Salary

Chenghao Xia

2023-12-12

Abstract

I selected salary as my data collection method because it can provide insightful information about the variables affecting individual salary outcomes. Additionally, knowing these connections might help decision-makers in organizational settings make well-informed choices. Employers can utilize the model to inform data-driven decisions about workforce planning, talent development, and wage structures. Additionally, the model might point out possible areas for improvement or action, such as addressing racial or gender-based differences in salary results.

Introduction

This study examines a large data collection that includes salary information for individuals with a variety of demographic and professional characteristics. The data set includes four categorical factors (gender, job title, nation, and race) and five numeric variables (age, education level, years of experience, salary, and senior status). Notably, job titles cover 129 different occupations, gender displays a binary distribution between male and female, and country includes five countries (UK, USA, Canada, China, and Australia). Ten categories make up race: African American, Welsh, White, Hispanic, Asian, Korean, Chinese, Australian, mixed, and black. Ages range from 21 to 62, education levels are broken down into four groups (high school, bachelor's, master's, and PhD degrees), and years of experience are 0 to 34, according to descriptive statistics. Salary ranges are expressed in US dollars; the range of salaries is 350-250,000 US dollars, and senior status is binary, meaning that a person is either in a senior position or not. The data set provides a wealth of resources for investigating the connections and trends among the variables, enabling examinations such as multilevel modeling, correlation analysis, and a more comprehensive understanding of the elements impacting salaries around the globe.

Variable	Description	Data Type	Possible Values
Gender	Gender of the employee	Categorical	Male, Female
Job Title	Job title of the employee	Categorical	(List of job titles)
Country	Country of residence	Categorical	UK, USA, Canada, China, Australia
Race	Race or ethnicity of the employee	Categorical	White, Hispanic, Asian, …
Age	Age of the employee	Numeric	21 to 62
Education Level	Education level of the employee	Categorical	High School, Bachelor, Master, PhD
Years of Experience	Number of years of work experience	Numeric	0 to 34
Salary	Salary of the employee (in USD)	Numeric	350 to 250000
Senior	Binary variable indicating senior position	Binary	0 (No), 1 (Yes)

Variable	Description	Data Type	Possible Values

Method

Data Cleaning

The original dataset has 6684 rows and 9 columns. The dataset was improved after identifying and deleting duplicate values, an essential step to preserve data integrity, resulting in a final dataset of 5148 rows and 9 columns. A subsequent check for null values revealed that there were no missing data points. This fortunate condition eliminates concerns about model failure owing to missing variables, resulting in a clean and complete dataset for future analysis and modeling efforts. The thorough data pretreatment, which includes duplicate removal and null value verification, adds to overall data quality and lays the groundwork for robust and trustworthy modeling results.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) serves as a crucial preliminary step in understanding the relationships within a dataset, particularly in the context of examining potential correlations with the variable of interest, which in this case is salary. To discern the influence of categorical variables, boxplots and analysis of variance (ANOVA) are employed. The boxplots provide visual insights into the distribution of salary across different categories, while ANOVA tests the statistical significance of variations among groups. Meanwhile, for numerical variables, a heatmap and variance inflation factor (VIF) analysis are conducted to gauge the degree of correlation among factors. The heatmap visually represents the strength and direction of correlations, while VIF quantifies the extent of multicollinearity. Variables exhibiting correlations with salary are identified and considered as potential factors in the subsequent modeling process. Furthermore, interactive effects among these variables are scrutinized to discern any nuanced relationships that might impact salary outcomes. This comprehensive analytical approach ensures a thorough exploration of both categorical and numerical factors, paving the way for a robust and informed modeling strategy.

