Final project: Is College Worth It?

Due date: December 9, 2019

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.

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 , 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)?

In the National Survey of College Graduates(NSCG), the population consists of individuals residing in the United States, under the age of 76, who hold a bachelor's degree or higher. Its sample is a two-stage sampling scheme : ACS households called simple random sample in 2010, and its sampling method is stratified systematic (sampling method). The samples have biases that scientists and engineers did not receive higher education in the United States. Also, there is no information about individuals without a science or engineering post-secondary degree, who are not currently working in a science or engineering occupation.

In the Survey of Doctorate Recipients (SDR), the popularation consists of individuals residing who earned a doctorate degree in science, engineering, or health in the United States. Its sample is a stratified sample from the eligible individuals in the Doctorate Records File, and its sampling method is stratified (sampling method). The samples have biases that the sample sizes of SDR has decreased since the more respondents of SDR currently live abroad, and their data is to manage a separate file:ISDR. Also, there is a non-response bias since the surveys used web surveys for data collection later.

We introduce an additional bias when we combine the surveys of NSCG and SDR. We do not know the ratio of populations between those who hold at least bachelor's degree and doctorate degree between the two surveys. Besides, as the sample size of SDR survey decreases, the ratio of populations about those who hold at least bachelor's degree in NSCG survey and doctorate degree in SDR do not balance. As the sample size of SDR survey decreases, the ratio of populations between those who only hold bachelor's degree to those who hold either Master's degree, Doctorate degree, or Professional degree may not be well representative.

Demographics

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

```
# show the demographic of the survey as mosaic plot
load("dataset.RData")
data.df <- dataset
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 3.5.3
```

```
## Loading required package: grid
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

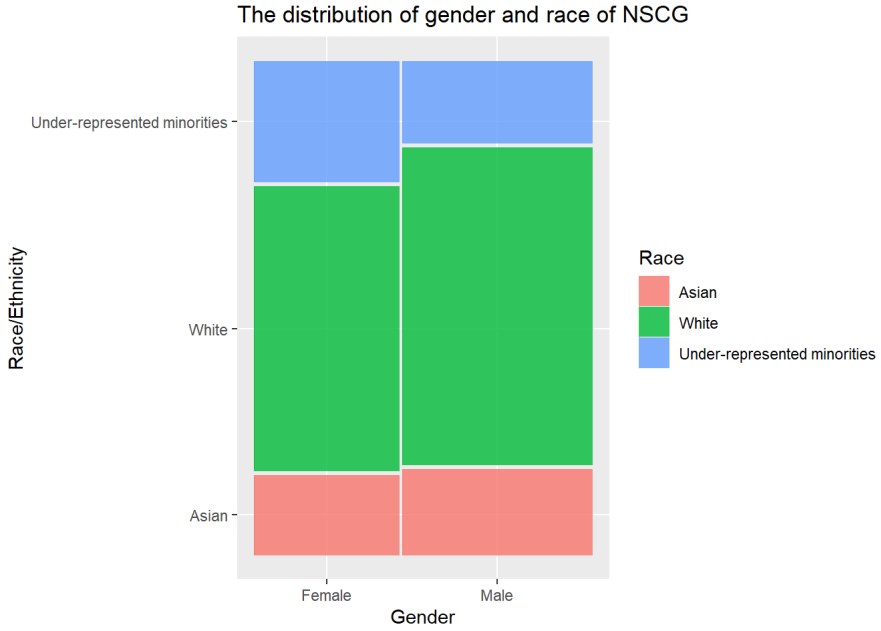
```
library(ggmosaic)
```

```
## Warning: package 'ggmosaic' was built under R version 3.5.3
```

```
##
## Attaching package: 'ggmosaic'
```

```
## The following objects are masked from 'package:vcd':
##
##     mosaic, spine
```

```
ggplot(data = data.df) + geom_mosaic(aes(x = product(RACETH,GENDER), fill=RACETH), na.rm = TRUE)+labs(x="Gender", y="Race/Ethnicity", title = 'The distribution of gender and race of NSCG')+scale_x_productlist(labels = c("Female","Male"))+scale_y_productlist(labels=c("Asian","White","Under-represented minorities"))+scale_fill_discrete(name = "Race", labels = c("Asian", "White", "Under-represented minorities"))
```




```
# show the table of demographic of the survey as matrix
demo.matrix <- 100*prop.table(table(data.df$GENDER, data.df$RACETH))
colnames(demo.matrix) <- c("Asian", "White", "Under-represented minorities")
rownames(demo.matrix) <- c("female", "male")
demo.matrix
```

```
##
##             Asian     White Under-represented minorities
```

```
##   female  7.076733 25.481103                  10.792691
##   male   10.013721 37.039739                   9.596012
```

```
# Sequentially, the sum of female population % and the sum of male population %.(43.35%, 56.64%)
sum(demo.matrix[1,1],demo.matrix[1,2],demo.matrix[1,3])
```

```
## [1] 43.35053
```

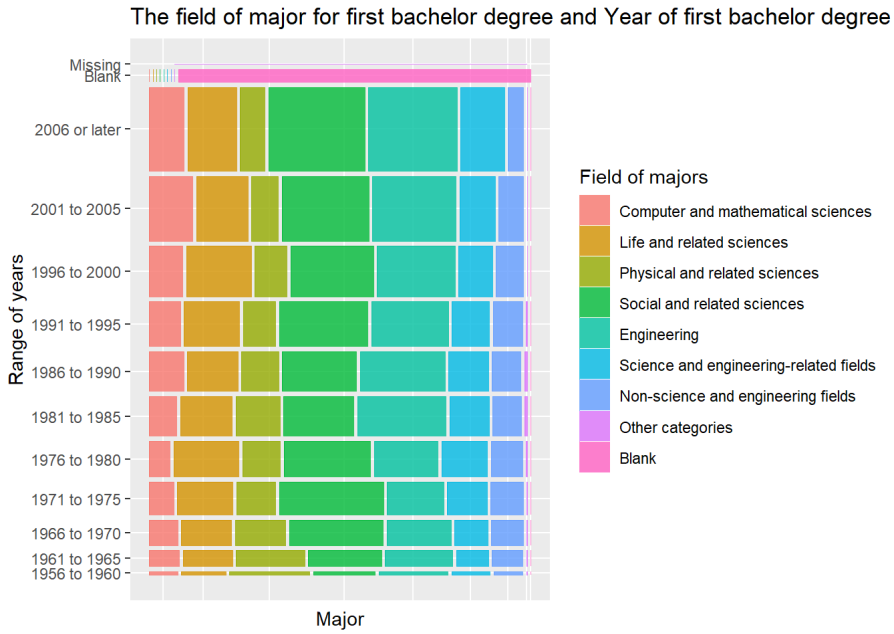
```
sum(demo.matrix[2,1],demo.matrix[2,2],demo.matrix[2,3])
```

```
## [1] 56.64947
```

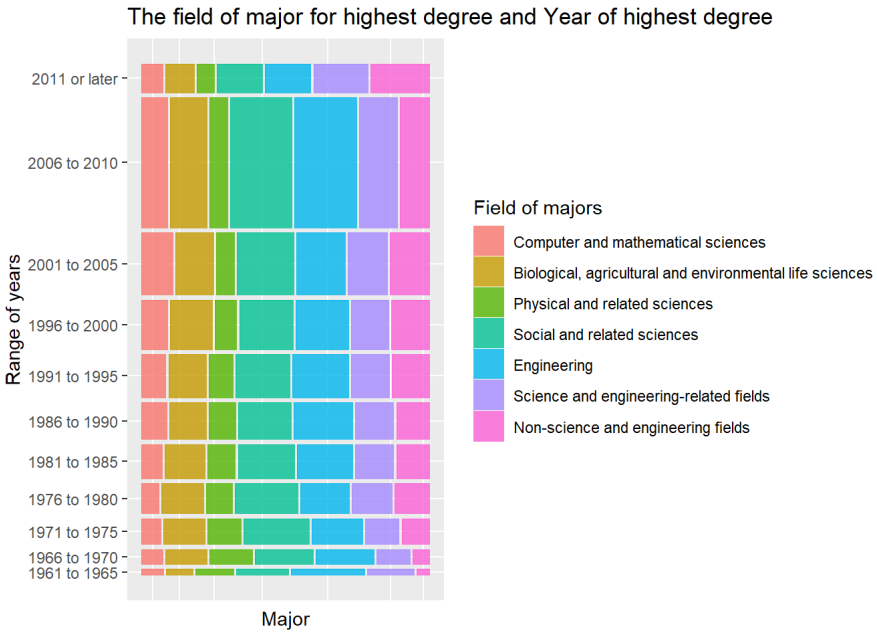
In the survey, the population of male is around 43.35% , and that of female is around 56.65%. As we observe, the population of male is around 13.3% more than that of female regardless of race. White occupies around 62.52%, Asian occupied 17.09%, and Under-represented minorities occupies 20.39% in the proportion table of race and ethnicity. We see that White is the highest population in both male and female. Also, the lowest population in male is Under-represented minorities(9.596%), and the Asian female is the lowest population in female(7.0767%). Considering overall race/ethnicity, the majority of race/ethnicity is White regardless of gender whereas the minority of race/ethnicity is Asian.
Education

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

```
# set the mosaic plot for the field of major for first bachelor degree and Year of first bachelor degree
  ggplot(data = data.df) + geom_mosaic(aes(x = product(NBAMEMG,BA03Y5), fill=NBAMEMG), na.rm = TRUE)+labs(x="Rang
e of years",y="Major", title = 'The field of major for first bachelor degree and Year of first bachelor degree')+
scale_x_productlist(labels=c("1956 to 1960","1961 to 1965","1966 to 1970","1971 to 1975","1976 to 1980","1981 to
 1985","1986 to 1990","1991 to 1995","1996 to 2000","2001 to 2005","2006 or later","Blank","Missing"))+scale_fill
_discrete(name = "Field of majors", labels = c("Computer and mathematical sciences", "Life and related sciences",
"Physical and related sciences", "Social and related sciences","Engineering","Science and engineering-related fie
lds","Non-science and engineering fields","Other categories","Blank"))+coord_flip()+theme(axis.text.x = element_b
lank(), axis.ticks.x = element_blank())
```




```
# set the mosaic plot for the field of major for highest degree and Year of highest degree
  ggplot(data = data.df) + geom_mosaic(aes(x = product(NDGMEMG,HD03Y5), fill=NDGMEMG), na.rm = TRUE)+labs(x="Rang
e of years",y="Major", title = 'The field of major for highest degree and Year of highest degree')+scale_x_produc
tlist(labels=c("1961 to 1965","1966 to 1970","1971 to 1975","1976 to 1980","1981 to 1985","1986 to 1990","1991 to
 1995","1996 to 2000","2001 to 2005","2006 to 2010","2011 or later"))+scale_fill_discrete(name = "Field of major
s", labels = c("Computer and mathematical sciences", "Biological, agricultural and environmental life sciences",
"Physical and related sciences", "Social and related sciences","Engineering","Science and engineering-related fie
lds","Non-science and engineering fields"))+coord_flip()+theme(axis.text.x = element_blank(), axis.ticks.x = elem
ent_blank())
```



In the field of major for first bachelor degree and Year of first bachelor degree, none of the sample of all majors obtained a major more than 2006 or later. We also observe that more population of all majors gradually obtain a major every range of years. Between 1956 and 1960, the people of sample obtained Physical and related sciences degree more than others. Between 1961 and 1980, the most people obtained Social and related science degrees more than others. Between 1981 and 1990, the most people obtained Engineering degrees more than others. Between 1991 and now, the most people obtained Social and related science degrees more than others.

In the field of major for highest degree and Year of highest degree, none of the people of sample in all majors obtained major degrees more than 2006 to 2010. The more people obtained the highest degree in all majors every range of years, but the size of sample is sharply decreased in 2011 or later. The most people of the sample obtained Engineering degrees between 1961 and 1965. Between 1966 and 2010, most of people obtained Social and related sciences or Engineering.

4. For those who obtained more than a bachelor degree, is there a significant association between field of major between their bachelor degree and their highest degree? State any tests you use, your p-value, and draw conclusions.

We set hypothesis tests that H_0 : There is not a significant difference in retention rates among different field of majors, and H_A : There is a significant difference in retention rates among different field of majors. To check the hypotheses, we would carry out chi-square test.

