Regression Analysis of Glassdoor Data Science Salaries

Matthew Chen, Jasper Tsai, Mark Faynboym

Abstract

We fit multiple regression models to study the relationship between average salary and given predictor variables scraped from Glassdoor. Some of the predictors used include different skills, job location, education level, and employee ratings of the company. Using stepwise regression we find subset of a pairwise interaction model that reasonably balances bias and variance. Further, our final model has only a multiple R^2 of 0.41, which indicates that it is not suited for prediction. To supplement this, we train a logistic regression model to predict if salaries are greater than $100k with reasonably high accuracy. Overall, we find that being in California, having a senior job position, having predictive modeling skills, and a higher education tend to correspond to higher average salary.

1 Introduction

According to the U.S. Bureau of Labor Statistics, the data science industry is projected to grow 36% from 2021 to 2031, which is much higher than the national average [1]. Many organizations are increasingly benefiting from data science and statistical knowledge as the availability of data grows exponentially. As such, we would like to explore salaries relating to this in-demand industry to better understand what factors may influence salary the most. This may give insight in understanding the current state of the industry.

Some of the research questions that we hope to answer are:

- What factors affect salaries of data related industries the most?
- What skills or education level do the highest paid data scientists have?
- Are there significant differences in the location or size of the company?
- How much do data science salaries vary naturally?

To answer this question, we use data sourced from Kaggle.com [2] (last updated in 2021) which originally scraped data-related job postings from Glassdoor.com. This data includes information, in no particular order, about the average salary, the company size, employee ratings (from a scale of 0-5), age of the company, the seniority of the role, the degree requirements, and the location of the role. We perform a thorough regression analysis on this data to answer our research questions.

2 Methods and Results

2.1 Data processing

In the original raw data, there were columns that were either difficult to interpret, difficult to process, or contained high amounts of missing data, and therefore were dropped from the overall data. There were also duplicate observations that were deleted from the data. Further, there were data entries that were nonsensical, in particular, containing negative ages, negative ratings, and “unknown” locations. These were also deleted from the data to make it more amenable for analysis. The final processed dataset has 433 observations.

While job location (state) is important, there are 50 potential categories which may be too numerous to use as dummy variables in multiple regression. However, we see that the only state with a significant difference in salary compared to the others is California (Figure 1). Thus, we create a binary dummy variable based on whether or not a job is in California. Similarly, we combine machine learning skills (keras, pytorch, scikit, tensorflow), data visualization skills (Tableau, Power BI), whether or not the position is senior standing or not into single dummy variables to reduce the total number of required classifiers. Table 1 contains a full description of the final variables selected.

2.2 Exploratory Data Analysis (EDA)

We conduct an exploratory data analysis to help us better understand the data collected and inform future decisions when we fit our models.

For the quantitative variables, we see in the scatterplot matrix (Figure 2) that rating and age are not very correlated with average salary. However, relationships between rating and age are also weak so there is little worry about multicollinearity. Further, we see that the distribution of average salary is slightly right skewed (this may suggest that a transformation is needed later), the distribution of rating is approximately symmetric, and that age is heavily right skewed. For average salary, we see that the distribution of the square-root transform is much more symmetric (Figure 3). Also, we see that the median average salary is approximately $100,000 (Figure 4).

For the qualitative variables, we find that PhD holders earn the highest median salary, followed by MS holders (Figure 5) and that senior positions have higher median pay (Figure 6). Further, the size of the company does not noticeably affect average salaries (Figure 7). In Figure 8, we plot side-by-side boxplots of the different skills and find that Python, machine learning, Spark, AWS, and Hadoop skills have noticeably higher median salaries.

2.3 First Order Multiple Regression

We initially fit a first order model based on all of the available predictor variables, and denote this as Model1. Here we have 22 regression coefficients, including the intercept. The R summary table including the estimates for each coefficient, their respective standard errors, the corresponding t-statistic and p-value, and the multiple R^2 and adjusted R^2_a are included in Table 2. We find that Model1 has a multiple R^2 of 0.39 and an adjusted R^2_a of 0.36. In Figure 9, we include plots of model diagnostics for Model 1, and find that the residuals have approximately equal spread (no sign of heteroskedasticity) and no systematic pattern. However, in the normal QQ plot, we see that the residuals have a heavy right tail, though this may be caused by outliers in the data which we will analyze later.

Since in our EDA, we find that average salaries are right-skewed. Thus, we check the Box-Cox Procedure to search for potential transformations that may be needed, and find that a square-root transformation in Y maximizes the log-likelihood (Figure 10). With this in mind, we fit another first order model with all terms using the square-root of the average salary, and denote this as

$\text{Model1}_{\text{sqrt}}$. The fitted regression coefficients and their standard errors are summarized in Table 3. Here, we obtain a multiple R^2 of 0.39 and an adjusted R^2_a of 0.36, which is not much different compared to Model 1. However, in the model diagnostics (Figure 11), we see that the residuals are less right skewed. Thus, we decided to move forward with the square root transformation.

2.4 Second Order Multiple Regression with Pairwise Interactions

In order to attempt to obtain an improved fit on the data, we explore fitting a second order model with all pairwise interactions. Fitting this model on the non-transformed salaries, we find that the Box-Cox Procedure still recommends a square-root transform. Thus, we continue to move forward with the square-root transformation of salaries and denote this model as $\text{Model2}_{\text{sqrt}}$. This model had 211 regression coefficients including the intercept, a multiple R^2 of 0.66, and an adjusted R^2_a of 0.33. Since the adjusted R^2_a is much lower than the multiple R^2, this model is likely to be overfitting. However, the model diagnostics appear to be reasonable (Figure 12).

