AS3: Statistical Analysis and Hypothesis Testing

By: Kevin Karnadi Kirmansjah 紀維鑫 - 109006241

---

1) Import the csv file into R and present the descriptive statistics of the numerical variables as well as the categorical variables in the dataset.

```
data = fread('banksalary.csv')
head(data)
```

```
##    Employee EducLev JobGrade YrsExper Age Gender YrsPrior PCJob   Salary
## 1:        1       3        1        3  26   Male        1    No $32,000
## 2:        2       1        1       14  38 Female        1    No $39,100
## 3:        3       1        1       12  35 Female        0    No $33,200
## 4:        4       2        1        8  40 Female        7    No $30,600
## 5:        5       3        1        3  28   Male        0    No $29,000
## 6:        6       3        1        3  24 Female        0    No $30,500
```

Descriptive Statistics:

EducLev

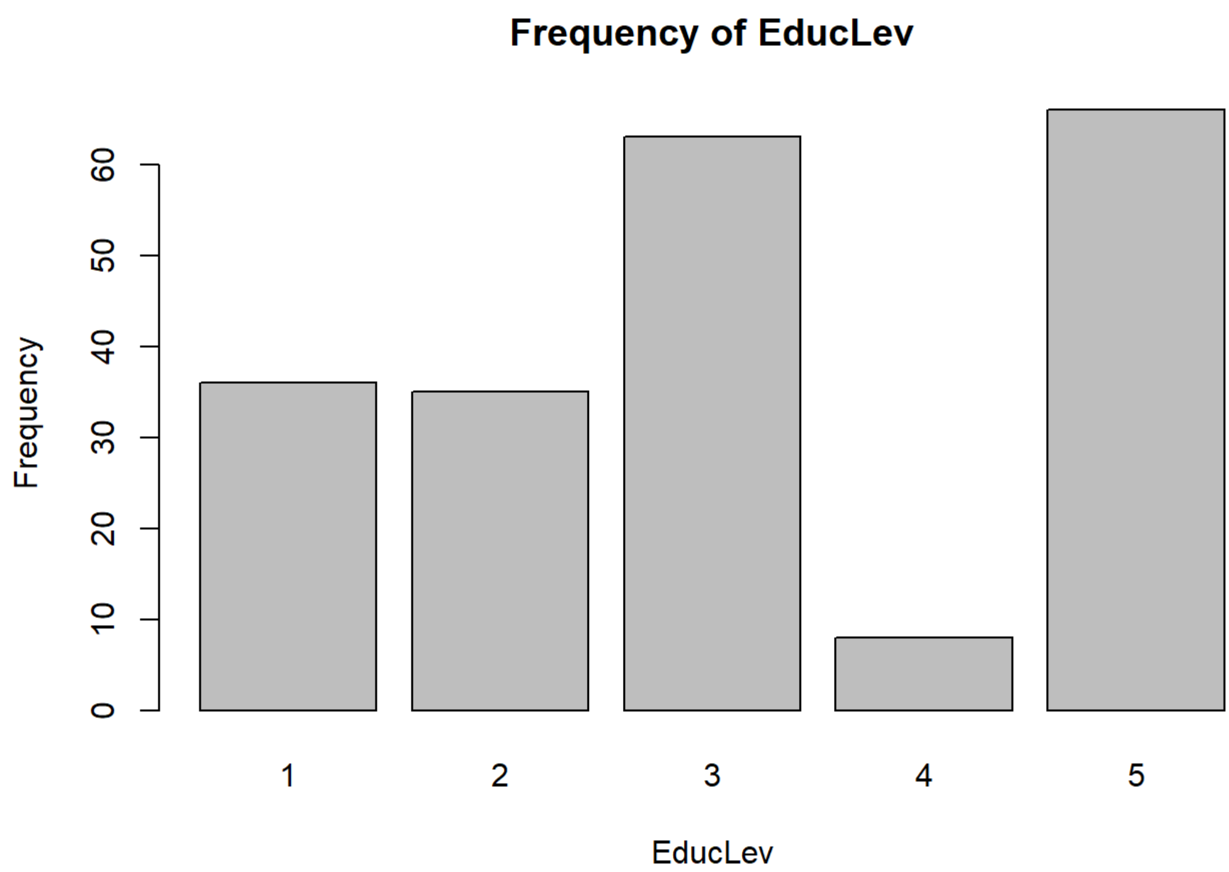
```
table(data$EducLev)
```

```
## 
##  1  2  3  4  5 
## 36 35 63  8 66
```

```
prop.table(table(data$EducLev))
```

```
## 
##          1          2          3          4          5 
## 0.17307692 0.16826923 0.30288462 0.03846154 0.31730769
```

```
EducLev.freq = table(data$EducLev)
barplot(EducLev.freq, main='Frequency of EducLev', xlab='EducLev', ylab='Frequency')
```