```
# Before doing Chi-Square, we need to make another variable that the people of the sample obtain same major in ba
chelor degree and at least master degree or not.
# initialize a new variable as 0
data.df$R.Rate <- 0
# 1 means the people in the sample have same degree from bachelors while 0 does not.
data.df$R.Rate[(data.df$DGRDG[] != 1) & (as.numeric((data.df$NBAMEMG[]) == as.numeric(data.df$NDGMEMG[])))] <- 1
# Next, we should consider excluding numerical value about 96("Blank") for bachelor's degree.
# test it what p-value outputs as chi-square.
chisq.test(data.df$R.Rate[data.df$NBAMEMG[] != 96], data.df$NBAMEMG[data.df$NBAMEMG!= 96])
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.df$R.Rate[data.df$NBAMEMG[] != 96] and data.df$NBAMEMG[data.df$NBAMEMG != 96]
## X-squared = 3828.2, df = 7, p-value < 2.2e-16
```

Through the Chi-Square test, the p-value is less than 2.2e-16 which is closed to 0. Since the p-value is less than significant level(0.05), we could contradict H_0. Therefore, 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.
- 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?
- What are the major reasons that led people to not work at the time of survey?

```
# find the % of people working
sum(data.df$LFSTAT=="1")*100/length(data.df$LFSTAT)
```

```
## [1] 85.14919
```

```
# find the % of people part-time working
round((length(which(data.df$PTWTFT=="0"))+length(which(data.df$PTWTFT=="1")))*100/length(which(data.df$LFSTAT==1
)),2)
```

```
## [1] 13.24
```

```
# find the % number of hours per week and weeks per year
# set the proportion table
work.time.matrix <- 100*prop.table(table(data.df$HRSWKGR,data.df$WKSWKGR))
rownames(work.time.matrix) <- c("20 or less", "21 - 35", "36 - 40", "over 40")
colnames(work.time.matrix) <- c("1-10 weeks", "11 - 20 weeks", "21 - 39 weeks", "40-52 weeks")
work.time.matrix
```

```
##
##              1-10 weeks 11 - 20 weeks 21 - 39 weeks 40-52 weeks
##   20 or less  0.26618800    0.46098459    1.39009291  5.37781359
##   21 - 35     0.11116664    0.12952443    0.82202119  7.20849354
##   36 - 40     0.13564370    0.10198774    1.25342934 36.26378109
##   over 40     0.13564370    0.07649081    2.84953749 43.41720125
```

```
# find the major reasons that led people to not work
# find the number of people: Reasons for not working about family responsibilities
table(data.df$NWFAM)[[2]]
```

```
## [1] 2766
```

```
# find the number of people: Reasons for not working about layoff
table(data.df$NWLAY)[[2]]
```

```
## [1] 1642
```

```
#find the number of people: Reasons for not working that they do not need or want to work
table(data.df$NWNOND)[[2]]
```

```
## [1] 3762
```

```
# find the number of people: Reasons for not working that there is no suitable job available
table(data.df$NWOCNA)[[2]]
```

```
## [1] 2584
```

```
# find the number of people: Reasons for not working: illness, retired or other (combined)
table(data.df$NWOTP)[[2]]
```

```
## [1] 10710
```

```
# find the number of people: Reasons for not working: student
table(data.df$NWSTU)[[2]]
```

```
## [1] 1948
```

In the general statistics, 85.14% of people are working, and 13.24% of working people have part-time job. Around 7.5% of people work 20 or less hours, around 8.27% of people work 21 ~ 35 hours, around 37.75% of people work 36 ~ 40 hours, around 46.48% of people work over 40 hours per week. We see that the majority of people work over 40 hours per week. Around 0.65% of people work 1 ~ 10 weeks, around 0.77% of people work 11 ~ 20 weeks, around 6.32% of people work 21 ~ 39 weeks, and around 92.27% of people work 40 ~ 52 weeks per year. We see that the majority of people work 40 ~ 52 weeks per year.
Through the matrix of the number of hours per week and weeks per year, we know that most people work with regular hours year-round. This is because around 79.6% of people work full time job and 40 ~ 52 weeks per year. Also, there is tiny rate around 0.136% of people work over 40 hours per week, and 1 ~ 10 weeks per year. It means really few people work in short bursts. Therefore, the majority of people work with regular hours year-round.
At the time of survey, 2766 people do not work due to family responsibilities, 1642 people do not work due to layoff. 3762 people do not work since they do not need or want to do. 2584 people do not work since they look for a suitable job. 10710 of people do not work since they are ill, retire, or have other(combined). 1948 people do not work since they are students. The majority of people do not work due to illness, retirement, or other(combined). As a result, illness, retirement, or other(combined) are the major reasons that led people to not work at the time of 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?).
- 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.

We set the first hypothesis test that H_0 : Degree relevance is independent for job type, and H_A : Degree relevance is dependent for job type. We set the second hypothesis test that H_0 : Degree relevance is independent for degree that they trained for, and H_A : Degree relevance is dependent for degree that they trained for. We set the third hypothesis test that H_0 : Degree relevance is independent for principal activity in people's job, and H_A : Degree relevance is dependent for principal activity in people's job. To conduct those three 3 hypothesis test, we use Chi-Square test method.

```
# set the chi-square test for degree relevance(OCEDRLP : Principal job related to highest degree) vs job type(NOC
PRMG :     Job code for principal job (major group))
chisq.test(data.df$OCEDRLP, data.df$NOCPRMG)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.df$OCEDRLP and data.df$NOCPRMG
## X-squared = 12359, df = 12, p-value < 2.2e-16
```

```
# set the chi-square test for degree relevance(OCEDRLP : Principal job related to highest degree) vs degree that
 they trained for(NDGMEMG : Field of major for highest degree (major group))
chisq.test(data.df$OCEDRLP, data.df$NDGMEMG)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.df$OCEDRLP and data.df$NDGMEMG
## X-squared = 3269, df = 12, p-value < 2.2e-16
```

```
# set the chi-square test for degree relevance(OCEDRLP : Principal job related to highest degree) vs principal ac
tivity in people's job(WAPRSM  : Summarized primary work activity)
chisq.test(data.df$OCEDRLP, data.df$WAPRSM)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.df$OCEDRLP and data.df$WAPRSM
## X-squared = 7890.5, df = 8, p-value < 2.2e-16
```

In the first hypothesis test: OCEDRLP vs NOCPRMG, the p-value is 2.2e-16 which is closed to 0, so we could contradict H_0. We conclude that degree relevance is dependent for job type. In the second hypothesis test: OCEDRLP vs NDGMEMG, the p-value is 2.2e-16 which is closed to 0, so we could contradict H_0. We conclude that degree relevance is dependent for degree that they trained for. In the third hypothesis test: OCEDRLP vs WAPRSM, the p-value is 2.2e-16 which is closed to 0, so we could contradict H_0. We conclude that degree relevance is dependent for principal activity in people's job.

7. Job satisfaction

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

```
#set the overall job satisfaction
ovl.sat.table <- table(data.df$JOBSATIS)
```

```
rownames(ovl.sat.table) <- c("Very satisfied", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfied"
)
ovl.sat.table
```

```
##
##        Very satisfied    Somewhat satisfied Somewhat dissatisfied
##                 43147                 44508                  8178
##     Very dissatisfied
##                  2218
```

```
# Since we need to integrate Very satisfied/somewhat satisfied and Very dissatisfied/somewhat dissatisfied, we cr
eate a new category to simplify satisfied or dissatisfied.
data.df$Ovl.JOBSATIS <- 0
data.df$Ovl.JOBSATIS[data.df$JOBSATIS == "3" | data.df$JOBSATIS == "4"] <- 1

# set a proportional table for the overall job satisfaction vs Satisfaction principal job's opportunities for adv
ancement
ovl.jopp.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATADV))
colnames(ovl.jopp.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.jopp.table) <- c("satistifed", "dissatisfied")
ovl.jopp.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed        24.4699187         41.6140580            18.3516741
##   dissatisfied       0.2927048          1.5848895             3.8102620
##
##                Very dissatisfied
##   satistifed           4.9617036
##   dissatisfied         4.9147892
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.jopp.table[1,1]*100/sum(ovl.jopp.table[1, ])
```

```
## [1] 27.37208
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.jopp.table[2,4]*100/sum(ovl.jopp.table[2, ])
```

```
## [1] 46.35437
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job benefits
ovl.jben.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATBEN))
colnames(ovl.jben.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.jben.table) <- c("satistifed", "dissatisfied")
ovl.jben.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed         36.342312          38.185230              9.732690
##   dissatisfied        1.813342           3.875534              2.328380
##
##                Very dissatisfied
##   satistifed            5.137123
##   dissatisfied          2.585389
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.jben.table[1,1]*100/sum(ovl.jben.table[1, ])
```

```
## [1] 40.65256
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.jben.table[2,4]*100/sum(ovl.jben.table[2, ])
```

```
## [1] 24.38438
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job's intellectual challeng
e
ovl.int.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATCHAL))
colnames(ovl.int.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.int.table) <- c("satistifed", "dissatisfied")
ovl.int.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed         45.820032          34.688070              7.556272
##   dissatisfied        1.232012           3.096348              3.427808
##
##                Very dissatisfied
##   satistifed            1.332980
##   dissatisfied          2.846478
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.int.table[1,1]*100/sum(ovl.int.table[1, ])
```

```
## [1] 51.25435
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.int.table[2,4]*100/sum(ovl.int.table[2, ])
```

```
## [1] 26.84686
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job's degree of independenc
e
```

```
ovl.ind.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATIND))
colnames(ovl.ind.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.ind.table) <- c("satistifed", "dissatisfied")
ovl.ind.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed         57.7311807          27.0359303              3.9948598
##   dissatisfied        2.4640238           4.1916962              2.3661156
##
##                Very dissatisfied
##   satistifed            0.6353836
##   dissatisfied          1.5808100
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.ind.table[1,1]*100/sum(ovl.ind.table[1, ])
```

```
## [1] 64.57818
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.ind.table[2,4]*100/sum(ovl.ind.table[2, ])
```

```
## [1] 14.90958
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job's job location
ovl.loc.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATLOC))
colnames(ovl.loc.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.loc.table) <- c("satistifed", "dissatisfied")
ovl.loc.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed          52.282996           27.813077              7.590948
##   dissatisfied         3.733771            3.706235              1.894932
##
##                Very dissatisfied
##   satistifed            1.710334
##   dissatisfied          1.267708
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.loc.table[1,1]*100/sum(ovl.loc.table[1, ])
```

```
## [1] 58.48383
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.loc.table[2,4]*100/sum(ovl.loc.table[2, ])
```

```
## [1] 11.95652
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job's level of responsibili
ty
ovl.res.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATRESP))
colnames(ovl.res.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.res.table) <- c("satistifed", "dissatisfied")
ovl.res.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed        47.4834525         35.9435396             5.3492570
##   dissatisfied       1.1769385          4.0682910             3.5348951
##
##                Very dissatisfied
##   satistifed           0.6211053
##   dissatisfied         1.8225209
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.res.table[1,1]*100/sum(ovl.res.table[1, ])
```

