­­­­­DOCUMENT APPROVAL SHEET

DATA PREDICTION PROJECT

SALARYDATA

TOTAL YEARLY SALARY PREDICTION

gold boundless bar

**DOCUMENT IDENTIFICATION:**

MSBA AY 21-22/FINAL PROJECT/TOTAL YEARLY SALARY PREDICTION

**DOCUMENT AUTHOR:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **AUTHORED BY/ ID** | **SECTION** | **DEPT./ ROLE** | **COHORT** | **DATE** | **CHECKED BY** | **DATE** |
| Teja Alluru  (1150077) | All | MSBA/ Student | 2021-2022 | 19-Dec-2021 | Teja Alluru (1150077) | 19-Dec-2021 |

**DOCUMENT APPROVAL:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **APPROVED BY/ ID** | **DEPT.** | **SCHOOL** | **SIGNATURE** | **DATE** |
| Teja Alluru (1150077) | MSBA/ Student | University of Washington, Tacoma, Milgard School of Business |  | 19-Dec-2021 |

*AUTHORED BY: Authored by means of a person/entity responsible for the creation of the document.*

*APPROVED BY: Approval means a person/entity responsible for the approval of the document content.*

*GRADED BY: Graded by means a person/entity having the authority to grade the assignment.*

*COPYRIGHT: This document was released for the viewing of the MSBA department only.*

INDEX OF CONTENTS

gold boundless bar

[DOCUMENT APPROVAL SHEET 1](#_Toc90824721)

[INDEX OF CONTENTS 3](#_Toc90824722)

[INDEX OF FIGURES 5](#_Toc90824723)

[INDEX OF TABLES 6](#_Toc90824724)

[1 ABSTRACT 7](#_Toc90824725)

[1.1 Purpose 7](#_Toc90824726)

[1.2 Benefit: 7](#_Toc90824727)

[1.3 Methodology 7](#_Toc90824728)

[1.3.1 Business/Project Methodology: 7](#_Toc90824729)

[1.3.2 Data Analysis Method: 9](#_Toc90824730)

[1.3.3 Software/ Programs: 9](#_Toc90824731)

[1.4 Conclusion/Recommendation 9](#_Toc90824732)

[2 INTRODUCTION / BUSINESS UNDERSTANDING 10](#_Toc90824733)

[2.1 PROJECT DETAILS 10](#_Toc90824734)

[2.1.1 Brief Project Background/Understanding 10](#_Toc90824735)

[2.1.2 Research Question/Project Scope 10](#_Toc90824736)

[2.2 PROJECT GOAL AND OBJECTIVES 10](#_Toc90824737)

[2.2.1 Goal: 10](#_Toc90824738)

[2.2.2 Objectives: 11](#_Toc90824739)

[2.3 MEETING/COMMUNICATIONS PLAN 11](#_Toc90824740)

[2.4 STAKEHOLDERS CONTACT INFORMATION 11](#_Toc90824741)

[2.5 RESPONSIBILITIES OF EACH PARTY 11](#_Toc90824742)

[2.6 PROJECT AGREEMENT SIGNATURES 12](#_Toc90824743)

[3 DATA UNDERSTANDING/DATA PREPARATION 13](#_Toc90824744)

[3.1 Initial Data Collection Report 13](#_Toc90824745)

[3.2 Data Description 13](#_Toc90824746)

[3.2.1 Data Dictionary 14](#_Toc90824747)

[3.3 Data Exploration 15](#_Toc90824748)

[3.3.1 Initial Trimming of Data 15](#_Toc90824749)

[3.3.2 Summary Statistics – Trimmed Data 16](#_Toc90824750)

[3.4 Condensing/Filtering Categorical Variables 18](#_Toc90824751)

[3.4.1 Location Data 18](#_Toc90824752)

[3.4.2 Education Data 19](#_Toc90824753)

[3.4.3 Company Data 19](#_Toc90824754)

[3.4.4 Title Data 20](#_Toc90824755)

[3.4.5 Gender Data 21](#_Toc90824756)

[3.5 Variable Summary – All 22](#_Toc90824757)

[3.6 Histograms – Univariate 23](#_Toc90824758)

[3.7 Descriptive Statistics 25](#_Toc90824759)

[3.8 Scatter Plots – Bivariate 26](#_Toc90824760)

[3.9 Correlation 30](#_Toc90824761)

[4 MODELING/EVALUATION 32](#_Toc90824762)

[4.1 Multiple Regression Model 32](#_Toc90824763)

[4.1.1 Model 1 – Output 32](#_Toc90824764)

[4.1.2 Model 1 – Residuals 33](#_Toc90824765)

[4.1.3 Model 2 – Output 36](#_Toc90824766)

[4.1.4 Model 2 – Residuals 37](#_Toc90824767)

[4.1.5 Model 3 (Log Model) – Output 40](#_Toc90824768)

[4.1.6 Model 3 – Residuals 41](#_Toc90824769)

[4.1.7 Test for collinearity (test with Variance Inflation Factor, VIF) 44](#_Toc90824770)

[4.2 Random Forest Model 45](#_Toc90824771)

[4.2.1 Model 1 – Output 45](#_Toc90824772)

[4.2.2 Model 1 – Variable Significance / Predicted Plots 46](#_Toc90824773)

[4.2.3 Model 2 – Output 47](#_Toc90824774)

[4.2.4 Model 2 – Variable Significance / Predicted Plots 47](#_Toc90824775)

[4.2.5 Model 3 (Log) – Output 49](#_Toc90824776)

[4.2.6 Model 3 – Variable Significance / Predicted Plots 49](#_Toc90824777)

[5 RESULTS 51](#_Toc90824778)

[6 LIMITATIONS 51](#_Toc90824779)

[7 CONCLUSION/RECOMMENDATIONS 52](#_Toc90824780)

[8 PREDICTIONS – MOST PARSIMONIOUS MODEL 54](#_Toc90824781)

[9 REFERENCES 59](#_Toc90824782)

INDEX OF FIGURES

gold boundless bar

[Figure 1‑1 Data Analytics Project CRISP-DM Methodology Ref. [3] 8](#_Toc90824787)

[Figure 1‑2 CRISP-DM – Stages Ref. [4] 8](#_Toc90824788)

[Figure 3‑1: Descriptive Statistics – Raw Data 15](#_Toc90824789)

[Figure 3‑2: Descriptive Statistics – Trimmed Data 16](#_Toc90824790)

[Figure 3‑3: NA Values – Trimmed Data 17](#_Toc90824791)

[Figure 3‑4: Salary Dataset – Null Values Removed 17](#_Toc90824792)

[Figure 3‑5: Condensed Location Data – Regions Ref.[5] 18](#_Toc90824793)

[Figure 3‑6: Condensed Location Data – Education 19](#_Toc90824794)

[Figure 3‑7: Condensed Location Data – Company 19](#_Toc90824795)

[Figure 3‑8: Condensed Role Data – Title 20](#_Toc90824796)

[Figure 3‑9: Condensed Gender Data 21](#_Toc90824797)

[Figure 3‑10: Variable Summary – All Variables 22](#_Toc90824798)

[Figure 3‑11: Histograms/Bar Plots – Numeric Data 23](#_Toc90824799)

[Figure 3‑12: Histograms/Bar Plots – Categorical Data 24](#_Toc90824800)

[Figure 3‑11: Univariate Statistics 25](#_Toc90824801)

[Figure 3‑14: Scatter Plots – Dependent Vs Independent 26](#_Toc90824802)

[Figure 3‑15: Scatter Plots – Dependent Vs Independent 27](#_Toc90824803)

[Figure 3‑16: Scatter Plots – Dependent Vs Independent – Log Total Salary 28](#_Toc90824804)

[Figure 3‑17: Scatter Plots – Dependent Vs Independent - Log Total Salary 29](#_Toc90824805)

[Figure 3‑14: Correlation Plots – Salary Data 30](#_Toc90824806)

[Figure 3‑15: Correlation Significance Table – Salary Data (P<0.0001 = “\*\*\*\*”) 31](#_Toc90824807)

[Figure 4‑1: Multi-Regression Model 1 - Output 32](#_Toc90824808)

[Figure 4‑2: Multi-Regression Model 1 – Model Residuals 33](#_Toc90824809)

[Figure 4‑3: Multi-Regression Model 1 – Variable Residuals 34](#_Toc90824810)

[Figure 4‑4: Multi-Regression Model 1 – Variable Residuals 35](#_Toc90824811)

[Figure 4‑5: Multi-Regression Model 2 - Output 36](#_Toc90824812)

[Figure 4‑6: Multi-Regression Model 2 – Model Residuals 37](#_Toc90824813)

[Figure 4‑7: Multi-Regression Model 2 – Variable Residuals 38](#_Toc90824814)

[Figure 4‑8: Multi-Regression Model 2 – Variable Residuals 39](#_Toc90824815)

[Figure 4‑9: Multi-Regression Model 3 - Output 40](#_Toc90824816)

[Figure 4‑10: Multi-Regression Model 3 – Model Residuals 41](#_Toc90824817)

[Figure 4‑11: Multi-Regression Model 3 – Variable Residuals 42](#_Toc90824818)

[Figure 4‑12: Multi-Regression Model 3 – Variable Residuals 43](#_Toc90824819)

[Figure 4‑13: Random Forest Model 1 - Output 45](#_Toc90824820)

[Figure 4‑14: Random Forest Model 1 – VariableSignificance 46](#_Toc90824821)

[Figure 4‑15: Random Forest Model 1 – Predicted vs Actual Plot 46](#_Toc90824822)

[Figure 4‑16: Random Forest Model 2 - Output 47](#_Toc90824823)

[Figure 4‑17: Random Forest Model 2 – Variable Significance 47](#_Toc90824824)

[Figure 4‑18: Random Forest Model 2 – Predicted vs Actual Plot 48](#_Toc90824825)

[Figure 4‑19: Random Forest Model 3 - Output 49](#_Toc90824826)

[Figure 4‑20: Random Forest Model 3 – Variable Significance 49](#_Toc90824827)

[Figure 4‑21: Random Forest Model 3 – Predicted vs Actual Plot 50](#_Toc90824828)

[Figure 7‑1: Yearlcompensations Vs Deciding Factors 52](#_Toc90824829)

[Figure 7‑2: Yearlcompensations Vs Top Companies 53](#_Toc90824830)

[Figure 8‑1: Predicted Vs Actual – Parsimonious Model 55](#_Toc90824831)

[Figure 8‑2: Predicted Vs Actual – Parsimonious Model – Cont… 56](#_Toc90824832)

[Figure 8‑3: Predicted Vs Actual – Parsimonious Model – Cont… 57](#_Toc90824833)

[Figure 8‑4: Predicted Vs Actual – Parsimonious Model – Cont… 58](#_Toc90824834)

INDEX OF TABLES

gold boundless bar

[Table 1‑1: Key Stakeholders Contact Information 11](#_Toc90824783)

[Table 3‑1: Data Dictionary: Salary Data 14](#_Toc90824784)

[Table 4‑1: Collinearity - VIF 44](#_Toc90824785)

[Table 5‑1: Model Results 51](#_Toc90824786)

# ABSTRACT

gold boundless bar

This section gives brief details of the project executed as a part of the TBANLT 540 final project. The main purpose/objective, the methodology used, the results, and the conclusion for executing the project are listed in the sections below.

## Purpose

gold boundless bar

Emphasizing the higher salaries associated with attracting new analytics professionals, it becomes necessary to understand the patterns in the total salary compensation offered by various companies. With record-breaking attrition rates that the world is seeing as a direct result of the pandemic, it gives a huge bargaining power for graduates of MSBA in the total salary compensation negotiations. Our team’s purpose is to create awareness through ML programs in identifying the total compensation that the market can offer to the students of MSBA to a high degree of accuracy. The project is also undertaken to fulfill the requirements of the TBANLT 540 subject final project.

## Benefit:

gold boundless bar

Having the ability to accurately predict the total salary compensation helps the MSBA students & Team 7 in the following ways-

1. To gain a better understanding of what variables affect the compensation and how they can improve upon those variables.
2. To identify what companies, offer better prospects in terms of total salary compensation.
3. To help Team 7 fulfill the requirements of the TBANLT 540 subject.

## Methodology

gold boundless bar

### Business/Project Methodology:

To execute the project, we will be following the CRISP-DM methodology of AGILE project management. The CRISP-DM methodology used in executing the project is shown in Figure 1‑1 and the various stages of the project are shown in Figure 1‑2.

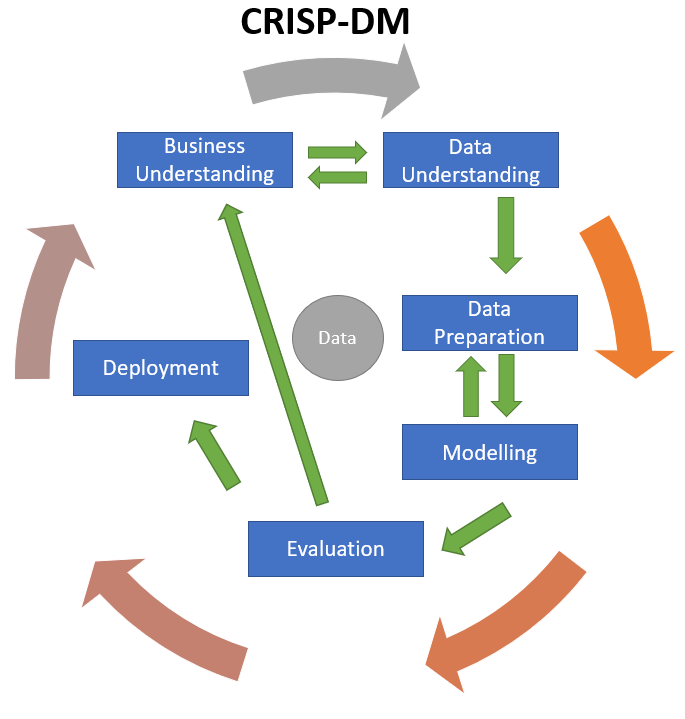


Figure 1‑1 Data Analytics Project CRISP-DM Methodology Ref. [3]

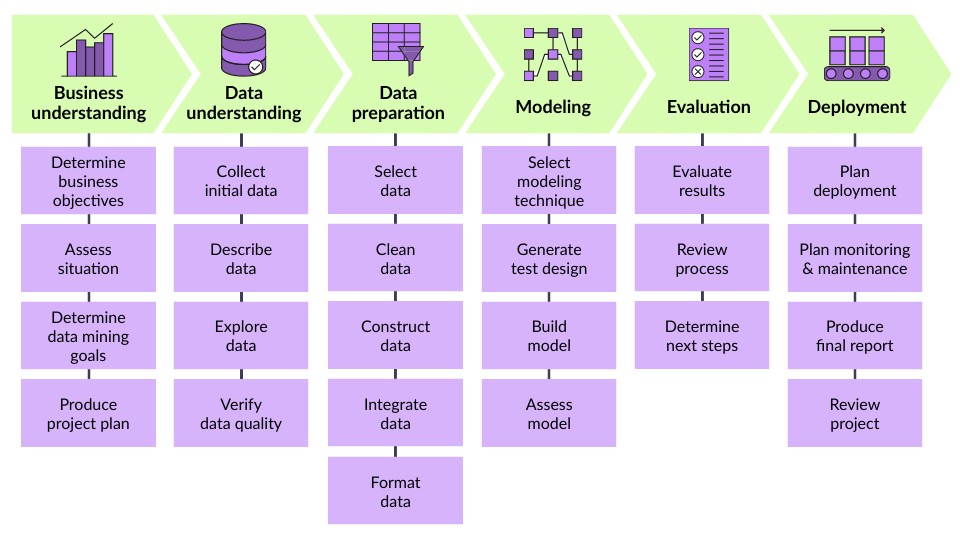


Figure 1‑2 CRISP-DM – Stages Ref. [4]

### Data Analysis Method:

The data analysis methods that we will use are as follows:

1. Plot Scatterplots, Histograms, and Residual Plots – these will check for any non-linearity in the data
2. Categorical Variable Cleaning – performing any dummy coding or collapsing any categories
3. Data Transformation – if any non-linearity is found (may use logarithmic transformation etc.)
4. Perform Ordinary Least Squares (OLS) Regression
5. Check for significant interaction and collinearity
6. Perform post regression analysis of model and individual variable residuals (heteroskedasticity, Cook’s distance, etc.)
7. Analyze explanatory variables significance using p-values and use VIF or partial F tests to determine if a variable is significant and could be excluded from the model
8. Determine parsimonious model

### Software/ Programs:

We will utilize .csv files for the original data source. Statistical/Data Analysis software R is used for data understanding, description, model analysis, and evaluating the model. Microsoft Word and Excel are used for the project documentation part.