JobGrade

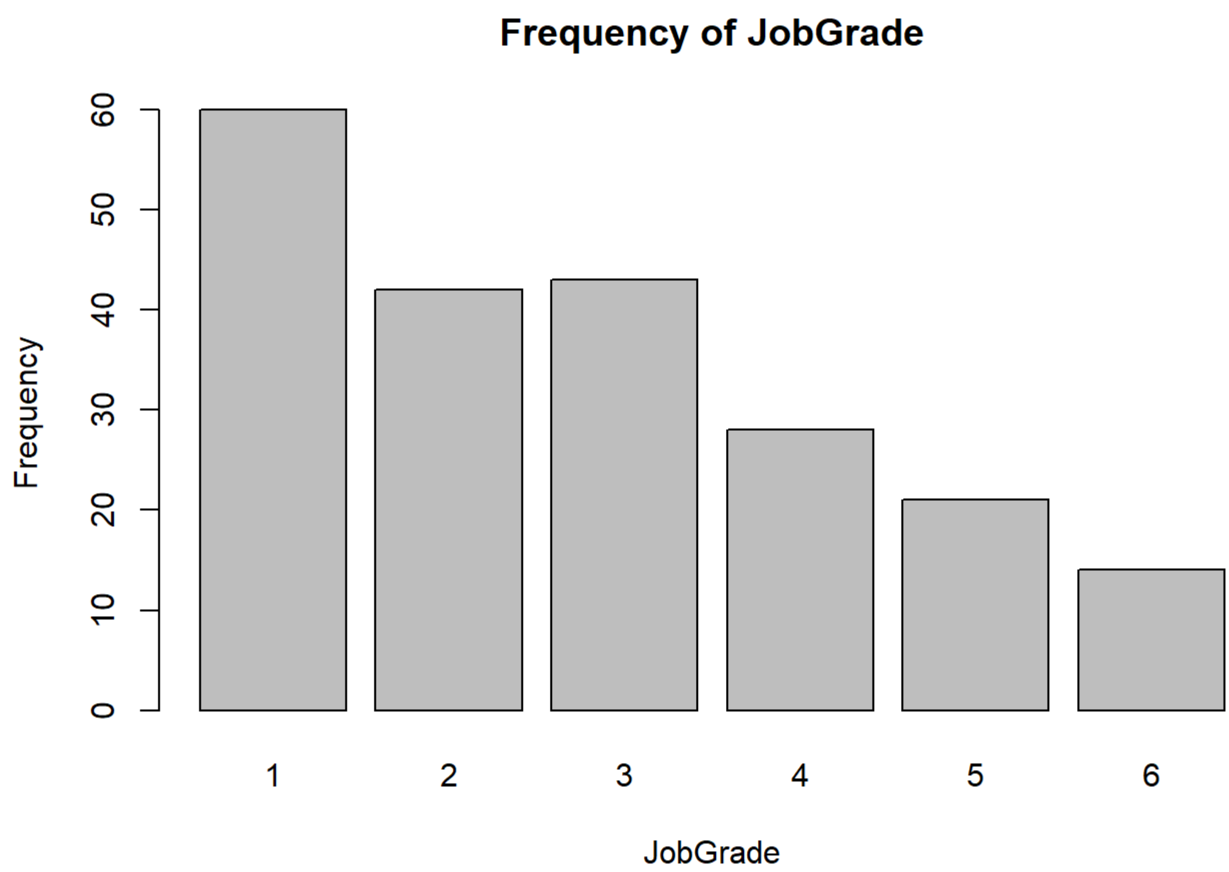
```
table(data$JobGrade)
```

```
##
##  1  2  3  4  5  6
## 60 42 43 28 21 14
```

```
prop.table(table(data$JobGrade))
```

```
##
##          1          2          3          4          5          6
## 0.28846154 0.20192308 0.20673077 0.13461538 0.10096154 0.06730769
```

```
JobGrade.freq = table(data$JobGrade)
barplot(JobGrade.freq, main='Frequency of JobGrade', xlab='JobGrade', ylab='Frequency')
```

