A Study of HR Analytics

Eric Lan, Xiaoyu Zhou, Shelby Ames

	Contribution	Email
Eric Lan	Model full&reduced&complex code, correlation matrix, K-fold-validation, and conclusions	ericlan@uw.edu
Xiaoyu Zhou	Coordination, models full&reduced code, correlation matrix, K-fold-validation, and conclusions	xz081302@uw.edu
Shelby Ames	Preliminary Presentation, Report Writing, further reduced model code, figure 6 code, Bonferroni Correction code	sah2003@uw.edu

Emanuela Furfaro

University of Washington

Stat 391- Quantitative Introductory Statistics for Data Science

March 4, 2024

Introduction

Human resources (HR) analytics is a new emerging “people analytics” term. The purpose of HR analytics is to establish business impacts that allow decisions to be made based on data. Companies can make these decisions by using visual and statistical analysis of HR processes of human capital, organizational performance, and external economic benchmarks (Boudreau & Marler, 2017). It is important to note that there are some criticisms about HR analytics that without the implementation of new methods, it is unlikely that these analytics will deliver transformation changes to the field. The current practice of HR analytics may have little influence on companies’ fates while actively damaging the trust and interest of employees (Angrave et al., 2016).

The HR analytics data set produced by IBM Corporation was used in this analysis (Tahir, 2023). The object of this project is to use the HR Analytics data set to find predictors of monthly income and build an appropriate model for it. We will answer the following questions in this report: Which numerical factors significantly influence the monthly income of employees? Is it possible to apply a linear regression model to predict the monthly income of employees? Can a more complex model be developed to address multicollinearity issues in predicting the monthly income of employees? Which model has demonstrated superior predictive performance?

Dataset Introduction

In this dataset, there are 1480 observations and 38 attributes. We will use monthly income as the outcome variable of the model, which is measured in US dollars. The attributes of HR Analytics are described in Appendix 3. This data site was obtained via Kaggle (see references). The dataset was originally created by IBM Corporations.

This project's primary emphasis is on analyzing several linear regressions to fit the dataset. Out of the 1480 observations, there were 57 missing variables. The data set was cleaned and missing variables were excluded for this project before any regression analysis.

Methodologies

This project utilizes R Studio to graph all figures and plot all diagrams and correlations. We first access the trends present in our data, we use correlation matrices for both numerical and categorical variables. We also used simple linear regression to fit models of a large number of attributes and observations. This project will fit reduced models from significant values from large-scale linear regression. Some complex regression models will also be looked at, including stepwise regressions and log formations. This project uses several Residual-Plots to analyze normality, non-constance, heteroscedasticity, and outliers. This project also uses ANOVA and K-Fold Cross Validation to analyze simple and complex regression and compare the performance of each.

Data Analyses of HR Analytics

Question 1: Correlation of Data

From both correlation matrices, we can see that our predictor variable, monthly income, has the greatest correlation with total working years. The daily rate has the least correlation with monthly income. The response variable has a low correlation with most individual variables, this fact leads us to begin predicting the model with several variables. Research question one seeks to find which numerical factors significantly influence employees' monthly income. Looking at Figure 1 and Figure 2, we can see a highly significant correlation between age, total working years, years at a company, and our predictive variable, monthly income. Making these numerical values significantly impacts the HR Analytics dataset.

JobLevel	TotalWorkingYears
0.949834613	0.771806849
YearsAtCompany	Age
0.514469115	0.497485937
AgeGroup	YearsInCurrentRole
0.471849811	0.364421788
YearsSinceLastPromotion	YearsWithCurrManager
0.346064588	0.343707335
NumCompaniesWorked	Education
0.148862759	0.092182082
Department	MonthlyRate
0.054078038	0.034263359
WorkLifeBalance	BusinessTravel
0.030430453	0.030319803
RelationshipSatisfaction	OverTime
0.022756565	0.007061915
StockOptionLevel	DailyRate
0.006517209	0.005963495
EnvironmentSatisfaction	JobSatisfaction
-0.006913358	-0.008831710
DistanceFromHome	HourlyRate
-0.015817082	-0.015980501
PerformanceRating	JobInvolvement
-0.017378899	-0.017496407
TrainingTimesLastYear	PercentSalaryHike
-0.021270056	-0.028259586
Gender	EducationField
-0.031134010	-0.042061659
MaritalStatus	JobRole
-0.076613083	-0.093236913
Attrition	SalarySlab
-0.157672485	-0.831154040

Figure 1: Correlation Matrix looking at Variables of interest. Value closest to 1 or -1, suggest high correlation.