## Conclusion/Recommendation

gold boundless bar

We recommend the students of MSBA to apply to the top 25 companies which pay higher salaries than others. Also, to check the years of experience to understand what salary they can negotiate for. By going through this report and running their statistics in R they should be able to predict what salary they can negotiate for to a ballpark figure.

# INTRODUCTION / BUSINESS UNDERSTANDING

gold boundless bar

## PROJECT DETAILS

gold boundless bar

This section gives details about the Project i.e., Project background understanding and the problems, challenges, and opportunities that must be addressed by the project and the scope of the project.

### Brief Project Background/Understanding

Being students of MSBA from the University of Washington Tacoma, we strongly believe that one of the most important aspects of any job search is to know what our experience and education, are worth in the market. A job seeker needs to know what are the prevailing trends in the market and how different variables like age, experience, gender, education, and race affect the total compensation.

Understanding the patterns behind the salary will equip the students with better bargaining power in job offers. This project is to help the students of MSBA to predict what salaries they can be offered in the market and what they can bargain for.

### Research Question/Project Scope

Total Yearly Salary Prediction: (MSBA Students)

To identify patterns from salary data and to create a machine learning model to predict the total salary compensation for the students of MSBA.

## PROJECT GOAL AND OBJECTIVES

gold boundless bar

### Goal:

To create a machine learning model to predict total salary compensation.

### Objectives:

1. To determine at least 5-6 variables responsible for indicating total salary compensation of employees using predictive analytics, by analyzing and digging deep-dive in historical salary data. To use any trends that they show as it links to a key in predicting what different employers might offer in the future. The key metrics identification will be quoted for further improvement tactics to be defined for a better model reliable for our goal.
2. To integrate the predictive model to existing MSBA student profiles in predicting the total compensation they might negotiate for in the market and use the model to increase the confidence of MSBA students in salary negotiations.

## MEETING/COMMUNICATIONS PLAN

The weekly status meeting has been scheduled for tracking the progress of the project, escalation of risks and issues from 24-Oct-2021 till the end of the project.

## STAKEHOLDERS CONTACT INFORMATION

The details of the Project's key stakeholders are in Table 1‑1 below.

| **Name** | **Email** |
| --- | --- |
| Kiran Tejaswi (Teja) Alluru | akteja@uw.edu |
| Shephali Jain | jainshep@uw.edu |
| Marion LaRocque | mtib76@uw.edu |
| Shilpi Karmakar​ | shilpikb@uw.edu |
| Shalini Bagadhi | shalinib@uw.edu |

Table 1‑1: Key Stakeholders Contact Information

## RESPONSIBILITIES OF EACH PARTY

All parties agree to share their data and insights on the project. Team 7 members agree to do their best and adhere to the deadlines/assignments posted on the TBANLT540 canvas page.

## PROJECT AGREEMENT SIGNATURES

| **Last Name, First Name** | **Signature** | **Date** |
| --- | --- | --- |
| Kiran Tejaswi (Teja) Alluru |  | 05-Nov-2021 |
| Shephali Jain |  | 05-Nov-2021 |
| Marion LaRocque |  | 05-Nov-2021 |
| Shilpi Karmakar​ |  | 05-Nov-2021 |
| Shalini Bagadhi |  | 05-Nov-2021 |
| Faculty Approval (Dr. Margo Bergman) |  |  |

# DATA UNDERSTANDING/DATA PREPARATION

## Initial Data Collection Report

Team 7 has been in continuous research to collect salary data from the global internet. Having worked through several days, team 7 was finally able to get the data for this project from Kaggle .com Ref.[1] and levels.fyi Ref. [2]. The data consists of location, demographics, education, company, and various other data. The dataset also includes the salary details of various companies across the globe.

The complete dataset consists of one .csv file “Levels\_Fyi\_Salary\_Data.csv”. There are no problems recorded while obtaining the data. A more in-depth understanding of the data is established in the chapters below.

## Data Description

The data is contained in one flat file and consists of mainly 62642 records. The number of features, records, and the detailed description of the features in the .csv file listed in section 2.1 is summarized in the data dictionary presented in section 2.2.1 below.

### Data Dictionary

The data dictionary for the salary data from “Levels\_Fyi\_Salary\_Data.xlsx” is presented in Table 3‑1 below. The initial dataset consists of 29 variables and 62642 observations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Description** | **Data Type** | **No of  Records** |
| timestamp | Date and Time the data was collected | TimeStamp | 62642 |
| company | Company for which the data is reported | Text/VarChar | 62642 |
| level | Designation/ level | Text/VarChar | 62642 |
| title | Job title | Text/VarChar | 62642 |
| totalyearlycompensation | Total yearly compensation | Numeric/Int | 62642 |
| location | Job location | Text/VarChar | 62642 |
| yearsofexperience | Job holder total experience | Numeric/Float | 62642 |
| yearsatcompany | Job holder experience in the company | Numeric/Float | 62642 |
| tag | Miscellaneous random data | Text/VarChar | 62642 |
| basesalary | The base salary for the job | Numeric/Int | 62642 |
| stockgrantvalue | Stock value granted for the job | Numeric/Int | 62642 |
| bonus | Bonus offered for the job | Numeric/Int | 62642 |
| gender | Gender of the job holder | Text/VarChar | 62642 |
| otherdetails | Miscellaneous random data | Text/VarChar | 62642 |
| cityid | City identification number | Numeric/Int | 62642 |
| dmaid | State identification number | Numeric/Int | 62642 |
| rowNumber | The row number of instances | Numeric/Int | 62642 |
| Masters\_Degree | Education of job holder: Masters 1=Yes, 0 = No | Numeric/Int | 62642 |
| Bachelors\_Degree | Education of job holder: Bachelors 1=Yes, 0 = No | Numeric/Int | 62642 |
| Doctorate\_Degree | Education of job holder: Doctorate 1=Yes, 0 = No | Numeric/Int | 62642 |
| Highschool | Education of job holder: Highschool 1=Yes, 0 = No | Numeric/Int | 62642 |
| Some\_College | Education of job holder: Some\_College 1=Yes, 0 = No | Numeric/Int | 62642 |
| Race\_Asian | Race of the job holder: Asian 1=Yes, 0 = No | Numeric/Int | 62642 |
| Race\_White | Race of the job holder: White 1=Yes, 0 = No | Numeric/Int | 62642 |
| Race\_Two\_Or\_More | Race of the job holder: Two or More 1=Yes, 0 = No | Numeric/Int | 62642 |
| Race\_Black | Race of the job holder: Black 1=Yes, 0 = No | Numeric/Int | 62642 |
| Race\_Hispanic | Race of the job holder: Hispanic 1=Yes, 0 = No | Numeric/Int | 62642 |
| Race | Race of the job holder | Text/VarChar | 62642 |
| Education | Education of job holder | Text/VarChar | 62642 |

Table 3‑1: Data Dictionary: Salary Data

## Data Exploration

The initial data exploration was carried out in R using summary statistics & visualizations and the summary statistics & univariate properties are listed in section 3.3.2 and section 3.3.3 respectively.

The summary descriptive statistics of the raw data are presented in Figure 3‑1 below.

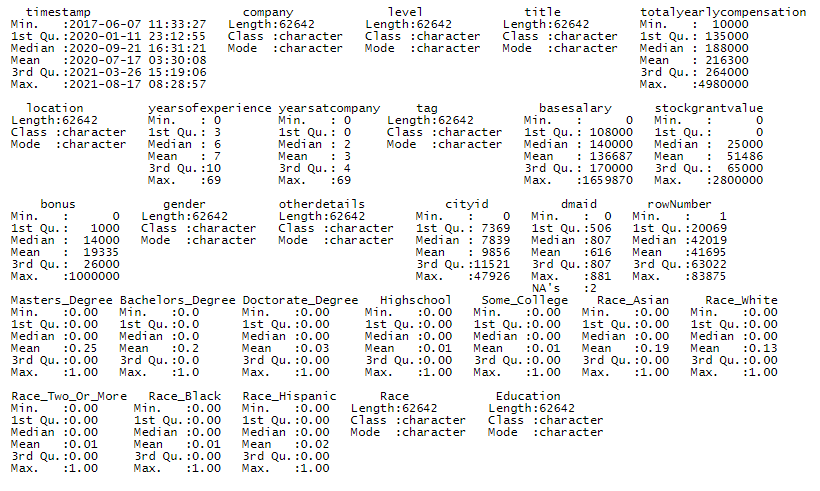


Figure 3‑1: Descriptive Statistics – Raw Data

### Initial Trimming of Data

Based on the initial visual data exploration from the .csv file and the data dictionary provided in Table 2‑1 the following observations were found.

* The level data is dropped outright as each company has a different level designated to their organizations and it mostly depends on the experience. So, dropping the column would not take out relevance/significance of levels in the prediction algorithm as it will be captured based on the total experience of the job holder.
* The location data is further split into three variables in total consisting of city, state, and country variables. The cityid and dmaid are also removed as they are the numeric identity numbers of the city and state of the position. The city and state details are already recorded.
* Row numbers do not make any sense as such are removed.
* Education variables consist of the required data and the data is repeated in five more columns as each of the categories in the education variable is again represented in a different column with 0’s and 1’s. Anyhow, we will be creating dummy variables during our analysis right now the other five columns are dropped and only the education variable is retained.
* Following the similar methodology described in the above point, for the race, only one column is retained, and the other five columns are dropped.
* All the initial data trimming is carried out in excel and no software is used, as for this project it seems cost & time effective

### Summary Statistics – Trimmed Data

The summary statistics of the trimmed data are presented in Figure 3‑2 below. As you can observe from Figure 3‑2 the NAs are not recognized as nulls and are part of the Class category as such R code is run and the NAs are presented in Figure 3‑3.

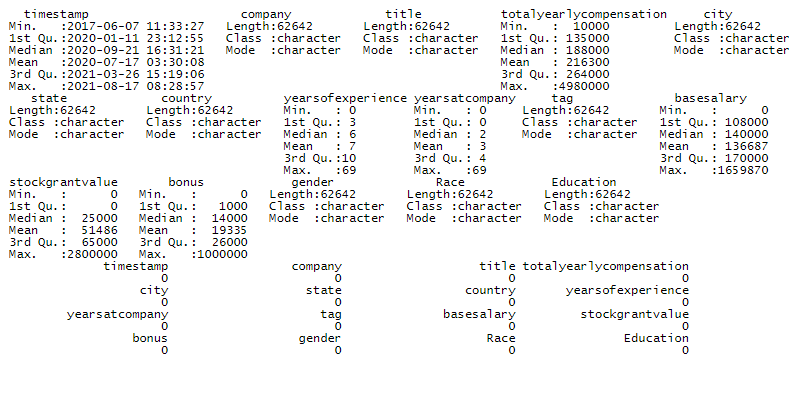


Figure 3‑2: Descriptive Statistics – Trimmed Data

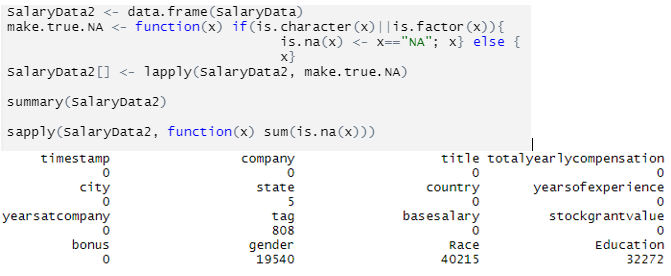


Figure 3‑3: NA Values – Trimmed Data

Since we have a vast number of records of 62642, we will go ahead and remove the NA rows to prepare our data for modeling. The R code is used to omit the NA values and the revised summary statistics of the new dataset are presented in Figure 3‑4

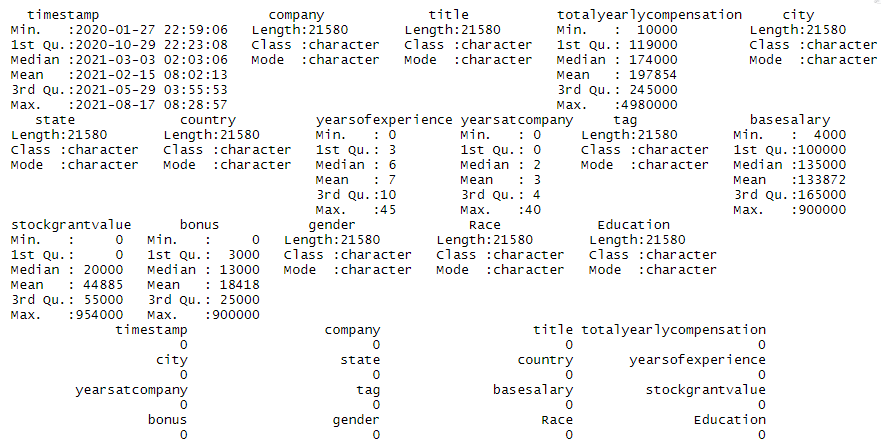


Figure 3‑4: Salary Dataset – Null Values Removed

## Condensing/Filtering Categorical Variables

Having observed the initial statistics for the data after the null values were removed. Based on the statistics shown in Figure 3‑4, we have decided to condense some of the categorical variables and filter some of the variables in the dataset. This is carried out to make sure that our regression model runs smoothly and efficiently.

### Location Data

Since the data we are interested in is only about salary details of companies located in the USA, we have filtered the data for country == USA. Further, the location data is condensed into nine categorical regions based on the geographic wealth of the different regions in the USA as shown in Figure 3‑5. The categorization of regions based on wealth is taken from Ref.[5].



Figure 3‑5: Condensed Location Data – Regions Ref.[5]

### Education Data

The education categorical variable has only 4 levels and each of the levels has enough observations as seen in Figure 3‑6. Therefore, there is no need to condense the data. However, we have decided to enforce the order based on levels of study.

