Basic analysis.

Population and sampling

The SESTAT is a subset of all three surveys(NSCG, NSRCG, SDR) conducted by the National Science Foundation(NSF). Individuals who have science/ engineering degrees or occupations are in this database. This implies that the NSCG survey can have individuals without a bachelor and/or occupation in S&E.

1. The surveys are 2013 SESTAT SDR and the 2013 SESTAT NSCG. The population of the NSCG is the entire field of non-institutionalized people under the age of 76 in the United States who hold at least a bachelor's degree in any field from any accredited instution. The relevant conditions must be true at the time the U.S Census was taken. Every 10 years, the survey uses a two-stage sampling design to use a stratified systematic sampling from only the American Community Survey (ACS,starting in 2010) and only the U.S Census(decennial) data in 1990 and 2000. The sub-samples from that selection were selected through stratified sampling with strata based on age, race, highest degree type, occupation, and sex. This new cohort is sent follow-ups, as part of the NSCG survey,for the rest of the decade. With the exception of the 2010 and the first NSCG survey of the decade, the sample of the ith year NSCG survey is the $(i-2)$th year NSRCG sample in addition to the $(i-2)$th year NSCG sample where i is initially 1995. The dataset used in this report will be the SESTAT version of the corresponding survey. A sampling bias that could occur is the non-response bias from the web-administered survey, mail survey, and the follow-up phone interview for initial respondents for the selected participants of the second stage sampling. Stratified sampling assumes that an exhaustive, list of the population is known. No census, despite repercussions of the law, can capture the entire population. In the second stage of the sample, the difficulty of the classification of observations into a strata such that an observation is only in one strata of the survey can cause overlap. Significant overlap is a cause for the sample to not to be representative of the population, and as a result, a sampling bias.

   The population of SDR is the total listing of doctorates from the Survey of Earned Doctorates(SED) database, an annual census conducted by the NSF and federal agencies which tracks all individuals receiving a research doctorates degree in science, engineering, or health from accredited US institutions in a given year. The sampling design is a stratified sampling of the SED database every two to three years and follow-up of individuals from previous SDR surveys. The sample of a SDR survey for a given year is a new cohort from the SED database selected using the sampling design and added to the study as well as recording the individuals from previous iterations of the survey with follow-ups until the age of 76. Unlike the NCSG survey, there SESTAT version of this survey is the same survey. A sampling bias that could occur is the unit non-response bias of selected individuals contacted by mail. Furthermore, the item non-response bias is caused by the SED not enforcing the completion of all questions (location after graduation being the most frequent culprit).

   Yes, the 2013 SESTAT SDR includes only individuals receiving a research doctorates degree in science, engineering, or health from accredited US institutions according to the sampling design, but the 2013 SESTAT NSCG will include individuals who recieved their bachelors from accredited instituions not in the US. The combination of both datsets introduces the question why individuals who have recieved doctorates from international instituitions are not considered when the research question investigates the impact of a college degree, no matter the origin, on the individual's relevant attributes.

Demographics

We will specifically describe the distributions of gender, minority, race/ethnicity

In [358]:

```
library(ggplot2)
library(gridExtra)
df <- read.csv("../input/data.formatted.csv")
head(df)
```