To reduce overfitting and increase model interpretability, we perform forward stepwise regression based on the AIC criterion to select a subset of $\text{Model2}_{\text{sqrt}}$ that balances the bias-variance tradeoff. The AIC procedure and the final selected model is summarized in Table 4; we denote this model as $\text{Model2}_{\text{AIC}}$. The fitted regression coefficients and the R summary of $\text{Model2}_{\text{AIC}}$ is shown in Table 5. Here, we obtain multiple R^2 of 0.41 and adjusted R^2_a of 0.38, which is a slight improvement over Model1. We also find that the residual plot shows no systematic pattern and approximately equal spread, and the residuals fit the Normal Q-Q line reasonably well (Figure 13). Thus, the model is reasonable.

2.5 Analysis of Outliers

We find one outlier in average salary after conducting the Bonferroni Outlier Test on Model, which uses the t-statistic on studentized deleted residuals to identify outliers in Y. Similarly, we find 22 outliers in X by identifying leverage values that are greater than 2p/n, which is a standard criterion. After identifying influential cases (where Cook's Distance > 4/(n-p)) on the identified outliers, we remove these from the data. Then, we refit $\text{Model2}_{\text{AIC}}$ and denote it as Model3, our final model. The model summary is shown in Table 6 and the model diagnostics are reasonable (Figure 14).

2.6 Internal Validation of Final Model and ANOVA

We validate the model internally using the Press_p criterion (Eq. 1), which is synonymous with Leave-One-Out-Cross-Validation (LOOCV). Here, $\hat{Y}_{i(i)}$ is the predicted value for the i^{th} case after fitting a model excluding case i, and Y_i is the i^{th} observed average salary.

$$Press_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)}\right)^2 \qquad (Eq. 1)$$

We obtain a Press_p score of 861 which is not far off from the residual sum of squares of Model 3 (822). Thus, we can conclude that our model is not severely overfitting the data.

An ANOVA table of the final model is shown in Table 7. We find that the regression sum of square is 545.7 and the residual sum of squares is 779.7 with a total sum of squares of 1325.4, which corresponds to a multiple R^2 of 0.41. Additionally, note that the reduction in residual sum of squares (given the variables already in the model) are much lower for the interaction terms than the first order terms. We also note that the interaction terms have been chosen by the stepwise regression algorithm to reduce AIC, and may be much less interpretable than the first order terms.

2.7 Logistic Regression Model

In sections 2.3-2.6, we see from our multiple regression models that the multiple R^2 is moderately low which prevents us from making good predictions. Thus, to fill in the gap of predictability, we train a logistic model to classify whether a salary is above \$100k (approximately the median salary) or not to reduce the prediction difficulty.

Logistic regression utilizes a log-linear model that models prediction probabilities given a binary Y. To do this, it fits a sigmoidal function on the observed y, in this case an indicator of whether or not a salary is greater than \$100k (Eq. 2). Note that it is possible to transform P(y) in Eq.2 to be linear in coefficients β, which indicates that it is a generalized linear model.

$$P(y) = \frac{1}{1 + \mathrm{e}^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\beta_{p-1}X_{p-1})}} \qquad (Eq. 2)$$

We fit the model based on our available X predictors (except for size, which was not found to be important in our EDA or the stepwise regression analyses). Then, the model is validated using k-fold cross-validation (k=10), and find that the mean training accuracy (0.73) and mean testing accuracy (0.70) is reasonably close which indicates that there is no severe overfitting. Note that our data is approximately balanced since we created our binary y variable using the median as the threshold, so using overall accuracy as a criterion is reasonable.

Training the model on all available data, we obtain an overall accuracy of 74%, which is very decent given the low predictability of our linear models. In Figure 15 we see through the trained coefficients that Python and machine learning skills, being in California, and having a senior position is highly rewarded in terms of prediction probability.

3 Discussion and Conclusion

From our final selected multiple regression model (Model 3), we see that of the first order terms, having a senior job position, being in California, having Python, SAS, AWS, machine learning, and Hadoop skills, and having a PhD tends to increase average salary. There is also a positive relationship with the age and rating of the company. Further, skills in SQL, data visualization, and having a degree lower than an MS tends to correspond with lower salaries. The most significant predictors are being senior status, being in California, and having Python and machine learning skills.

In terms of skills, we see that abilities relating to predictive modeling (big data analytics, machine learning, statistical modeling, cloud computing, etc…) positively impact salary outcomes. However, skills relating more to data analytics (visualization, database queries, etc…) tend to correspond to lower salaries. This difference is reasonable, since “data analyst” careers make a lower mean salary ($81,946) [3] compared to “data scientists” ($139,202) [4] in the United States. We also see that education is important, with PhD level jobs tending to have higher salaries than MS or lower. These results are consistent with what we found in our exploratory data analysis.

In the regression analysis, we obtained a multiple R^2 of 0.41. This means that given our chosen set of predictors, the model only explains 41% of the variation in average salary. There are several potential reasons for this, including potentially noisy data (high error variance) or the lack of important unknown X-variables. Overall, this indicates the multiple regression model is not suited for making predictions about salary. To fill in this gap, we were able to train a logistic regression model to classify salaries on whether or not they are greater than $100k with approximately 74% accuracy. We also find that the variables that tend to predict salaries above $100k are largely consistent with what we found in our multiple regression analysis.

It is important to note that there are caveats in doing analysis on this dataset. Given that the data is scrapped from Glassdoor, we do not have access to all variables that may have been significant in our model. Additionally, skills are difficult to classify and may be inconsistent between job postings, which may hide the true relationship of these skills with higher salary in our model. Just from anecdotal evidence, we are also aware that reported Glassdoor salaries may differ from real-life pay and may not account for other forms of compensation. All of these factors may negatively impact the quality of our model. However, it is still interesting to see how these salary estimates varied with our given predictors, and how the important factors that we found to be significant are reasonable with common sense.

In the future, further study can be done including more predictors, such as a greater variety of skills. Additionally, work can be done to analyze how salaries change with time, as we were not given a definite time interval of when the data was sourced (though we did know it was somewhat recent).

