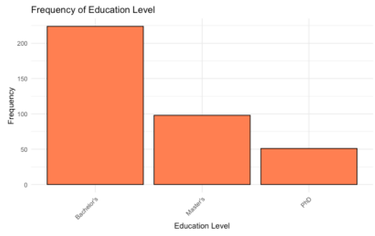


Figure 8: Frequency of Education Level

The "Frequency of Gender" bar plot displays the distribution of genders within the dataset. Two bars represent the frequency of female and male individuals, respectively. The bars are of similar height, which suggest a relatively balanced gender representation in the dataset.

The "Frequency of Education Level" bar plot displays the number of individuals with different levels of education attainment. Bachelor's, Master's and PhD. The tallest bar is for Bachelor's degree holders, followed by a smaller but significant number of Master's degree holders and the shortest bar for PhDs. This indicates that the dataset likely includes more individuals with Bachelor's degrees than other categories.

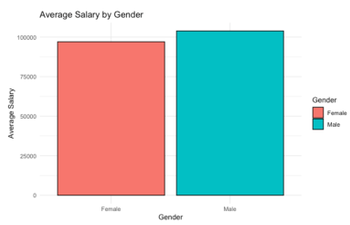


Figure 9: Average Salary (Gender)

Figure 10: Average Salary (Education Level)

The "Average Salary by Gender" bar chart shows the comparison of average salaries between genders. The two bars, one for each gender indicate the average salary level for females and males. The height of each bar gives a visual representation of the average salary, where one can observe any apparent disparities between the genders.

The "Average Salary by Education Level" bar chart is the average salary plotted against the level of education. Three bars represent the average salaries for individuals with Bachelor's, Master's and PhD degrees. The increasing height of the bars from Bachelor's to PhD suggests a trend where higher education correlates with higher average salaries.

Figure 11: Average Salary by Years of Experience

The bar chart visualizes the average salaries across different categories of years of experience. From this chart that there is a general upward trend in salary with increasing experience. Each bar represents a different experience bracket.

- 0-5 years: Individuals with this category, presumably entry-level or early career professionals, earn the lowest average salaries.
- 6-10 years: There is a notable jump in average salary as individuals gain more experience, moving into mid-level positions.
- 11-15 years: Continuing with the trend, average salaries increase further for those who have solidified their careers.
- 16-20 years: The average salary in this group is higher yet, likely reflecting senior-level expertise.
- 20+ years: This category shows the highest average salaries, aligning with the expectation that extensive commands top compensation.

A tibble: 5 × 2

Experience_Category <fctr>	Average_Salary <chr>
20+ years	$173,659
16-20 years	$151,343
11-15 years	$109,627
6-10 years	$85,904
0-5 years	$46,494

5 rows

Figure 12: Average Salaries

Correlation Heatmap:

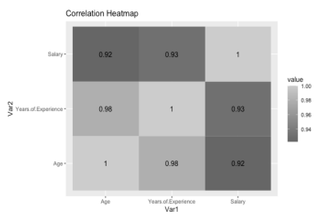


Figure 13: Correlation Heatmap

This correlation heatmap is a visual representation of the strength and the direction of linear relationships between three variables: Age, Years of Experience and Salary. In a correlation heatmap, the colors and the numerical variables within cells indicate the magnitude of the correlation coefficients with 1 being the strongest positive correlation and values closer to 0 indicating weaker correlations.

In this heatmap:

- The diagonal from the top left to the bottom right shows perfect correlations of 1, as these compare each variable to itself.
- The correlation between Age and Years of Experience is extremely high (0.98), suggesting a strong positive linear relationship, as would be expected since more years typically pass as one ages.
- The correlation between Age and Salary is also high (0.92), indicating that older individuals tend to have higher salaries, potentially due to more experience and seniority.

- The correlation between Years of Experience and Salary is similarly strong (0.93), supporting the idea that more experience can often lead to higher salaries.

Modelling and Diagnostics:

We hypothesize that salary can be predicted from all the predictors such as Age, Years of Experience, Education Level and Gender.

