STEM Salary Prediction

Group 2
Angeli De Los Reyes
Nay Zaw Lin
Meyliani Sanjaya

Team Member

Neo

Geli

Mey

What is the best model to predict the salaries of STEM employees?

01 Data Information

Source : Kaggle
Author : Jack Ogozaly
Dataset name : 2022 Data Science and STEM salary
Responses captured from 2017 to 2019

Variable Information

Variable Name	Type	Unit	Description
totalyearlycompensation	Numerical	USD	Total compensation received (all forms)
basesalary	Numerical	USD	Fixed amount of money received on regular basis
stockgrantvalue	Numerical	USD	Shares of company stock
bonus	Numerical	USD	One-time or irregular payment
yearsofexperience	Numerical	year	Total no. of years experience in field
yearsatcompany	Numerical	year	Total no. of years employed at company

Variable Information

Variable Name	Type	Levels	Transformation
gender	Categorical	2 (Male/Female)	Reclassified 'Other'
race	Categorical	2 (White/Non-W)	Simplified (from 5)
country	Categorical	4 (US, UK, IN, CA)	Retain high-response only
education	Categorical	3 (PhD, Masters, College/below)	Simplified (from 5)
fortune_500	Categorical	2 (Yes/No)	Derived from 'company'
title	Categorical	2 (Management/Non-Mgt)	Simplified (from 15)

02 Literature Review

information Published: 12 October 2022 MDPI

Article

Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations

Yasser T. Matbouli [1,*] and Suliman M. Alghamdi [2]

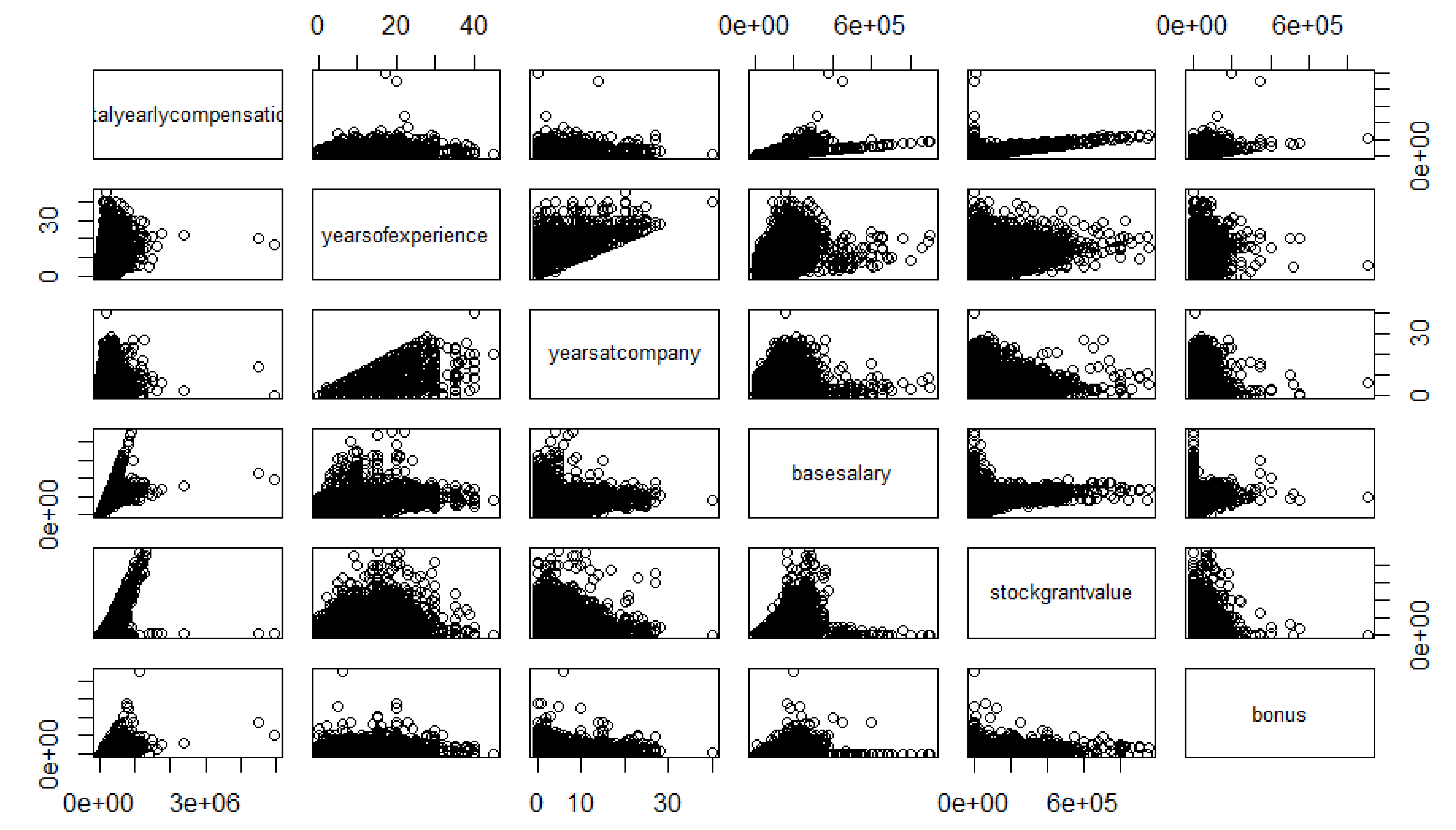
Citation: Matbouli, Y.T.; Alghamdi, S.M. Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. *Information* **2022**, *13*, 495. https://doi.org/10.3390/info13100495

Key Takeaway

- Salary prediction models in literature are mostly concerned with the problem of unequal pay based on gender, race, or other biases that are not related to job content or job performance
- The performance of each regression model is given based on root-mean-square error (RMSE), R-squared (R 2), and mean absolute error (MAE).
- In this study, when cover the broader salary estimates, minor groups can also be featured in the prediction model to capture the occupational characteristics