4 Appendix A: Figures and Tables

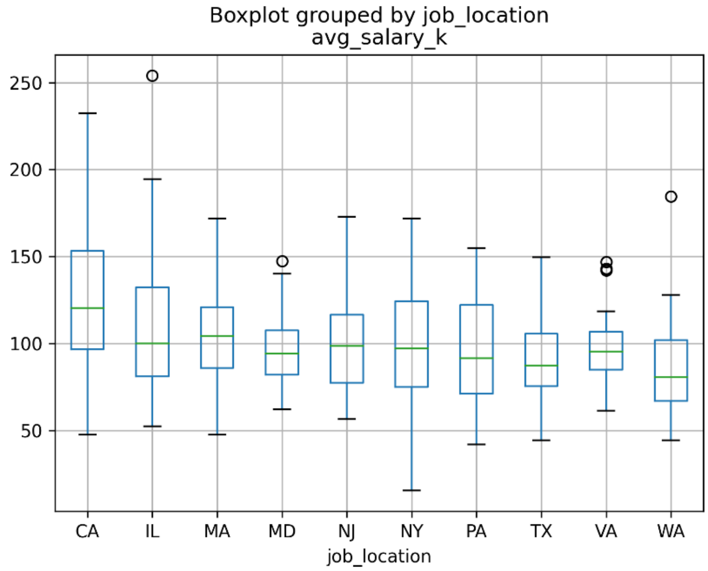

Figure 1. Salary of the top 10 states based on occurrence frequency in the data

Table 1. Description of selected variables after data cleaning

Variable	**Description**	**Data type**
avg_salary_k	average reported salary in thousands of dollars	quantitative
rating	Glassdoor employee rating between 0-5	quantitative
size	size of the company in 7 categories	categorical
age	age of the company in years	quantitative
python visual_software ML_software Spark AWS Excel SQL SAS Hadoop	all skills are binary variables: 1 if the respective skill is possessed, 0 otherwise	categorical
degree	whether the job requires a PhD, MS, or other	categorical
in_CA	whether or not the job is in CA, binary variable	categorical
senior_status	whether or not the job is a senior position or not, binary variable	categorical

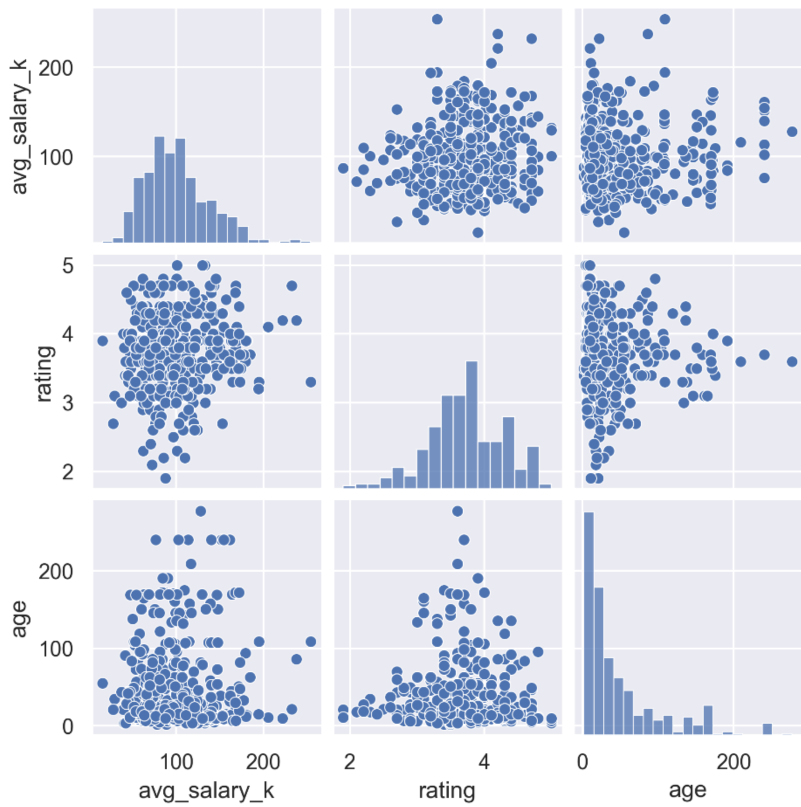

Figure 2. Scatterplot matrix of quantitative variables

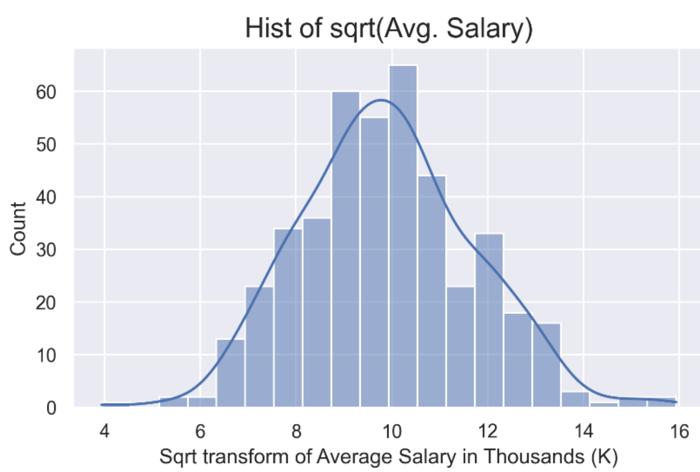

Figure 3. Histogram of square-root transformed average salary.

