### AS3  
  
install.packages('plyr', repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/danie/Documents/R/win-library/4.0'  
## (as 'lib' is unspecified)

## package 'plyr' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'plyr'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:  
## \Users\danie\Documents\R\win-library\4.0\00LOCK\plyr\libs\x64\plyr.dll to C:  
## \Users\danie\Documents\R\win-library\4.0\plyr\libs\x64\plyr.dll: Permission  
## denied

## Warning: restored 'plyr'

##   
## The downloaded binary packages are in  
## C:\Users\danie\AppData\Local\Temp\RtmpKAMEoR\downloaded\_packages

options(repos = list(CRAN="http://cran.rstudio.com/"))  
require(plyr)

## Loading required package: plyr

## Warning: package 'plyr' was built under R version 4.0.5

# 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.  
bs <- read.csv(file = "banksalary.csv")  
  
# Convert the Salary variable into numeric:  
bs$Salary <- gsub("[,$]", "", bs$Salary)  
bs$Salary <- as.numeric(bs$Salary)  
  
summary(bs)

## Employee EducLev JobGrade YrsExper   
## Min. : 1.00 Min. :1.000 Min. :1.00 Min. : 2.000   
## 1st Qu.: 52.75 1st Qu.:2.000 1st Qu.:1.00 1st Qu.: 5.000   
## Median :104.50 Median :3.000 Median :3.00 Median : 8.000   
## Mean :104.50 Mean :3.159 Mean :2.76 Mean : 9.673   
## 3rd Qu.:156.25 3rd Qu.:5.000 3rd Qu.:4.00 3rd Qu.:13.000   
## Max. :208.00 Max. :5.000 Max. :6.00 Max. :39.000   
## Age Gender YrsPrior PCJob   
## Min. :22.00 Length:208 Min. : 0.000 Length:208   
## 1st Qu.:32.00 Class :character 1st Qu.: 0.000 Class :character   
## Median :38.50 Mode :character Median : 1.000 Mode :character   
## Mean :40.39 Mean : 2.375   
## 3rd Qu.:47.25 3rd Qu.: 4.000   
## Max. :65.00 Max. :18.000   
## Salary   
## Min. :26700   
## 1st Qu.:33000   
## Median :37000   
## Mean :39922   
## 3rd Qu.:44000   
## Max. :97000

# 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.  
  
# First, let's check the normality and variances within the Gender variable:  
require(data.table)

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 4.0.5

bs <- data.table(bs)  
  
# Normality test:  
bs[, shapiro.test(Salary), Gender] # p-values are close to 0, the distribution is probably not normal.

## Gender statistic p.value method data.name  
## 1: Male 0.8329482 2.744032e-07 Shapiro-Wilk normality test Salary  
## 2: Female 0.9202464 4.814479e-07 Shapiro-Wilk normality test Salary

# So, we shall use an Ansari-Bradley Test for the equality of variances:  
ansari.test(Salary ~ Gender, bs) # The p-value is very small, indicating that the variances are probably not equal.

##   
## Ansari-Bradley test  
##   
## data: Salary by Gender  
## AB = 8024, p-value = 0.0009319  
## alternative hypothesis: true ratio of scales is not equal to 1

t.test(Salary ~ Gender, data = bs, var.equal = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: Salary by Gender  
## 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:  
## -12282.943 -4308.082  
## sample estimates:  
## mean in group Female mean in group Male   
## 37209.93 45505.44

# The p-value is close to zero, therefore, we can reject the H\_0 and confirm the lawyer's claims. There is a significant difference in average salary between males and females.

# 3) Transform EducLev into several dummy variables. The number of dummy variables you create will depend on your logical judgment. Also transform JobGrade, Gender, and PCJob into dummy variables.  
install.packages("fastDummies")

## Installing package into 'C:/Users/danie/Documents/R/win-library/4.0'  
## (as 'lib' is unspecified)

## package 'fastDummies' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\danie\AppData\Local\Temp\RtmpKAMEoR\downloaded\_packages

require(fastDummies)

## Loading required package: fastDummies

## Warning: package 'fastDummies' was built under R version 4.0.5

bs <- dummy\_cols(bs, select\_columns = c("EducLev", "JobGrade", "Gender", "PCJob"), remove\_first\_dummy = TRUE)  
head(bs)