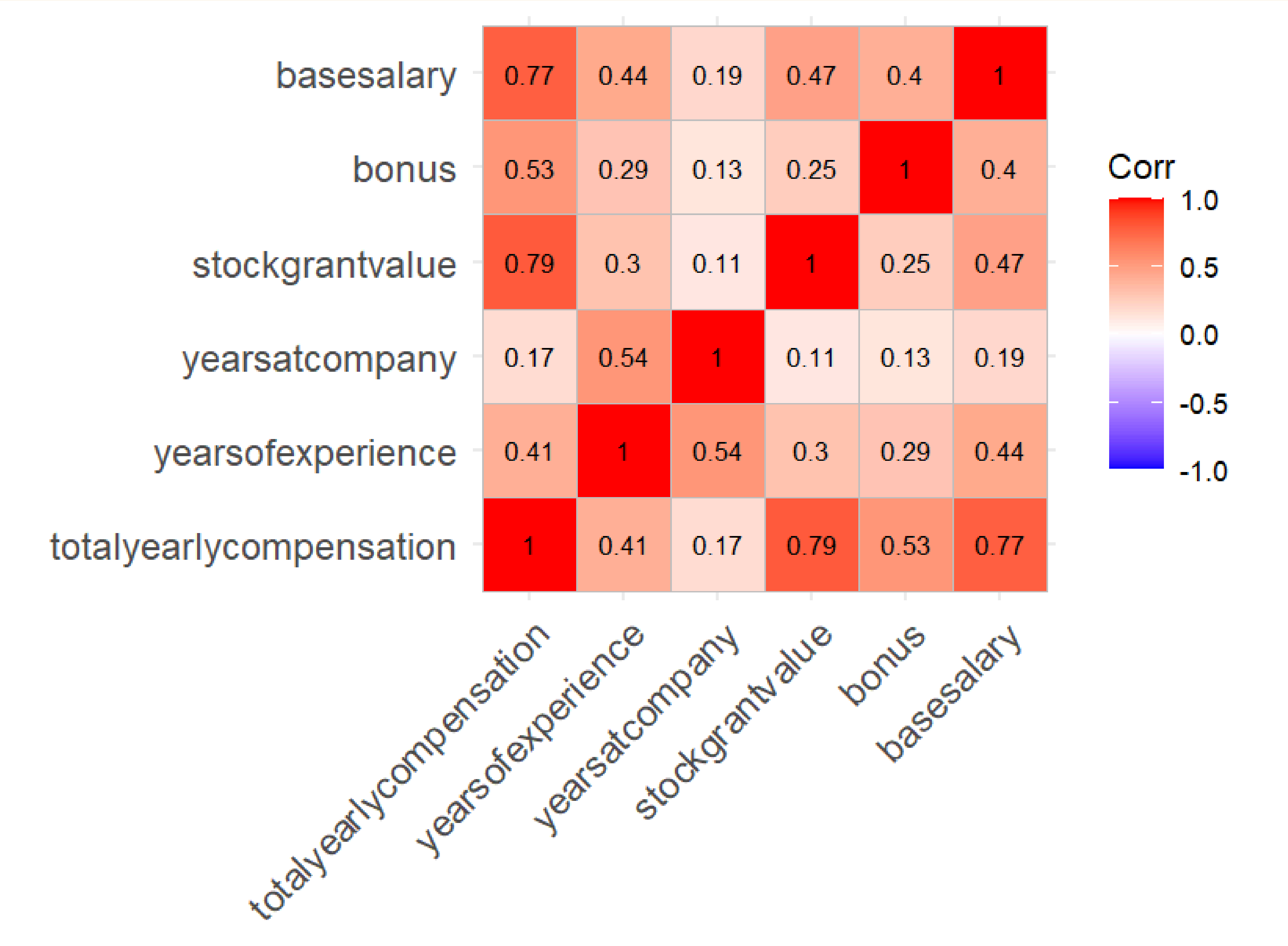
03 Analysis of Variables

Scatterplot Matrix

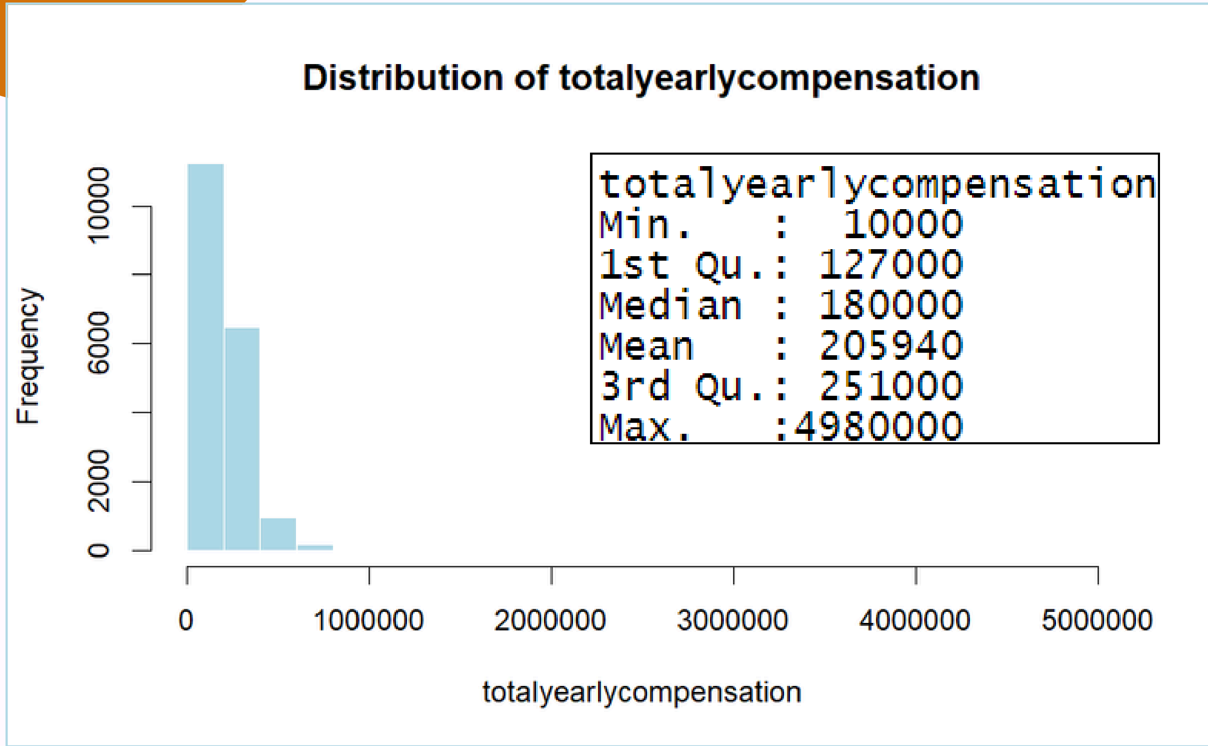


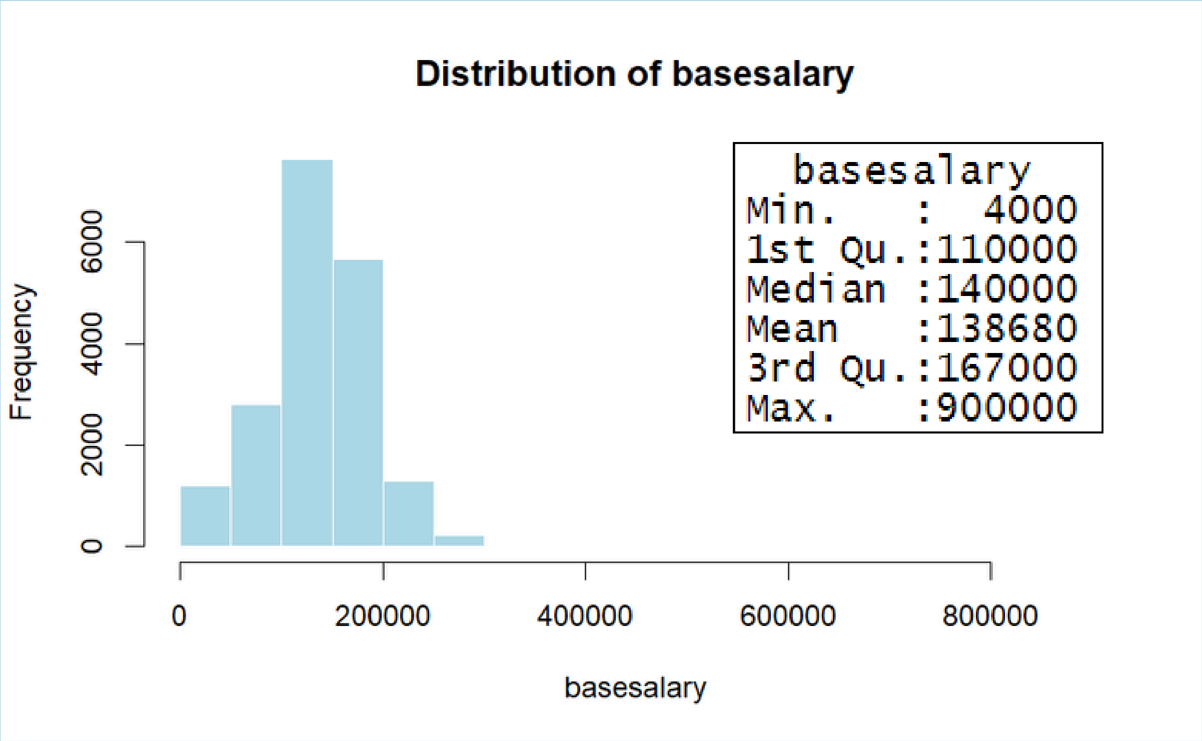
03 Analysis of Variables

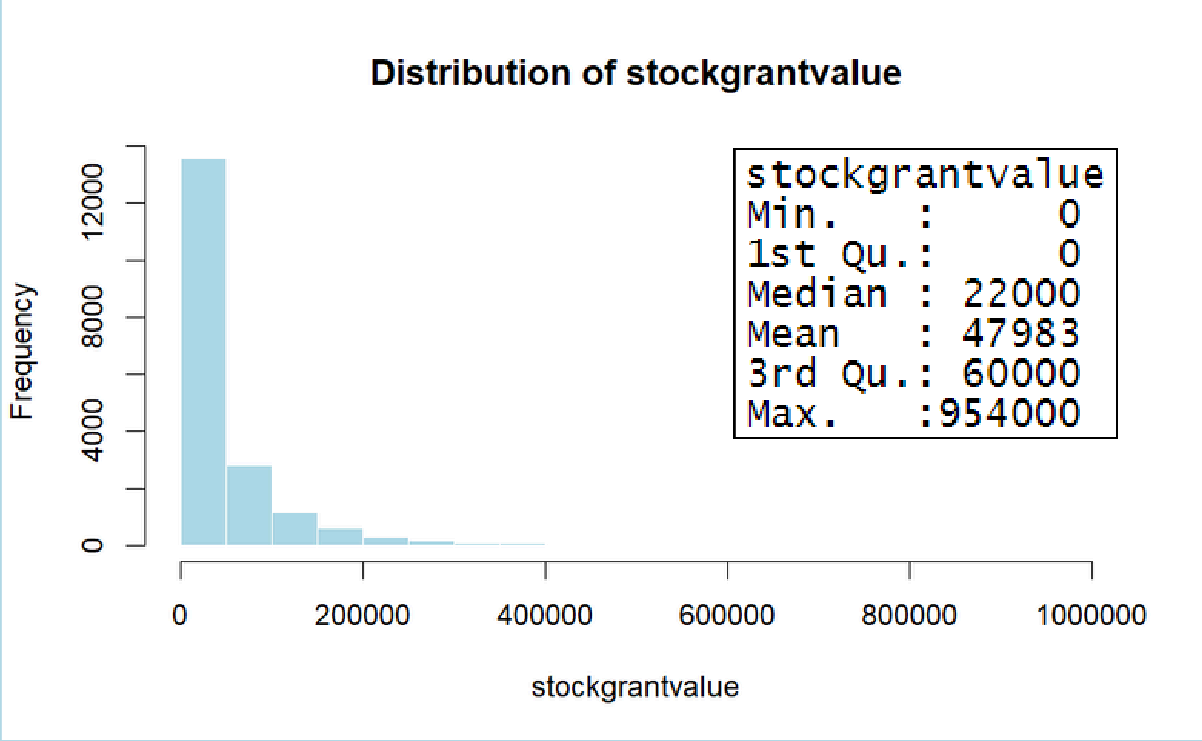
Correlation Heatmap

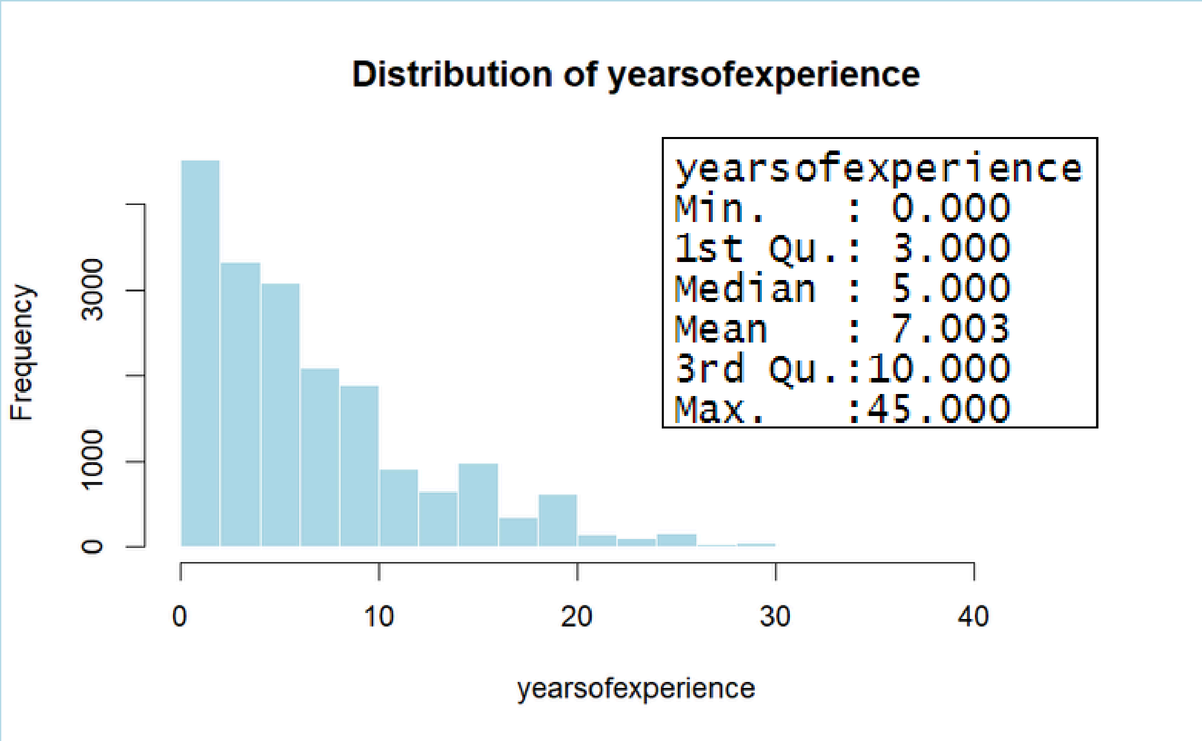


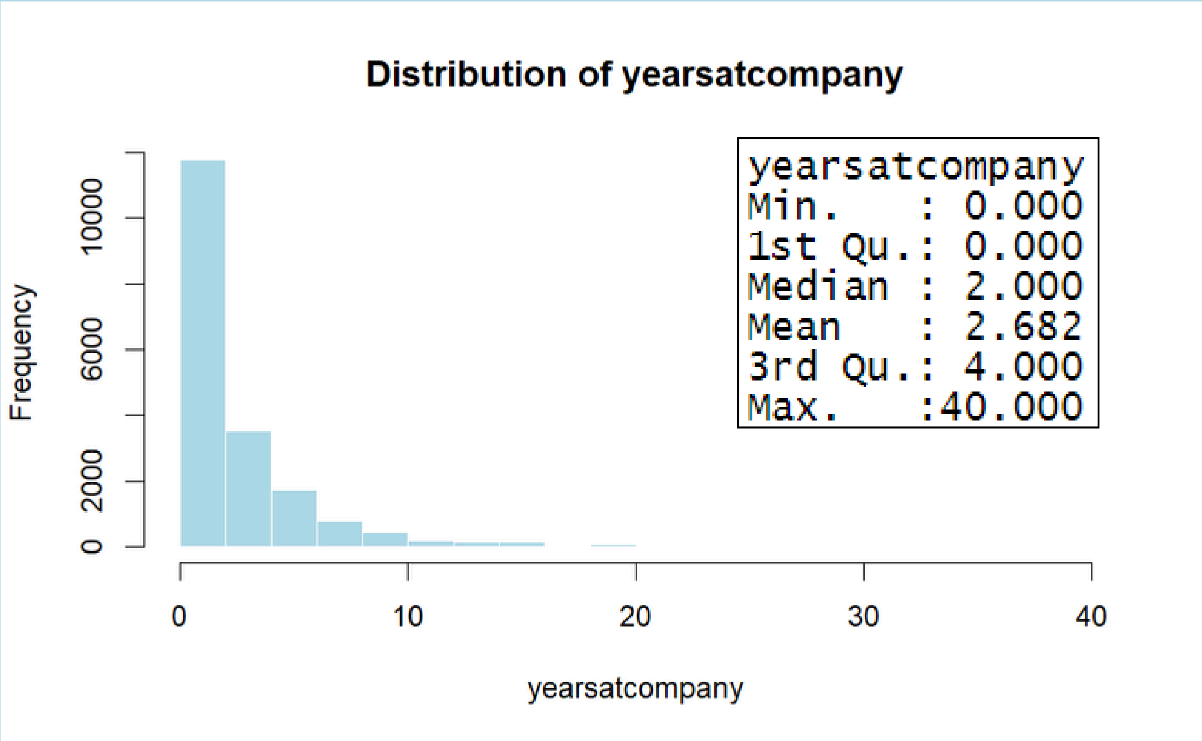
Numerical Variable Distribution

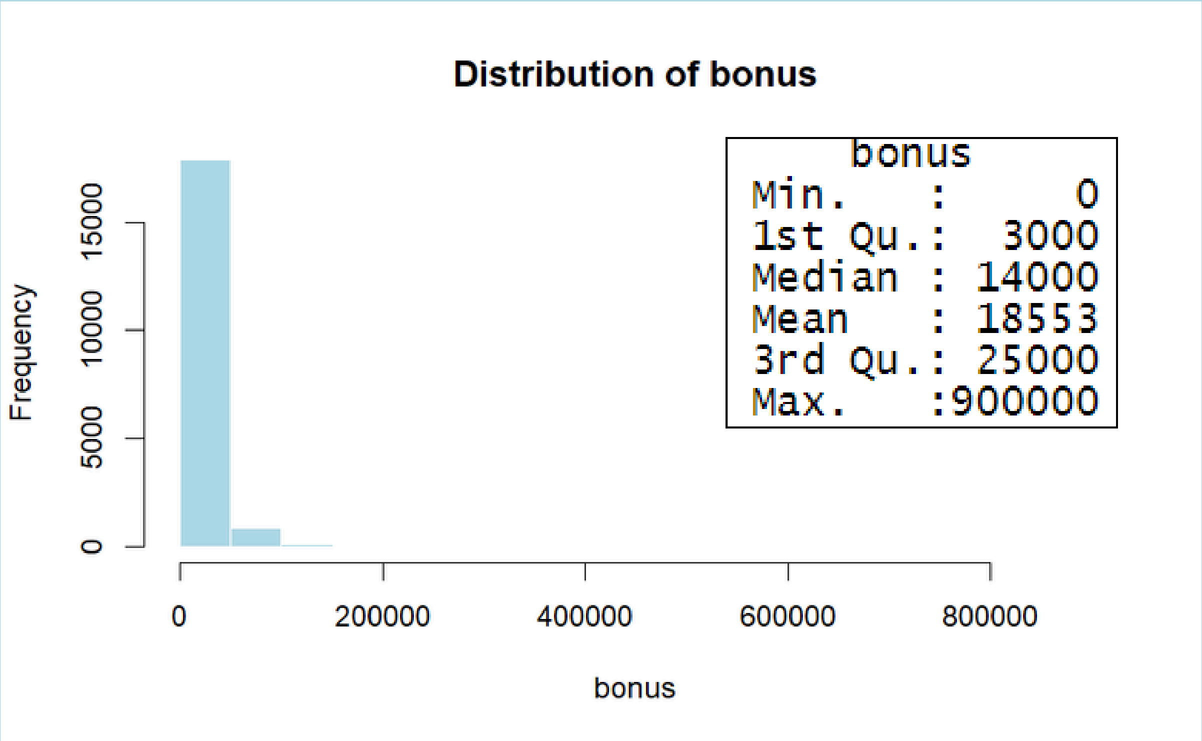






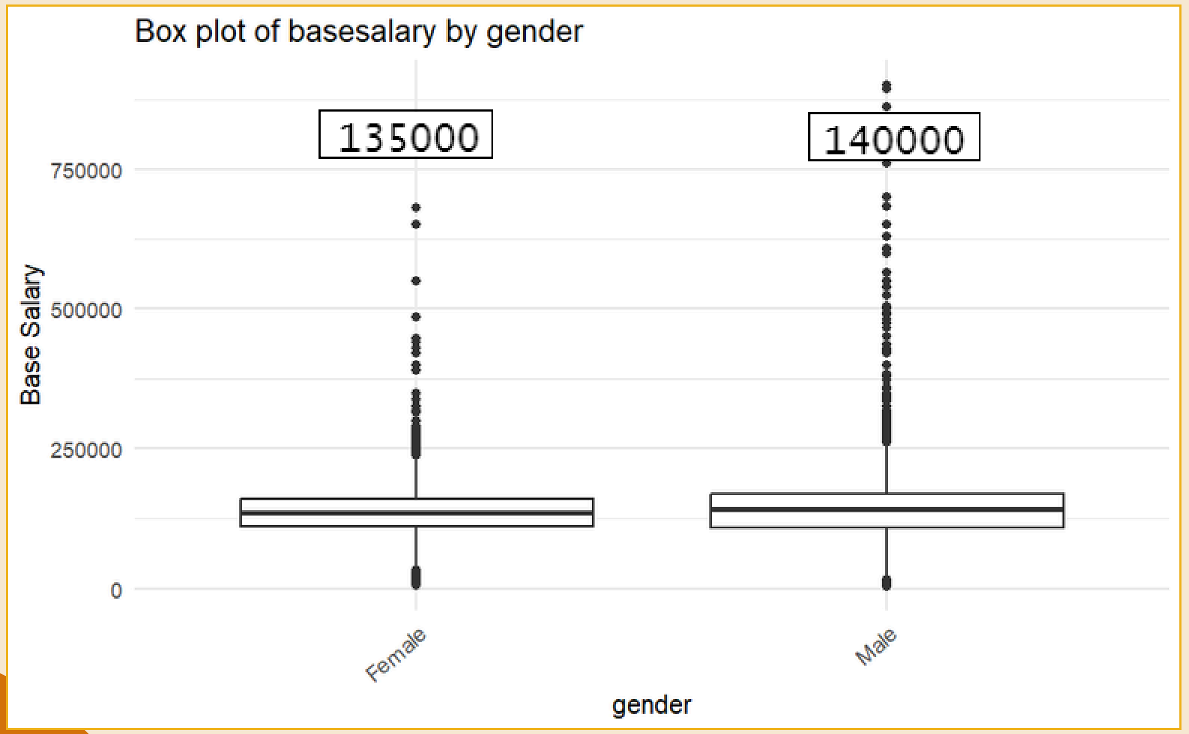




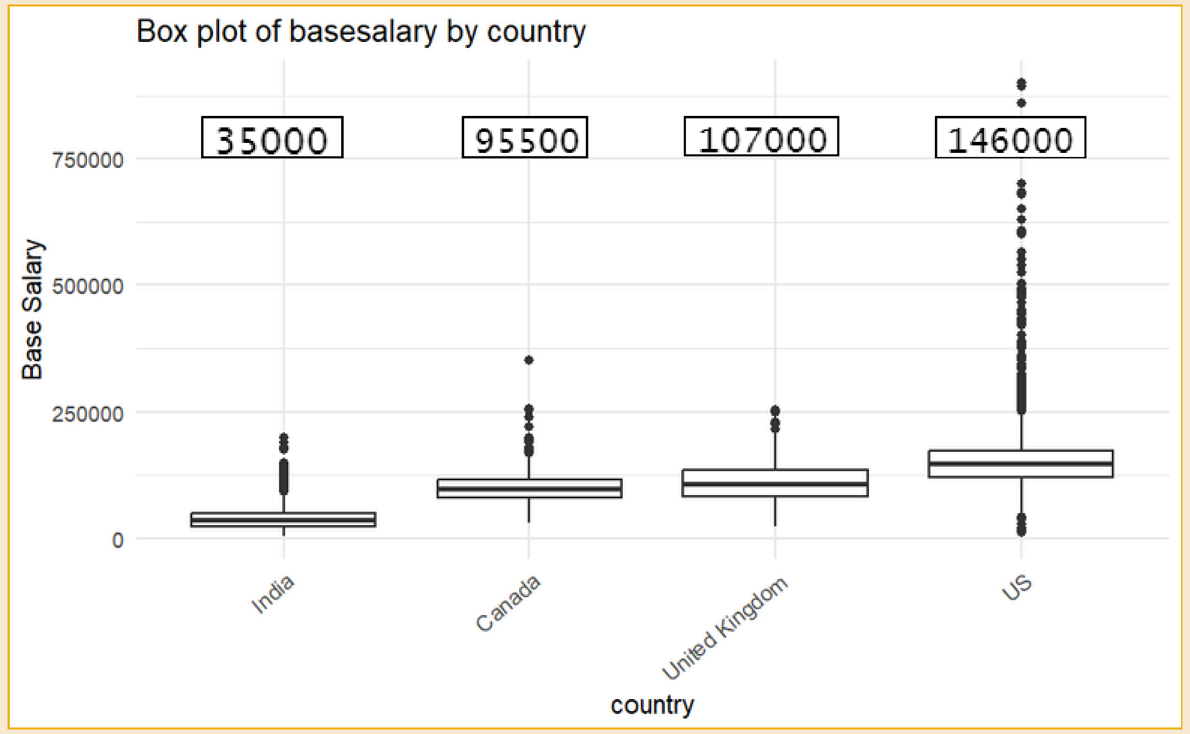


Categorical Variable Distribution

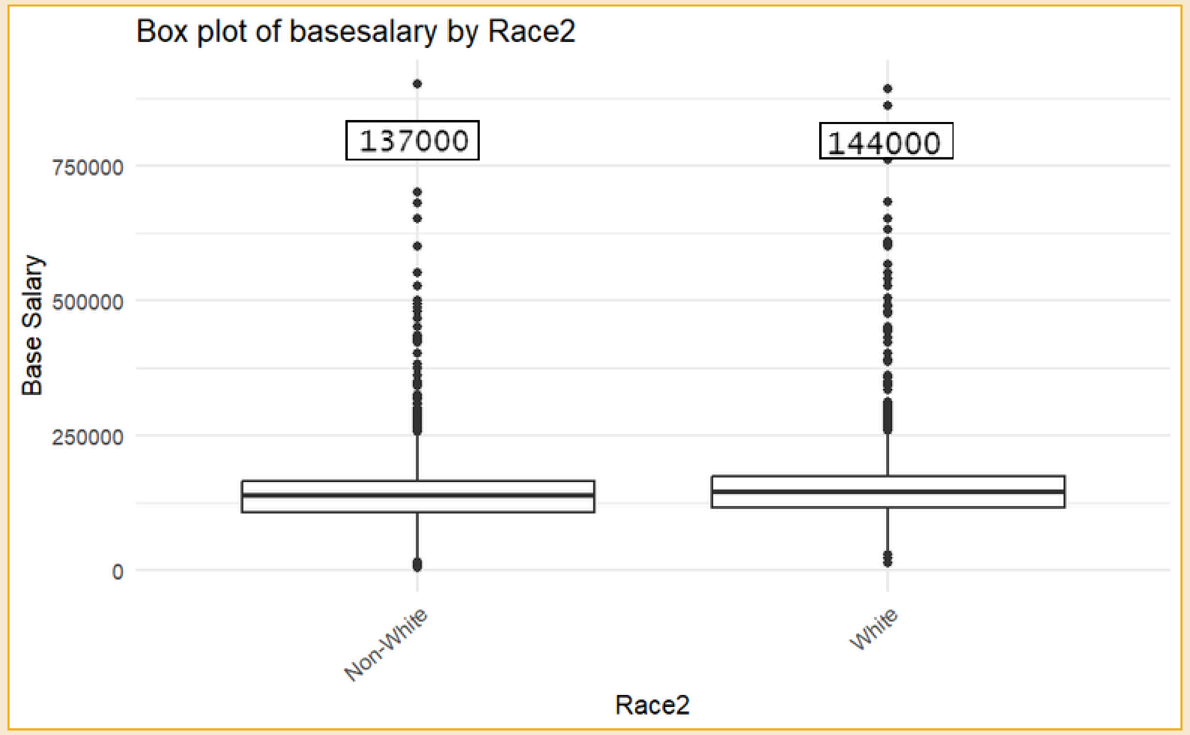
Gender



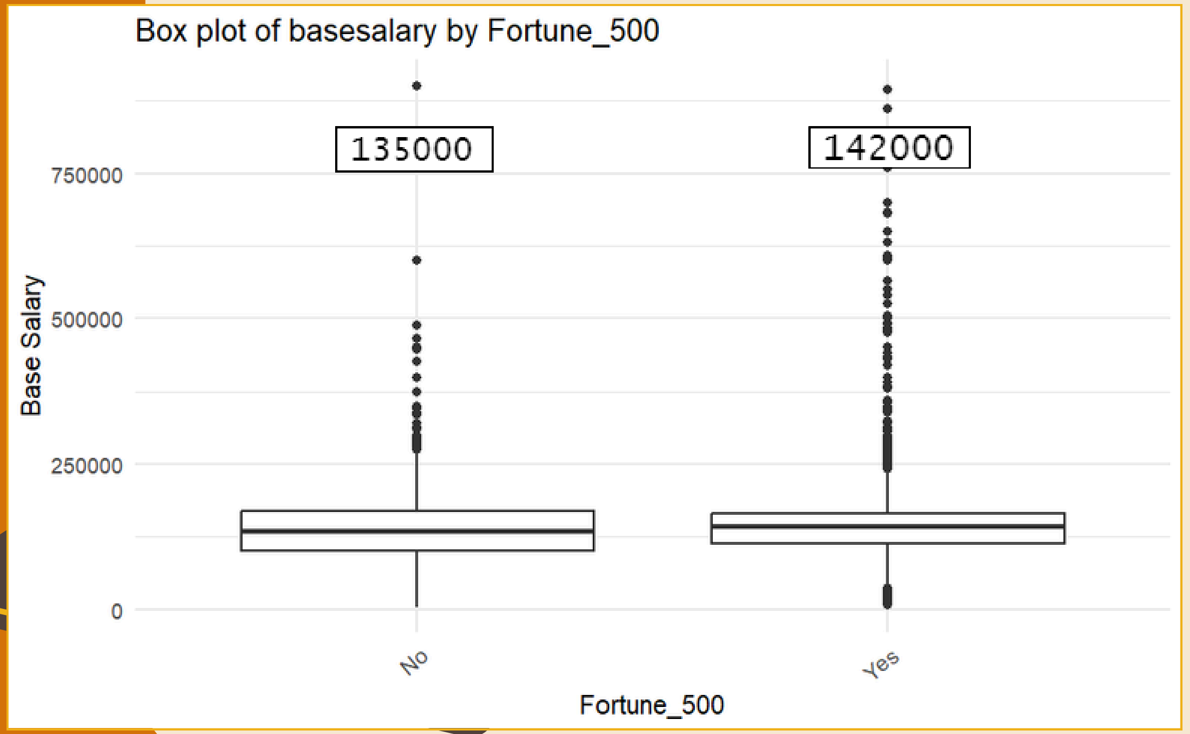
Country



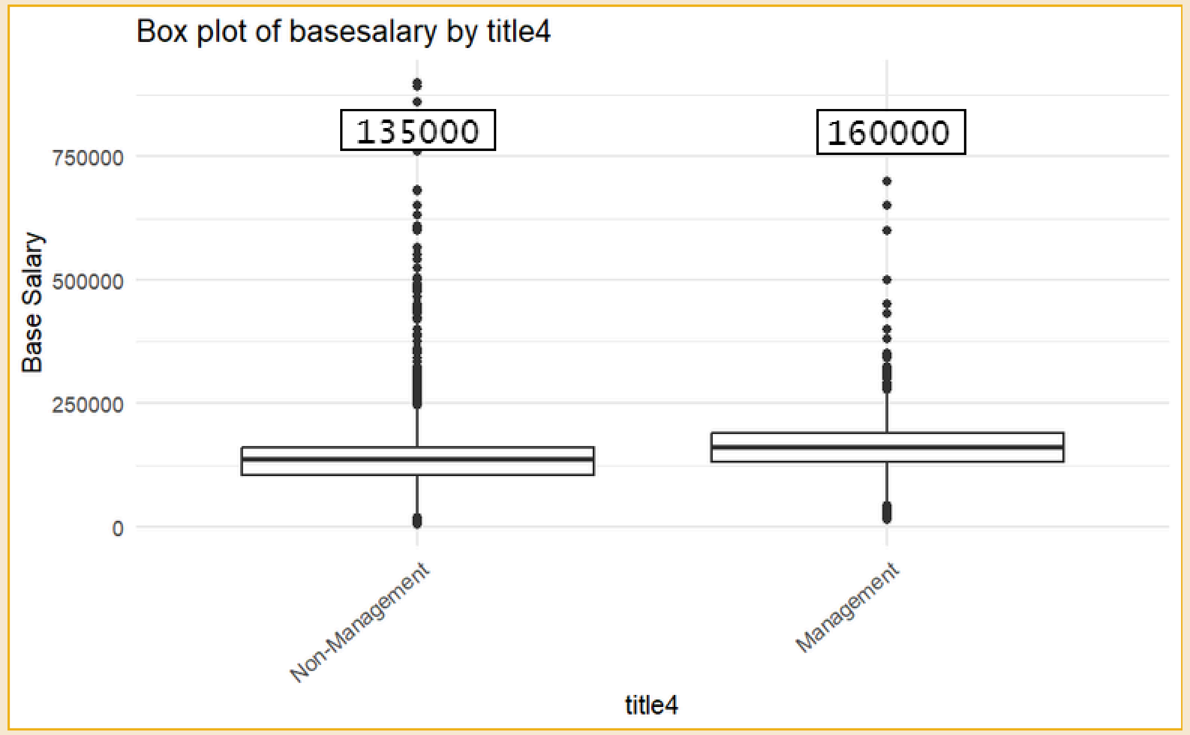
Race



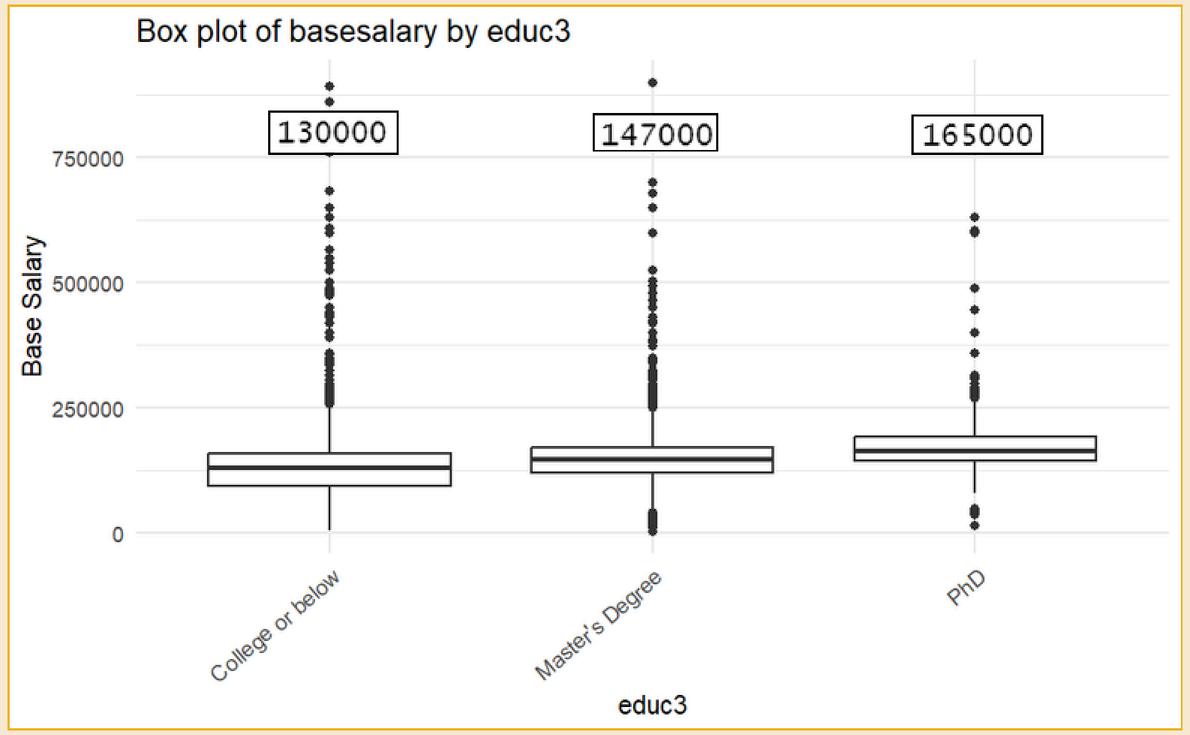
Fortune 500



Title



Education



Multicollinearity

```
lm(formula = totalyearlycompensation ~ basesalary + yearsofexperience +
    yearsatcompany + stockgrantvalue + bonus + gender + Race2 +
    educ3 + country + Fortune_500 + title4, data = STEM2)

Residuals:
    Min      1Q  Median      3Q     Max
-185329   -2878     141    2560 4339340

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -7.165e+03  1.945e+03  -3.683 0.000231 ***
basesalary             1.072e+00  9.485e-03 113.056  < 2e-16 ***
yearsofexperience     -2.282e+02  8.348e+01  -2.733 0.006275 **
yearsatcompany         5.077e+01  1.234e+02   0.412 0.680705
stockgrantvalue        9.592e-01  5.267e-03 182.124  < 2e-16 ***
bonus                  1.203e+00  1.562e-02  77.055  < 2e-16 ***
genderFemale          -3.890e+02  9.085e+02  -0.428 0.668530