```
## [1] 53.11505
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.res.table[2,4]*100/sum(ovl.res.table[2, ])
```

```
## [1] 17.1893
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job salary
ovl.sal.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATSAL))
colnames(ovl.sal.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.sal.table) <- c("satistifed", "dissatisfied")
ovl.sal.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed        28.0629468         45.6996869            12.0172155
##   dissatisfied       0.8995319          2.8831934             3.3870129
##
##                Very dissatisfied
##   satistifed           3.6175052
##   dissatisfied         3.4329074
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.sal.table[1,1]*100/sum(ovl.sal.table[1, ])
```

```
## [1] 31.39125
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.sal.table[2,4]*100/sum(ovl.sal.table[2, ])
```

```
## [1] 32.37784
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job's job security
ovl.sec.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATSEC))
colnames(ovl.sec.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.sec.table) <- c("satistifed", "dissatisfied")
ovl.sec.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed         42.096460          35.430541              8.633262
##   dissatisfied        1.840879           3.714394              2.463004
##
##                Very dissatisfied
##   satistifed            3.237091
##   dissatisfied          2.584369
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.sec.table[1,1]*100/sum(ovl.sec.table[1, ])
```

```
## [1] 47.08916
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.sec.table[2,4]*100/sum(ovl.sec.table[2, ])
```

```
## [1] 24.37476
```

```
# set a proportional table for the overall job satisfaction vs Satisfaction principal job's contribution to socie
ty
ovl.cont.table <- 100*prop.table(table(data.df$Ovl.JOBSATIS,data.df$SATSOC))
colnames(ovl.cont.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissatisfie
d")
rownames(ovl.cont.table) <- c("satistifed", "dissatisfied")
ovl.cont.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   satistifed         48.461515          33.436681              6.237570
##   dissatisfied        2.002019           3.932647              2.656781
##
##                Very dissatisfied
```

```
##   satistifed            1.261588
##   dissatisfied          2.011198
```

```
# compute the percentage of somewhat/very satisfied people are the most satisfied with
ovl.cont.table[1,1]*100/sum(ovl.cont.table[1, ])
```

```
## [1] 54.20912
```

```
# compute the percentage of somewhat/very dissatisfied people are the most dissatisfied with.
ovl.cont.table[2,4]*100/sum(ovl.cont.table[2, ])
```

```
## [1] 18.96883
```

In the overall job satisfaction(JOBSATIS), 43147 people are very satisfied with their jobs, 44508 people are somewhat satisfied, 8178 are somewhat dissatisfied, and 2218 people are very dissatisfied. We observe that the majority of people are somewhat satisfied with their jobs.

Sequentially, we arranged 9 proportional tables about Overall Satisfaction vs Satisfaction principal job's opportunities for advancement, Overall Satisfaction vs Satisfaction principal job benefits, Overall Satisfaction vs Satisfaction principal job's intellectual challenge, Overall Satisfaction vs Satisfaction principal job's degree of independence, Overall Satisfaction vs Satisfaction principal job's job location, Overall Satisfaction vs Satisfaction principal job's level of responsibility, Overall Satisfaction vs Satisfaction principal job salary, Overall Satisfaction vs Satisfaction principal job's job security, and Overall Satisfaction vs Satisfaction principal job's contribution to society. Those who are somewhat/very satisfied with jobs, are the most satisfied with the aspect of principal job's degree of independence since the aspect is the highest rate(around 64.58%) among others. Those who are somewhat/very dissatisfied with jobs, are the least somewhat dissatisfied with the aspect of principal job's opportunities for advancement since the aspect is the highest dissatisfied rate (around 46.35%) among others.

Based on above, the most important job satisfaction is degree of independence. This is because around 64.58% of the highest somewhat/very satisfied proportion is reported in the degree of independence while around 46.35% of the highest somewhat/very dissatisfied proportion is reported in the principal job's opportunities for advancement. It means depending on the aspect for degree of independence, it has the greatest influence on the job satisfaction.

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?

I considered some variables which might correlate with the salary. I added a variable one by one in the model, and compared adjusted R^2 values. When I gathered enough variables, I tested summary of my rough model. Finally, I created an appropriate model with 15 variables:NDGMEMG(field of major for highest degree),DGRDP(type of highest degree),HRSWKGR(working hours per week), WKSWKGR(working weeks per year),BA03Y5(year of first bachelor degree),BADGRUS(location of school for awarding first bachelor degree),HD03Y5(year of highest degree),JOBINS(job health insurance),JOBVAC(available benefits: paid vacation, sick, or personal days),OCEDRLP(principal job related to highest degree),NOCPRMG(job code for principal job),EMSEC(employer sector),WAPRSM(summarized primary work activity),JOBSATIS(job satisfaction),SATADV(satisfaction of job's opportunities for advancement). Since the others of variables do not raise adjusted R^2 value, I substracted them. I also transformed some variables about HD03Y5 as factor since the variable is numerical. I created a new employed data frame: emp.data.df because I should consider excluding unemployed people and no labor force in my final model. In other words, the salary of all unemployed people and no labor force are 0. In the diagnostic tests, I test a method : stepAIC. When I run my model as stepAIC, it filtered unnecessary information about variables. In my model, when I run my model with stepAIC, any factor is not excluded. When I summarize my final model, there is 9999("missing") factor which is unnecessary to analyze in the summary. I also create a new variable about newBA03Y5 to exclude it as factor function instead of BA03Y5.

```
#Before running my final regression model, I need to trim my selected variables.
# change HD03Y5 to factor from numerical as mentioned
data.df$HD03Y5.factor<- as.factor(data.df$HD03Y5)
# set a new variable instead of BA03Y5 to exclude missing values
data.df$newBA03Y5 <- factor(data.df$BA03Y5, exclude = 9999)
# set a new employed data frame
emp.data.df <- subset(data.df, LFSTAT == 1)
```

2. Call your final regression model . 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.

```
# call the MASS to run stepAIC to remove outliers
library(MASS)
# set my final regression model with 15 variables
model.1 <- lm(SALARY ~ NDGMEMG+DGRDG+HRSWKGR+WKSWKGR+newBA03Y5+BADGRUS+HD03Y5.factor+JOBINS+JOBVAC+OCEDRLP+NOCPRM
G+EMSEC+WAPRSM+JOBSATIS+SATADV,data = emp.data.df)
# show the summary of the model to check R^2 values and other factors' coefficient and p-value.

summary(stepAIC(model.1))
```

```
## Start:  AIC=1949480
## SALARY ~ NDGMEMG + DGRDG + HRSWKGR + WKSWKGR + newBA03Y5 + BADGRUS +
##     HD03Y5.factor + JOBINS + JOBVAC + OCEDRLP + NOCPRMG + EMSEC +
##     WAPRSM + JOBSATIS + SATADV
##
##                 Df  Sum of Sq        RSS     AIC
## <none>                        7.7382e+13 1949480
## - BADGRUS        1 1.4492e+10 7.7397e+13 1949496
## - JOBVAC         1 2.7594e+11 7.7658e+13 1949816
## - SATADV         3 2.8072e+11 7.7663e+13 1949818
## - WKSWKGR        3 4.4428e+11 7.7826e+13 1950018
## - newBA03Y5     10 4.6551e+11 7.7848e+13 1950030
## - JOBSATIS       3 5.4749e+11 7.7930e+13 1950144
## - NOCPRMG        6 7.4960e+11 7.8132e+13 1950384
## - NDGMEMG        6 7.6940e+11 7.8151e+13 1950408
## - OCEDRLP        2 1.2449e+12 7.8627e+13 1950993
## - WAPRSM         4 1.3287e+12 7.8711e+13 1951090
## - JOBINS         1 1.4273e+12 7.8809e+13 1951215
## - HD03Y5.factor 10 1.7396e+12 7.9122e+13 1951572
## - EMSEC          3 3.6035e+12 8.0985e+13 1953799
## - HRSWKGR        3 7.4925e+12 8.4875e+13 1958255
## - DGRDG          3 8.7590e+12 8.6141e+13 1959662
```