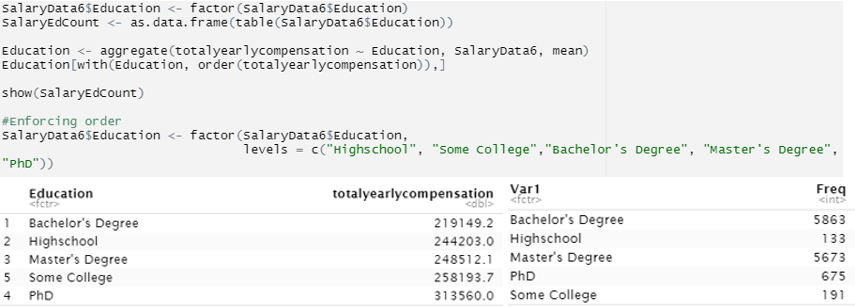


Figure 3‑6: Condensed Location Data – Education

### Company Data

The company data is filtered based on the frequency, i.e., only companies with observation greater than or equal to 25 are selected as shown in Figure 3‑6. This is done as fewer observations don’t generate much information related to companies. The total number of companies is now reduced to 96 from 1085. The total number of observations in the dataset after this filter is 12535.

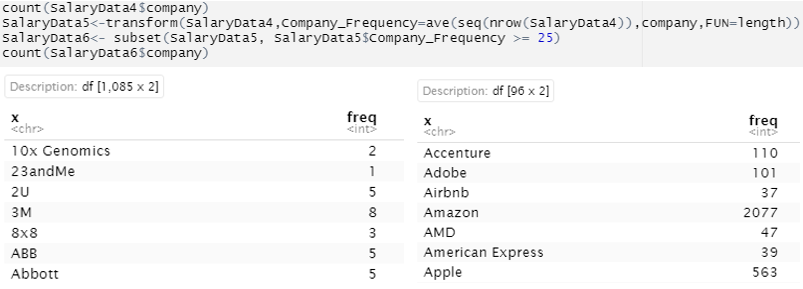


Figure 3‑7: Condensed Location Data – Company

### Title Data

The title variable has 15 levels and each of the levels has enough observations as seen in Figure 3‑8. Therefore, there is no need to condense the data. However, we have decided to put recruiter at number 1, because we need a business analyst, whichever we put here first will act as a baseline and will not show a coefficient.

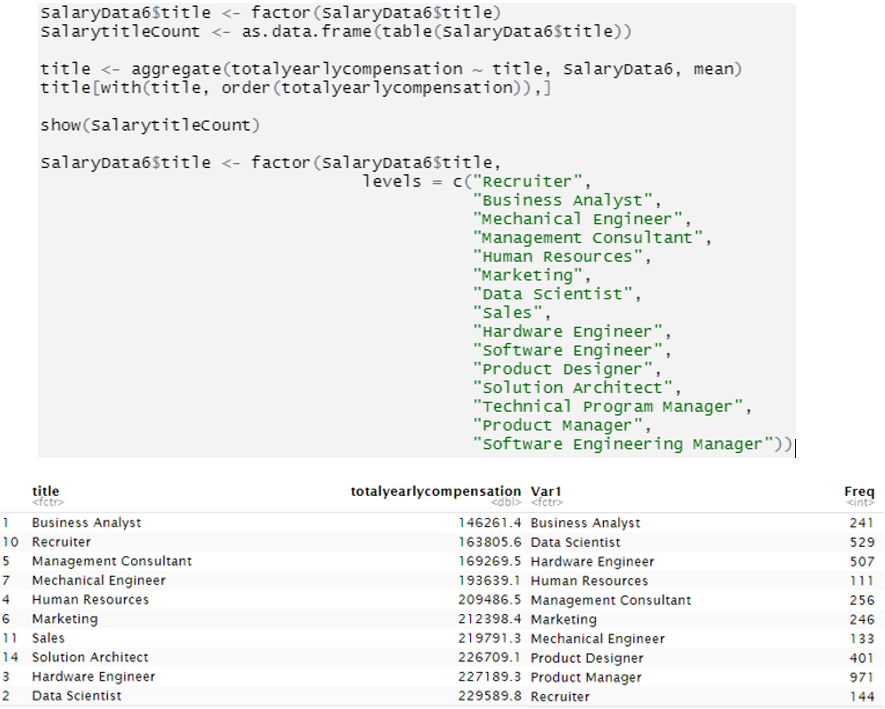


Figure ‑: Condensed Role Data – Title

### Gender Data

The gender variable has 3 levels namely male, female and other. The other instance has very few observations as compared to the other two, therefore we have enforced the variable order with others at first so the data for dummy variables created for male and female are retained. The code runs for enforcing the order, the ranking based on total compensation, and the number of observations for each level is presented for reference in Figure 3‑9.

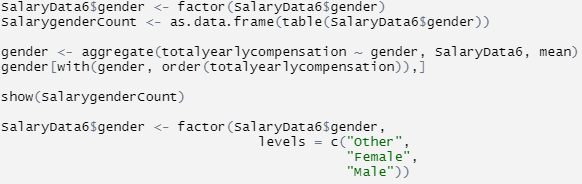






Figure 3‑9: Condensed Gender Data

## Variable Summary – All

The variable summary for all the variables in consideration is presented in Figure 3‑10.

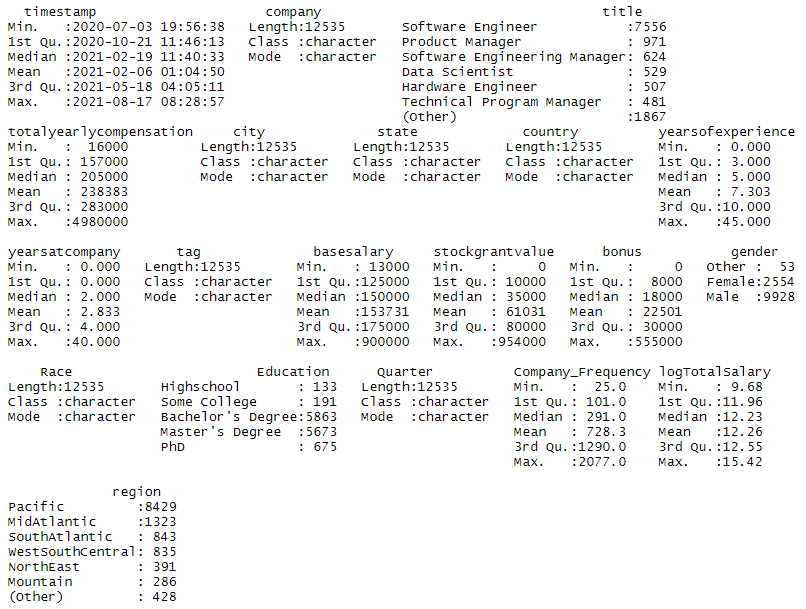


Figure 3‑10: Variable Summary – All Variables

## Histograms – Univariate

The univariate histograms/bar plots for all the variables in consideration are presented in Figure 3‑11 and Figure 3‑12.

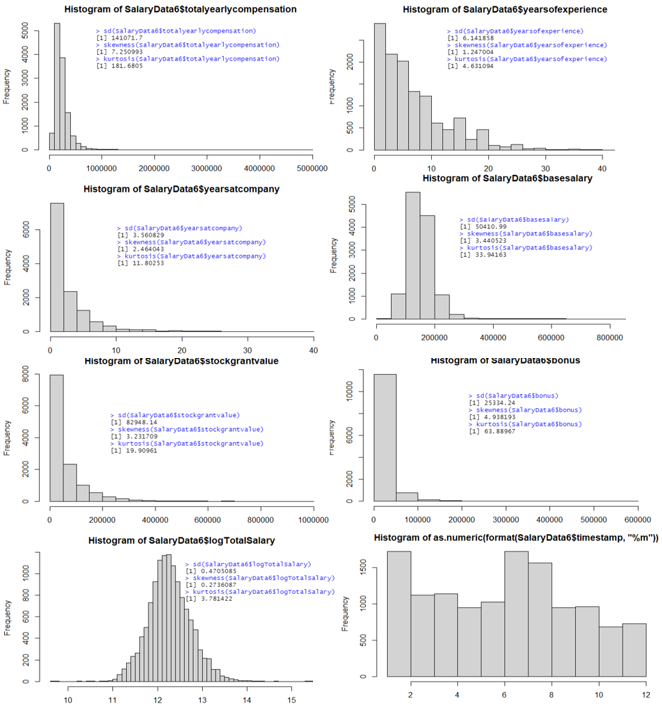


Figure ‑: Histograms/Bar Plots – Numeric Data

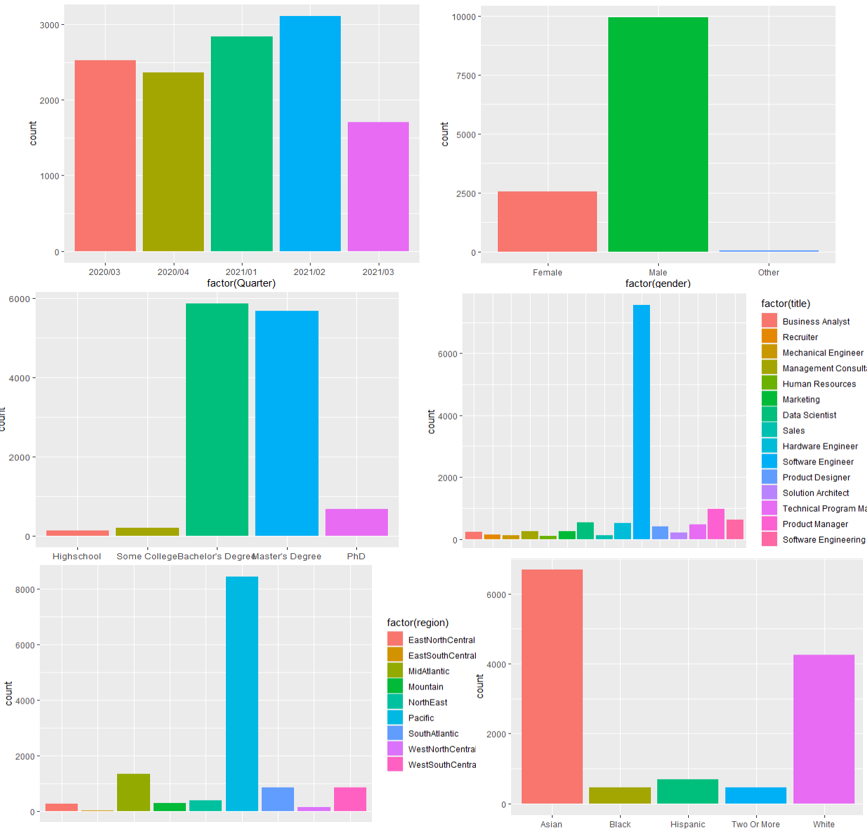


Figure ‑: Histograms/Bar Plots – Categorical Data

## Descriptive Statistics

The univariate statistics for the features and label for the condensed data are presented in Figure 3‑11 below. The visualization and complete summary of the univariate properties are presented in “Univariate Statistics.xlsx”. The total number of observations in the dataset is now 12535.

The basic deductions from Figure 3‑11 below are

1. No missing values as all the missing values have been removed from the steps followed in the chapters above.
2. Skewness and kurtosis problem for all variables.



Figure 3‑11: Univariate Statistics

## Scatter Plots – Bivariate

The bivariate scatter plots for dependent variable vs independent variables in consideration are presented in Figure 3‑14 and Figure 3‑15. The bivariate statistics for the log of the dependent variable and independent variable are presented in Figure 3‑16 and Figure 3‑17.