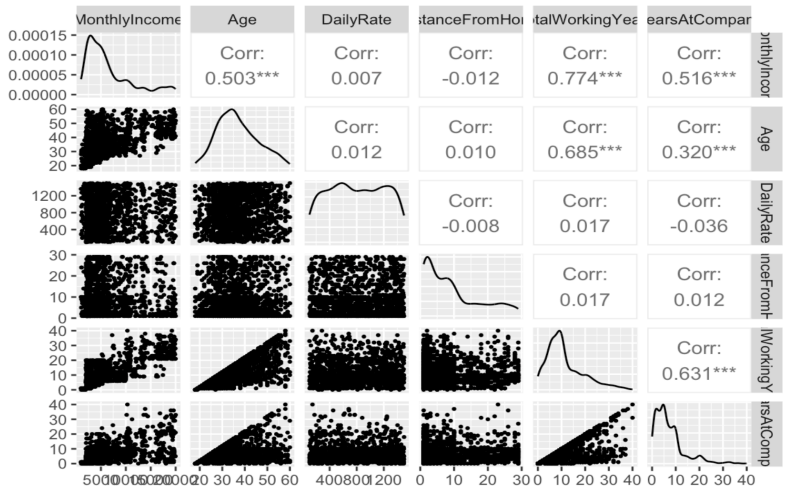


Figure 2: Distribution and correlation of core numerical variable. Bivariate scatterplots are on the bottom and correlation on the top.

When looking at correlation matrices, any negative value has a negative correlation, whereas all positive values are positively correlated to monthly income (Figure 1). Job level, total working years, and years at the company all have a positive correlation to the dependent variable, monthly income. Whereas, the salary slab has a strong negative correlation to the dependent variable.

Question 1 Conclusion: Using Figure 2 and some of our primary numerical variables being used for correlation. We can see that years at a company, total working years, and age are some top significant predictors for our dependent variable, monthly income.

Question 2: Linear Regressions

To further investigate this correlation, we can examine a full model of our HR Analytics data set to find if these variables remain good predictors, and if more variables will increase the performance of our model, including categorical variables.

The full linear regression model in Figure 3 allows us to see that the dependent variables: total working years, job role, job level, distance from home, and department, are the most significant predictors for monthly income at a significance level of 95%. **A Bonferroni correction was applied to the linear regression in Figure 3 to consider multiple testing adjustments.** Analyzing the residual plots in Figure 4, although the regression appears relatively normal, with only a few high leverage points, the linear regression indicates heteroscedasticity. The Residual vs. Fitted and Scale-Location plots also appear to be grouped, instead of uniformly distributed points. **The data is being clustered into several specific groups on the x-axis because of non-equal variance.** The most probable attribute causing this, since we are observing monthly income, would be job role or job level as those are typically associated with **income categorization.**