```
##
## Call:
## lm(formula = SALARY ~ NDGMEMG + DGRDG + HRSWKGR + WKSWKGR + newBA03Y5 +
##     BADGRUS + HD03Y5.factor + JOBINS + JOBVAC + OCEDRLP + NOCPRMG +
##     EMSEC + WAPRSM + JOBSATIS + SATADV, data = emp.data.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -137518  -18636   -1966   17825  152163
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10479.64    2314.44   4.528 5.96e-06 ***
## NDGMEMG2           -7774.39     484.55 -16.045  < 2e-16 ***
## NDGMEMG3           -3180.00     554.76  -5.732 9.94e-09 ***
## NDGMEMG4           -7327.39     445.02 -16.465  < 2e-16 ***
## NDGMEMG5            2693.26     434.14   6.204 5.54e-10 ***
## NDGMEMG6           -5496.25     460.67 -11.931  < 2e-16 ***
## NDGMEMG7           -3940.30     448.23  -8.791  < 2e-16 ***
## DGRDG2             12776.16     274.58  46.529  < 2e-16 ***
## DGRDG3             32234.99     336.72  95.731  < 2e-16 ***
## DGRDG4             33158.06     582.61  56.913  < 2e-16 ***
## HRSWKGR2           15859.64     483.59  32.796  < 2e-16 ***
## HRSWKGR3           27798.09     435.71  63.800  < 2e-16 ***
## HRSWKGR4           37616.80     431.89  87.097  < 2e-16 ***
## WKSWKGR2            5154.30    1559.28   3.306 0.000948 ***
## WKSWKGR3           15559.13    1217.13  12.783  < 2e-16 ***
## WKSWKGR4           20033.11    1163.94  17.212  < 2e-16 ***
## newBA03Y51961      -3403.14    1931.20  -1.762 0.078040 .
## newBA03Y51966       1708.99    1976.40   0.865 0.387205
## newBA03Y51971       2961.26    2000.25   1.480 0.138758
## newBA03Y51976       5070.99    2023.43   2.506 0.012208 *
## newBA03Y51981       5879.63    2040.69   2.881 0.003963 **
## newBA03Y51986       5768.02    2056.39   2.805 0.005034 **
## newBA03Y51991       4699.60    2070.16   2.270 0.023200 *
## newBA03Y51996       1972.59    2087.26   0.945 0.344629
## newBA03Y52001      -2123.71    2107.45  -1.008 0.313594
## newBA03Y52006      -4417.91    2149.16  -2.056 0.039820 *
## BADGRUS1            1167.62     276.89   4.217 2.48e-05 ***
## HD03Y5.factor1966   -212.63    1681.40  -0.126 0.899370
## HD03Y5.factor1971    -91.76    1703.64  -0.054 0.957045
## HD03Y5.factor1976   -979.77    1733.91  -0.565 0.572030
## HD03Y5.factor1981  -1088.60    1755.94  -0.620 0.535290
## HD03Y5.factor1986  -2759.91    1773.58  -1.556 0.119681
## HD03Y5.factor1991  -5609.55    1788.75  -3.136 0.001713 **
## HD03Y5.factor1996  -8077.81    1806.30  -4.472 7.76e-06 ***
## HD03Y5.factor2001 -13195.24    1828.74  -7.215 5.41e-13 ***
## HD03Y5.factor2006 -22488.01    1857.00 -12.110  < 2e-16 ***
## HD03Y5.factor2011 -30169.67    1911.91 -15.780  < 2e-16 ***
## JOBINS1            15430.85     368.74  41.848  < 2e-16 ***
## JOBVAC1             6310.56     342.96  18.400  < 2e-16 ***
## OCEDRLP2           -2972.58     232.21 -12.801  < 2e-16 ***
## OCEDRLP3          -12693.79     324.81 -39.081  < 2e-16 ***
## NOCPRMG2          -13316.92     546.54 -24.366  < 2e-16 ***
## NOCPRMG3          -10504.25     619.93 -16.944  < 2e-16 ***
## NOCPRMG4           -6288.68     555.41 -11.323  < 2e-16 ***
## NOCPRMG5           -2653.75     461.87  -5.746 9.18e-09 ***
## NOCPRMG6           -2047.54     428.98  -4.773 1.82e-06 ***
## NOCPRMG7           -5906.09     408.53 -14.457  < 2e-16 ***
## EMSEC2               141.53     446.60   0.317 0.751316
## EMSEC3             13642.81     504.06  27.066  < 2e-16 ***
## EMSEC4             19090.29     440.19  43.368  < 2e-16 ***
## WAPRSM2           -10398.56     394.61 -26.352  < 2e-16 ***
## WAPRSM3             3457.58     286.27  12.078  < 2e-16 ***
## WAPRSM4             3098.02     452.13   6.852 7.32e-12 ***
## WAPRSM5            -4671.76     299.78 -15.584  < 2e-16 ***
## JOBSATIS2          -5186.28     221.93 -23.369  < 2e-16 ***
```

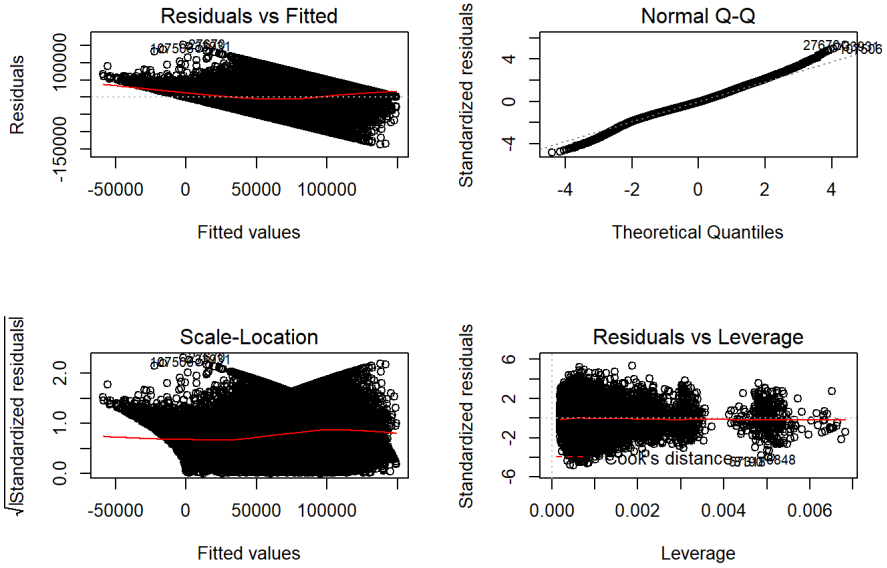
```
## JOBSATIS3          -7440.18     400.58 -18.573  < 2e-16 ***
## JOBSATIS4          -8806.35     692.79 -12.711  < 2e-16 ***
## SATADV2            -2302.45     248.40  -9.269  < 2e-16 ***
## SATADV3            -4166.34     308.32 -13.513  < 2e-16 ***
## SATADV4            -7314.22     415.91 -17.586  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28550 on 94947 degrees of freedom
##   (3044 observations deleted due to missingness)
## Multiple R-squared:  0.5163, Adjusted R-squared:  0.516
## F-statistic:  1717 on 59 and 94947 DF,  p-value: < 2.2e-16
```

In the equation,
$Y = 10479.64 - 7774.39\beta_1 - 3180.00\beta_2 - 7327.39\beta_3 + 2693.26\beta_4 - 5496.25\beta_5 - 3940.30\beta_6 + 12776.16\beta_7 + 32234.99\beta_8 + 33158.06\beta_9 + 15859.64\beta_10 + 27798.$
Intercept(β_0) coefficient is 10479.64 with p-value 5.96e-06 which means the standard of NDGMEMG1(Computer and mathematical sciences), DGRDG1: Bachelor's, HRSWKGR1(working 20 hours or less per week), WKSWKGR(working 1-10 weeks per year), newBA03YY1956(year of first bachelor degree between 1956 and 1960), BADGRUS0(awarding first bachelor degree in location of non-US school), HD03Y5.factor1961(year of highest degree between 1961 and 1965), JOBINS0(no health insurance),JOBVAC0(no paid vacation/sick/personal days), OCEDRLP1(closely related job with highest degree), NOCPRMG1(Job code for principal job for Computer and mathematical scientists), EMSEC1(employer sector for years college or other school system), WAPRSM1(Research and Development in Summarized primary work activity), JOBSATIS1(very satisfied job satisfaction), and SATADV1(very satisfied in satisfaction principal job's opportunities for advancement).
The slope of NDGMEMG2(β_1) coefficient is -7774.39 with p-value: less than 2e-16 closed to 0. The slope of NDGMEMG3(β_2) coefficient is -3180.00 with p-value: 9.94e-09 closed to 0. The slope of NDGMEMG4(β_3) coefficient is -7327.39 with p-value: less than 2e-16 closed to 0. The slope of NDGMEMG5(β_4) coefficient is 2693.26 with p-value: 5.54e-10 closed to 0. The slope of NDGMEMG6(β_5) coefficient is -5496.25 with p-value: less than 2e-16 closed to 0. The slope of NDGMEMG7(β_6) coefficient is -3940.30 with p-value: less than 2e-16 closed to 0. The slope of DGRDG2(β_7) coefficient is 12776.16 with p-value: less than 2e-16 closed to 0. The slope of DGRDG3(β_8) coefficient is 32234.99 with p-value: less than 2e-16 closed to 0. The slope of DGRDG4(β_9) coefficient is 33158.06 with p-value: less than 2e-16 closed to 0. The slope of HRSWKGR2(β_10) coefficient is 15859.64 with p-value: less than 2e-16 closed to 0. The slope of HRSWKGR3(β_11) coefficient is 27798.09 with p-value: less than 2e-16 closed to 0. The slope of HRSWKGR4(β_12) coefficient is 37616.80 with p-value: less than 2e-16 closed to 0. The slope of WKSWKGR2(β_13) coefficient is 5154.30 with p-value: 0.000948. The slope of WKSWKGR3(β_14) coefficient is 15559.13 with p-value: less than 2e-16. The slope of WKSWKGR4(β_15) coefficient is 20033.11 with p-value: less than 2e-16. The slope of newBA03Y51961(β_16) coefficient is -3403.14 with p-value: 0.078040. The slope of newBA03Y51966(β_17) coefficient is 1708.99 with p-value: 0.387205. The slope of newBA03Y51971(β_18) coefficient is 2961.26 with p-value: 0.138758. The slope of newBA03Y51976(β_19) coefficient is 5070.99 with p-value: 0.012208. The slope of newBA03Y51981(β_20) coefficient is 5879.63 with p-value: 0.003963. The slope of newBA03Y51986(β_21) coefficient is 5768.02 with p-value: 0.005034. The slope of newBA03Y51991(β_22) coefficient is 4699.60 with p-value: 0.023200. The slope of newBA03Y51996(β_23) coefficient is 1972.59 with p-value: 0.344629. The slope of newBA03Y2001(β_24) coefficient is -2123.71 with p-value: 0.313594. The slope of newBA03Y52006(β_25) coefficient is -4417.91 with p-value: 0.039820. The slope of BADGRUS1(β_26) coefficient is 1167.62 with p-value: 2.48e-05. The slope of HD03Y5.factor1966(β_27) coefficient is -212.63 with p-value 0.899370. The slope of HD03Y5.factor1971(β_28) coefficient is -91.76 with p-value: 0.957045. The slope of HD03Y5.factor1976(β_29) coefficient is -979.77 with p-value: 0.572030. The slope of HD03Y5.factor1981(β_30) coefficient is -1088.60 with p-value: 0.535290. The slope of HD03Y5.factor1986(β_31) coefficient is -2759.91. with p-value: 0.119681. The slope of HD03Y5.factor1991(β_32) coefficient is -5609.55 with p-value : 0.001713. The slope of HD03Y5.factor1996(β_33) coefficient is -8077.81 with p-value : 7.76e-06. The slope of HD03Y5.factor2001(β_34) coefficient is -13195.24 with p-value : 5.41e-13. The slope of HD03Y5.factor2006(β_35) coefficient is -22488.01 with p-value: less than 2e-16. The slope of HD03Y5.factor2011(β_36) coefficient is -30169.67 with p-value: less than 2e-16. The slope of JOBINS1(β_37) coefficient is 15430.85 with p-value: less than 2e-16. The slope of JOBVAC1(β_38) coefficient is 6310.56 with p-value: less than 2e-16. The slope of OCEDRLP2(β_39) coefficient is -2972.58 with p-value: less than 2e-16. The slope of OCEDRLP3(β_40) coefficient is -12693.79 with p-value: less than 2e-16. The slope of NOCPRMG2(β_41) coefficient is -13316.92 with p-value: less than 2e-16. The slope of NOCPRMG3(β_42) coefficient is -10504.25 with p-value: less than 2e-16. The slope of NOCPRMG4(β_43) coefficient is -6288.68 with p-value: less than 2e-16. The slope of NOCPRMG5(β_44) coefficient is -2653.75 with p-value: 9.18e-09. The slope of NOCPRMG6(β_45) coefficient is -2047.54 with p-value: 1.82e-06. The slope of NOCPRMG7(β_46) coefficient is -5906.09 with p-value: less than 2e-16. The slope of EMSEC2(β_47) coefficient is 141.53 with p-value: 0.751316. The slope of EMSEC3(β_48) coefficient is 13642.81 with p-value: less than 2e-16. The slope of EMSEC4(β_49) coefficient is 19090.29 with p-value: less than 2e-16. The slope of

WAPRSM2($\beta_5 0$) coefficient is -10398.56 with p-value: less than 2e-16. The slope of WAPRSM3($\beta_5 1$) coefficient is 3457.58 with p-value: less than 2e-16. The slope of WAPRSM4($\beta_5 2$) coefficient is 3098.02 with p-value: 7.32e-12. The slope of WAPRSM5($\beta_5 3$) coefficient is -4671.76 with p-value: less than 2e-16. The slope of JOBSATIS2($\beta_5 4$) coefficient is -5186.28 with p-value: less than 2e-16. The slope of JOBSATIS3($\beta_5 5$) coefficient is -7440.18 with p-value: less than 2e-16. The slope of JOBSATIS4($\beta_5 6$) coefficient is -8806.35 with p-value: less than 2e-16. The slope of SATADV2($\beta_5 7$) coefficient is -2302.45 with p-value: less than 2e-16. The slope of SATADV3($\beta_5 8$) coefficient is -4166.34 with p-value: less than 2e-16. The slope of SATADV4($\beta_5 9$) coefficient is -7314.22 with p-value: less than 2e-16.
To interpret p-value in each factor, the p-values of newBA03Y51961, newBA03Y51966, newBA03Y51971, newBA03Y51996, HD03Y5.factor1966, HD03Y5.factor1971, HD03Y5.factor1976, HD03Y5.factor1981, HD03Y5.factor1986, and EMSEC2 are higher than the significant level(0.05). It means they are irrelevant factors to analyze the summary.

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?