The model is lm(Salary ~ Age + Years.of.Experience + Education.Level + Gender, data = cleaned_data)

The summary of the model is

```
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
```

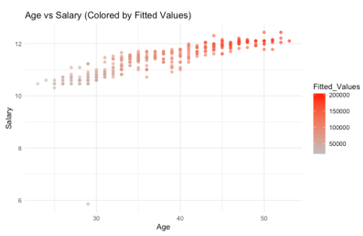


Figure 14: Scatterplot with fitted values

Figure 15: Scatterplot with fitted values

The second model is log-transformed model,

lm(log(Salary) ~ Age + Years.of.Experience + Education.Level + Gender, data = cleaned_data)

The summary of the log-transformed model is

```
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
```

The third model is polynomial transformation model,

lm(poly(Salary,2) ~ Age + Years.of.Experience + Education.Level + Gender, data = cleaned_data)

The summary of the polynomial transformation model is

```
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
```

The fourth model is the interaction model,

lm(Salary ~ Age * Years.of.Experience + Education.Level + Gender, data = cleaned_data)

The summary of the model is

```
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
```

- Out of these four models, first model (Linear model) has the highest Multiple R-squared value: 0.9031 and Adjusted R-squared value: 0.9017.
- We can see that for first model and the polynomial transformation model (Response 1) has the same R-squared values.
- The F-statistic value for the first model is 683.1 is the highest, indicating that the model is significant.

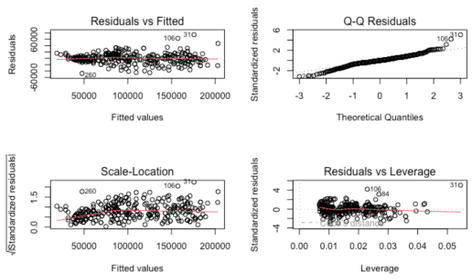


Figure 16: Assessment Plots for the first model

- Residuals vs Fitted: This plot helps to check the assumption of linearity and homoscedasticity (constant variance) of residuals. Ideally, the points should be randomly dispersed around the horizontal line, which would indicate a linear relationship and homoscedasticity. In this graph, there is a slight pattern to the residuals, which may suggest some non-linearity in the relationship between the variables or the presence of heteroscedasticity.
- Normal Q-Q: This plot is used to assess if the residuals are normally distributed. Points following the line closely indicate the residuals are normally distributed. In the Q-Q plot presented, most points follow the line but deviate at the ends, suggesting the presence of outliers or that the tails of the distribution are heavier than expected under normality.

- Scale-Location: This plot is another way to visualize homoscedasticity. Ideally, the points should be evenly spread across the horizontal axis, indicating equal variance of errors. A funnel shape (points fanning out with fitted values) would indicate heteroscedasticity. The points in this graph seem to show a consistent spread, suggesting homoscedasticity is a reasonable assumption.
- Residuals vs Leverage: This plot helps to identify influential cases (outliers) that might have an undue influence on the model's fit. The points with the highest leverage and largest residuals are of particular interest as potential influential outliers. The Cook's distance lines help determine if any points are unduly influencing the regression fit.

Model Selection:

Train Test Split:

The dataset was divided into two parts: 80% training data and the rest 20% as testing data. We build the model on these training data using cross-validation.

Our linear regression model with cross validation:

```
# 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)
```

The summary of this model is

```
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
```

- We can notice that the model has scored an R-squared value: 0.8973575
- The values of RMSE: 15658.61 and MAE: 11291.03

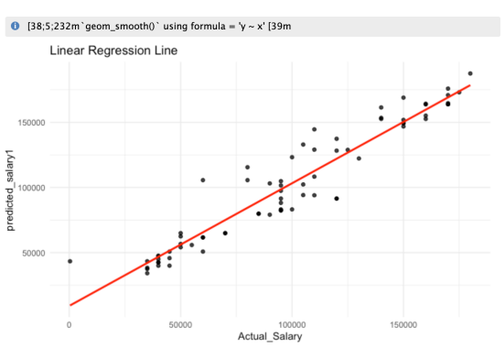
Below is the data frame showing the actual salary and the predicted salary