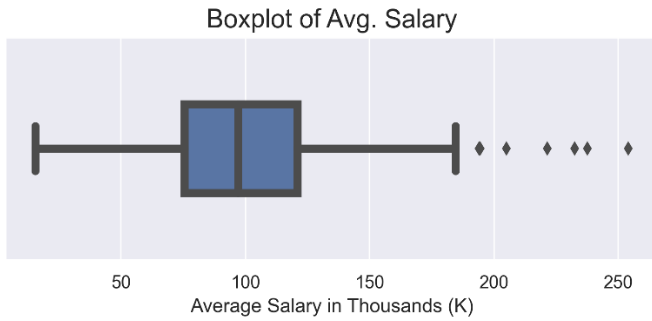

Figure 4. Boxplot of average salary

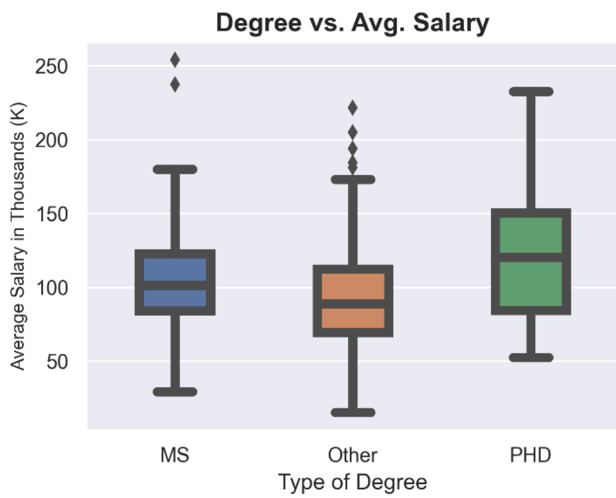

Figure 5. Side-by-side boxplots of average salary vs. education

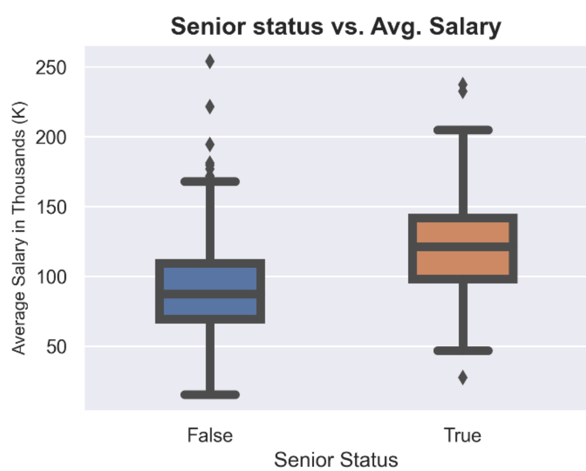

Figure 6. Side-by-side boxplots of average salary and senior status

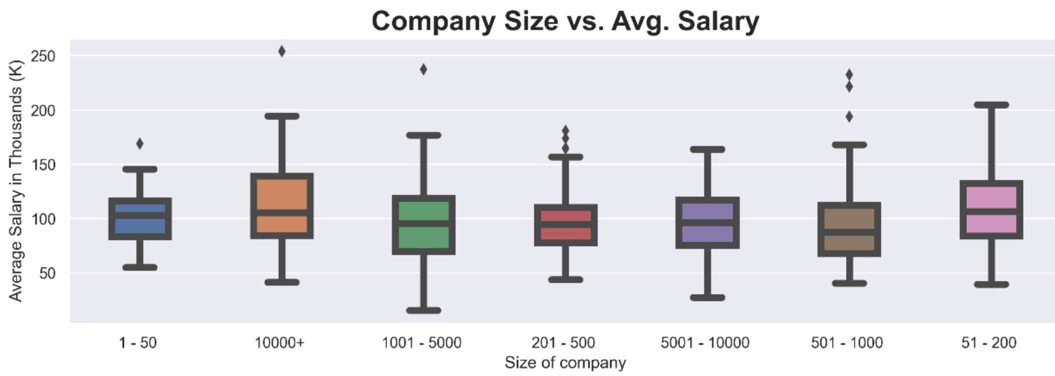

Figure 7. Size of the company vs. average salary

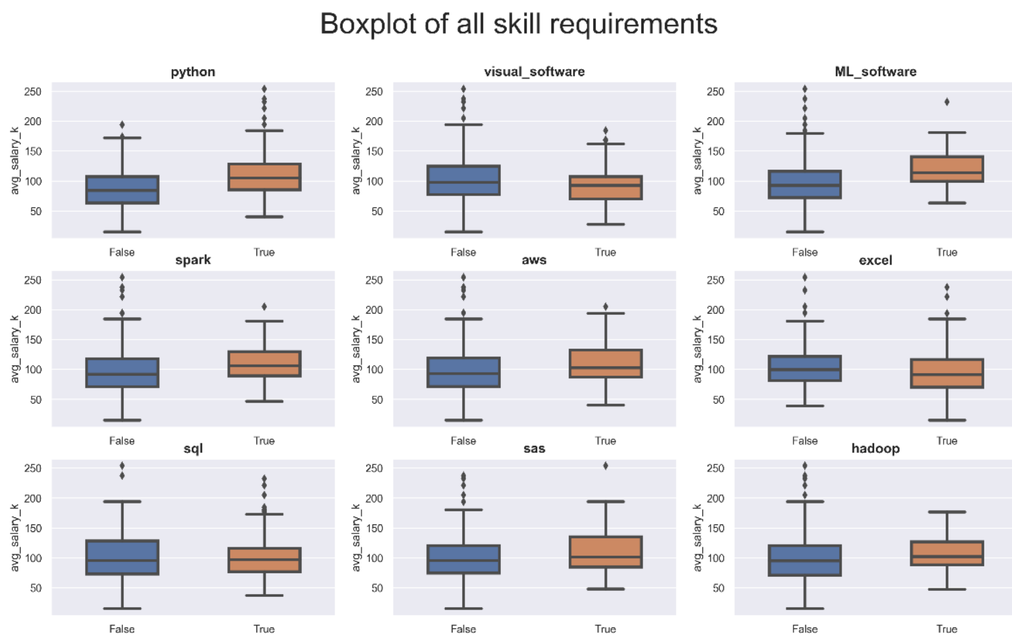

Figure 8. Side-by-side boxplots of different skills vs. average salary

Table 2. Summary of the fitted regression line for average salary vs. all first order terms (Model 1)

```
Call:
lm(formula = avg_salary_k ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-68.237 -19.300  -2.686  15.645 129.433 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       50.88530   14.00955   3.632 0.000317 ***
rating             6.06497    2.71308   2.235 0.025924 *  
size10000+        10.23128    9.39165   1.089 0.276615    
size1001 - 5000    2.08715    8.96246   0.233 0.815973    
size201 - 500      1.76830    9.05893   0.195 0.845333    
size5001 - 10000  -5.66351    9.81860  -0.577 0.564381    
size501 - 1000     1.08277    9.02291   0.120 0.904540    
size51 - 200       3.65827    9.15903   0.399 0.689794    
age                0.07354    0.03402   2.162 0.031189 *  
python1           18.72444    3.26697   5.731 1.93e-08 ***
visual_software1  -8.44936    3.85267  -2.193 0.028859 *  
ML_software1       6.97582    4.42777   1.575 0.115918    
spark1             0.62159    4.27003   0.146 0.884332    
aws1               5.06345    3.59485   1.409 0.159731    
excel1            -2.69535    2.94687  -0.915 0.360913    
sql1              -5.76717    3.49817  -1.649 0.099989 .  
sas1              16.69562    5.13433   3.252 0.001241 ** 
hadoop1            6.00225    4.59322   1.307 0.192023    
degreeOther       -2.81089    3.33942  -0.842 0.400428    
degreePHD         10.11828    4.86816   2.078 0.038287 *  
in_CATrue         28.07331    3.68637   7.615 1.82e-13 ***
senior_statusTrue 24.42671    3.27889   7.450 5.56e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.35 on 411 degrees of freedom
