Final project: Is College Worth It?

Due date: December 9, 2019

```
rm(list = ls())
load("dataset.RData")
dim(dataset)
```

```
## [1] 115152     63
```

```
head(dataset)
```

```
##            PERSONID YEAR  WEIGHT SAMPLE SURID AGE GENDER MINRTY RACETH CHTOT
## 1 20000000003230008 2013 30.9312   1002     2  56      1      0      2    NA
## 2 20000000003250708 2013 31.1697   1002     2  57      1      0      2    NA
## 3 20000000008220908 2013 31.1697   1002     2  59      1      0      2    NA
## 4 20000000008250308 2013 31.1697   1002     2  58      1      0      2    NA
## 5 20000000008270008 2013 32.7715   1002     2  58      1      0      2    NA
## 6 20000000103230700 2013 30.3566   1002     2  61      2      0      2    NA
##   CHU2IN CH25IN CH611IN CH1218IN CH19IN BA03Y5 NBAMEMG BADGRUS DGRDG HD03Y5
## 1     NA     NA      NA       NA     NA   1976       3       1     3   1981
## 2     NA     NA      NA       NA     NA   1976       3       1     3   1981
## 3     NA     NA      NA       NA     NA   1971       3       1     3   1981
## 4     NA     NA      NA       NA     NA   1976       3       1     3   1981
## 5     NA     NA      NA       NA     NA   1976       1       1     3   1981
## 6     NA     NA      NA       NA     NA   1971       2       1     3   1981
##   NDGMEMG HDDGRUS LFSTAT HRSWKGR WKSWKGR JOBINS JOBPENS JOBPROFT JOBVAC LOOKWK
## 1       3       1      1       1       4      0       1        0      0   <NA>
## 2       3       1      3    <NA>    <NA>   <NA>    <NA>     <NA>   <NA>      0
## 3       3       1      1       4       4      1       1        1      1   <NA>
## 4       3       1      1       4       3      1       1        0      1   <NA>
## 5       3       1      3    <NA>    <NA>   <NA>    <NA>     <NA>   <NA>      0
## 6       2       1      1       4       4      1       1        0      1   <NA>
##   FTPRET PTWTFT PTFAM PTNOND PTOCNA PTOTP OCEDRLP NOCPRMG EMSEC WAPRSM WASCSM
## 1      1      0     0      1      0     1       3       7     4      3      3
## 2   <NA>   <NA>  <NA>   <NA>   <NA>  <NA>    <NA>    <NA>  <NA>   <NA>   <NA>
## 3      0   <NA>  <NA>   <NA>   <NA>  <NA>       2       3     4      3      3
## 4      0   <NA>  <NA>   <NA>   <NA>  <NA>       1       3     2      1      2
## 5   <NA>   <NA>  <NA>   <NA>   <NA>  <NA>    <NA>    <NA>  <NA>   <NA>   <NA>
## 6      0   <NA>  <NA>   <NA>   <NA>  <NA>       1       2     2      1      2
##   SALARY NRREA NRSEC JOBSATIS SATADV SATBEN SATCHAL SATIND SATLOC SATRESP
## 1  45000     5     4        1      1      2       1      1      1       1
## 2     NA  <NA>  <NA>     <NA>   <NA>   <NA>    <NA>   <NA>   <NA>    <NA>
## 3 115000  <NA>  <NA>        1      2      2       1      1      1       1
## 4 139000  <NA>  <NA>        1      2      1       1      1      1       1
## 5     NA  <NA>  <NA>     <NA>   <NA>   <NA>    <NA>   <NA>   <NA>    <NA>
## 6 142000  <NA>  <NA>        2      3      3       2      1      1       2
```

```
##   SATSAL SATSEC SATSOC MGRNAT MGROTH MGRSOC NWFAM NWLAY NWNOND NWOCNA NWOTP
## 1      1      1      1      0      1      0  <NA>  <NA>   <NA>   <NA>  <NA>
## 2   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>     0     0      1      0     1
## 3      1      1      1      0      1      0  <NA>  <NA>   <NA>   <NA>  <NA>
## 4      3      1      1      1      0      0  <NA>  <NA>   <NA>   <NA>  <NA>
## 5   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>     0     0      1      0     1
## 6      4      1      1      1      0      0  <NA>  <NA>   <NA>   <NA>  <NA>
##   NWSTU
## 1  <NA>
## 2     0
## 3  <NA>
## 4  <NA>
## 5     0
## 6  <NA>
```

Data description

This dataset is an extract of two surveys conducted by the National Science Foundation (NSF) in 2013: the National Survey of College Graduates (SESTAT 2013) and Survey of Doctorate Recipients (SESTAT 2013). Information on the survey and sampling methods can be found here.

https://highered.ipums.org/highered/survey_designs.shtml

The dataset is public and can be cited in your report as: Minnesota Population Center. IPUMS Higher Ed: Version 1.0 [dataset]. Minneapolis, MN: University of Minnesota, 2016. https://doi.org/10.18128/D100.V1.0

On Canvas, you would find the following files:

- **data.formatted.csv**: the dataset downloaded from IPUMS Higher Ed, with missing or logical skips recoded to NA, the error in the variable CHTOT fixed.
- **dataset.RData**: an R workspace that contains data.formatted.csv pre-loaded as a dataframe called `dataset`, with each variable given the correct type.

It is recommended that you start with this file.

For regression you may find it convenient to recode some yes/no variables as a binary 1/0 numeric variable.

- **codebook-basic.txt**: a list of variables and the meaning of their values. Note that missing or logical skips have been recoded to NA.
- **codebook.xml**: an XML version of the codebook, with more detailed explanations on the variables and hyperlinks. You can open this in your browser.
- **final-project.rmd / final-project.pdf**: instructions and questions

The goals of this analysis are following.

- Give a general description of the work landscape for those with a college degree in the US, as surveyed in 2013
- Build a regression model to predict annual salary
- Build a regression model to predict job satisfaction
- Use our analysis to fact-check news outlets.
- Convey our findings in a technical report and in plain terms.

General instructions on formatting

You should hand in two files in total: an rmd file and a pdf file.

However, it should look less like homework and more like a professional report.

A good standard are the PEW research reports, such as this:

https://www.pewsocialtrends.org/2014/02/11/the-rising-cost-of-not-going-to-college/

Here is what the lay summary from that article looks like

https://www.pewresearch.org/fact-tank/2014/02/11/6-key-findings-about-going-to-college/

Please answer all questions asked and write in full sentences with good formatting (eg: clear paragraphs).

For hypothesis testing, use 95% significance level unless otherwise specified.

The Report

Your report should contain the same headings as the sections below. Under each heading, put answers to these questions.

For each question/bullet, summarize in ONE paragraph, with appropriate plots and/or numbers/tables.

Basic analysis.

Population and sampling

1. This dataset consists of two different surveys. Briefly describe the population, the sample, and the sampling method for each of the surveys. Name TWO possible biases that each sample can have. Do we introduce further biases when we analyze the results of these surveys together (ie: treat it as one big dataset)?

National Survey of College Graduates -

Population - The target population included non-institutionalized individuals living in the US under the age of 76 who had completed at least a bachelor's degree by the date of the Census.

Sample - Stratified Random Selection of Census long-form households in 1990 and 2000, and ACS households in 2010, then taking another stratified into groups of age, race, highest degree type, occupation, and sex.

Sampling Method - A two-stage, stratified systematic sampling method chooses between Census long-form households in 1990 and 2000, and ACS households in 2010, then, from long-form or ACS households, uses a stratification sampling scheme using age, race, highest degree type, occupation, and sex.

Bias 1 - Non response bias, meaning that not all households in this sample were responsive to this survey / and or completed a response.

Bias 2 - Selective Bias, in this situation the survey administrators chose who and what kind of people to survey, and that initial screening of who to choose has the possibility to skew the answers.

The Survey of Doctorate Recipients -

Population - Individuals who have earned a doctorate degree in science, engineering, or health in the United States

Sample - The pool of possible respondents comes from a stratified sample from the Survey of Earned Doctorates

Sampling Method - Stratified sampling

Bias 1 - Non response bias, meaning that not all households in this sample were responsive to this survey / and or completed a response - therefore this causes bias.

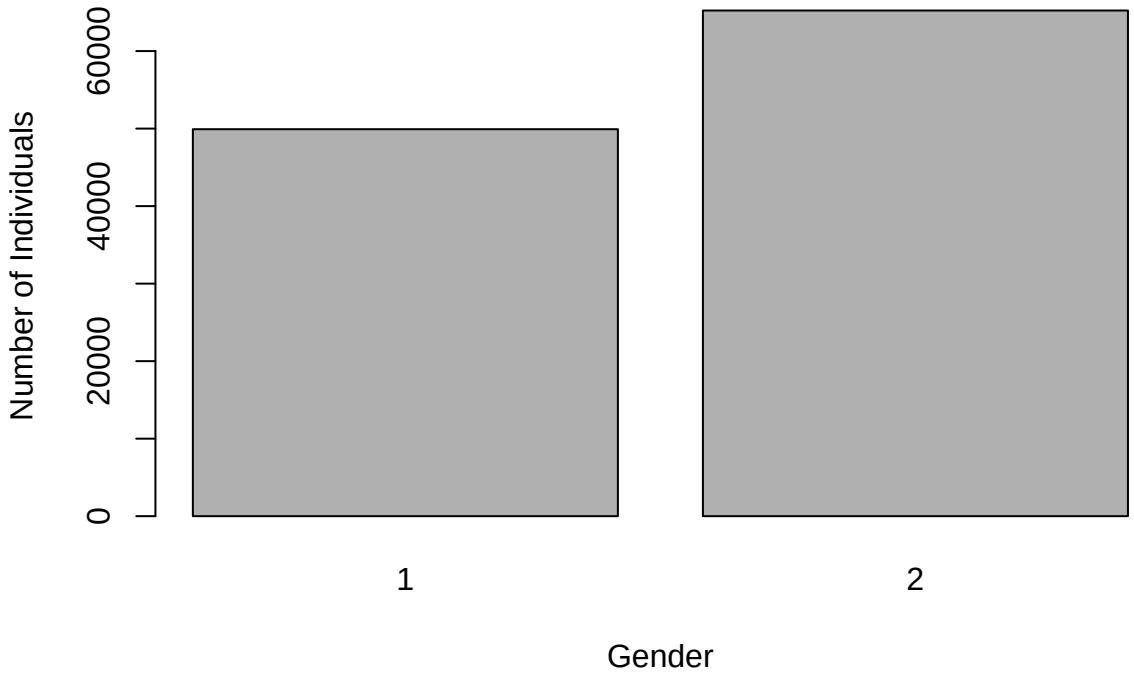
Bias 2 - Selective Bias, in this situation the survey administrators chose who and what kind of people to survey, and that initial screening of who to choose has the possibility to skew the answers.

We do introduce further biases when analyzing the two surveys together as they will provide results from seperate subsets of individuals with different credentials. National Survey of College Graduates simply surveys individuals in the US under the age of 76 who had completed at least a bachelor's degree, while The Survey of Doctorate Recipients surveyed those who have earned a doctorate degree in science, engineering, or health in the United States. Also, the sample in the The Survey of Doctorate Recipients will potentially overlap in the National Survey of college Graduate. Due to the differences in samples, there will be further biases when analyzing the two surveys together, when it comes to looking at the results of how college affects the characteristics of one's life.

Demographics

2. Summarize the demographics of the survey. Specifically, you should describe the distribution of gender, minority, race/ethnicity, and total number of children.

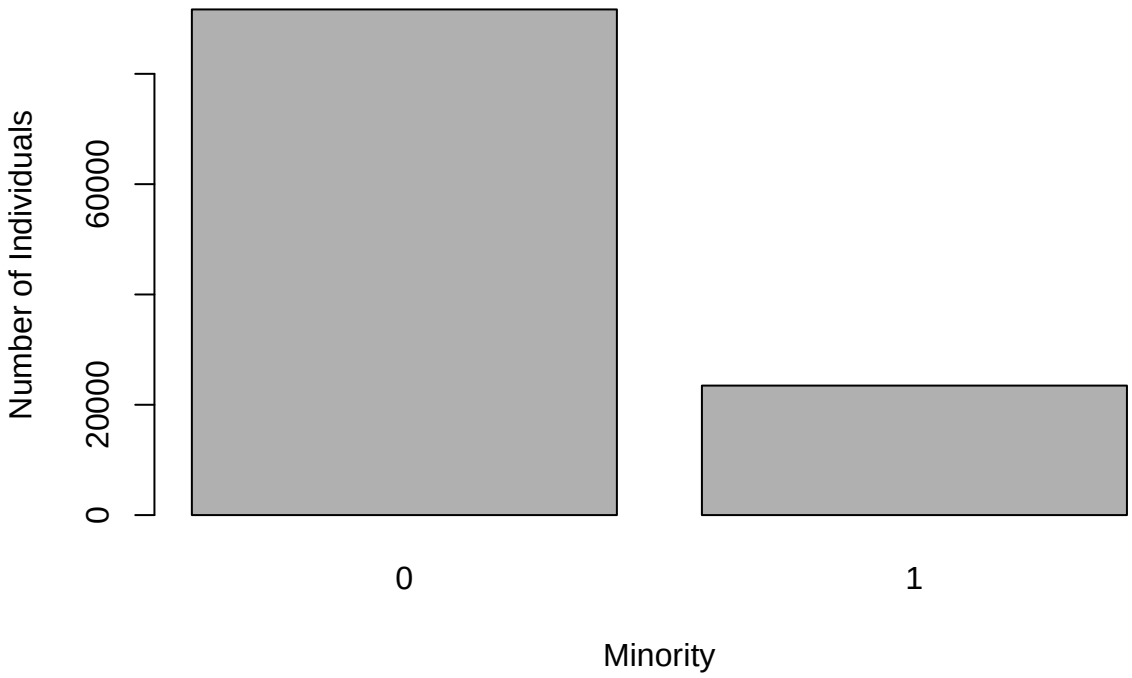
```
gender_plot <- plot(dataset$GENDER, xlab="Gender", ylab="Number of Individuals")
```




```
gender_table <-table(dataset$GENDER)
gender_table
```

```
##
##     1     2
## 49919 65233
```

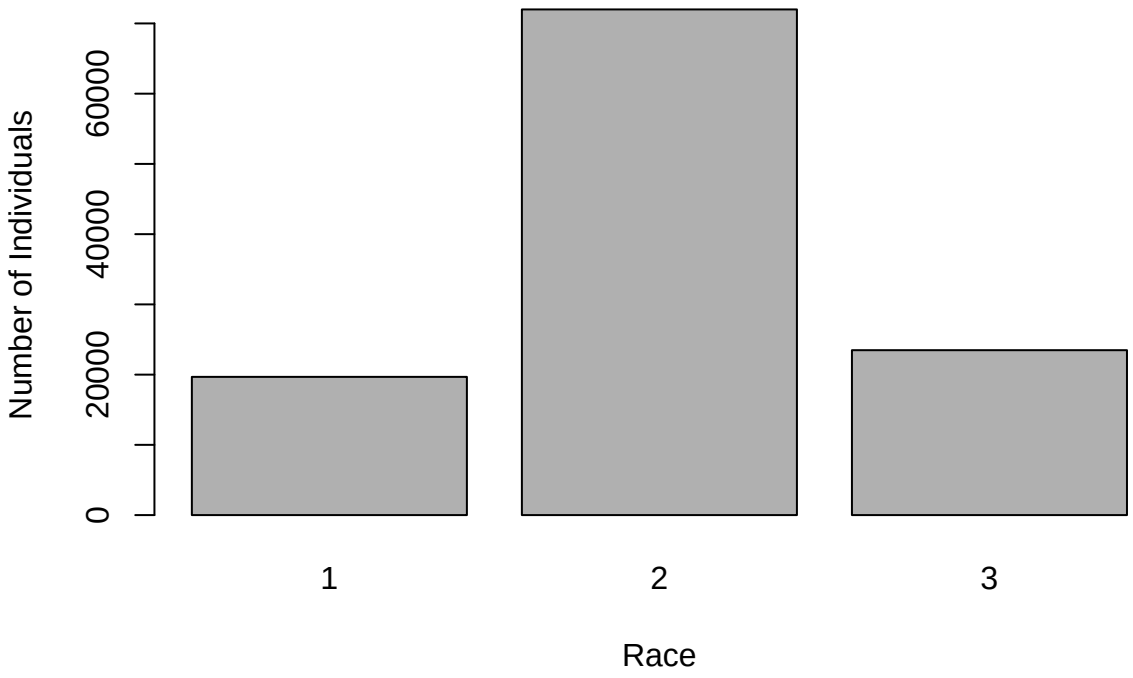
```
minority_plot <- plot(dataset$MINRTY, xlab="Minority", ylab="Number of Individuals")
```




```
minority_table <- table(dataset$MINRTY)
minority_table
```

```
##
##     0     1
## 91674 23478
```

```
race_plot <- plot(dataset$RACETH, xlab="Race", ylab="Number of Individuals")
```




```
race_table <- table(dataset$RACETH)
race_table
```

```
##
##     1     2     3
## 19680 71994 23478
```

Summary of gender - 49919 individuals were of option 1, which was female, while 65233 individuals were of option 2, which was male.

Proportion of Females - .43350 Proportion of Males - .56649

Summary of minority - 91674 individuals were of option 0, which was "NO", while 23478 individuals were of option 1, which was "YES".

Proportion of "NO" Minority - .796112 Proportion of "YES" Minority - .203887

Summary of Race - 19680 individuals were of option 1, which was "Asian", 71994 individuals were of option 2, which was "White", and 23478 individuals were of option 3, which was "Under Represented Minorities"

Proportion of "Asian" - .17090454 Proportion of "White" - .62520842 Proportion of "Under Represented Minorities" - .20388704

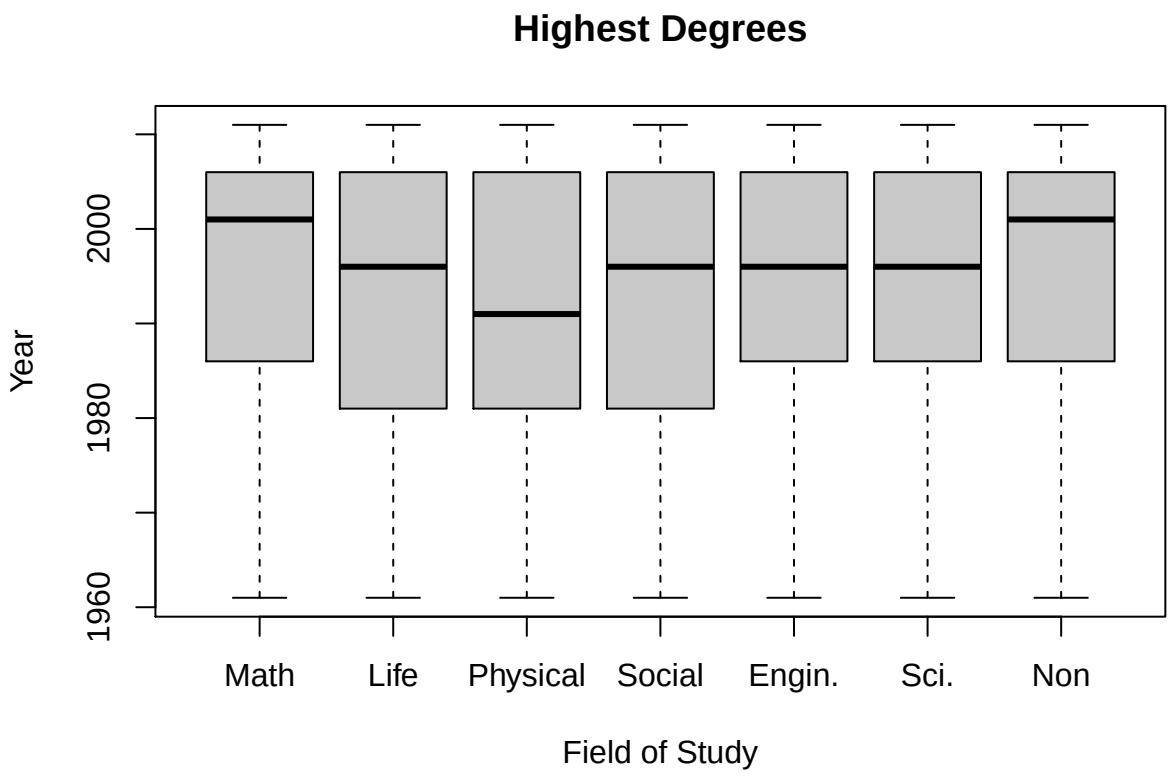
Education

3. Summarize the distribution of highest degrees and bachelor degrees by field and year obtained obtain.

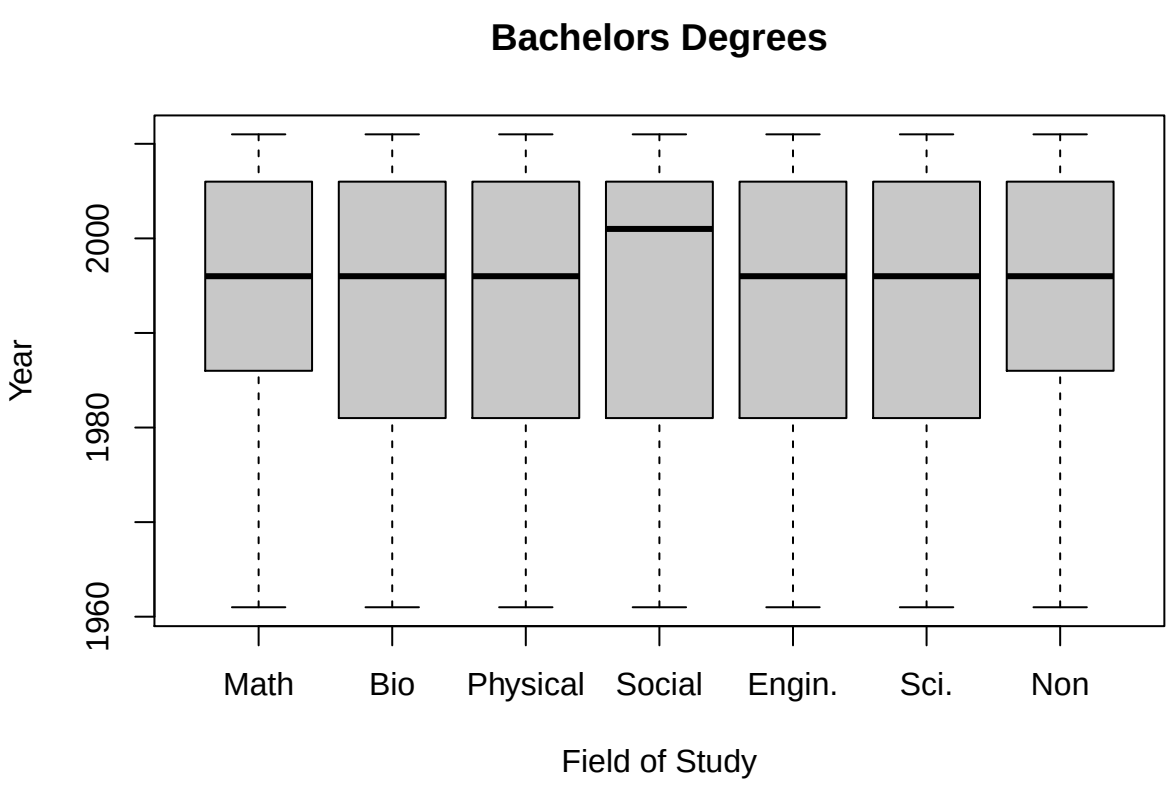
```r
table(x = dataset$HD03Y5, y = dataset$NDGMEMG)
```

```
##       y
## x        1    2    3    4    5    6    7
##   1961  120  144  204  284  396  253   70
##   1966  286  560  583  782  791  452  226
##   1971  458  975  772 1541 1195  794  655
##   1976  471 1174  708 1740 1354 1106  967
##   1981  649 1238  864 1758 1719 1204 1033
##   1986  867 1248  904 1794 1984 1312 1127
##   1991  964 1485  957 2150 2223 1523 1478
##   1996 1179 1930  974 2443 2400 1694 1750
##   2001 1808 2171 1032 3243 2784 2255 2267
##   2006 3127 4358 2198 7359 7404 4595 3504
##   2011  554  756  450 1197 1201 1426 1551
```

```r
plot(x = dataset$NDGMEMG, y = dataset$HD03Y5, main = "Highest Degrees", xlab="Field of Study", ylab="Ye
```




```r
bachelor_deg <- subset(dataset, DGRDG == 1)
plot(x=bachelor_deg$NDGMEMG, y = bachelor_deg$HD03Y5,main = "Bachelors Degrees", xlab="Field of Study",
```



In terms of Highest Degrees - For the categories 1,5,6 and 7, 50 percent of the data that is recorded of degree paths inbetween 1961 and 2011, are concentrated at later years (1990 onward). Therefore this can show that majors 1,5,6, and 7 have shown more popularity in recent years. Category 1,5,6 and 7 represent computer and mathematic sciences, Engineerings, Science and Engineering Related Fields, and non science fields. On the other hand, 50 percent of the data that is recorded between 1961 and 2011, for majors 2,3 and 4 have a wider quartile range. These majors of category 2,3 and 4 are Biological, agricultural and environmental life sciences, physical and related sciences, and Social and related sciences. Therefore, this shows that these majors have had significant popularity in earlier years (before 1990). Physical and Related Science had a median close to 1990, while computer and mathematic science had a median closer to 2000. This shows that 2,3, and 4 were more populated earlier than 1,5,6,7.

In terms of all Bachelors Degrees - For the categories 1 and 7, 50 percent of the data that is recorded of highest degree paths inbetween 1961 and 2011, are concentrated at later years (1985 Onwards), in terms of quartile range. Therefore this can show that majors 1 and 7 have shown more of their respective popularity in recent years. Category 1 and 7 represent computer and mathematic sciences and non science and engineering fields. On the other hand, 50 percent of the data that is recorded between 1961 and 2011, for majories 2,3 and 4, 5 and 6 have a wider quartile range. Therefore, this shows that these majors have had significant and independent popularity in earlier years in terms of bachelors degrees. These majors of category 2,3 and 4, 5 and 6 are Biological, agricultural and environmental life sciences, physical and related sciences, Social and related sciences, Engineering, and Science and engineering-related fields. 4 has a higher median, therefore that shows that it's popularity has grown over time. Category 4 is Social and related Sciences.

4. The retention rate of a field is the rate at which people with a bachelor degree in this field would do a higher degree in the same field. Is there a significant difference in retention rates among different field of majors?