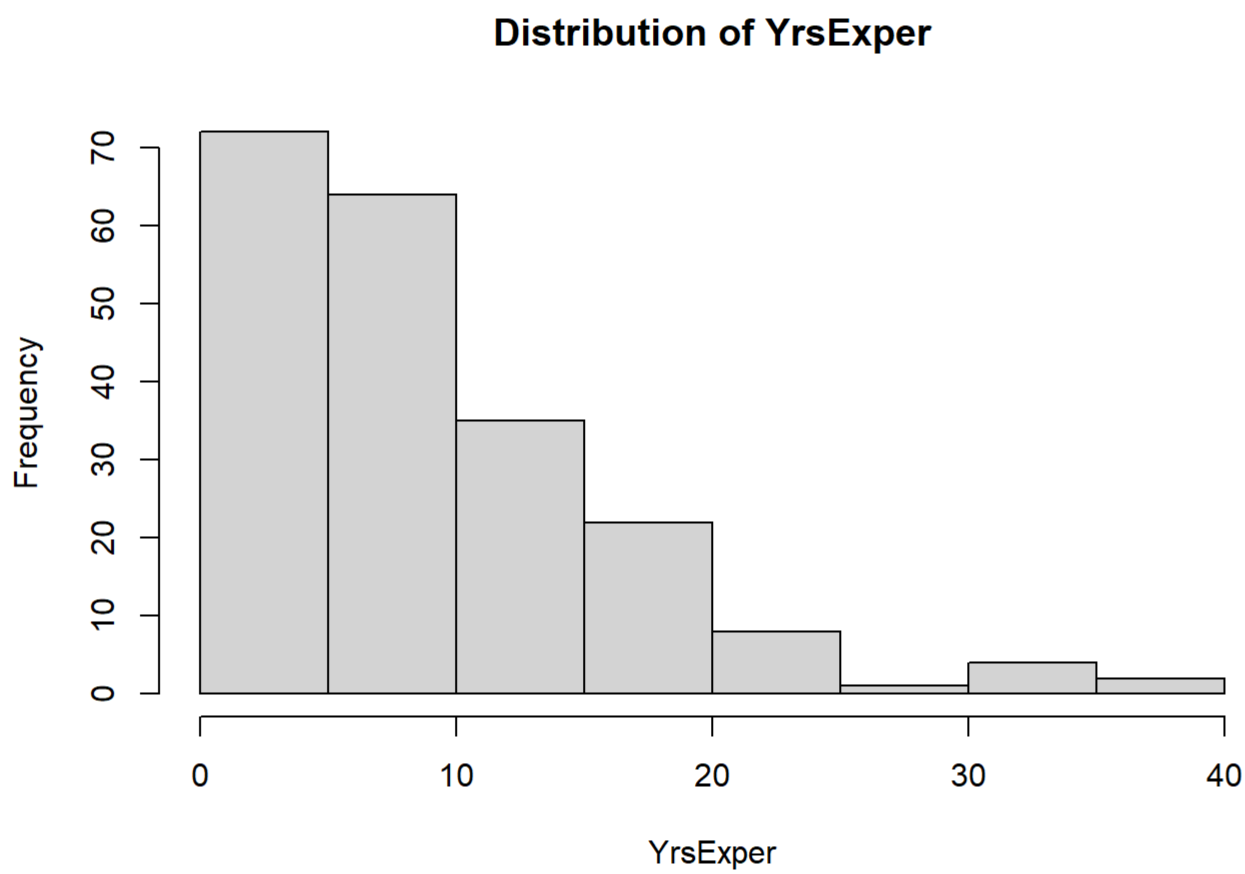


YrsExper

```
summary(data$YrsExper)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   5.000   8.000   9.673  13.000  39.000
```

```
hist(data$YrsExper, main='Distribution of YrsExper', xlab='YrsExper', ylab='Frequency')
```