```
# show the final regression model's dignostic plots
par(mfrow = c(2,2))
plot(model.1)
```



The Multiple R^2 value is 0.5163, and adjusted R^2 value is 0.516. R^2 means this model could explain 51.63% the variance in this data. Adjusted R^2 controls again the model increase or decrease for the number of predictors in the model compared to the multiple R^2. In this case, since adjusted R^2 decreases, it means a predictor improves the model by less than expected by chance. Through the diagnostic plot, there are somewhat violated in independence, and normality in the right tail, but it is somewhat tolerable to qualify a model. Therefore, my model is a good fit. Besides, it is easy to interpret since the salary is significantly increased or decreased depending on a factor such as DGRDG and EMSEC in the summary.

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).

Through my model, I would obtain professional Engineering since the slopes of NDGMEMG5(the highest of degree for Engineering) and DGRDG4(highest degree for professional) are the highest around 2693.26 and 33158.06 respectively. I should also select the Business or Industry because EMSEC4(Employer Sector for Business or Industry) is the highest slope around 19090.29 to increase salary. To maximize more salary, I should also work full time over 40 hours per week and 40 ~ 52 weeks per year since the slopes of HRSWKGR4 and WKSWKGR4 are 37616.80 and 20033.11 each. Also, I had better do primary work activity about management or administration in the field since the slope of WAPRSM3 is 3457.58.

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.

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?

To consider the job satisfaction, I must exclude SATADV(Satisfaction principal job’s opportunities for advancement), SATBEN(Satisfaction principal job benefits), SATCHAL(Satisfaction principal job’s intellectual challenge), SATIND(Satisfaction principal job’s degree of independence), SATLOC(Satisfaction principal job’s job location), SATRESP(Satisfaction principal job’s level of responsibility), SATSAL(Satisfaction principal job salary), SATSEC(Satisfaction principal job’s job security), SATSOC(Satisfaction principal job’s contribution to society). This is because the factors are really closed relation to the job satisfaction, and make ROC Curve ridiculous between 0.92~1. Next, I should add more variables such as NBAMEMG, NDGMEMG, and so on to find the correlation between job satisfaction. Finally, I find an ideal model with 16 variables(NDGMEMG+DGRDG+Full.Part+MGRNAT+MGROTH+MGRSOC+WAPRSM+NOCPRMG+AGE.factor+EMSEC+GENDER+RACETH+JOBINS+JOBPENS+JOBPROFT+JOBVAC). The other variables are irrelevant to the job satisfaction or make the model sharply drop the AUROC value when we plotROC test. I create other transformed variable about HRSWKGR to distinguish whether it is part-time or full-time instead of specific range time. I recode JOBSATIS into two categories: “satisfied” = “somewhat/very satisfied”, and “not satisfied” = “somewhat/very dissatisfied.” as instruction. Besides, I transformed AGE variable as factor from numerical since I want to check the correlation of age and job satisfaction in detail. In reference, we use for diagnostic tests as stepAIC. In my model, there is no missing values like my regression 1 such as missing or blank values.

2. Call your final regression model . 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.

```
# recode overall job satisfaction as job integration with two categories "somewhat/very satisfied" and "somewhat/
very dissatisfied"
emp.data.df$JOB_Inte[emp.data.df$JOBSATIS == 1 | emp.data.df$JOBSATIS == 2] <- 0
emp.data.df$JOB_Inte[emp.data.df$JOBSATIS == 3 | emp.data.df$JOBSATIS == 4] <- 1
# Create other variable about part or full time job using HRSWKGR
emp.data.df$Full.Part <- emp.data.df$HRSWKGR
# 1 is part-time, and 2 is full-time
emp.data.df$Full.Part[emp.data.df$HRSWKGR == 1 | emp.data.df$HRSWKGR == 2] <- 1
emp.data.df$Full.Part[emp.data.df$HRSWKGR == 3 | emp.data.df$HRSWKGR == 4] <- 2
# set age as factor from numerical
emp.data.df$AGE.factor <- as.factor(emp.data.df$AGE)
# call the information value to run ROC Curve
library(InformationValue)
```

```
## Warning: package 'InformationValue' was built under R version 3.5.3
```

```
# set my final regression model
model.2 <- glm(JOB_Inte ~ NDGMEMG+DGRDG+Full.Part+MGRNAT+MGROTH+MGRSOC+WAPRSM+WASCSM+NOCPRMG+AGE.factor+EMSEC+GEN
DER+RACETH+JOBINS+JOBPENS+JOBPROFT+JOBVAC,data = emp.data.df)
```

```r
# show the summary of the model as stepAIC
summary(stepAIC(model.2))
```

```
## Start:  AIC=45175.68
## JOB_Inte ~ NDGMEMG + DGRDG + Full.Part + MGRNAT + MGROTH + MGRSOC +
##     WAPRSM + WASCSM + NOCPRMG + AGE.factor + EMSEC + GENDER +
##     RACETH + JOBINS + JOBPENS + JOBPROFT + JOBVAC
##
##              Df Deviance   AIC
## - JOBINS      1   9084.2 45174
## - Full.Part   1   9084.2 45174
## - JOBVAC      1   9084.3 45175
## <none>            9084.2 45176
## - GENDER      1   9084.4 45177
## - DGRDG       3   9086.5 45194
## - MGRSOC      1   9086.6 45200
## - WAPRSM      4   9087.7 45206
## - EMSEC       3   9088.1 45212
## - JOBPENS     1   9088.5 45221
## - NDGMEMG     6   9090.0 45227
## - MGROTH      1   9092.9 45267
## - NOCPRMG     6   9094.1 45271
## - RACETH      2   9094.9 45287
## - WASCSM      5   9098.6 45322
## - MGRNAT      1   9100.0 45345
## - JOBPROFT    1   9104.0 45388
## - AGE.factor 49   9131.2 45584
##
## Step:  AIC=45174.03
## JOB_Inte ~ NDGMEMG + DGRDG + Full.Part + MGRNAT + MGROTH + MGRSOC +
##     WAPRSM + WASCSM + NOCPRMG + AGE.factor + EMSEC + GENDER +
##     RACETH + JOBPENS + JOBPROFT + JOBVAC
##
##              Df Deviance   AIC
## - Full.Part   1   9084.2 45173
## <none>            9084.2 45174
## - JOBVAC      1   9084.4 45174
## - GENDER      1   9084.5 45175
## - DGRDG       3   9086.5 45193
## - MGRSOC      1   9086.6 45198
## - WAPRSM      4   9087.8 45205
## - EMSEC       3   9088.3 45212
## - NDGMEMG     6   9090.0 45225
## - JOBPENS     1   9089.4 45228
## - MGROTH      1   9092.9 45266
## - NOCPRMG     6   9094.2 45270
## - RACETH      2   9094.9 45286
## - WASCSM      5   9098.7 45321
## - MGRNAT      1   9100.1 45344
## - JOBPROFT    1   9104.2 45388
## - AGE.factor 49   9131.3 45583
##
## Step:  AIC=45172.61
## JOB_Inte ~ NDGMEMG + DGRDG + MGRNAT + MGROTH + MGRSOC + WAPRSM +
##     WASCSM + NOCPRMG + AGE.factor + EMSEC + GENDER + RACETH +
##     JOBPENS + JOBPROFT + JOBVAC
```

```
##
##              Df Deviance   AIC
## <none>            9084.2 45173
## - GENDER      1   9084.5 45173
## - JOBVAC      1   9084.5 45174
## - DGRDG       3   9086.5 45191
## - MGRSOC      1   9086.7 45197
## - WAPRSM      4   9087.8 45203
## - EMSEC       3   9088.3 45211
## - NDGMEMG     6   9090.1 45223
## - JOBPENS     1   9089.6 45229
## - MGROTH      1   9093.0 45265
## - NOCPRMG     6   9094.2 45268
## - RACETH      2   9095.0 45284
## - WASCSM      5   9099.0 45322
## - MGRNAT      1   9100.3 45344
## - JOBPROFT    1   9104.3 45386
## - AGE.factor 49   9131.8 45586
```

```
##
## Call:
## glm(formula = JOB_Inte ~ NDGMEMG + DGRDG + MGRNAT + MGROTH +
##     MGRSOC + WAPRSM + WASCSM + NOCPRMG + AGE.factor + EMSEC +
##     GENDER + RACETH + JOBPENS + JOBPROFT + JOBVAC, data = emp.data.df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.28234  -0.12693  -0.09573  -0.06141   1.05309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1619459  0.0169296   9.566  < 2e-16 ***
## NDGMEMG2      0.0032484  0.0050657   0.641 0.521354
## NDGMEMG3      0.0105323  0.0057813   1.822 0.068492 .
## NDGMEMG4      0.0124570  0.0049063   2.539 0.011118 *
## NDGMEMG5     -0.0110364  0.0045445  -2.429 0.015161 *
## NDGMEMG6     -0.0149017  0.0049681  -2.999 0.002705 **
## NDGMEMG7      0.0010115  0.0048042   0.211 0.833241
## DGRDG2       -0.0041034  0.0025970  -1.580 0.114090
## DGRDG3        0.0081354  0.0030264   2.688 0.007186 **
## DGRDG4       -0.0178883  0.0057665  -3.102 0.001922 **
## MGRNAT1      -0.0369413  0.0028094 -13.149  < 2e-16 ***
## MGROTH1      -0.0230408  0.0023755  -9.699  < 2e-16 ***
## MGRSOC1      -0.0145990  0.0028544  -5.115 3.15e-07 ***
## WAPRSM2       0.0054230  0.0042202   1.285 0.198798
## WAPRSM3       0.0174989  0.0030013   5.830 5.55e-09 ***
## WAPRSM4      -0.0005260  0.0048354  -0.109 0.913374
## WAPRSM5       0.0114134  0.0032141   3.551 0.000384 ***
## WASCSM2      -0.0022403  0.0044700  -0.501 0.616248
## WASCSM3      -0.0030513  0.0026085  -1.170 0.242108
## WASCSM4      -0.0061431  0.0051410  -1.195 0.232114
## WASCSM5       0.0075433  0.0036937   2.042 0.041132 *
## WASCSM6       0.0328004  0.0032219  10.181  < 2e-16 ***
## NOCPRMG2      0.0124224  0.0057388   2.165 0.030418 *
## NOCPRMG3      0.0003443  0.0065114   0.053 0.957832
## NOCPRMG4     -0.0315891  0.0060628  -5.210 1.89e-07 ***
```