Description: df [6 × 3]

	Actual_Salary <int>	Predicted_Salary <dbl>	Error <dbl>
7	120000	128287	8287
16	125000	128905	3905
26	45000	45806	806
35	170000	170879	879
36	45000	39993	-5007
43	60000	50832	-9168

6 rows

Linear Regression Plot:



The Linear Regression plot depicts a model's fit. The line of best fit through data points suggest a strong linear relationship between actual and predicted salaries, indicating that the model has good predictive accuracy. The data points are relatively close to the line, though some variance is visible, implying that while the model is generally accurate, there may be some predictions that are not as close to the actual values.

Our Random Forest model with cross validation:

```{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)
```

The summary of the model is

```
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.
```

- We can notice that for the tree 3, it has the highest R-squared value: 0.9071382
- The RMSE value of mtry = 3 is low compared to mtry = 2 and mtry = 5 i.e. 15065.59
- The MAE value of mtry = 3 is low compared to mtry = 2 and mtry =5 i.e. 10307.93

Below is the data frame showing the actual salary and the predicted salary

Description: df [6 × 3]

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

6 rows

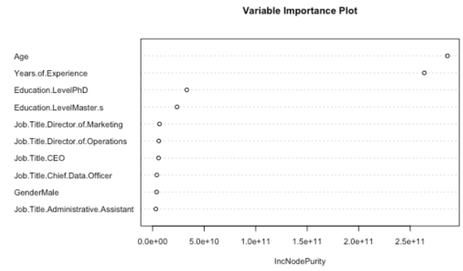
Our Random Forest model without cross-validation:

```
# 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)
```

The summary of this model is,

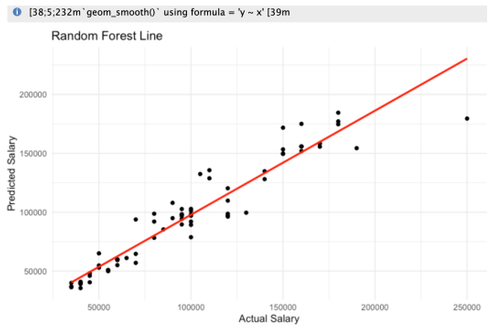
```
Call:
 randomForest(x = X_train, y = y_train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 59

          Mean of squared residuals: 221682732
                    % Var explained: 90.45
```

Variable Importance Plot:



The Variable Importance Plot displays the relative significance of different predictors in a model. “Age” and “Years of Experience” stand out as the most influential variables, suggesting they are key factors in the model’s decisions. The importance of “Education Level” and specific “Job Titles” varies with “Gender” having the least impact. This visualization helps identify which features most affect the model’s output.

Random Forest Plot:



The Random Forest plot shows the fit of the random forest model. This line also indicates a positive linear relationship between actual and predicted salaries. The line is relatively smooth, suggesting that the random forest model may have a slightly different set of predictions with possibly less variance in the residuals.

Conclusion:

Model <chr>	Dataset <chr>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>
Linear Regression	Training	15658.61	0.8973575	11291.030
Random Forest	Training	15065.59	0.9071382	10307.930
Linear Regression	Testing	14352.31	0.8825102	9470.211
Random Forest	Testing	14197.53	0.9162342	8840.320

4 rows

The performance metrics for Linear regression and Random Forest models indicate that the random forest outperforms the linear regression across all three considered metrics – RMSE, R-squared and MAE on both training and testing datasets. The random forest’s higher R-squared and lower RMSE and MAE on testing data suggest better generalization error and predictive accuracy. These results underscore the effectiveness of ensemble methods like random forest in capturing complex patterns in the data. Our study concludes that experience and advanced education significantly influence salary outcomes, with the random forest model providing a more nuanced understanding of these relationships than linear regression.

References:

1. “Employee Salaries Analysis and Prediction with Machine Learning” 2022,
   Link: https://ieeexplore.ieee.org/document/9943146
2. “Employee Salary Prediction”, 2022
   Link: https://www.ijariit.com/manuscript/employees-salary-prediction/