Analysis of factors influencing annual wage

Tanya Sharma
Statistics 202A, Fall'2022

Introduction

Wages are an essential component of an individual's life. It has the ability to impact not just the economic aspects like purchasing power and standard of living for any person but in many cases may also be a proxy for physical and mental health. Having worked as a full-time employee myself, I have witnessed the curiosity among people to better comprehend the factors that influence their monthly pay. Therefore, in this paper, I have investigated various factors that have an effect on the annual wages of workers. Although the determination of wages is a complex phenomenon and depends on a variety of features, to name a few, market demand and supply, geographical location, bargaining power of both employers and employees etc. In this analysis my focus has been to study how a person's age or gender, level of education and job specific features like job title, type of department, and seniority level impact wages. In the next few sections I will discuss about the dataset I utilized for this analysis, modeling approach, results, and its limitations.

About the data

The dataset that I used for this project is a Kaggle dataset. It is collected from Glassdoor website. Glassdoor is a digital platform that gathers information from current and former employees about salary, reviews, job openings, interview experience etc. The dataset comprised of 1000 rows (observations) and 9 columns (variables). It consisted of the following variables: Age, Job title, Gender, Education, Department, Seniority, Performance score, bonus, and base pay. There are no missing values in the dataset. There were 6 categorical variables namely Job title, gender, education, department and seniority and the rest are numerical variables.

In order to conduct any analysis, it is imperative to have a clear understanding of our independent and response variables. In this study the response variable is base pay. Although, there are two variables that determine total wages of workers – base pay and bonus. But for the purpose of this analysis, only base pay is considered. We do this because we are interested in studying the impact of our independent variables on the wage level decided between the employee and employer beforehand. Bonuses are usually based on employee's performance. The employer after evaluating the employee for an entire year decides on the bonus amount. My assumption is that the dynamics may change drastically when the employee and employer have known each other for a certain amount of time and the relationship thus formed may overweigh the actual relationship between our independent and response variables. Therefore, I decided against factoring in bonus, or any performance-based metric in the scope of this study.

I have studied the impact of 7 variables on base pay in this study. The population in this dataset is aged between 18 years and 65 years, 41 being the median age. Gender has 2 categories -Males and Females. Proportion of females is 46.8% while that of males is 53.2%. Education is divided into 4 levels – High school, College, Masters, and PhD. The distribution of data points among these categories is fairly uniform. There are 6 categories for the Job title - Data Scientist, Driver, Financial Analyst, Graphic Designer, IT, Manager, Marketing Associate, Sales Associate, Software Engineer, Warehouse Associate. Department has 5 categories - Admin, Engineering, Management, Sales, Operations. Seniority and performance evaluation score each has 5 levels with 1 being the lowest and 5 being the highest. The distribution of data between various categories of these variables can be seen in the bar graphs (Figure 1) attached in the appendix below.

Another important aspect that we need to discuss with respect to categorical variables is the encoding technique. For modelling purposes, I have performed ordinal encoding on categorical variables expect education level. So, for instance, gender has 2 categories, then they are coded as 0 (Male) and 1 (Female). Similarly, department has 5

categories, so they are coded as 0,1,2,3,4 respectively. For education level, I adopted a slightly different approach. Each education level is converted to number of years of education so for high school its 12 years. Assuming the average length of an undergraduate degree to be 4 years, years of education for college is 16 years. A master's program is usually 2 years so 18 years for a master's degree and 23 years for a PhD.

Exploratory Data Analysis

I performed a detailed exploratory data analysis on all the variables with an objective to build a stronger understanding of the dataset. In this section, I will study the distribution of data and talk about the patterns and associations that I identified in the dataset.

Figure 2A and 2B show kernel density estimates for Base Pay and Age respectively. Base pay appears to follow a normal distribution. Distribution of age resembles uniform distribution i.e. the dataset captures an equal proportion of people in each age group. However, this might not always be a true representation of real-world demography. Next, we consider a series of boxplots to scrutinize the relationship between various categories of categorical variables. Figure 3 shows a comparison of males and female's base pay. Ideally there should not be a difference in the base pay for males and females, however given societal norms one would expect females to earn less than males. This is what the boxplot depicts - a gender pay gap. Figure 4 represents an increase in base pay as education level increases which aligns with what we expected. However, the increase in wages as the level of education increase is not noticeably significant. Figure 5 shows the distribution of base pay for various job titles and we can see how certain job titles like Manager and software engineer are associated with considerably higher base pay than others like marketing associates or drivers. Similarly, we see in figure 6 how certain departments like Management, Sales or Engineering are associated with higher base pay than Operations and Administration. Figure 7 depicts a relationship between seniority level and base pay and as one might expect the relationship is positive. Figure 8 and figure 9 displays a comparison of relationship of performance evaluation score with base pay and bonus respectively. We see a positive relationship between bonus and performance score. This is fair because bonus amount is majorly dependent on an individual's performance. However, no clear trends are observed for the relationship between performance score and base pay.

Gender pay gap is a pressing issue and I wanted to understand the depth with which it existed. As I continued my analysis some interesting trends emerged. Figure 10 and 11, illustrate a concerning reality. At all education levels and across all departments, we can see that the basic pay for females is less than that for males. Figure 12 shows the distribution of base pay for males and females for various job titles. Another troubling trend emerge here. Manager and software engineering jobs are the highest paid. While Marketing associates fall in the lowest income bracket. If we closely look at the distribution of women for these positions, it can be concluded that there are significantly less proportion of women employed in high paying jobs while a considerably higher proportion of women are engaged in low paying jobs.

Model Results

1. Analysis of Variance

I performed a multi-way anova analysis with an objective to study if there is a significant difference in the base pay for different classes of categorical independent variables. The results can be found in the table below.

```
Analysis of Variance Table

Response: base
             Df     Sum Sq    Mean Sq F value    Pr(>F)
gender        1 1.8051e+10 1.8051e+10 41.1479 2.183e-10 ***
jobt          1 1.6576e+09 1.6576e+09  3.7786 0.0521960 .
edu           1 1.6402e+10 1.6402e+10 37.3900 1.390e-09 ***
dept          1 4.9580e+09 4.9580e+09 11.3020 0.0008039 ***
senior        4 1.6344e+11 4.0861e+10 93.1451 < 2.2e-16 ***
performance   1 2.5366e+09 2.5366e+09  5.7824 0.0163702 *
Residuals   990 4.3430e+11 4.3868e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output we can see that all the factors namely gender, job title, education, department, seniority level and performance evaluation score explain a significant amount of variation in the base pay. In the next step we have attempted to quantify the correlation between the independent variables and the response using a linear regression analysis. It then enables us to do wage predictions.

2. Linear Regression

Since it is a regression style data, I performed a linear multivariate regression analysis. Linear regression allows us to study linear relationship between are independent and response variables. The table below displays the regression results.

```
Call:
lm(formula = base ~ age + gender + jobt + edu + dept + senior +
    performance)