Multiple R-squared:  0.3865, Adjusted R-squared:  0.3552 
F-statistic: 12.33 on 21 and 411 DF,  p-value: < 2.2e-16
```

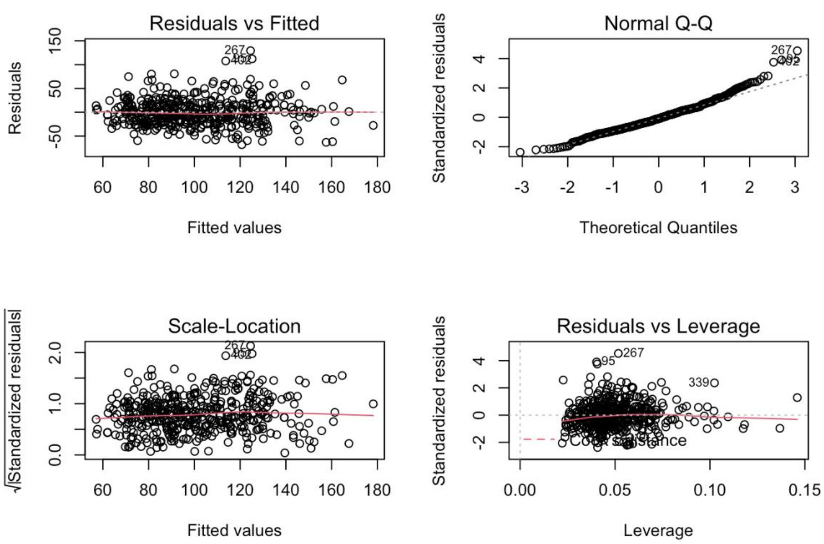

Figure 9. Model diagnostics for Model 1: all first order terms

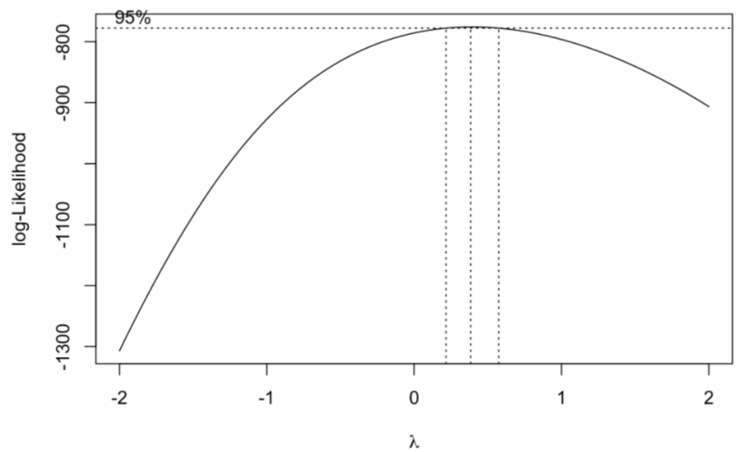

Figure 10. Log-likelihood of Box-Cox power transformations for Model 1

Table 3. R summary output for Model1$_{sqrt}$

```
Call:
lm(formula = avg_salary_k ~ ., data = df_sqrtY)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5815 -1.0102 -0.0118  0.8430  4.8772 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        7.618047   0.688072  11.072  < 2e-16 ***
rating             0.261840   0.133252   1.965   0.0501 .  
size10000+         0.428366   0.461266   0.929   0.3536    
size1001 - 5000    0.018265   0.440187   0.041   0.9669    
size201 - 500      0.062090   0.444925   0.140   0.8891    
size5001 - 10000  -0.342500   0.482236  -0.710   0.4780    
size501 - 1000    -0.026790   0.443156  -0.060   0.9518    
size51 - 200       0.159912   0.449841   0.355   0.7224    
age                0.003654   0.001671   2.187   0.0293 *  
python1            0.950128   0.160456   5.921 6.74e-09 ***
visual_software1  -0.398770   0.189222  -2.107   0.0357 *  
ML_software1       0.382377   0.217468   1.758   0.0794 .  
spark1             0.054051   0.209720   0.258   0.7967    
aws1               0.264160   0.176559   1.496   0.1354    
excel1            -0.168254   0.144734  -1.163   0.2457    
sql1              -0.263934   0.171811  -1.536   0.1253    
sas1               0.801324   0.252170   3.178   0.0016 ** 
hadoop1            0.351854   0.225594   1.560   0.1196    
degreeOther       -0.171737   0.164014  -1.047   0.2957    
degreePHD          0.465469   0.239097   1.947   0.0522 .  
in_CATrue          1.317280   0.181054   7.276 1.76e-12 ***
senior_statusTrue  1.195941   0.161041   7.426 6.50e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.441 on 411 degrees of freedom
Multiple R-squared:  0.3887, Adjusted R-squared:  0.3574 
F-statistic: 12.44 on 21 and 411 DF,  p-value: < 2.2e-16
```