```
## NOCPRMG5      0.0041627  0.0048737   0.854 0.393047
## NOCPRMG6     -0.0046633  0.0045599  -1.023 0.306467
## NOCPRMG7      0.0116292  0.0044499   2.613 0.008966 **
## AGE.factor27  0.0263961  0.0154336   1.710 0.087212 .
## AGE.factor28  0.0283228  0.0159912   1.771 0.076539 .
## AGE.factor29  0.0080549  0.0159792   0.504 0.614200
## AGE.factor30  0.0044923  0.0160436   0.280 0.779475
## AGE.factor31  0.0101717  0.0162293   0.627 0.530824
## AGE.factor32  0.0074739  0.0161848   0.462 0.644234
## AGE.factor33  0.0026753  0.0161906   0.165 0.868759
## AGE.factor34  0.0085513  0.0163159   0.524 0.600204
## AGE.factor35  0.0048033  0.0164496   0.292 0.770287
## AGE.factor36  0.0072936  0.0164968   0.442 0.658403
## AGE.factor37  0.0025305  0.0165309   0.153 0.878336
## AGE.factor38  0.0045876  0.0164576   0.279 0.780433
## AGE.factor39 -0.0059047  0.0165525  -0.357 0.721299
## AGE.factor40 -0.0027537  0.0165467  -0.166 0.867829
## AGE.factor41 -0.0054610  0.0165281  -0.330 0.741095
## AGE.factor42 -0.0067000  0.0164691  -0.407 0.684141
## AGE.factor43 -0.0043148  0.0165476  -0.261 0.794286
## AGE.factor44 -0.0049155  0.0165814  -0.296 0.766890
## AGE.factor45 -0.0080611  0.0166307  -0.485 0.627882
## AGE.factor46  0.0011760  0.0166723   0.071 0.943767
## AGE.factor47 -0.0017080  0.0167107  -0.102 0.918590
## AGE.factor48 -0.0057435  0.0165970  -0.346 0.729301
## AGE.factor49  0.0041028  0.0165127   0.248 0.803776
## AGE.factor50 -0.0092184  0.0165889  -0.556 0.578418
## AGE.factor51  0.0093860  0.0166232   0.565 0.572326
## AGE.factor52  0.0088799  0.0166702   0.533 0.594254
## AGE.factor53  0.0002430  0.0166388   0.015 0.988347
## AGE.factor54 -0.0103160  0.0166211  -0.621 0.534827
## AGE.factor55 -0.0126123  0.0166794  -0.756 0.449553
## AGE.factor56 -0.0117188  0.0166427  -0.704 0.481346
## AGE.factor57 -0.0042322  0.0167949  -0.252 0.801049
## AGE.factor58 -0.0042898  0.0167524  -0.256 0.797896
## AGE.factor59 -0.0166234  0.0168149  -0.989 0.322856
## AGE.factor60 -0.0119892  0.0168930  -0.710 0.477884
## AGE.factor61 -0.0306371  0.0169534  -1.807 0.070744 .
## AGE.factor62 -0.0233987  0.0172310  -1.358 0.174484
## AGE.factor63 -0.0523239  0.0173106  -3.023 0.002506 **
## AGE.factor64 -0.0369649  0.0175089  -2.111 0.034757 *
## AGE.factor65 -0.0530947  0.0176400  -3.010 0.002614 **
## AGE.factor66 -0.0595649  0.0181243  -3.286 0.001015 **
## AGE.factor67 -0.0668585  0.0190995  -3.501 0.000465 ***
## AGE.factor68 -0.0867837  0.0194447  -4.463 8.09e-06 ***
## AGE.factor69 -0.0742438  0.0198119  -3.747 0.000179 ***
## AGE.factor70 -0.0916283  0.0199286  -4.598 4.27e-06 ***
## AGE.factor71 -0.0919303  0.0214386  -4.288 1.80e-05 ***
## AGE.factor72 -0.0794529  0.0222620  -3.569 0.000359 ***
## AGE.factor73 -0.0900155  0.0243023  -3.704 0.000212 ***
## AGE.factor74 -0.0935300  0.0248294  -3.767 0.000165 ***
## AGE.factor75 -0.1115647  0.0266714  -4.183 2.88e-05 ***
## EMSEC2       -0.0109122  0.0047030  -2.320 0.020328 *
## EMSEC3       -0.0139239  0.0053452  -2.605 0.009190 **
## EMSEC4        0.0042655  0.0047377   0.900 0.367941
## GENDER2       0.0036238  0.0021620   1.676 0.093722 .
```

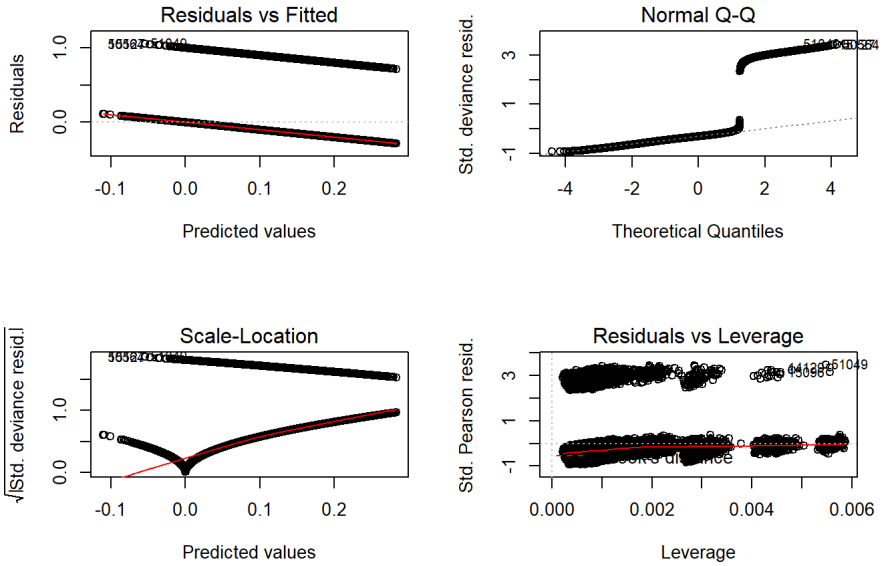
```
## RACETH2      -0.0144723  0.0027468  -5.269 1.38e-07 ***
## RACETH3       0.0117192  0.0032612   3.594 0.000326 ***
## JOBPENS1     -0.0206773  0.0027112  -7.627 2.43e-14 ***
## JOBPROFT1    -0.0362523  0.0024679 -14.690  < 2e-16 ***
## JOBVAC1      -0.0054258  0.0030949  -1.753 0.079581 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09272947)
##
##     Null deviance: 9293.7  on 98050  degrees of freedom
## Residual deviance: 9084.2  on 97965  degrees of freedom
## AIC: 45173
##
## Number of Fisher Scoring iterations: 2
```

In the equation,
$Y = 0.1619459 + 0.0032484\beta_1 + 0.0105323\beta_2 + 0.0124570\beta_3 - 0.0110364\beta_4 - 0.0149017\beta_5 + 0.0010115\beta_6 - 0.0041034\beta_7 + 0.0081354\beta_8 - 0.0178883\beta_9 - 0.0$
Intercept(β_0) coefficient is 0.1619459 with p-value: less than 2e-16 which means the standard of NDGMEMG1(highest degree in Computer and mathematical sciences), DGRDG1: Bachelor's, MGRNAT0(no technical expertise required in natural sciences), MGROTH0(no technical expertise required in other), MGRSOC0(no technical expertise required in social sciences) WAPRSM1(Research and Development), WASCSM1(awarding first bachelor degree in location of non-US school), NOCPRMG1(Computer and mathematical scientists job), AGE26.factor(26 years old),EMSEC1(2 year college or other school system in employer sector), GENDER1(female), RACETH1(Asian), EMSEC1(employer sector for years college or other school system), JOBPENS0(No pension/retirement plan), JOBPROFT0(no profit-sharing plan), and JOBVAC0(no paid vacation/sick/personal days).
The slope of NDGMEMG2(β_1) coefficient is 0.0032484 with p-value: 0.521354. The slope of NDGMEMG3(β_2) coefficient is 0.0105323 with p-value: 0.068492. The slope of NDGMEMG4(β_3) coefficient is 0.0124570 with p-value: 0.011118. The slope of NDGMEMG5(β_4) coefficient is -0.0110364 with p-value: 0.015161. The slope of NDGMEMG6(β_5) coefficient is -0.0149017 with p-value: 0.002705. The slope of NDGMEMG7(β_6) coefficient is 0.0010115 with p-value: 0.833241. The slope of DGRDG2(β_7) coefficient is -0.0041034 with p-value: 0.114090. The slope of DGRDG3(β_8) coefficient is 0.0081354 with p-value: 0.007186. The slope of DGRDG4(β_9) coefficient is -0.0178883 with p-value: 0.001922. The slope of MGRNAT1(β_10) coefficient is -0.0369413 with p-value: less than 2e-16 closed to 0. The slope of MGROTH1(β_11) coefficient is -0.0230408 with p-value: less than 2e-16 closed to 0. The slope of MGRSOC1(β_12) coefficient is -0.0145990 with p-value: 3.15e-07. The slope of WAPRSM2(β_13) coefficient is 0.0054230 with p-value: 0.198798. The slope of WAPRSM3(β_14) coefficient is 0.0174989 with p-value: 5.55e-09. The slope of WAPRSM4(β_15) coefficient is -0.0005260 with p-value: 0.913374. The slope of WAPRSM5(β_16) coefficient is 0.0114134 with p-value: 0.000384. The slope of WASCSM2(β_17) coefficient is -0.0022403 with p-value: 0.616248. The slope of WASCSM3(β_18) coefficient is -0.0030513 with p-value: 0.242108. The slope of WASCSM4(β_19) coefficient is -0.0061431 with p-value: 0.232114. The slope of WASCSM5(β_20) coefficient is 0.0075433 with p-value: 0.041132. The slope of WASCSM6(β_21) coefficient is 0.0328004 with p-value: less than 2e-16. The slope of NOCPRMG2(β_22) coefficient is 0.0124224 with p-value: 0.030418. The slope of NOCPRMG3(β_23) coefficient is 0.0003443 with p-value: less than 2e-16. The slope of NOCPRMG4(β_24) coefficient is -0.0315891 with p-value: 1.89e-07. The slope of NOCPRMG5(β_25) coefficient is 0.0041627 with p-value: 0.393047. The slope of NOCPRMG6(β_26) coefficient is -0.0046633 with p-value: 0.306467. The slope of NOCPRMG7(β_27) coefficient is 0.0116292 with p-value 0.008966. The slope of AGE.factor27(β_28) coefficient is 0.0263961 with p-value: 0.087212. The slope of AGE.factor28(β_29) coefficient is 0.0283228 with p-value: 0.076539. The slope of AGE.factor29(β_30) coefficient is 0.0080549 with p-value: 0.614200. The slope of AGE.factor30(β_31) coefficient is 0.0044923. with p-value: 0.779475. The slope of AGE.factor31(β_32) coefficient is 0.0101717 with p-value : 0.530824. The slope of AGE.factor32(β_33) coefficient is 0.0074739 with p-value : 0.644234. The slope of AGE.factor33(β_34) coefficient is 0.0026753 with p-value : 0.868759. The slope of AGE.factor34(β_35) coefficient is 0.0085513 with p-value: 0.600204. The slope of AGE.factor35(β_36) coefficient is 0.0048033 with p-value: 0.770287. The slope of AGE.factor36(β_37) coefficient is 0.0072936 with p-value: 0.658403. The slope of AGE.factor37(β_38) coefficient is 0.0025305 with p-value: 0.878336. The slope of AGE.factor38(β_39) coefficient is 0.0045876 with p-value: 0.780433. The slope of AGE.factor39(β_40) coefficient is -0.0059047 with p-value: 0.721299. The slope of AGE.factor40(β_41) coefficient is -0.0027537 with p-value: 0.867829. The slope of AGE.factor41(β_42) coefficient is -0.0054610 with p-value: 0.741095. The slope of AGE.factor42(β_43) coefficient is -0.0067000 with p-value: 0.684141. The slope of AGE.factor43(β_44) coefficient is -0.0043148 with p-value: 0.794286. The slope of AGE.factor44(β_45) coefficient is -0.0049155 with p-value: 0.766890. The slope of AGE.factor45(β_46) coefficient is -0.0080611 with p-value: 0.627882. The slope of AGE.factor46(β_47) coefficient is 0.0011760 with p-value: 0.943767. The slope of AGE.factor47(

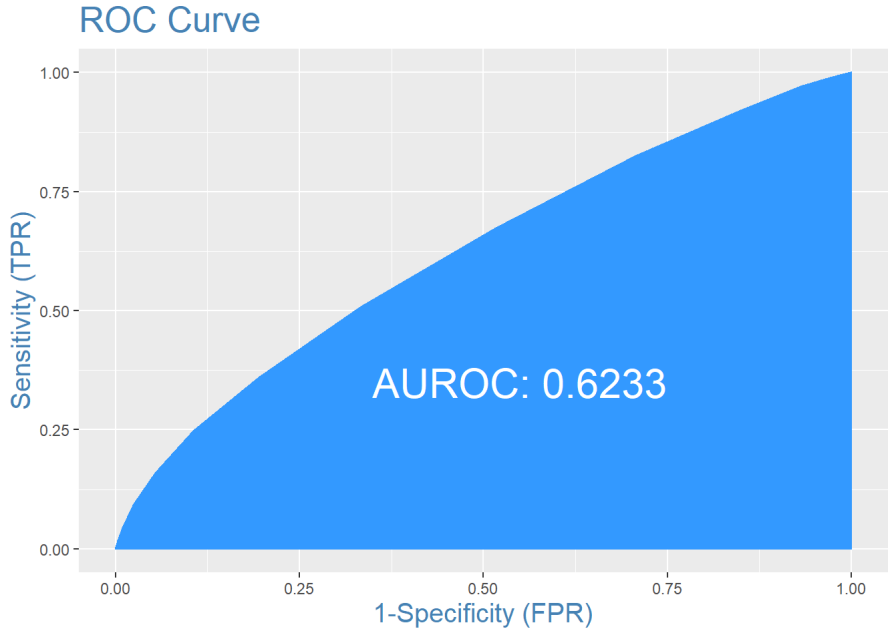
$\beta_4 8$) coefficient is -0.0017080 with p-value: 0.918590. The slope of AGE.factor48($\beta_4 9$) coefficient is -0.0057435 with p-value: 0.729301. The slope of AGE.factor49($\beta_5 0$) coefficient is 0.0041028 with p-value: 0.803776. The slope of AGE.factor50($\beta_5 1$) coefficient is -0.0092184 with p-value: 0.578418. The slope of AGE.factor51($\beta_5 2$) coefficient is 0.0093860 with p-value: 0.572326. The slope of AGE.factor52($\beta_5 3$) coefficient is 0.0088799 with p-value: 0.594254. The slope of AGE.factor53($\beta_5 4$) coefficient is 0.0002430 with p-value: 0.988347. The slope of AGE.factor54($\beta_5 5$) coefficient is -0.0103160 with p-value: 0.534827. The slope of AGE.factor55($\beta_5 6$) coefficient is -0.0126123 with p-value: less than 0.449553. The slope of AGE.factor56($\beta_5 7$) coefficient is -0.0117188 with p-value: 0.481346. The slope of AGE.factor57($\beta_5 8$) coefficient is -0.0042322 with p-value: 0.801049. The slope of AGE.factor58($\beta_5 9$) coefficient is -0.0042898 with p-value: 0.797896. The slope of AGE.factor59($\beta_6 0$) coefficient is -0.0166234 with p-value: 0.322856. The slope of AGE.factor60($\beta_6 1$) coefficient is -0.0119892 with p-value: 0.477884. The slope of AGE.factor61($\beta_6 2$) coefficient is -0.0306371 with p-value: 0.070744. The slope of AGE.factor62($\beta_6 3$) coefficient is -0.0233987 with p-value: 0.174484. The slope of AGE.factor63($\beta_6 4$) coefficient is -0.0523239 with p-value: 0.002506. The slope of AGE.factor64($\beta_6 5$) coefficient is -0.0369649 with p-value: 0.034757. The slope of AGE.factor65($\beta_6 6$) coefficient is -0.0530947 with p-value: 0.002614. The slope of AGE.factor66($\beta_6 7$) coefficient is -0.0595649 with p-value: 0.001015. The slope of AGE.factor67($\beta_6 8$) coefficient is -0.0668585 with p-value: 0.000465.The slope of AGE.factor68($\beta_6 9$) coefficient is -0.0867837 with p-value: 8.09e-06. The slope of AGE.factor69($\beta_7 0$) coefficient is -0.0916283 with p-value: 4.27e-06. The slope of AGE.factor70($\beta_7 1$) coefficient is -0.0916283 with p-value: 4.27e-06. The slope of AGE.factor71($\beta_7 2$) coefficient is -0.0919303 with p-value: 1.80e-05. The slope of AGE.factor72($\beta_7 3$) coefficient is -0.0794529 with p-value: 0.000359. The slope of AGE.factor73($\beta_7 4$) coefficient is -0.0900155 with p-value: 0.000212. The slope of AGE.factor74($\beta_7 5$) coefficient is -0.0935300 with p-value: 0.000165. The slope of AGE.factor75($\beta_7 6$) coefficient is -0.1115647 with p-value: 2.88e-05. The slope of EMSEC2($\beta_7 7$) coefficient is -0.0109122 with p-value: 0.020328. The slope of EMSEC3($\beta_7 8$) coefficient is -0.0139239 with p-value: 0.009190. The slope of EMSEC4($\beta_7 9$) coefficient is 0.0042655 with p-value: 0.367941. The slope of GENDER2($\beta_8 0$) coefficient is 0.0036238 with p-value: 0.093722. The slope of RACETH2($\beta_8 1$) coefficient is -0.0144723 with p-value: 1.38e-07. The slope of RACETH3($\beta_8 2$) coefficient is 0.0117192 with p-value: 0.000326. The slope of JOBPENS1($\beta_8 3$) coefficient is -0.0206773 with p-value: 2.43e-14. The slope of JOBPROFT1($\beta_8 4$) coefficient is -0.0362523 with p-value: less than 2e-16. The slope of JOBVAC1($\beta_8 5$) coefficient is -0.0054258 with p-value: 0.079581.
To analyze p-values, p-values of NDGMEMG2, NDGMEMG3, DGRDG2, WAPRSM2, WAPRSM4, WASCSM2, WASCSM3, WASCSM4, NOCPRMG3, NOCPRMG5, NOCPRMG6, AGE.factor27 ~ AGE.factor62 are higher than significant level(0.05). It means they are not useful to analyze the summary.

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

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