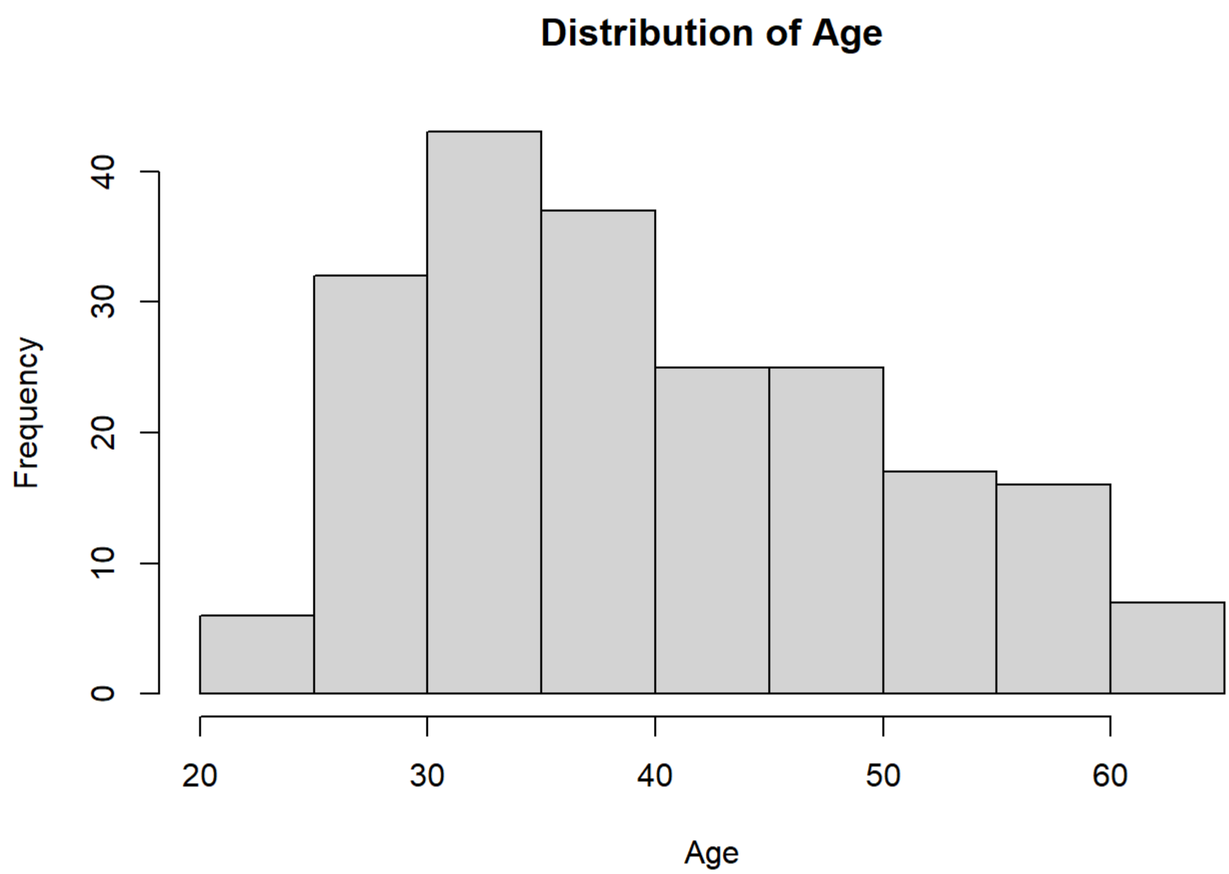


Age

```
summary(data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   32.00   38.50   40.39   47.25   65.00
```

```
hist(data$Age, main='Distribution of Age', xlab='Age', ylab='Frequency')
```



Gender

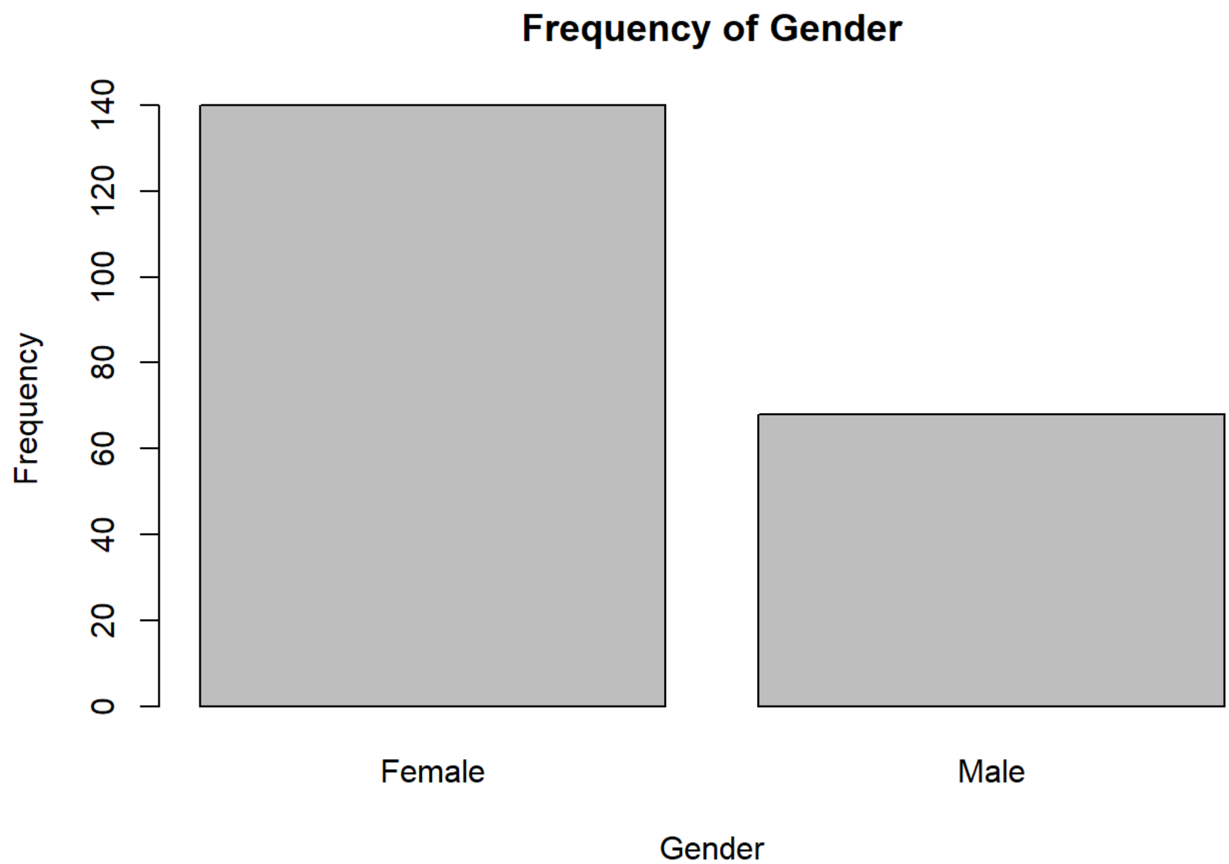
```
table(data$Gender)
```

```
##
## Female   Male
##    140     68
```

```
prop.table(table(data$Gender))
```

```
##
##    Female      Male
## 0.6730769 0.3269231
```

```
Gender.freq = table(data$Gender)
barplot(Gender.freq, main='Frequency of Gender', xlab='Gender', ylab='Frequency')
```