Race2White             1.858e+01  7.813e+02   0.024 0.981026
educ3Master's Degree  -4.331e+01  7.446e+02  -0.058 0.953617
educ3PhD              -3.373e+03  1.712e+03  -1.970 0.048832 *
countryIndia           4.479e+03  2.182e+03   2.052 0.040149 *
countryUnited Kingdom -1.457e+03  2.711e+03  -0.537 0.590936
countryUS             -3.842e+03  1.803e+03  -2.132 0.033047 *
Fortune_500Yes         3.245e+02  7.171e+02   0.452 0.650942
title4Management       2.900e+03  9.799e+02   2.960 0.003082 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47160 on 18910 degrees of freedom
Multiple R-squared:  0.8793,    Adjusted R-squared:  0.8792
F-statistic:  9842 on 14 and 18910 DF,  p-value: < 2.2e-16
```

- Global utility of the model is highly significant, but t-tests for individual beta's are insignificant.
- Negative values for education (PhD, Masters), country (US) even when we expect a positive relationship against the totalyearlycompensation (Y).

	GVIF	Df	GVIF^(1/(2*Df))
basesalary	2.253280	1	1.501093
yearsofexperience	1.995267	1	1.412539
yearsatcompany	1.454977	1	1.206224
stockgrantvalue	1.340114	1	1.157633
bonus	1.252679	1	1.119231
gender	1.032173	1	1.015959
Race2	1.107654	1	1.052451
educ2	1.170617	3	1.026603
country2	1.486755	1	1.219326
Fortune_500	1.038533	1	1.019084
title2	1.119979	1	1.058291

- Although VIF values did not exceed 10, suggesting that multicollinearity is not a significant concern, it is important to note that basesalary, stockgrantvalue, and bonus together form the totalyearlycompensation.