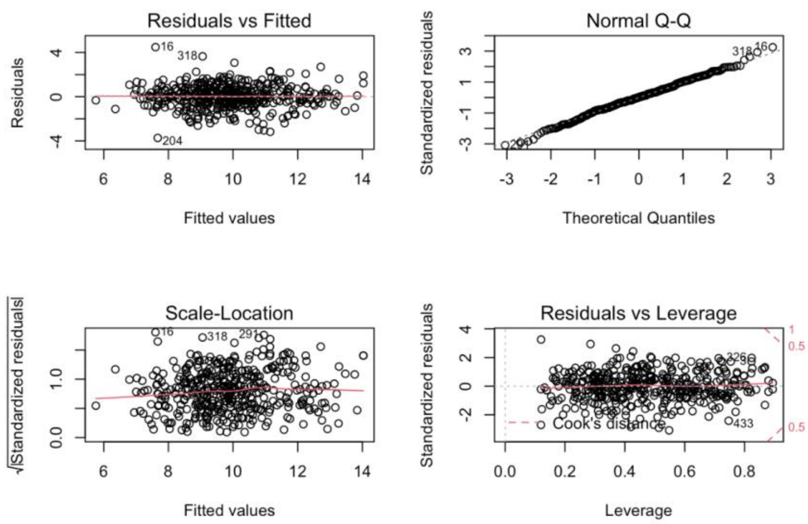

Figure 11. Model diagnostics for Model1$_{sqrt}$

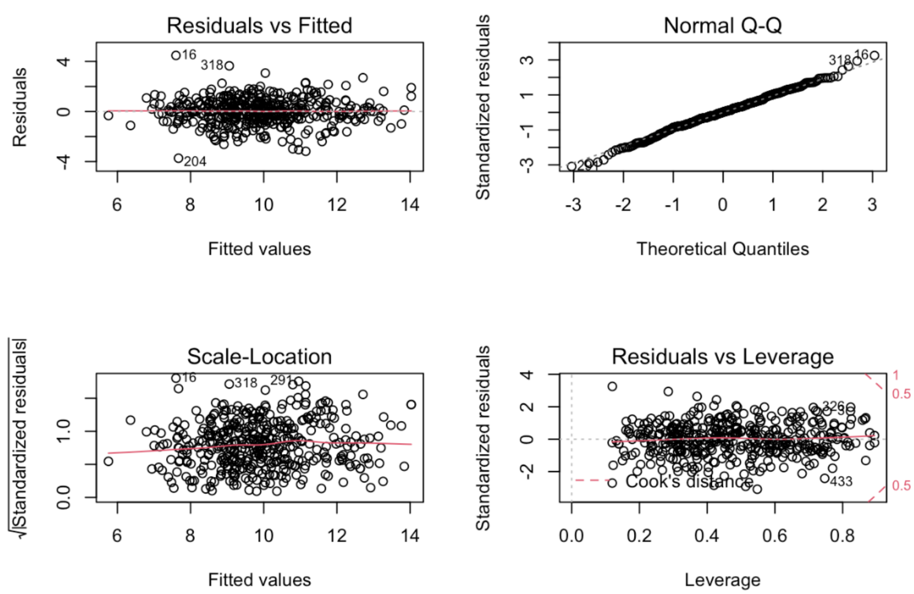

Figure 12. Model diagnostics for Model2$_{sqrt}$

Table 4. Forward stepwise AIC procedure on Model2$_{sqrt}$. The final model we obtain we denote as Model2$_{AIC}$

```
Initial Model:
avg_salary_k ~ 1

Final Model:
avg_salary_k ~ senior_status + in_CA + python + degree + visual_software +
    age + rating + sas + aws + ML_software + hadoop + sql + python:aws +
    in_CA:ML_software + age:ML_software + age:rating + python:rating +
    visual_software:hadoop + python:hadoop + sas:sql


                      Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                              432  1396.8593 509.1486
2          + senior_status  1 162.064783       431  1234.7945 457.7503
3                  + in_CA  1 135.443012       430  1099.3515 409.4426
4                 + python  1 100.779407       429   998.5721 369.8099
5                 + degree  2  38.304043       427   960.2680 356.8736
6        + visual_software  1  21.200060       426   939.0680 349.2070
7                    + age  1  12.259848       425   926.8081 345.5168
8                 + rating  1  13.826643       424   912.9815 341.0084
9                    + sas  1  12.537375       423   900.4441 337.0211
10                   + aws  1   8.764848       422   891.6793 334.7857
11            + python:aws  1   7.097963       421   884.5813 333.3251
12           + ML_software  1   6.270283       420   878.3110 332.2449
13     + ML_software:in_CA  1   9.998189       419   868.3128 329.2876
14       + age:ML_software  1   5.446500       418   862.8663 328.5631
15                 + excel  1   5.042879       417   857.8235 328.0250
16            + rating:age  1   4.824547       416   852.9989 327.5829
17         + rating:python  1   4.698548       415   848.3004 327.1912
18                + hadoop  1   4.288238       414   844.0121 326.9968
19 + visual_software:hadoop  1   5.712531       413   838.2996 326.0562
20         + python:hadoop  1   5.514417       412   832.7852 325.1985
21                   + sql  1   4.068588       411   828.7166 325.0778
22               + sql:sas  1   6.865174       410   821.8514 323.4759
23                 - excel  1   3.346980       411   825.1984 323.2357
```