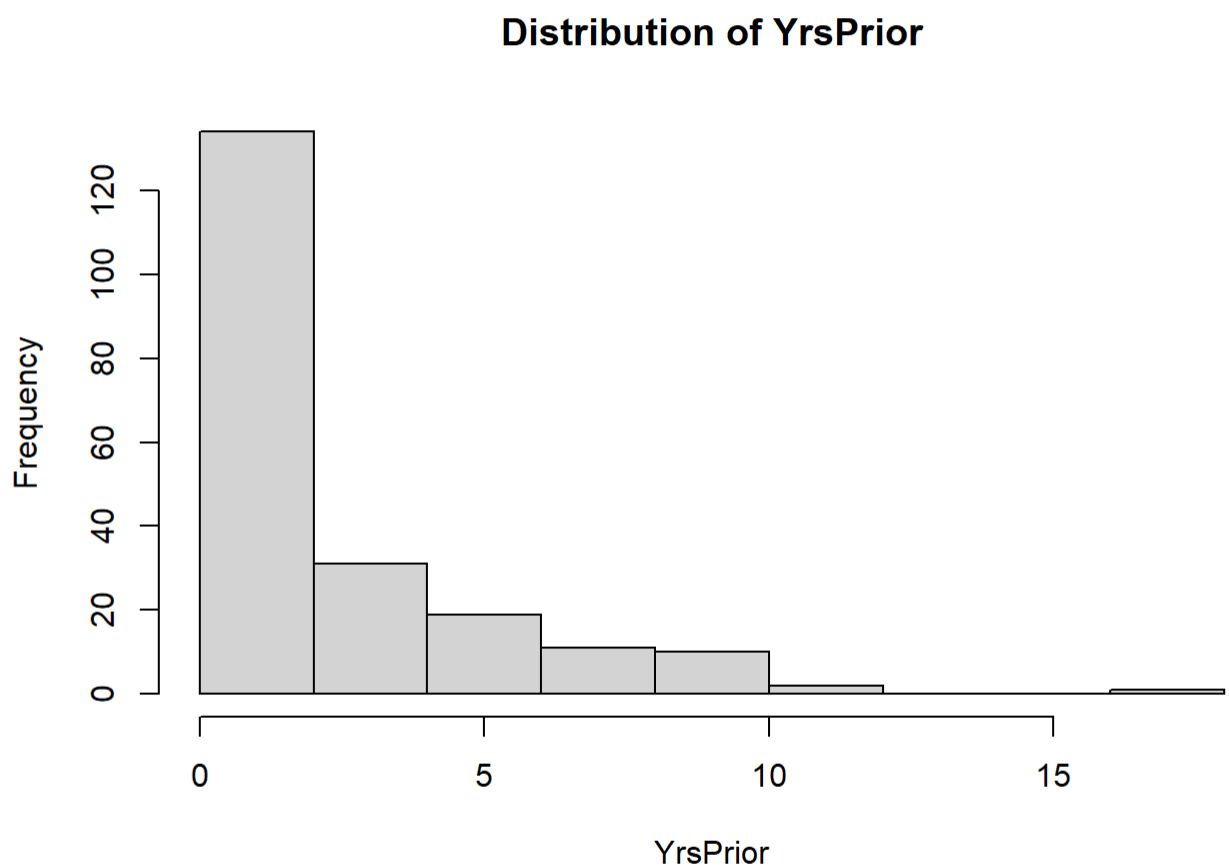


YrsPrior

```
summary(data$YrsPrior)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   1.000   2.375   4.000  18.000
```

```
hist(data$YrsPrior, main='Distribution of YrsPrior', xlab='YrsPrior', ylab='Frequency')
```



PCJob

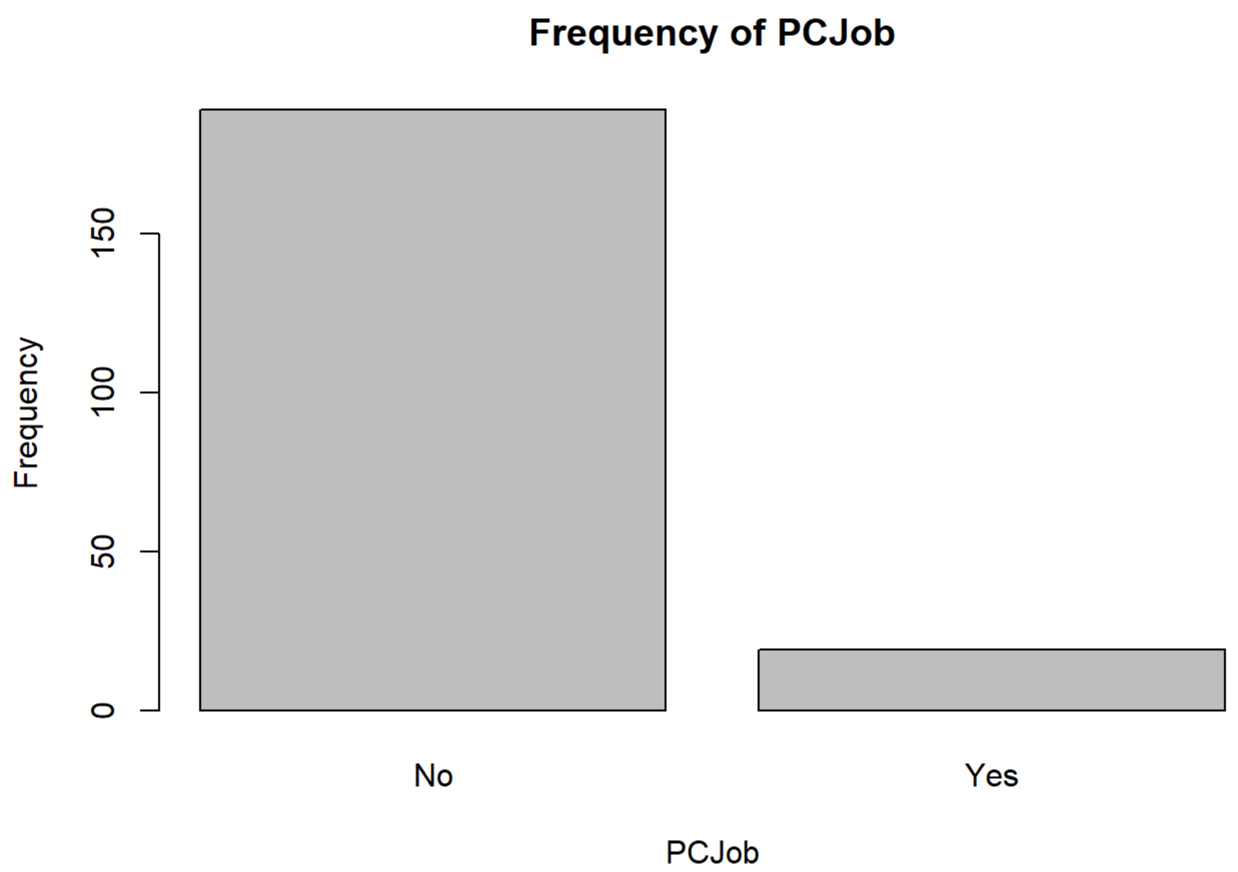
```
table(data$PCJob)
```

```
##
##  No Yes
## 189  19
```

```
prop.table(table(data$PCJob))
```

```
##
##         No        Yes
## 0.90865385 0.09134615
```

```
PCJob.freq = table(data$PCJob)
barplot(PCJob.freq, main='Frequency of PCJob', xlab='PCJob', ylab='Frequency')
```