04 Model Building

- Variable Transformations
- Interaction Terms
- Main Effects
- Stepwise Regression
- All-Possible-Regression Selection

4.1 Model with Quantitative Variables

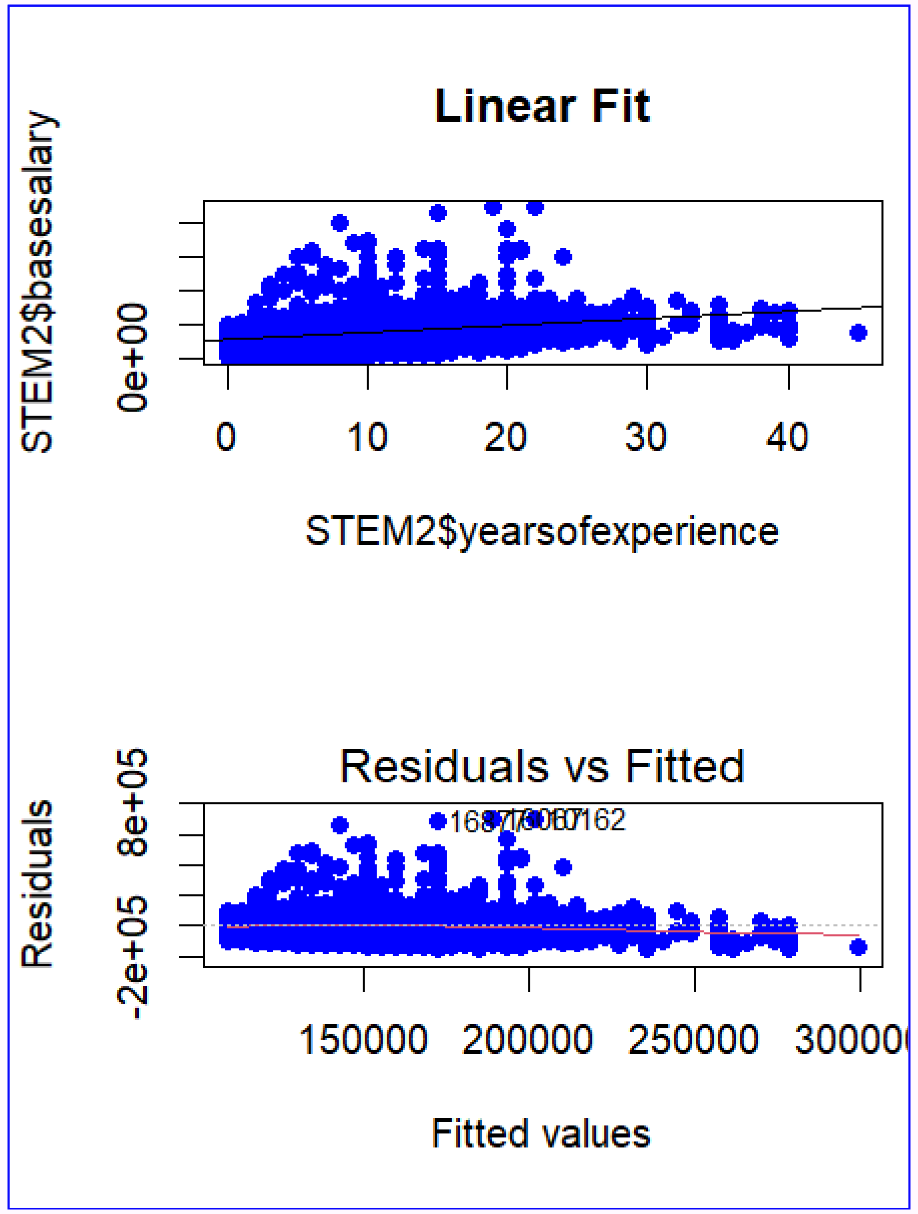
lm(basesalary~yearsofexperience+<variable transformation>)

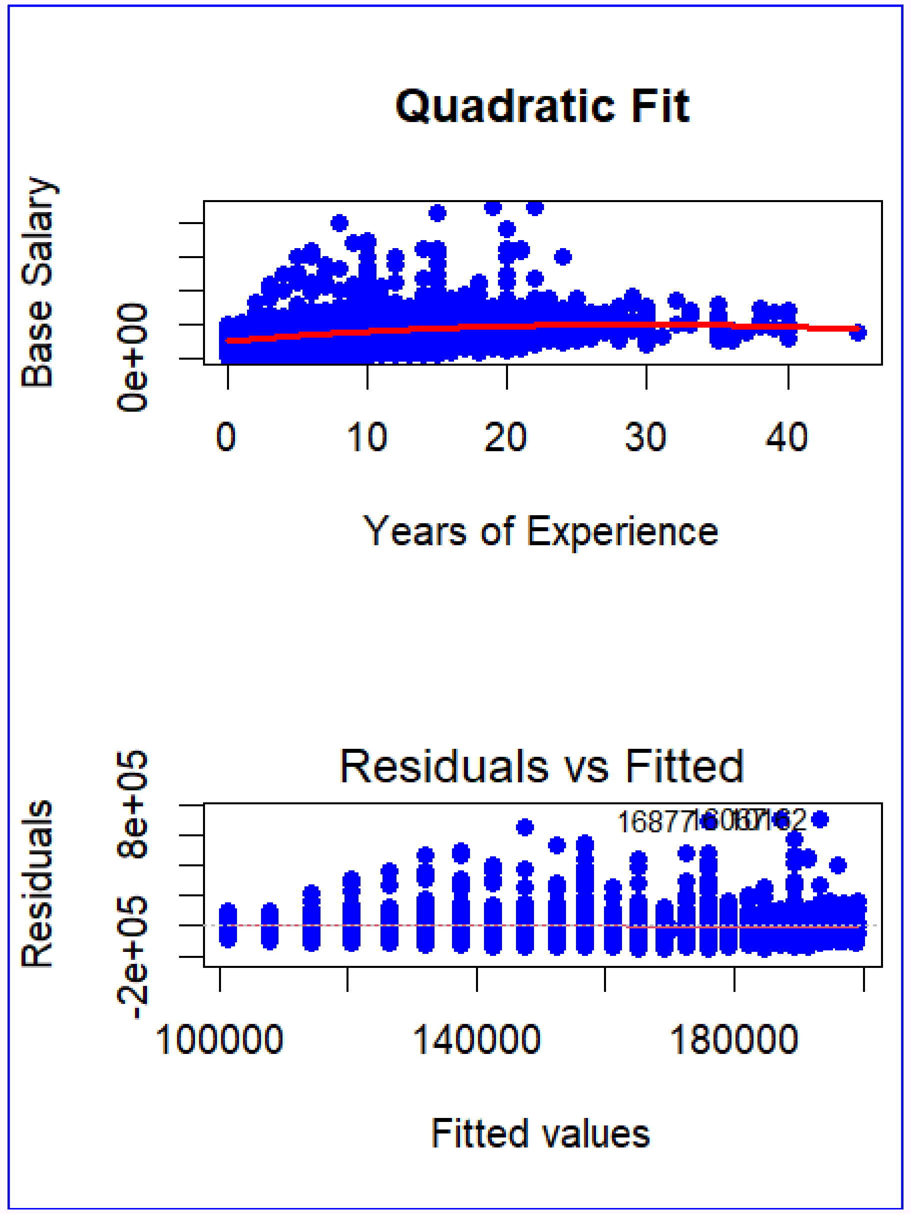
Term	Pr(>|t|)	RSE	Adj. R-squared
Main Effect	Significant	50,700	19.4%
Quadratic	Significant	50,360	20.48%
Log of X	Significant	50,630	19.62%
Square Root	Significant	50,480	20.28%

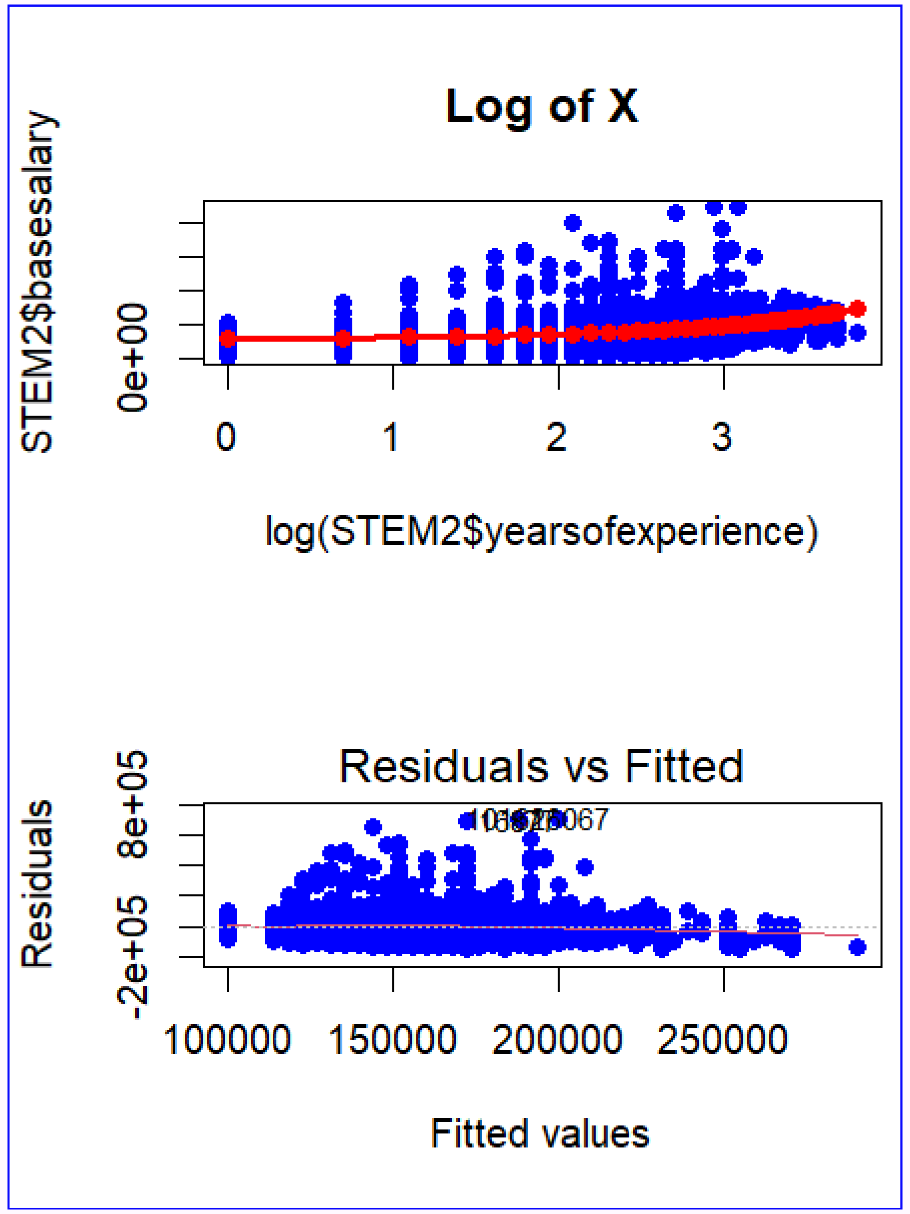
ANOVA tests conclude that all transformations contribute to the prediction of y.

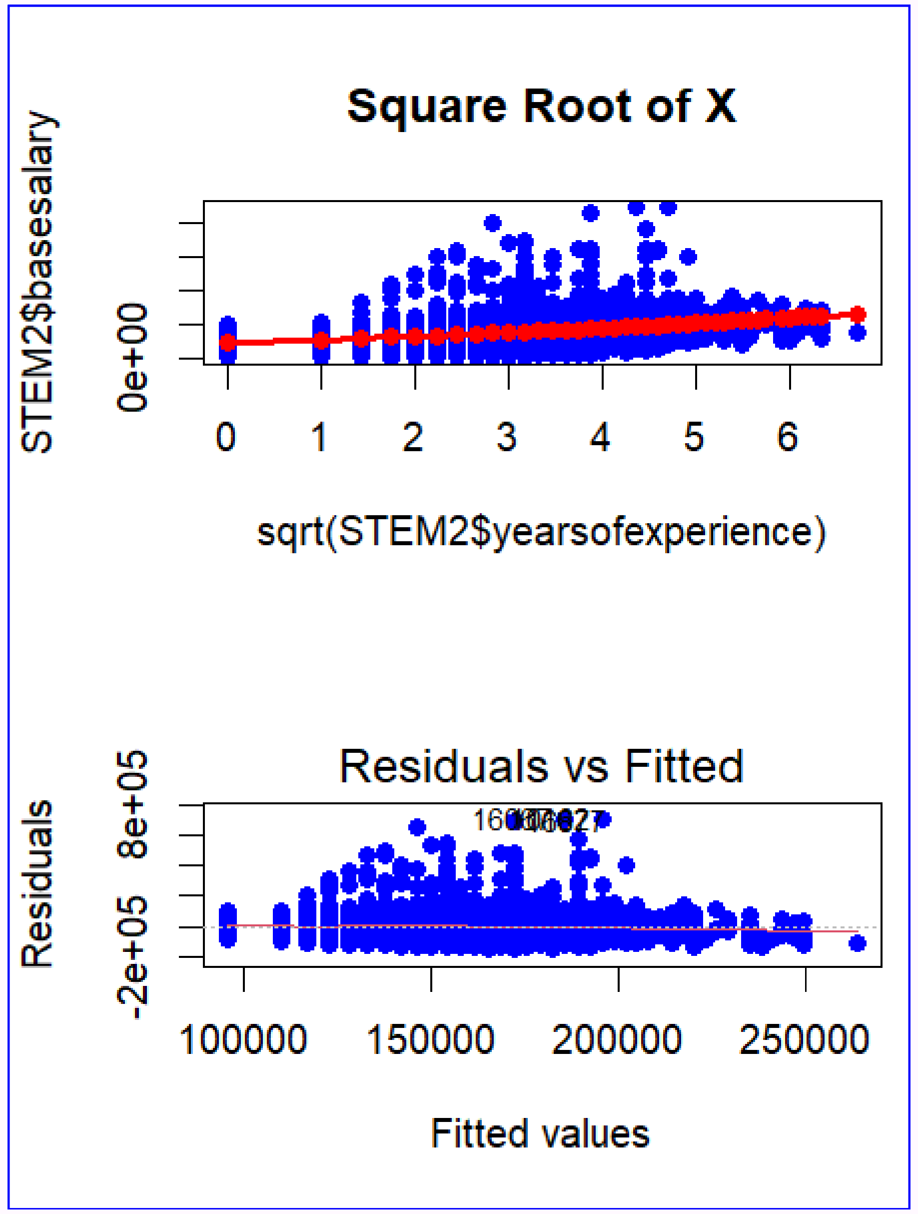
4.1 Model with Quantitative Variables

lm(basesalary~yearsofexperience+<variable transformation>)









4.1 Model with Quantitative Variables

lm(basesalary~yearsatcompany+<variable transformation>)

Term	Pr(>|t|)	RSE	Adj. R-squared
Main Effect	Significant	55,420	3.68%
Quadratic	Significant	55,410	3.71%
Log of X	Significant	55,340	3.96%
Square Root	Significant	55,390	3.80%

ANOVA tests conclude that all transformations contribute to the prediction of y.

4.2 Model with Qualitative Variables

lm(basesalary~<predictor 1>+<predictor 2>)
vs.
lm(basesalary~<predictor 1>*<predictor 2>)

Term	Pr(>|t|)	Diff. in RSE	Diff. in Adj. R-sq
yearsofexperience:yearsatcompany	Significant	-44	0.15%
yearsofexperience:title4	Significant	-37	0.13%
yearsofexperience:Race2	Significant	-6	0.03%
yearsofexperience:Fortune_500	Significant	-12	0.03%
yearsofexperience:educ3	Significant	-58	0.19%
yearsofexperience:country	Significant	-2	0.01%

ANOVA tests conclude that all transformations contribute to the prediction of y.

4.2 Model with Qualitative Variables

lm(basesalary~<predictor 1>+<predictor 2>)
vs.
lm(basesalary~<predictor 1>*<predictor 2>)

Term	Pr(>|t|)	Diff. in RSE	Diff. in Adj. R-sq
yearsatcompany:title4	Significant	-55	0.17%
yearsatcompany:gender	Significant	-6	0.02%
yearsatcompany:Fortune_500	Significant	-30	0.10%
yearsatcompany:educ3	Significant	-73	0.24%
Race2:title4	Significant	-8	0.04%
Race2:educ3	Significant	-89	0.29%

ANOVA tests conclude that all transformations contribute to the prediction of y.