Residuals:
   Min     1Q Median     3Q    Max
-39766 -10058  -1346   9344  47584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10740.2     3104.4   3.460 0.000564 ***
age           1019.4       33.3  30.613  < 2e-16 ***
gender       10298.9      962.1  10.704  < 2e-16 ***
jobt           480.0      158.9   3.021 0.002581 **
edu            857.6      132.0   6.499 1.28e-10 ***
dept          1183.6      403.4   2.934 0.003425 **
senior        9434.4      341.7  27.610  < 2e-16 ***
performance   -528.5      335.5  -1.575 0.115592
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15010 on 992 degrees of freedom
Multiple R-squared:  0.6517,   Adjusted R-squared:  0.6492
F-statistic: 265.1 on 7 and 992 DF,  p-value: < 2.2e-16
```

From the above table we can conclude that age, gender, job title, department and seniority level are significant variables since the p value is very small. However, the performance evaluation score is not significant. Also, the coefficients are positive which implies that all the variables are positively correlated with the response variable. An increase in age by 1 year increases the annual base pay on an average by $1019. In this model, age acts as a proxy for years of experience. So higher age indicates more experience and therefore it is associated with higher base pay. It will be interesting to see how the relationship between age and base pay change if we add individual's years of experience to the model. My assumption is that age and years of experience will be highly correlated and will introduce multicollinearity in the model. However, that analysis is currently out of scope. If we look at the coefficient for gender, there is a difference of $10,289 between the annual wages of males and females. It again indicates a gender pay gap prevalent in the society we live. R^2 is a measure of goodness of fit. The R^2 value for the model is 0.6517 which means that 65% of the response variability is explained by the model. The result is reasonable because wages depend on a plethora of factors while we have considered only a subset of those variables in our current analysis. The above results are consistent with the exploratory data analysis that we performed earlier. Our final model is as follows:

Base = 10740.2+1090.4**age*+10298.9**gender*+480**jobt*+857.6**edu*+1183.6**dept*+9434.4**senior*

3. Residual Analysis

According to the assumptions the expected value of errors should be zero. The variance should be constant, and the errors must be normally distributed. If we look at the fitted vs residual plot (Figure 13), we can conclude that the errors appear to be distributed randomly. This implies that some values are greater than zero while others are less than zero, so on an average the expected value seems to be zero. The variance appears to be constant so the assumption of homoskedasticity holds. If we look at the QQ plot, the residuals are distributed along the straight line, so it is safe to conclude that the assumption of normality also hold. The residual vs leverage plot is used to identify any influential points in the dataset. If we look at the residual vs leverage plot for this model, we see clustering happening in the middle of the plot and it is safe to conclude that there are no influential points in our regression model as we don't encounter any points with significantly high leverage.

Limitations

There are certain limitations in the current analysis which leaves scope for future research in this area. Wage determination is a complex procedure and there are several other important features that play a key role in wage prediction and are not discussed in this study. Geographical location is one such factor. An individual's previous pay scale also has a significant influence on the current wage. Number of years of relevant work experience is another

important factor influencing wages. Collecting data on such variables may significantly improve model results. Another critical thing to note here is that for this study, I applied ordinal encoding on categorical variables. It imposes a relationship between the categories even in scenarios where it does not exist. For instance, coding gender categories as 1 for male and 0 for female imposes a natural order on gender. This order does not exist in reality. So, using a different coding scheme for categorical variables may give improved prediction results.