```
plotROC(emp.data.df$JOB_Inte, model.2$fitted.values)
```



The ROC Curve value(pseudo-R^2) is 0.6233. In the diagnostic plots, independence, normality, and constant variance are totally violated. Nevertheless, since ROC Curve value is higher than 0.6, it is good fit model. Besides, it is easy to interpret that job satisfaction has positively been infuenced by the highest degree of field major(NDGMEMG), primary work activity(WAPRSM), secondary work activity(WASCSM), and Job code for principal job(NOCPRMG) in broad outlines.

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).

In this model, I should choose Social and related sciences(NDGMEMG4) with Doctorate(DGRDG3). This is because the slopes of NDGMEMG4 and DGRDG3 is the highest positive rate. To maximize my job satisfaction, I had better do management and administration(WAPRSM3) as primary work activity and no secondary activity(WASCSM6). Lastly, I would enter to the business or industry(EMSEC4) after graduation as much as I could. In the business or industry, I could work my major job Social and related sciences with field of Biological, agricultural and other life scientists(NOCPRMG2). ## 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?

This article claims that: a. 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.

a.

```
# show the proportional table to check a correlation between job satisfaction and Educational level
edu.jobsatis.table <- 100*prop.table(table(emp.data.df$DGRDG, emp.data.df$JOBSATIS))
colnames(edu.jobsatis.table) <- c("Very satistifed", "Somewhat satisfied", "Somewhat dissatisfied", "Very dissati
sfied")
rownames(edu.jobsatis.table) <- c("Bachelor's", "Master's", "Doctorate", "Professional")
edu.jobsatis.table
```

```
##
##                Very satistifed Somewhat satisfied Somewhat dissatisfied
##   Bachelor's       15.19923305        17.69283332            3.37477435
##   Master's         12.77294469        13.48481912            2.43036787
##   Doctorate        13.98455906        12.84331623            2.29064466
##   Professional      2.04791384         1.37173512            0.24477058
##
##                Very dissatisfied
##   Bachelor's          1.04435447
##   Master's            0.60478730
##   Doctorate           0.53135613
##   Professional        0.08159019
```

```
# show the bachelor's percentage about job satisfaction sequentially "Very satistifed", "Somewhat satisfied", "So
mewhat dissatisfied", "Very dissatisfied"
edu.jobsatis.table[1,1]*100/sum(edu.jobsatis.table[1, ])
```

```
## [1] 40.73639
```

```
edu.jobsatis.table[1,2]*100/sum(edu.jobsatis.table[1, ])
```

```
## [1] 47.41964
```

```
edu.jobsatis.table[1,3]*100/sum(edu.jobsatis.table[1, ])
```

```
## [1] 9.044938
```

```
edu.jobsatis.table[1,4]*100/sum(edu.jobsatis.table[1, ])
```

```
## [1] 2.799038
```

```
# show the master's percentage about job satisfaction sequentially "Very satistifed", "Somewhat satisfied", "Some what dissatisfied", "Very dissatisfied"
edu.jobsatis.table[2,1]*100/sum(edu.jobsatis.table[2, ])
```

```
## [1] 43.60421
```

```
edu.jobsatis.table[2,2]*100/sum(edu.jobsatis.table[2, ])
```

```
## [1] 46.0344
```

```
edu.jobsatis.table[2,3]*100/sum(edu.jobsatis.table[2, ])
```

```
## [1] 8.296776
```

```
edu.jobsatis.table[2,4]*100/sum(edu.jobsatis.table[2, ])
```

```
## [1] 2.064619
```

```
# show the doctorate's percentage about job satisfaction sequentially "Very satistifed", "Somewhat satisfied", "S omewhat dissatisfied", "Very dissatisfied"
edu.jobsatis.table[3,1]*100/sum(edu.jobsatis.table[3, ])
```

```
## [1] 47.16566
```

```
edu.jobsatis.table[3,2]*100/sum(edu.jobsatis.table[3, ])
```

```
## [1] 43.31659
```

```
edu.jobsatis.table[3,3]*100/sum(edu.jobsatis.table[3, ])
```

```
## [1] 7.725647
```

```
edu.jobsatis.table[3,4]*100/sum(edu.jobsatis.table[3, ])
```

```
## [1] 1.792102
```

```
# show the professional's percentage about job satisfaction sequentially "Very satistifed", "Somewhat satisfied",
 "Somewhat dissatisfied", "Very dissatisfied"
edu.jobsatis.table[4,1]*100/sum(edu.jobsatis.table[4, ])
```

```
## [1] 54.66921
```

```
edu.jobsatis.table[4,2]*100/sum(edu.jobsatis.table[4, ])
```

```
## [1] 36.61857
```

```
edu.jobsatis.table[4,3]*100/sum(edu.jobsatis.table[4, ])
```

```
## [1] 6.534168
```

```
edu.jobsatis.table[4,4]*100/sum(edu.jobsatis.table[4, ])
```

```
## [1] 2.178056
```

First of all, we need to test the correlation between education level vs job satisfaction as proportable table method. In the varialbles, I selected DGRDG(Type of highest certificate or degree) which is educational level and JOBSATIS(overall job satisfaction). When I compute job satisfaction distribution each degree: Bachelor, Master, Doctorate, and Professional, there is not prominently different percentage between degrees. There is slightly difference within 2~14% to compare each degree. Therefore, the claim is true that Education level has very little to do with job satisfaction.

b.

In the question, the having opportunity to do what you best means the job's degree of independence. It is related to question number 7 in the Basic Analysis section. Using the question number 7 proportion table, the most important job satisfaction is degree of independence. since around 64.58% of the highest somewhat/very satisfied proportion is reported in the degree of independence while around 46.35% of the highest somewhat/very dissatisfied proportion is reported in the principal job's opportunities for advancement. It means depending on the aspect for degree of independence, it has the greatest influence on the job satisfaction. Therefore, the claim is true.

2. Diverse Education: College-educated Americans More Likely Experience Job Satisfaction, Lead Healthier Lives, Study Says

This article claims that: a. Certain race groups earn less than others when they have the same education level. b. STEM (science, technology, engineering and mathematics) careers, in which minorities are underrepresented, tend to pay more than careers in social sciences.

a.