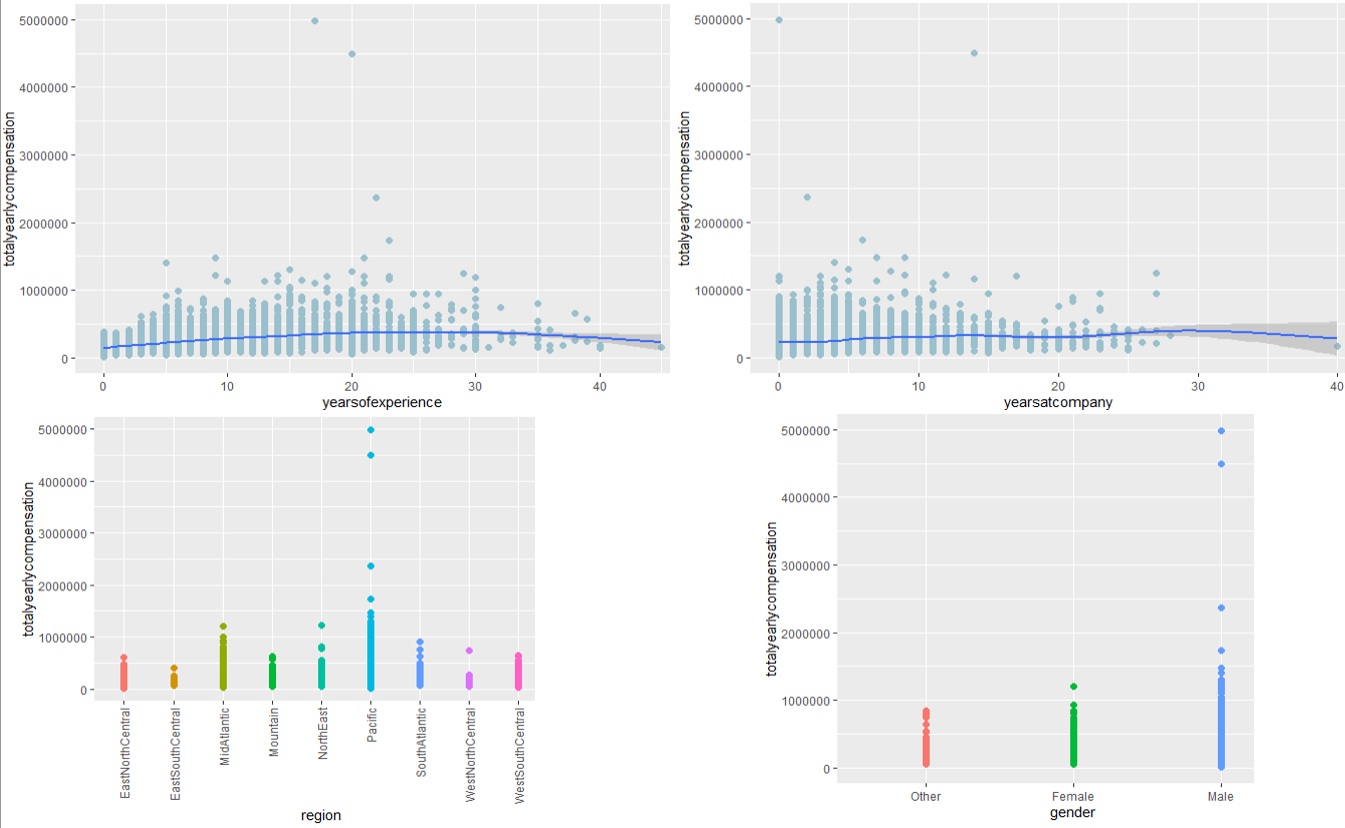


Figure ‑: Scatter Plots – Dependent Vs Independent

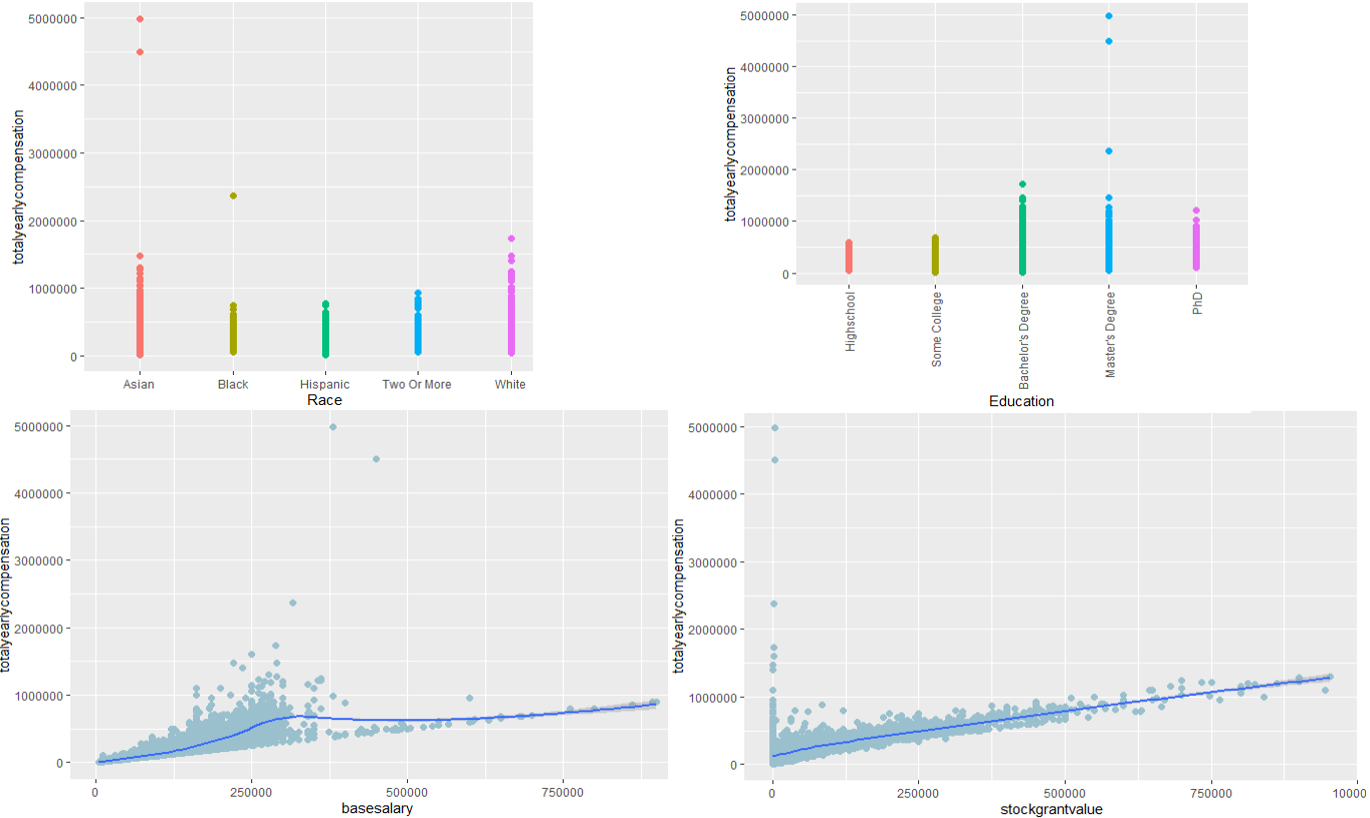


Figure ‑: Scatter Plots – Dependent Vs Independent

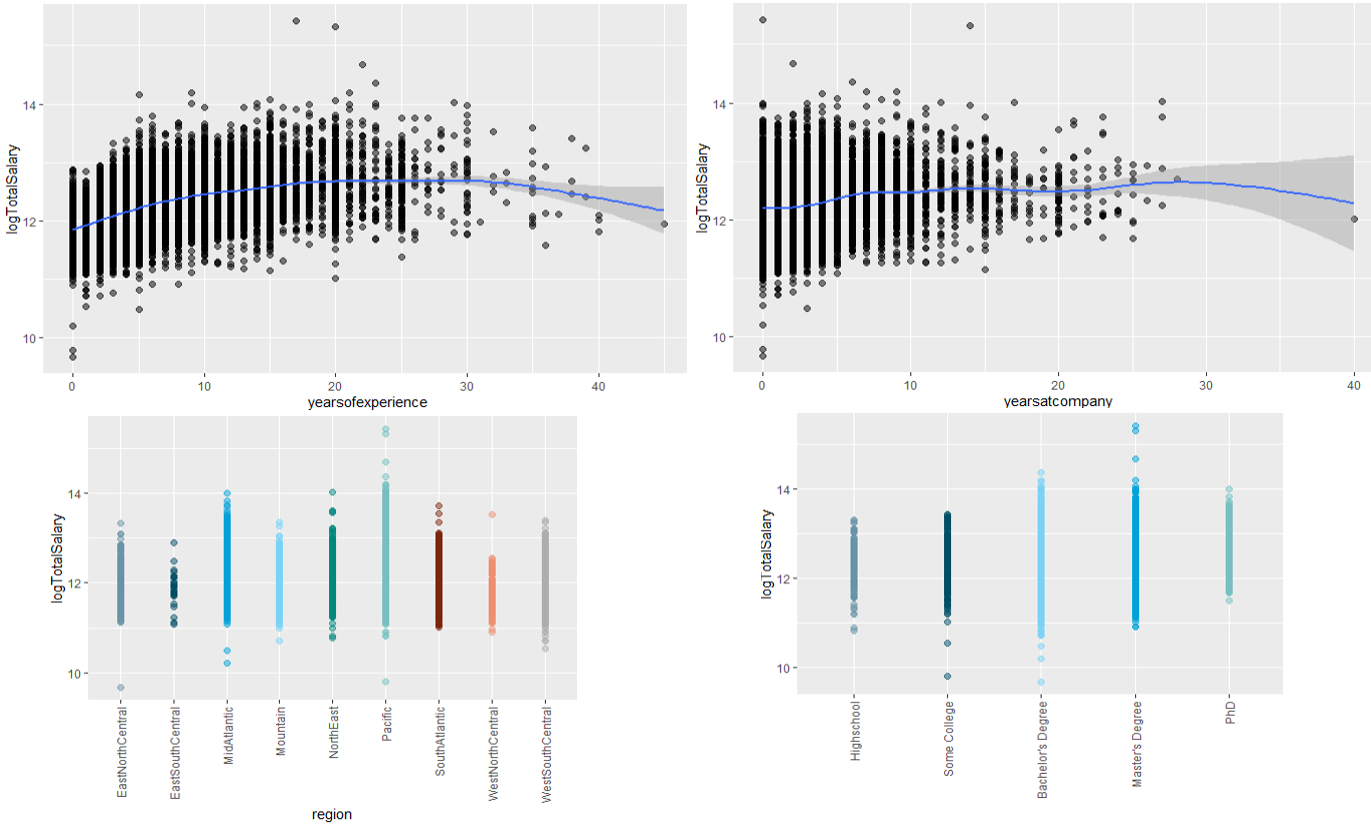


Figure 3‑16: Scatter Plots – Dependent Vs Independent – Log Total Salary

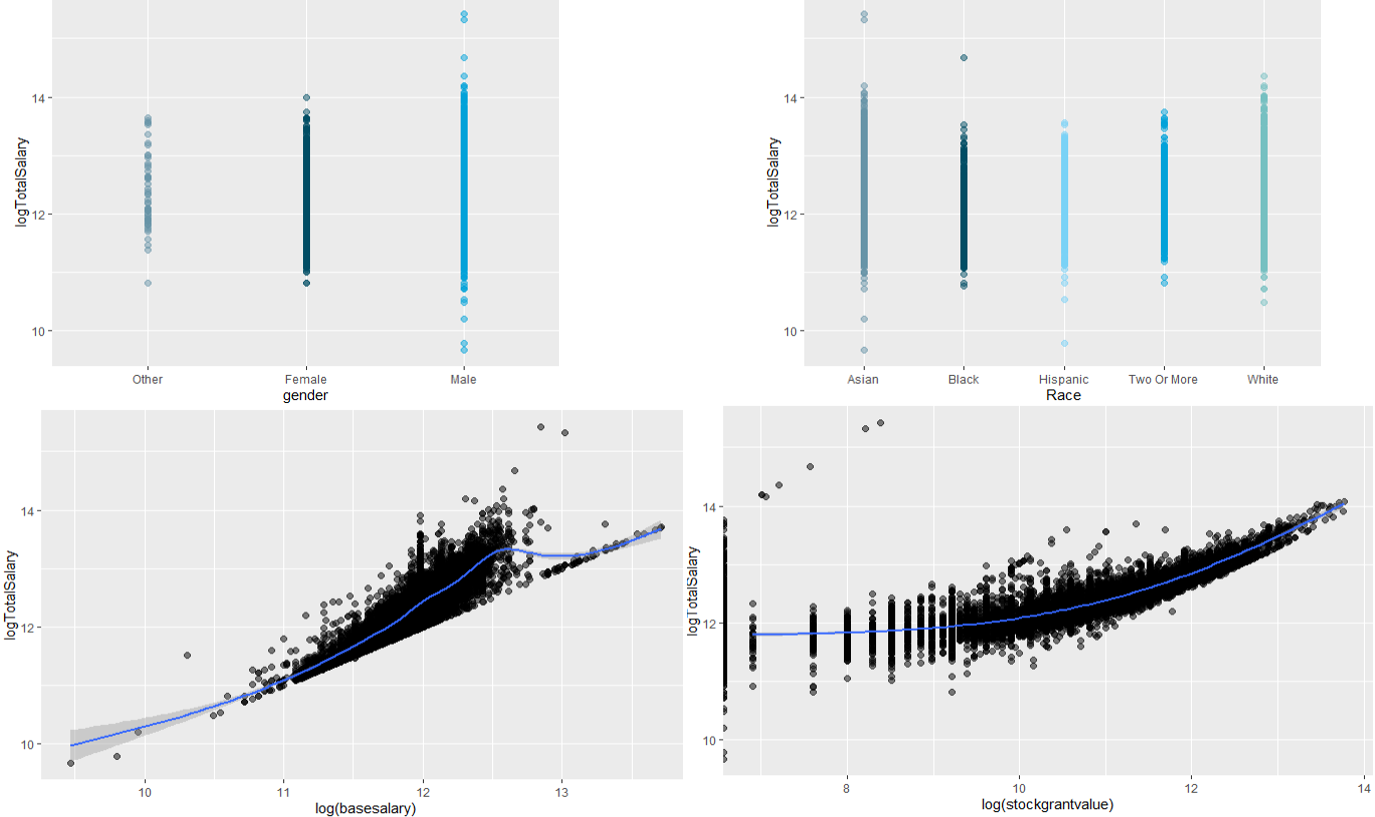


Figure 3‑17: Scatter Plots – Dependent Vs Independent - Log Total Salary

## Correlation

Since we are predicting total salary compensation which is a combination of other salary components like base salary, stock grants, and bonus we can ignore the other variables. To cross verify a simple correlation plot and correlation significance table are created in Figure 3‑14 and Figure 3‑15 respectively. As stated above it's clear from Figure 3‑14 that a strong correlation exists between base salary, stock grants, bonus, and totalyearly compensation as total yearly compensation is derived by adding the others.

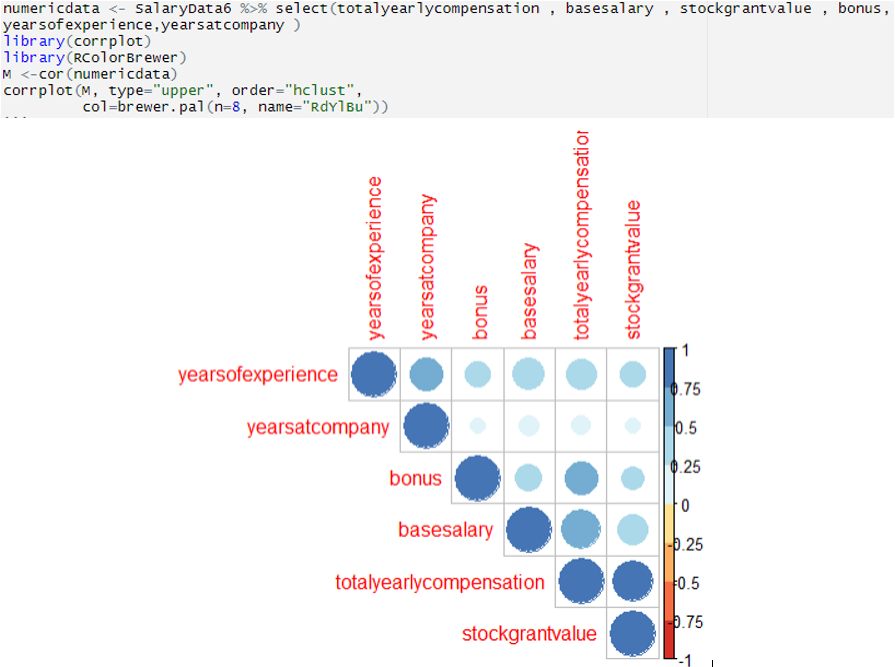


Figure ‑: Correlation Plots – Salary Data



Figure ‑: Correlation Significance Table – Salary Data (P<0.0001 = “\*\*\*\*”)

# MODELING/EVALUATION

Different models were built using different variables and parameters. For this model as per the content learned in the course, we have selected to go with multiple linear regression. As an additional point of the study, we have researched random forest and have executed random forest models as well.

To summarize the following model algorithms were used which are presented in the section below.

* + - 1. Multiple Regression Model
      2. Random Forrest Regression

## Multiple Regression Model

For multiple regression three models are run with different combinations of dependent variables. The output and the residual plots for all three models are presented in sections below.

### Model 1 – Output

The regression model 1 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, tag, and the output is presented in Figure 4‑1. From Figure 4‑1, the model is not a good fit with the R-Squared value of 0.386, indicating it can only explain 38.6 % of the variation in the data.

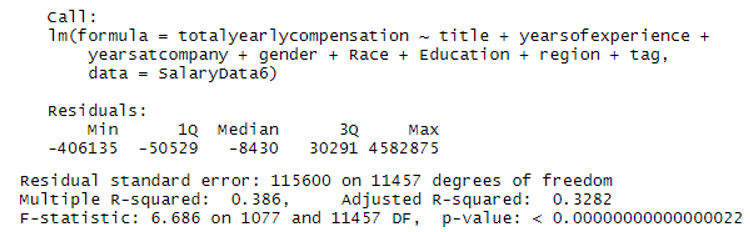


Figure ‑: Multi-Regression Model 1 - Output

### Model 1 – Residuals

The regression model 1 model residual plots and the variable residual plots are presented in Figure 4‑2 to Figure 4‑4. From the residuals, it can be concluded that the model is not a good fit and we can see that there are many exceptions.

Interpretation of the residuals:

**Top left:** The residual across the fitted values shows an asymmetrical distribution, with substantial number points on the above y = 0 line, which indicates the prediction is biased and much more salaries were predicted lower than actual.

**Top right:** the normal Q-Q plot shows is diverging towards the end tail on the higher theoretical quantile side (>2), indicating the deviation from a normal distribution for the errors.

**Bottom left:** The following scale-location plot indicates the “homoscedasticity” or validity of the ‘assumption of equal variance’ of the residuals. The red mean line deviates sharply from horizontal toward higher salary side, indicating the equivalence of variances assumption could be violated.