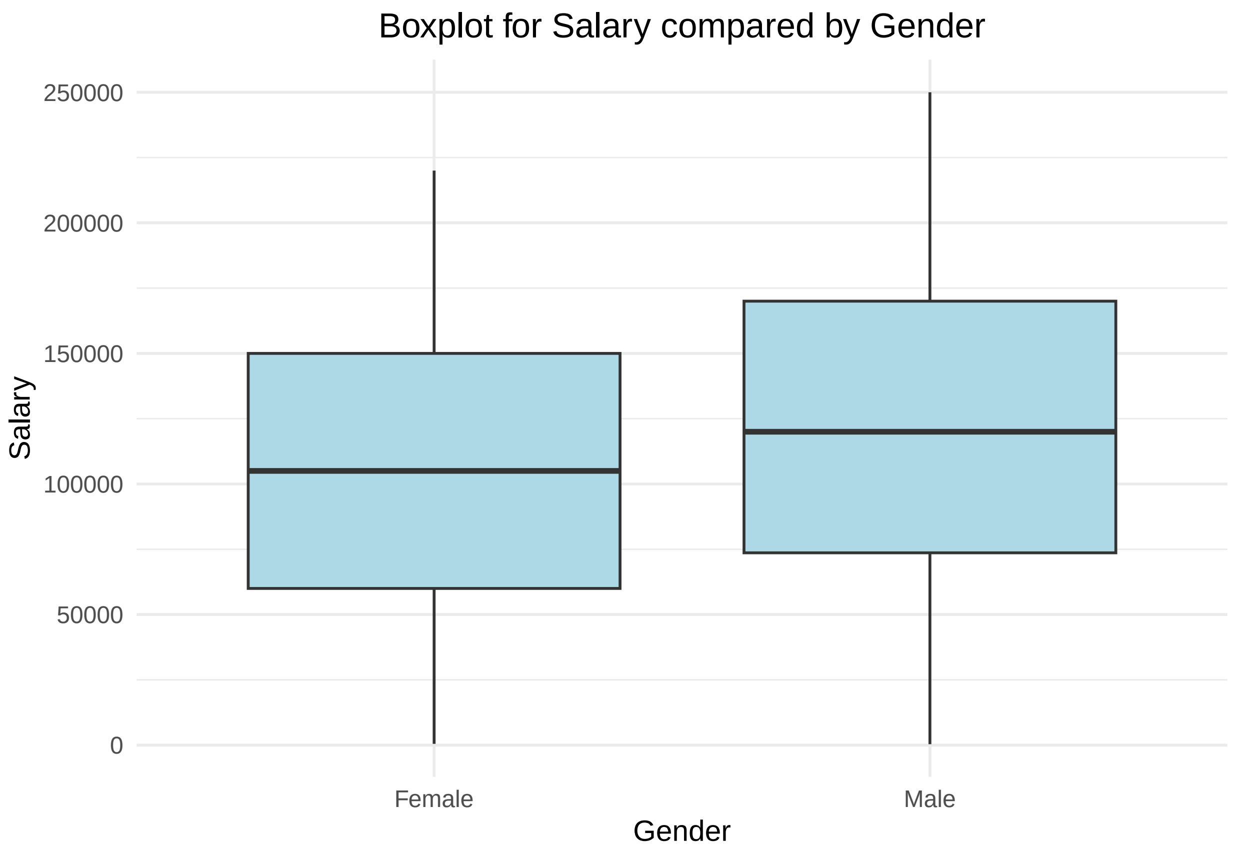
Multilevel Linear Model

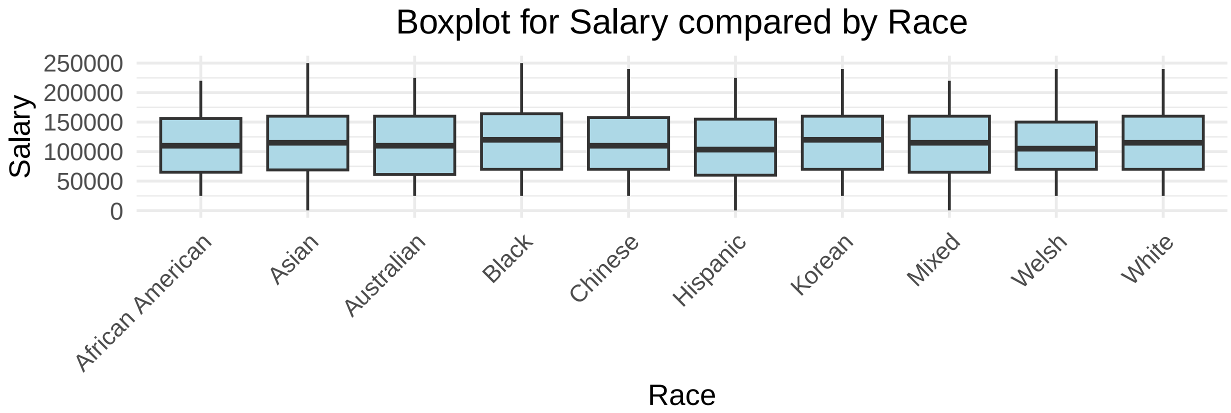
There are two distinct models under consideration: one that includes the interaction term and one that does not. The reasoning behind fitting both models is to investigate the term of the interaction effect on the outcomes. To ensure the validity and dependability of the multilevel models, their assumptions are thoroughly examined. This entails a thorough analysis of essential assumptions, such as linearity, normality, and variance homogeneity. In both cases, the goal is to examine whether the data meets the key requirements for multilevel modeling. We hope to determine the suitability of the models and get insight into potential deviations from the underlying assumptions that could affect the trustworthiness of the outcomes through this thorough review.