```
# set the pair of groups : bachelor, master, doctorate, and professional
bach = emp.data.df[emp.data.df$DGRDG == 1,]
mast = emp.data.df[emp.data.df$DGRDG == 2,]
doct = emp.data.df[emp.data.df$DGRDG == 3,]
prof = emp.data.df[emp.data.df$DGRDG == 4,]
# Next, we would do anova test
anova(lm(SALARY ~ RACETH, data = bach))
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
RACETH	2	9.008668e+11	450433414063	326.8255	2.064034e-141
Residuals	36581	5.041622e+13	1378207818	*NA*	*NA*

2 rows

```
anova(lm(SALARY ~ RACETH, data = mast))
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
RACETH	2	6.617319e+11	330865940472	220.7426	7.279299e-96
Residuals	28719	4.304623e+13	1498876454	*NA*	*NA*

2 rows

```
anova(lm(SALARY ~ RACETH, data = doct))
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
RACETH	2	4.891961e+11	244598050663	149.5991	2.301513e-65
Residuals	29069	4.752850e+13	1635023427	*NA*	*NA*

2 rows

```
anova(lm(SALARY ~ RACETH, data = prof))
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
RACETH	2	5.237909e+10	26189545102	12.23408	5.06338e-06
Residuals	3670	7.856385e+12	2140704321	*NA*	*NA*

2 rows

To verify the claim, we need to use variables to match the same education level about RACETH and DGRDG. After that we need to use salary to compare it. We would use 4 hypothesis test as ANOVA method. As First hypothesis test, H_0 : Salary and race with same bachelor degree are independent, H_A : Salary and race with same bachelor degree are dependent. As second hypothesis test, H_0 : Salary and race with same master degree are independent, H_A : Salary and race with same master degree are dependent. As third hypothesis test, H_0 : Salary and race with same doctorate are independent, H_A : Salary and race with same doctorate degree are dependent. As fourth hypothesis test, H_0 : Salary and race with same professional degree are independent, H_A : Salary and race with same professional degree are dependent. In ANOVA test, all p-values in each tests are less than significant level(0.05). Therefore, we could contradict H_0 in all four hypothesis test. We conclude that H_A. Based on H_A, since salary and race with same degree are dependent, it is true that certain race groups earn less than others when they have the same education level.

b.

```
# check the chi-square test between MINRTY and NDGMEMG(the highest degree major) as first step.
chisq.test(emp.data.df$MINRTY, emp.data.df$NDGMEMG)
```

```
##
##  Pearson's Chi-squared test
##
## data:  emp.data.df$MINRTY and emp.data.df$NDGMEMG
## X-squared = 963.51, df = 6, p-value < 2.2e-16
```

```
# Since we know that MINRTY and NDGMEMG are dependent, we check ANOVA test between STEM and social science of NDG
MEMG(the highest degree major) as second step
anova(lm(SALARY ~ NDGMEMG, data=emp.data.df))
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
NDGMEMG	6	7.593481e+12	1.265580e+12	785.1769	0
Residuals	98044	1.580313e+14	1.611841e+09	*NA*	*NA*

2 rows

To verify the claim, we need to two steps of hypothesis tests by chi-square and ANOVA. In the chi-square test, we would test that H_0 : minority and highest degree of major(career) are independent, and H_A : minority and highest degree of major(career) are dependent. Since p-value is 2.2e-16 which is less than significant level, we could contradict H_0. Therefore, minority and highest degree of major(career) are dependent. Next, we would do ANOVA test to check the correlation between SALARY and NDGMEMG. We would test that H_0 : salary and highest degree of major(career) are independent, and H_A : salary and highest degree of major(career) are dependent. Since p-value is 2.2e-16 which is less than significant level, we could contradict H_0. Therefore, minority and highest degree of major(career) are dependent. As a result, we could conclude that the claim is true that STEM (science, technology, engineering and mathematics) careers, in which minorities are underrepresented, tend to pay more than careers in social sciences.

3. PEW: 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 ?€쌚ery closely?€? related to their college or graduate field of study.

```
# set another variable to test it exactly as stem NDGMEMG.
emp.data.df$NDGMEMG.STEM <- (emp.data.df$NDGMEMG == 1 | emp.data.df$NDGMEMG == 2 | emp.data.df$NDGMEMG == 3 | em
p.data.df$NDGMEMG == 4 | emp.data.df$NDGMEMG == 5 | emp.data.df$NDGMEMG == 6)

# set chi-square test NDGMEMG.STEM and OCEDRLP
chisq.test(emp.data.df$NDGMEMG.STEM, emp.data.df$OCEDRLP)
```

```
##
##  Pearson's Chi-squared test
##
## data:  emp.data.df$NDGMEMG.STEM and emp.data.df$OCEDRLP
## X-squared = 64.033, df = 2, p-value = 1.246e-14
```

To check the claim, we need to use variables NDGMEMG and OCEDRLP. However, we need to create a new variable as NDGMEMG stem because we want to check only STEM major in NDGMEMG. Next, we would do a hypothesis test as chi-square. H_0 : the highest degree in STEM majors and principal job related to highest degree are independent. H_A : the highest degree in STEM majors and principal job related to highest degree are dependent. Since p-value is 1.246e-14, we could contract H_0. Therefore, the highest degree in STEM majors and principal job related

to highest degree are dependent. In conclusion, 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

Basic analysis In question 1, there are two different surveys (NSCG and SDR). To compare NSCG and SDR, the population of NSCG is individuals residing in the US, while SDR is doctorate degree in STEM in the US. Its sample of NSCG is ACES households, and the sample of SDR is a stratified sample from the eligible individuals in the Doctorate Records File. Those surveys used the similar sampling method as stratified. When we think of both survey biases with combination, the main points of biases are we do not know the ratio of populations. In question 2, the population of male is around 13.3% more than that of female regardless of race. Considering overall race/ethnicity, the majority of race/ethnicity is White regardless of gender whereas the minority of race/ethnicity is Asian. In question 3, in the field of major for first bachelor degree and Year of first bachelor degree, between 1981 and 1990, Engineering degrees were popular more than others. Between 1991 and now, Social and related science degrees have been popular more than others. The more people obtained the highest degree in all majors every range of years, but the size of sample is sharply decreased in 2011 or later. In the field of major for highest degree and Year of highest degree, the most people of the sample obtained Engineering degrees between 1961 and 1965. Between 1966 and 2010, most of people obtained Social and related sciences or Engineering. In question 4, through the Chi-Square test, we concluded that there is a significant difference in retention rates among different field of majors. In question 5, the majority of people work with regular hours year-round. Also, we observed that illness, retirement, or other(combined) are the major reasons that led people to not work at the time of survey. In question 6, we tested 3 hypothesis tests as Chi-square for Degree relevance vs job type, degree that they trained for, and principal activity in people’s job. We conclude that Degree relevance is dependent on all three of them through the tests. In question 7, the most important job satisfaction is independence since its ratio(64.58%) is the most highest. People who were very satisfied with their jobs resulted from the satisfaction of independence. However, people who were very dissatisfied with their jobs resulted from the satisfaction of principal job’s opportunities for advancement.

Regression 1: SALARY vs other variables In my model, I created an appropriate model with 15 variables: NDGMEMG, DGRDP, HRSWKGR, WKSWKGR, BA03Y5, BADGRUS, HD03Y5, JOBINS, JOBVAC, OCEDRLP, NOCPRMG, EMSEC, WAPRSM, JOBSATIS, SATADV. This is because I think they are the most useful variables to predict salary. Besides, I used stepAIC method in my model to exclude an outlier. In the equation,
$Y = 10479.64 - 7774.39\beta_1 - 3180.00\beta_2 - 7327.39\beta_3 + 2693.26\beta_4 - 5496.25\beta_5 - 3940.30\beta_6 + 12776.16\beta_7 + 32234.99\beta_8 + 33158.06\beta_9 + 15859.64\beta_10 + 27798.$
To interpret p-value in each factor, the p-values of newBA03Y51961, newBA03Y51966, newBA03Y51971, newBA03Y51996, HD03Y5.factor1966, HD03Y5.factor1971, HD03Y5.factor1976, HD03Y5.factor1981, HD03Y5.factor1986, and EMSEC2 are higher than the significant level(0.05). It means they are irrelevant factors to analyze the summary. The model is a good fit since there is no problem about violation through diagnostic test. In this model, I should make my career professional Engineering and obtain a job in a business or industry with full-time and year-round working(40~52 years) since NDGMEMG5 and EMSEC4

Regression 2: job satisfaction vs other variables I find an ideal model with 16 variables(NDGMEMG+DGRDG+Full.Part+MGRNAT+MGROTH+MGRSOC+WAPRSM+NOCPRMG+AGE.factor+EMSEC+GENDER+RACETH+JOBINS+JOBPENS+JOBPROFT+JOBVAC). The other variables are irrelevant to the job satisfaction or make the model sharply drop the AUROC value when we plotROC test. In the equation,
$Y = 0.1619459 + 0.0032484\beta_1 + 0.0105323\beta_2 + 0.0124570\beta_3 - 0.0110364\beta_4 - 0.0149017\beta_5 + 0.0010115\beta_6 - 0.0041034\beta_7 + 0.0081354\beta_8 - 0.0178883\beta_9 - 0.0$
Since the ROC Curve value(pseudo-R^2) of model is 0.6233 over than 0.6, and it is good fit model. Besides, it is easy to interpret that job

satisfaction has positively been infuenced by the highest degree of field major(NDGMEMG), primary work activity(WAPRSM), secondary work activity(WASCSM), and Job code for principal job(NOCPRMG) in broad outlines. I should choose Social and related sciences(NDGMEMG4) with Doctorate(DGRDG3). This is because the slopes of NDGMEMG4 and DGRDG3 is the highest positive rate. To maximize my job satisfaction, I had better do management and administration(WAPRSM3) as primary work activity and no secondary activity(WASCSM6). Lastly, I would enter to the business or industry(EMSEC4) after graduation as much as I could. In the business or industry, I could work my major job Social and related sciences with field of Biological, agricultural and other life scientists(NOCPRMG2).

Fact Check Through the proportion table: education level vs job satisfaction, we cannot find that there is correlation between the two variables. This is because the difference of percentages are small between 3~14%. Therefore, the claim: Education level has very little to do with job satisfaction is true. In 1-(b), referred from Basic Analysis number 7, job for degree of independence has the greatest influence on the job satisfaction. Thus, the claim: Having the opportunity to do what you do best is the one factor that correlates most highly with overall job satisfaction is, is true. In Diverse education, we did four hypothesis tests dividing 4 degree(bachelor, master, doctoral, professional) as ANOVA. We conclude that Salary and race with same degree are dependent through the tests. Therefore, the claim: Certain race groups earn less than others when they have the same education level is true. In 2-(b), through Chi-square and ANOVA sequentially, we found that minority and the highest degree major are dependent, and the highest degree major and salary are also dependent. Therefore, the claim is true that STEM (science, technology, engineering and mathematics) careers, in which minorities are underrepresented, tend to pay more than careers in social sciences. In question 3, we created a new variable about STEM NDGMEMG and did hypothesis test with OCEDRLP variable as Chi-square test. Through the test, we could conclude that 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.