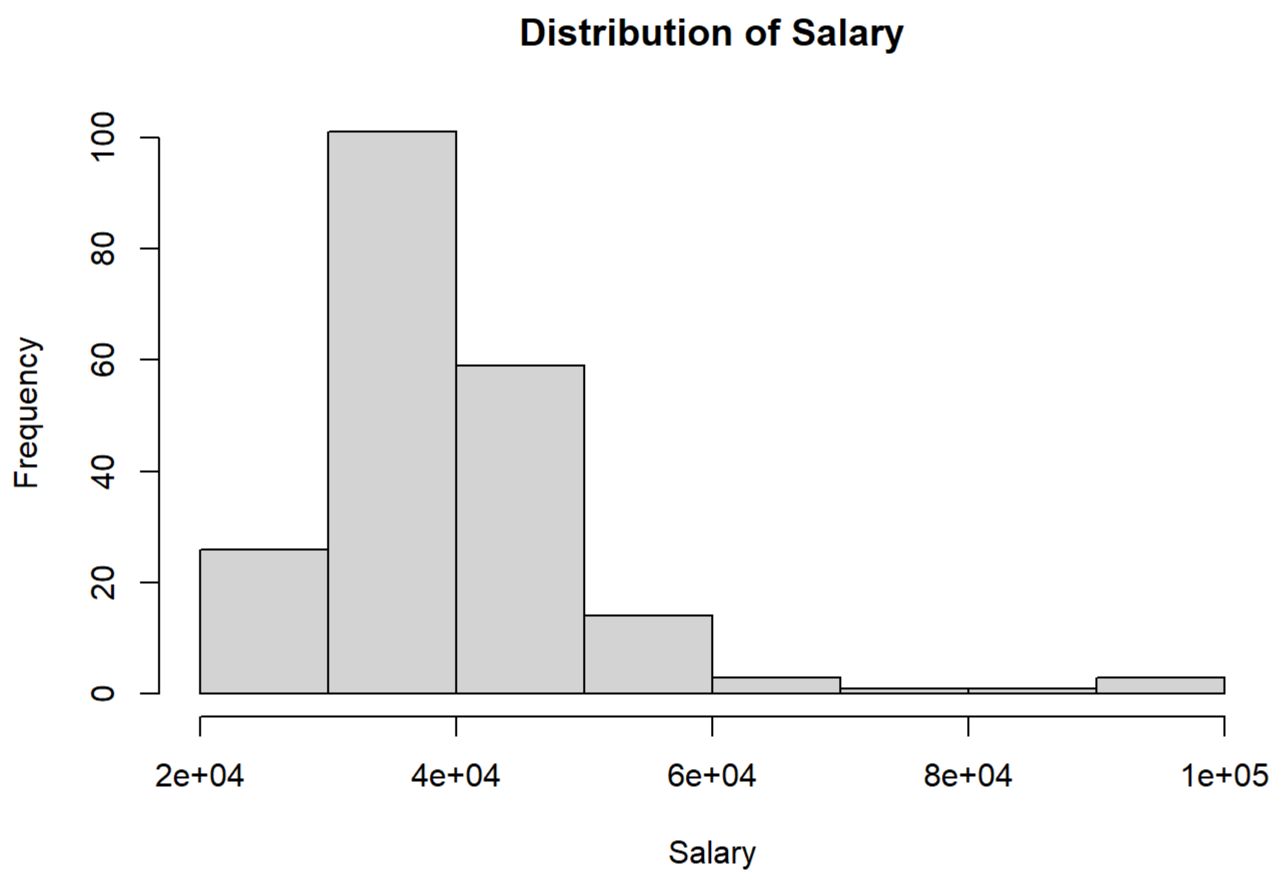


Salary

```
data$Salary.Numeric = as.integer(parse_number(data$Salary))
summary(data$Salary.Numeric)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26700   33000   37000   39922   44000   97000
```

```
hist(data$Salary.Numeric, main='Distribution of Salary', xlab='Salary', ylab='Frequency')
```



2) A plaintiff's lawyer claims that there is a significant difference in average salary between female employees and male employees. As an analyst for the plaintiff, how would you support this claim? Use a t-test and explain the results as well as your interpretation.