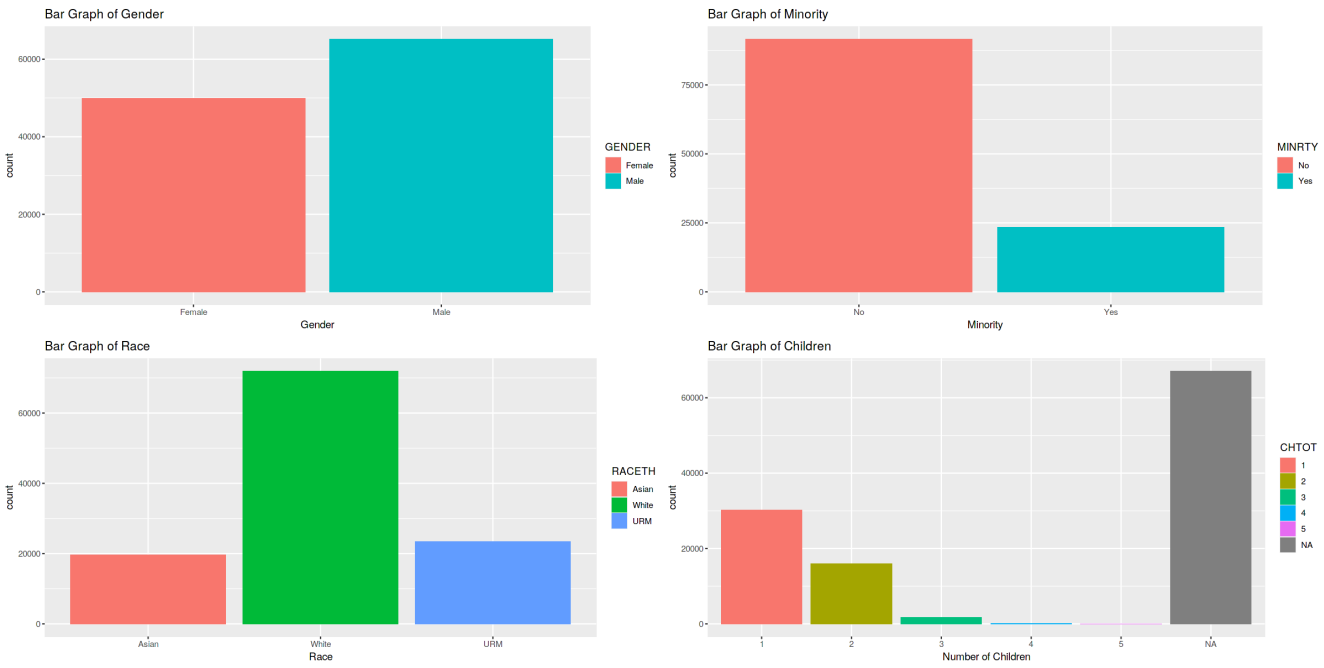
A data.frame: 6 × 63

	PERSONID	YEAR	WEIGHT	SAMPLE	SURID	AGE	GENDER	MINRTY	RACETH	CHTOT	⋯	SATSOC	MGRNAT	MGROTH	MGF
	<dbl>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	⋯	<int>	<int>	<int>	
1	2e+16	2013	30.9312	1002	2	56	1	0	2	NA	⋯	1	0	1	
2	2e+16	2013	31.1697	1002	2	57	1	0	2	NA	⋯	NA	NA	NA	
3	2e+16	2013	31.1697	1002	2	59	1	0	2	NA	⋯	1	0	1	
4	2e+16	2013	31.1697	1002	2	58	1	0	2	NA	⋯	1	1	0	
5	2e+16	2013	32.7715	1002	2	58	1	0	2	NA	⋯	NA	NA	NA	
6	2e+16	2013	30.3566	1002	2	61	2	0	2	NA	⋯	1	1	0	

In [359]:

```
temp<-df
cols <- c("GENDER","CHTOT","RACETH","MINRTY")
temp[,cols] <- lapply(temp[,cols], factor)

levels(temp$GENDER) = c("Female","Male")
levels(temp$MINRTY) = c("No","Yes")
levels(temp$RACETH) = c("Asian","White","URM","other")

saved <- options(repr.plot.width=15, repr.plot.height=10)
p1 <- ggplot(temp, aes (x=GENDER))+geom_bar(aes(fill=GENDER)) + xlab("Gender") + ggtitle("Bar Graph
of Gender")
p2 <- ggplot(temp, aes (x=MINRTY))+geom_bar(aes(fill=MINRTY)) + xlab("Minority")+ ggtitle("Bar Grap
h of Minority")
p3 <- ggplot(temp, aes (x=RACETH))+geom_bar(aes(fill=RACETH)) + xlab("Race")+ ggtitle("Bar Graph of
Race")
p4 <- ggplot(temp, aes (x=CHTOT))+geom_bar(aes(fill=CHTOT)) + xlab("Number of Children")+ ggtitle("
Bar Graph of Children")
options(saved)

prop.table(table(temp$RACETH))
prop.table(table(temp$MINRTY))
prop.table(table(temp$CHTOT))
saved <- options(repr.plot.width=20, repr.plot.height=10)
grid.arrange(p1, p2, p3, p4, nrow = 2)
saved <- options(repr.plot.width=20, repr.plot.height=10)
```

```
    Asian     White       URM     other
0.1709045 0.6252084 0.2038870 0.0000000


      No      Yes
0.796113 0.203887


           1            2            3            4            5
0.6288676967 0.3317577979 0.0364802399 0.0022279599 0.0006663058
```



GENDER: There are about 30% more males than females.