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.408e+03  6.189e+02   3.891  0.00345 **
Age                       -8.730e+00  1.300e+01  -0.672  1.00000
AgeGroup                  -1.898e+01  1.224e+02  -0.155  1.00000
Attrition                 -2.901e+00  1.071e+02  -0.027  1.00000
BusinessTravel             7.511e+01  5.220e+01   1.439  1.00000
DailyRate                  4.914e-02  8.758e-02   0.561  1.00000
Department                -5.881e+02  9.157e+01  -6.422 6.01e-09 ***
DistanceFromHome          -1.369e+01  4.337e+00  -3.156  0.05395 .
Education                 -1.527e+01  3.509e+01  -0.435  1.00000
EducationField            -1.192e+00  2.646e+01  -0.045  1.00000
EnvironmentSatisfaction   -1.192e+01  3.249e+01  -0.367  1.00000
Gender                     5.972e+01  7.197e+01   0.830  1.00000
HourlyRate                 1.857e+00  1.733e+00   1.072  1.00000
JobInvolvement             1.196e+01  4.994e+01   0.239  1.00000
JobLevel                   3.323e+03  6.375e+01  52.122  < 2e-16 ***
JobRole                    7.426e+01  1.945e+01   3.817  0.00464 **
JobSatisfaction           -3.178e+01  3.216e+01  -0.988  1.00000
MaritalStatus              2.984e+00  6.516e+01   0.046  1.00000
SalarySlab                -9.096e+02  6.190e+01 -14.694  < 2e-16 ***
MonthlyRate               -3.748e-03  4.946e-03  -0.758  1.00000
NumCompaniesWorked        -1.485e+01  1.586e+01  -0.937  1.00000
OverTime                  -3.325e+01  8.186e+01  -0.406  1.00000
PercentSalaryHike          1.566e+01  1.512e+01   1.036  1.00000
PerformanceRating         -9.164e+01  1.532e+02  -0.598  1.00000
RelationshipSatisfaction   4.719e+00  3.266e+01   0.144  1.00000
StockOptionLevel          -5.034e+01  5.533e+01  -0.910  1.00000
TotalWorkingYears          3.696e+01  9.861e+00   3.748  0.00611 **
TrainingTimesLastYear     -1.684e+01  2.742e+01  -0.614  1.00000
WorkLifeBalance           -2.348e+01  4.997e+01  -0.470  1.00000
YearsAtCompany             1.085e+01  1.222e+01   0.888  1.00000
YearsInCurrentRole        -2.583e+01  1.603e+01  -1.611  1.00000
YearsSinceLastPromotion    9.683e+00  1.411e+01   0.686  1.00000
YearsWithCurrManager      -3.864e+01  1.658e+01  -2.330  0.65782
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1340 on 1447 degrees of freedom
Multiple R-squared:  0.9204,    Adjusted R-squared:  0.9187
F-statistic: 523.1 on 32 and 1447 DF,  p-value: < 2.2e-16
```

Figure 3: Linear Regression of full model, utilizes Bonferroni correction.

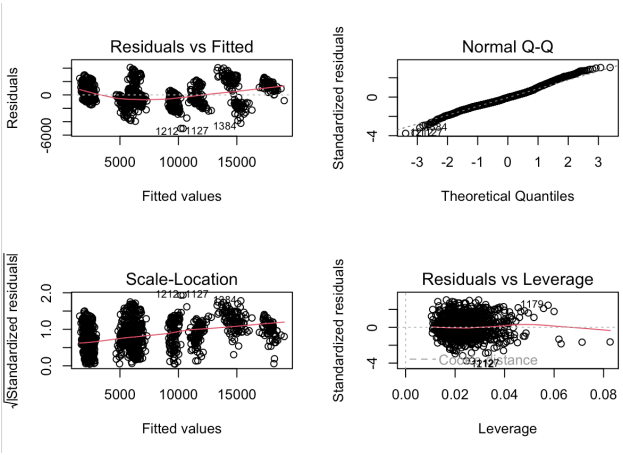


Figure 4: Residual/diagnostic plots for the full linear regression model.