4.2 Model with Qualitative Variables

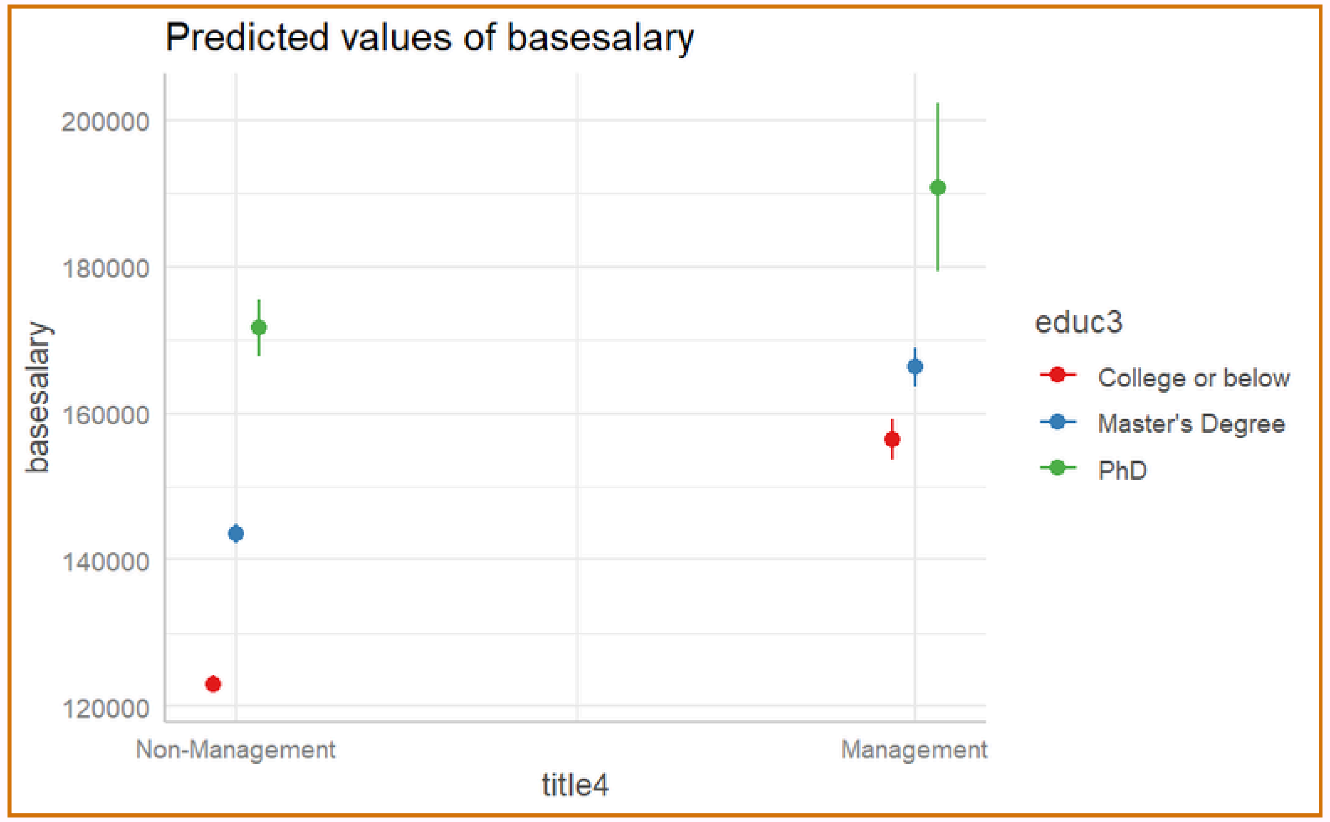
lm(basesalary~<predictor 1>+<predictor 2>)
vs.
lm(basesalary~<predictor 1>*<predictor 2>)

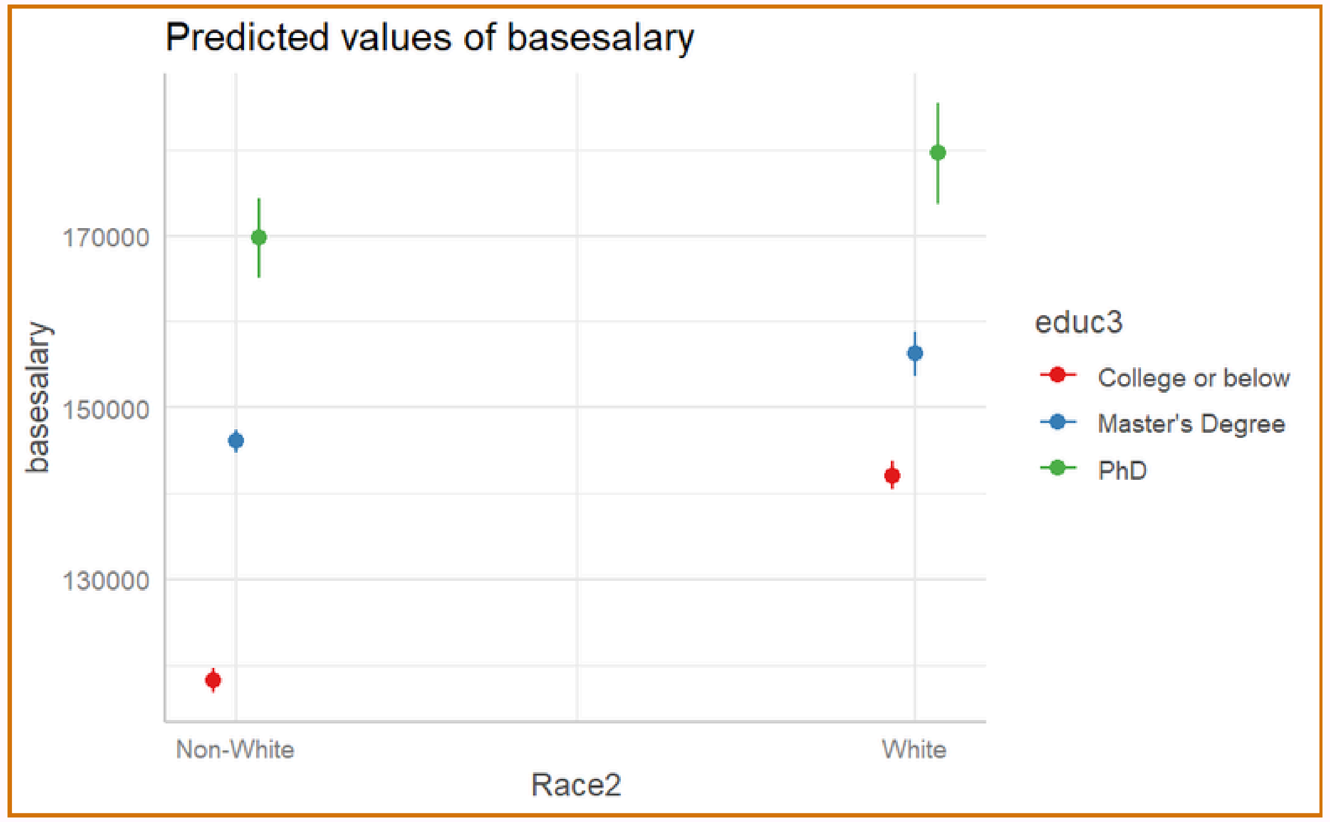
Term	Pr(>|t|)	Diff. in RSE	Diff. in Adj. R-sq
Race2:gender	Significant	-25	0.08%
gender:title4	Significant	-15	0.04%
educ3:gender	Significant	-9	0.02%
educ3:title4	Significant	-33	0.12%
educ3:Fortune_500	Significant	-14	0.04%
educ3:country	Significant	-3	0.01%

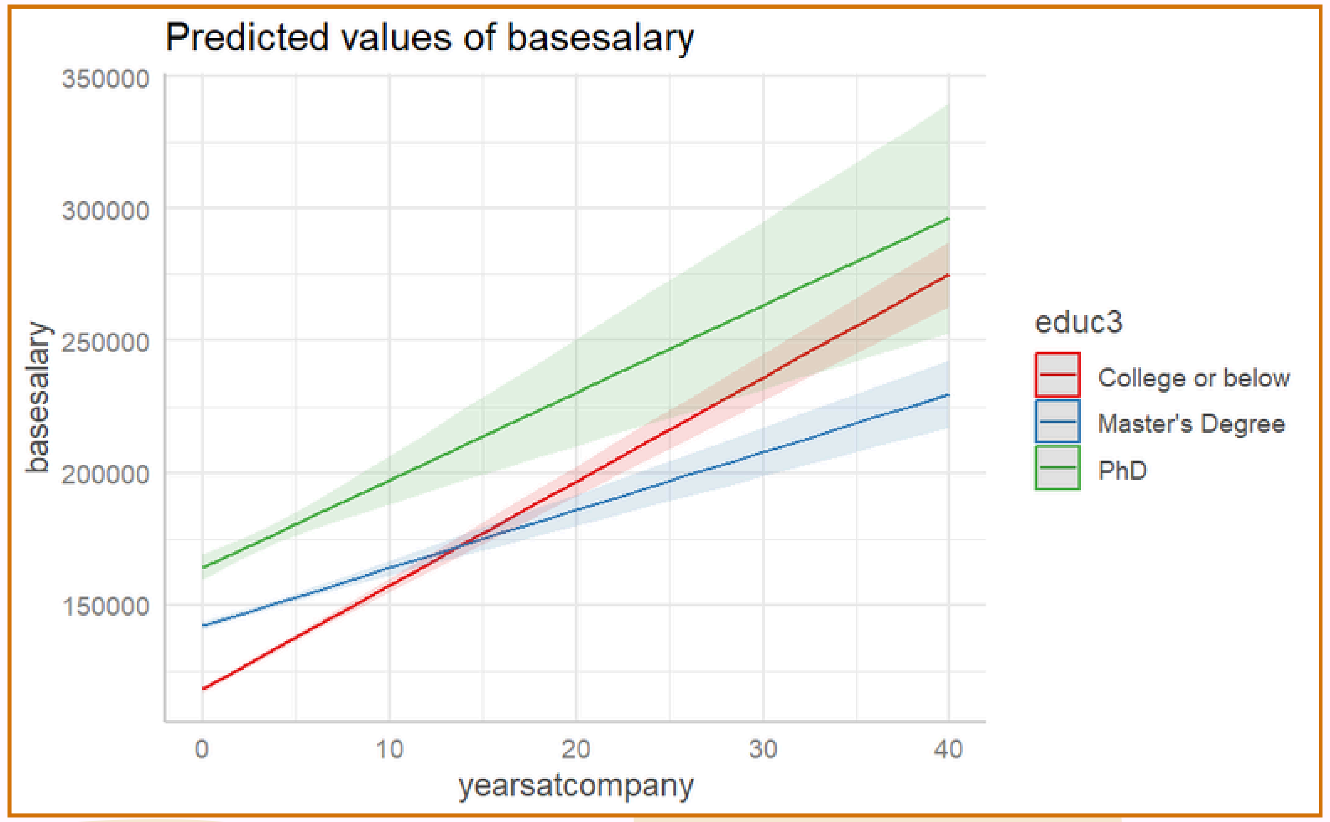
ANOVA tests conclude that all transformations contribute to the prediction of y.