Null Hypothesis - There is no significant difference in retention rates among different field of majors.

Alternative Hypothesis - There is a significant difference in retention rates among different field of majors.

```
subset1 <- dataset
subset1$NBAMEMG[subset1$NBAMEMG == "96" | subset1$NBAMEMG == "9"] <- NA
subset1$NBAMEMG <- droplevels(subset1$NBAMEMG)
subset <- subset1[subset1$DGRDG != 1,]


#Compute intial D
data.deviation <- function(){

  retention1 <- sum(subset$NBAMEMG == 1 & subset$NDGMEMG == subset$NBAMEMG, na.rm = TRUE) / sum(subset$

  retention2 <- sum(subset$NBAMEMG == 2 & subset$NDGMEMG == subset$NBAMEMG, na.rm = TRUE) / sum(subset$

  retention3 <- sum(subset$NBAMEMG == 3 & subset$NDGMEMG == subset$NBAMEMG, na.rm = TRUE) / sum(subset$

  retention4 <- sum(subset$NBAMEMG == 4 & subset$NDGMEMG == subset$NBAMEMG, na.rm = TRUE) / sum(subset$

  retention5 <- sum(subset$NBAMEMG == 5 & subset$NDGMEMG == subset$NBAMEMG, na.rm = TRUE) / sum(subset$

  retention6 <- sum(subset$NBAMEMG == 6 & subset$NDGMEMG == subset$NBAMEMG, na.rm = TRUE) / sum(subset$

  retention7 <- sum(subset$NBAMEMG == 7 & subset$NDGMEMG == subset$NBAMEMG, na.rm = TRUE) / sum(subset$

  average <- (retention1 + retention2 + retention3 + retention4 + retention5 + retention6 + retention7)

  D <- abs(average - retention1) + abs(average - retention2) + abs(average - retention3) + abs(average -

  return(D)
}

#--- permutation test: shuffle the highest degree column, and then compute d.

shuffle <-function(){
  subset$shuffle <- sample(subset$NDGMEMG)

  ret1 <- sum(subset$NBAMEMG == 1 & subset$shuffle == subset$NBAMEMG, na.rm = TRUE) / sum(subset$NBAMEMG

  ret2 <- sum(subset$NBAMEMG == 2 & subset$shuffle == subset$NBAMEMG, na.rm = TRUE) / sum(subset$NBAMEMG

  ret3 <- sum(subset$NBAMEMG == 3 & subset$shuffle == subset$NBAMEMG, na.rm = TRUE) / sum(subset$NBAMEMG

  ret4 <- sum(subset$NBAMEMG == 4 & subset$shuffle == subset$NBAMEMG, na.rm = TRUE) / sum(subset$NBAMEMG

  ret5 <- sum(subset$NBAMEMG == 5 & subset$shuffle == subset$NBAMEMG, na.rm = TRUE) / sum(subset$NBAMEMG

  ret6 <- sum(subset$NBAMEMG == 6 & subset$shuffle == subset$NBAMEMG, na.rm = TRUE) / sum(subset$NBAMEMG

  ret7 <- sum(subset$NBAMEMG == 7 & subset$shuffle == subset$NBAMEMG, na.rm = TRUE) / sum(subset$NBAMEMG

  average <- (ret1 + ret2 + ret3 + ret4 + ret5 + ret6 + ret7) / 7

  difference <- abs(average - ret1) + abs(average - ret2) + abs(average - ret3) + abs(average - ret4) +
```

```
  return(difference)

}

#run many times
m <- 10
d <- replicate(m,shuffle())

#calculate the actual data deviation
D <- data.deviation()
p.value <- sum(d>=D)/m
p.value
```

```
## [1] 0
```

We are using a permutation test to make our conclusion. With this test we are getting our P value. The p value is 0. Therefore it is less than our significance level of 0.05. Therefore we can reject the null hypothesis. Hence, we can conclude that there is a significant difference in retention rates among different field of majors.

Job status

5. What does the labor force look like?

- Describe general statistics: % of people working, % working part-time, number of hours per week and number of weeks per year.

```
table.employed <- table(dataset$LFSTAT)
prop.table(table.employed)*100
```

```
##
##         1         2         3
## 85.149194  2.930909 11.919897
```

```
#The percentage of people working is roughly 85.15%

table.partTime <- table(dataset$HRSWKGR)
prop.table(table.partTime)*100
```

```
##
##         1         2         3         4
##  7.495079  8.271206 37.754842 46.478873
```

```
#The percentage of people working part time, where part time is defined as 35 hours or less, is approxi
#Approximately 37.75% of the sample works 36-40 hours per week, and another 46.47% work overtime.

table.wksPerYear <- table(dataset$WKSWKGR)
prop.table(table.wksPerYear)*100
```

```
##
##          1          2          3          4
##  0.6486420  0.7689876  6.3150809 92.2672895
```

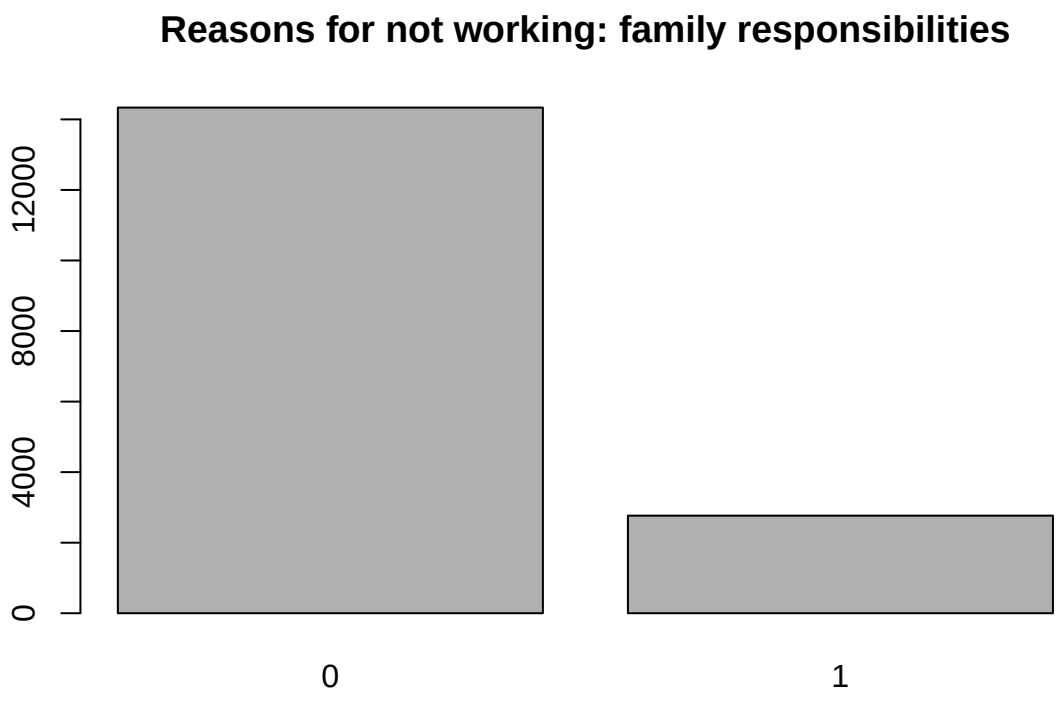
```
#.648% work 1-10 weeks a year, .768% work 11-20 weeks a year, 6.31% work 21-39 weeks a year, and 92.267
```

- Do most people work in short bursts (few weeks but high number of hours per week), or do most people work with regular hours year-round?

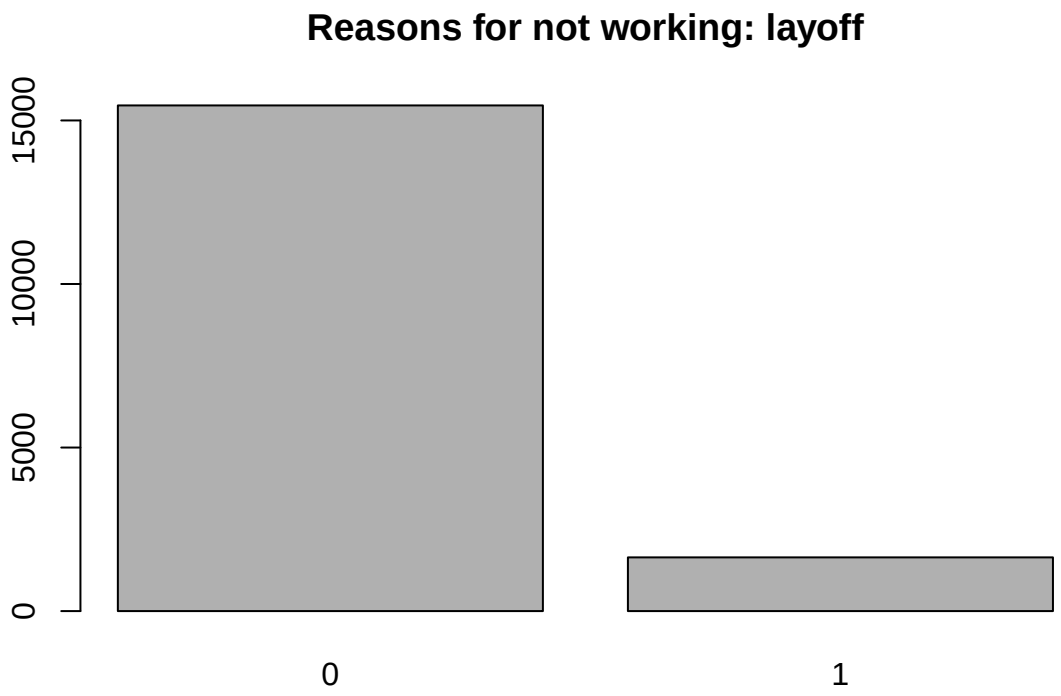
According to the data above, most people work overtime hours of above 40 hours per week. This is shown by the data point showoing that 15% of people work 35 hours or less, while 37.75% of people work normal hours, and 46.47% work overtime. Also, people work with regular hours year-round, as 92.267% of people work 40-52 weeks a year.

- What are the major reasons that led people to not work at the time of survey?

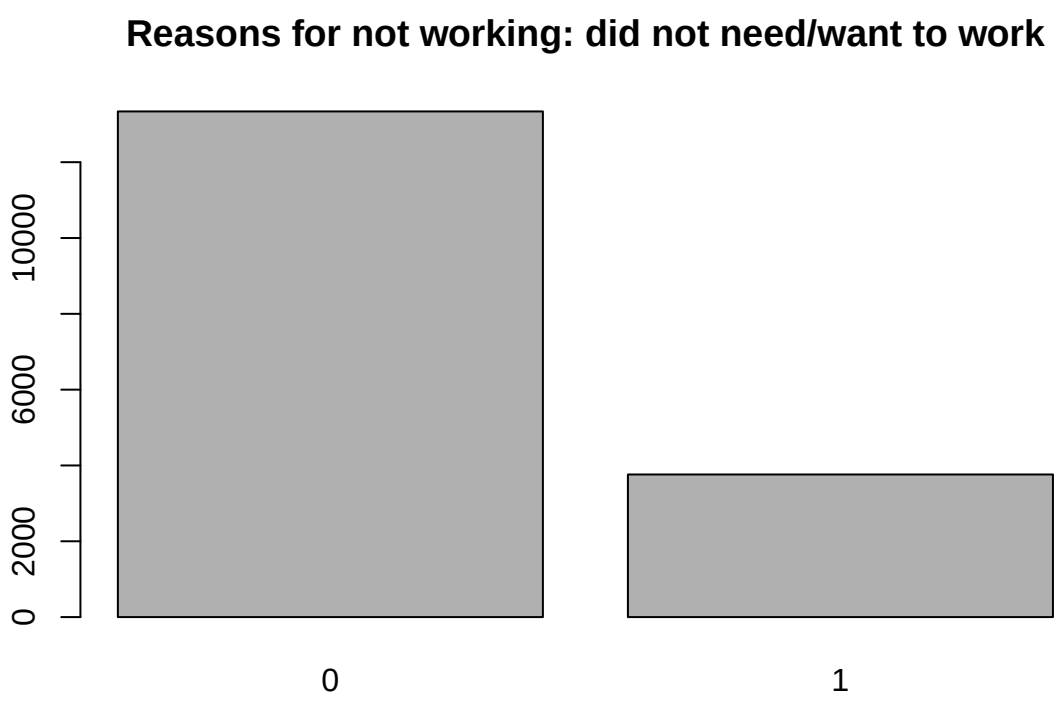
```
library(ggplot2)
plot(dataset$NWFAM, main = "Reasons for not working: family responsibilities")
```



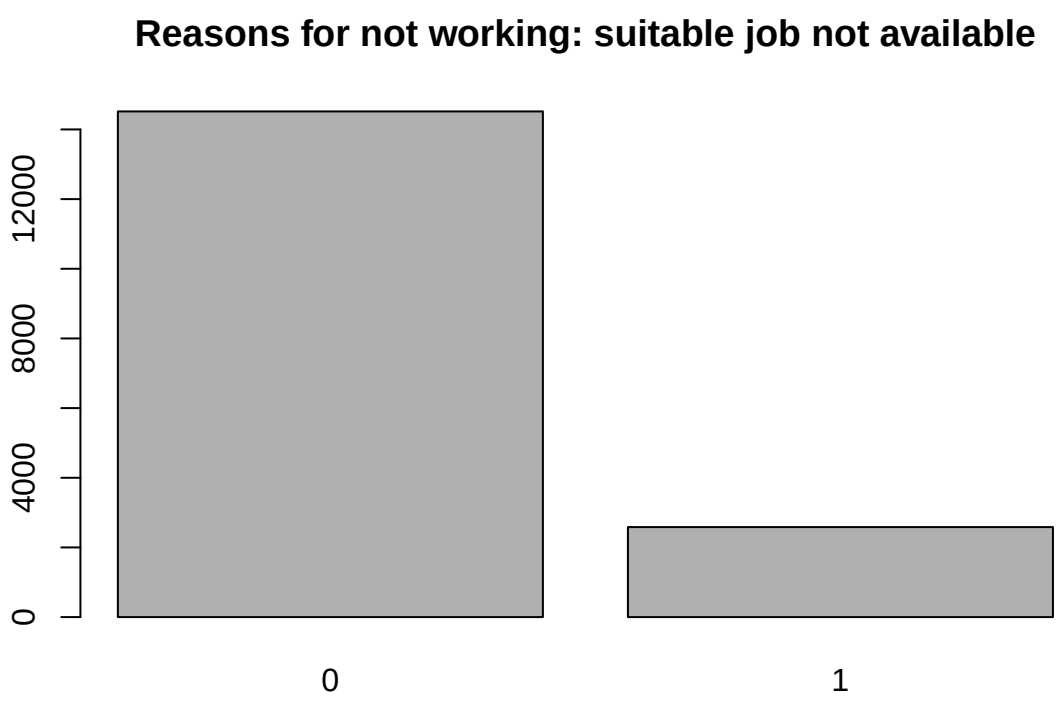

```
plot(dataset$NWLAY, main = "Reasons for not working: layoff")
```



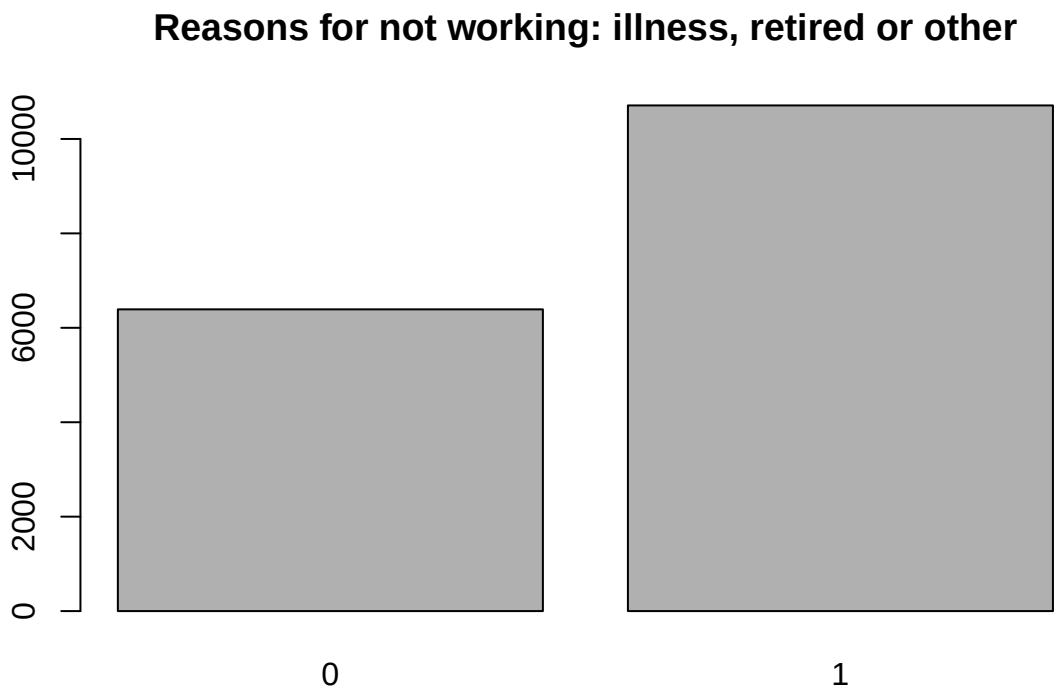

```
plot(dataset$NWNOND, main = "Reasons for not working: did not need/want to work")
```



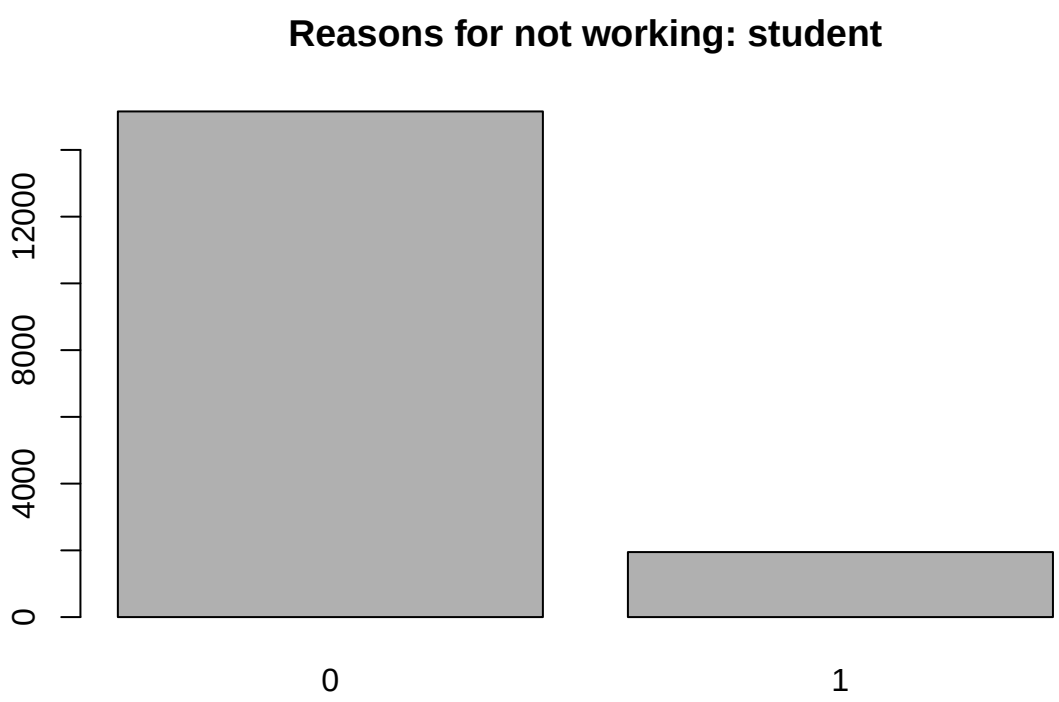

```
plot(dataset$NWOCNA, main = "Reasons for not working: suitable job not available")
```




```
plot(dataset$NWOTP, main = "Reasons for not working: illness, retired or other")
```




```
plot(dataset$NWSTU, main = "Reasons for not working: student")
```



Upon looking at the different distributions plotted above for why people were not working at the time of the survey, it becomes apparent that the main reason was due to "Illness, retired or other". Hence, most people were dealing with illness at the time of this survey.

6. Degree relevance

- How relevant are the people's degree to their principle job? (Do people work in the field that they were trained for, or do they work in unrelated areas?).

```
table.relevance <- table(dataset$OCEDRLP)
prop.table(table.relevance)*100
```

```
##
##        1        2        3
## 62.82649 24.88195 12.29156
```

Using the prop.table above, we are able to see the proportion of job relevance that is "closely related to their principle job" has a percentage of 62.8%. Therefore there is enough data to assume that there is a significant relevance to people's degree with their principle job.

- Is there a statistically significant difference in relevance of degree vs
  - job type
  - the degree that they are trained for, and
  - the type of job that people do?

Note: state the tests you use, p-value and draw conclusions. You may find the variables MGRNAT, MGROTH, MGRSOC, NOCPRMG, OCEDRLP, NDGMEMG, WAPRSM and WASCSM relevant.

H Null - There is no statistical difference in relevance of degree vs job type H Alternative - There is a statistical difference in relevance of degree vs job type

H Null - There is no statistical difference in relevance of degree vs the degree that they are trained for H Alternative - There is a statistical difference in relevance of degree vs the degree that they are trained for

H Null - There is no statistical difference in relevance of degree vs the type of job that people do H Alternative - There is a statistical difference in relevance of degree vs the type of job that people do

```
#Add OCEDRLP_UPDATED - Recode for OCEDRLP
dataset$OCEDRLP_UPDATED <- 0
dataset$OCEDRLP_UPDATED[dataset$OCEDRLP[] == 1 | dataset$OCEDRLP[] ==2 ] <- 1

#Chisq test for relevance of degree vs job type
chisq.test(dataset$OCEDRLP_UPDATED, dataset$NOCPRMG)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dataset$OCEDRLP_UPDATED and dataset$NOCPRMG
## X-squared = 8834.8, df = 6, p-value < 2.2e-16
```

```
#Chisq test for relevance of degree vs the degree that they are trained for
chisq.test(dataset$OCEDRLP_UPDATED, dataset$NDGMEMG)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dataset$OCEDRLP_UPDATED and dataset$NDGMEMG
## X-squared = 2071.6, df = 6, p-value < 2.2e-16
```

```
#Chisq test for relevance of degree vs the type of job that people do
chisq.test(dataset$OCEDRLP_UPDATED, dataset$WAPRSM)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dataset$OCEDRLP_UPDATED and dataset$WAPRSM
## X-squared = 4181.4, df = 4, p-value < 2.2e-16
```

To find these values we used a chi-squared test. All three p values for each test are much below the significance level of 0.05. (2.2e-16)

Hence reject all of the Null Hypothesis above.

Hence,

There is a statistical difference in relevance of degree vs job type There is a statistical difference in relevance of degree vs the degree that they are trained for There is a statistical difference in relevance of degree vs the type of job that people do

7. Job satisfaction

- Summarize overall job satisfaction

```
subset3 <- dataset
subset3$SATISFACTION[subset3$JOBSATIS == 1 | subset3$JOBSATIS == 2] <- 0
subset3$SATISFACTION[subset3$JOBSATIS == 3 | subset3$JOBSATIS == 4] <- 1
table.satisfaction <- table(subset3$SATISFACTION)
prop.table(table.satisfaction)*100
```

```
##
##        0        1
## 89.39735 10.60265
```

Looking at the proportion of Job Satisfaction, the proportion table above shows the columns "0" and "1".

"0" stands for those who are "Very Satisfied / Somewhat Satisfied" and "1" stands for those who are "Somewhat Dissatisfied / very dissatisfied"

The table shows that there is an 89.43% proportion of those who are "Very Satisfied / Somewhat Satisfied" while there is a 10.56% proportion of those who are "Somewhat Dissatisfied / very dissatisfied". Therefore this can be summarized to say that a much larger proportion of respondants believe that they are satisfied with their job.

- Among those who reported "somewhat/very satisfied", which aspects of their jobs are they most satisfied with? Among those who reported "somewhat/very dissatisfied", which aspects of their jobs are they least satisfied with?

Those who reported "Somewhat / Very Satisfied", they are most satisfied with Independence, Location, Social Impact.

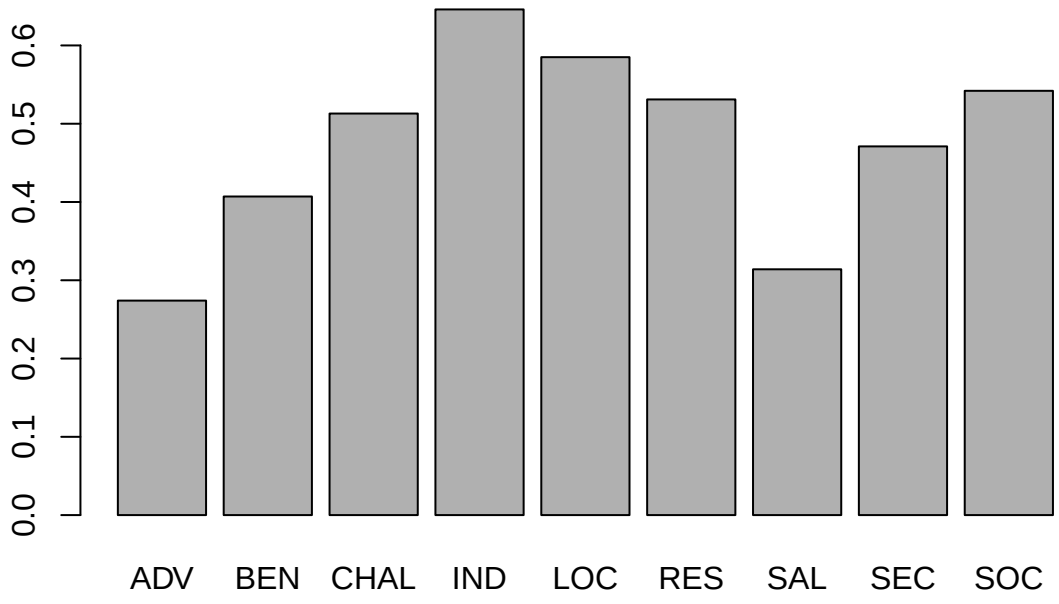
Those who reported "Somewhat / Very Dissatisfied", they are most dissatisfied with their career advancement, salary, and intellectual challenge.

- Base on the above, which factors are most important to job satisfaction?