Conclusion

Based on the analysis, our understanding about the factors influencing annual wages has improved. Linear model seems to fit our data well. Some interesting trends emerged but at the same time, data also reflected the deep-rooted biases of our society which are still prevalent. While this study focuses on a subset of factors, there is still a variety of important features which are not a part of this study, and they define the future course of research in this area.

Appendices

Figure 1: Bar Graphs for categorical variables

Figure 1A: Gender

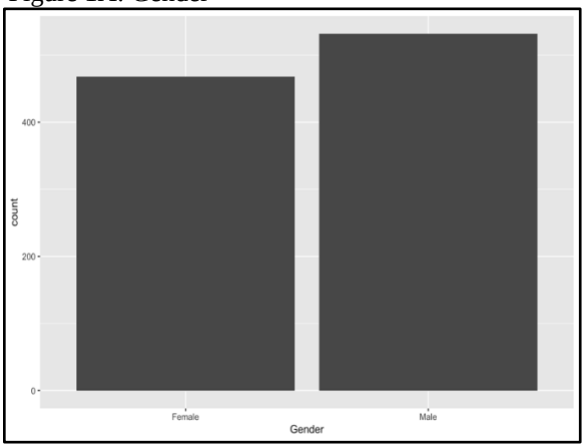


Figure 1B: Education

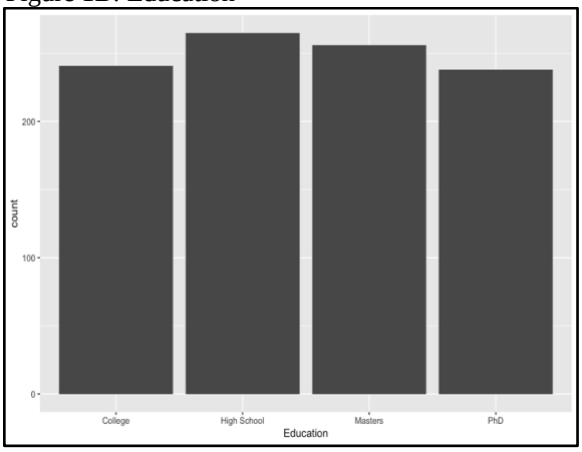


Figure 1C: Department

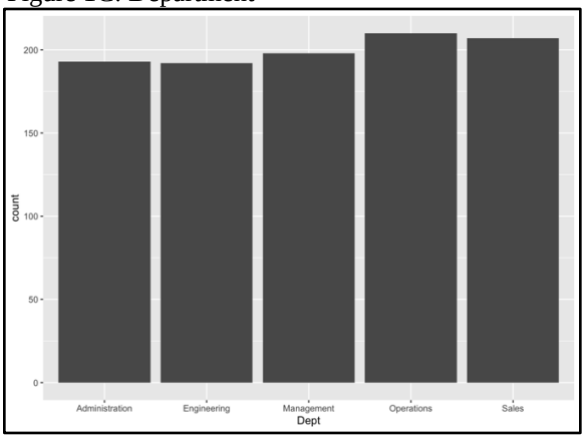


Figure 1D: Seniority Level

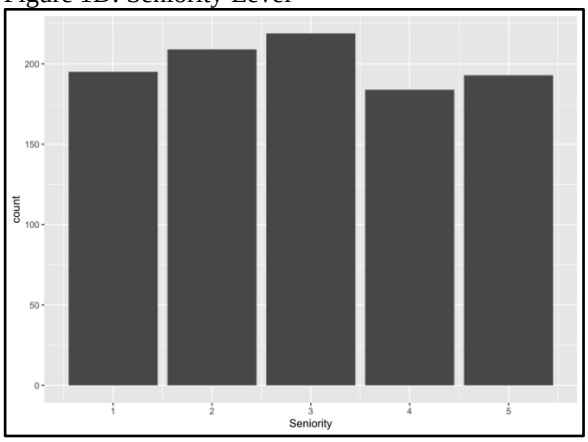


Figure 1E: Job Title

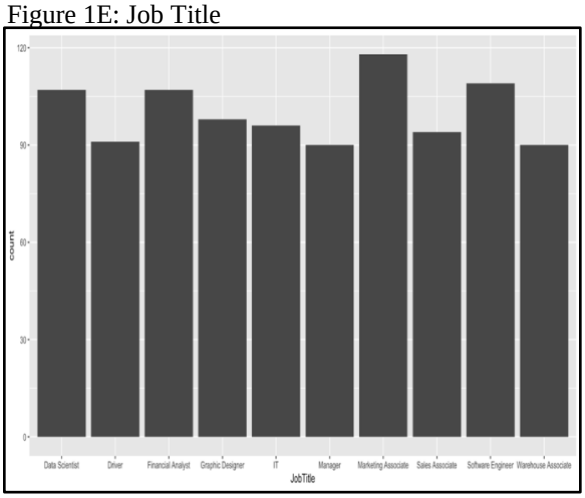


Figure 1F: Performance Evaluation Score

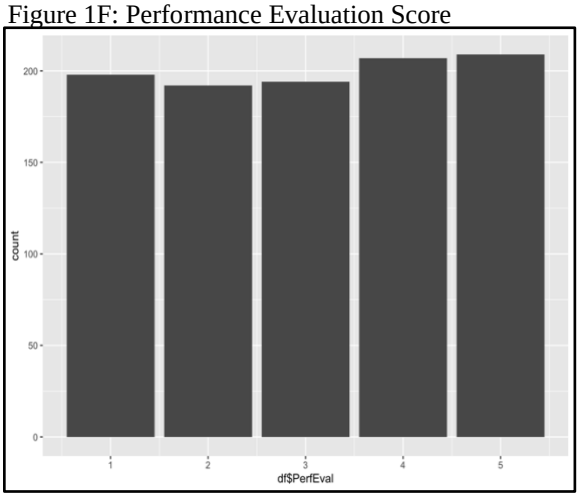


Figure 2: Kernel Density plots

Figure 2A: Base Pay

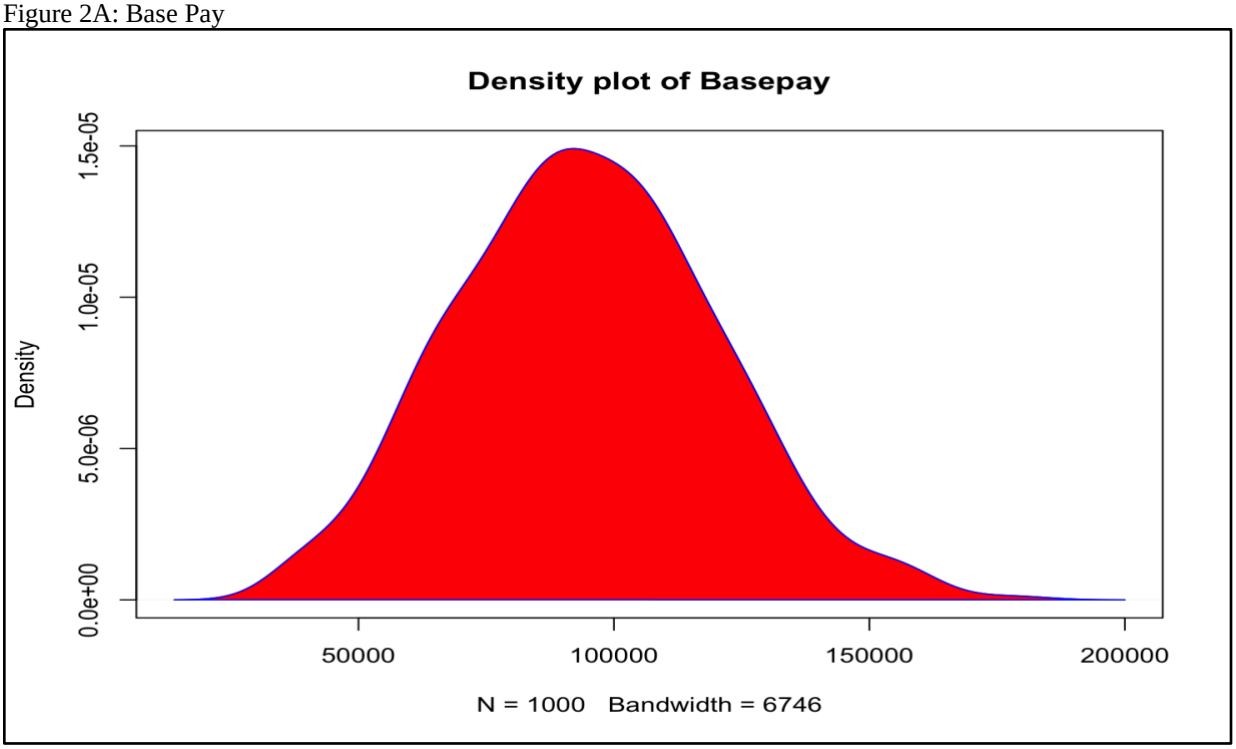


Figure 2B: Age

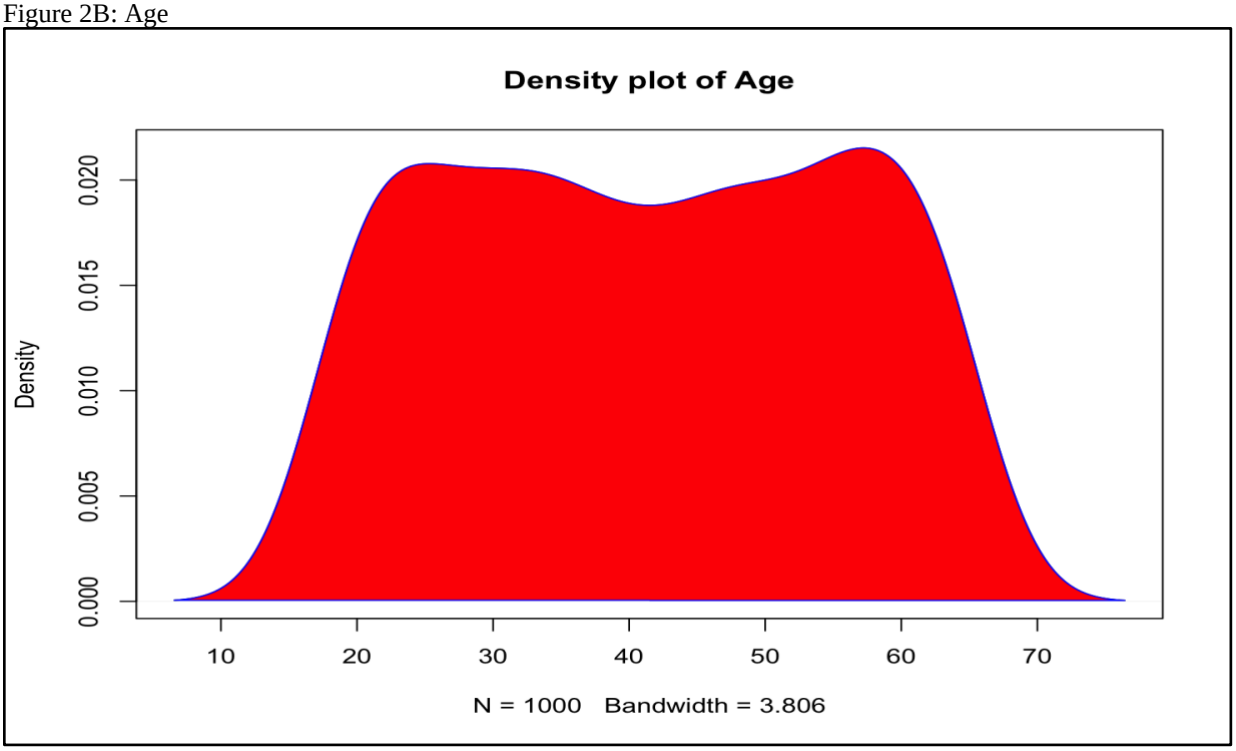