Table 5. R summary of Model2$_{AIC}$

```
Call:
lm(formula = avg_salary_k ~ senior_status + in_CA + python +
    degree + visual_software + age + rating + sas + aws + ML_software +
    hadoop + sql + python:aws + in_CA:ML_software + age:ML_software +
    age:rating + python:rating + visual_software:hadoop + python:hadoop +
    sas:sql, data = df_sqrtY)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5140 -0.8801  0.0026  0.9061  4.4492

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.876354   0.865080   6.793 3.87e-11 ***
senior_statusTrue        1.209110   0.158084   7.649 1.46e-13 ***
in_CATrue                1.585530   0.192990   8.216 2.78e-15 ***
python1                  2.573401   0.958289   2.685  0.00754 **
degreeOther             -0.223158   0.159973  -1.395  0.16378
degreePHD                0.574757   0.238056   2.414  0.01620 *
visual_software1        -0.607935   0.203312  -2.990  0.00296 **
age                      0.029976   0.014968   2.003  0.04587 *
rating                   0.686972   0.229374   2.995  0.00291 **
sas1                     1.736194   0.601294   2.887  0.00409 **
aws1                     0.794445   0.294290   2.700  0.00723 **
ML_software1             1.002828   0.321239   3.122  0.00192 **
hadoop1                  0.727856   0.409442   1.778  0.07620 .
sql1                    -0.184384   0.173065  -1.065  0.28732
python1:aws1            -0.706984   0.362891  -1.948  0.05207 .
in_CATrue:ML_software1  -1.183797   0.464795  -2.547  0.01123 *
age:ML_software1        -0.007241   0.004431  -1.634  0.10303
age:rating              -0.007133   0.004068  -1.753  0.08028 .
python1:rating          -0.377096   0.254662  -1.481  0.13943
visual_software1:hadoop1  0.721045   0.407674   1.769  0.07769 .
python1:hadoop1         -0.798905   0.457253  -1.747  0.08135 .
sas1:sql1               -1.238629   0.653602  -1.895  0.05878 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.417 on 411 degrees of freedom
Multiple R-squared:  0.4092, Adjusted R-squared:  0.3791
F-statistic: 13.56 on 21 and 411 DF,  p-value: < 2.2e-16
```

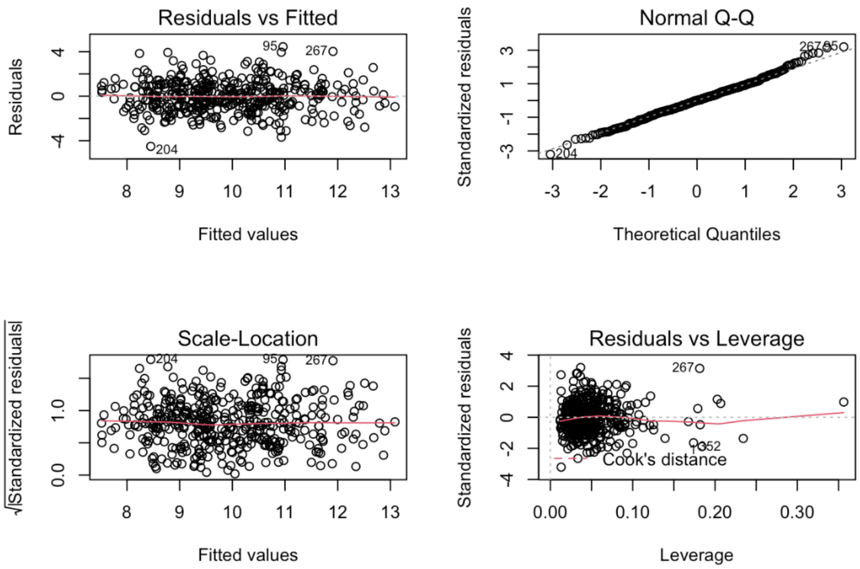

Figure 13. Model diagnostics for Model2$_{AIC}$

Table 6. Model summary of Model3 (final model)

```
Call:
lm(formula = avg_salary_k ~ senior_status + in_CA + python +
    degree + visual_software + age + rating + sas + aws + ML_software +
    hadoop + sql + python:aws + in_CA:ML_software + age:ML_software +
    age:rating + python:rating + visual_software:hadoop + python:hadoop +
    sas:sql, data = df_out)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5567 -0.8897  0.0012  0.8643  4.3200

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.244234   0.861244   7.250 2.15e-12 ***
senior_statusTrue         1.249977   0.157140   7.955 1.85e-14 ***
in_CATrue                 1.594628   0.191053   8.347 1.14e-15 ***
python1                   2.495199   0.949534   2.628 0.008923 **
degreeOther              -0.233769   0.158646  -1.474 0.141393
degreePHD                 0.466519   0.240680   1.938 0.053282 .
visual_software1         -0.551567   0.201633  -2.735 0.006504 **
age                       0.020052   0.015187   1.320 0.187476
rating                    0.609791   0.228560   2.668 0.007940 **
sas1                      2.384939   1.407004   1.695 0.090841 .
aws1                      0.765368   0.295652   2.589 0.009982 **
ML_software1              1.171472   0.339264   3.453 0.000613 ***
hadoop1                   0.755939   0.430062   1.758 0.079552 .
sql1                     -0.259689   0.171894  -1.511 0.131638
python1:aws1             -0.705968   0.362511  -1.947 0.052178 .
in_CATrue:ML_software1   -1.223246   0.458751  -2.666 0.007975 **
age:ML_software1         -0.012591   0.005913  -2.129 0.033835 *
age:rating               -0.004652   0.004127  -1.127 0.260283
python1:rating           -0.358031   0.252850  -1.416 0.157555
visual_software1:hadoop1  0.714515   0.403397   1.771 0.077277 .
python1:hadoop1          -0.832109   0.472214  -1.762 0.078805 .
sas1:sql1                -1.853342   1.429958  -1.296 0.195691
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.393 on 402 degrees of freedom
Multiple R-squared:  0.4117, Adjusted R-squared:  0.381
F-statistic:  13.4 on 21 and 402 DF,  p-value: < 2.2e-16
```