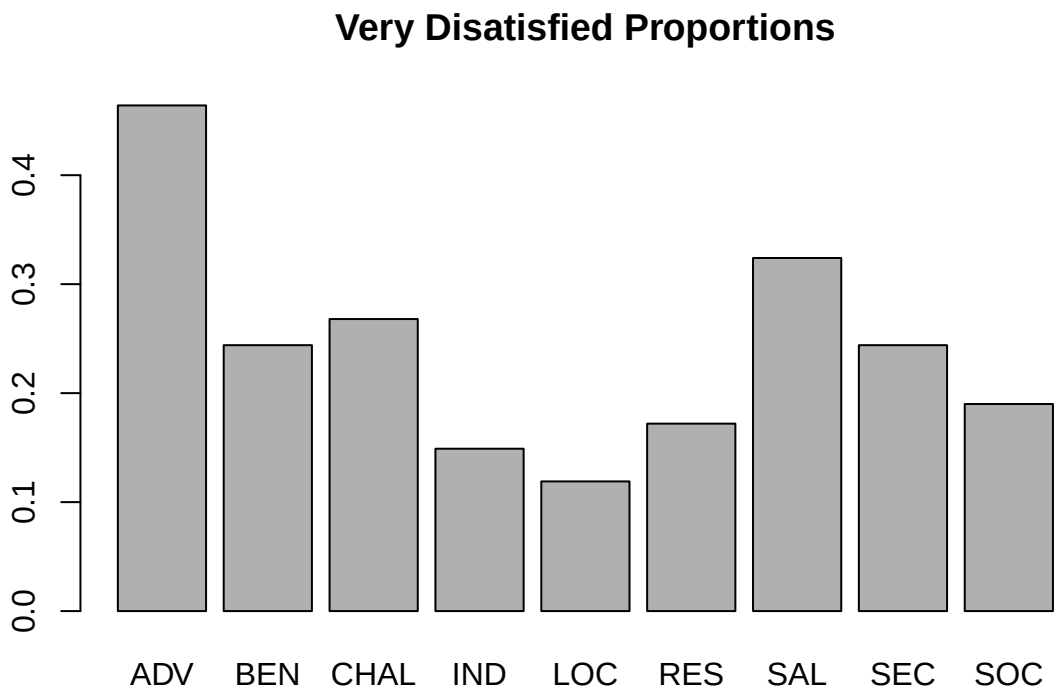
The most important factors to job satisfaction are in the order of importance :

Independence, Location, Social Impact, Responsibility, Security, Benefits, Salary, Career Advancement.

```
#Those who reported Satisfied / Somewhat Satisfied
dataset.satified <- subset(dataset , dataset$JOBSATIS == 1 | dataset$JOBSATIS == 2)

# table(dataset.satified$SATADV)/length(dataset.satified$SATADV)
# table(dataset.satified$SATBEN)/length(dataset.satified$SATBEN)
# table(dataset.satified$SATCHAL)/length(dataset.satified$SATCHAL)
# table(dataset.satified$SATIND)/length(dataset.satified$SATIND)
# table(dataset.satified$SATLOC)/length(dataset.satified$SATLOC)
# table(dataset.satified$SATRESP)/length(dataset.satified$SATRESP)
# table(dataset.satified$SATSAL)/length(dataset.satified$SATSAL)
# table(dataset.satified$SATSEC)/length(dataset.satified$SATSEC)
# table(dataset.satified$SATSOC)/length(dataset.satified$SATSOC)


satisfaction.categories <- c("ADV" , "BEN", "CHAL", "IND", "LOC", "RES", "SAL", "SEC","SOC")
satisfaction.proportion <- c(0.274, 0.407, 0.513, 0.646, 0.585, 0.531, 0.314, 0.471, 0.542)
barplot(satisfaction.proportion, names.arg = satisfaction.categories, main = "Very Satisfied Proportions
```




```r
#Those who reported very / somewhat dissatisfied
dataset.notSatified <- subset(dataset , dataset$JOBSATIS == 3 | dataset$JOBSATIS == 4)

# table(dataset.notSatified$SATADV)/length(dataset.notSatified$SATADV)
# table(dataset.notSatified$SATBEN)/length(dataset.notSatified$SATBEN)
# table(dataset.notSatified$SATCHAL)/length(dataset.notSatified$SATCHAL)
# table(dataset.notSatified$SATIND)/length(dataset.notSatified$SATIND)
# table(dataset.notSatified$SATLOC)/length(dataset.notSatified$SATLOC)
# table(dataset.notSatified$SATRESP)/length(dataset.notSatified$SATRESP)
# table(dataset.notSatified$SATSAL)/length(dataset.notSatified$SATSAL)
# table(dataset.notSatified$SATSEC)/length(dataset.notSatified$SATSEC)
# table(dataset.notSatified$SATSOC)/length(dataset.notSatified$SATSOC)

not.satisfaction.categories <- c("ADV" , "BEN", "CHAL", "IND", "LOC", "RES", "SAL", "SEC","SOC")
not.satisfaction.proportion <- c(0.464, 0.244, 0.268, 0.149, 0.119, 0.172, 0.324, 0.244, 0.190)
barplot(not.satisfaction.proportion, names.arg = not.satisfaction.categories, main = "Very Disatisfied
```



Regression 1: SALARY vs other variables

Build a linear regression model to predict SALARY based on the other relevant variables.

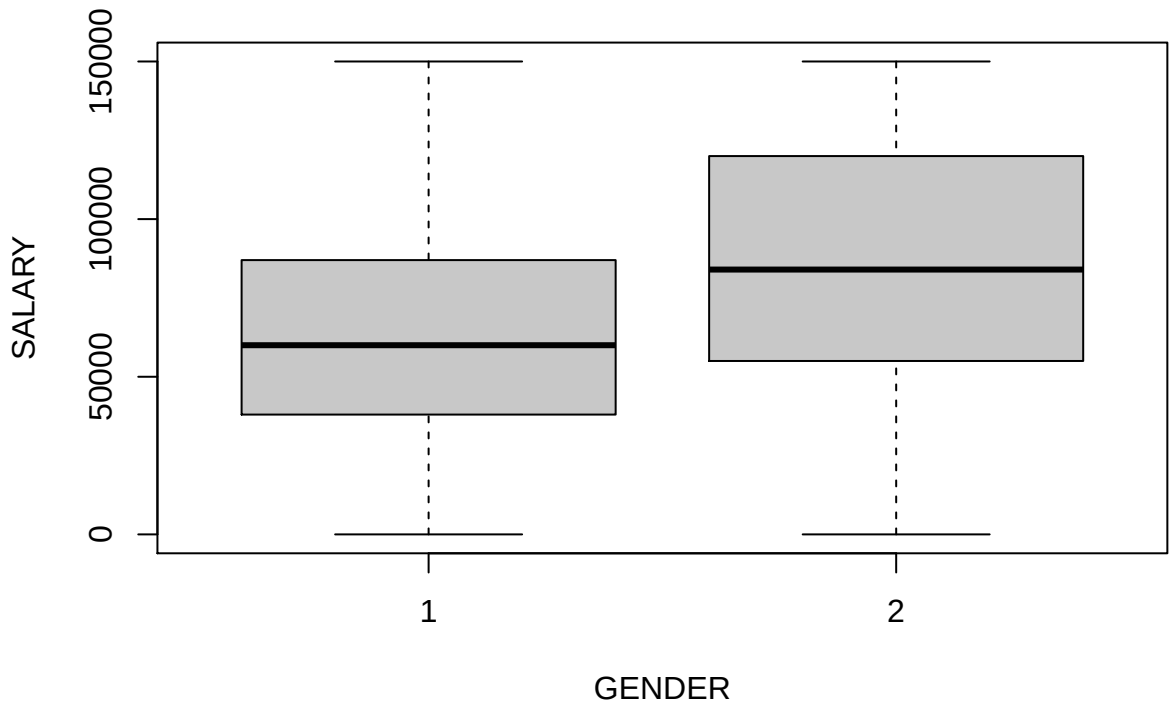
1. Detail how you did variable selection: which models did you run, why did you discard certain models or variables, any variable transformations or recoding you did and why, which diagnostic tests did you run and what they showed, justifications if you removed outliers. How did you decide to deal with missing values in this dataset?

```
library(MASS)
```

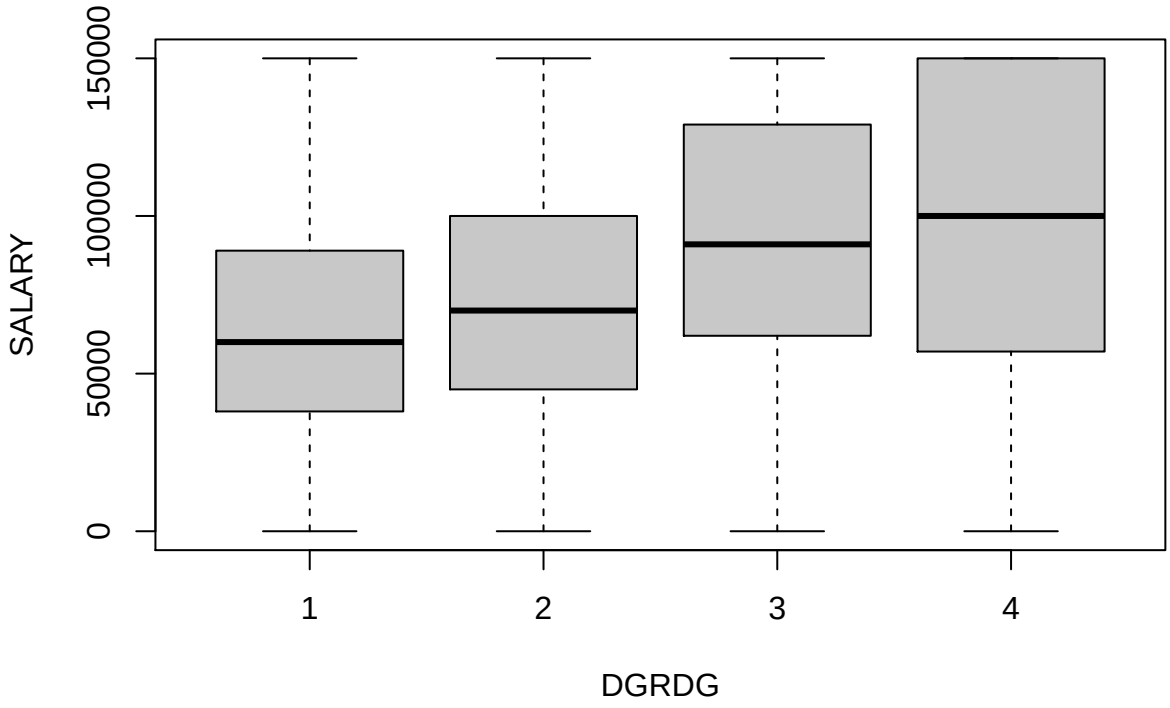
```
#Remove Outliers
dataset <- subset(dataset, LFSTAT == 1)
dataset <- subset(dataset, (EMSEC == 1 & SALARY < 125000) | (EMSEC == 1 & SALARY > 0 & SALARY < 2000) |
dataset <- subset(dataset, (SATSAL == 4 & SALARY < 125000) | SATSAL == 1 | SATSAL == 2 | SATSAL == 3)
dataset <- subset(dataset, (HRSWKGR == 1 & SALARY < 80000) | HRSWKGR == 2 | HRSWKGR == 3 | HRSWKGR == 4

#p1 <- pairs(SALARY ~ WEIGHT, data = dataset)
# p2 <- pairs(SALARY ~ SAMPLE, data = dataset)
# p3 <- pairs(SALARY ~ SURID, data = dataset)
# p4 <- pairs(SALARY ~ AGE, data = dataset)
# p5 <- pairs(SALARY ~ GENDER, data = dataset)
# p6 <- pairs(SALARY ~ MINRTY, data = dataset)
# p7 <- pairs(SALARY ~ RACETH, data = dataset)
```

```
# p8 <- pairs(SALARY ~ CHU2IN, data = dataset)
# p9 <- pairs(SALARY ~ CH25IN, data = dataset)
# p10 <- pairs(SALARY ~ CH611IN, data = dataset)
# p11 <- pairs(SALARY ~ CH1218IN, data = dataset)
# p12 <- pairs(SALARY ~ CH19IN, data = dataset)
# p13 <- pairs(SALARY ~ BA03Y5, data = dataset)
# p14 <- pairs(SALARY ~ NBAMEMG, data = dataset)
# p15 <- pairs(SALARY ~ BADGRUS, data = dataset)
# p16 <- pairs(SALARY ~ DGRDG, data = dataset)
# p17 <- pairs(SALARY ~ HD03Y5, data = dataset)
# p18 <- pairs(SALARY ~ NDGMEMG, data = dataset)
# p19 <- pairs(SALARY ~ HDDGRUS, data = dataset)
# p20 <- pairs(SALARY ~ LFSTAT, data = dataset)
# p21 <- pairs(SALARY ~ HRSWKGR, data = dataset)
# p22 <- pairs(SALARY ~ WKSWKGR, data = dataset)
# p23 <- pairs(SALARY ~ JOBINS, data = dataset)
# p24 <- pairs(SALARY ~ JOBPENS, data = dataset)
# p25 <- pairs(SALARY ~ JOBPROFT, data = dataset)
# p26 <- pairs(SALARY ~ JOBVAC, data = dataset)
# p27 <- pairs(SALARY ~ LOOKWK, data = dataset)
# p28 <- pairs(SALARY ~ FTPRET, data = dataset)
# p29 <- pairs(SALARY ~ PTWTFT, data = dataset)
# p30 <- pairs(SALARY ~ PTFAM, data = dataset)
# p31 <- pairs(SALARY ~ PTNOND, data = dataset)
# p32 <- pairs(SALARY ~ PTOCNA, data = dataset)
# p33 <- pairs(SALARY ~ PTOTP, data = dataset)
# p34 <- pairs(SALARY ~ OCEDRLP, data = dataset)
# p35 <- pairs(SALARY ~ NOCPRMG, data = dataset)
# p36 <- pairs(SALARY ~ EMSEC, data = dataset)
# p37 <- pairs(SALARY ~ WAPRSM, data = dataset)
# p38 <- pairs(SALARY ~ WASCSM, data = dataset)
# p39 <- pairs(SALARY ~ NRREA, data = dataset)
# p40 <- pairs(SALARY ~ NRSEC, data = dataset)
# p41 <- pairs(SALARY ~ JOBSATIS, data = dataset)
# p42 <- pairs(SALARY ~ SATADV, data = dataset)
# p43 <- pairs(SALARY ~ SATBEN, data = dataset)
# p44 <- pairs(SALARY ~ SATCHAL, data = dataset)
# p45 <- pairs(SALARY ~ SATIND, data = dataset)
# p46 <- pairs(SALARY ~ SATLOC, data = dataset)
# p47 <- pairs(SALARY ~ SATRESP, data = dataset)
# p48 <- pairs(SALARY ~ SATSAL, data = dataset)
# p49 <- pairs(SALARY ~ SATSEC, data = dataset)
# p50 <- pairs(SALARY ~ SATSOC, data = dataset)
# p51 <- pairs(SALARY ~ MGRNAT, data = dataset)
# p52 <- pairs(SALARY ~ MGROTH, data = dataset)
# p53 <- pairs(SALARY ~ MGRSOC, data = dataset)
# p54 <- pairs(SALARY ~ NWFAM, data = dataset)
# p55 <- pairs(SALARY ~ NWLAY, data = dataset)
# p56 <- pairs(SALARY ~ NWNOND, data = dataset)
# p57 <- pairs(SALARY ~ NWOCNA, data = dataset)
# p58 <- pairs(SALARY ~ NWOTP, data = dataset)
# p59 <- pairs(SALARY ~ NWSTU, data = dataset)
```

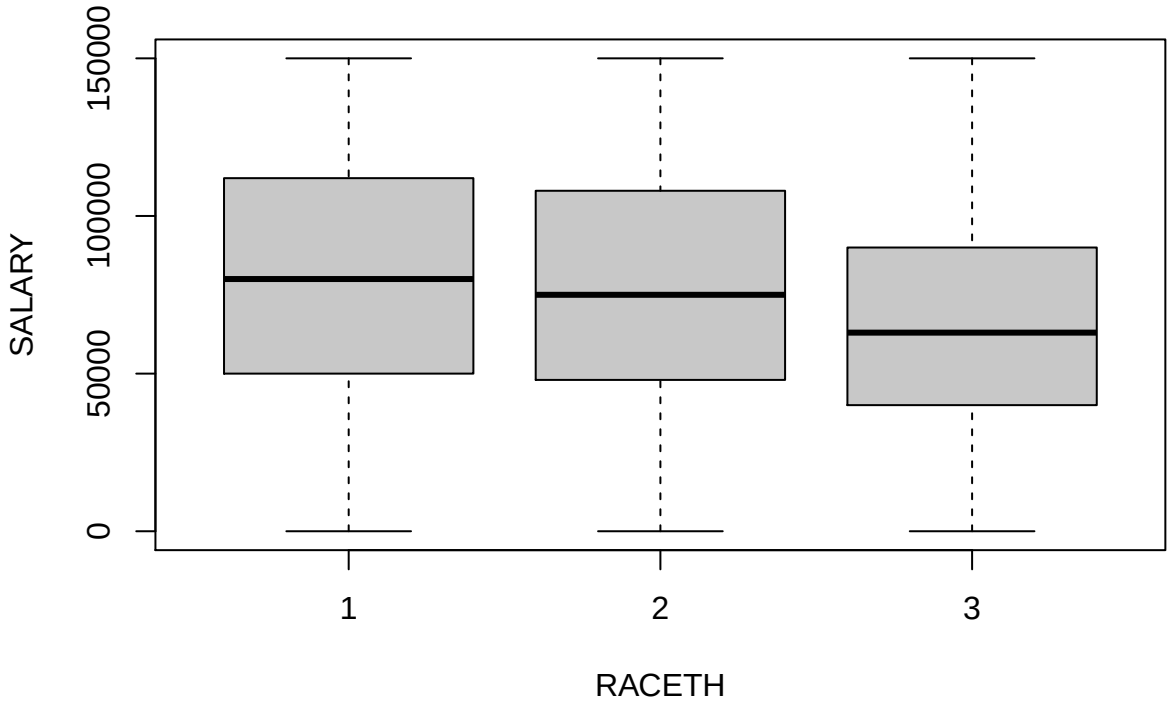
```r
total_lm <- lm(SALARY ~ WEIGHT + AGE + GENDER + MINRTY + RACETH + CHU2IN + CH25IN + CH611IN + CH1218IN +

final_model <- lm(SALARY ~ AGE + as.factor(GENDER) + as.factor(DGRDG) + as.factor(RACETH) + as.factor(N

boxplot(SALARY ~ GENDER, data = dataset)
```



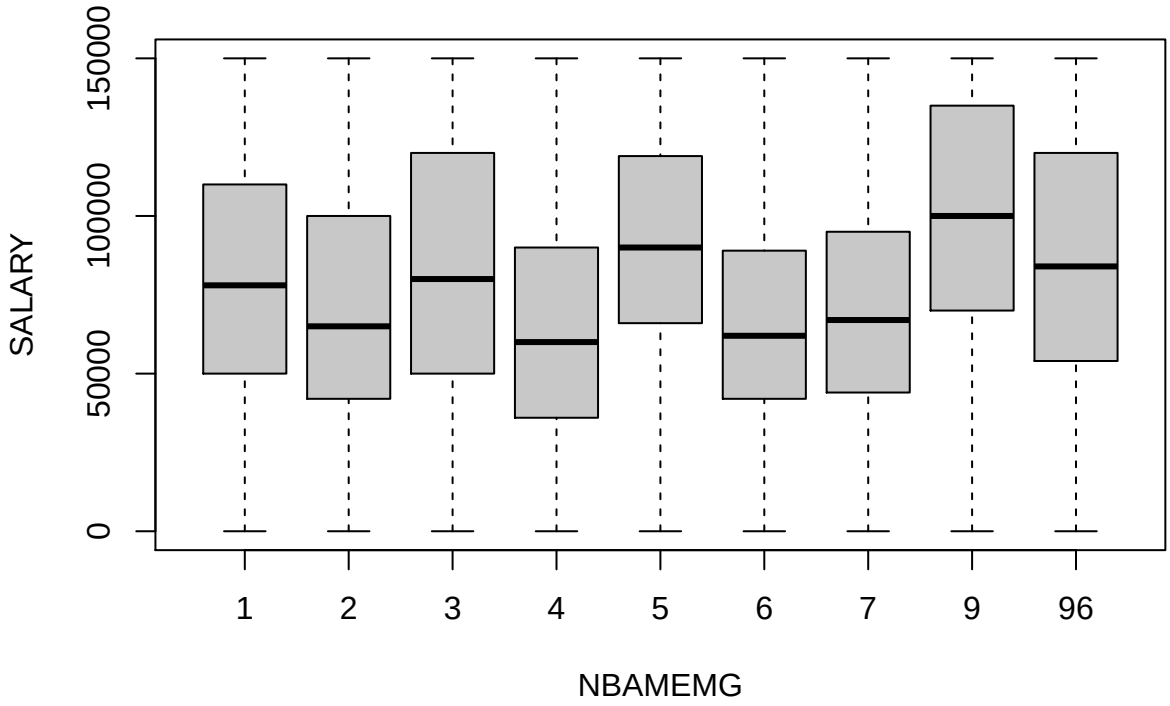

```r
boxplot(SALARY ~ DGRDG, data = dataset)
```



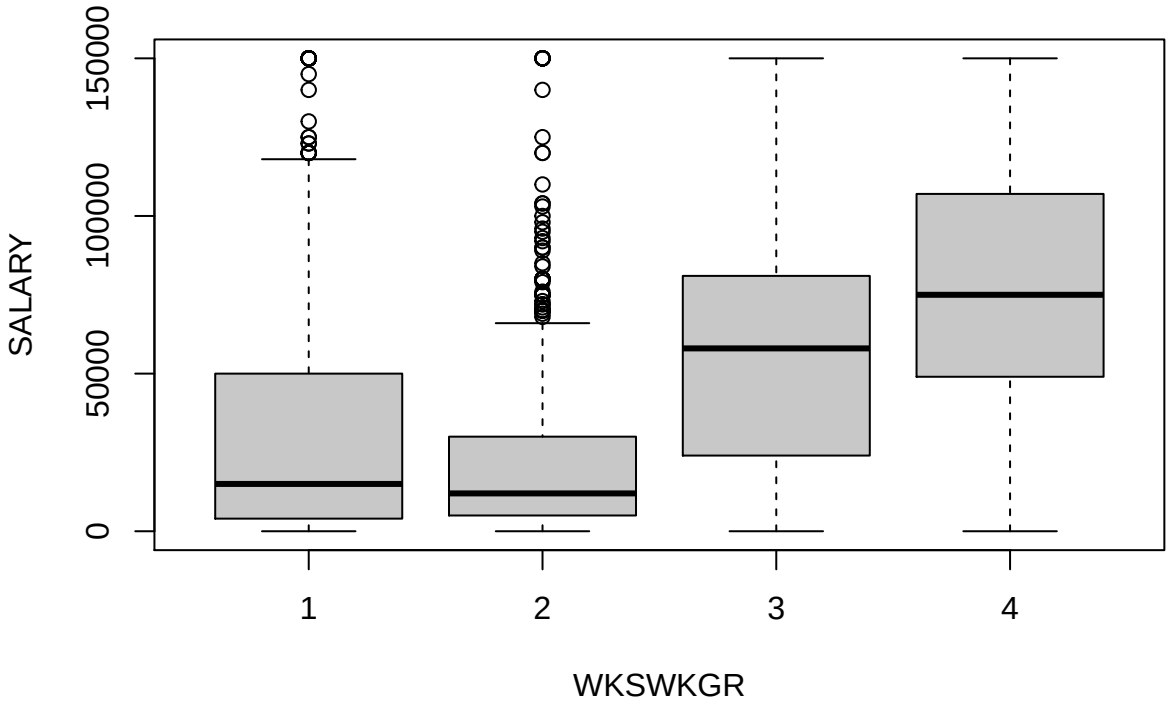

```
boxplot(SALARY ~ RACETH, data = dataset)
```



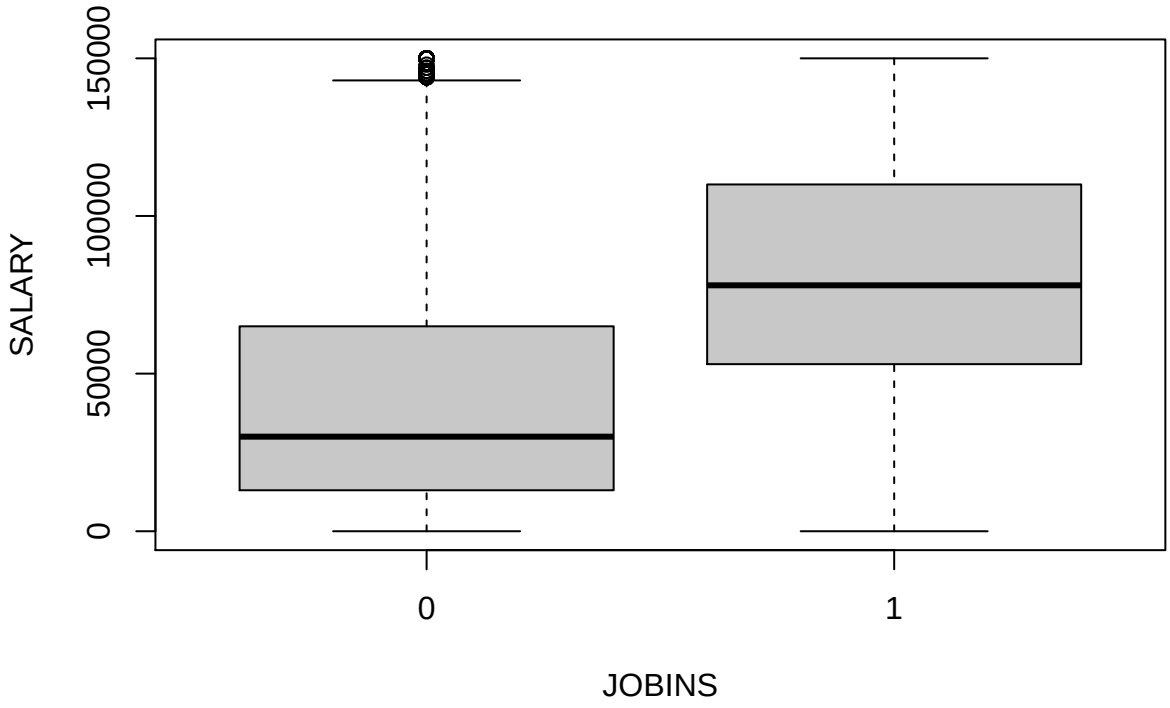

```
boxplot(SALARY ~ NBAMEMG, data = dataset)
```