Reading the linear model code, we view positive slopes by positive coefficients, and negative slopes by negative coefficients. Reducing the model to only significant values from our Bonferroni correction leaves five variables: department (negative correlation), job level (positive correlation), job role (positive correlation), salary slab (negative correlation), and total working years (positive correlation). All of these dependent variables are significant (p-value of 0.01).

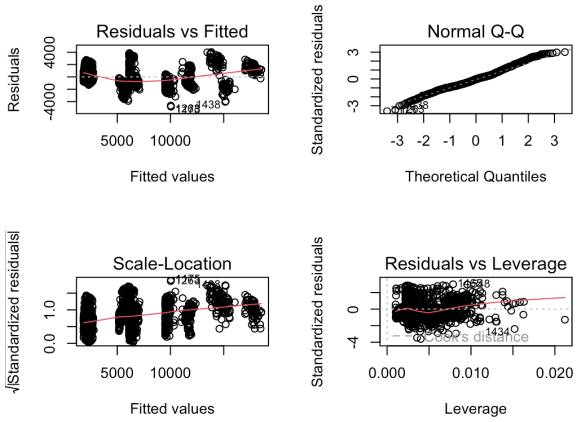
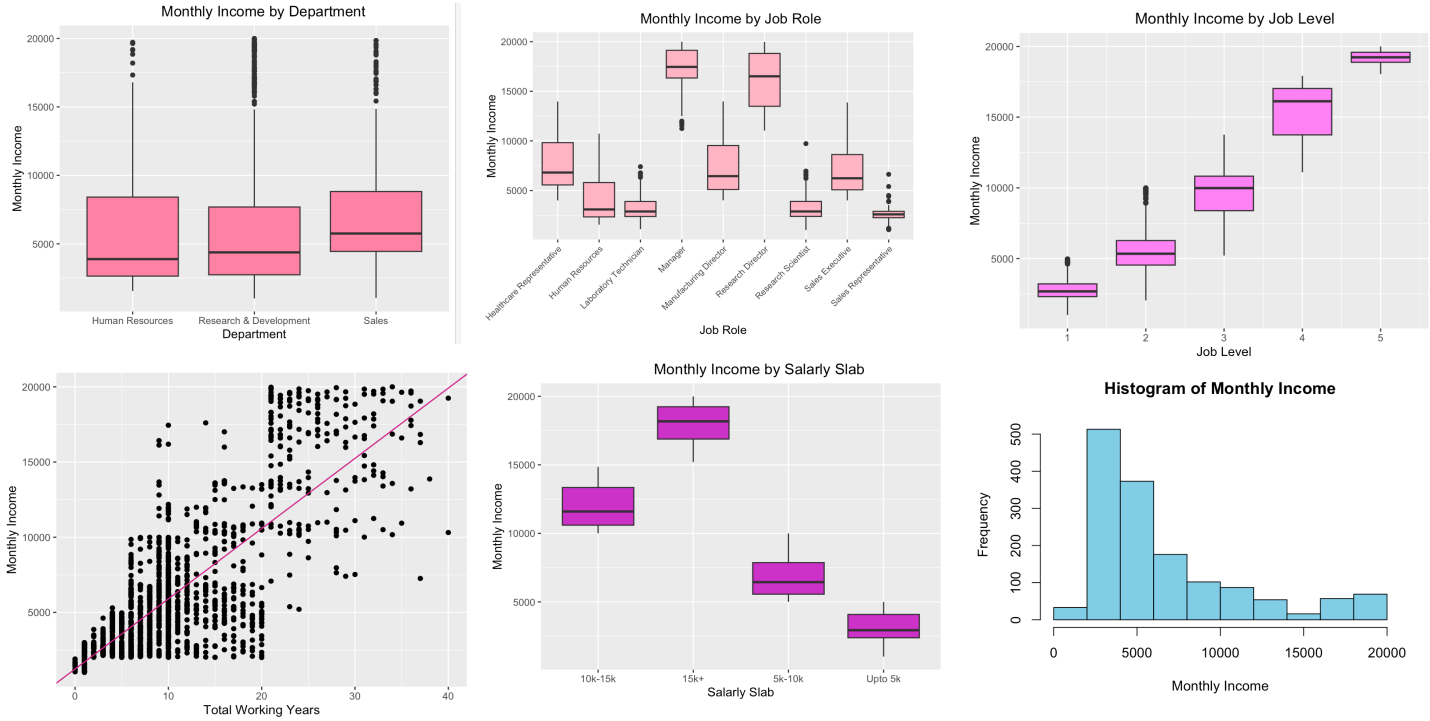


Figure 5: Residual/diagnostic plots for the reduced linear regression model.

Figure 5 showcases very similar diagnostic issues in the reduced linear regression model of these variables that the initial full model linear regression showcased. Residual groupings and heteroscedasticity still appear in this particular set of data.

To analyze this unique pattern of residuals, we have chosen to plot all of the significant variables in relationship to monthly income and also plot a histogram of monthly income to rule out any discrete variable overdispersion. We can tell some pretty obvious grouping in the box plots of job roles: it has low roles, medium roles, and high paying roles amongst the

Figure 6: Comparison of 5 significant predictor variables of monthly income, and monthly income distribution



several seen in HR_Analytics. Job level and salary slab also have pretty distinguishable groupings, which may be causing the pattern in the residual plots. **As salary slab deals with brackets of salary levels within an organization, further analysis of the variable, salary slab, reveals a right-skewed distribution (App 1).** The histogram of monthly income is left skewed, but not overdispersed so it is not the cause behind the groupings. By further testing and through backward stepwise regression, the categories salary slab and job level are the main distributors to this grouping in the residual plots and likely the cause of the non-equal variance. By further reducing this model, we obtain the residual plot, Figure 7, where grouping is diminished, the regression appears normal, with less leverage, but there is a decrease in R^2, and only 57% of the variability in monthly income can be explained by the 3 remaining variables- job role, department, and total working years. This makes sense as salary slab and job level were main contributors to both the negative correlation (salary slab: -0.83) and the positive correlation (job level: 0.95) and hence a reduced explanation of variability in the further reduced model. The Anova pictured in Figure 8 compares the first reduced model to a further reduced model, which has no grouping and is deemed significant compared to the reduced model, which was not.

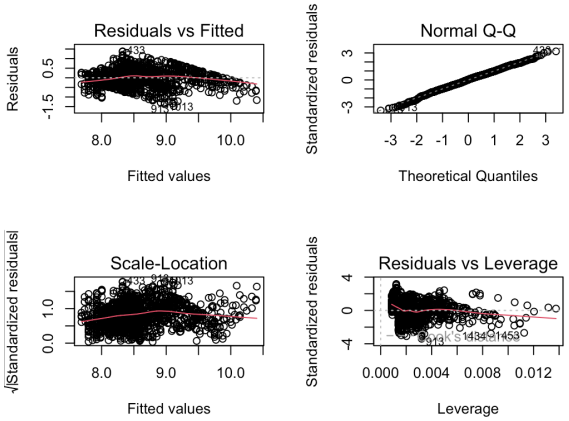


Figure 7: Residual plots for further reduced linear regression.