**Bottom Right:** For model1, a couple of points lie across/ at the Cook’s distance reference line.

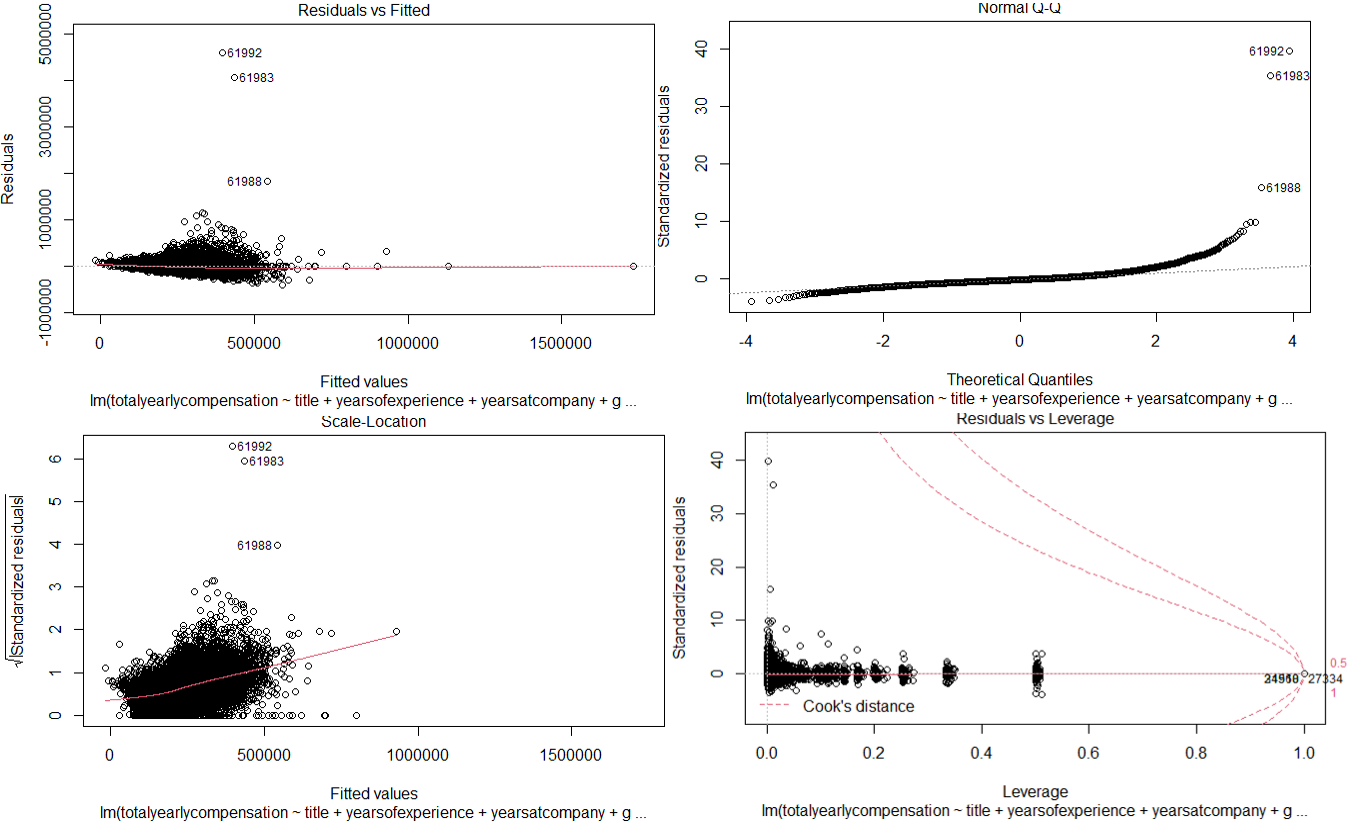


Figure ‑: Multi-Regression Model 1 – Model Residuals

The individual variable residuals are also densely distributed towards the positive residuals side, indicating a biased prediction ( towards lower salary) from the model.

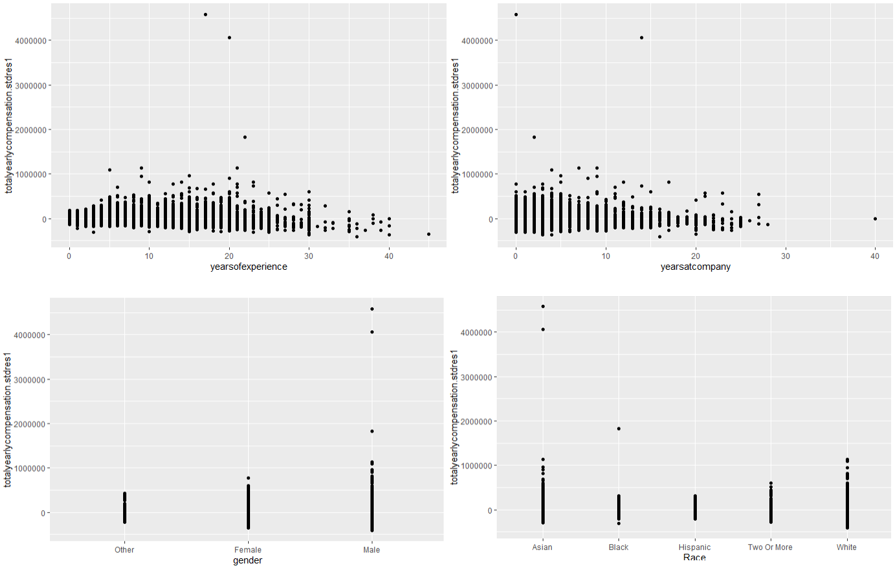


Figure ‑: Multi-Regression Model 1 – Variable Residuals

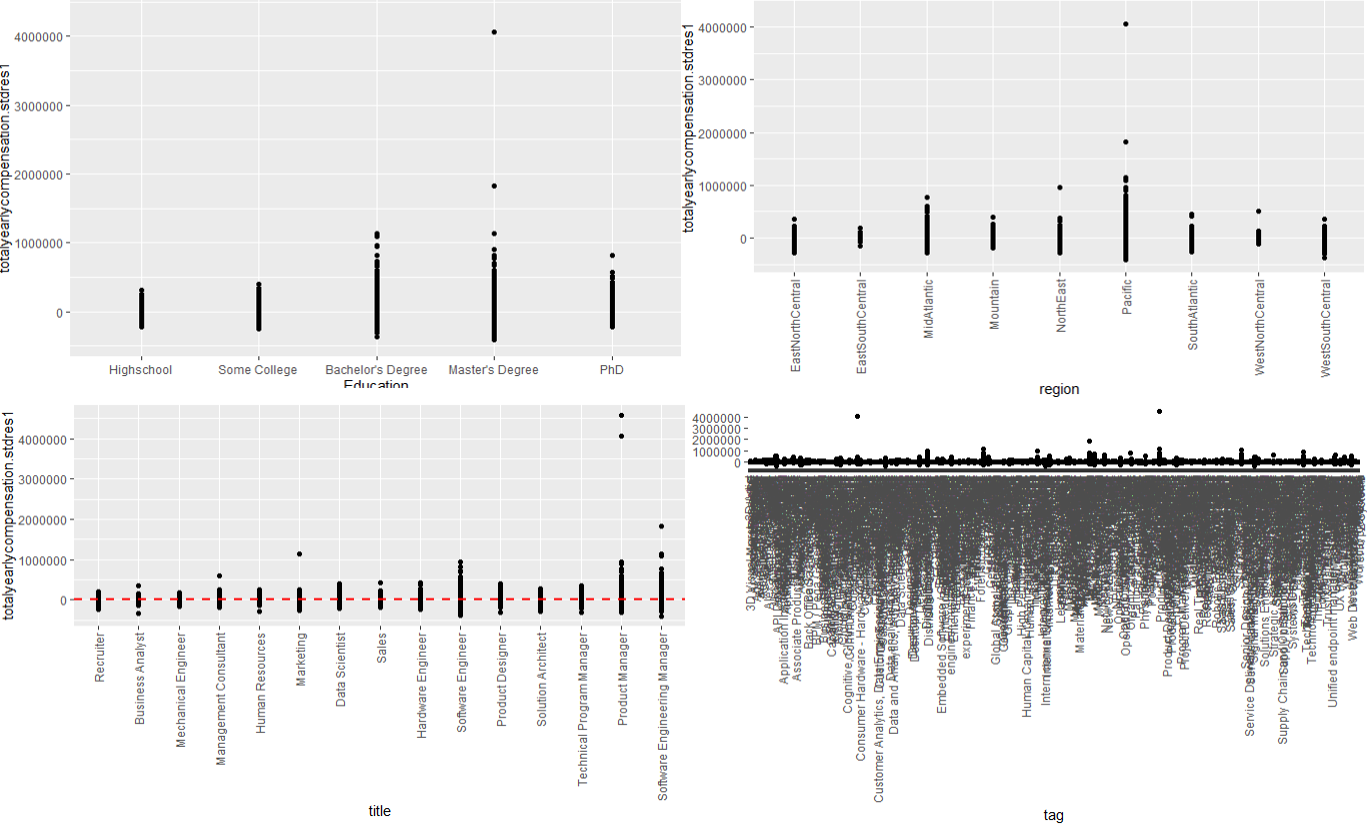


Figure ‑: Multi-Regression Model 1 – Variable Residuals

### Model 2 – Output

The regression model 2 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, and the output is presented in Figure 4‑5. From Figure 4‑5, the model does not seem to be a very good fit with an R-Squared value of 0.48 (Adj R –Squared 0.475), however, it is better than model 1 in explaining the variation of the data. This model explains 48% of the variation in the data. The F-stat value was 88.99 (p < 0.00000000000000022), indicating the model itself is significant.

The numerical variables in model 2, yearsofexperiece and yearsatcompany were highly significant (p< 0.001) and multiple subcategories for all the categorical variables in the model were found to be statistically significant (p< 0.05). For example, in the ‘region’ category, the pacific and Mid-Atlantic were Statistically highly significant (p< 0.001).

Since all the variables were found statistically significant and the R-squared value was not very high, we did not run the model with the reduced number of variables. However, as seen from section 3.6 (univariate analysis histograms), the histogram of log(totalcompensation) is much more normally distributed, so next (section 4.1.5), we ran the linear regression using the log of totalcompensation variable.

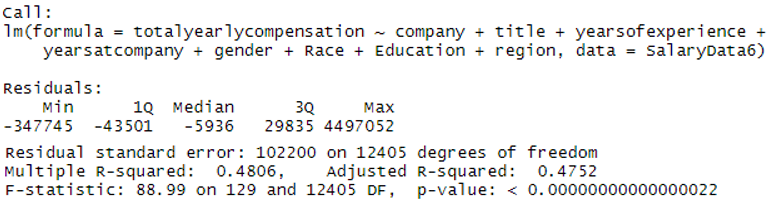


Figure ‑: Multi-Regression Model 2 - Output

### Model 2 – Residuals

The regression model 2 model residual plots and the variable residual plots are presented in Figure 4‑6 to Figure 4‑8. From the residuals, it can be concluded that the model is a fair fit.

Interpretation of the residuals:

**Top left:** The residual across the fitted values shows a slightly asymmetrical distribution, with more points on above y = 0, which indicates the prediction is biased and more salaries were predicted lower than actual.

**Top right:** the normal Q-Q plot shows is diverging towards the end tail on the higher theoretical quantile side (>2), indicating the deviation from the normal distribution.

**Bottom left:** The following scale-location plot indicates the “homoscedasticity” or validity of the ‘assumption of equal variance’ of the residuals. The red mean line deviates from horizontal toward higher salary side. No drastic deviation was observed, it could be inferred that the assumption of equal variance is not violated.

**Bottom Right:** For model2, no influential points lie across the Cook’s distance. There are a couple of some observations (61992, 61983) that are away from the residual cluster; however, they are still well below Cook’s distance limit (dashed red line).

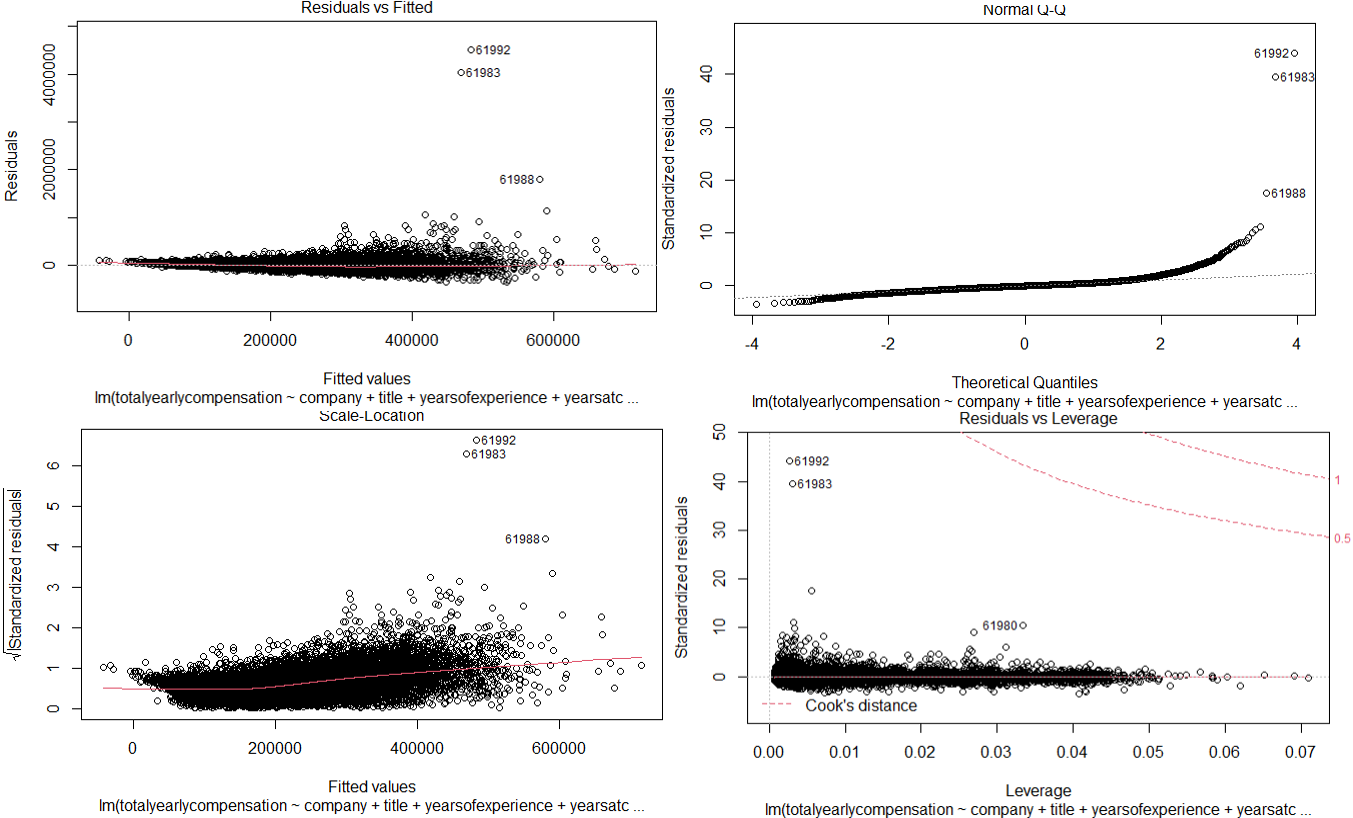


Figure ‑: Multi-Regression Model 2 – Model Residuals

**Residuals of individual variables**

The residuals for the variables as shown below are denser and more spread (distant) on the positive side from the y = 0 axis (shown by the dashed lines), which indicates that the model predicts salaries lower than the actual salaries.