Figure 3: Boxplot for Gender Vs Base Pay

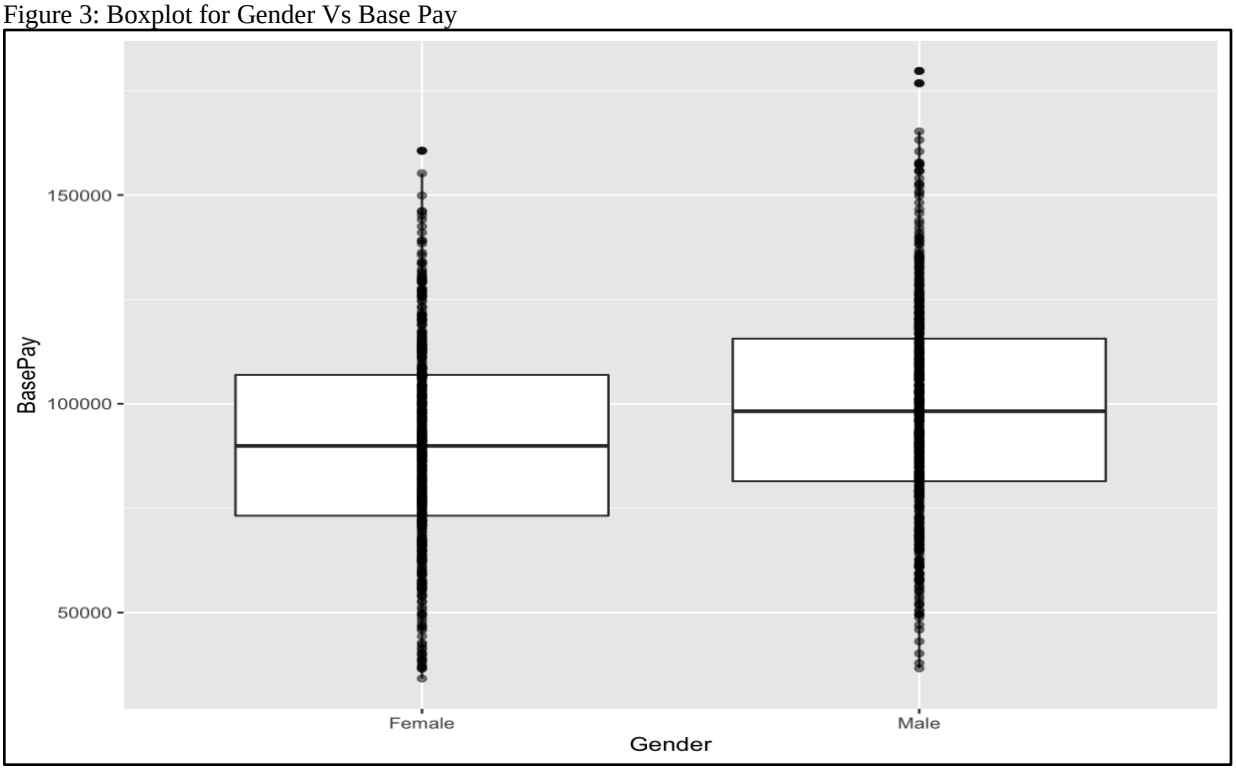


Figure 4: Boxplot for Education Level Vs Base Pay

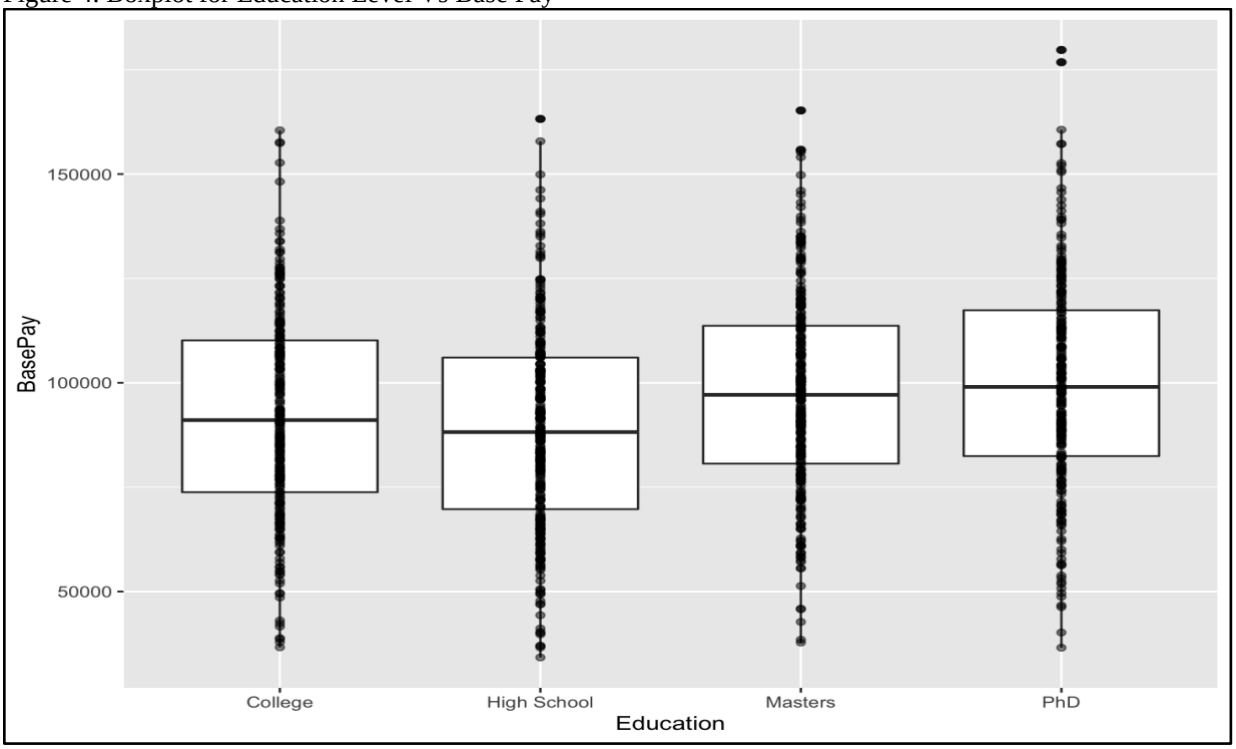


Figure 5: Boxplot for Job Title vs Base Pay

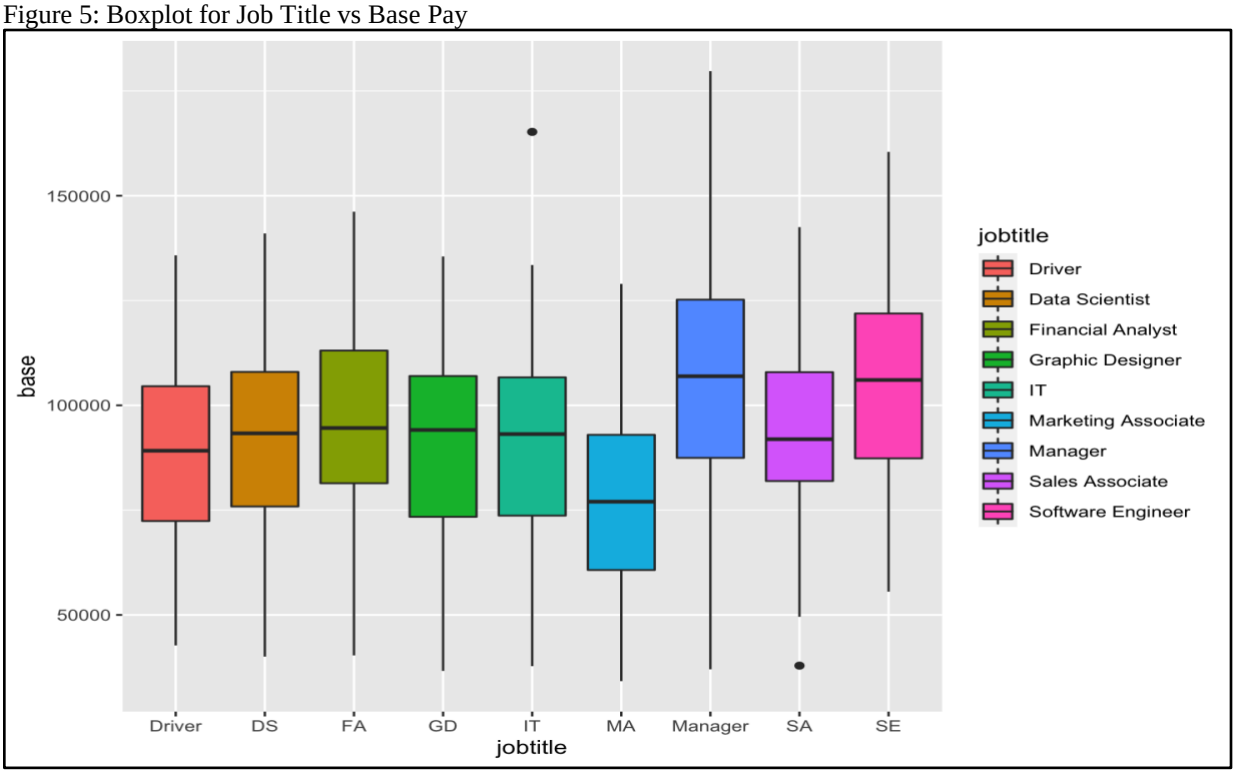


Figure 6: Boxplot for Department Vs Base Pay

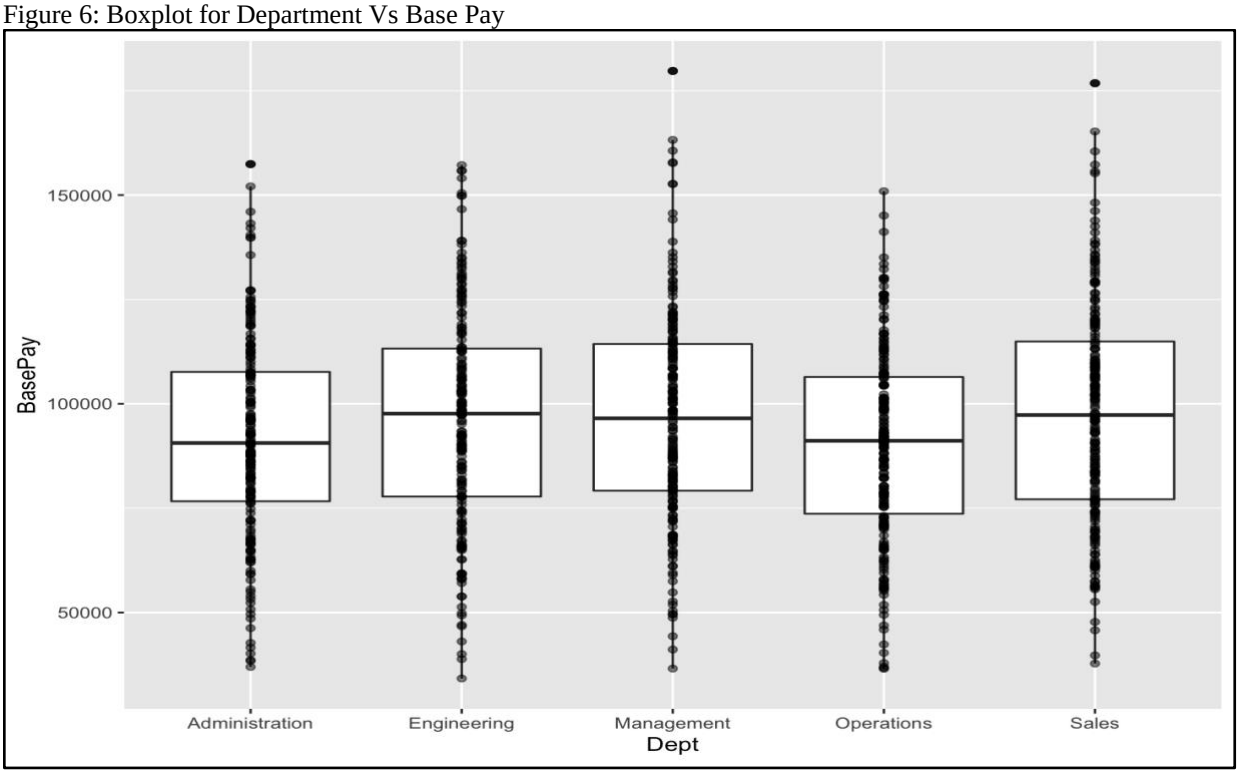


Figure 7: Plot for Seniority Level Vs Base Pay

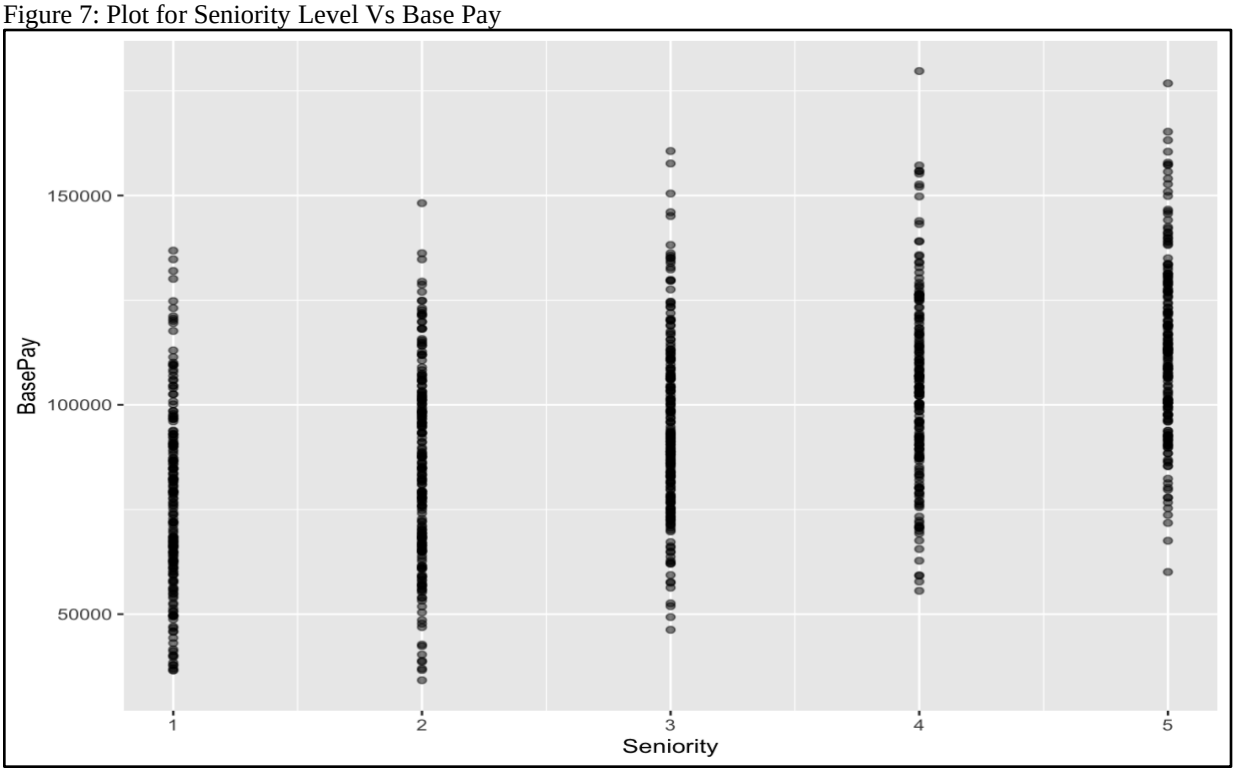


Figure 8: Plot for Performance Evaluation Score Vs Base Pay

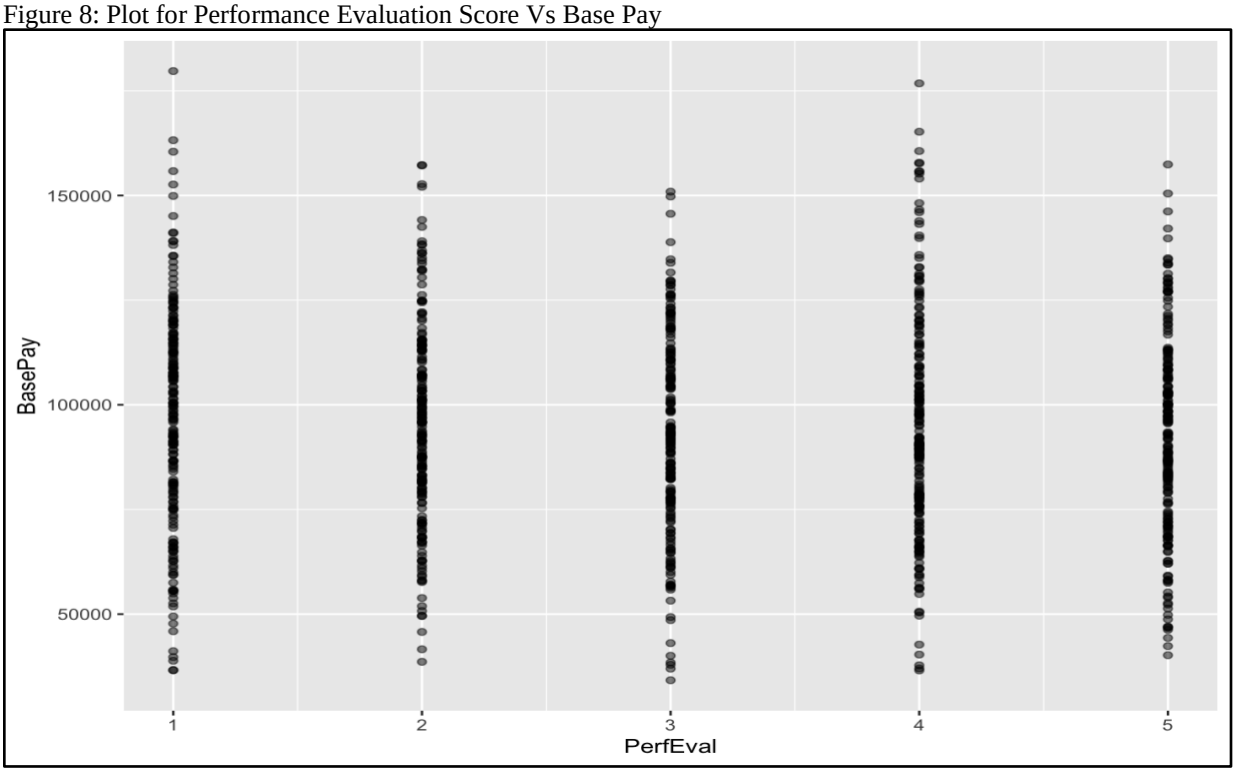


Figure 9: Plot for Performance Evaluation Score Vs Bonus

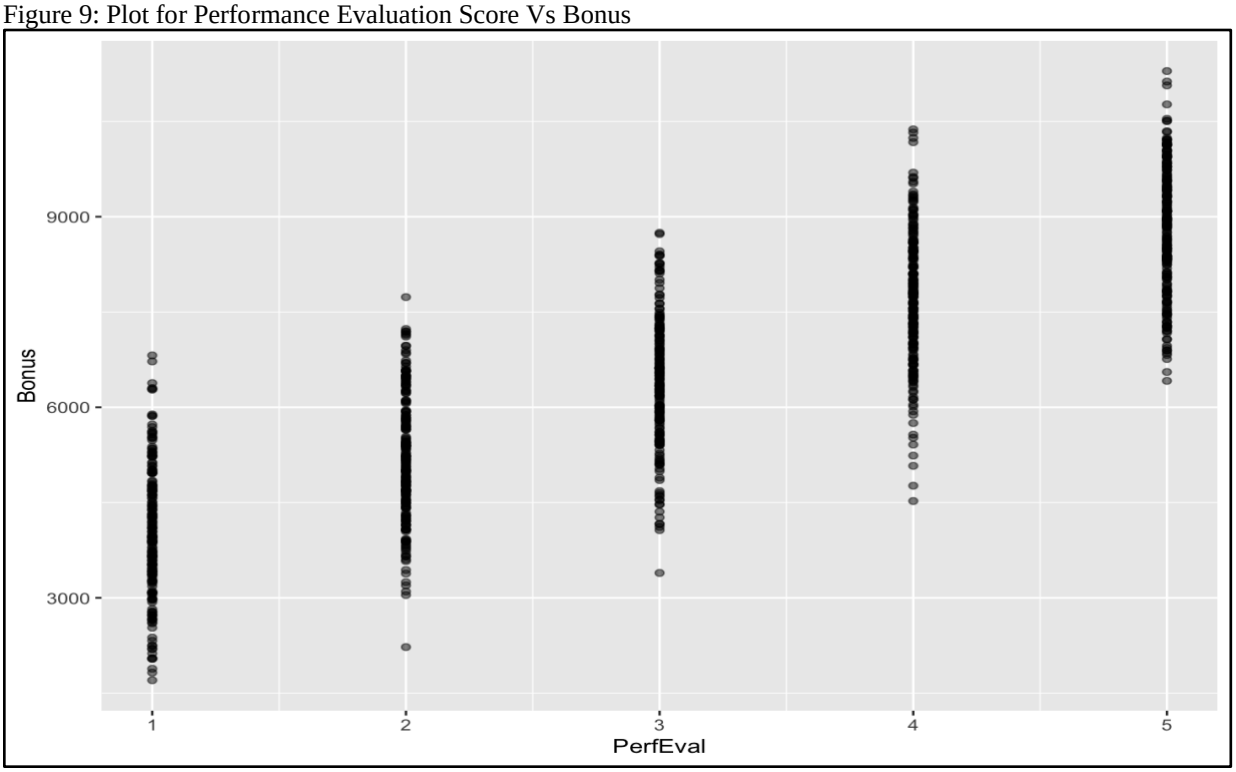


Figure 10: Boxplot for Education Vs Base pay for males and females

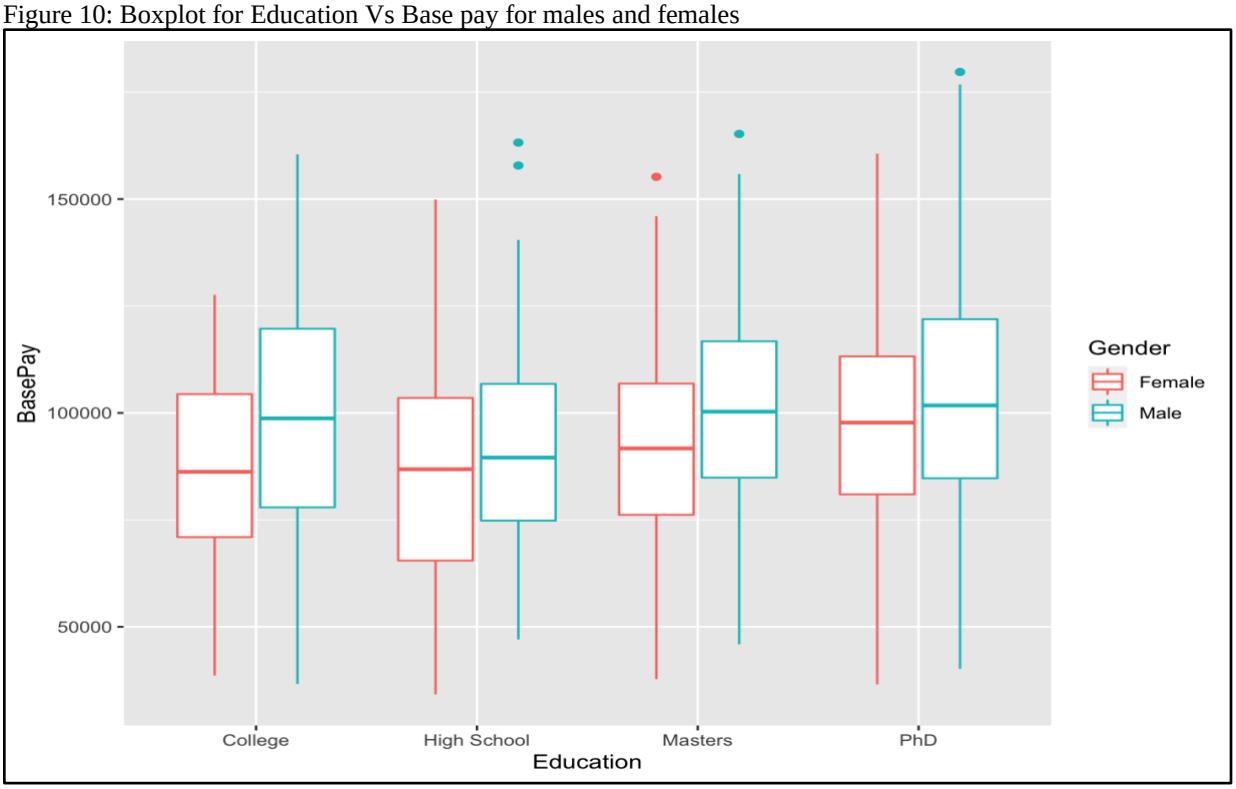


Figure 11: Boxplot for Department Vs Base pay for males and females

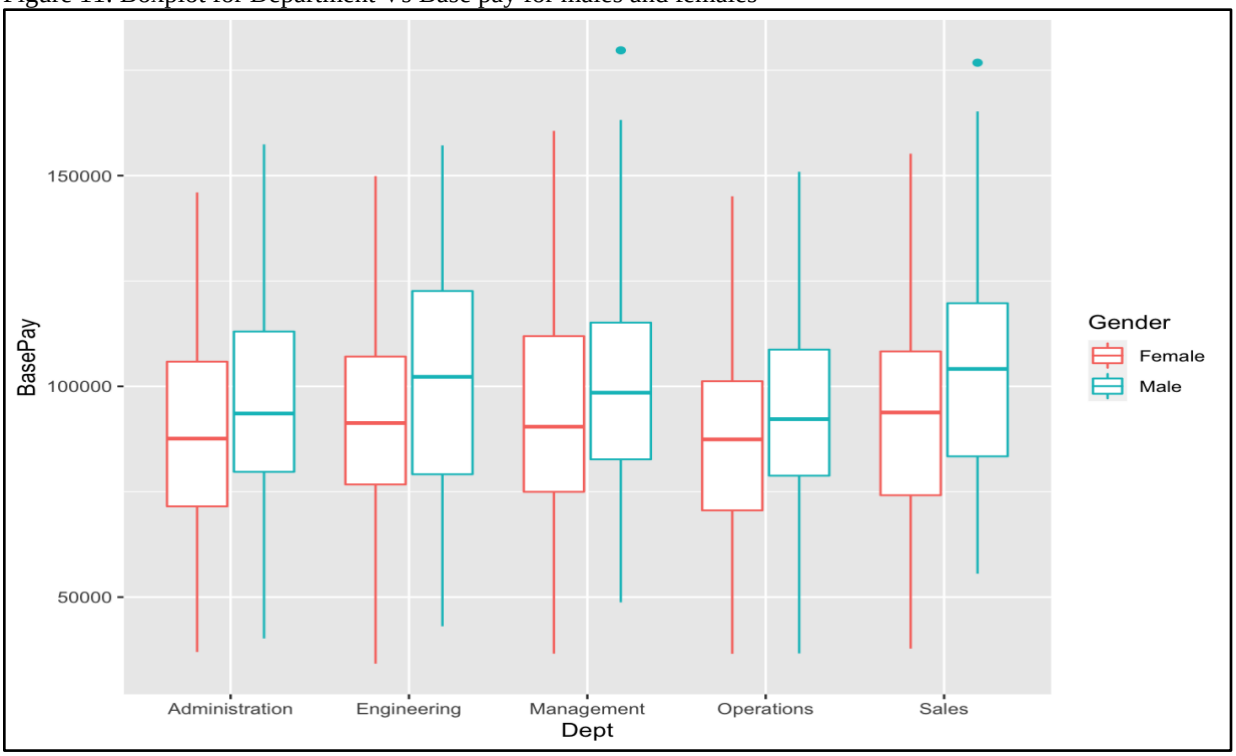


Figure 12: Age vs Base pay distribution of Males and Females across job titles

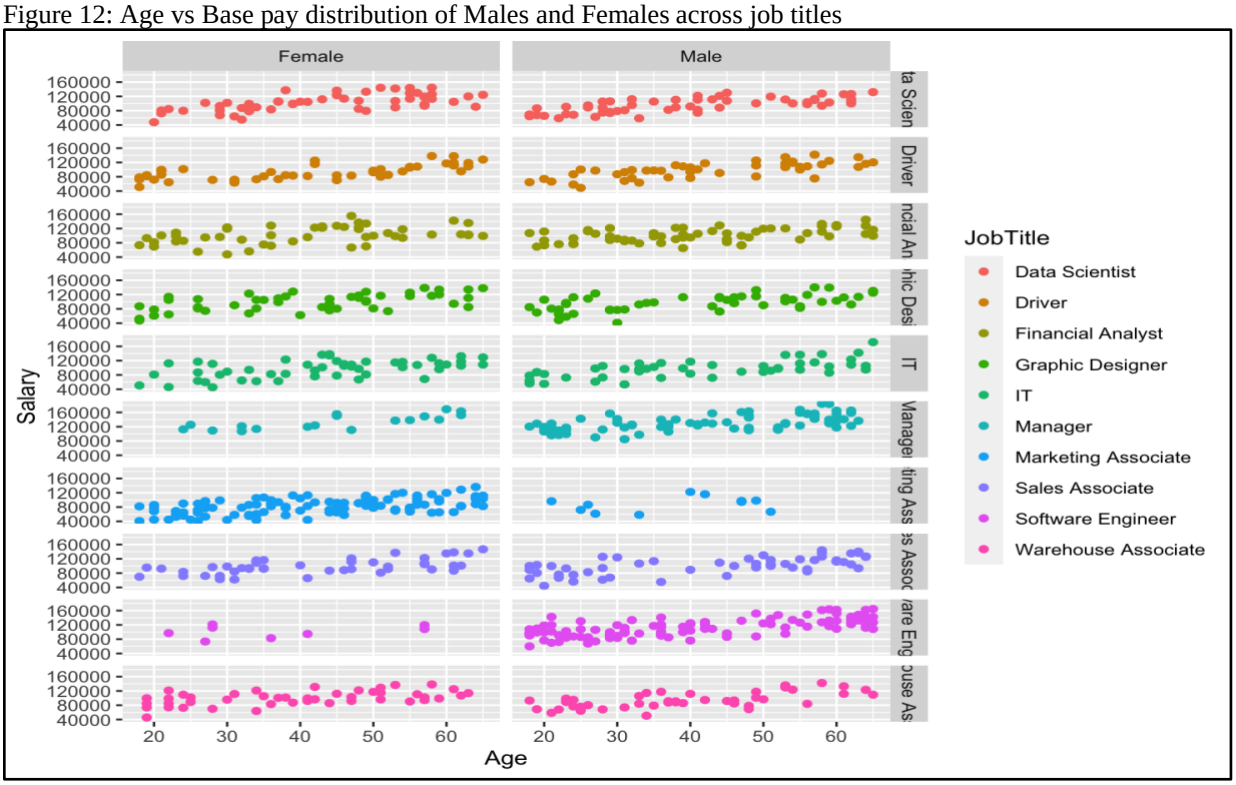
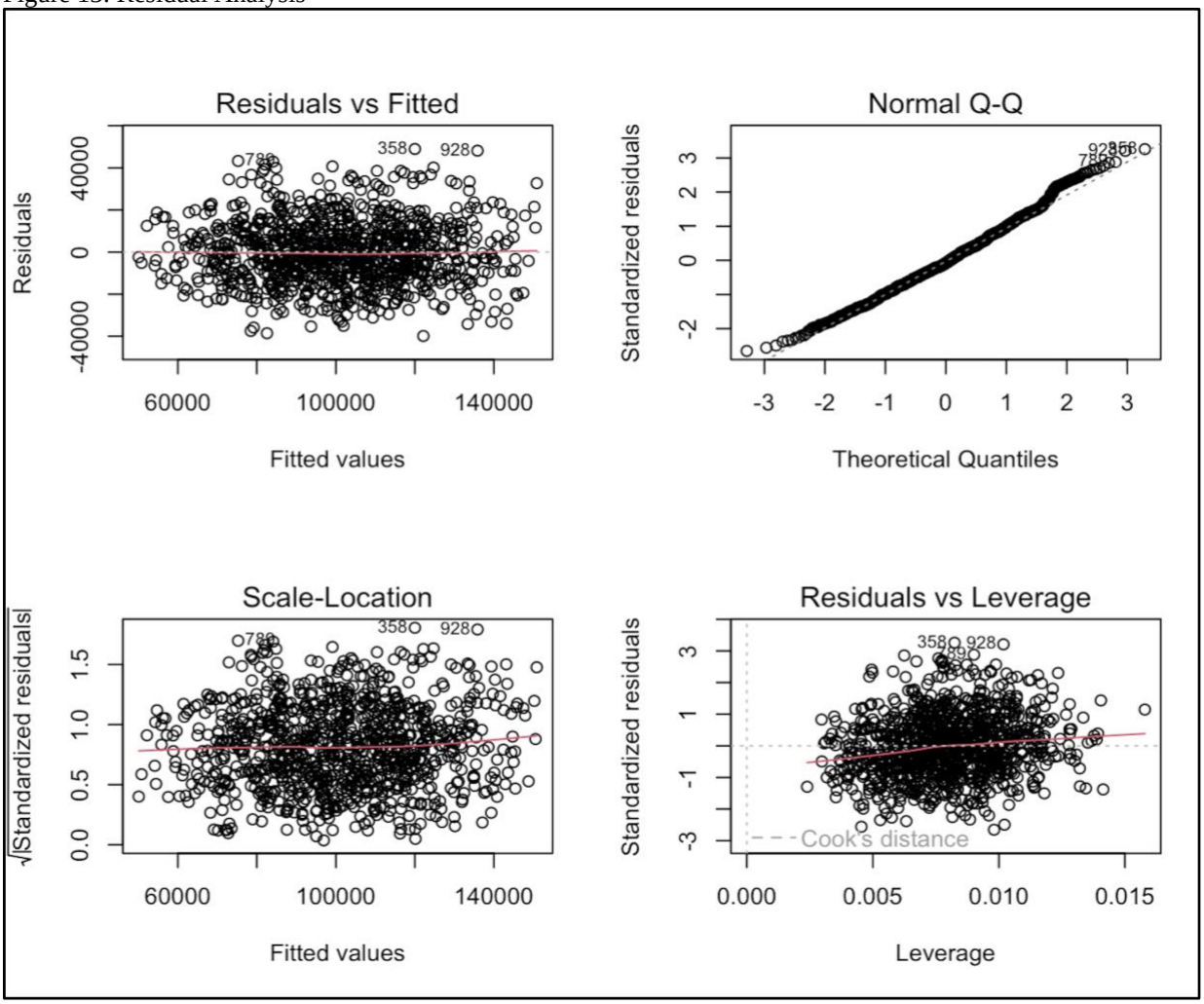


Figure 13: Residual Analysis



Code

R Code