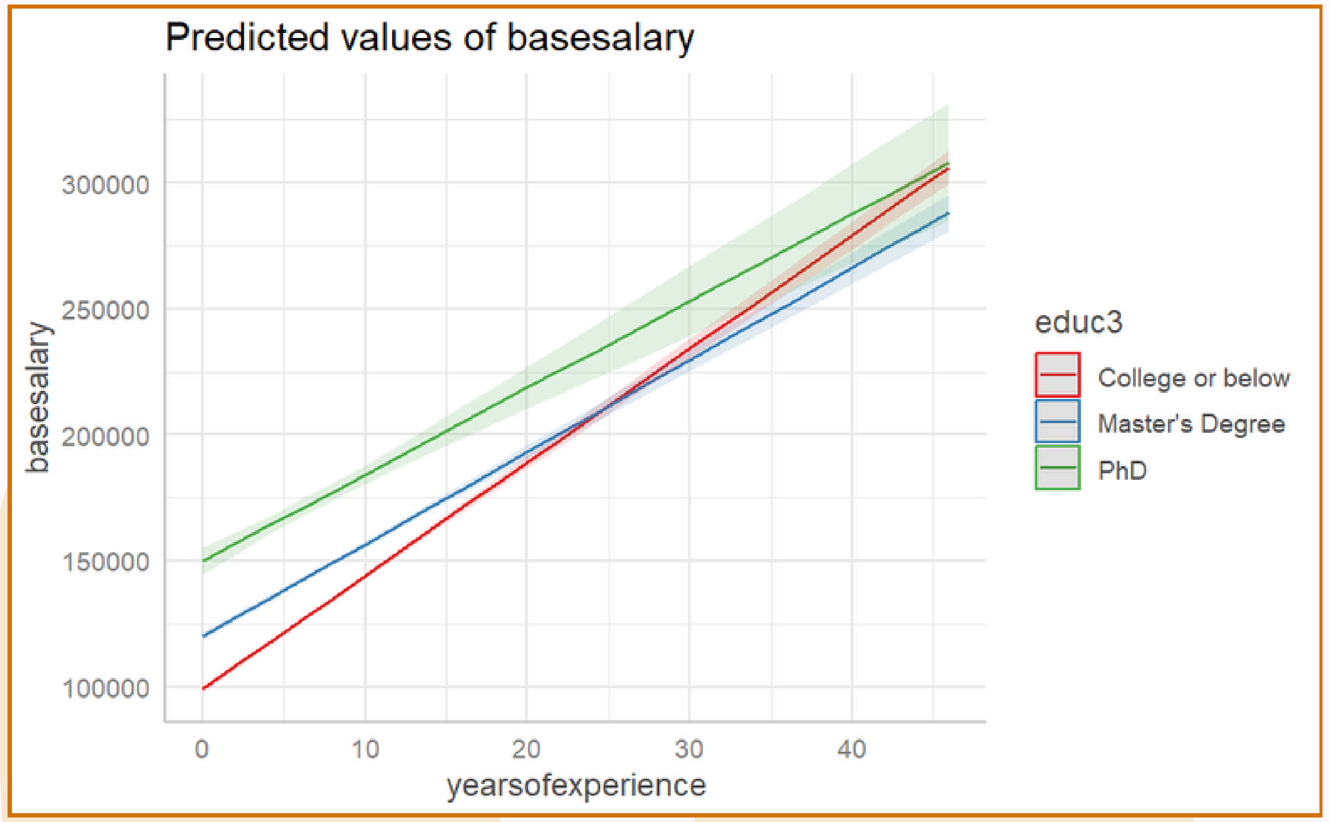
4.3 Interaction Plots

Show interactions with > 0.10% improvement



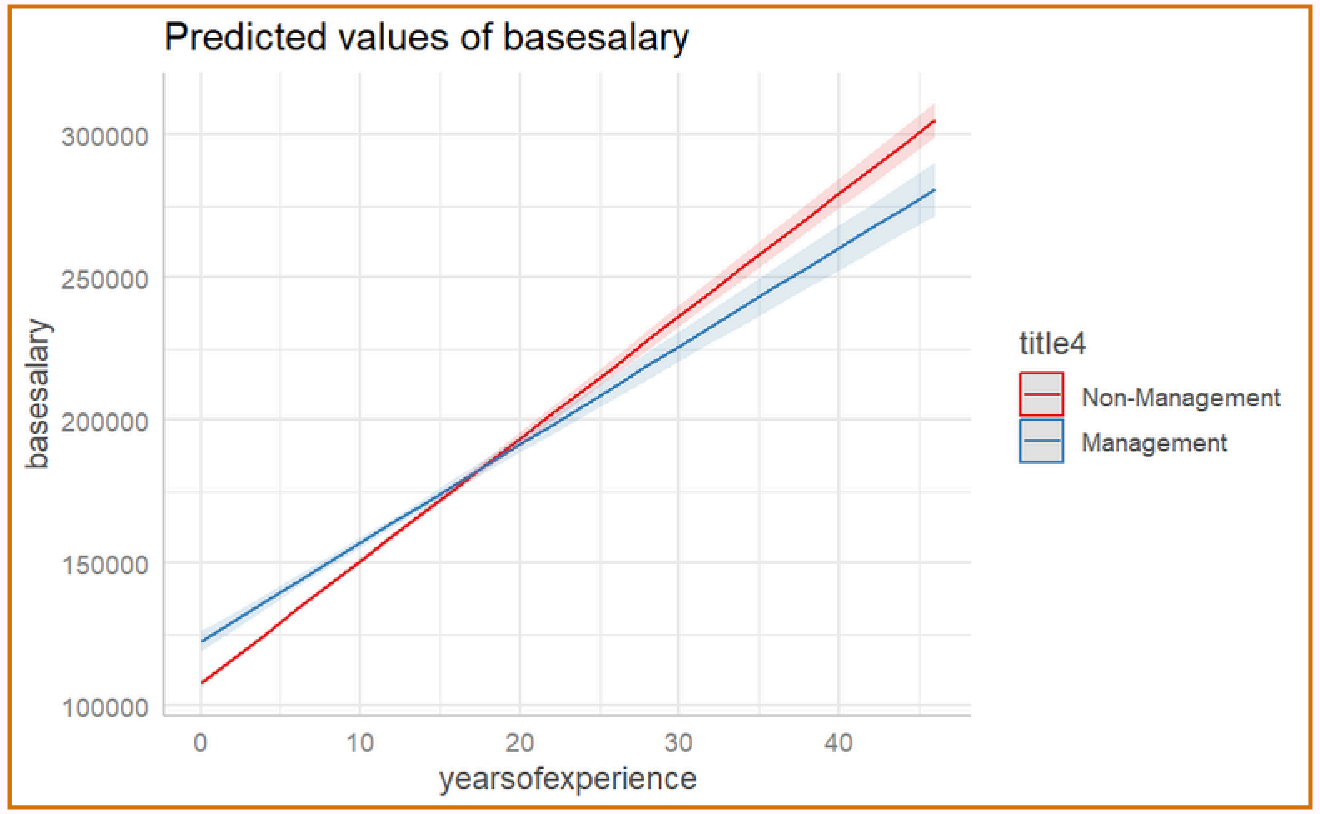


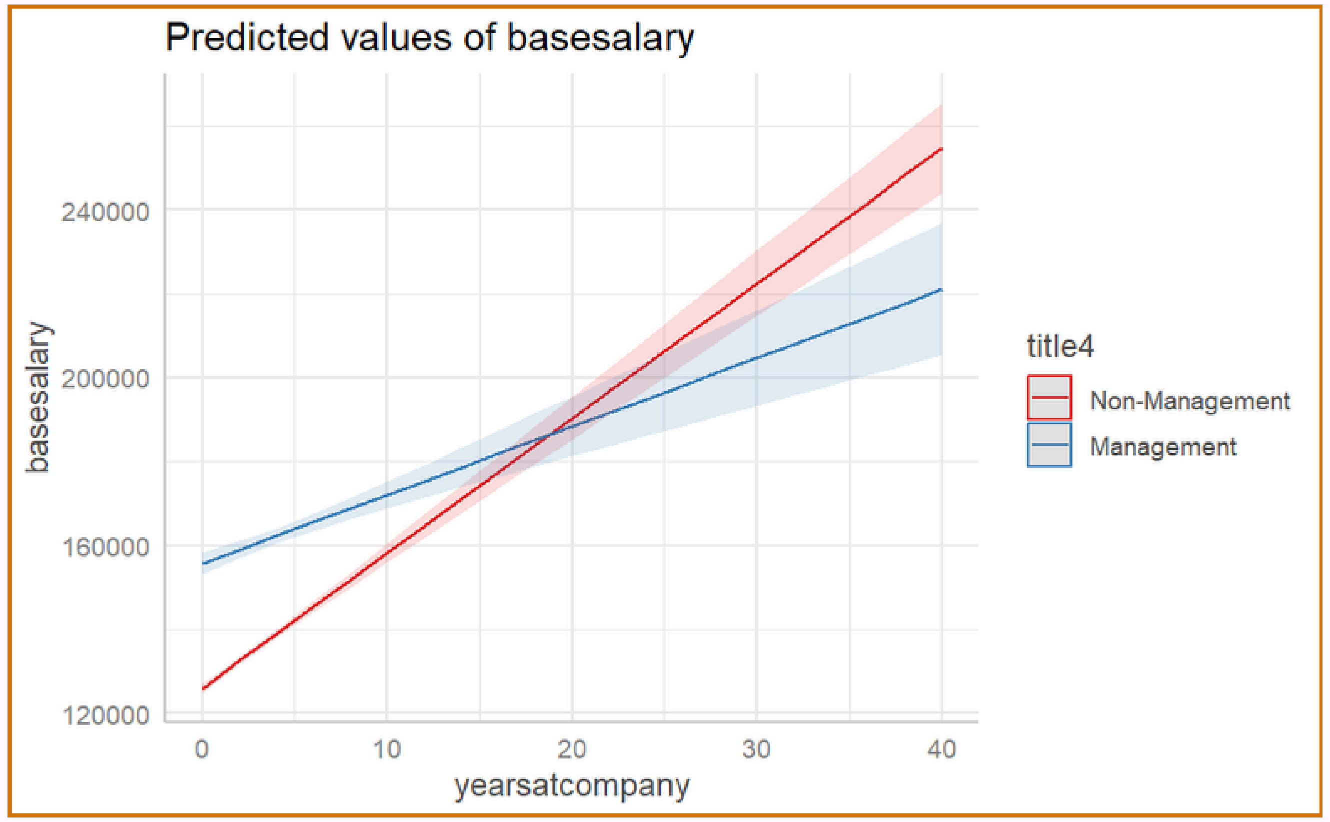


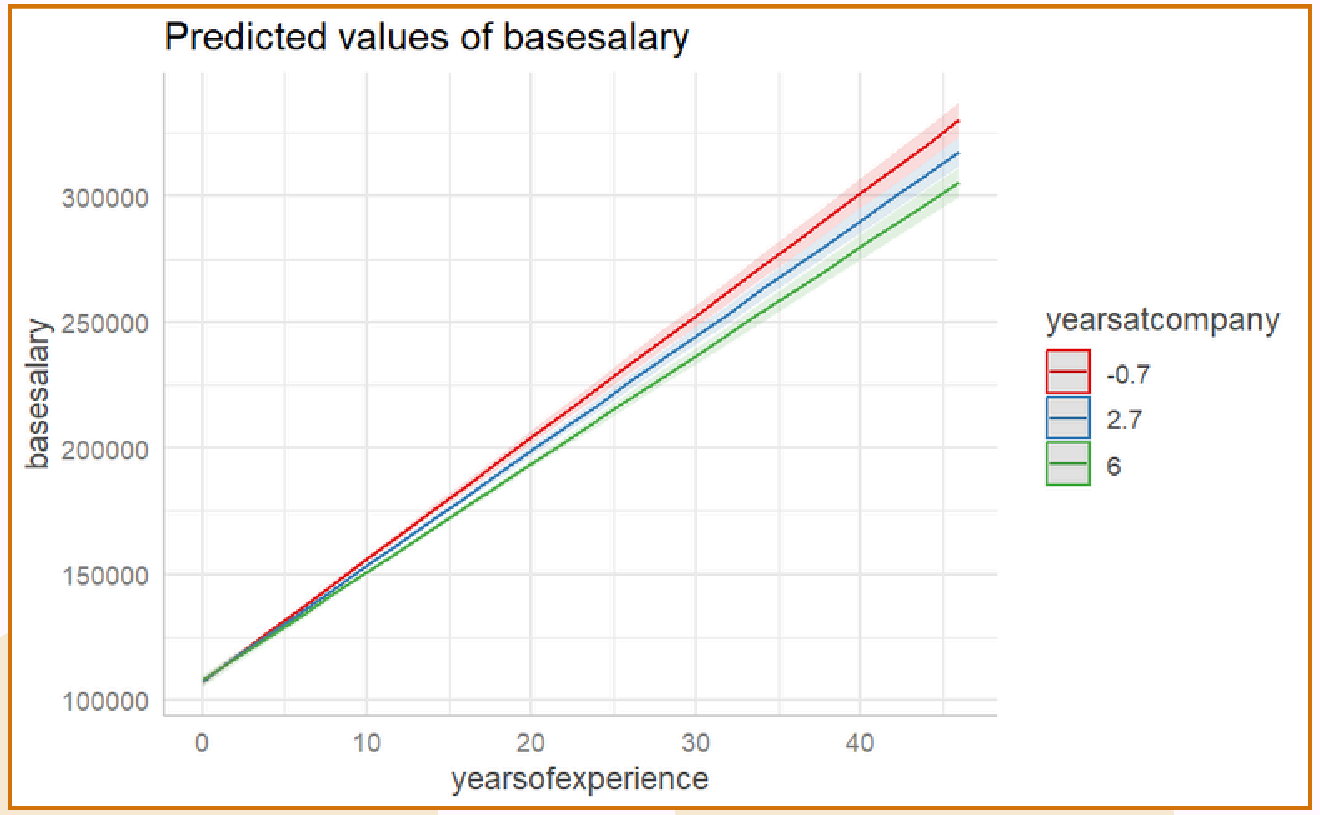


4.3 Interaction Plots

Show interactions with > 0.10% improvement





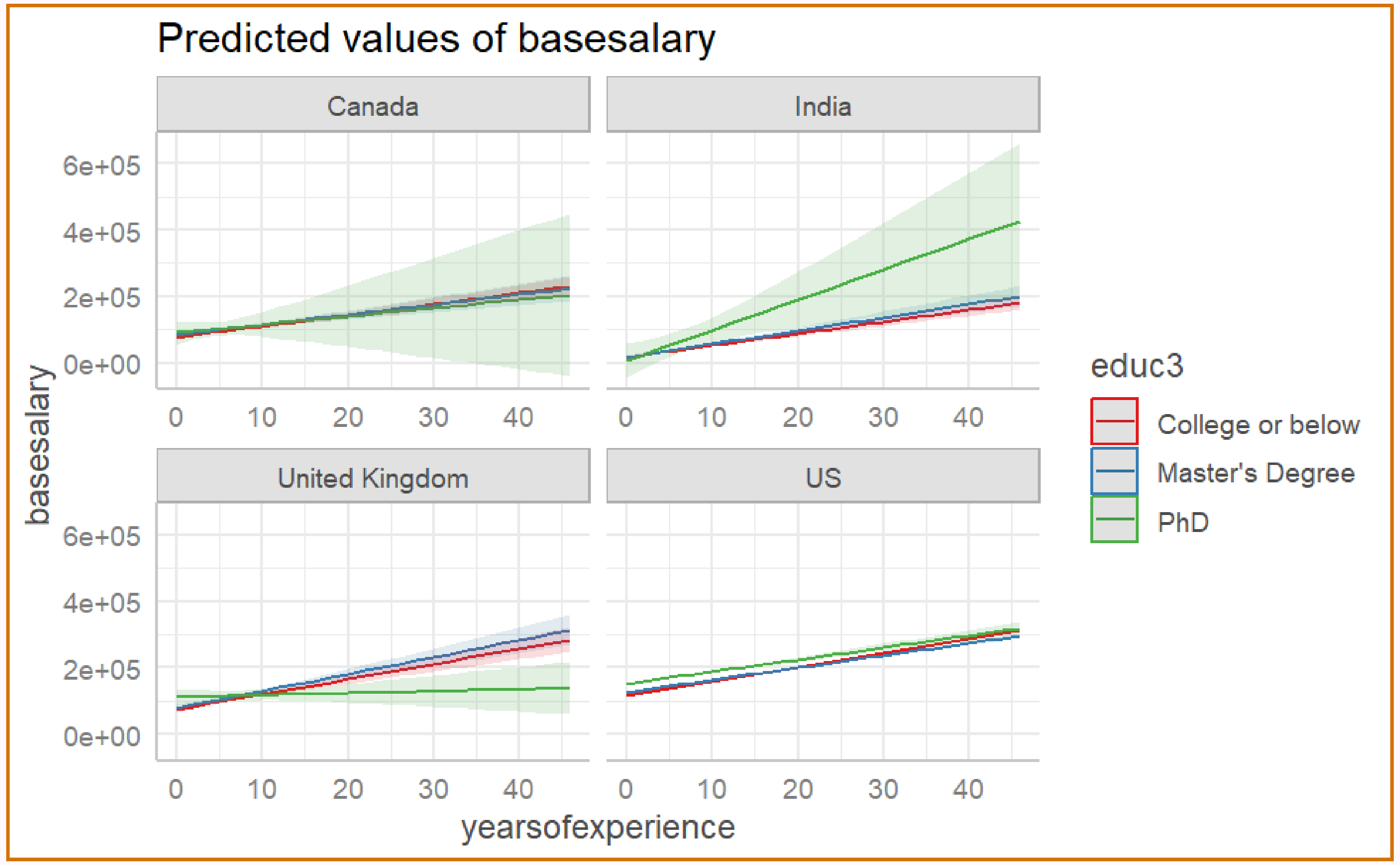


4.3 Interaction Plots

Exploring 3-way and 4-way interactions (Limitation: Rank deficiency issue)

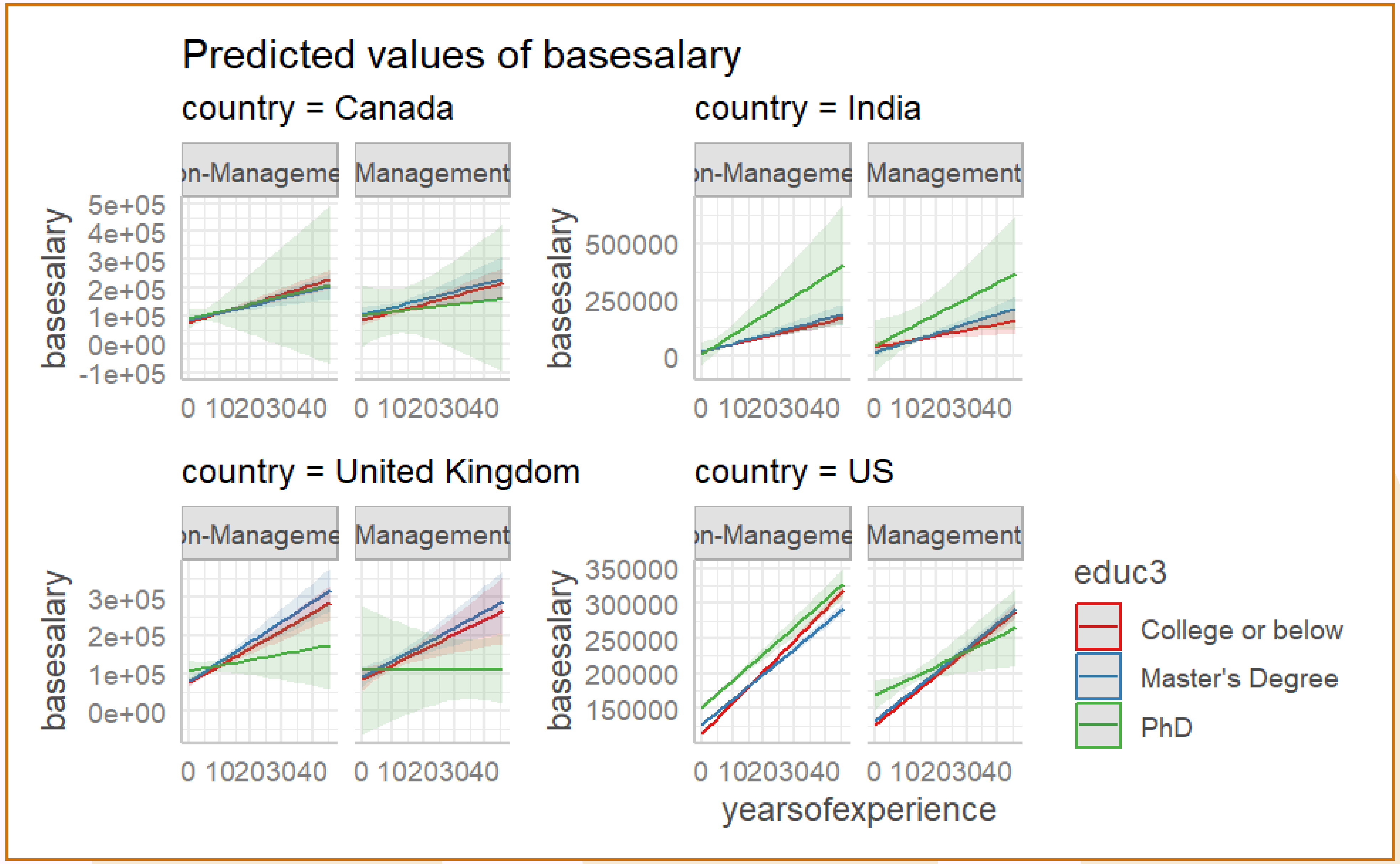
Three-way interaction

lm(basesalary ~ yearsofexperience*educ3*country)