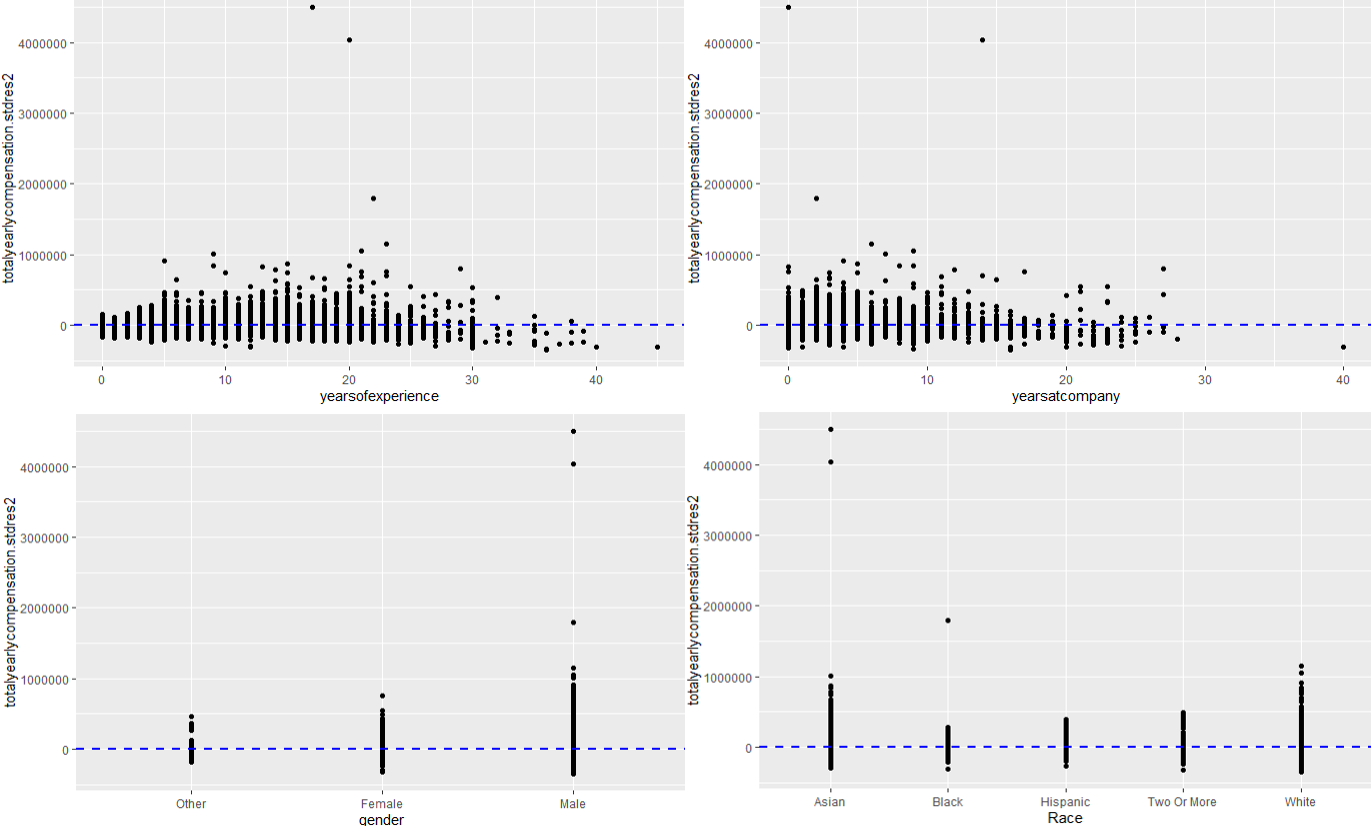


Figure ‑: Multi-Regression Model 2 – Variable Residuals

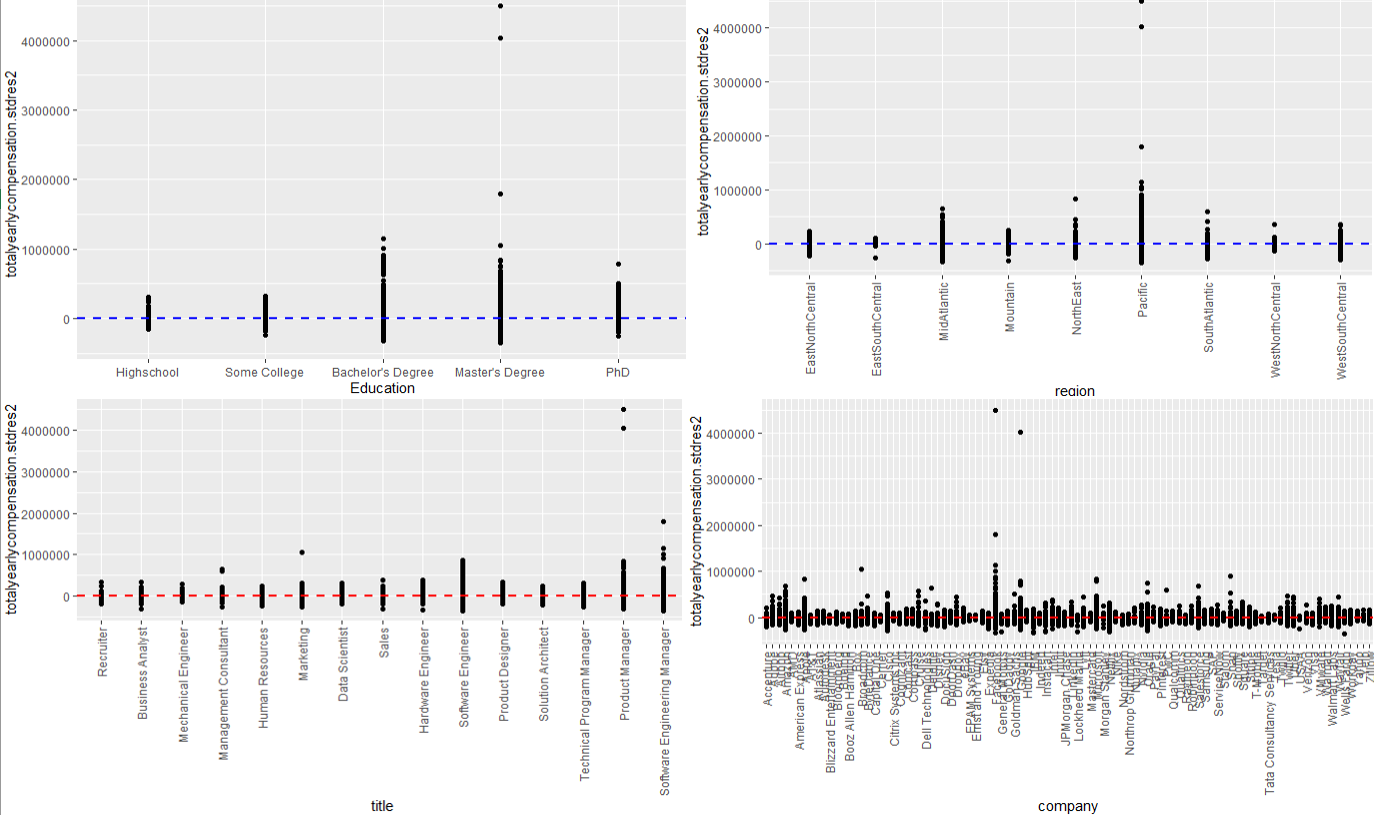


Figure ‑: Multi-Regression Model 2 – Variable Residuals

### Model 3 (Log Model) – Output

The regression model 3 was executed with log value of y variable and the following x variables in consideration; title, yearsofexperience, yearsatcompany, company, gender, Race, Education, region, and the output is presented in Figure 4‑9. The model is a better fit with an R-Squared value of 0.681 (adj R Squared = 0.678). This means it explains 68% of the variance in the data. The F-stat value was 205.6 (p < 0.00000000000000022). The higher R squared and F-statistics value of model 3 as compared to model 1 and model 2 indicates, it is better in explaining the variation of the dataset with the model variables than models 1 and 2.

For model3 as well, the numerical variables yearsofexperiece and yearsatcompany were highly statistically significant (p< 0.001), and multiple subcategories for all the categorical variables in the model were found to be statistically significant (p< 0.05). For example, all the categories in the job ‘title’ variable were statistically highly significant (p< 0.001).

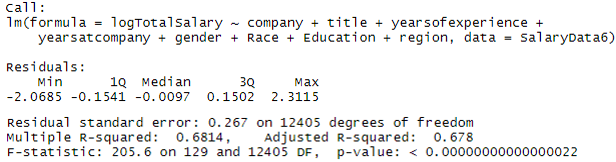


Figure ‑: Multi-Regression Model 3 - Output

### Model 3 – Residuals

The regression model 3 model residual plots and the variable residual plots are presented in Figure 4‑10 to Figure 4‑12. From the residuals, it can be concluded that the model is a better fit. The model is better than models 1 & 2 and is the most parsimonious of all.

Interpretation of the residuals:

**Top left: The** residual across the fitted values shows a symmetrical distribution. The distribution is not curved, it has equally spread residuals around the horizontal line without any noticeable distinct pattern, indicating that it is not a non-linear relationship.

**Top right:** the normal Q-Q plot the normal Q-Q plot shows a fairly good alignment to the dotted straight line in the quantile region (-2,2) showing closeness to a normal distribution of the errors in that region, the plot diverges towards the end tail on both sides.

**Bottom left:** The following scale-location plot indicates the “homoscedasticity” or validity of the ‘assumption of equal variance’ of the residuals. The red mean line deviates from horizontal toward higher salary side. No drastic deviation was observed, it could be inferred that the assumption of equal variance is not violated.