Subset into Male salary and Female salary:

```
salary.male = data[Gender == 'Male',]$Salary.Numeric
salary.female = data[Gender == 'Female',]$Salary.Numeric
```

Check if the sample follows a normal distribution:

```
shapiro.test(salary.male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salary.male
## W = 0.83295, p-value = 2.744e-07
```

```
shapiro.test(salary.female)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salary.female
## W = 0.92025, p-value = 4.814e-07
```

The low p-values indicates that the sample data likely does not follow a normal distribution.

Check the equality of variance:

```
ansari.test(salary.male, salary.female)
```

```
##
##  Ansari-Bradley test
##
## data:  salary.male and salary.female
## AB = 2897, p-value = 0.0009484
## alternative hypothesis: true ratio of scales is not equal to 1
```

The p-value, which is smaller than 0.1, shows that the variances are likely not equal. So we cannot use the conventional t-test, and we need to use the Welch t-test instead.

Perform Hypothesis Testing:

```
t.test(salary.male, salary.female)
```

```
##
##  Welch Two Sample t-test
##
## data:  salary.male and salary.female
## t = 4.141, df = 78.898, p-value = 8.604e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4308.082 12282.943
## sample estimates:
## mean of x mean of y
##  45505.44  37209.93
```

The p-value = 8.604e-05 is very small when compared to the significance level alpha = 0.05, meaning that we can reject the null hypothesis H0. So, we can conclude that the mean salary between male and female employees is not equal, indicating that they have a significant difference.

3) Transform EducLev into several dummy variables. The number of dummy variables you create will need to depend on your logical judgement. Also transform JobGrade, Gender, and PCJob into

dummy variables.

```
data$EducLev_1 = ifelse(data$EducLev == 1, 1, 0)
data$EducLev_2 = ifelse(data$EducLev == 2, 1, 0)
data$EducLev_3 = ifelse(data$EducLev == 3, 1, 0)
data$EducLev_4 = ifelse(data$EducLev == 4, 1, 0)

data$JobGrade_1 = ifelse(data$JobGrade == 1, 1, 0)
data$JobGrade_2 = ifelse(data$JobGrade == 2, 1, 0)
data$JobGrade_3 = ifelse(data$JobGrade == 3, 1, 0)
data$JobGrade_4 = ifelse(data$JobGrade == 4, 1, 0)
data$JobGrade_5 = ifelse(data$JobGrade == 5, 1, 0)

data$Gender_Male = ifelse(data$Gender == 'Male', 1, 0)

data$PCJob_Yes = ifelse(data$PCJob == 'Yes', 1, 0)
```

For each variable, I created k-1 dummy variables, where k is the number of unique values of that variable. We only need to use k-1 dummy variables, not k, because the last value of each variable can be represented when all the values of the k-1 dummy variables are 0.

4) The defense counsel tries to counter against the plaintiff's argument by showing that the mean difference between the two groups is biased because he or she did not control for several other factors/variables. Estimate a multiple regression model to strengthen/bolster the plaintiff's justification, then write a report explaining your results.