Four-way interaction

lm(basesalary ~ yearsofexperience*educ3*title4*country)



Main Effects Model

```
Call:
lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
    gender + Race2 + educ3 + country + title4 + Fortune_500,
    data = STEM2)

Residuals:
    Min      1Q  Median      3Q     Max
-172208  -21622   -1892   16809  704778

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             74650.07    1534.80  48.638  < 2e-16 ***
yearsofexperience        4360.86      62.58  69.690  < 2e-16 ***
yearsatcompany          -1537.64     103.72 -14.824  < 2e-16 ***
genderFemale            -3285.68     767.68  -4.280 1.88e-05 ***
Race2White              -2462.59     660.46  -3.729 0.000193 ***
educ3Master's Degree     6258.08     628.59   9.956  < 2e-16 ***
educ3PhD                31353.53    1425.72  21.991  < 2e-16 ***
countryIndia           -59812.66    1791.92 -33.379  < 2e-16 ***
countryUnited Kingdom    5644.57    2293.34   2.461 0.013853 *
countryUS               45787.96    1487.98  30.772  < 2e-16 ***
title4Management         4051.01     826.15   4.903 9.49e-07 ***
Fortune_500Yes           -308.92     602.81  -0.512 0.608334
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39920 on 18913 degrees of freedom
Multiple R-squared:  0.5005,	Adjusted R-squared:  0.5002
F-statistic:  1723 on 11 and 18913 DF,  p-value: < 2.2e-16
```

Insignificant

```
Call:
lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
    gender + Race2 + educ3 + country + title4, data = STEM2)

Residuals:
    Min      1Q  Median      3Q     Max
-172329  -21606   -1925   16777  704653

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            74517.36    1512.77  49.259  < 2e-16 ***
yearsofexperience       4360.82      62.57  69.691  < 2e-16 ***
yearsatcompany         -1540.55     103.57 -14.875  < 2e-16 ***
genderFemale           -3287.76     767.66  -4.283 1.85e-05 ***
Race2White             -2444.29     659.48  -3.706 0.000211 ***
educ3Master's Degree    6242.27     627.82   9.943  < 2e-16 ***
educ3PhD               31305.07    1422.55  22.006  < 2e-16 ***
countryIndia          -59832.75    1791.45 -33.399  < 2e-16 ***
countryUnited Kingdom   5622.27    2292.88   2.452 0.014213 *
countryUS              45743.20    1485.38  30.796  < 2e-16 ***
title4Management        4027.31     824.84   4.883 1.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39920 on 18914 degrees of freedom
Multiple R-squared:  0.5005,	Adjusted R-squared:  0.5003
F-statistic:  1895 on 10 and 18914 DF,  p-value: < 2.2e-16
```

Individual t-tests: Significant

Prediction Interval

Moderate Fit

Global F-test: Significant

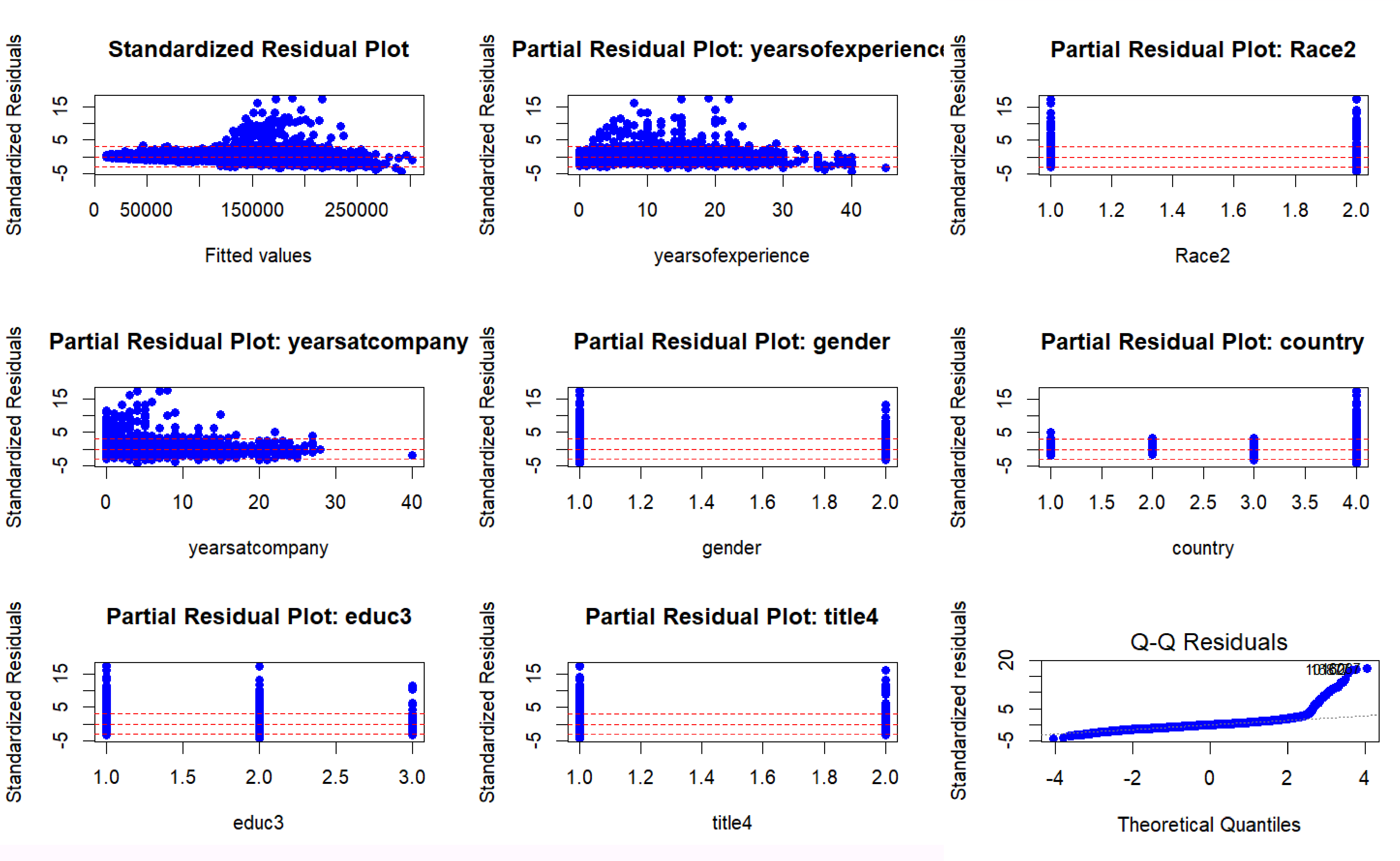
Residuals Analysis

Detecting Lack of Fit & Outliers

OUTLIERS

|stan. res.| > 2

2.5% of dataset



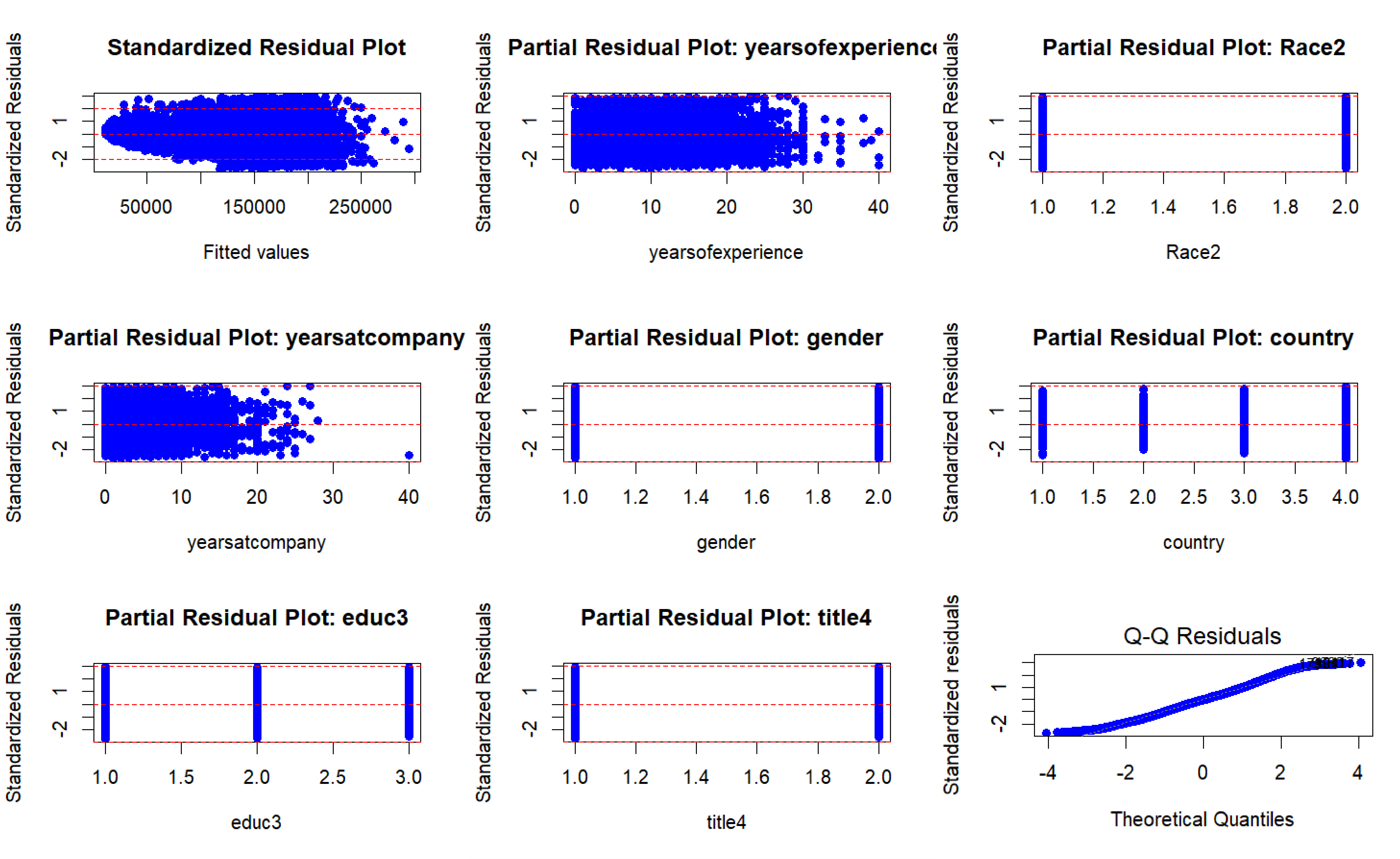
IMPROVEMENT

RSE: 28,720

R^2: 64.25%

Residuals Analysis

Detecting Lack of Fit & Outliers



Stepwise Regression

01 Main Effects

- Residual standard error: 28,710
- Adjusted R-squared: 64.27%
- Global F-test: < 2.2e-16

```
Call:
lm(formula = basesalary ~ yearsofexperience + yearsatcompany +
    gender + Race2 + educ3 + country + title4 + Fortune_500,
    data = STEM3)

Residuals:
   Min     1Q Median     3Q    Max
-79061 -19167   -232  17814  87203

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           75176.35    1110.88  67.673  < 2e-16 ***
yearsofexperience      4212.46      46.77  90.074  < 2e-16 ***
yearsatcompany        -1588.79      77.12 -20.601  < 2e-16 ***
genderFemale          -3142.30     558.45  -5.627 1.86e-08 ***
Race2White            -2704.48     482.13  -5.609 2.06e-08 ***
educ3Master's Degree   6950.29     457.93  15.178  < 2e-16 ***
educ3PhD              30642.04    1045.79  29.300  < 2e-16 ***
countryIndia         -59192.02    1293.49 -45.762  < 2e-16 ***
countryUnited Kingdom  5937.97    1661.99   3.573 0.000354 ***
countryUS             44403.05    1074.98  41.306  < 2e-16 ***
title4Management       3590.58     608.32   5.902 3.64e-09 ***
Fortune_500Yes        -1486.06     439.22  -3.383 0.000718 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28710 on 18441 degrees of freedom
Multiple R-squared:  0.6429,	Adjusted R-squared:  0.6427
F-statistic:  3019 on 11 and 18441 DF,  p-value: < 2.2e-16
```

Stepwise Regression

02 Main Effects + Transformation + Interaction

- Residual standard error: 28,180
- Adjusted R-squared: 65.6%
- Global F-test: < 2.2e-16

```
Call:
lm(formula = basesalary ~ yearsofexperience + sqrt(yearsofexperience) +
    yearsatcompany + gender + Race2 + educ3 + country + title4 +
    Fortune_500 + yearsofexperience:country + educ3:country +
    yearsofexperience:educ3 + yearsofexperience:yearsatcompany +
    yearsofexperience:title4 + yearsofexperience:Fortune_500 +
    yearsatcompany:educ3 + yearsatcompany:title4 + gender:title4 +
    yearsatcompany:Fortune_500, data = STEM3)

Residuals:
   Min     1Q Median     3Q    Max
-83208 -19176   -188  17678  97006

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               67898.03    1821.55  37.275  < 2e-16 ***
yearsofexperience                          1020.82     262.77   3.885 0.000103 ***
sqrt(yearsofexperience)                   14030.71     702.41  19.975  < 2e-16 ***
yearsatcompany                            -3554.52     199.65 -17.804  < 2e-16 ***
genderFemale                              -2660.86     613.27  -4.339 1.44e-05 ***
Race2White                                -2831.87     476.34  -5.945 2.81e-09 ***
educ3Master's Degree                       5742.24    2357.98   2.435 0.014892 *
educ3PhD                                  13558.40    7697.43   1.761 0.078184 .
countryIndia                             -62083.99    2083.78 -29.794  < 2e-16 ***
countryUnited Kingdom                     -4524.64    2908.93  -1.555 0.119861
countryUS                                 37250.47    1726.75  21.573  < 2e-16 ***
title4Management                           4505.53    1207.63   3.731 0.000191 ***
Fortune_500Yes                             -647.79     685.33  -0.945 0.344554
yearsofexperience:countryIndia              397.68     255.20   1.558 0.119176
yearsofexperience:countryUnited Kingdom    1501.10     328.45   4.570 4.90e-06 ***
yearsofexperience:countryUS                1038.82     205.48   5.056 4.33e-07 ***
educ3Master's Degree:countryIndia         -1584.23    2872.90  -0.551 0.581338
educ3PhD:countryIndia                     18382.48   13839.81   1.328 0.184118
educ3Master's Degree:countryUnited Kingdom  2805.59    3486.54   0.805 0.421009
educ3PhD:countryUnited Kingdom             3305.81    9720.46   0.340 0.733794
educ3Master's Degree:countryUS             1963.14    2339.34   0.839 0.401376
educ3PhD:countryUS                        20574.80    7715.21   2.667 0.007665 **
yearsofexperience:educ3Master's Degree      -80.91      90.46  -0.894 0.371109
yearsofexperience:educ3PhD                 -908.74     213.91  -4.248 2.16e-05 ***
yearsofexperience:yearsatcompany             45.70      10.10   4.526 6.06e-06 ***
yearsofexperience:title4Management         -181.32     107.14  -1.692 0.090598 .
yearsofexperience:Fortune_500Yes           -656.99      91.40  -7.188 6.84e-13 ***
yearsatcompany:educ3Master's Degree        -433.93     155.39  -2.792 0.005237 **
yearsatcompany:educ3PhD                    1088.35     389.17   2.797 0.005170 **
yearsatcompany:title4Management             584.61     166.22   3.517 0.000437 ***
genderFemale:title4Management             -2587.60    1366.74  -1.893 0.058339 .
yearsatcompany:Fortune_500Yes              1641.17     170.07   9.650  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28180 on 18421 degrees of freedom
Multiple R-squared:  0.6565,	Adjusted R-squared:  0.656
F-statistic:  1136 on 31 and 18421 DF,  p-value: < 2.2e-16
```