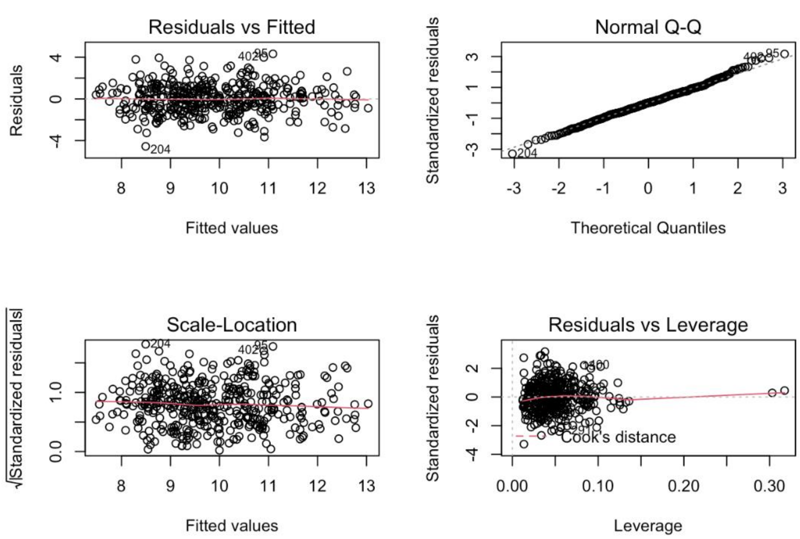

Figure 14. Model diagnostics for Model3 (final model)

Table 7. ANOVA of Model 3 (final model)

Analysis of Variance Table

Response: avg_salary_k

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
senior_status	1	167.91	167.909	86.5714	< 2.2e-16	***
in_CA	1	147.48	147.479	76.0379	< 2.2e-16	***
python	1	90.74	90.738	46.7832	2.964e-11	***
degree	2	27.63	13.815	7.1227	0.0009125	***
visual_software	1	17.62	17.622	9.0859	0.0027391	**
age	1	5.01	5.010	2.5832	0.1087866	
rating	1	15.46	15.458	7.9702	0.0049918	**
sas	1	6.96	6.962	3.5897	0.0588564	.
aws	1	7.06	7.059	3.6394	0.0571406	.
ML_software	1	7.22	7.221	3.7233	0.0543621	.
hadoop	1	3.80	3.802	1.9603	0.1622563	
sql	1	4.19	4.186	2.1583	0.1425822	
python:aws	1	7.25	7.253	3.7394	0.0538465	.
in_CA:ML_software	1	8.94	8.938	4.6084	0.0324133	*
age:ML_software	1	6.83	6.827	3.5201	0.0613530	.
age:rating	1	1.50	1.502	0.7746	0.3793257	
python:rating	1	5.60	5.600	2.8870	0.0900693	.
visual_software:hadoop	1	5.20	5.199	2.6805	0.1023658	
python:hadoop	1	6.06	6.056	3.1225	0.0779792	.
sas:sql	1	3.26	3.258	1.6798	0.1956909	
Residuals	402	779.70	1.940			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

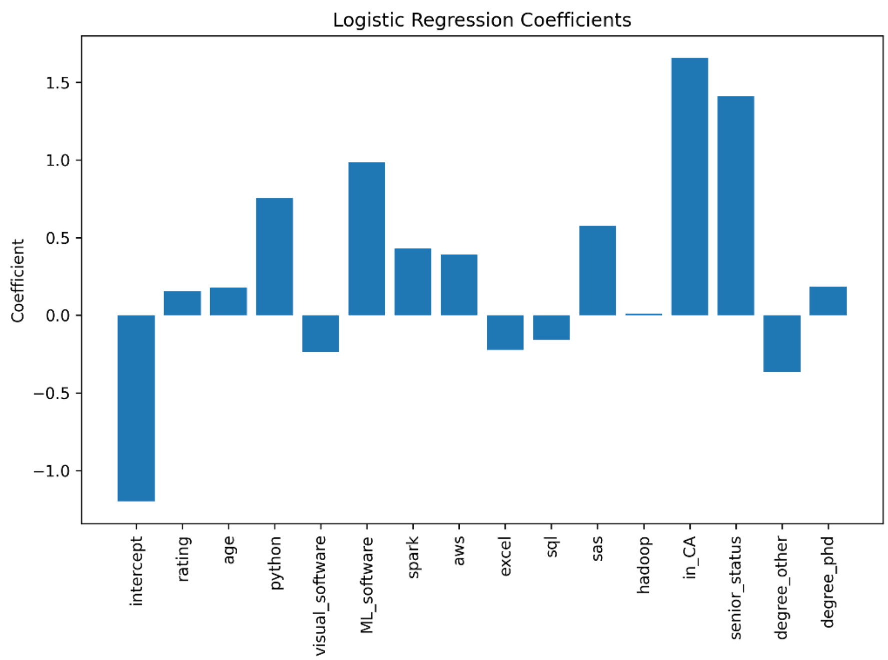

Figure 15. Coefficients from the Logistic Regression model

5 Appendix B: Code and Notebooks

Github repository:
https://github.com/jttsai99/data_science_salary

Notebook for data cleaning:
https://github.com/jttsai99/data_science_salary/blob/main/data_cleaning.ipynb

Notebook for exploratory data analysis:
https://github.com/jttsai99/data_science_salary/blob/main/EDA.ipynb

Rmd file for multiple regression:
https://github.com/jttsai99/data_science_salary/blob/main/Linear_regression.Rmd

Notebook for logistic regression:
https://github.com/jttsai99/data_science_salary/blob/main/logistic_regression.ipynb

6 References

1. https://www.bls.gov/ooh/math/data-scientists.htm
2. https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor
3. https://www.salary.com/research/salary/listing/data-analyst-salary
4. https://www.salary.com/research/salary/listing/data-scientist-salary