- Also discuss about: what R-squared is and what it means, what the meaning of the t-values and the coefficients are (or estimates).

```
model = lm(Salary.Numeric ~
             EducLev_1+EducLev_2+EducLev_3+EducLev_4 +
             JobGrade_1+JobGrade_2+JobGrade_3+JobGrade_4+JobGrade_5 +
             YrsExper +
             Age +
             Gender_Male +
             YrsPrior +
             PCJob_Yes,
           data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Salary.Numeric ~ EducLev_1 + EducLev_2 + EducLev_3 +
##     EducLev_4 + JobGrade_1 + JobGrade_2 + JobGrade_3 + JobGrade_4 +
##     JobGrade_5 + YrsExper + Age + Gender_Male + YrsPrior + PCJob_Yes,
##     data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40117  -2359   -397   1778  23958
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53658.652   3524.035  15.226  < 2e-16 ***
## EducLev_1    -2690.801   1620.891  -1.660   0.0985 .
## EducLev_2    -3176.353   1561.221  -2.035   0.0433 *
## EducLev_3    -2162.886   1169.702  -1.849   0.0660 .
## EducLev_4    -2405.625   2166.128  -1.111   0.2681
## JobGrade_1  -23832.391   2799.888  -8.512 4.75e-15 ***
## JobGrade_2  -22267.894   2742.129  -8.121 5.38e-14 ***
## JobGrade_3  -18613.032   2624.817  -7.091 2.45e-11 ***
## JobGrade_4  -15237.558   2455.646  -6.205 3.27e-09 ***
## JobGrade_5  -10172.981   2369.738  -4.293 2.79e-05 ***
## YrsExper       515.583     97.980   5.262 3.77e-07 ***
## Age             -8.962     57.699  -0.155   0.8767
## Gender_Male   2554.474   1011.974   2.524   0.0124 *
## YrsPrior       167.727    140.442   1.194   0.2338
## PCJob_Yes     4922.846   1473.825   3.340   0.0010 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5648 on 193 degrees of freedom
## Multiple R-squared:  0.7652, Adjusted R-squared:  0.7482
## F-statistic: 44.94 on 14 and 193 DF,  p-value: < 2.2e-16
```

```
stargazer(model, type='text')
```

```
##
## ===============================================
##                          Dependent variable:
##                      ---------------------------
##                            Salary.Numeric
## -----------------------------------------------
## EducLev_1                    -2,690.801*
##                              (1,620.891)
##
## EducLev_2                   -3,176.353**
##                              (1,561.221)
##
## EducLev_3                    -2,162.886*
##                              (1,169.702)
##
## EducLev_4                    -2,405.625
##                              (2,166.128)
##
## JobGrade_1                 -23,832.390***
##                              (2,799.888)
##
## JobGrade_2                 -22,267.890***
##                              (2,742.129)
##
## JobGrade_3                 -18,613.030***
##                              (2,624.817)
##
## JobGrade_4                 -15,237.560***
##                              (2,455.646)
##
## JobGrade_5                 -10,172.980***
##                              (2,369.738)
##
## YrsExper                     515.583***
##                               (97.980)
##
## Age                            -8.962
##                               (57.699)
##
## Gender_Male                  2,554.474**
##                              (1,011.974)
##
## YrsPrior                       167.727
##                              (140.442)
##
## PCJob_Yes                   4,922.846***
##                              (1,473.825)
##
## Constant                    53,658.650***
##                              (3,524.035)
##
## -----------------------------------------------
```

```
## Observations                   208
## R2                            0.765
## Adjusted R2                   0.748
## Residual Std. Error    5,648.080 (df = 193)
## F Statistic          44.939*** (df = 14; 193)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

We can see that the R-squared value is 70%++, meaning that a high percentage of the variation in the dependent variable is explained by the independent variable. From the asterisks (more asterisks means higher significance), we can see that the variables with the highest significant levels are the JobGrade and YrsExper variables.

R-squared is a statistical measure which explains how much proportion (in percentage) does an independent variable explain the variation of a dependent variable. A higher R-squared value means that a large proportion of the dependent variables can be explained by the independent variables.

The t-values are the ratio of the Estimates and Std. Error (Estimate / Std.Error). For example if Estimate = 5 and Std. Error = 2, t-value will be 5/2 = 2.5. It is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far from 0 as this would mean that we could reject the null hypothesis. In general, t-values are also used to calculate p-values.

The coefficients (or estimates), indicates the expected change in the output due to a unit change in the feature / variable. The intercept tells us that when all the features are at 0, the expected output is the intercept value itself. For the numerical variables, a coefficient C means that for every one unit increase in that specific feature, the output will go up by C. For example, in the linear model above, a 1 year increase in age will result in a -8.962 increment of Salary. For the dummy variables, a coefficient C means that a sample with that feature will have its output go up by C, when compared to the baseline value (the redundant variable which isn't expressed as a dummy variable). For example, in the linear model above, a male worker will have 2,554.474 more salary than a female worker.

5) Do these data provide evidence that there is discrimination against female employees in terms of salary?

No, there isn't a discrimination against female employees in terms of salary. We can see that the model has a high R-squared value, meaning that all the independent variables (not only the Gender) explains the variation of the dependent variable really well. Furthermore, we can see from the model summary above that for the Gender variable, there is one asterisk on it. A higher number of asterisks indicates that there is a high significance between that specific variable and the response variable. We can see that there are other variables with a higher level of significance towards the response variable than Gender, such as JobGrade, YrsExper, and PCJob, meaning that the salary level is not necessarily highly impacted by the Gender of the employee.

Extra Credit:

a. You may get more interesting results to talk about by including interaction terms in your regression model. Explain what an interaction term is, how we can estimate a regression model with interaction terms and how we could interpret the results.

An interaction effect is when the effect of an independent variable towards the dependent variable changes according to the values of other independent variables.

We can estimate a regression model with interaction terms by including the product of two or more independent variables to a regression equation.
Example: $Y = b0 + b1*X1 + b2*X2 + b3*(X1*X2)$
Interaction term: $(X1*X2)$

To interpret the results, we need to know whether if the interaction term contributes meaningfully to the explanatory power of the model or not.

b. How would you determine whether the interaction terms contribute in a meaningful way to the explanatory power of your estimation model?

We can determine this by two ways:
- Assessing the statistical significance of the interaction term
- Comparing the coefficient of determination with and without the interaction term

If the interaction term is statistically significant, the interaction term is probably important. And if the coefficient of determination is also much bigger with the interaction term, it is definitely important. If neither of these outcomes are observed, the interaction term can be removed from the regression equation.