All-Possible-Regression Selection

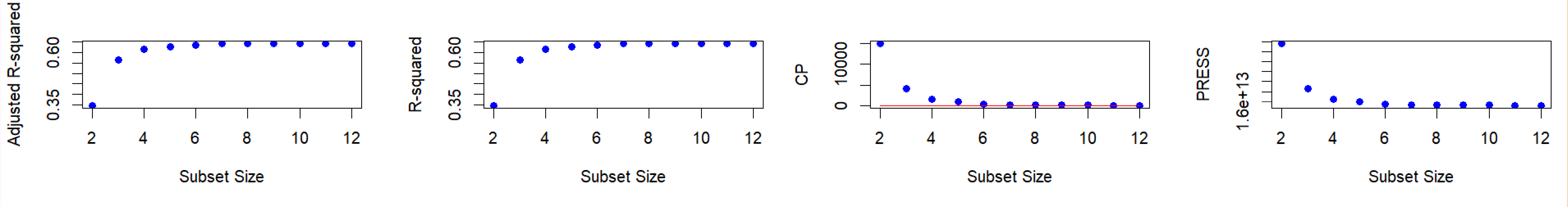
01 “Best” model for each value of p

```
Selection Algorithm: exhaustive
          yearsofexperience yearsatcompany genderFemale Race2White Fortune_500Yes title4Management Masters PhD India UK  US 
1  ( 1 )  " "               " "            " "          " "        " "            " "              " "     " " "*"   " " " "
2  ( 1 )  "*"               " "            " "          " "        " "            " "              " "     " " "*"   " " " "
3  ( 1 )  "*"               " "            " "          " "        " "            " "              " "     " " "*"   " " "*"
4  ( 1 )  "*"               " "            " "          " "        " "            " "              " "     "*" "*"   " " "*"
5  ( 1 )  "*"               "*"            " "          " "        " "            " "              " "     "*" "*"   " " "*"
6  ( 1 )  "*"               "*"            " "          " "        " "            " "              "*"     "*" "*"   " " "*"
7  ( 1 )  "*"               "*"            " "          " "        " "            "*"              "*"     "*" "*"   " " "*"
8  ( 1 )  "*"               "*"            "*"          " "        " "            "*"              "*"     "*" "*"   " " "*"
9  ( 1 )  "*"               "*"            "*"          "*"        " "            "*"              "*"     "*" "*"   " " "*"
10  ( 1 ) "*"               "*"            "*"          "*"        " "            "*"              "*"     "*" "*"   "*" "*"
11  ( 1 ) "*"               "*"            "*"          "*"        "*"            "*"              "*"     "*" "*"   "*" "*"
```

All-Possible-Regression Selection

02 "Best subset" model along with their criteria

	Rsq	AdRsq	CP	BIC	RSS	AIC	PRESS	MSE	MSE1
1	0.3492	0.3492	15163.0673	-7907.048	2.770980e+13	-7922.694	2.771415e+13	1501804865	1501804865
2	0.5631	0.5631	4116.6439	-15251.580	1.860148e+13	-15275.049	1.860675e+13	1008210131	1008210131
3	0.6129	0.6128	1549.7924	-17472.136	1.648371e+13	-17503.428	1.648997e+13	893474700	893474700
4	0.6263	0.6263	855.8252	-18116.059	1.590996e+13	-18155.174	1.591780e+13	862421896	862421896
5	0.6352	0.6351	399.9836	-18549.288	1.553251e+13	-18596.226	1.554257e+13	842007608	842007608
6	0.6409	0.6407	109.8884	-18827.786	1.529171e+13	-18882.547	1.530335e+13	828998750	828998750
7	0.6414	0.6412	84.9217	-18844.810	1.526948e+13	-18907.394	1.528297e+13	827838416	827838416
8	0.6420	0.6418	56.6986	-18865.122	1.524456e+13	-18935.529	1.525959e+13	826532401	826532401
9	0.6425	0.6423	31.7761	-18882.185	1.522237e+13	-18960.414	1.523920e+13	825373785	825373785
10	0.6427	0.6425	21.4472	-18884.687	1.521220e+13	-18970.740	1.523042e+13	824867408	824867408
11	0.6429	0.6427	12.0000	-18886.315	1.520277e+13	-18980.191	1.522274e+13	824400395	824400395



05 Model Fitting

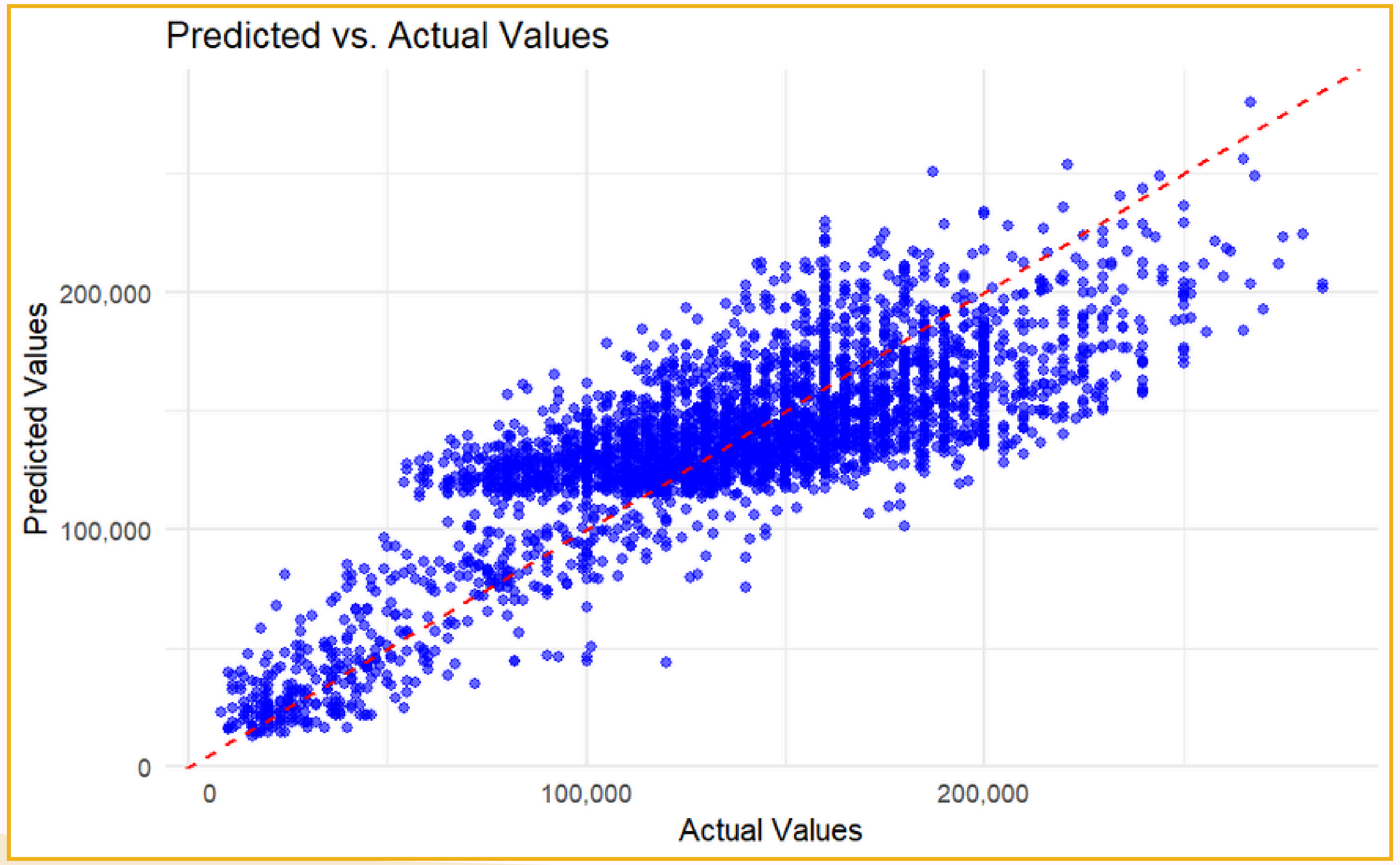
Stratified Random Sampling
(by country)

Method : 80% Training
20% Test

Model Fitting

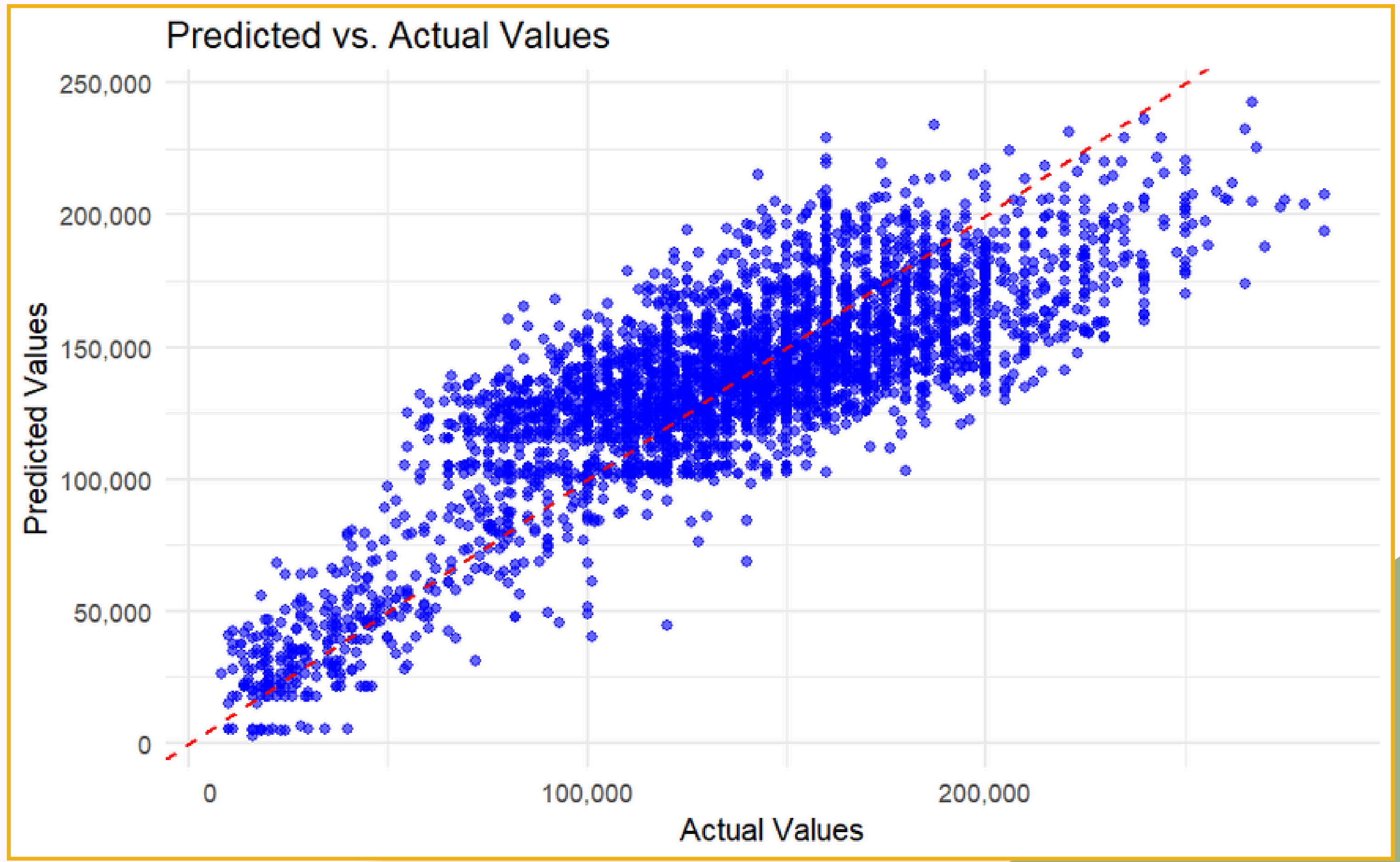
01 Main Effects

- **MSE:** 804454948
- **RMSE:** 28,363
- **MAE:** 22,388
- **Adj. R^2:** 64.07%



02 Main Effects + Transformation + Interaction

- **MSE:** 772722685
- **RMSE:** 27,798
- **MAE:** 22,107
- **Adj. R^2:** 65.41%



Model Fitting

03 Predicting the salary of a male, non-White, college graduate in Canada who landed a non-managerial role at a start-up company (non-Fortune 500)

Prediction	Lower	Fit	Upper
Confidence interval	$72,830	$75,274	$77,718
Prediction interval	$18,768	$75,274	$131,779

Best Model

to predict the salaries of STEM employees

$$\begin{aligned}\hat{\text{BaseSalary}} = {} & 75,274 + 4,198 \cdot \text{YearsOfExperience} - 1,552 \cdot \text{YearsAtCompany} - 3,394 \cdot \text{GenderFemale} \\ & - 2,561 \cdot \text{RaceWhite} + 6,679 \cdot \text{EducMaster'sDegree} + 30,651 \cdot \text{EducPhD} \\ & - 58,958 \cdot \text{CountryIndia} + 5,548 \cdot \text{CountryUK} + 44,372 \cdot \text{CountryUS} \\ & + 3,711 \cdot \text{TitleManagement} - 1,492 \cdot \text{Fortune500Yes} + \epsilon\end{aligned}$$

06 Limitations & Future Improvements

- **Other contributing factors unaccounted for:** The current model excludes significant predictors of base salary, such as:
  - **Tech Stack:** Top-paying programming languages and frameworks
  - **Industry:** Sector-specific variations (e.g., FinTech, AI/ML)
  - **Company Profitability:** Impact of employer financial health
  - **Prestige of Previous Employers:** Influence of working for high-profile companies (e.g., FAANG)

- **Inestimable parameters:** Some parameter combinations cannot be estimated due to missing data or sparse representation, limiting the model's ability to generalize for these cases.

06 Limitations & Future Improvements

- **Insufficient data for certain countries:** The dataset is heavily skewed, with the US accounting for 90% of responses. This imbalance reduces the model's ability to accurately represent salary trends in less-represented countries like the UK, Canada, and India.

- **Income disparity and outliers:** Income disparity within the US is pronounced, with several extreme outliers significantly inflating salaries compared to other countries. This likely affects the model’ s performance and skews predictions.

Thank You

May the force be with you!