**Bottom Right:** For model3, no influential points lie across the Cook’s distance.

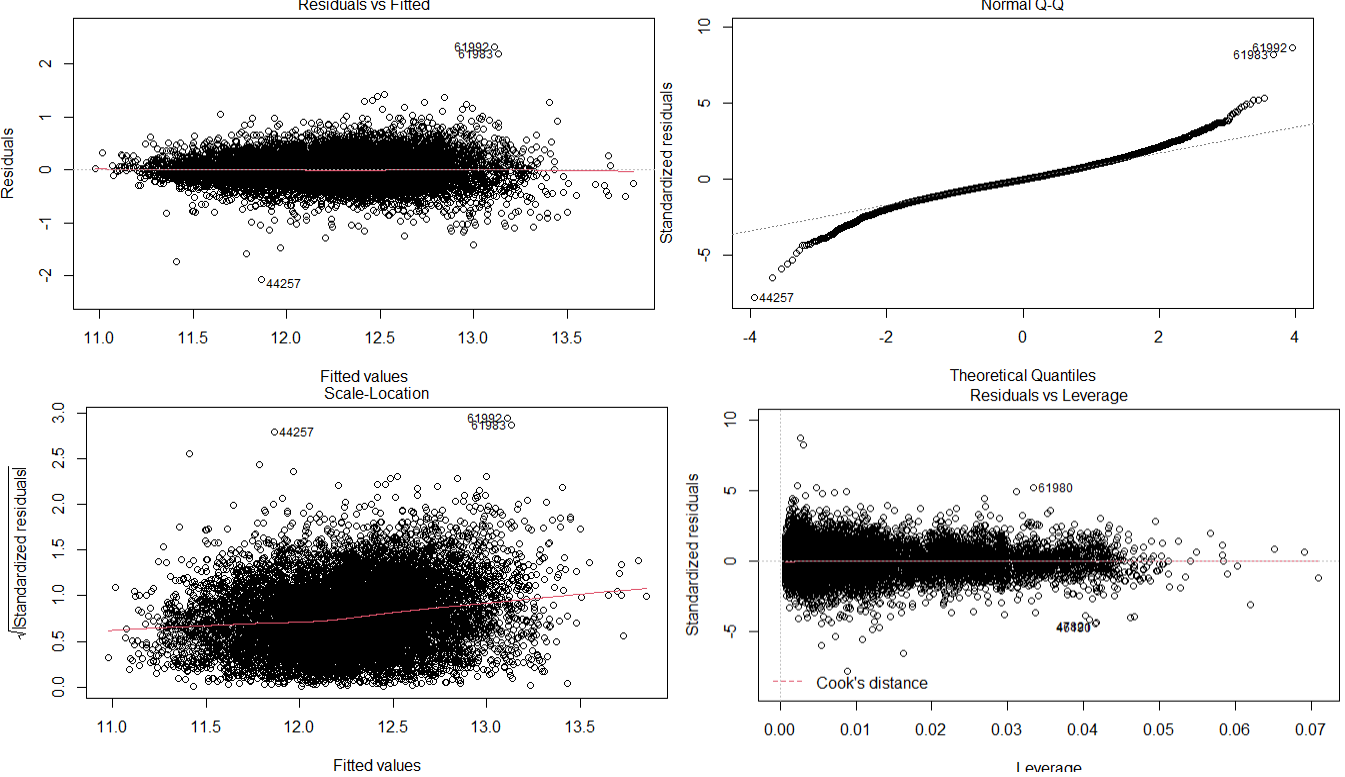


Figure ‑: Multi-Regression Model 3 – Model Residuals

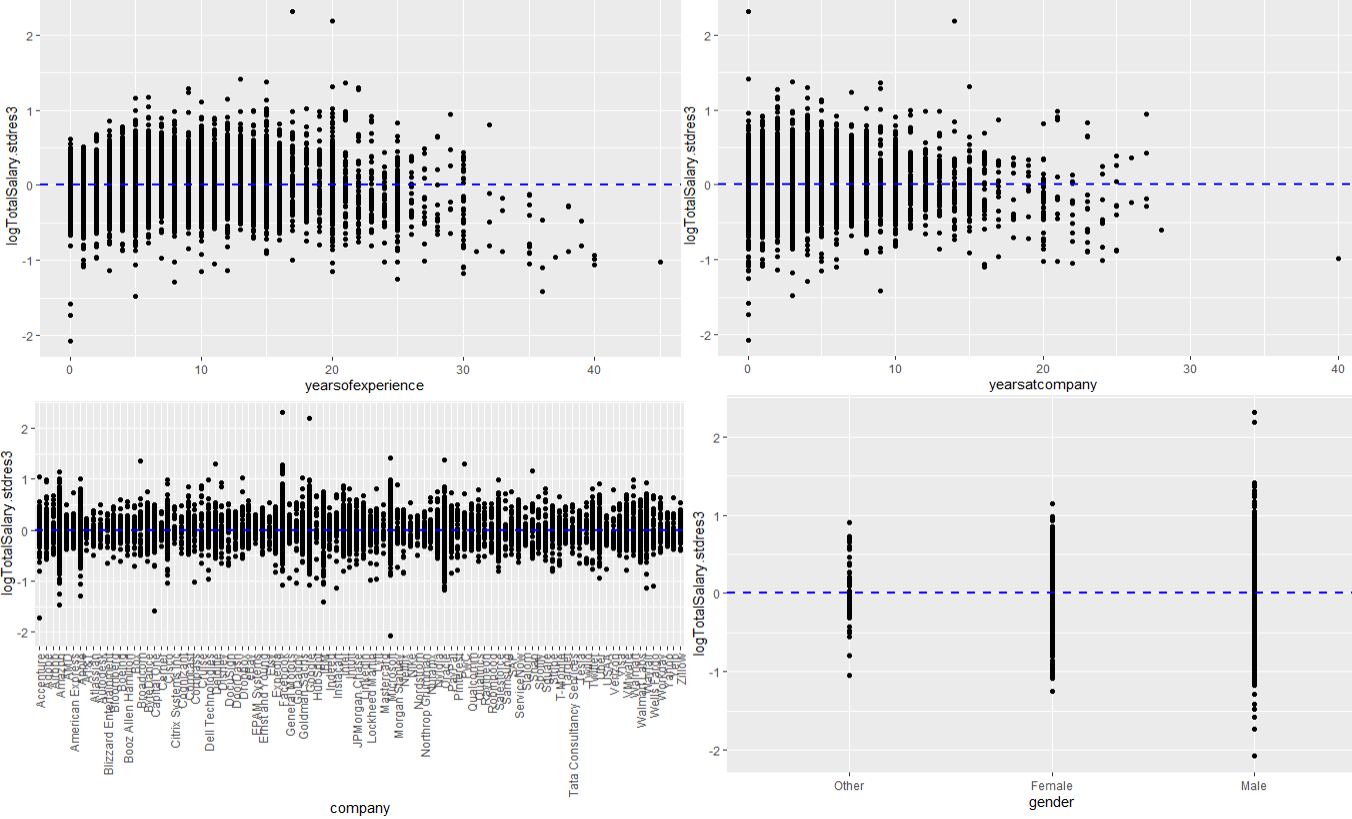


Figure ‑: Multi-Regression Model 3 – Variable Residuals

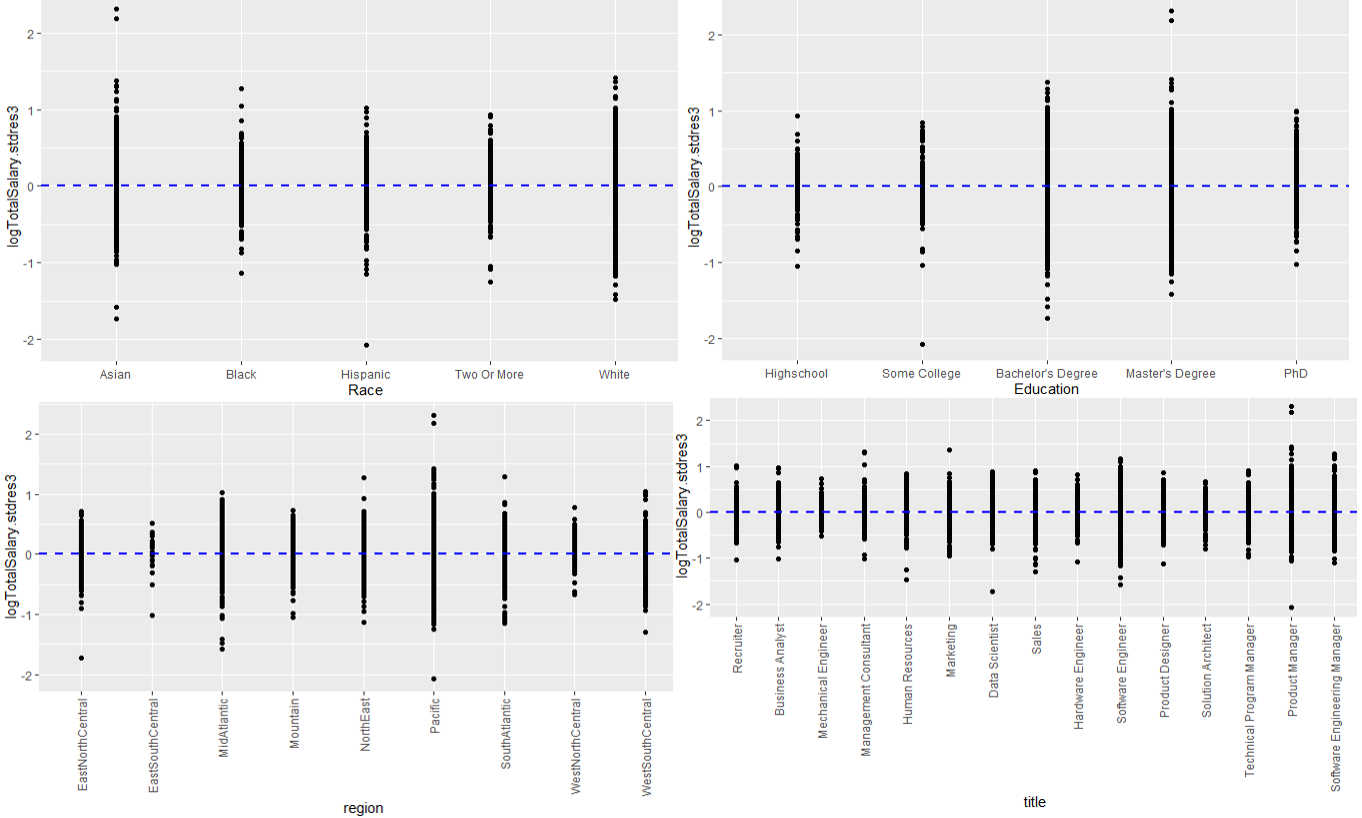


Figure ‑: Multi-Regression Model 3 – Variable Residuals

### Test for collinearity (test with Variance Inflation Factor, VIF)

Collinear variables inflate the model, because they are likely are explaining the same variance in the model and can result in an apparent better fit, rather than the real fit. Also, since collinear variables compete with each other sometimes, neither of them shows statistical significance in the model, even though they might be. We used the ‘VIF’ function in the ‘car’ package for this test.

Models 2 and 3 were checked for collinearity using VIF regression (models 2 and 3 have the same independent variables). A GVIF value (proportional change of the standard error and confidence interval of their coefficients due to the level of collinearity) of greater than 5, would indicate collinearity between the variables. The variables ‘company ‘and ‘region’ had a very high GVIF value, much greater than 5, therefore we can interpret that there is collinearity of these variables with other variables. An accurate contribution of these variables in the model is hence difficult to assess. It is also possible that the two variables ‘company’ and ‘region’ are highly correlated with each other. One possibility could be that a large number of software/ IT companies are located in the pacific region and also has higher salaries than several other regions.[6]. The results of the collinearity test are presented in Table 4‑1.

**## GVIF Df GVIF^(1/(2\*Df))**  
 ## company 44.449976 95 1.020171  
 ## title 4.724101 14 1.057019  
 ## yearsofexperience 1.765986 1 1.328904  
 ## yearsatcompany 1.629641 1 1.276574  
 ## gender 1.154667 2 1.036607  
 ## Race 1.375968 4 1.040701  
 ## Education 1.412016 4 1.044071  
 ## region 11.210974 8 1.163062

Table 4‑1: Collinearity - VIF

## Random Forest Model

For random forest three models are run with different combinations of dependent variables. The output and the prediction/variable significance plots for all three models are presented in the sections below.

### Model 1 – Output

The random forest model 1 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, tag, and the output is presented in Figure 4‑13. From Figure 4‑13 the model is not a good fit with a % variance explained value of 33.39.

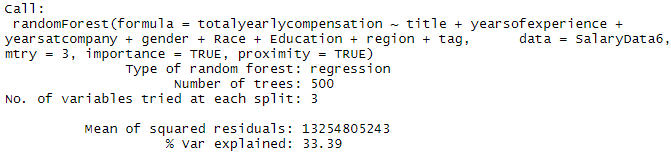


Figure ‑: Random Forest Model 1 - Output

### Model 1 – Variable Significance / Predicted Plots

The random forest model 1 variable significance and the predicted plot are presented in Figure 4‑2 to Figure 4‑4. From the predicted vs actual plot, the model is not a good fit.

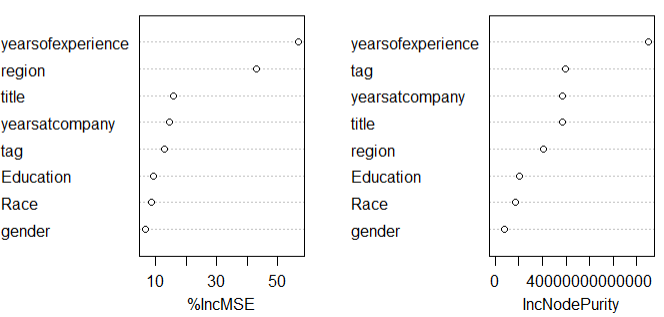


Figure ‑: Random Forest Model 1 – VariableSignificance

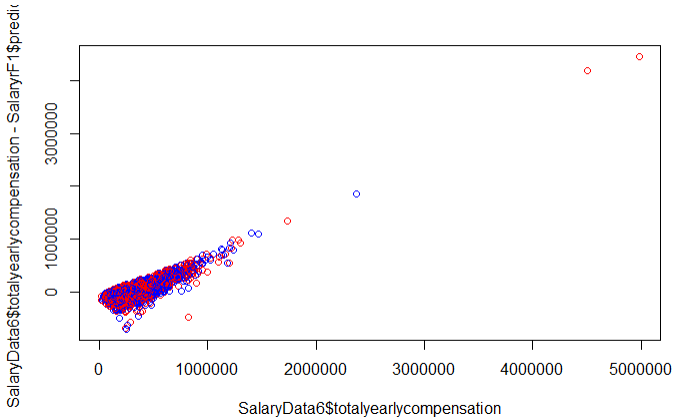


Figure ‑: Random Forest Model 1 – Predicted vs Actual Plot

### Model 2 – Output

The random forest model 2 was executed with the following variables in consideration; title, yearsofexperience, yearsatcompany, gender, Race, Education, region, and the output is presented in Figure 4‑16. From Figure 4‑16, the model is a better fit compared to model 1 with a % variance explained value of 46.09.

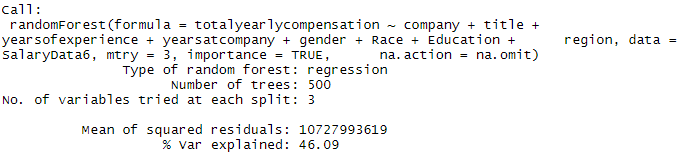


Figure ‑: Random Forest Model 2 - Output

### Model 2 – Variable Significance / Predicted Plots

The random forest model 2 variable significance and the predicted plot are presented in Figure 4‑17 to Figure 4‑18. From the predicted vs actual plot, the model is ok to fit, and we can see that the predictions are all over the place.

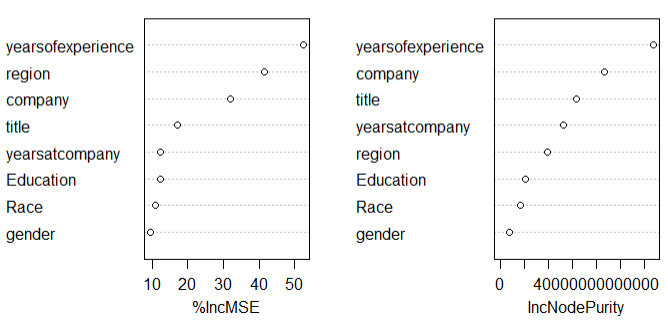


Figure ‑: Random Forest Model 2 – Variable Significance

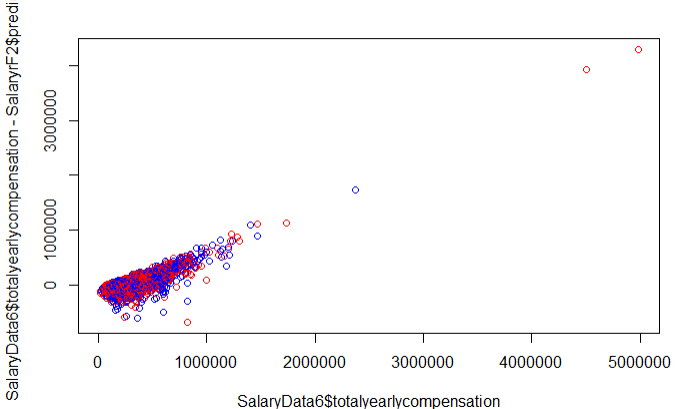


Figure ‑: Random Forest Model 2 – Predicted vs Actual Plot

### Model 3 (Log) – Output

The random forest model 3 was executed by taking the log of y variable and the following x variables into consideration; title, yearsofexperience, yearsatcompany, company, gender, Race, Education, region, and the output is presented in Figure 4‑19. From Figure 4‑19, the model is a good fit with a % variance explained value of 62.99.

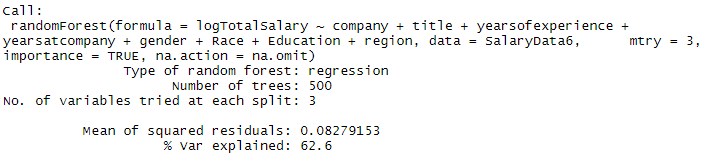


Figure ‑: Random Forest Model 3 - Output

### Model 3 – Variable Significance / Predicted Plots

The random forest model 3 variable significance and the predicted plot are presented in Figure 4‑20 to Figure 4‑21. From the predicted vs actual plot, the model is an ok fit compared to model 2 and we can see that the predictions are all over the place.

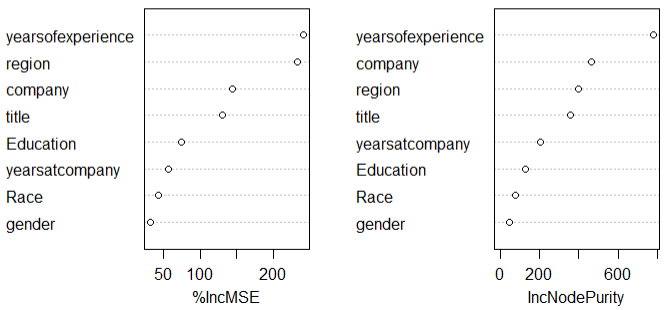


Figure ‑: Random Forest Model 3 – Variable Significance

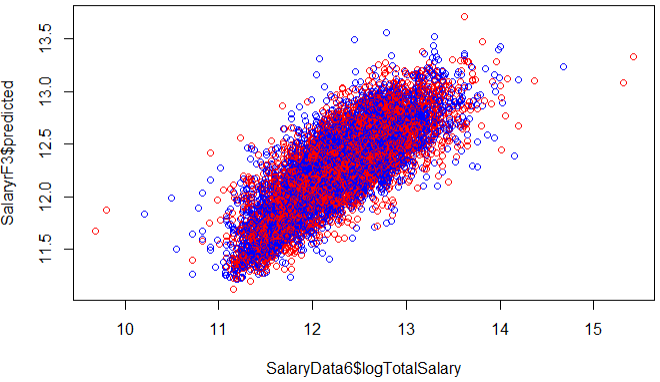


Figure ‑: Random Forest Model 3 – Predicted vs Actual Plot

# RESULTS

The results from the three models and the two different regression algorithms are summarized and presented in Table 5‑1 below. From the result, the most parsimonious model is model 3, a linear regression model with an R squared value of 0.68. (Explains 68% of the variance in the data)

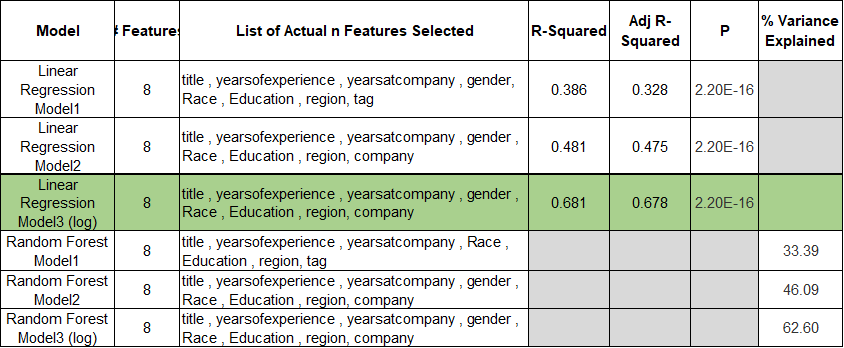


Table 5‑1: Model Results

# LIMITATIONS

One of the key limitations we had was that we were predicting numerical value dependent on many categorical variables and as such running the code took a lot of time. There was also a strong correlation between bonus, base salary, stockgrants, and the total yearly compensation which has introduced bias into our model.