```
Analysis of Variance Table

Model 1: MonthlyIncome ~ Department + JobLevel + JobRole + SalarySlab +
    TotalWorkingYears
Model 2: MonthlyIncome ~ Department + JobRole + TotalWorkingYears
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1474  77.69
2   1476 278.96 -2   -201.27 1909.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8: ANOVA of reduced and further Reduced linear regression.

Question 2 Conclusion: Based on this analysis, it is unlikely that a simple linear regression will be able to accurately predict monthly income. None of the regressions fit any suitable linear models that supplied both a high R^2, to explain monthly income predictable by the remaining dependent variables. Nor did any simple linear regression model (with reasonable R^2 value) not contain groupings in the fitted values. **Additional log transformation on the reduced model does not alter either of these issues (Appendix 2)**. By searching for a more complex model to assist in this multicollinearity issue, can potentially find a regression that fits monthly income.

Question 3: More Complex Regression

As simple linear regression was not able to fit a model for monthly income, attempting to use a stepwise method to reduce parameters on both sides and apply a linear regression model to the reduced variables. The stepwise regression found more several attributes significant than the reduced models. The most significant variables in the stepwise not previously recorded in the reduced mode included, attrition, education, number of companies worked for, and years in current roles.

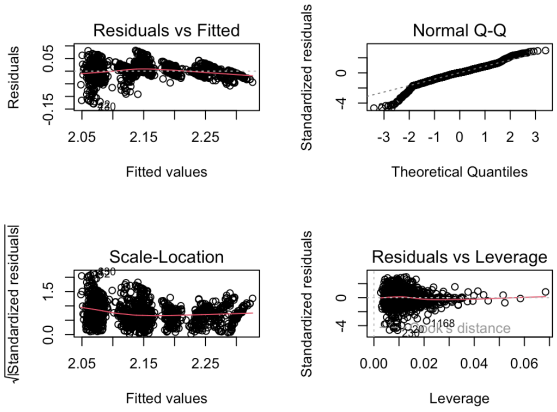


Figure 9: Residual plots for stepwise regression.

The residual plots for the stepwise regression have some high leverage points, there is still a grouping within residual vs fitted and scale-location plots, however, fitted values are much closer on this stepwise regression compared to linear regressions. The lower tail of the QQ plot is below the line and demonstrates a right-skewed distribution. Residual vs fitted also illustrates a funnel shape, leading towards probabilities of non-constance.

Question 3 Conclusions: The stepwise regression is a better predictor for monthly income. However, this model still faces issues, including the grouping of fitted values in residual plots. Right-skewed distributions, and some higher leverage points, It does offer an R^2 of 87%, meaning 87% of the variation in monthly income can be predicted by the dependent variables. It is difficult to say that this model fully addresses multicollinearity issues but it is closer than simple linear regression models where.

Question 4: Performance of Models

To test the performance of each model we discussed in this project, we will use K-Fold Validation. K Fold Validation will inform us of the root mean squared error and the mean absolute error. The highest-performing model will obtain the lowest values of these errors. The highest-performing model will also have a large R^2, which helps explain the variability of the independence variable, monthly income.

In Table 2, the RMSE and the MSE of the models are … Further Reduced model > Full model > Reduced model > Stepwise model. The R^2 shows us the Further Reduced mode < Full model < Stepwise model = Reduced mode. This means that the overall performance of these models is Further Reduced Model < Full model < Reduced model < Stepwise model.