RACETH: There are only 3 categories - Asian, White, and Other. Whites comprise approximately 63% of the data while Asians and Other comprise around 20% each.

MINTRY: There are only 2 categories - being part of a minority and not. As expected, there are fewer tagged with the MINRTY flag that those who are not. It is about a 1:5 ratio.

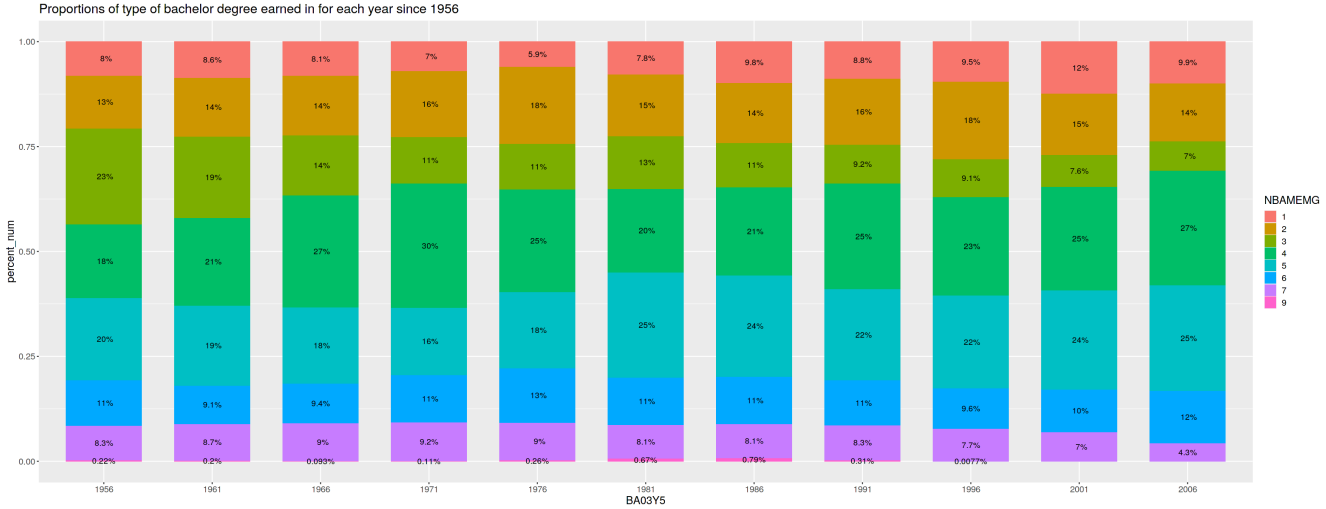
CHTOT: There are 5 valid categories - zero children, one child, one to three children, two or more children, and more than three