```
# Library and Packages
install.packages("dslabs")
install.packages("ggridges")
library(dplyr)
library(ggplot2)

# Importing dataset
destfile <- "/Users/tanyasharma/Desktop/Statistics and Data Science
Books/Statistical Programming/Final Project/glassdoor_data.csv"
df <- read.csv(destfile)
head(df,5)

age <- df$Age
jobtitle <- df$jobtitle
performance <- as.ordered(df$PerfEval)
senior <- as.ordered(df$Seniority)
base <- df$BasePay

# Bar Graphs for Categorical Variables
df %>% ggplot(aes(Gender))+geom_bar()
df %>% ggplot(aes(Education))+geom_bar()
df %>% ggplot(aes(Dept))+geom_bar()
df %>% ggplot(aes(JobTitle))+geom_bar()
df %>% ggplot(aes(PerfEval))+geom_bar()
df %>% ggplot(aes(Seniority))+geom_bar()

# Exploratory Data Analysis

# Kernel Density Plots
system("R CMD SHLIB kernel_density.c")
dyn.load("kernel_density.so")


den_fn = function(data1,xgrid1,e1){
  c = length(data1)
  d = length(xgrid1)
  fn = .C("kernel_density", x = as.double(data1), n = as.integer(c),
g=as.double(xgrid1), m = as.integer(d), bw = as.double(e1), y = double(d))
  fn$y
}

age_sample = seq(min(age), max(age), length.out = 500)
e1 <- bw.nrd(age)

k_estimates <- den_fn(age,age_sample,e1)


# Plotting the estimates
plot.new()
```

```
par(mar = c(1, 1, 1, 1))
plot(k_estimates, main="Density plot of Age")
polygon(k_estimates, col="red", border="blue")

base_sample = seq(min(base), max(base), length.out = 500)
e1 <- bw.nrd(base)

k_estimates <- den_fn(base,base_sample,e1)

# Plotting the estimates
plot.new()
par(mar = c(1, 1, 1, 1))
plot(k_estimates, main="Density plot of BasePay")
polygon(k_estimates, col="red", border="blue")

# Boxplots

p <- df %>%
  ggplot(aes(Gender, BasePay))+
  geom_boxplot()
p+geom_point(alpha=0.5)

p <- df %>%
  ggplot(aes(Education, BasePay))+
  geom_boxplot()
p+geom_point(alpha=0.5)

p <- df %>%
  ggplot(aes(Dept, BasePay))+
  geom_boxplot()
p+geom_point(alpha=0.5)


base <- df$BasePay
df_temp <- data.frame(base,jobtitle)
pg_plot <- ggplot(df_temp, aes(x = jobtitle, y = base, fill = jobtitle)) +
  geom_boxplot()
# Change the legend labels
pg_plot +
  scale_fill_discrete(labels = c("Driver", "Data Scientist", "Financial
Analyst", "Graphic Designer", "IT","Marketing Associate","Manager","Sales
Associate","Software Engineer"))

#Plots focussing on Gender Pay Gap

p <- df %>%
  ggplot(aes(Dept, BasePay, color=Gender))+
  geom_boxplot()
p

p <- df %>%
  ggplot(aes(Education, BasePay, color=Gender))+
  geom_boxplot()
p
```

```
filter(df, Gender%in%c("Female","Male")) %>%
  ggplot(aes(Age, Salary, col=JobTitle))+
  geom_point()+
  facet_grid(JobTitle~Gender)


# Coding Categorical Variables
gender <- c()
for (i in df$Gender){
  if (i == "Male"){
    gender = c(gender,0)
  }
  else{
    gender = c(gender,1)
  }
}

jobt <- c()
for (i in df$JobTitle){
  if (i == "Graphic Designer"){
    jobt = c(jobt,0)
  } else if (i == "Software Engineer"){
    jobt = c(jobt,1)
  } else if (i == "Warehouse Associater"){
    jobt = c(jobt,2)
  } else if (i == "IT"){
    jobt = c(jobt,3)
  } else if (i == "Sales Associate"){
    jobt = c(jobt,4)
  } else if (i == "Driver"){
    jobt = c(jobt,5)
  } else if (i == "Financial Analyst"){
    jobt = c(jobt,6)
  } else if (i == "Marketing Associate"){
    jobt = c(jobt,7)
  } else if (i == "Data Scientist"){
    jobt = c(jobt,8)
  } else{
    jobt = c(jobt,9)
  }
}

jobtitle <- c()
for (i in df$JobTitle){
  if (i == "Graphic Designer"){
    jobtitle = c(jobtitle,"GD")
  } else if (i == "Software Engineer"){
    jobtitle = c(jobtitle,"SE")
  } else if (i == "Warehouse Associater"){
    jobtitle = c(jobtitle,"WA")
  } else if (i == "IT"){
    jobtitle = c(jobtitle,"IT")
  } else if (i == "Sales Associate"){
```

```
    jobtitle = c(jobtitle,"SA")
  } else if (i == "Driver"){
    jobtitle = c(jobtitle,"Driver")
  } else if (i == "Financial Analyst"){
    jobtitle = c(jobtitle,"FA")
  } else if (i == "Marketing Associate"){
    jobtitle = c(jobtitle,"MA")
  } else if (i == "Data Scientist"){
    jobtitle = c(jobtitle,"DS")
  } else{
    jobtitle = c(jobtitle,"Manager")
  }
}


dept <- c()
for (i in df$Dept){
  if (i == "Operations"){
    dept = c(dept,0)
  } else if (i == "Management"){
    dept = c(dept,1)
  } else if (i == "Administration"){
    dept = c(dept,2)
  } else{
    dept = c(dept,3)
  }
}

edu <- c()
for (i in df$Education){
  if (i == "High School"){
    edu = c(edu,12)
  } else if (i == "College"){
    edu = c(edu,16)
  } else if (i == "Masters"){
    edu = c(edu,18)
  } else{
    edu = c(edu,22)
  }
}


senior <- c()
for (i in df$Seniority){
  if (i == 1){
    senior = c(senior,0)
  } else if (i == 2){
    senior = c(senior,1)
  } else if (i == 3){
    senior = c(senior,2)
  } else if (i == 4){
    senior = c(senior,3)
  } else{
    senior = c(senior,4)
```

```
  }
}

# Multi-Way Anova
model1 <- lm(base~gender+jobt+edu+dept+senior+performance)
anova(model1)

#Linear Reression
model2 <- lm(salary~age+gender+jobt+edu+dept+senior+performance)
summary(model2)

# Residual Plots
diagRegressionPlots(model)
```

C Code

```
#include <stdio.h>
#include <Rmath.h>
#include <R.h>
#include <math.h>

void kernel_density (double *x, int *n, double *g, int *m, double *bw,
double *y)
{
  int i;
  int j;
  double a;
  for(i = 0; i < *m; i++){
    a = 0.0;
    for(j=0; j < *n; j++){
      double temp = x[j] - g[i];
      a += dnorm(temp, 0.0 , *bw, 0);
    }
    y[i] = a / ((double) *n);
  }
}
```