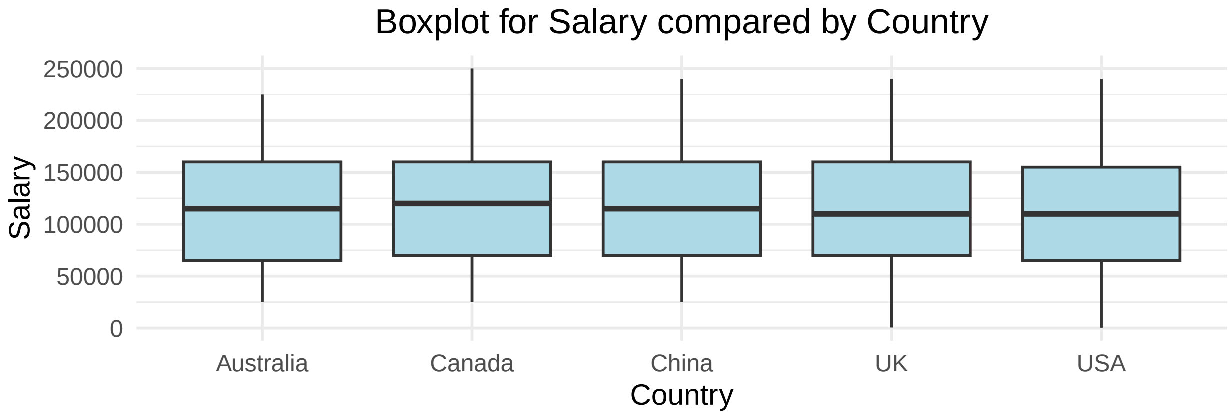
Result

Visualization

categorical variables

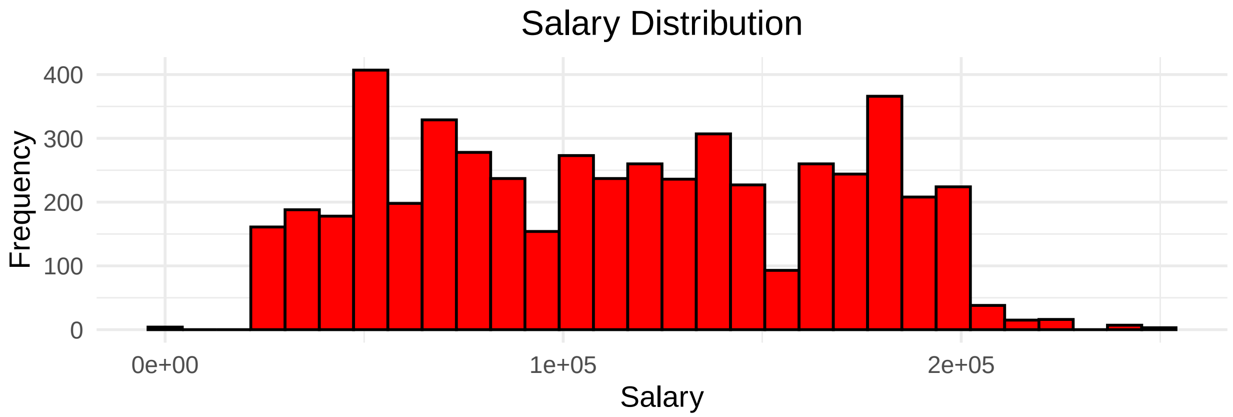


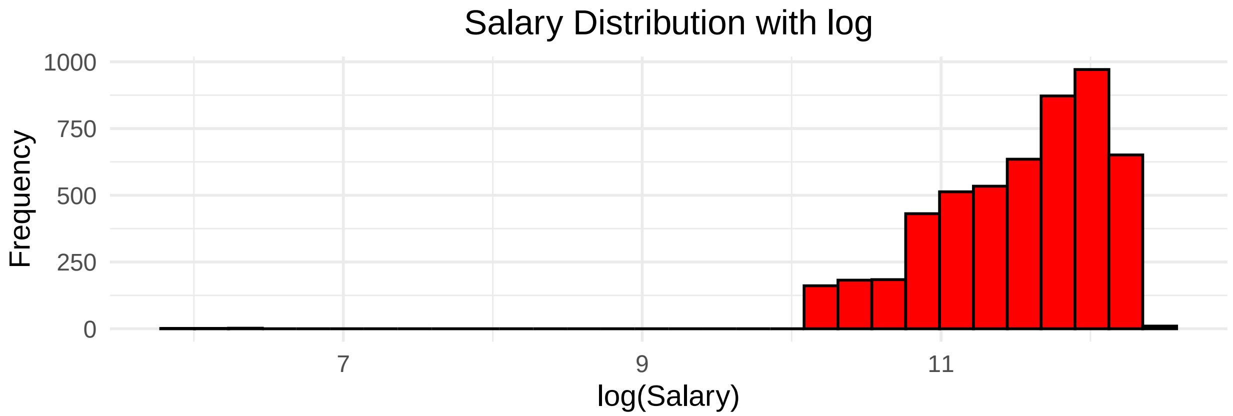




Figures above show the distribution of the salary given different factors. For gender, we can see there is a clear difference between male and female, whether in 25% value, median, 75% value, or max value, which shows that gender is an important factor for salary. And for ANOVA analysis, the p-value is small, which indicates that there are significant differences in salary across genders. For country and race, it seems like we cannot see a significant difference in salary across countries and races. And in order to check by ANOVA, it gives us a high p-value, which proves that there is no significant difference in salary across country and race. For job titles, since we have 129 different job titles, boxplot cannot work well here, so we just use ANOVA to check the correlation. The p-value is small enough to say that there are significant differences in salary across job titles.