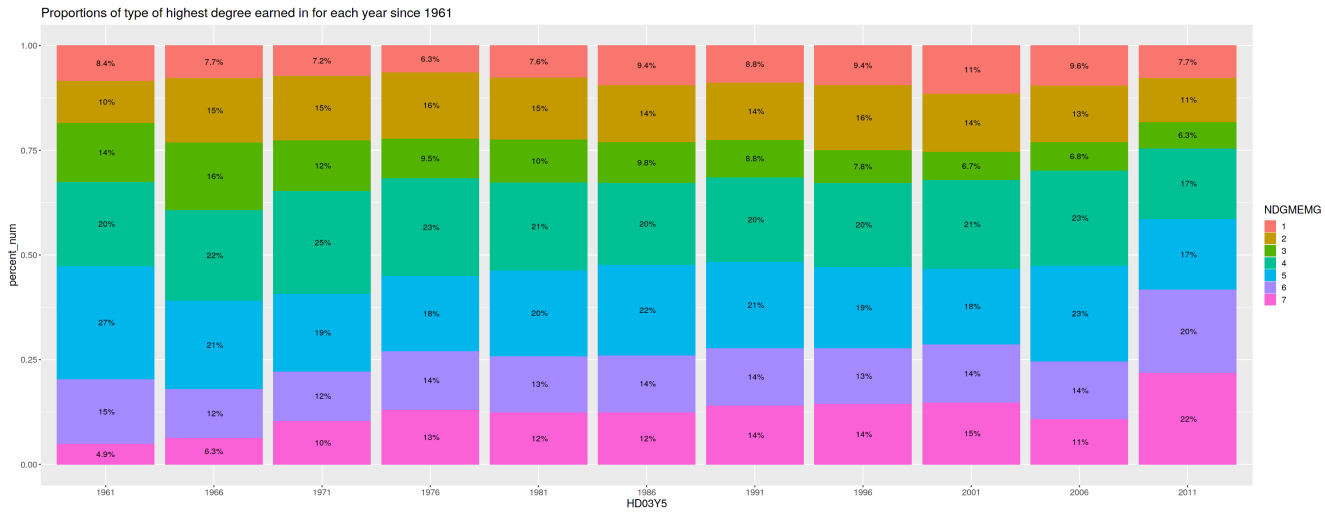
OTHER: There are 5 valid categories - zero children, one child, one to three children, two or more children, and more than three children. The use of NA(the category with the greatest number of responses) was ignored in the calculation of proportions. Most of the data falls in the first bucket(62%) followed by the second category(33%) and the remaining categories with small percentages.

Education

In [360]:

```
library(data.table)
library(scales)
library(dplyr)

cols <- c("BA03Y5","NBAMEMG","HD03Y5","NDGMEMG")
temp[,cols] <- lapply(temp[,cols], factor)
temp <- temp[temp$NBAMEMG != "96",]
temp <- temp[temp$NDGMEMG != "96",]
temp <- temp[temp$BA03Y5 != "9999",]
# levels(temp$NBAMEMG) <- c("Computer and mathematical sciences","Life and related sciences", "Physical and related sciences", "Social and related sciences", "Engineering", "Science and engineering-related fields", "Non-science and engineering fields","Other categories")
# levels(temp$NDGMEMG) <- c("Computer and mathematical sciences","Life and related sciences", "Physical and related sciences", "Social and related sciences", "Engineering", "Science and engineering-related fields", "Non-science and engineering fields")


#copied from STACK OVERFLOW
dt <- setDT(temp)[,list(count = .N), by = .(BA03Y5,NBAMEMG)][,list(NBAMEMG = NBAMEMG, count = count,
                percent_fmt = paste0(formatC(count*100/sum(count), digits = 2), "%"),
                percent_num = count/sum(count),
                cum_pct = cumsum(count/sum(count)),
                label_y = (cumsum(count/sum(count)) + cumsum(ifelse(is.na(shift(count/sum(count))),0,shift(count/sum(count))))) / 2
), by = BA03Y5]

dt1 <- setDT(temp)[,list(count = .N), by = .(HD03Y5,NDGMEMG)][,list(NDGMEMG = NDGMEMG, count = count,
                percent_fmt = paste0(formatC(count*100/sum(count), digits = 2), "%"),
                percent_num = count/sum(count),
                cum_pct = cumsum(count/sum(count)),
                label_y = (cumsum(count/sum(count)) + cumsum(ifelse(is.na(shift(count/sum(count))),0,shift(count/sum(count))))) / 2
), by = HD03Y5]

#create plots that
saved <- options(repr.plot.width=26, repr.plot.height=10)
ggplot(dt, aes (x=BA03Y5,y=percent_num,fill = NBAMEMG))+geom_bar(position=position_fill(reverse=FALSE), stat = "identity",width=0.7) +geom_text(aes(label = percent_fmt),position = position_stack(vjust = 0.5))+theme(text = element_text(size=15))+ggtitle("Proportions of type of bachelor degree earned in for each year since 1956")
ggplot(dt1, aes (x=HD03Y5,y=percent_num,fill = NDGMEMG))+geom_bar(position=position_fill(reverse=FALSE), stat = "identity",) +geom_text(aes(label = percent_fmt),position = position_stack(vjust = 0.5))+theme(text = element_text(size=15))+ggtitle("Proportions of type of highest degree earned in for each year since 1961")
options(saved)
```





Summarizing the distribution of highest degrees and bachelor degrees by field and year obtained relies on knowing how the ratio of field of major within a given year changes from 1956 to 2006 for each type of degree.