## Employee EducLev JobGrade YrsExper Age Gender YrsPrior PCJob Salary  
## 1: 1 3 1 3 26 Male 1 No 32000  
## 2: 2 1 1 14 38 Female 1 No 39100  
## 3: 3 1 1 12 35 Female 0 No 33200  
## 4: 4 2 1 8 40 Female 7 No 30600  
## 5: 5 3 1 3 28 Male 0 No 29000  
## 6: 6 3 1 3 24 Female 0 No 30500  
## EducLev\_2 EducLev\_3 EducLev\_4 EducLev\_5 JobGrade\_2 JobGrade\_3 JobGrade\_4  
## 1: 0 1 0 0 0 0 0  
## 2: 0 0 0 0 0 0 0  
## 3: 0 0 0 0 0 0 0  
## 4: 1 0 0 0 0 0 0  
## 5: 0 1 0 0 0 0 0  
## 6: 0 1 0 0 0 0 0  
## JobGrade\_5 JobGrade\_6 Gender\_Male PCJob\_Yes  
## 1: 0 0 1 0  
## 2: 0 0 0 0  
## 3: 0 0 0 0  
## 4: 0 0 0 0  
## 5: 0 0 1 0  
## 6: 0 0 0 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).  
reg <- lm(Salary ~ YrsExper + Age + YrsPrior + EducLev\_2 + EducLev\_3 + EducLev\_4 + EducLev\_5 + JobGrade\_2 + JobGrade\_3 + JobGrade\_4 + JobGrade\_5 + JobGrade\_6 + Gender\_Male + PCJob\_Yes, data = bs)  
  
install.packages("stargazer")

## Installing package into 'C:/Users/danie/Documents/R/win-library/4.0'  
## (as 'lib' is unspecified)

## package 'stargazer' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\danie\AppData\Local\Temp\RtmpKAMEoR\downloaded\_packages

require(stargazer)

## Loading required package: stargazer

## Warning: package 'stargazer' was built under R version 4.0.3

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

stargazer(reg, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## Salary   
## -----------------------------------------------  
## YrsExper 515.583\*\*\*   
## (97.980)   
##   
## Age -8.962   
## (57.699)   
##   
## YrsPrior 167.727   
## (140.442)   
##   
## EducLev\_2 -485.552   
## (1,398.657)   
##   
## EducLev\_3 527.915   
## (1,357.519)   
##   
## EducLev\_4 285.176   
## (2,404.727)   
##   
## EducLev\_5 2,690.801\*   
## (1,620.891)   
##   
## JobGrade\_2 1,564.497   
## (1,185.771)   
##   
## JobGrade\_3 5,219.358\*\*\*   
## (1,262.395)   
##   
## JobGrade\_4 8,594.833\*\*\*   
## (1,496.018)   
##   
## JobGrade\_5 13,659.410\*\*\*   
## (1,874.269)   
##   
## JobGrade\_6 23,832.390\*\*\*   
## (2,799.888)   
##   
## Gender\_Male 2,554.474\*\*   
## (1,011.974)   
##   
## PCJob\_Yes 4,922.846\*\*\*   
## (1,473.825)   
##   
## Constant 27,135.460\*\*\*   
## (2,455.280)   
##   
## -----------------------------------------------  
## 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

# 5) Do these data provide evidence that there is discrimination against female employees in terms of salary?  
  
# Yes, males earn 2,554.47 more than female on average, other conditions being equal. So, there is definitely an evidence that there is a discrimination against female employees in terms of salary.

### Extra credit  
# 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.  
  
inter <- lm(Salary ~ YrsExper + Age + YrsPrior + EducLev\_2 + EducLev\_3 + EducLev\_4 + EducLev\_5 + JobGrade\_2 + JobGrade\_3 + JobGrade\_4 + JobGrade\_5 + JobGrade\_6 + Gender\_Male + PCJob\_Yes + Gender\_Male \* Age, data = bs)  
  
stargazer(inter, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## Salary   
## -----------------------------------------------  
## YrsExper 506.000\*\*\*   
## (95.286)   
##   
## Age -108.835\*   
## (62.937)   
##   
## YrsPrior 90.005   
## (138.320)   
##   
## EducLev\_2 -209.890   
## (1,361.918)   
##   
## EducLev\_3 308.225   
## (1,321.138)   
##   
## EducLev\_4 206.791   
## (2,337.744)   
##   
## EducLev\_5 2,516.268   
## (1,576.460)   
##   
## JobGrade\_2 1,823.770   
## (1,155.070)   
##   
## JobGrade\_3 5,435.729\*\*\*   
## (1,228.733)   
##   
## JobGrade\_4 8,923.928\*\*\*   
## (1,457.320)   
##   
## JobGrade\_5 13,511.530\*\*\*   
## (1,822.469)   
##   
## JobGrade\_6 20,643.410\*\*\*   
## (2,870.382)   
##   
## Gender\_Male -8,710.489\*\*   
## (3,367.160)   
##   
## PCJob\_Yes 4,815.672\*\*\*   
## (1,433.034)   
##   
## Age:Gender\_Male 298.985\*\*\*   
## (85.469)   
##   
## Constant 31,456.900\*\*\*   
## (2,687.525)   
##   
## -----------------------------------------------  
## Observations 208   
## R2 0.779   
## Adjusted R2 0.762   
## Residual Std. Error 5,490.503 (df = 192)   
## F Statistic 45.201\*\*\* (df = 15; 192)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01