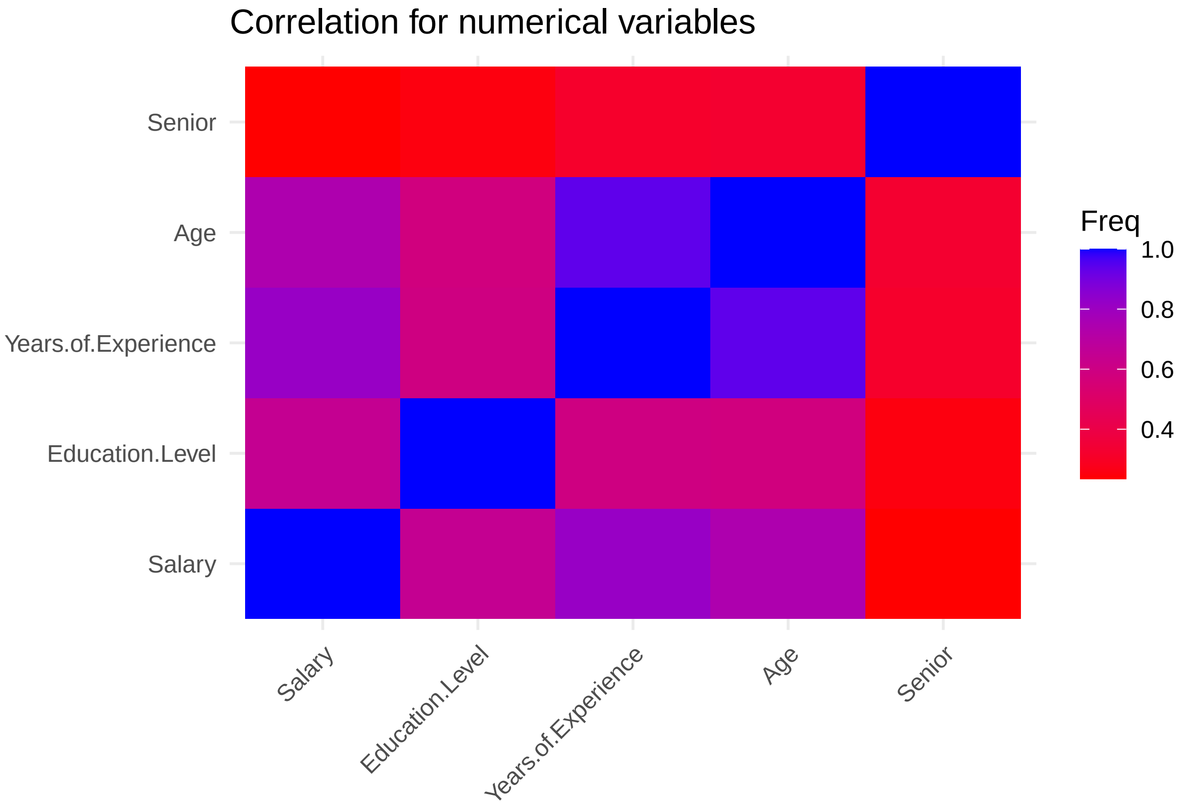
numerical variables


Salary Distribution


Salary Distribution with log

First, we look at the most important variable, which is salary. The histogram of the raw salary data suggests a relatively even distribution, indicating that salaries are spread across different ranges without a strong skew. This suggests a more uniform distribution of salaries without a pronounced concentration in a particular salary range.

Since the range for salary is wide, we use a logarithmic transformation. After applying the logarithm transformation to the salary data, the distribution becomes right-skewed. The right skewness indicates that there are relatively more high salary values, and the transformation helps in highlighting differences within the high salary range.



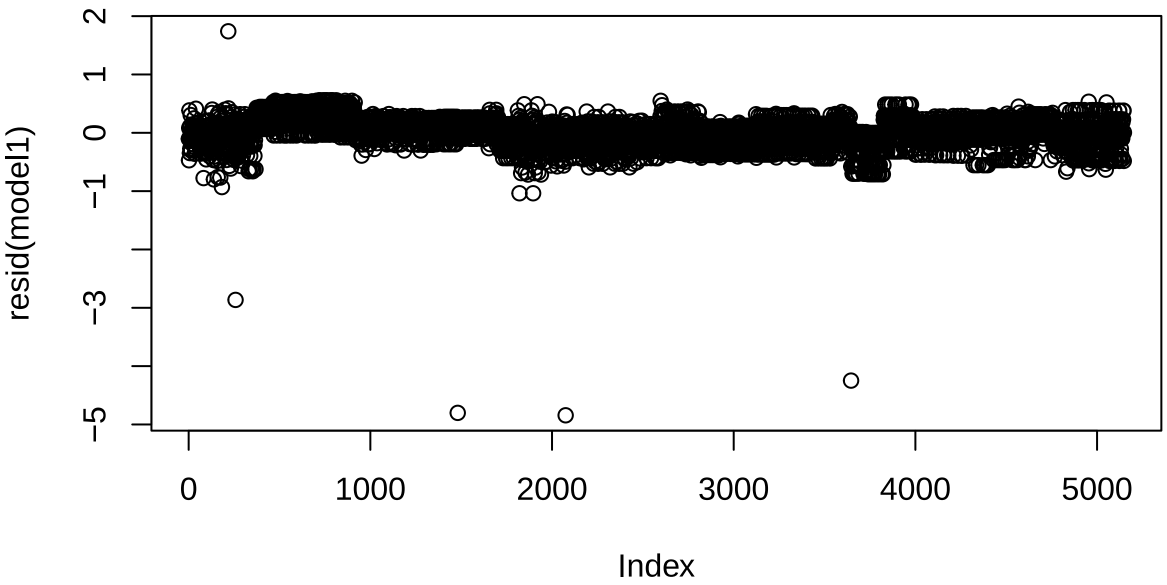
Now we are looking at the correlation between these numerical variables. We use the heat map to see if there is a high correlation between these variables. We find that age and years of Experience has not only a high correlation with each other but also a high correlation with salary. But senior is a variable that has a very low correlation with all other variables.