	Full Model	Reduced Model	More Reduced	Stepwise
RMSE	1366	1349	2971	1351
R^2	0.916	0.918	0.604	0.916
MSE	1063	1054	2168	1054

Table 2: K Fold Validation Results

Question 4 Conclusions: The stepwise regression demonstrates the greatest performance. However, we still need to consider the non-constance, non-normality; and fitted value groupings before declaring this is a reliable or accurate model.

Conclusions

Using simple linear regression, a more complex stepwise regression, ANOVA, K Fold validation, and several residual plots we were able to draw insightful conclusions about monthly income from HR Analytics regarding our dependent variable monthly income. From our correlation tables, we know several factors that are correlated with monthly income, including the top three attributes correlated to monthly income; job level, salary slab, and total working years. Our simple linear regressions, Full model, and Reduced model, quickly have an issue with the fitted values in the residual plots. Instead of a uniform distribution of residuals, we see “groupings” of residuals into distinct fitted values. After examining both of these models using Figure 6, we could conclude that this **extreme fitting originated from error variances not being equal.** The reduced model saw better accounts of residual groupings but it had a R^2 of 0.604 compared to 0.916 from full and 0.918 from reduced models. Our stepwise model plots showed better groupings, but it was not completely gone and was paired with fanning, indicating non-constance.

We concluded the best model for predictive performance was stepwise, however, these issues are taken into consideration and acknowledged. Because of this residual problem, our recommendation is for the variables salary slab and job level to be further analyzed within each category and use “groups” of these categories to find a more predictive model. From our research questions, we could not formally find one model that explained variation significantly, obeyed laws of equal variance, heteroscedasticity, and obtained a normal distribution. However, with further exploration we are confident that this can be done. Finding a predictive linear model of monthly income was incredibly difficult with so many attributions and such broad research questions. Narrowing down research questions to smaller amounts of variables will give us more meaningful conclusions.

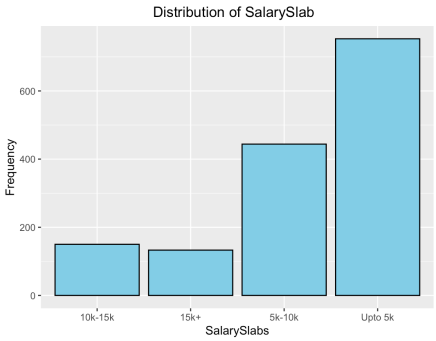
References

Angrave D., Charlwood A., Kirkpatrick I., Lawrence M., Stuart M. HR, and Analytics (2016) Why HR Is Set to Fail the Big Data Challenge: Why HR Is Set to Fail the Big Data Challenge. *Hum. Resour. Manag. J*, 26:1–11. doi: 10.1111/1748-8583.12090.

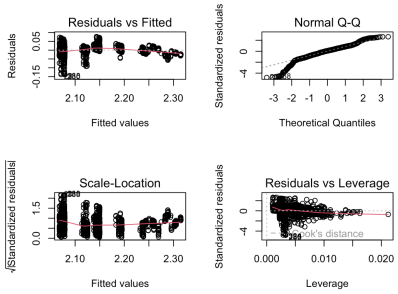
John W. Boudreau & Janet H. Marler (2017) An evidence-based review of HR Analytics, The International Journal of Human Resource Management, 28:1, 3-26, doi: 10.1080/09585192.2016.1244699.

Tahir, M. (2023). *HR Analytics*. Kaggle. https://www.kaggle.com/datasets/mohammadkaiftahir/hr-analytics

Appendix:



App 1: Distribution of SalarySlabs



App 2: Log Transformation of the reduced Model.

App 3: Attribute description of all values found significant in all models.

Attribute	Contribution
Age	Age of the employee (Numerical discrete; years)
Distance from home	Distance of the employee's home from the workplace. (Numerical continuous; miles)
Department	Department in which the employee works. (Categorical)
Job level	Job level or position of the employee. (Categorical (0,1,2, etc.)
Job role	Role or position in the job. (Categorical)
Salary slab	Categorized salary slab of the employee. (Categorical)
Total Working Years	Total number of years the employee has worked. (Numerical discrete)
Years with curr manager	Number of years with the current manager. (Numerical discrete)