Bachelors

- Computer and mathematical sciences - Starting at ~8%, the ratio dips to 5% thoughout the 80s, and returns to about the 1956 levels by 2006.
- Life and related sciences - Starting at ~13%, the ratio steadily increases to 18% in 1976, and returns to about the 1956 levels by 2006.
- Physical and Related Sciences - Starting at 23% in 1956(all time high), the ratio steadily decreases to ~6% until 2006.
- Social and related sciences - Starting at ~18% in 1956, the ratio steadily increases to ~ 27% in 2006. However, the ratio in 1971 to 1986 saw strict decreases.
- Engineering - Aside from the spike from 1966 to 1981 (~18% to ~25%), the Engineering ratio remains around 20%.
- Science and engineering-related fields - Starting at 11%, the greatest increase in the ratio occur between 1966 and 1976. Otherwise, the ratio is around ~11%.
- Non-science and engineering fields - The ratio remains around ~8% until 1991 where after it drops ~4% by 2006.
- Other category is neglible for the period of time during 1956 to 2006.
- The missing category was dropped in order for the percentages to be correct.

Highest Degree

- Computer and mathematical sciences - The ratio remains around ~8% until the period of time between 1991 and 1996 where it grows to ~10% and remains there until 2006.
- Life and related sciences - The ratio remains around ~14% until the period of time between between 1996 and 2001 where it drops to ~12% and remains there until 2006.
- Physical and Related Sciences - The ratio steadily decreases to ~5% until 2006. The greatest decrease occurs between 1961 and 1966.
- Social and related sciences - The ratio remains around 20% with the greatest increase during the period of time between 1956 and 1971.
- Engineering - The ratio remains around 16% during the period of time between 1956 to 1976. The greatest increase to ~23% is during the period of time between 1976 and 1981. The ratio during the remaining years is around 21%.
- Science and engineering-related fields - Starting at ~11%, the ratio increases to ~13% and remains around ~13% during the period of time between 1966 to 1996. The ratio increases to ~15% during the period of time between 1996 and 2006.
- Non-science and engineering fields - The greatest increase of the ratio is the period of time between 1961 and 1966 from 9.9% to 14%. The ratio remains around ~14% during the period of time between 1966 to 2001. There is a drop to ~9% during the period of time between 2001 and 2006.

For those who obtained more than a bachelor degree, is there a significant difference in retention rates among different field of majors?

The test is the permutation test.

$H_O: $ There is no difference in retention rate among different field of majors between their bachelor degree and their highest degree.

$H_A: $ There is a difference in retention rate among different field of majors between their bachelor degree and their highest degree.

Let d_i be the observed $\sum_{i=1}^{7} |r_i - \bar{r}| $ where r is the rate of people with a bachelor degree would do a higher degree in the same field of major where i is the degree code in NDGMEMG and NBAMEMG.

Forming a contingency table would reveal this rate is along the diagonals.