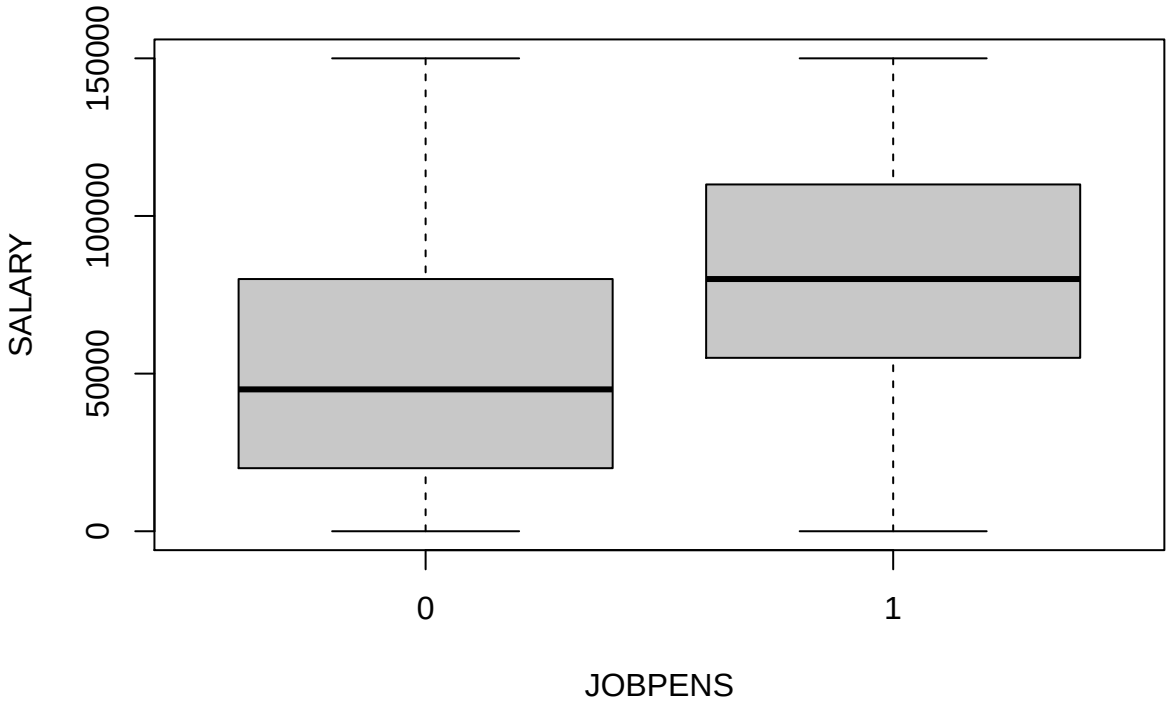

```
boxplot(SALARY ~ WKSWKGR, data = dataset)
```



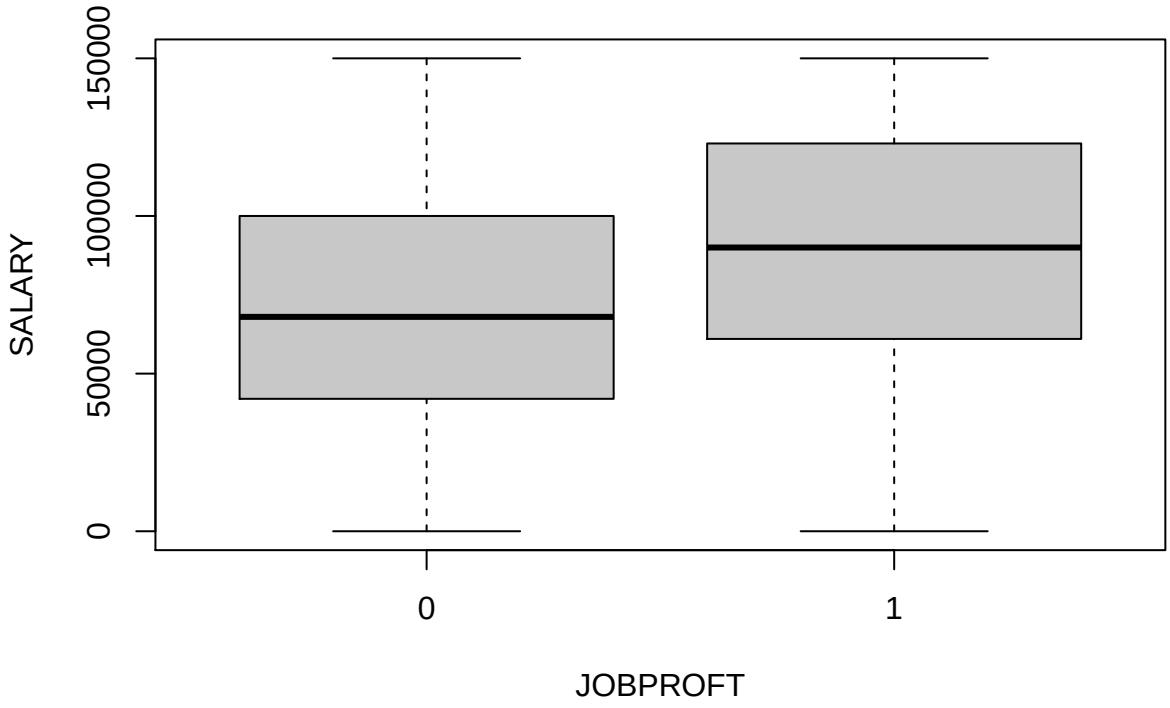

```
boxplot(SALARY ~ JOBINS, data = dataset)
```



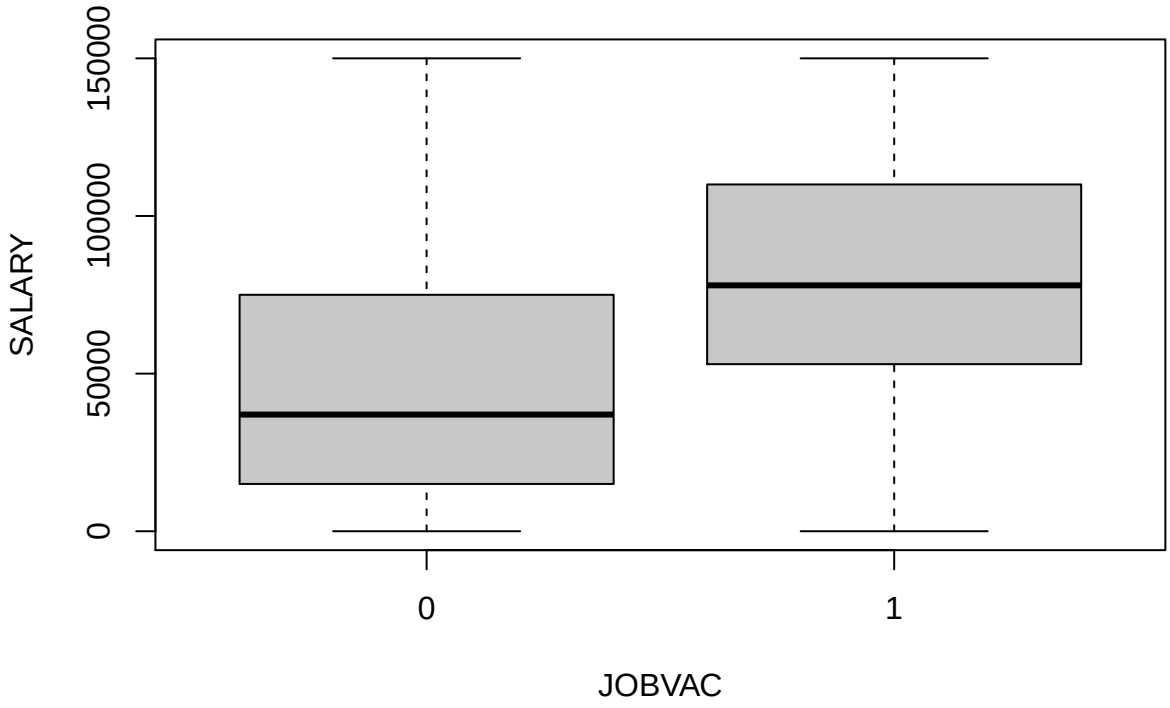

```
boxplot(SALARY ~ JOBPENS, data = dataset)
```



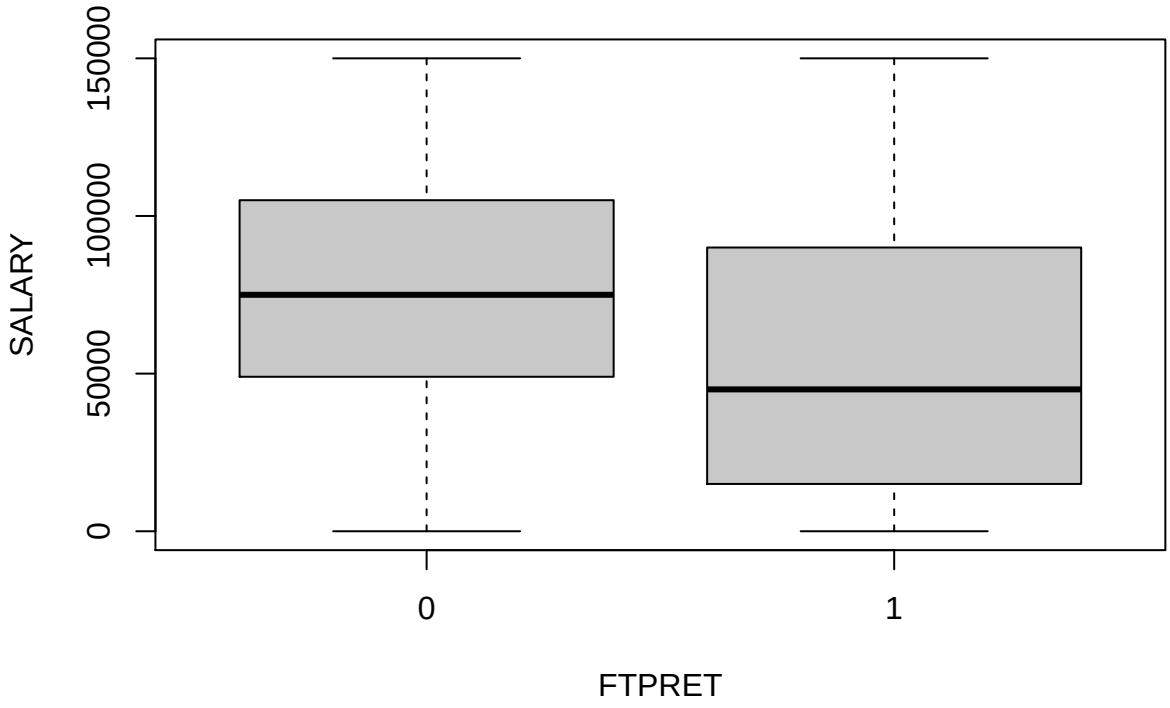

```
boxplot(SALARY ~ JOBPROFT, data = dataset)
```



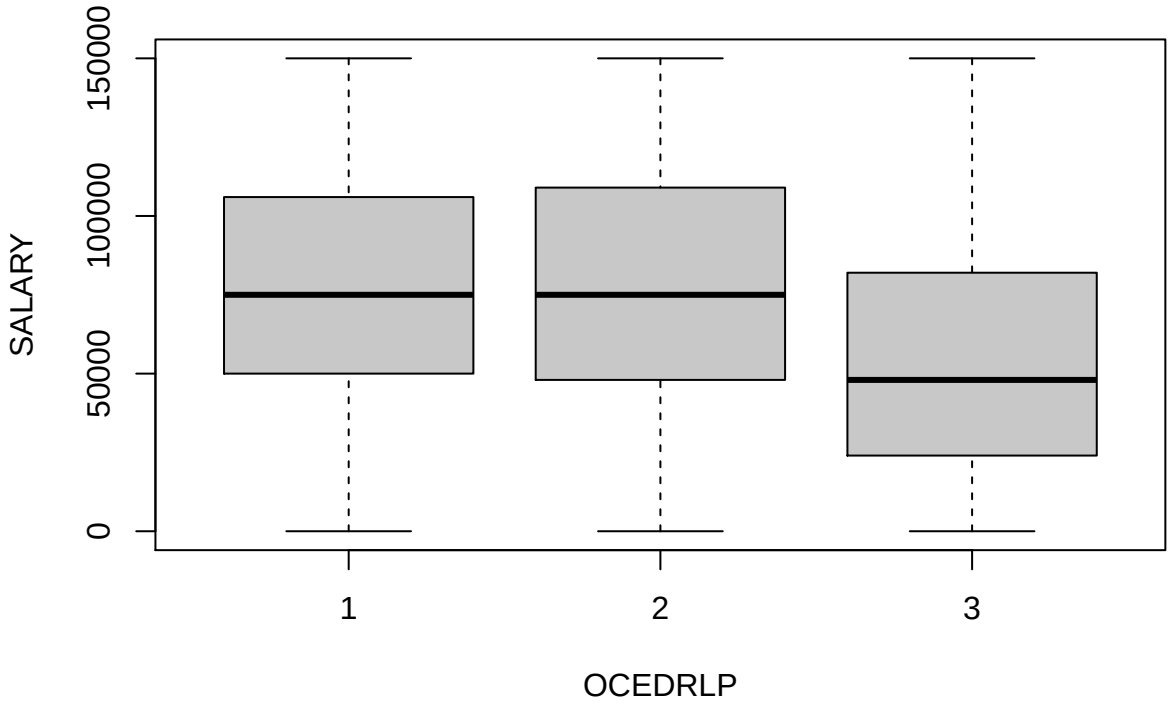

```
boxplot(SALARY ~ JOBVAC, data = dataset)
```



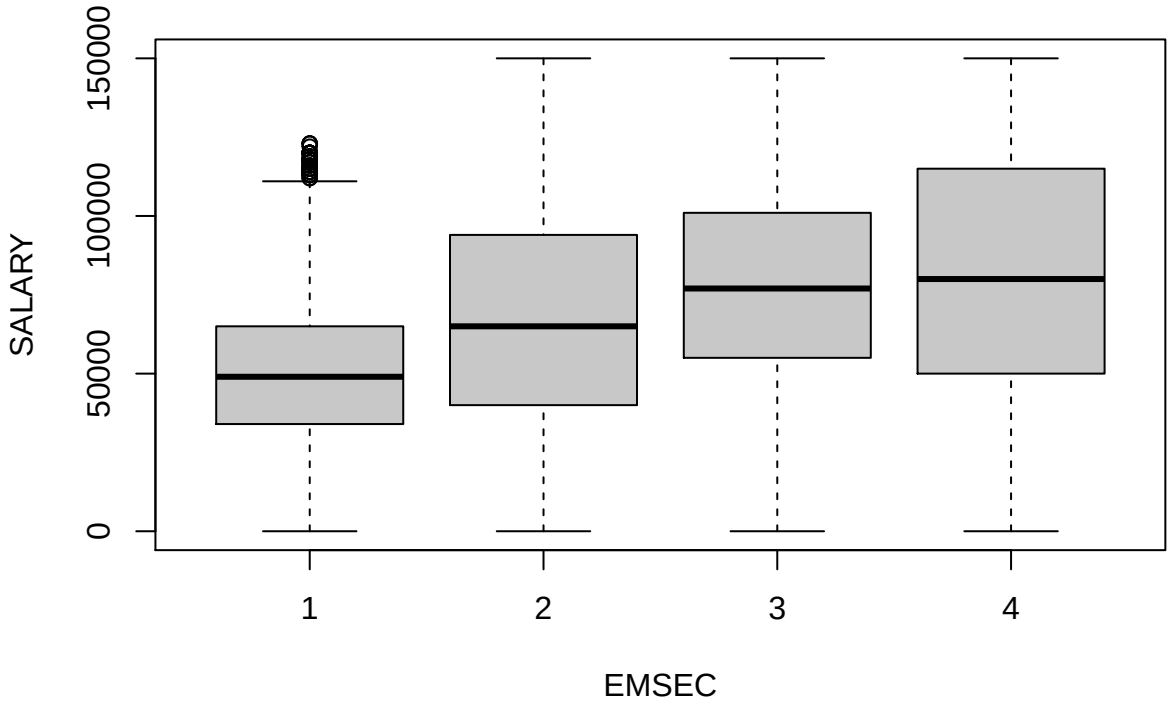

```
boxplot(SALARY ~ FTPRET, data = dataset)
```



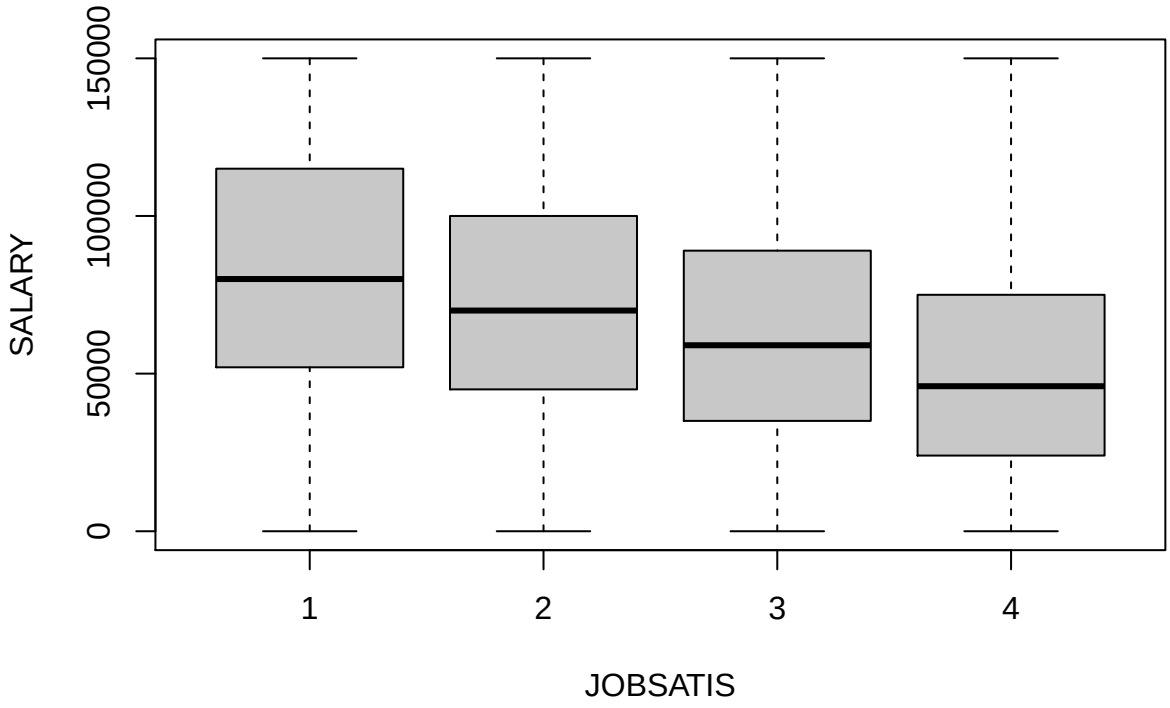

```
boxplot(SALARY ~ OCEDRLP, data = dataset)
```



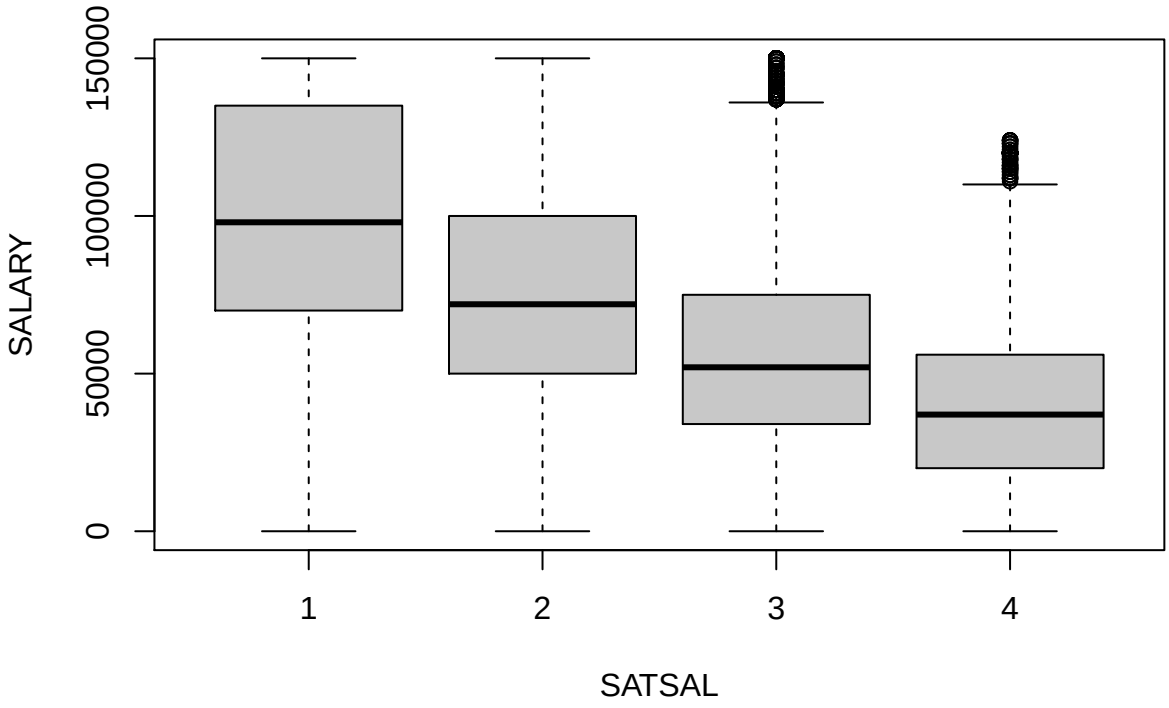

```
boxplot(SALARY ~ EMSEC, data = dataset)
```



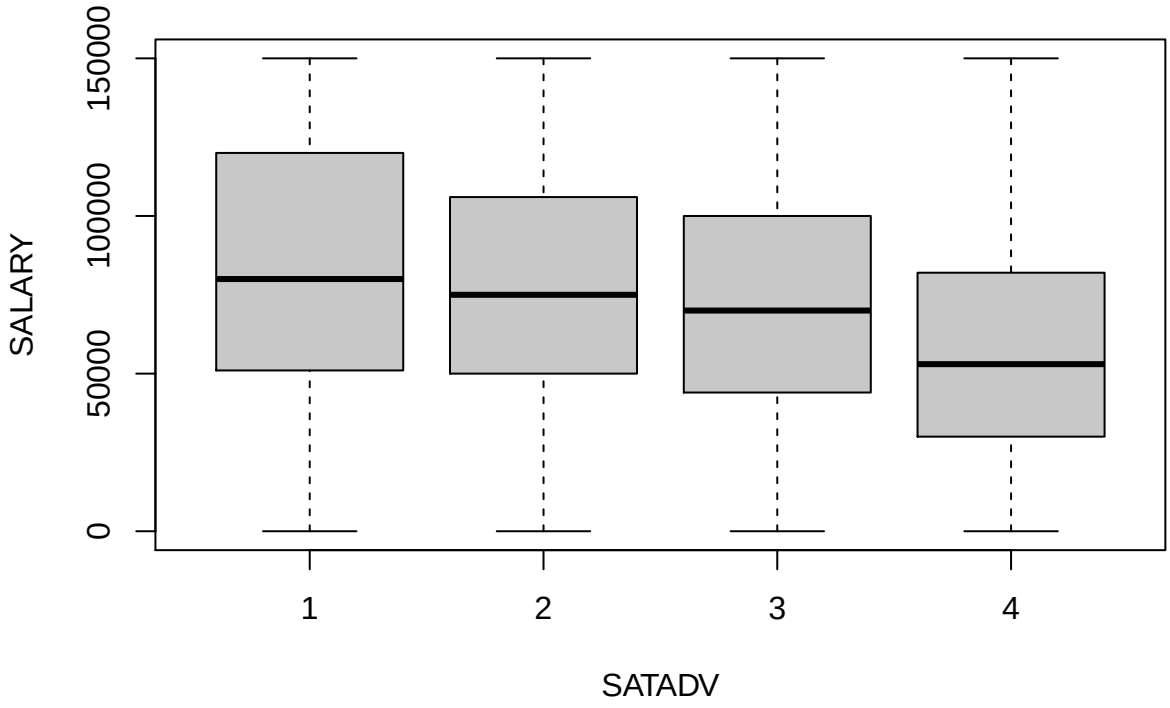

```
boxplot(SALARY ~ JOBSATIS, data = dataset)
```