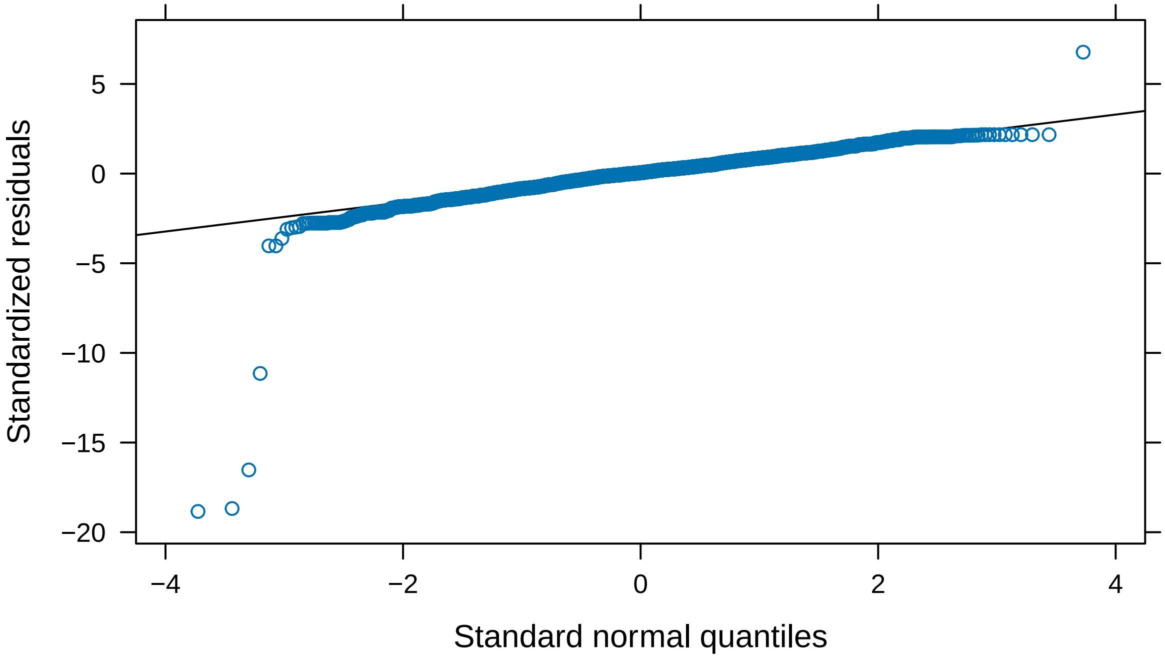
Then we are going to check for collinearity. Years.of. Experience and age have a VIF value that is very high. So we can choose age and years of Experience as our variables and put an interaction term for these two variables in the model.

Multilevel Model

Without interaction effects

The model includes age, year of experience, and random effects grouped by gender and job title. And we normalize the salary by log (salary), as mentioned before.