Another limitation was quite a bit of null values and though the data was acquired from levels. fyi which is mostly accurate, we had a few entries where people reported quite conflicting salaries due to which we had few outliers and exceptions in the model.

# CONCLUSION/RECOMMENDATIONS

To conclude we say that the top FAANG / MAMAA companies tend to be on the higher spectrum salaries, and we identified that the top 20 companies tend to pay more salaries. From Figure 7‑1 & Figure 7‑2, we recommend applying to the top 25 companies which pay higher salaries than others. We also recommend checking the years of experience to understand what salary they can negotiate for.

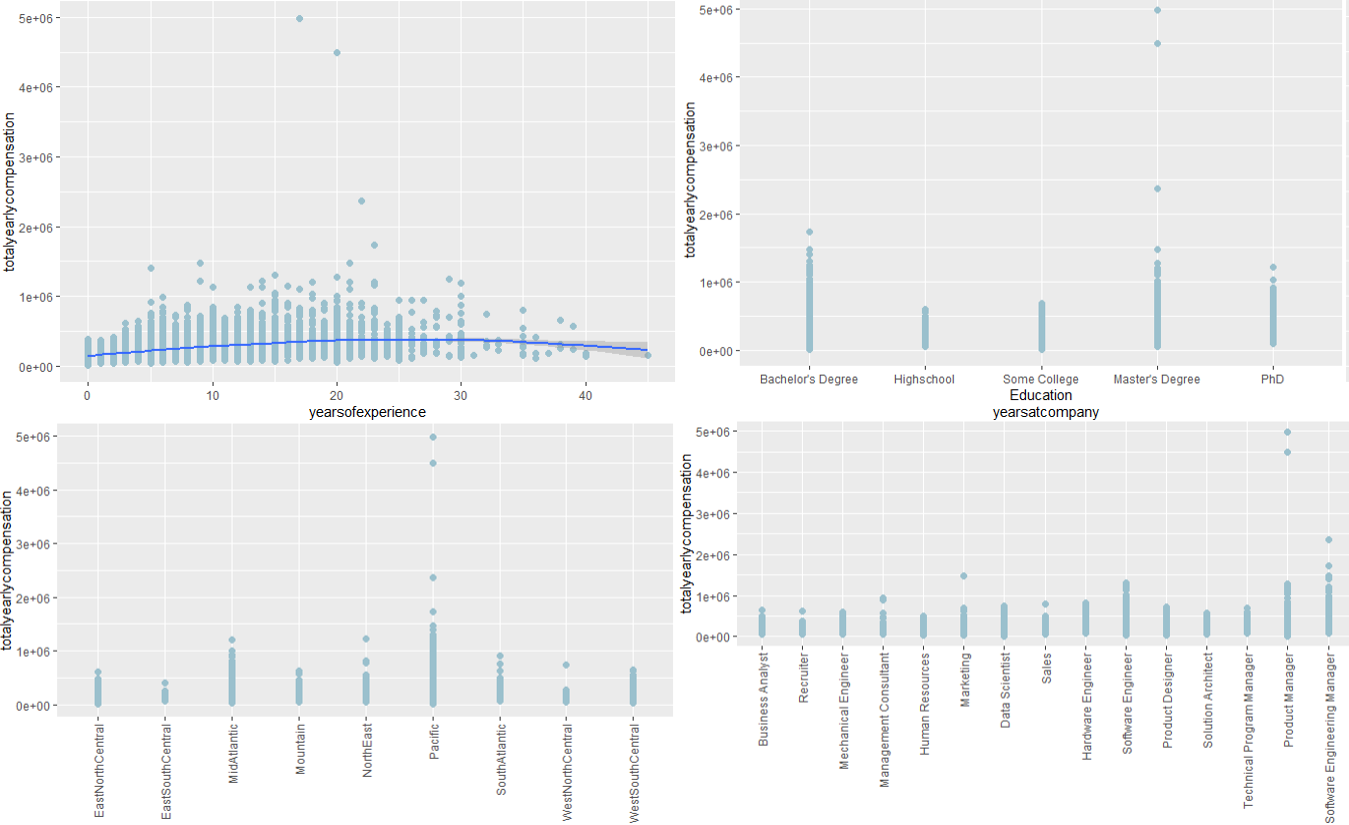


Figure ‑: Yearlcompensations Vs Deciding Factors

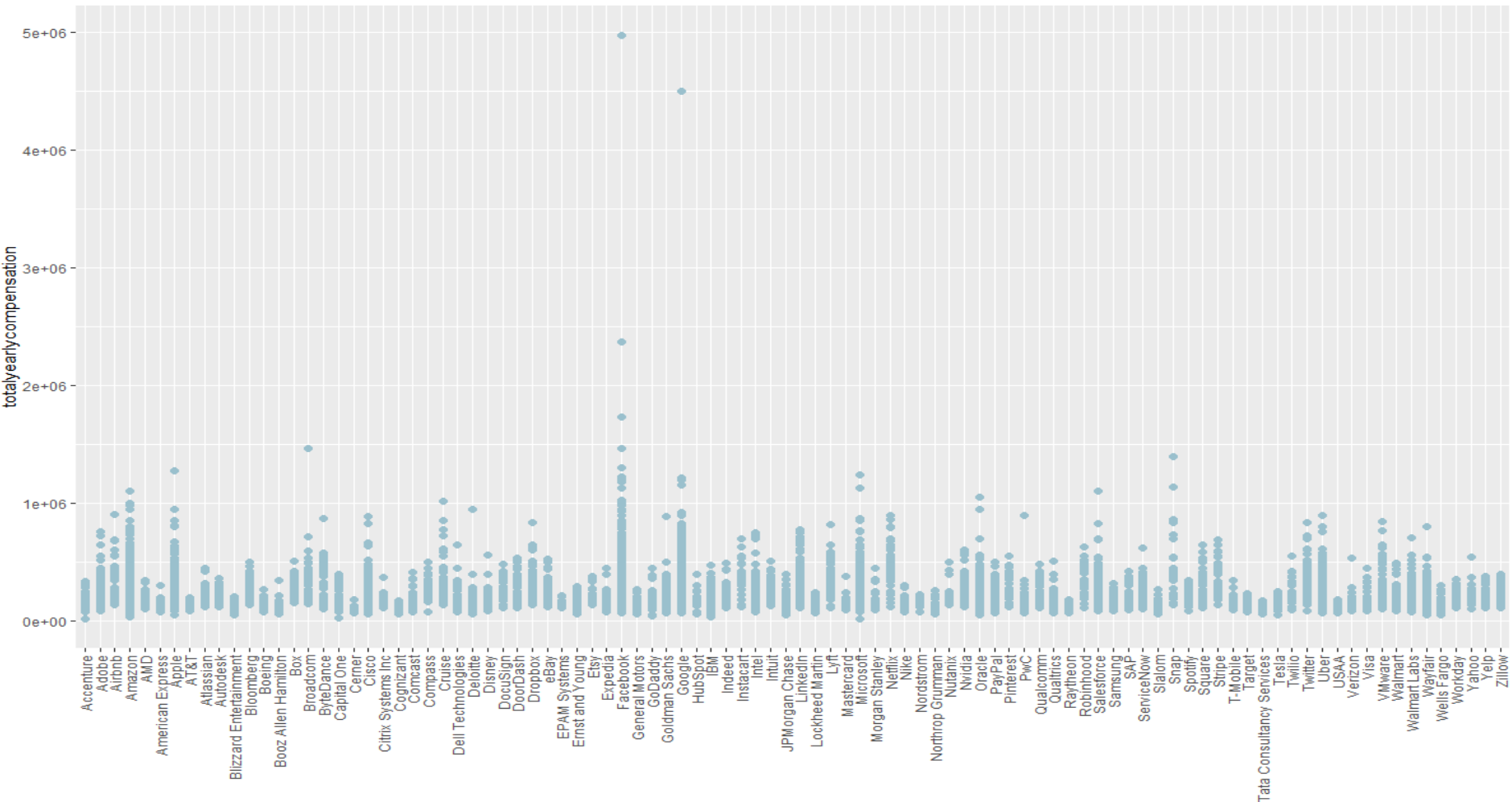


Figure ‑: Yearlcompensations Vs Top Companies

# PREDICTIONS – MOST PARSIMONIOUS MODEL

From Table 5‑1 the most parsimonious model is established as model 3 with linear regression. The prediction algorithm is run for this model and the predicted and actual values are plotted for each variable. The plots are presented in Figure 8‑1 to Figure 8‑4.

It can be seen from the following plots that the model is a good fit for the average salaries and does not predict the higher or lower ranges of the salary well. One reason could be that the number of observations (ref sec 3.6, histogram) in the lower/ higher end of the range is lower in number.

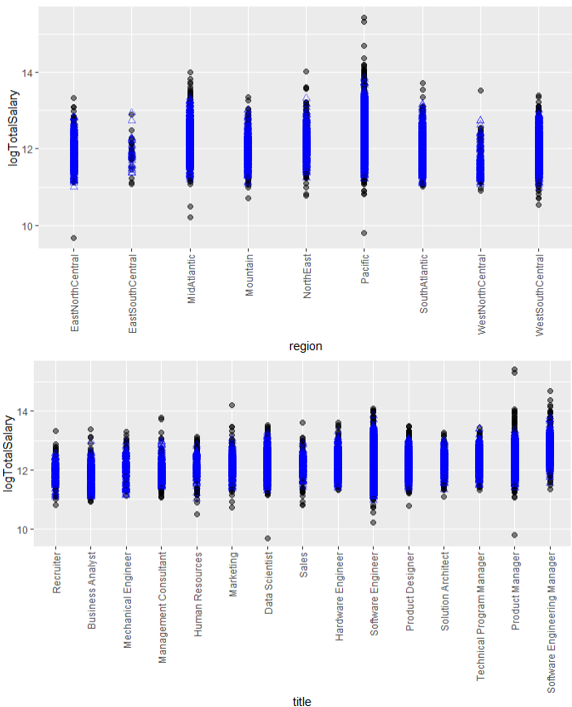


Figure 8‑1: Predicted Vs Actual – Parsimonious Model

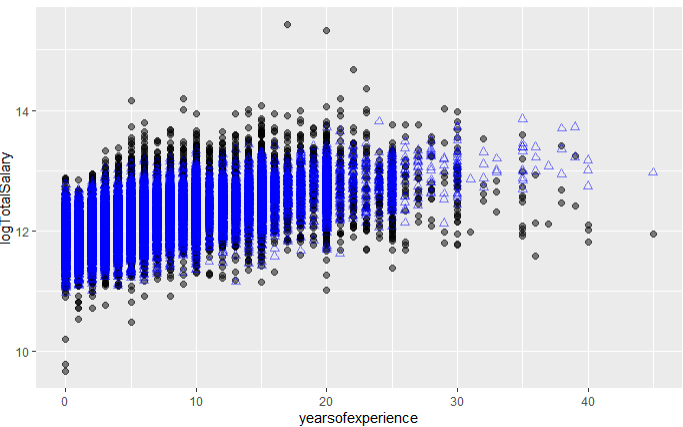
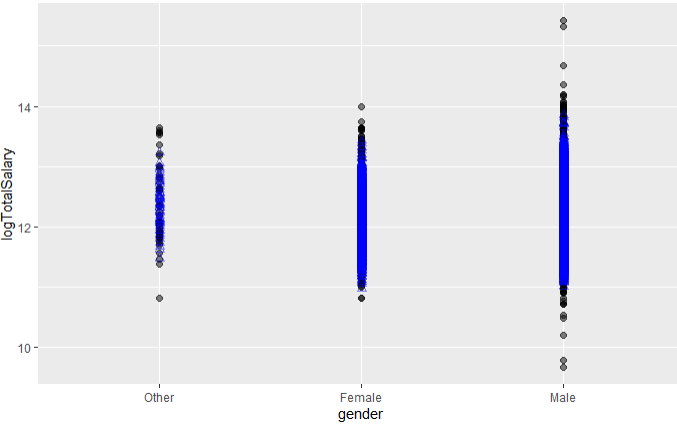




Figure 8‑2: Predicted Vs Actual – Parsimonious Model – Cont…



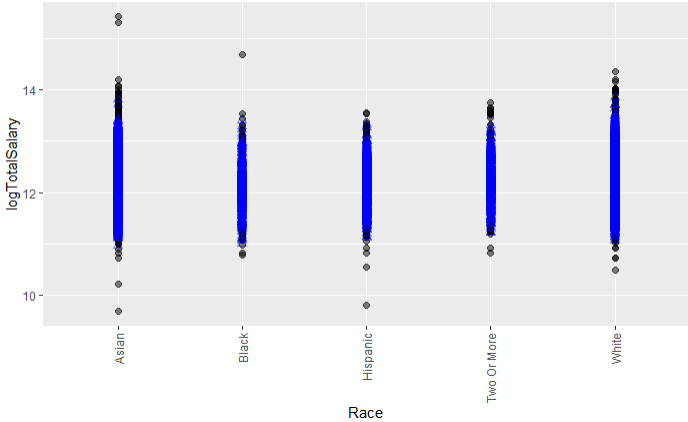
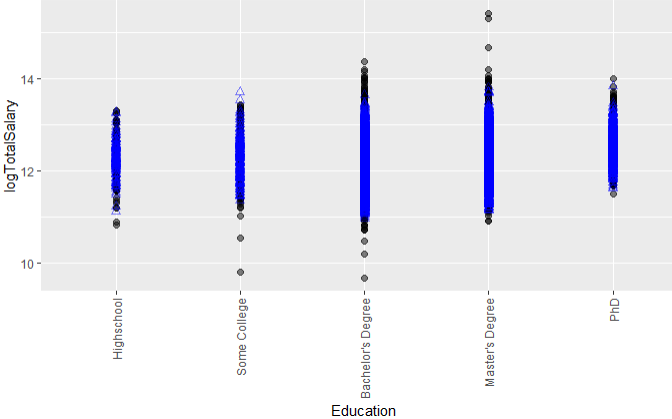


Figure 8‑3: Predicted Vs Actual – Parsimonious Model – Cont…



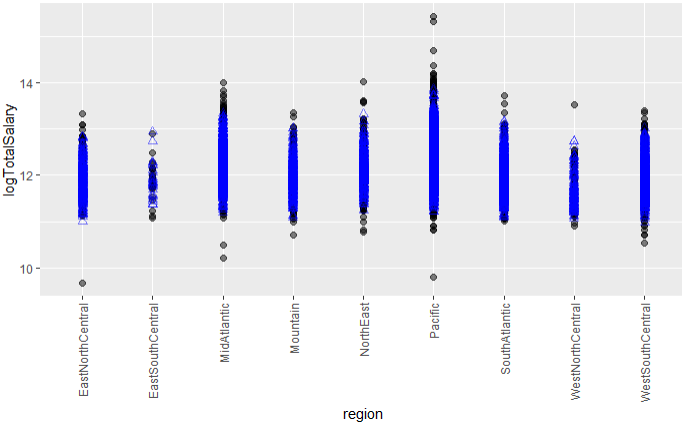


Figure 8‑4: Predicted Vs Actual – Parsimonious Model – Cont…

# REFERENCES

1. <https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries>
2. <https://www.levels.fyi/comp.html?track=Software%20Engineer&region=819>
3. <https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c>
4. <https://view.genial.ly/608626ff91d129103155b087/interactive-content-311-crisp-dm-phases-and-tasks-ver2>
5. <https://www.visualcapitalist.com/comparing-wealth-u-s-geographic-regions-time/>
6. <https://www.business2community.com/brandviews/upwork/why-silicon-valley-techies-are-rushing-to-the-pacific-northwest-02076366>