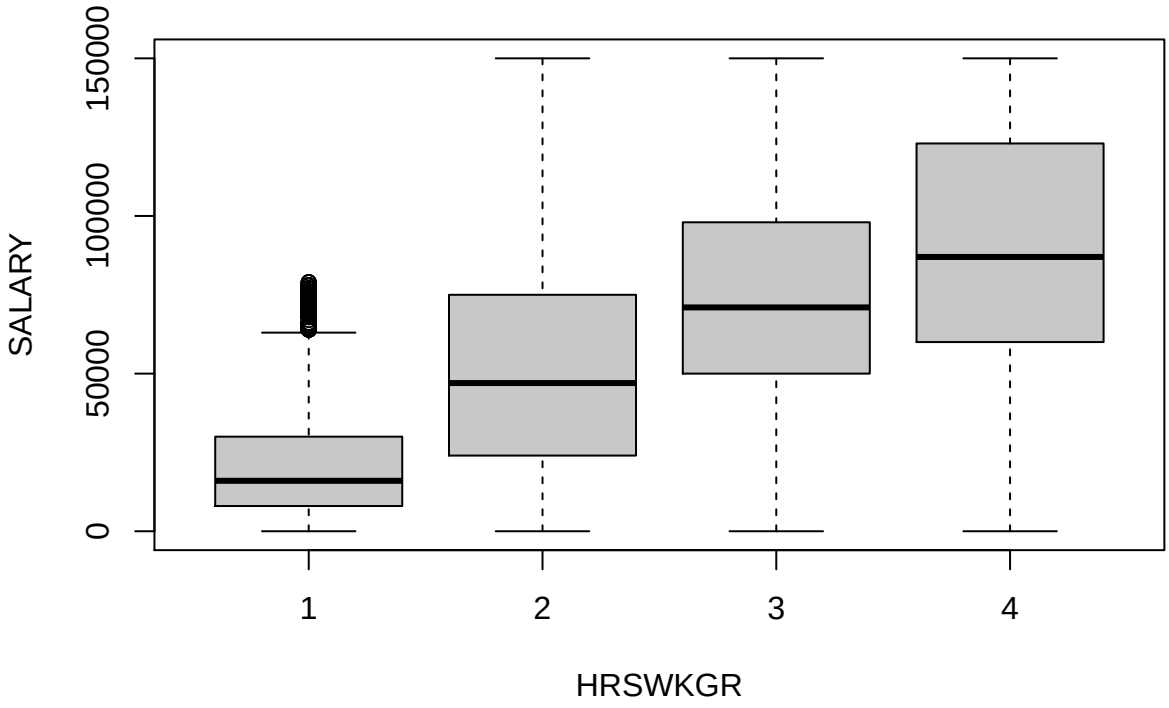

```
boxplot(SALARY ~ SATSAL, data = dataset)
```




```
boxplot(SALARY ~ SATADV, data = dataset)
```




```
boxplot(SALARY ~ HRSWKGR, data = dataset)
```




```
final_model.AIC <- stepAIC(final_model)
```

```
## Start:  AIC=1984286
## SALARY ~ AGE + as.factor(GENDER) + as.factor(DGRDG) + as.factor(RACETH) +
##     as.factor(NOCPRMG) + as.factor(NBAMEMG) + as.factor(WKSWKGR) +
##     as.factor(HRSWKGR) + as.factor(JOBINS) + as.factor(JOBPENS) +
##     as.factor(JOBPROFT) + as.character(JOBVAC) + as.factor(FTPRET) +
##     as.factor(OCEDRLP) + as.factor(EMSEC) + as.factor(JOBSATIS) +
##     as.factor(SATSAL) + as.factor(SATADV) + as.factor(MGRNAT) +
##     as.factor(MGROTH)
##
##                         Df  Sum of Sq        RSS     AIC
## <none>                                7.1006e+13 1984286
## - as.factor(SATADV)      3 4.1078e+10 7.1047e+13 1984336
## - as.factor(JOBSATIS)    3 5.5276e+10 7.1062e+13 1984356
## - as.character(JOBVAC)   1 1.1727e+11 7.1124e+13 1984445
## - as.factor(JOBPROFT)    1 1.6754e+11 7.1174e+13 1984513
## - as.factor(MGROTH)      1 1.7599e+11 7.1182e+13 1984525
## - as.factor(RACETH)      2 3.0346e+11 7.1310e+13 1984697
## - as.factor(JOBINS)      1 4.2760e+11 7.1434e+13 1984868
## - as.factor(WKSWKGR)     3 4.3620e+11 7.1443e+13 1984876
## - as.factor(NBAMEMG)     8 5.4374e+11 7.1550e+13 1985012
## - as.factor(JOBPENS)     1 5.4289e+11 7.1549e+13 1985025
## - as.factor(MGRNAT)      1 5.6702e+11 7.1573e+13 1985057
## - as.factor(OCEDRLP)     2 6.5966e+11 7.1666e+13 1985181
```

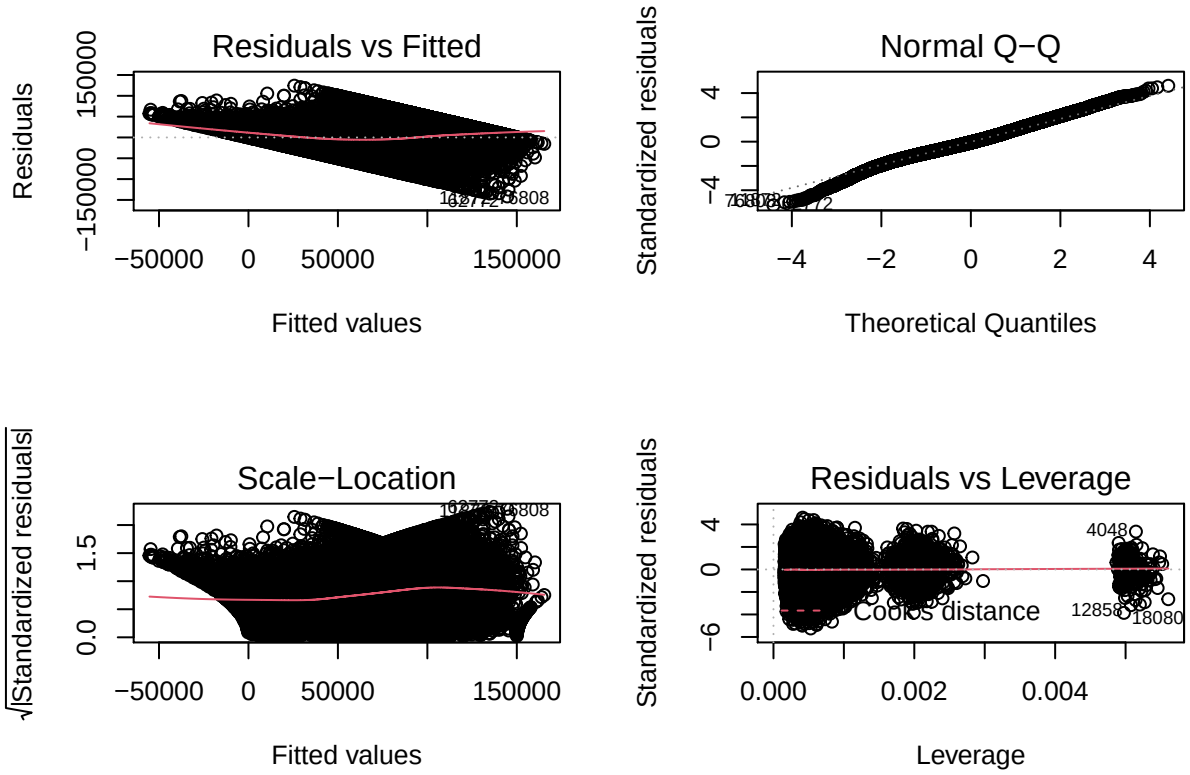
```
## - as.factor(GENDER)     1 7.7031e+11 7.1777e+13 1985333
## - as.factor(NOCPRMG)    6 7.9885e+11 7.1805e+13 1985362
## - as.factor(FTPRET)     1 1.1360e+12 7.2142e+13 1985827
## - as.factor(EMSEC)      3 4.1402e+12 7.5147e+13 1989790
## - AGE                   1 6.8819e+12 7.7888e+13 1993278
## - as.factor(DGRDG)      3 7.5342e+12 7.8541e+13 1994084
## - as.factor(HRSWKGR)    3 7.6166e+12 7.8623e+13 1994186
## - as.factor(SATSAL)     3 8.2409e+12 7.9247e+13 1994955
```

```
summary(final_model.AIC)
```

```
##
## Call:
## lm(formula = SALARY ~ AGE + as.factor(GENDER) + as.factor(DGRDG) +
##     as.factor(RACETH) + as.factor(NOCPRMG) + as.factor(NBAMEMG) +
##     as.factor(WKSWKGR) + as.factor(HRSWKGR) + as.factor(JOBINS) +
##     as.factor(JOBPENS) + as.factor(JOBPROFT) + as.character(JOBVAC) +
##     as.factor(FTPRET) + as.factor(OCEDRLP) + as.factor(EMSEC) +
##     as.factor(JOBSATIS) + as.factor(SATSAL) + as.factor(SATADV) +
##     as.factor(MGRNAT) + as.factor(MGROTH), data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141654  -17683   -1732   16841  124344
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -34341.859   1298.252 -26.452  < 2e-16 ***
## AGE                      765.949      7.893  97.045  < 2e-16 ***
## as.factor(GENDER)2      6315.537    194.516  32.468  < 2e-16 ***
## as.factor(DGRDG)2       8815.289    222.665  39.590  < 2e-16 ***
## as.factor(DGRDG)3      25042.198    263.336  95.096  < 2e-16 ***
## as.factor(DGRDG)4      27377.966    503.112  54.417  < 2e-16 ***
## as.factor(RACETH)2     -3350.684    246.684 -13.583  < 2e-16 ***
## as.factor(RACETH)3     -5928.154    290.921 -20.377  < 2e-16 ***
## as.factor(NOCPRMG)2   -12547.931    461.224 -27.206  < 2e-16 ***
## as.factor(NOCPRMG)3   -11571.562    523.458 -22.106  < 2e-16 ***
## as.factor(NOCPRMG)4    -3239.871    483.383  -6.702 2.06e-11 ***
## as.factor(NOCPRMG)5    -4952.859    389.982 -12.700  < 2e-16 ***
## as.factor(NOCPRMG)6    -2303.462    365.244  -6.307 2.86e-10 ***
## as.factor(NOCPRMG)7    -2540.381    353.462  -7.187 6.66e-13 ***
## as.factor(NBAMEMG)2    -4551.103    418.929 -10.864  < 2e-16 ***
## as.factor(NBAMEMG)3     -744.047    464.919  -1.600  0.10952
## as.factor(NBAMEMG)4    -2175.921    399.204  -5.451 5.03e-08 ***
## as.factor(NBAMEMG)5     3906.878    389.392  10.033  < 2e-16 ***
## as.factor(NBAMEMG)6    -4053.354    442.777  -9.154  < 2e-16 ***
## as.factor(NBAMEMG)7    -4677.029    453.104 -10.322  < 2e-16 ***
## as.factor(NBAMEMG)9     2943.816   1881.822   1.564  0.11774
## as.factor(NBAMEMG)96   -1934.262    607.942  -3.182  0.00146 **
## as.factor(WKSWKGR)2     2165.526   1471.824   1.471  0.14121
## as.factor(WKSWKGR)3    13480.365   1141.525  11.809  < 2e-16 ***
## as.factor(WKSWKGR)4    18135.003   1092.693  16.597  < 2e-16 ***
## as.factor(HRSWKGR)2    19824.940    462.577  42.858  < 2e-16 ***
## as.factor(HRSWKGR)3    29191.549    423.248  68.970  < 2e-16 ***
```

```
## as.factor(HRSWKGR)4    39603.517    421.300  94.003  < 2e-16 ***
## as.factor(JOBINS)1      8867.633    366.578  24.190  < 2e-16 ***
## as.factor(JOBPENS)1     7027.754    257.832  27.257  < 2e-16 ***
## as.factor(JOBPROFT)1    3347.925    221.106  15.142  < 2e-16 ***
## as.character(JOBVAC)1   4132.994    326.244  12.668  < 2e-16 ***
## as.factor(FTPRET)1    -15991.373    405.580 -39.428  < 2e-16 ***
## as.factor(OCEDRLP)2    -2389.788    215.694 -11.080  < 2e-16 ***
## as.factor(OCEDRLP)3    -9246.112    308.569 -29.964  < 2e-16 ***
## as.factor(EMSEC)2       4995.566    398.071  12.549  < 2e-16 ***
## as.factor(EMSEC)3      17103.501    421.679  40.560  < 2e-16 ***
## as.factor(EMSEC)4      21425.442    358.058  59.838  < 2e-16 ***
## as.factor(JOBSATIS)2    1045.879    216.993   4.820 1.44e-06 ***
## as.factor(JOBSATIS)3    3197.113    391.449   8.167 3.19e-16 ***
## as.factor(JOBSATIS)4    3486.731    674.870   5.167 2.39e-07 ***
## as.factor(SATSAL)2    -15645.280    222.554 -70.299  < 2e-16 ***
## as.factor(SATSAL)3    -28389.939    310.769 -91.354  < 2e-16 ***
## as.factor(SATSAL)4    -36887.379    431.440 -85.498  < 2e-16 ***
## as.factor(SATADV)2      1475.363    236.436   6.240 4.39e-10 ***
## as.factor(SATADV)3      1973.156    293.980   6.712 1.93e-11 ***
## as.factor(SATADV)4       865.671    398.624   2.172  0.02988 *
## as.factor(MGRNAT)1      7060.576    253.465  27.856  < 2e-16 ***
## as.factor(MGROTH)1      3265.142    210.395  15.519  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27030 on 97172 degrees of freedom
## Multiple R-squared:  0.5651, Adjusted R-squared:  0.5648
## F-statistic:  2630 on 48 and 97172 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(final_model.AIC)
```



To begin, I started by qualitatively removing variables that did not necessarily affect the model. These were variables such as PERSONID, YEAR, CHTOT, and SALARY. Next, I plotted each variable against Salary, to see if there were any obvious non-correlations that could be removed. Then I created a subset where LFSTAT == 1 to get rid of NA values. After this initial cleaning, I started by creating the linear model to predict salary using as many of the variables from the original data left. I used this large model as a control to test the available R^2 as well as the P-values for each coefficient, then started adjusting. I switched all variables to the format of "as.factor()".

From this point, is when I used backward elimination to slowly move away the coefficients that had a p value higher than 0.05. This type of p value less than the significance level accepts the Null, stating that this coefficient will become 0 during analysis, showing that it will have little effect on the model. To arrive to final_model, all of the coefficients in analysis have a p-value less than 0.05, showing that they will be significant for analysis. At this point, we then use boxplots to plot the remaining values against Salary. We then clean the outliers from these variables by creating further subsets. Once this is done, the diagnastic plot states the following.

Independence Assessment - The line on the Residuals vs. Fitted plot is horizontal, therefore, independence is clear.

Constant Variance Assessment - The line on the Scale-location plot is somewhat horizontal, therefore, constant variance is somewhat clear.

Normal Residual Distrubution Assessment - 80%-90% of the points fall on the qq plot, therefore normality of the distribution is not violated.

Hence this shows that the model passes the conditions, making it a good fit.

Therefore this model has been stripped of unnessecary values, all of the p values for coefficients are below 0.05, the models maintains a relatively strong R^2 at ~0.56, and also maintains a good of fit, therefore it is a good model.

2. Call your final regression model `model.lm`. Clearly show your final regression model: the R command, and the R output summary. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients.

```
final_model <- lm(SALARY ~ AGE + as.factor(GENDER) + as.factor(DGRDG) + as.factor(RACETH) + as.factor(NO

model.lm <- stepAIC(final_model)
```

```
## Start:  AIC=1984286
## SALARY ~ AGE + as.factor(GENDER) + as.factor(DGRDG) + as.factor(RACETH) +
##     as.factor(NOCPRMG) + as.factor(NBAMEMG) + as.factor(WKSWKGR) +
##     as.factor(HRSWKGR) + as.factor(JOBINS) + as.factor(JOBPENS) +
##     as.factor(JOBPROFT) + as.character(JOBVAC) + as.factor(FTPRET) +
##     as.factor(OCEDRLP) + as.factor(EMSEC) + as.factor(JOBSATIS) +
##     as.factor(SATSAL) + as.factor(SATADV) + as.factor(MGRNAT) +
##     as.factor(MGROTH)
##
##                         Df  Sum of Sq        RSS     AIC
## <none>                                7.1006e+13 1984286
## - as.factor(SATADV)      3 4.1078e+10 7.1047e+13 1984336
## - as.factor(JOBSATIS)    3 5.5276e+10 7.1062e+13 1984356
## - as.character(JOBVAC)   1 1.1727e+11 7.1124e+13 1984445
## - as.factor(JOBPROFT)    1 1.6754e+11 7.1174e+13 1984513
## - as.factor(MGROTH)      1 1.7599e+11 7.1182e+13 1984525
## - as.factor(RACETH)      2 3.0346e+11 7.1310e+13 1984697
## - as.factor(JOBINS)      1 4.2760e+11 7.1434e+13 1984868
## - as.factor(WKSWKGR)     3 4.3620e+11 7.1443e+13 1984876
## - as.factor(NBAMEMG)     8 5.4374e+11 7.1550e+13 1985012
## - as.factor(JOBPENS)     1 5.4289e+11 7.1549e+13 1985025
## - as.factor(MGRNAT)      1 5.6702e+11 7.1573e+13 1985057
## - as.factor(OCEDRLP)     2 6.5966e+11 7.1666e+13 1985181
## - as.factor(GENDER)      1 7.7031e+11 7.1777e+13 1985333
## - as.factor(NOCPRMG)     6 7.9885e+11 7.1805e+13 1985362
## - as.factor(FTPRET)      1 1.1360e+12 7.2142e+13 1985827
## - as.factor(EMSEC)       3 4.1402e+12 7.5147e+13 1989790
## - AGE                    1 6.8819e+12 7.7888e+13 1993278
## - as.factor(DGRDG)       3 7.5342e+12 7.8541e+13 1994084
## - as.factor(HRSWKGR)     3 7.6166e+12 7.8623e+13 1994186
## - as.factor(SATSAL)      3 8.2409e+12 7.9247e+13 1994955
```

```
summary(model.lm)
```

```
##
## Call:
## lm(formula = SALARY ~ AGE + as.factor(GENDER) + as.factor(DGRDG) +
##     as.factor(RACETH) + as.factor(NOCPRMG) + as.factor(NBAMEMG) +
##     as.factor(WKSWKGR) + as.factor(HRSWKGR) + as.factor(JOBINS) +
##     as.factor(JOBPENS) + as.factor(JOBPROFT) + as.character(JOBVAC) +
##     as.factor(FTPRET) + as.factor(OCEDRLP) + as.factor(EMSEC) +
##     as.factor(JOBSATIS) + as.factor(SATSAL) + as.factor(SATADV) +
##     as.factor(MGRNAT) + as.factor(MGROTH), data = dataset)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -141654  -17683   -1732   16841  124344
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -34341.859   1298.252 -26.452  < 2e-16 ***
## AGE                      765.949      7.893  97.045  < 2e-16 ***
## as.factor(GENDER)2      6315.537    194.516  32.468  < 2e-16 ***
## as.factor(DGRDG)2       8815.289    222.665  39.590  < 2e-16 ***
## as.factor(DGRDG)3      25042.198    263.336  95.096  < 2e-16 ***
## as.factor(DGRDG)4      27377.966    503.112  54.417  < 2e-16 ***
## as.factor(RACETH)2     -3350.684    246.684 -13.583  < 2e-16 ***
## as.factor(RACETH)3     -5928.154    290.921 -20.377  < 2e-16 ***
## as.factor(NOCPRMG)2   -12547.931    461.224 -27.206  < 2e-16 ***
## as.factor(NOCPRMG)3   -11571.562    523.458 -22.106  < 2e-16 ***
## as.factor(NOCPRMG)4    -3239.871    483.383  -6.702 2.06e-11 ***
## as.factor(NOCPRMG)5    -4952.859    389.982 -12.700  < 2e-16 ***
## as.factor(NOCPRMG)6    -2303.462    365.244  -6.307 2.86e-10 ***
## as.factor(NOCPRMG)7    -2540.381    353.462  -7.187 6.66e-13 ***
## as.factor(NBAMEMG)2    -4551.103    418.929 -10.864  < 2e-16 ***
## as.factor(NBAMEMG)3     -744.047    464.919  -1.600  0.10952
## as.factor(NBAMEMG)4    -2175.921    399.204  -5.451 5.03e-08 ***
## as.factor(NBAMEMG)5     3906.878    389.392  10.033  < 2e-16 ***
## as.factor(NBAMEMG)6    -4053.354    442.777  -9.154  < 2e-16 ***
## as.factor(NBAMEMG)7    -4677.029    453.104 -10.322  < 2e-16 ***
## as.factor(NBAMEMG)9     2943.816   1881.822   1.564  0.11774
## as.factor(NBAMEMG)96   -1934.262    607.942  -3.182  0.00146 **
## as.factor(WKSWKGR)2     2165.526   1471.824   1.471  0.14121
## as.factor(WKSWKGR)3    13480.365   1141.525  11.809  < 2e-16 ***
## as.factor(WKSWKGR)4    18135.003   1092.693  16.597  < 2e-16 ***
## as.factor(HRSWKGR)2    19824.940    462.577  42.858  < 2e-16 ***
## as.factor(HRSWKGR)3    29191.549    423.248  68.970  < 2e-16 ***
## as.factor(HRSWKGR)4    39603.517    421.300  94.003  < 2e-16 ***
## as.factor(JOBINS)1      8867.633    366.578  24.190  < 2e-16 ***
## as.factor(JOBPENS)1     7027.754    257.832  27.257  < 2e-16 ***
## as.factor(JOBPROFT)1    3347.925    221.106  15.142  < 2e-16 ***
## as.character(JOBVAC)1   4132.994    326.244  12.668  < 2e-16 ***
## as.factor(FTPRET)1    -15991.373    405.580 -39.428  < 2e-16 ***
## as.factor(OCEDRLP)2    -2389.788    215.694 -11.080  < 2e-16 ***
## as.factor(OCEDRLP)3    -9246.112    308.569 -29.964  < 2e-16 ***
## as.factor(EMSEC)2       4995.566    398.071  12.549  < 2e-16 ***
## as.factor(EMSEC)3      17103.501    421.679  40.560  < 2e-16 ***
## as.factor(EMSEC)4      21425.442    358.058  59.838  < 2e-16 ***
## as.factor(JOBSATIS)2    1045.879    216.993   4.820 1.44e-06 ***
## as.factor(JOBSATIS)3    3197.113    391.449   8.167 3.19e-16 ***
## as.factor(JOBSATIS)4    3486.731    674.870   5.167 2.39e-07 ***
## as.factor(SATSAL)2    -15645.280    222.554 -70.299  < 2e-16 ***
## as.factor(SATSAL)3    -28389.939    310.769 -91.354  < 2e-16 ***
## as.factor(SATSAL)4    -36887.379    431.440 -85.498  < 2e-16 ***
## as.factor(SATADV)2      1475.363    236.436   6.240 4.39e-10 ***
## as.factor(SATADV)3      1973.156    293.980   6.712 1.93e-11 ***
## as.factor(SATADV)4       865.671    398.624   2.172  0.02988 *
## as.factor(MGRNAT)1      7060.576    253.465  27.856  < 2e-16 ***
## as.factor(MGROTH)1      3265.142    210.395  15.519  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27030 on 97172 degrees of freedom
## Multiple R-squared:  0.5651, Adjusted R-squared:  0.5648
## F-statistic:  2630 on 48 and 97172 DF,  p-value: < 2.2e-16
```

Model Equation -

salary = -34341.859 + 765.949(AGE) + 6315.537(GENDER2) + 8815.289(DGRDG2) + 25042.198(DGRDG3) + 27377.966(DGRDG4) + -3350.684(RACETH2) + -5928.154(RACETH3) + -12547.931(NOCPRMG2) + -11571.562(NOCPRMG3) + -3239.871(NOCPRMG4) + -4952.859(NOCPRMG5) + -2303.462(NOCPRMG6) + -2540.381(NOCPRMG7) + -4551.103(NBAMEMG2) + -744.047(NBAMEMG3) + -2175.921(NBAMEMG4) + 3906.878(NBAMEMG5) + -4053.354(NBAMEMG6) + -4677.029(NBAMEMG7) + 2943.816(NBAMEMG9) + -1934.262(NBAMEMG96) + 2165.526(WKSWKGR2) + 13480.365(WKSWKGR3) + 18135.003(WKSWKGR4) + 19824.940(HRSWKGR2)+ 29191.549(HRSWKGR3) + 39603.517(HRSWKGR4) + 8867.633(JOBINS1) + 7027.754(JOBPENS1) + 3347.925(JOBPROFIT1) + 4132.994(JOBVAC1) + -15991.373(FTPRET1) + -2389.788(OCEDRLP2) + -9246.112(OCEDRLP3) + 4995.566(EMSEC2) + 17103.501(EMSEC3) + 21425.442(EMSEC4) + 1045.879(JOBSATIS2)+ 3197.113(JOBSATIS3) + 3486.731(JOBSATIS4) + -15645.280(SATSAL2) + -28389.939(SATSAL3) + -36887.379(SATSAL4) + 1475.363(SATADV2) + 1973.156(SATADV3) + 865.671(SATADV4) + 7060.576(MGRNAT1) + 3265.142(MGROTH1)

If we hold all other variables constant, for every 1 increase in each variable, there will be a increase in salary based on each coefficient.

If all the variables are 0, then salary will be -34341.859.

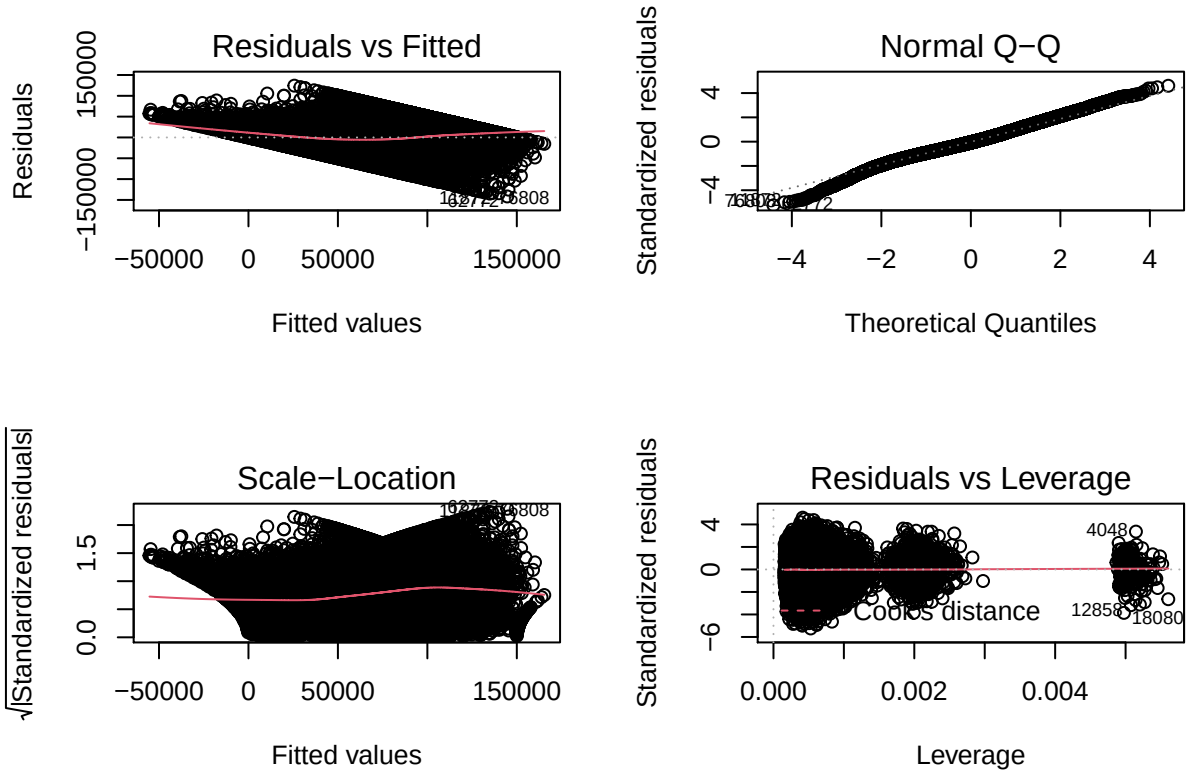
If the p value for a variable is large (larger than 0.05), this means that that variable is irrelevant and insignificant in this model.