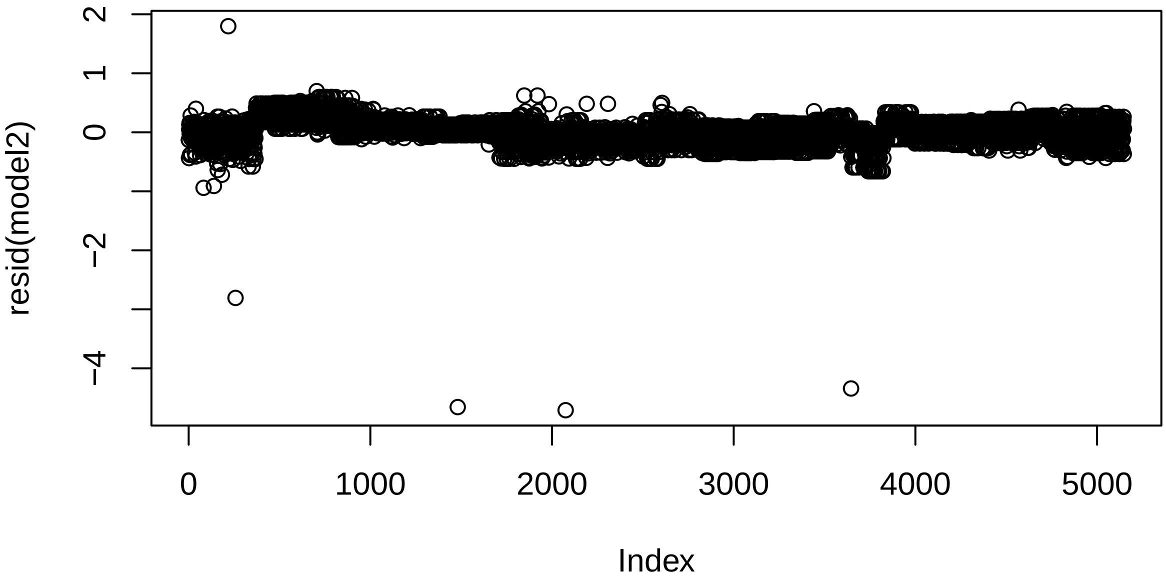


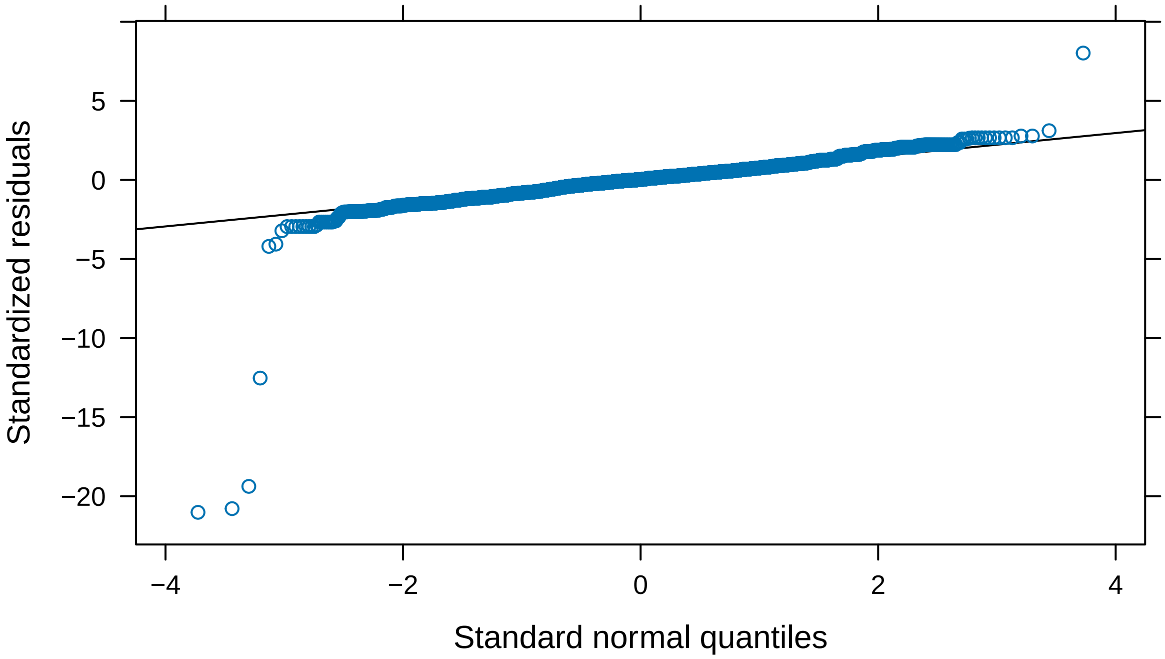


In these two figures, it shows the residual plot and the qq plot, which generally confirm the linearity and normality of the model. Although there are some outliers that are away from 0, most of the residuals are randomly distributed around 0.

Include interaction effects

The model includes age, year of experience, the interaction effect between age and year of experience, and the random effects grouped by gender and job title. And we normalize the salary by log (salary),as mentioned before.





In the figure above, the residual plot and the qq plot generally confirm the linearity and normality of the model. The model helps to improve the homogeneity of variability as the points are more normally distributed in the residual and qq plots.

Discussion

There are two models, where one contains the interaction term and the other doesn't. The result shows that the homogeneity is improving while adding the interaction term between age and year of experience.

For the model, y equals the salary, and x is the variable that includes Age, year of experience, the interaction effect between age and year of experience, and the random effects grouped by gender and job title. The interaction term is much more statistically significant than other fixed effects.The model, using restricted maximum likelihood (REML), effectively explores the relationship between log-transformed salary and demographic factors like age, years of experience, and their interaction. The model demonstrates robust fit with a convergence criterion of -398.4, and the scaled residuals display a symmetrical distribution around zero (-21.0209 to 8.0234). Random effects analysis indicates minimal salary variability linked to gender, while job title introduces some variance (0.1096). Fixed effects reveal positive impacts of age and years of experience on log (salary), supported by significant t-values. The interaction term, age and years of experience, underscores a moderating effect.

However, there are some challenges for me with this study. Firstly, I think it would be better to classify 129 job titles into some groups. "Manager" would include "Marketing Manager" and "Financial Manager," but I think it would be hard when one job title can fit more than one group, so I tried but gave up on it.

Secondly, it is surprising that in the residual plot, most of the outliers are below 0. I would think that outliers should be equally above and below 0 since there is a saying that "1 percent of the population holds 99% of the money." But there is only one outlier above the residual plot. Maybe the reason is that people who have salaries much higher than other people are not included in the study, so this study may not represent the population.

Lastly, in the race part, I tried to put "Korean" and "Chinese" into "Asian," but then I found that in the country part, there is China. So I think maybe it was not correct to classify "Chinese" as "Asian."

What's more, a future study can be done to see which factors might affect the salary the most, and a model might be fitted to predict its popularity.