3. Report the R^2 and adjusted R^2 of your model. What are the meaning of these values? Run a diagnostic plot for your model. Is your model a good fit? Is it easy to interpret?

Multiple R-squared: 0.5651 Adjusted R-squared: 0.5648

Meaning of R^2: Means this model can explain 56.51% of the variance in this data

```
par(mfrow = c(2,2))
plot(model.lm)
```



Independence Assessment - The line on the Residuals vs. Fitted plot is horizontal, therefore, independence is clear.

Constant Variance Assessment - The line on the Scale-location plot is horizontal, therefore, constant variance is somewhat clear.

Normal Residual Distrubution Assessment - 80-90% of the points fall on the Q-Q plot, therefore normality of the distribution is not violated.

Therefor our model is a good fit, since all the conditions are met. Based on the clarity of the diagnostic plots, this fit is easy to interpret.

4. Suppose you want to choose a career path to maximize your SALARY. Which career path would you choose base on your model? (Detail which highest degree you should obtain in which major, which sector should your employer be, etc).

To maximize salary, one should get a professional degree as their highest degree, work in the field of Science and Engineering related occupation for their principle job, obtain a bachelors in engineering, work for 40-52 weeks, work greater than 40 hours per week, have their principal job "somewhat related" to highest degree, work in the employer sector of Business or Industry, and have technical expertise in natural sciences.

Regression 2: job satisfaction vs other variables

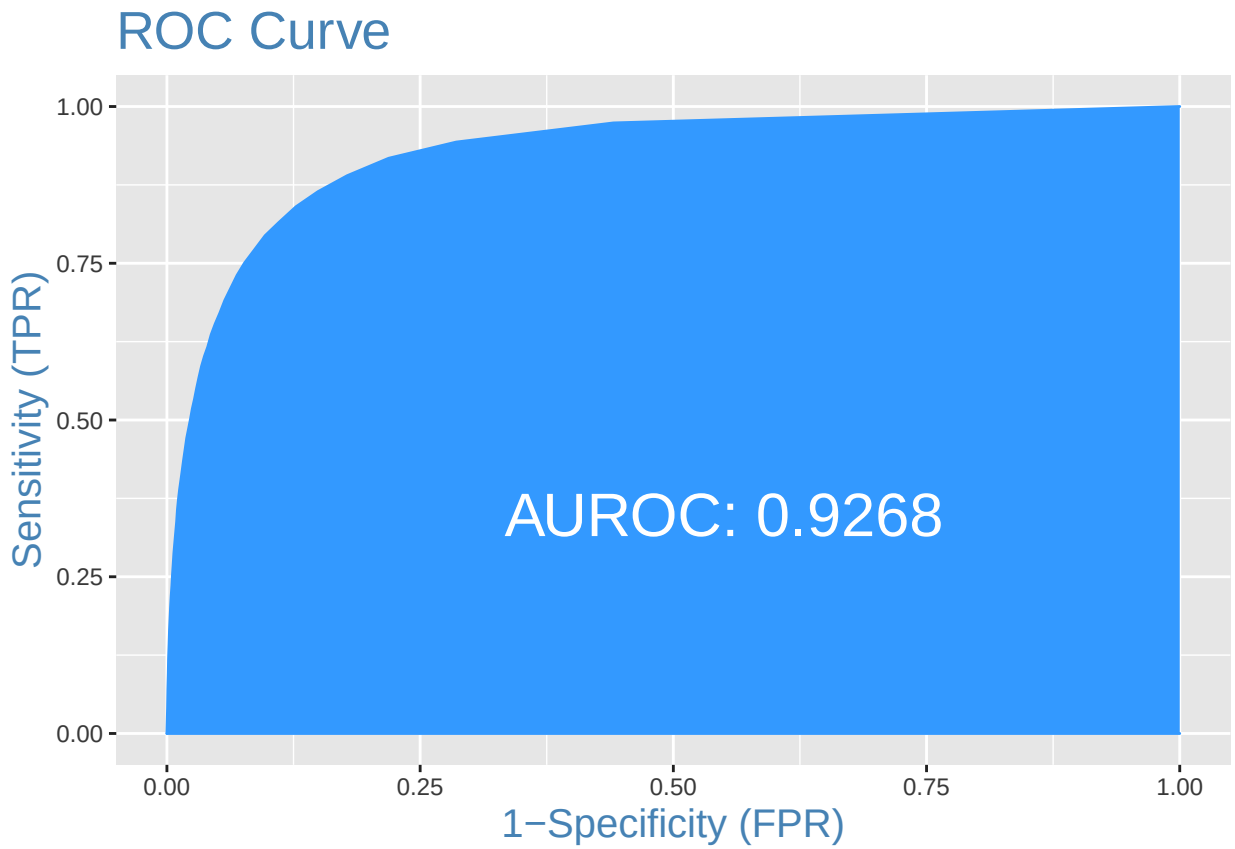
Recode JOBSATIS into two categories: "satisfied" = "somewhat/very satisfied", and "not satisfied" = "somewhat/very dissatisfied". Build a logistic regression model to predict the recoded job satisfaction based on the other variables.

```
dataset$JOBSATIS_UPDATED[dataset$JOBSATIS[] == 1 | dataset$JOBSATIS[] == 2 ] <- 0
dataset$JOBSATIS_UPDATED[dataset$JOBSATIS[] == 3 | dataset$JOBSATIS[] == 4 ] <- 1
dataset$is.satis <- 0
dataset$is.satis[dataset$JOBSATIS_UPDATED == 1] <- 1
```

1. Detail how you did variable selection: which models did you run, why did you discard certain models or variables, any variable transformations you did and why, which diagnostic tests did you run and what they showed, justifications if you removed outliers. How did you decide to deal with missing values in this dataset?

```
library(InformationValue)
```

```
updated_model <- glm(is.satis ~ AGE + as.factor(GENDER) + as.factor(DGRDG) + as.factor(NDGMEMG) + as.fa

plotROC(dataset$is.satis, updated_model$fitted.values)
```




```
# p1 <- pairs(is.satis ~ WEIGHT, data = dataset)
# p1 <- pairs(is.satis ~ WEIGHT, data = dataset)
# p2 <- pairs(is.satis ~ SAMPLE, data = dataset)
# p3 <- pairs(is.satis ~ SURID, data = dataset)
# p4 <- pairs(is.satis ~ AGE, data = dataset)
# p5 <- pairs(is.satis ~ GENDER, data = dataset)
# p6 <- pairs(is.satis ~ MINRTY, data = dataset)
# p7 <- pairs(is.satis ~ RACETH, data = dataset)
```

```
# p8 <- pairs(is.satis ~ CHU2IN, data = dataset)
# p9 <- pairs(is.satis ~ CH25IN, data = dataset)
# p10 <- pairs(is.satis ~ CH611IN, data = dataset)
# p11 <- pairs(is.satis ~ CH1218IN, data = dataset)
# p12 <- pairs(is.satis ~ CH19IN, data = dataset)
# p13 <- pairs(is.satis ~ BA03Y5, data = dataset)
# p14 <- pairs(is.satis ~ NBAMEMG, data = dataset)
# p15 <- pairs(is.satis ~ BADGRUS, data = dataset)
# p16 <- pairs(is.satis ~ DGRDG, data = dataset)
# p17 <- pairs(is.satis ~ HD03Y5, data = dataset)
# p18 <- pairs(is.satis ~ NDGMEMG, data = dataset)
# p19 <- pairs(is.satis ~ HDDGRUS, data = dataset)
# p20 <- pairs(is.satis ~ LFSTAT, data = dataset)
# p21 <- pairs(is.satis ~ HRSWKGR, data = dataset)
# p22 <- pairs(is.satis ~ WKSWKGR, data = dataset)
# p23 <- pairs(is.satis ~ JOBINS, data = dataset)
# p24 <- pairs(is.satis ~ JOBPENS, data = dataset)
# p25 <- pairs(is.satis ~ JOBPROFT, data = dataset)
# p26 <- pairs(is.satis ~ JOBVAC, data = dataset)
# p28 <- pairs(is.satis ~ FTPRET, data = dataset)
# p29 <- pairs(is.satis ~ PTWTFT, data = dataset)
# p30 <- pairs(is.satis ~ PTFAM, data = dataset)
# p31 <- pairs(is.satis ~ PTNOND, data = dataset)
# p32 <- pairs(is.satis ~ PTOCNA, data = dataset)
# p33 <- pairs(is.satis ~ PTOTP, data = dataset)
# p34 <- pairs(is.satis ~ OCEDRLP, data = dataset)
# p35 <- pairs(is.satis ~ NOCPRMG, data = dataset)
# p36 <- pairs(is.satis ~ EMSEC, data = dataset)
# p37 <- pairs(is.satis ~ WAPRSM, data = dataset)
# p38 <- pairs(is.satis ~ WASCSM, data = dataset)
# p39 <- pairs(is.satis ~ NRREA, data = dataset)
# p40 <- pairs(is.satis ~ NRSEC, data = dataset)
# p41 <- pairs(is.satis ~ JOBSATIS, data = dataset)
# p42 <- pairs(is.satis ~ SATADV, data = dataset)
# p43 <- pairs(is.satis ~ SATBEN, data = dataset)
# p44 <- pairs(is.satis ~ SATCHAL, data = dataset)
# p45 <- pairs(is.satis ~ SATIND, data = dataset)
# p46 <- pairs(is.satis ~ SATLOC, data = dataset)
# p47 <- pairs(is.satis ~ SATRESP, data = dataset)
# p48 <- pairs(is.satis ~ SATSAL, data = dataset)
# p49 <- pairs(is.satis ~ SATSEC, data = dataset)
# p50 <- pairs(is.satis ~ SATSOC, data = dataset)
# p51 <- pairs(is.satis ~ MGRNAT, data = dataset)
# p52 <- pairs(is.satis ~ MGROTH, data = dataset)
# p53 <- pairs(is.satis ~ MGRSOC, data = dataset)
# p54 <- pairs(is.satis ~ NWFAM, data = dataset)
# p55 <- pairs(is.satis ~ NWLAY, data = dataset)
# p56 <- pairs(is.satis ~ NWNOND, data = dataset)
# p57 <- pairs(is.satis ~ NWOCNA, data = dataset)
# p58 <- pairs(is.satis ~ NWOTP, data = dataset)
# p59 <- pairs(is.satis ~ NWSTU, data = dataset)
```

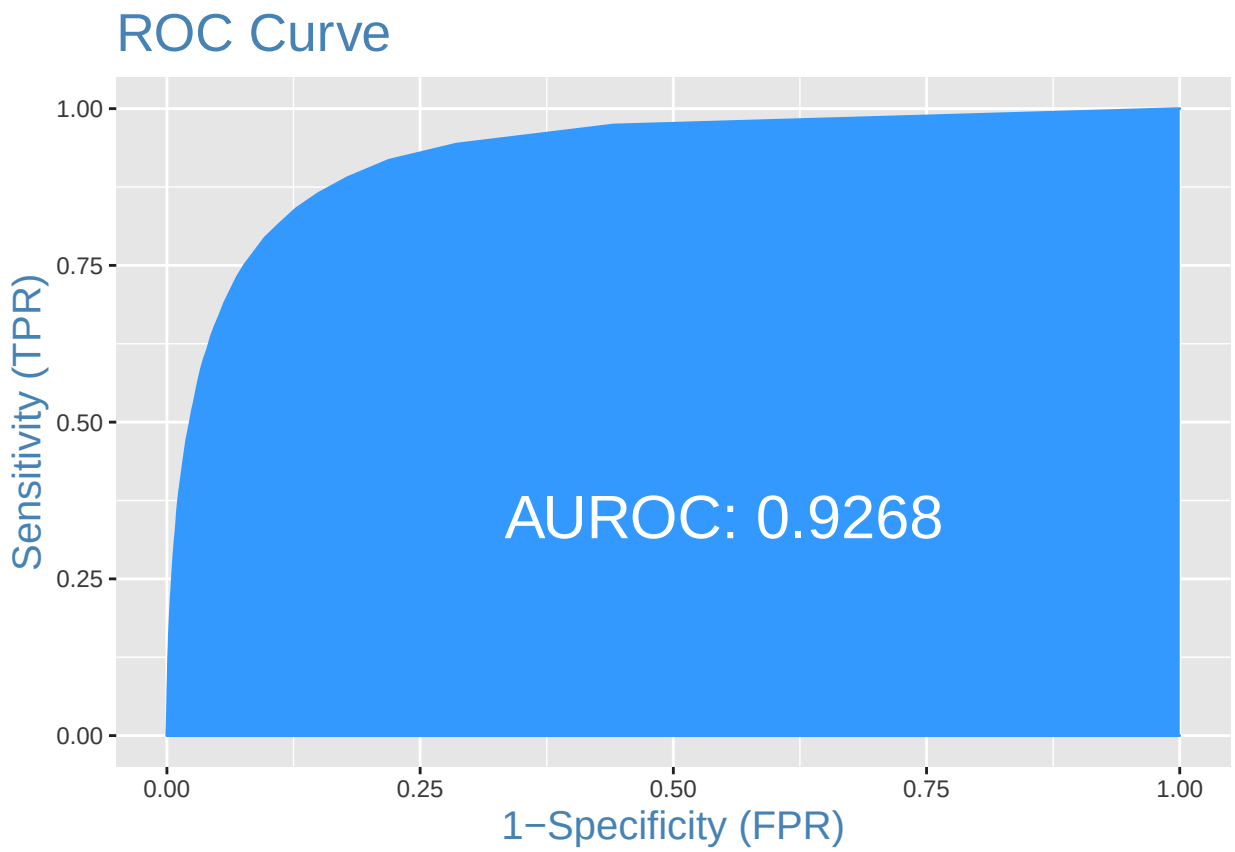
The way I conducted variable selection was as such - To begin, I started by qualitatively removing variables

that did not necessarily affect the model. These were variables such as PERSONID, YEAR, CHTOT, and SALARY. Next, I plotted each variable against is.satis, which is a simple recode of job satsifaction, to see if there were any obvious non-correlations that could be removed. Then I used the subset where LFSTAT == 1 to not use NA values. After this initial cleaning, I started by creating the logistic model to predict Job Satisfaction using as many of the variables from the data left after cleaning. I used this large model as a control to test the available AUROC as well as the P-values for each coefficient, then started adjusting further from there. I switched all variables to the format of "as.factor()".

From this point, is when I used backward elimination to slowly move away the coefficients that had a p value higher than 0.05. This type of p value less than the significance level accepts the Null, stating that this coefficient will become 0 during analysis, showing that it will have little effect on the model. To arrive to updated_model, I cleared unneccesary variables to get the highest AUROC value. Once this is done, the logistic model is complete.

2. Call your final regression model `model.lm`. Clearly show your final regression model: the R command, and the R output summary. Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients.

```
model.lm <- glm(is.satis ~ AGE + as.factor(GENDER) + as.factor(DGRDG) + as.factor(NDGMEMG) + as.factor(

plotROC(dataset$is.satis, model.lm$fitted.values)
```




```
summary(model.lm)
```

```
##
```

```
## Call:
## glm(formula = is.satis ~ AGE + as.factor(GENDER) + as.factor(DGRDG) +
##     as.factor(NDGMEMG) + as.factor(NOCPRMG) + as.factor(RACETH) +
##     as.factor(NBAMEMG) + +as.factor(WKSWKGR) + as.factor(HRSWKGR) +
##     as.factor(JOBINS) + as.factor(JOBPENS) + as.factor(JOBPROFT) +
##     as.factor(JOBVAC) + as.factor(FTPRET) + as.factor(OCEDRLP) +
##     as.factor(EMSEC) + as.factor(SATADV) + as.factor(SATIND) +
##     as.factor(SATLOC) + as.factor(SATRESP) + as.factor(SATSAL) +
##     as.factor(SATSEC) + as.factor(SATSOC), family = "binomial",
##     data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3850  -0.2907  -0.1598  -0.0985   3.6111
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.180580   0.219858 -28.112  < 2e-16 ***
## AGE                    -0.011258   0.001313  -8.575  < 2e-16 ***
## as.factor(GENDER)2     -0.020402   0.030716  -0.664 0.506561
## as.factor(DGRDG)2       0.087689   0.037326   2.349 0.018809 *
## as.factor(DGRDG)3       0.216511   0.043044   5.030 4.91e-07 ***
## as.factor(DGRDG)4       0.386990   0.091895   4.211 2.54e-05 ***
## as.factor(NDGMEMG)2    -0.140667   0.099536  -1.413 0.157588
## as.factor(NDGMEMG)3     0.098606   0.110198   0.895 0.370894
## as.factor(NDGMEMG)4     0.072773   0.089683   0.811 0.417111
## as.factor(NDGMEMG)5    -0.056962   0.089521  -0.636 0.524578
## as.factor(NDGMEMG)6     0.103245   0.093614   1.103 0.270082
## as.factor(NDGMEMG)7    -0.068130   0.083389  -0.817 0.413920
## as.factor(NOCPRMG)2     0.113804   0.078454   1.451 0.146897
## as.factor(NOCPRMG)3     0.016329   0.089812   0.182 0.855730
## as.factor(NOCPRMG)4     0.121243   0.081144   1.494 0.135130
## as.factor(NOCPRMG)5     0.176681   0.066340   2.663 0.007738 **
## as.factor(NOCPRMG)6     0.225972   0.060542   3.732 0.000190 ***
## as.factor(NOCPRMG)7     0.180400   0.055779   3.234 0.001220 **
## as.factor(RACETH)2      0.051950   0.039012   1.332 0.182978
## as.factor(RACETH)3      0.072634   0.045146   1.609 0.107644
## as.factor(NBAMEMG)2     0.077284   0.092358   0.837 0.402713
## as.factor(NBAMEMG)3    -0.093117   0.100554  -0.926 0.354428
## as.factor(NBAMEMG)4     0.005443   0.084369   0.065 0.948561
## as.factor(NBAMEMG)5    -0.053496   0.086093  -0.621 0.534348
## as.factor(NBAMEMG)6    -0.049926   0.095177  -0.525 0.599888
## as.factor(NBAMEMG)7     0.039368   0.087428   0.450 0.652499
## as.factor(NBAMEMG)9    -0.424849   0.333140  -1.275 0.202208
## as.factor(NBAMEMG)96   -0.426859   0.112875  -3.782 0.000156 ***
## as.factor(WKSWKGR)2    -0.061870   0.246635  -0.251 0.801925
## as.factor(WKSWKGR)3     0.175471   0.188131   0.933 0.350972
## as.factor(WKSWKGR)4     0.294585   0.179552   1.641 0.100867
## as.factor(HRSWKGR)2     0.258167   0.072883   3.542 0.000397 ***
## as.factor(HRSWKGR)3     0.137876   0.067377   2.046 0.040724 *
## as.factor(HRSWKGR)4     0.304363   0.067396   4.516 6.30e-06 ***
## as.factor(JOBINS)1      0.092297   0.057445   1.607 0.108120
## as.factor(JOBPENS)1     0.046432   0.039669   1.170 0.241807
## as.factor(JOBPROFT)1   -0.155912   0.037548  -4.152 3.29e-05 ***
```

```
## as.factor(JOBVAC)1    0.035078   0.052331   0.670 0.502655
## as.factor(FTPRET)1   -0.300602   0.074600  -4.030 5.59e-05 ***
## as.factor(OCEDRLP)2   0.203572   0.033882   6.008 1.88e-09 ***
## as.factor(OCEDRLP)3   0.387413   0.044112   8.783  < 2e-16 ***
## as.factor(EMSEC)2     0.031914   0.063594   0.502 0.615785
## as.factor(EMSEC)3     0.153634   0.067258   2.284 0.022357 *
## as.factor(EMSEC)4     0.240553   0.056876   4.229 2.34e-05 ***
## as.factor(SATADV)2    0.371115   0.071634   5.181 2.21e-07 ***
## as.factor(SATADV)3    1.232631   0.071217  17.308  < 2e-16 ***
## as.factor(SATADV)4    2.029525   0.073561  27.590  < 2e-16 ***
## as.factor(SATIND)2    0.458395   0.035338  12.972  < 2e-16 ***
## as.factor(SATIND)3    1.171849   0.045648  25.671  < 2e-16 ***
## as.factor(SATIND)4    1.463978   0.073103  20.026  < 2e-16 ***
## as.factor(SATLOC)2    0.118866   0.032774   3.627 0.000287 ***
## as.factor(SATLOC)3    0.455889   0.042061  10.839  < 2e-16 ***
## as.factor(SATLOC)4    0.849456   0.060024  14.152  < 2e-16 ***
## as.factor(SATRESP)2   0.421909   0.043549   9.688  < 2e-16 ***
## as.factor(SATRESP)3   1.284904   0.049985  25.706  < 2e-16 ***
## as.factor(SATRESP)4   1.546687   0.076279  20.277  < 2e-16 ***
## as.factor(SATSAL)2   -0.032565   0.046882  -0.695 0.487295
## as.factor(SATSAL)3    1.137915   0.049232  23.113  < 2e-16 ***
## as.factor(SATSAL)4    2.075284   0.055519  37.380  < 2e-16 ***
## as.factor(SATSEC)2    0.052622   0.037797   1.392 0.163854
## as.factor(SATSEC)3    0.625316   0.043986  14.216  < 2e-16 ***
## as.factor(SATSEC)4    1.094416   0.050980  21.468  < 2e-16 ***
## as.factor(SATSOC)2    0.476159   0.037131  12.824  < 2e-16 ***
## as.factor(SATSOC)3    1.321776   0.045445  29.085  < 2e-16 ***
## as.factor(SATSOC)4    1.734693   0.062006  27.976  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65609  on 97220  degrees of freedom
## Residual deviance: 36060  on 97156  degrees of freedom
## AIC: 36190
##
## Number of Fisher Scoring iterations: 7
```

If we hold all other variables constant, for every 1 increase in each variable, there will be a increase in salary based on each coefficient.

If the p value for a variable is large (larger than 0.05), this means that that variable is irrelevant and largely insignificant in this model.

3. Report your model's ROC curve, and report any diagnostic plots or statistics that you used. Is your model a good fit? Is it easy to interpret?

AUROC - .9268

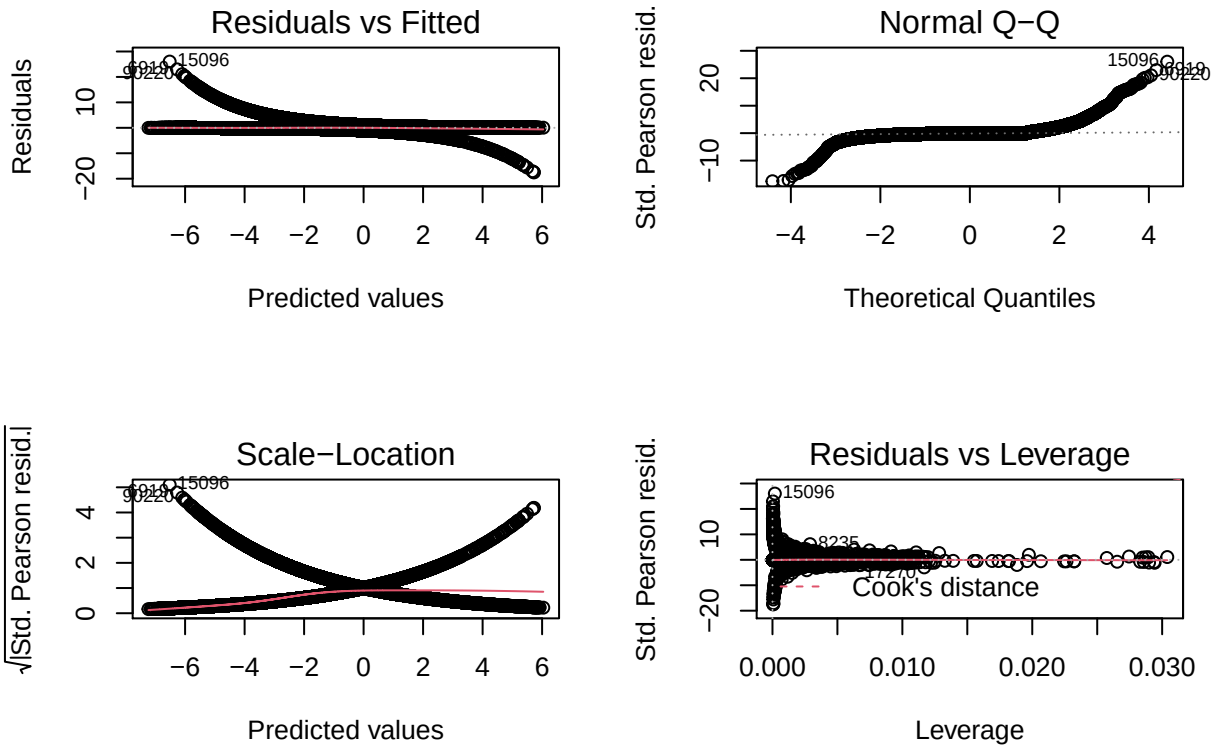
```
cutoff = optimalCutoff(dataset$is.satis, model.lm$fitted.values)
cutoff
```

```
## [1] 0.487623
```

```
table <- confusionMatrix(dataset$is.satis, model.lm$fitted.values)
accuracy <- (table[2,2] + table[1,1])/sum(table)
accuracy
```

```
## [1] 0.9265488
```

```
par(mfrow = c(2,2))
plot(model.lm)
```



To conclude this we use a confusion matrix. After calculation, the optimal accuracy is .926, which is obtained with a .457 cutoff value. Because this value is far above the cutoff, and is in the ~.9 range, therefore because the diagnostic plots pass the conditions, and the accuracy is high, this means that this model is a good fit, as well as clear to interperate.

4. Suppose you want to choose a career path to maximize your job satisfaction. Which career path would you choose base on your model? (Detail which highest degree you should obtain in which major, which sector should your employer be, etc).

```
summary(updated_model)
```

```
##
## Call:
## glm(formula = is.satis ~ AGE + as.factor(GENDER) + as.factor(DGRDG) +
##     as.factor(NDGMEMG) + as.factor(NOCPRMG) + as.factor(RACETH) +
```

```
##     as.factor(NBAMEMG) + +as.factor(WKSWKGR) + as.factor(HRSWKGR) +
##     as.factor(JOBINS) + as.factor(JOBPENS) + as.factor(JOBPROFT) +
##     as.factor(JOBVAC) + as.factor(FTPRET) + as.factor(OCEDRLP) +
##     as.factor(EMSEC) + as.factor(SATADV) + as.factor(SATIND) +
##     as.factor(SATLOC) + as.factor(SATRESP) + as.factor(SATSAL) +
##     as.factor(SATSEC) + as.factor(SATSOC), family = "binomial",
##     data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3850  -0.2907  -0.1598  -0.0985   3.6111
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.180580   0.219858 -28.112  < 2e-16 ***
## AGE                    -0.011258   0.001313  -8.575  < 2e-16 ***
## as.factor(GENDER)2     -0.020402   0.030716  -0.664 0.506561
## as.factor(DGRDG)2       0.087689   0.037326   2.349 0.018809 *
## as.factor(DGRDG)3       0.216511   0.043044   5.030 4.91e-07 ***
## as.factor(DGRDG)4       0.386990   0.091895   4.211 2.54e-05 ***
## as.factor(NDGMEMG)2    -0.140667   0.099536  -1.413 0.157588
## as.factor(NDGMEMG)3     0.098606   0.110198   0.895 0.370894
## as.factor(NDGMEMG)4     0.072773   0.089683   0.811 0.417111
## as.factor(NDGMEMG)5    -0.056962   0.089521  -0.636 0.524578
## as.factor(NDGMEMG)6     0.103245   0.093614   1.103 0.270082
## as.factor(NDGMEMG)7    -0.068130   0.083389  -0.817 0.413920
## as.factor(NOCPRMG)2     0.113804   0.078454   1.451 0.146897
## as.factor(NOCPRMG)3     0.016329   0.089812   0.182 0.855730
## as.factor(NOCPRMG)4     0.121243   0.081144   1.494 0.135130
## as.factor(NOCPRMG)5     0.176681   0.066340   2.663 0.007738 **
## as.factor(NOCPRMG)6     0.225972   0.060542   3.732 0.000190 ***
## as.factor(NOCPRMG)7     0.180400   0.055779   3.234 0.001220 **
## as.factor(RACETH)2      0.051950   0.039012   1.332 0.182978
## as.factor(RACETH)3      0.072634   0.045146   1.609 0.107644
## as.factor(NBAMEMG)2     0.077284   0.092358   0.837 0.402713
## as.factor(NBAMEMG)3    -0.093117   0.100554  -0.926 0.354428
## as.factor(NBAMEMG)4     0.005443   0.084369   0.065 0.948561
## as.factor(NBAMEMG)5    -0.053496   0.086093  -0.621 0.534348
## as.factor(NBAMEMG)6    -0.049926   0.095177  -0.525 0.599888
## as.factor(NBAMEMG)7     0.039368   0.087428   0.450 0.652499
## as.factor(NBAMEMG)9    -0.424849   0.333140  -1.275 0.202208
## as.factor(NBAMEMG)96   -0.426859   0.112875  -3.782 0.000156 ***
## as.factor(WKSWKGR)2    -0.061870   0.246635  -0.251 0.801925
## as.factor(WKSWKGR)3     0.175471   0.188131   0.933 0.350972
## as.factor(WKSWKGR)4     0.294585   0.179552   1.641 0.100867
## as.factor(HRSWKGR)2     0.258167   0.072883   3.542 0.000397 ***
## as.factor(HRSWKGR)3     0.137876   0.067377   2.046 0.040724 *
## as.factor(HRSWKGR)4     0.304363   0.067396   4.516 6.30e-06 ***
## as.factor(JOBINS)1      0.092297   0.057445   1.607 0.108120
## as.factor(JOBPENS)1     0.046432   0.039669   1.170 0.241807
## as.factor(JOBPROFT)1   -0.155912   0.037548  -4.152 3.29e-05 ***
## as.factor(JOBVAC)1      0.035078   0.052331   0.670 0.502655
## as.factor(FTPRET)1     -0.300602   0.074600  -4.030 5.59e-05 ***
## as.factor(OCEDRLP)2     0.203572   0.033882   6.008 1.88e-09 ***
```

```
## as.factor(OCEDRLP)3   0.387413   0.044112   8.783  < 2e-16 ***
## as.factor(EMSEC)2     0.031914   0.063594   0.502 0.615785
## as.factor(EMSEC)3     0.153634   0.067258   2.284 0.022357 *
## as.factor(EMSEC)4     0.240553   0.056876   4.229 2.34e-05 ***
## as.factor(SATADV)2    0.371115   0.071634   5.181 2.21e-07 ***
## as.factor(SATADV)3    1.232631   0.071217  17.308  < 2e-16 ***
## as.factor(SATADV)4    2.029525   0.073561  27.590  < 2e-16 ***
## as.factor(SATIND)2    0.458395   0.035338  12.972  < 2e-16 ***
## as.factor(SATIND)3    1.171849   0.045648  25.671  < 2e-16 ***
## as.factor(SATIND)4    1.463978   0.073103  20.026  < 2e-16 ***
## as.factor(SATLOC)2    0.118866   0.032774   3.627 0.000287 ***
## as.factor(SATLOC)3    0.455889   0.042061  10.839  < 2e-16 ***
## as.factor(SATLOC)4    0.849456   0.060024  14.152  < 2e-16 ***
## as.factor(SATRESP)2   0.421909   0.043549   9.688  < 2e-16 ***
## as.factor(SATRESP)3   1.284904   0.049985  25.706  < 2e-16 ***
## as.factor(SATRESP)4   1.546687   0.076279  20.277  < 2e-16 ***
## as.factor(SATSAL)2   -0.032565   0.046882  -0.695 0.487295
## as.factor(SATSAL)3    1.137915   0.049232  23.113  < 2e-16 ***
## as.factor(SATSAL)4    2.075284   0.055519  37.380  < 2e-16 ***
## as.factor(SATSEC)2    0.052622   0.037797   1.392 0.163854
## as.factor(SATSEC)3    0.625316   0.043986  14.216  < 2e-16 ***
## as.factor(SATSEC)4    1.094416   0.050980  21.468  < 2e-16 ***
## as.factor(SATSOC)2    0.476159   0.037131  12.824  < 2e-16 ***
## as.factor(SATSOC)3    1.321776   0.045445  29.085  < 2e-16 ***
## as.factor(SATSOC)4    1.734693   0.062006  27.976  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65609  on 97220  degrees of freedom
## Residual deviance: 36060  on 97156  degrees of freedom
## AIC: 36190
##
## Number of Fisher Scoring iterations: 7
```

To maximize job satisfaction, get a proffesional degree as a highest degree, the field of major for highest degree should be in Science and engineering related fields, have a life and related science bachelors degree, work greater than 40 hours per week, work 40-52 weeks out of the year, be in a science and engineering related occupation, and be in the business or industry employer sector.

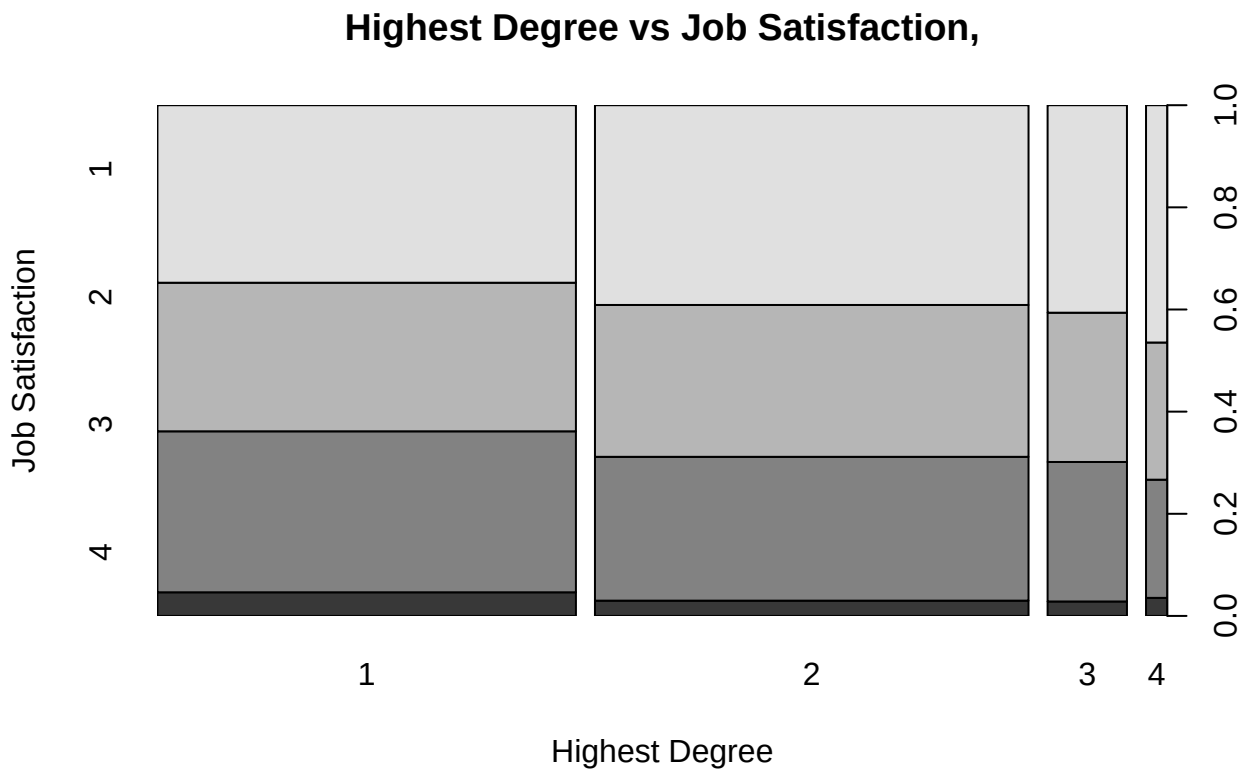
Fact-check news outlets

News outlets regularly examine relationships between degrees, job satisfaction and income. Here are various claims from three different outlets.

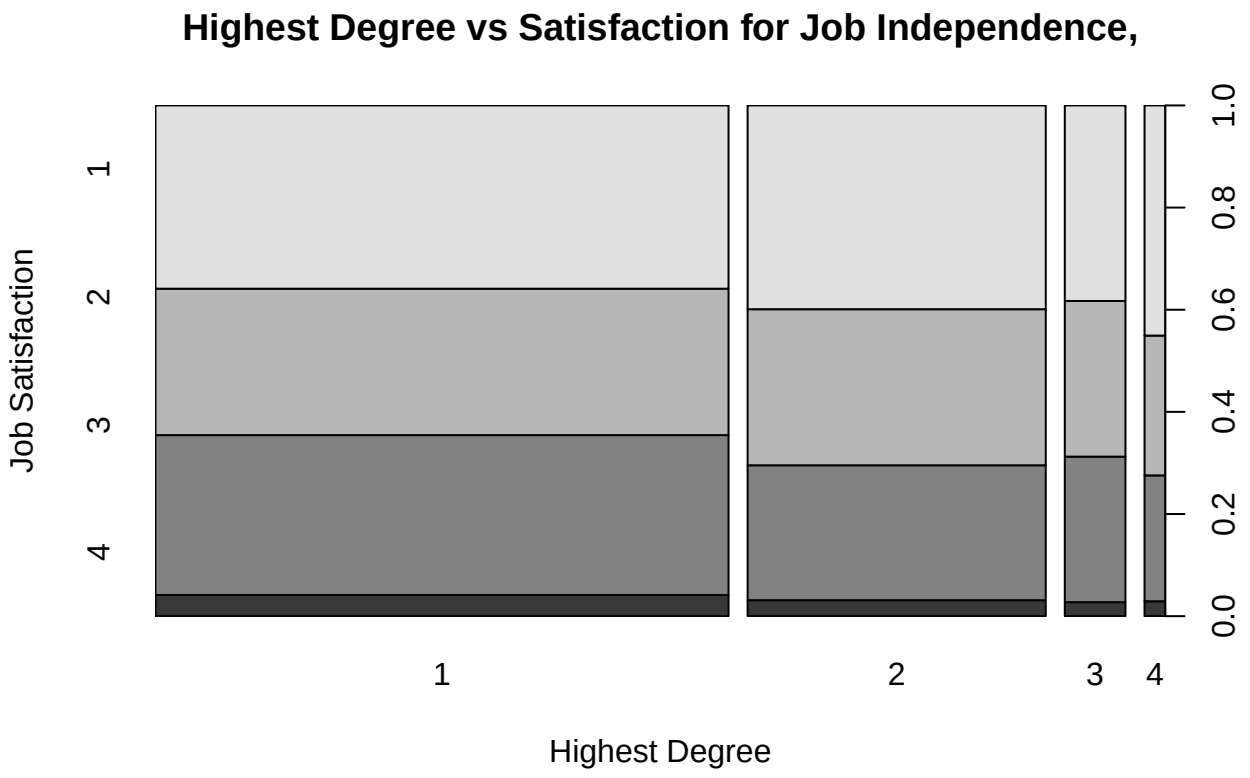
1. Gallup: Does Higher Learning = Higher Job Satisfaction? https://news.gallup.com/poll/6871/does-higher-learning-higher-job-satisfaction.aspx

This article claims that: a. Education level has very little to do with job satisfaction, or satisfaction with income and time flexibility.

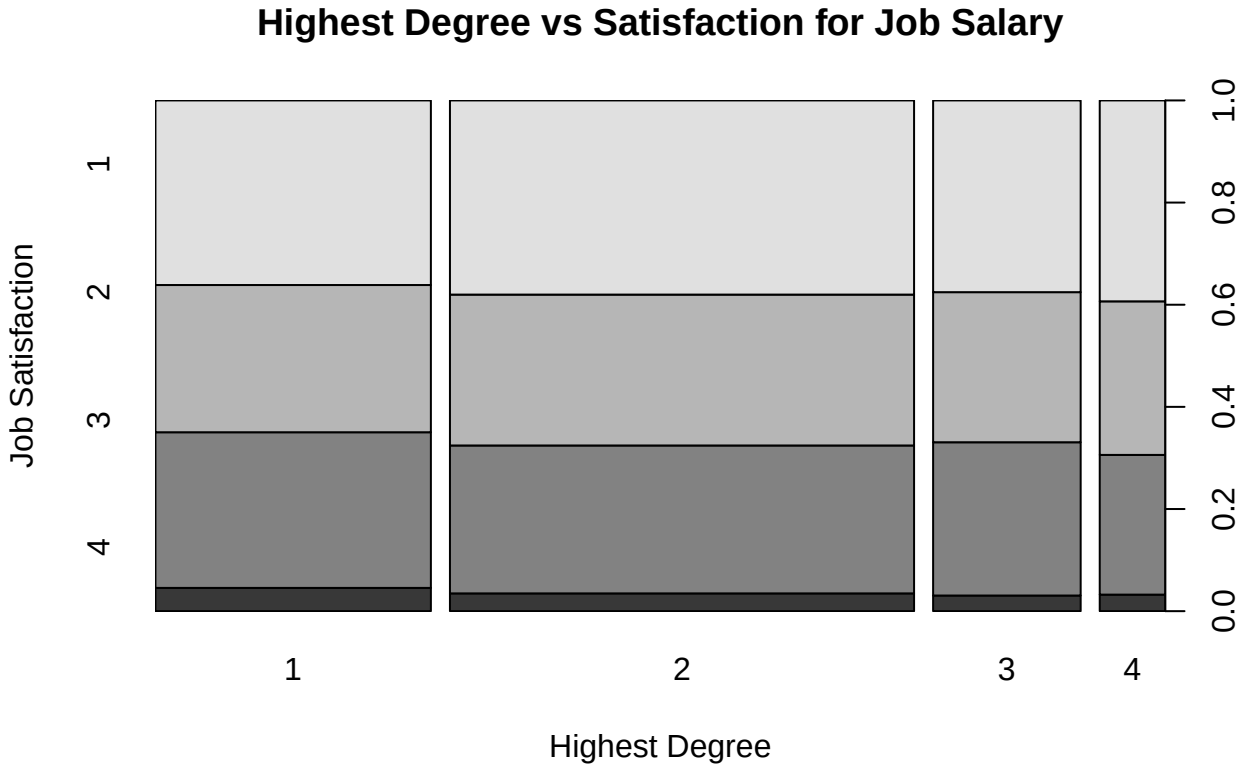
```
plot(dataset$JOBSATIS, dataset$DGRDG, main = "Highest Degree vs Job Satisfaction,", xlab = "Highest Degr
```




```
plot(dataset$SATIND, dataset$DGRDG, main = "Highest Degree vs Satisfaction for Job Independence,", xlab
```




```
plot(dataset$SATSAL, dataset$DGRDG, main = "Highest Degree vs Satisfaction for Job Salary", xlab = "High
```



Looking at the plots above, we are able to see the relationship between "Highest Degree" or Highest Education Level, versus job satisfaction, or satisfaction with income and time flexibility.

Each plot maintains the same scale, therefore we can look at each x value, which represents each of the degree types. Within each degree level, we see the different proportions of job satisfaction. We can clearly see that the proportions of job satisfaction, for each type of satisfaction variable, does not change significantly based on the degree type. Therefore this shows that education level has a very low correlation to job satisfaction, or satisfaction with income and time flexibility,

Therefore from this data we can see that education level has very little to do with job satisfaction, or satisfaction with income and time flexibility

b. Having the opportunity to do what you do best is the one factor that correlates most highly with overall job satisfaction is.

Based on the data analysis from question 7, we are able to see what factors correlate to what degree with overall job satisfaction. Looking at that data, we are able to see that "Degree of Independence" has the highest correlation to overall job satisfaction. Therefore it is acceptable to say that having the opportunity to "do what you do best", is indeed, one of the key factors that correlates most highly with overall job satisfaction.

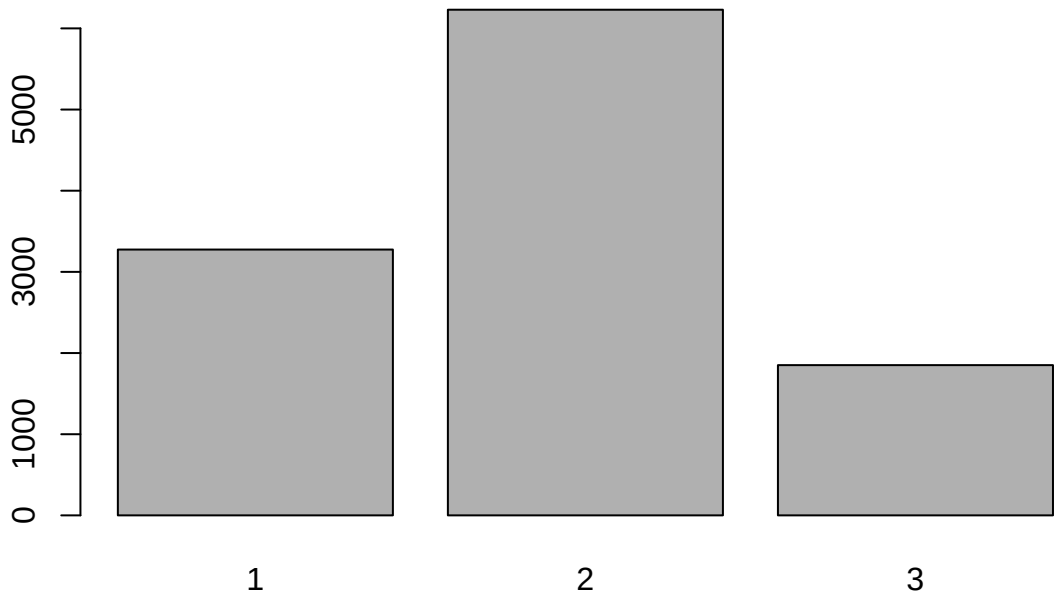
2. Diverse Education: College-educated Americans More Likely Experience Job Satisfaction, Lead Healthier Lives, Study Says https://diverseeducation.com/article/14156/

This article claims that: a. Certain race groups earn less than others when they have the same education level.

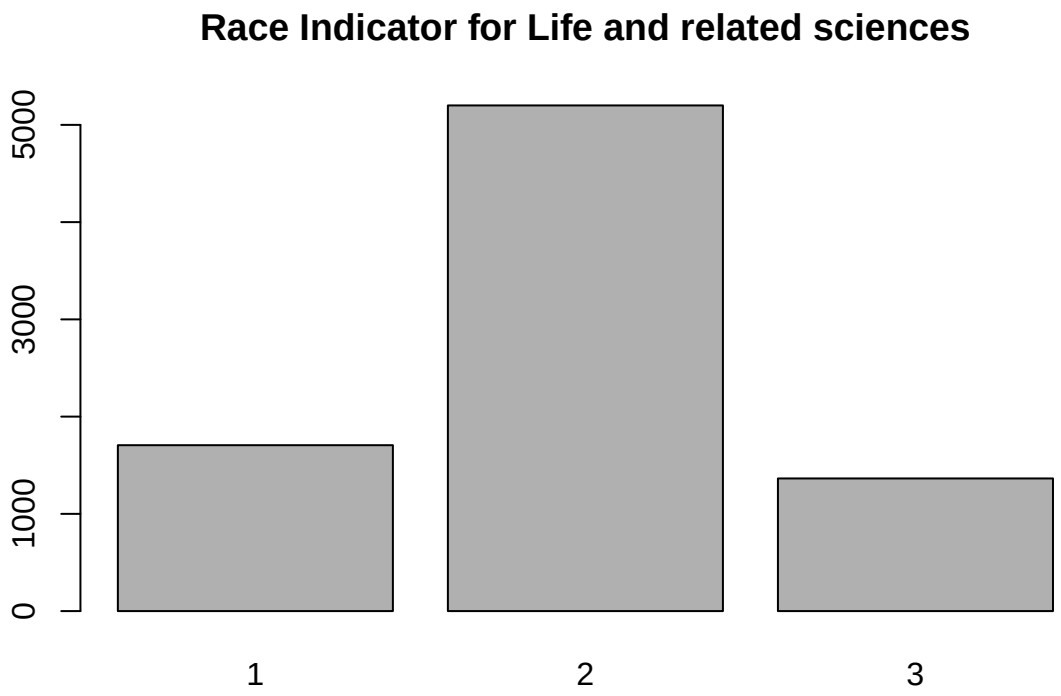
Based on the model created for salary above, each category for RACETH has different coefficients in relation to Salary. Therefore, when you hold education level constant, and solely adjust by race, you will see that certain groups will have less salary than others. For example, based on Linear Regression Model 1, people who classified as white will earn more than those who classified as an Under-represented minority. Hence it is true that certain race groups earn less than others when they have the same education level, looking at the coefficients from linear model 1.

b. STEM (science, technology, engineering and mathematics) careers, in which minorities are underrepresented, tend to pay more than careers in social sciences.