Appendix

```
## Summary Statistics:
```

```
##       Age           Gender          Education.Level  Job.Title
##  Min.   :21.00   Length:6684        Min.   :0.000   Length:6684
##  1st Qu.:28.00   Class :character   1st Qu.:1.000   Class :character
##  Median :32.00   Mode  :character   Median :1.000   Mode  :character
##  Mean   :33.61                      Mean   :1.622
##  3rd Qu.:38.00                      3rd Qu.:2.000
##  Max.   :62.00                      Max.   :3.000
##  Years.of.Experience     Salary         Country              Race
##  Min.   : 0.000      Min.   :   350   Length:6684        Length:6684
##  1st Qu.: 3.000      1st Qu.: 70000   Class :character   Class :character
##  Median : 7.000      Median :115000   Mode  :character   Mode  :character
##  Mean   : 8.078      Mean   :115307
##  3rd Qu.:12.000      3rd Qu.:160000
##  Max.   :34.000      Max.   :250000  
##      Senior      
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.1435  
##  3rd Qu.:0.0000  
##  Max.   :1.0000  

## 
## Number of Unique Values:

##                Age              Gender     Education.Level           Job.Title 
##                 41                   2                   4                 129 
## Years.of.Experience              Salary             Country                Race 
##                 37                 437                   5                  10 
##              Senior 
##                   2 

## 
## Count of Missing Values:

##                Age              Gender     Education.Level           Job.Title 
##                   0                   0                   0                   0 
## Years.of.Experience              Salary             Country                Race 
##                   0                   0                   0                   0 
##              Senior 
##                   0 

## Gender ANOVA:

##               Df    Sum Sq   Mean Sq F value Pr(>F)    
## Gender         1 2.384e+11 2.384e+11   88.08 <2e-16 ***
## Residuals   5146 1.393e+13 2.706e+09                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

## 
## Race ANOVA:

##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Race           9 2.222e+10 2.469e+09   0.897  0.527
## Residuals   5138 1.414e+13 2.752e+09               

## 
## Job ANOVA:

##               Df    Sum Sq   Mean Sq F value Pr(>F)    
## Job.Title    128 8.473e+12 6.619e+10   58.37 <2e-16 ***
## Residuals   5019 5.692e+12 1.134e+09                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## 
## Country ANOVA:
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## Country        4 1.141e+10 2.851e+09   1.036  0.387
## Residuals   5143 1.415e+13 2.752e+09
```

```
## 
##  VIF_values with Salary:
```

```
##     Education.Level Years.of.Experience                 Age              Senior 
##            1.564813            8.408709            8.294776            1.127902
```

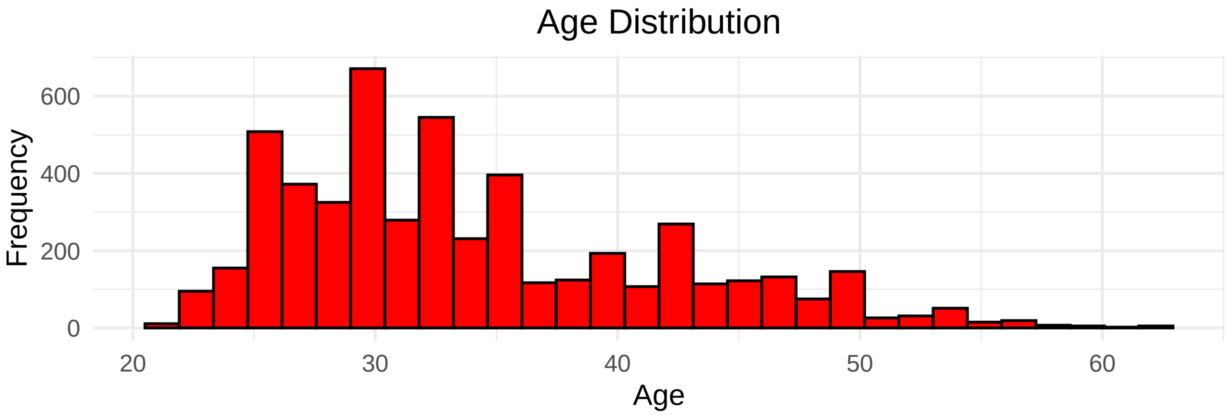
```
## 
## Model 1 Summary:
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(Salary) ~ Age + Years.of.Experience + (1 | Gender) + (1 |  
##     Job.Title)
##    Data: salary
## 
## REML criterion at convergence: 991.2
## 
## Scaled residuals: 
##      Min       1Q   Median       3Q      Max 
## -18.8418  -0.5198   0.0498   0.5797   6.7712 
## 
## Random effects:
##  Groups    Name        Variance  Std.Dev.
##  Job.Title (Intercept) 0.1332817 0.36508 
##  Gender    (Intercept) 0.0001572 0.01254 
##  Residual              0.0660610 0.25702 
## Number of obs: 5148, groups:  Job.Title, 129; Gender, 2
## 
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)         10.654549   0.052721 202.091
## Age                  0.009310   0.001484   6.273
## Years.of.Experience  0.039208   0.001841  21.301
## 
## Correlation of Fixed Effects:
##             (Intr) Age   
## Age         -0.709       
## Yrs.f.Exprn  0.590 -0.910
```

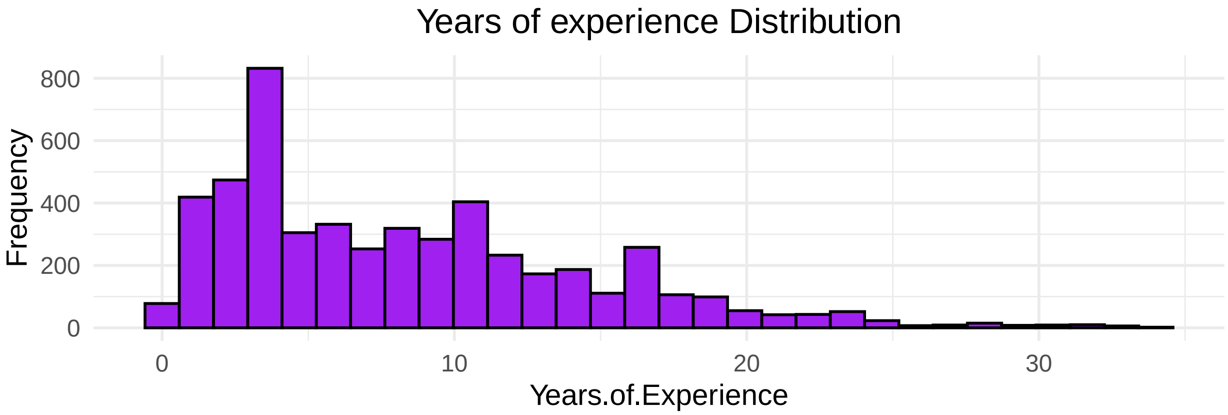
```
## 
## Model 2 Summary:
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(Salary) ~ Age + Years.of.Experience + Age * Years.of.Experience +  
##     (1 | Gender) + (1 | Job.Title)
##    Data: salary
```