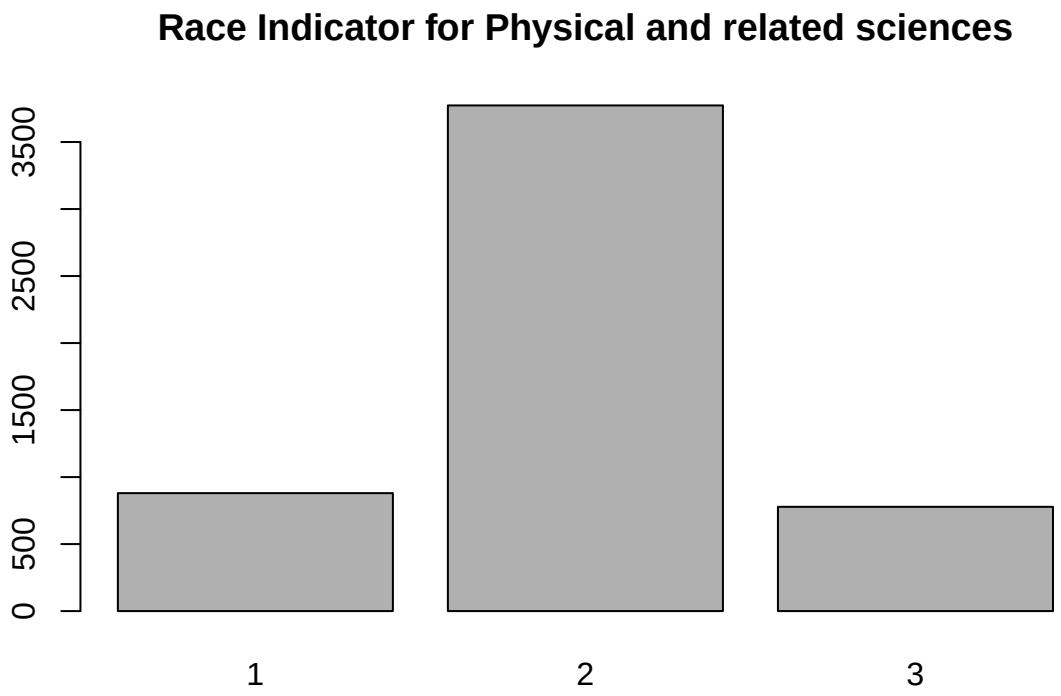
```
#Create Subsets for Each Field of Highest Study
subset.field1 <- subset(dataset, dataset$NOCPRMG == 1)
subset.field2 <- subset(dataset, dataset$NOCPRMG == 2)
subset.field3 <- subset(dataset, dataset$NOCPRMG == 3)
subset.field5 <- subset(dataset, dataset$NOCPRMG == 5)
subset.field6 <- subset(dataset, dataset$NOCPRMG == 6)


#Plots race with each field to check proportions of minorities and other races
plot(subset.field1$RACETH, main = "Race Indicator for Computer and mathematical sciences")
```



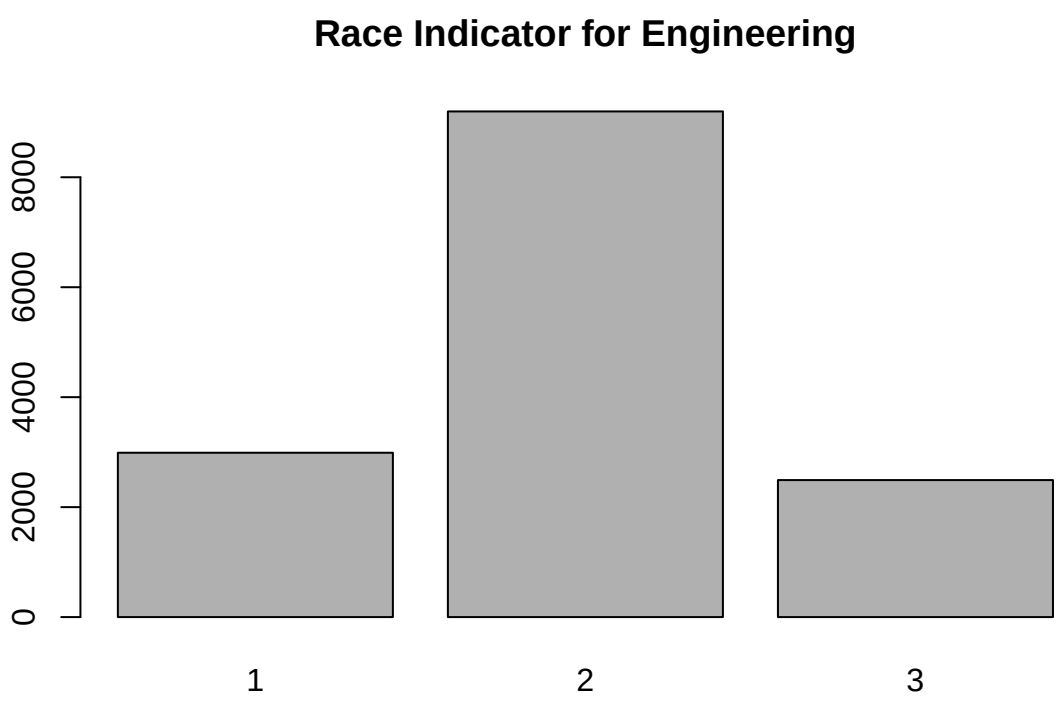

```
plot(subset.field2$RACETH, main = "Race Indicator for Life and related sciences")
```



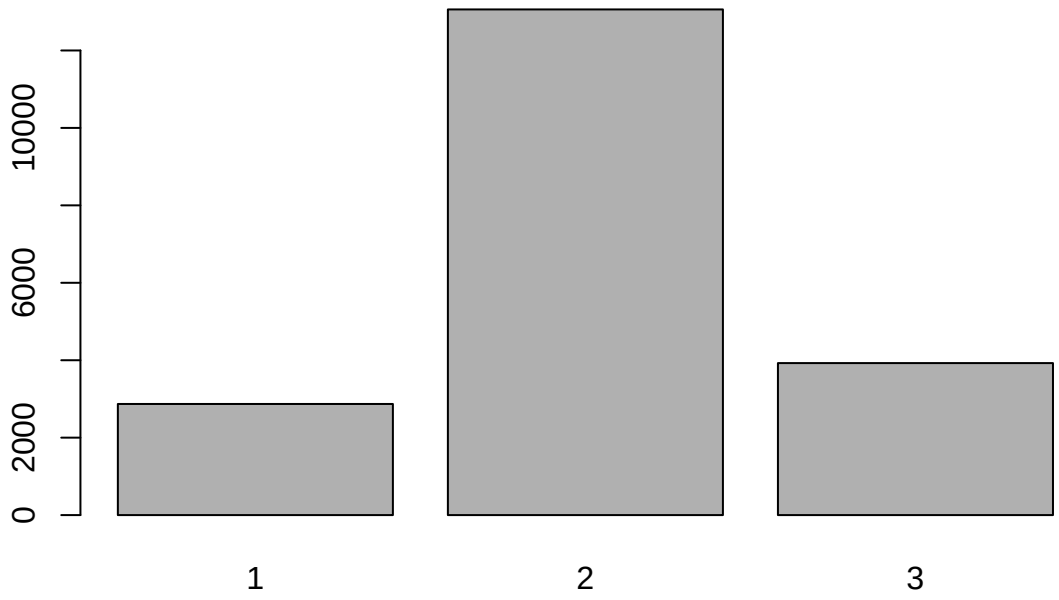

```
plot(subset.field3$RACETH, main = "Race Indicator for Physical and related sciences")
```




```
plot(subset.field5$RACETH, main = "Race Indicator for Engineering")
```




```
plot(subset.field6$RACETH, main = "Race Indicator for Science and engineering-related fields")
```




```
#Create a subset of all stem fields where salary != NA
sub.stem <- subset(dataset, !is.na(dataset$SALARY) & (dataset$NOCPRMG == 1 | dataset$NOCPRMG == 2 | data

#Create a subset of all social fields where salary != NA
sub.social <- subset(dataset, !is.na(dataset$SALARY) & (dataset$NOCPRMG == 4 | dataset$NOCPRMG == 7))


#Find average of stem related salary compared to average of social related salary
stem.salary.total <- sum(sub.stem$SALARY)
stem.salary.average <- stem.salary.total / nrow(sub.stem)

social.salary.total <- sum(sub.social$SALARY)
social.salary.average <- social.salary.total / nrow(sub.social)

print("Stem Salary Average")
```

```
## [1] "Stem Salary Average"
```

```
stem.salary.average
```

```
## [1] 81330.6
```

```
print("Social Salary Average")
```

```
## [1] "Social Salary Average"
```

```
social.salary.average
```

```
## [1] 69880.23
```

In this question, we are asked to understand two portions.

1. minorities are underrepresented in STEM.
2. STEM careers pay better than social sciences

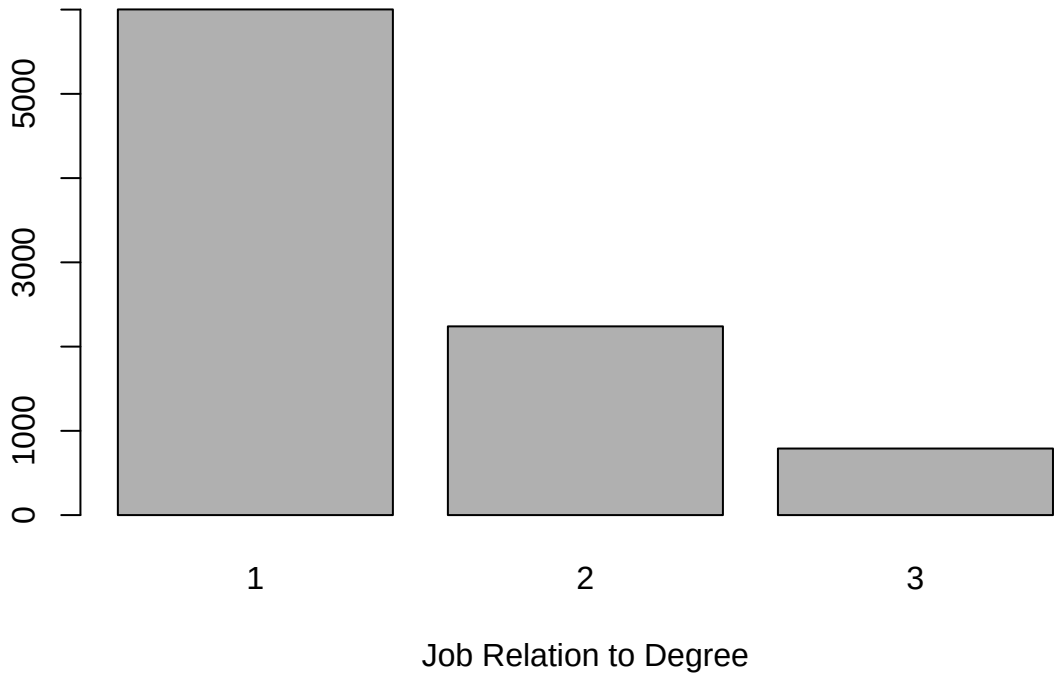
To address that minorities that are underrespresented in STEM, we have created bar plots to show the RACETH distribution for every type of stem related field. In this analysis, we want to look at choice "3", or "Underrepresented Minorities". In each field, choice 3 was the smallest proportion, or close to the smallest proportion, of race types in each of the STEM fields. Therefore it can be fairly concluded that minorities are underrepresented in STEM related fields.

To address the idea that STEM careers pay better than social careers, we take a subset of stem related fields, then a subset of social related fields, then get average salaries for both subsets. Then in comparison of the averages, the Stem Salary Average is 81330.6 while the Social Salary Average is 69880.23. Therefore it can be concluded that STEM careers pay better than social science.

3. PEW: the rising cost of not going to college https://www.pewsocialtrends.org/2014/02/11/the-rising-cost-of-not-going-to-college/

This article claims that: a. Those who studied science or engineering are the most likely to say that their current job is "very closely" related to their college or graduate field of study.

```
field1.data <- subset(dataset, dataset$NDGMEMG == 1)
field1.plot <- plot(field1.data$OCEDRLP, xlab="Job Relation to Degree", main = "Field 1 : Computer and I
```

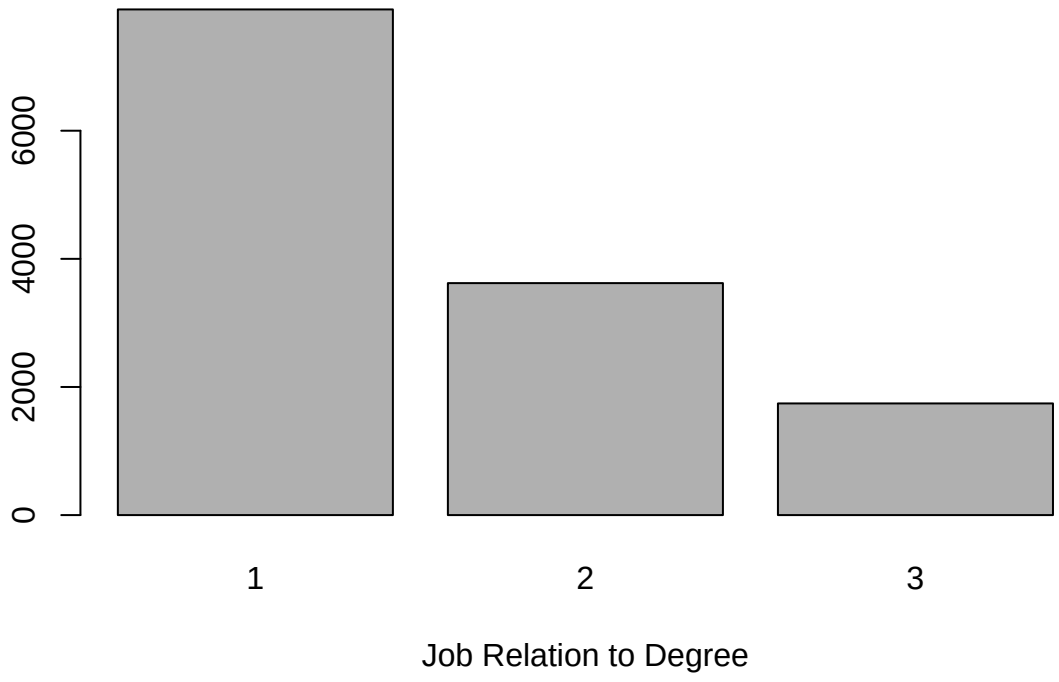



```
field1.table <- table(field1.data$OCEDRLP)
field1.table
```

```
##
##    1    2    3
## 6002 2240  790
```

```
field2.data <- subset(dataset, dataset$NDGMEMG == 2)
field2.plot <- plot(field2.data$OCEDRLP, xlab="Job Relation to Degree", main = "Field 2 : Biological, ag
```

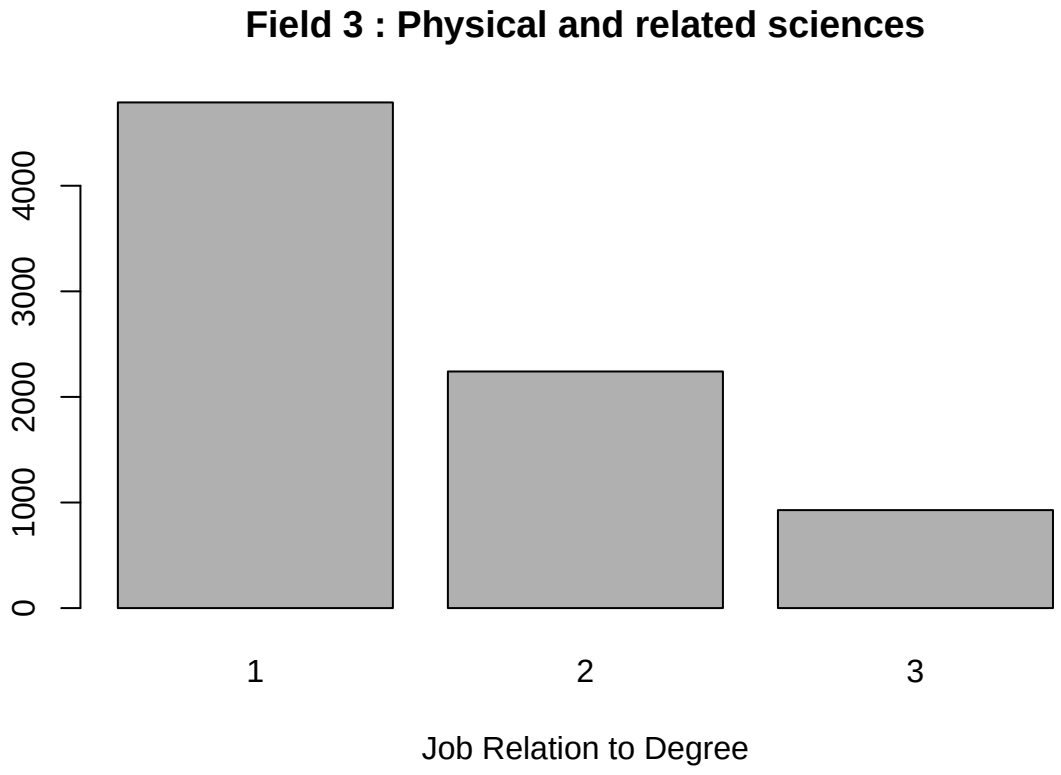
Field 2 : Biological, agricultural and environmental life sciences




```
field2.table <- table(field2.data$OCEDRLP)
field2.table
```

```
##
##    1    2    3
## 7893 3621 1743
```

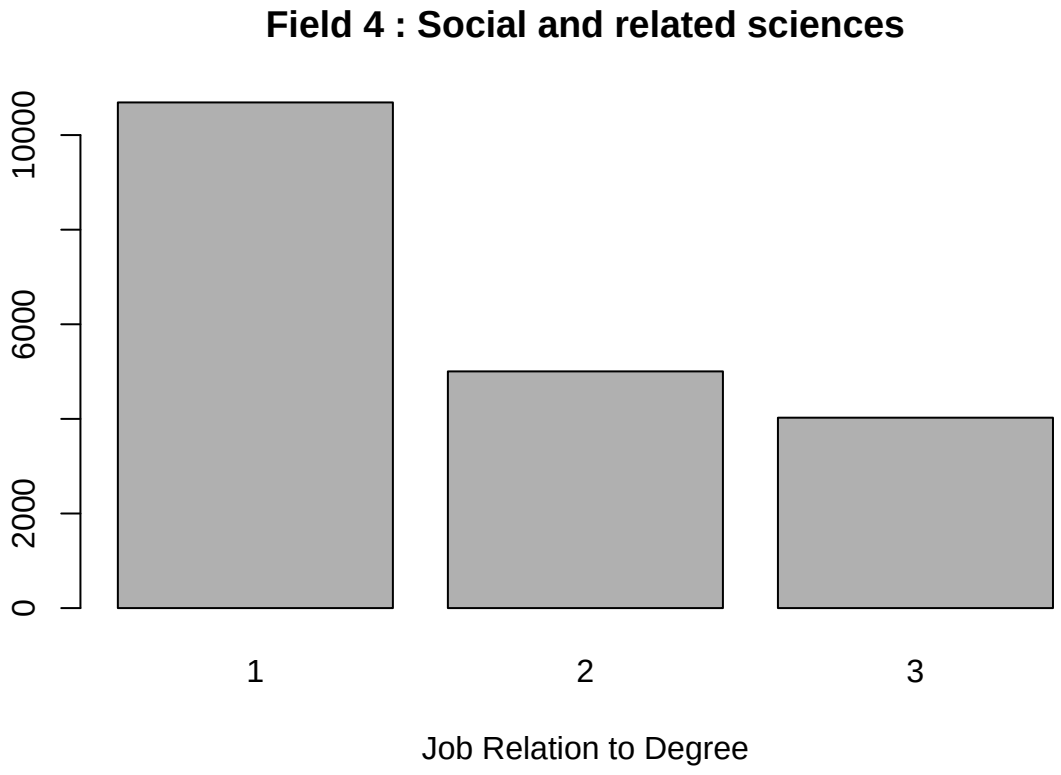
```
field3.data <- subset(dataset, dataset$NDGMEMG == 3)
field3.plot <- plot(field3.data$OCEDRLP, xlab="Job Relation to Degree", main = "Field 3 : Physical and
```




```
field3.table <- table(field3.data$OCEDRLP)
field3.table
```

```
##
##    1    2    3
## 4789 2241  928
```

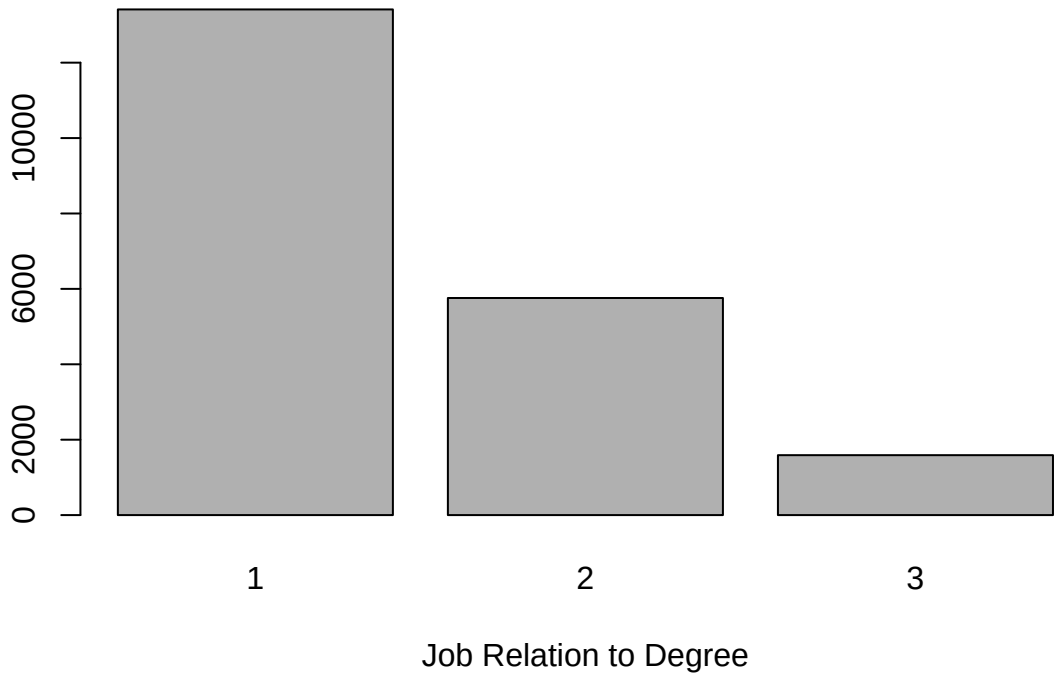
```
field4.data <- subset(dataset, dataset$NDGMEMG == 4)
field4.plot <- plot(field4.data$OCEDRLP, xlab="Job Relation to Degree", main = "Field 4 : Social and rel
```




```
field4.table <- table(field4.data$OCEDRLP)
field4.table
```

```
##
##     1     2     3
## 10690  5004  4025
```

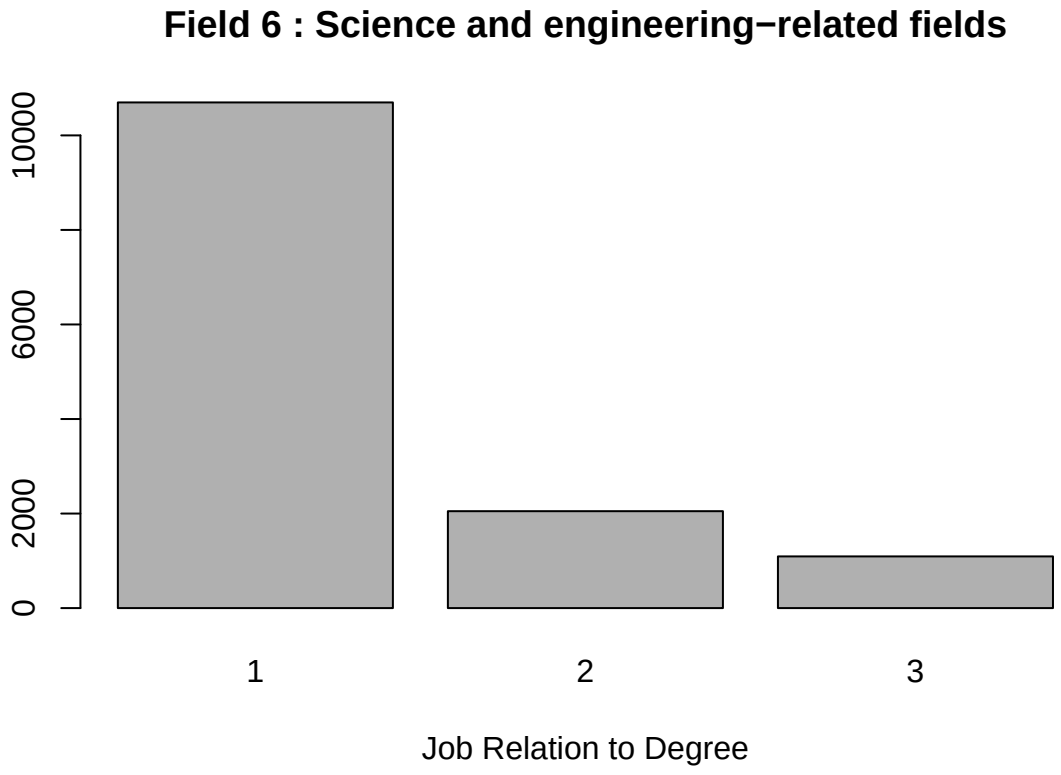
```
field5.data <- subset(dataset, dataset$NDGMEMG == 5)
field5.plot <- plot(field5.data$OCEDRLP, xlab="Job Relation to Degree", main = "Field 5 : Engineering")
```




```
field5.table <- table(field5.data$OCEDRLP)
field5.table
```

```
##
##     1     2     3
## 13412  5758  1589
```

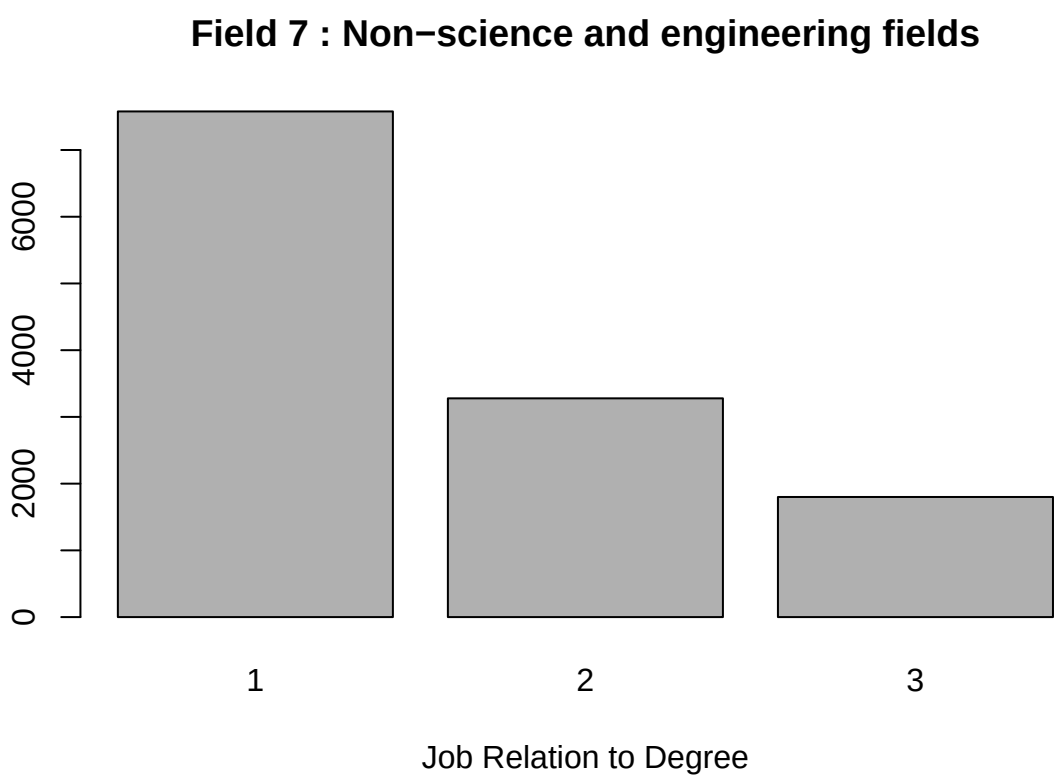
```
field6.data <- subset(dataset, dataset$NDGMEMG == 6)
field6.plot <- plot(field6.data$OCEDRLP, xlab="Job Relation to Degree", main = "Field 6 : Science and e
```




```
field6.table <- table(field6.data$OCEDRLP)
field6.table
```

```
##
##     1     2     3
## 10697  2050  1093
```

```
field7.data <- subset(dataset, dataset$NDGMEMG == 7)
field7.plot <- plot(field7.data$OCEDRLP, xlab="Job Relation to Degree", main = "Field 7 : Non-science ar
```




```
field7.table <- table(field7.data$OCEDRLP)
field7.table
```

```
##
##    1    2    3
## 7578 3278 1800
```

Upon plotting the relationship between each highest field and job relation to highest degree, we are able to see the distribution for relationship to each degree field.

Based on the data above, those who studied scientific majors have the highest amount of individuals choosing choice 1, which means that is the most "closely related" field of study to current job.

Hence, it is true that those who studied science or engineering are the most likely to say that their current job is "very closely" related to their college or graduate field of study.

1. For each of the claim above, use your analysis above to verify or disprove it.

2. If you disprove any claims, explain why your conclusions could be different from theirs. For example, you could elaborate on major differences between the dataset you are using and the survey used by the article, or your method of analysis vs theirs.

Lay summary

Give a two to three-page summary to highlight the findings in the technical report for the general public. Your summary should contain four sections:

- highlights from the basic analysis
- highlights from the salary model
- highlights from the job satisfaction model
- highlights from the fact-check section