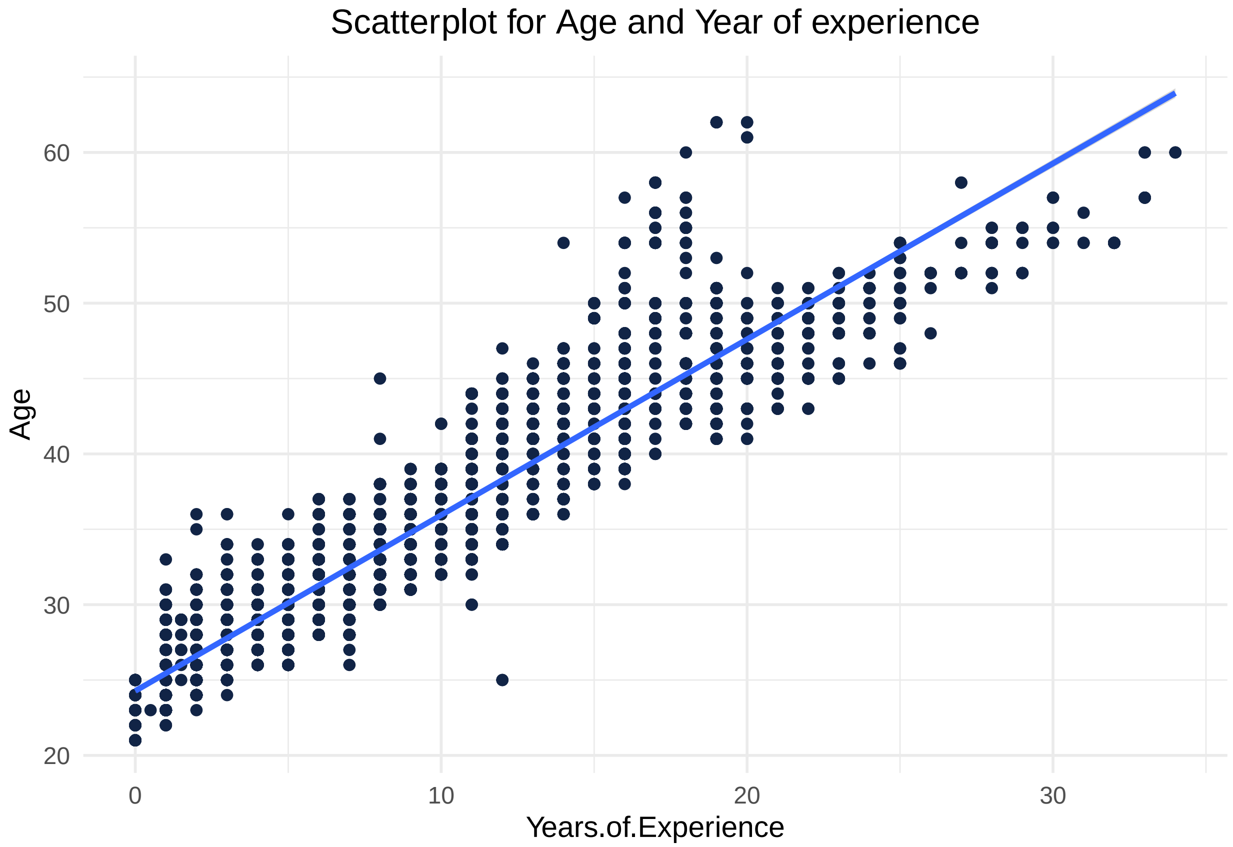
```
##
## REML criterion at convergence: -398.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -21.0209  -0.4839   0.0231   0.5114   8.0234
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Job.Title (Intercept) 0.1096   0.3311
##  Gender    (Intercept) 0.0000   0.0000
##  Residual              0.0502   0.2241
## Number of obs: 5148, groups:  Job.Title, 129; Gender, 2
##
## Fixed effects:
##                           Estimate Std. Error t value
## (Intercept)              9.970e+00  4.894e-02  203.72
## Age                      2.695e-02  1.360e-03   19.82
## Years.of.Experience      1.592e-01  3.388e-03   46.98
## Age:Years.of.Experience -2.741e-03  6.806e-05  -40.28
##
## Correlation of Fixed Effects:
##             (Intr) Age    Yrs..E
## Age         -0.739
## Yrs.f.Exprn -0.041 -0.129
## Ag:Yrs.f.Ex  0.344 -0.317 -0.881
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```



Age Distribution



Years of experience Distribution


Scatterplot for Age and Year of experience
Age
60
50
40
30
20
0
10
20
30
Years.of.Experience