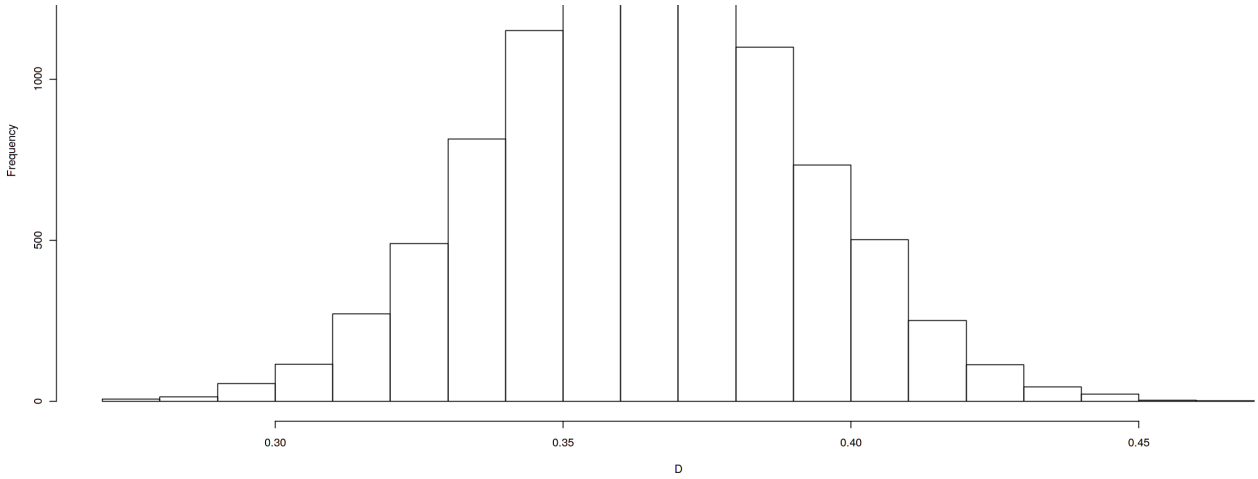
Form a distribution D from the permuation test such that we can find $P(E|H_0)$ where $E=D \geq d$

This means we are looking for the probability that an event is as extreme or more extreme than what we saw in the observed data.

In [361]:

```
df_train <- df[sample(nrow(df),nrow(df)*.1,replace=FALSE),]
temp<-df_train
#only include those who highest degree is not a bachelors
temp<- temp[temp$DGRDG >1,]
temp$NDGMEMG <- as.factor(temp$NDGMEMG)
temp$NBAMEMG <- as.factor(temp$NBAMEMG)

# only include those who have a higher degree than a bachelors
temp <- temp[temp$NDGMEMG != "9" & temp$NDGMEMG != "96",]
temp <- temp[temp$NBAMEMG != "96" & temp$NBAMEMG != "98" & temp$NBAMEMG != "99" & temp$NBAMEMG !=
"9",]
#compute the data deviation
data.deviation <- function(rates){
    r.bar <- mean(rates)
    s <- sum(abs(rates-r.bar))
    return (s)
}

tb <- prop.table(table(temp$NDGMEMG,temp$NBAMEMG),margin=2)
tb
d<- data.deviation(diag(tb))

shuffle <- function(){
    NDGMEMG.shuffle <- sample(temp$NDGMEMG)
    tb <- prop.table(table(temp$NBAMEMG,NDGMEMG.shuffle),margin=2)
    test_statistic <- data.deviation(diag(tb))
    return (test_statistic)
}

num.exp <- 10**4
D <- replicate(num.exp,shuffle())
p.hat <-sum(D>=d)/num.exp

print(paste("Observed Test Statistic",d,"; p-value: ", p.hat))
hist(D)
```

```
             1           2           3           4           5           6
1 0.580139373 0.009900990 0.030786773 0.012113870 0.064371257 0.025862069
2 0.029616725 0.625562556 0.092360319 0.018170806 0.012724551 0.074137931
3 0.027874564 0.033303330 0.671607754 0.004239855 0.022455090 0.010344828
4 0.047038328 0.030603060 0.017103763 0.569351908 0.014221557 0.050000000
5 0.083623693 0.016201620 0.084378563 0.006056935 0.731287425 0.027586207
6 0.054006969 0.198019802 0.045610034 0.068443368 0.026197605 0.639655172
7 0.177700348 0.086408641 0.058152794 0.321623259 0.128742515 0.172413793

             7 9 96
1 0.064406780
2 0.067796610
3 0.033898305
4 0.450847458
5 0.023728814
6 0.171186441
7 0.188135593


[1] "Observed Test Statistic 0.774019158548528 ; p-value:  0"
```



phat(~0) is less than any α > 0, so we reject H_O and the results are significant. There is sufficient evidence to believe that for those who obtained more than a bachelor degree, there is a difference in retention rate among different field of majors between their bachelor degree and their highest degree.

Job status

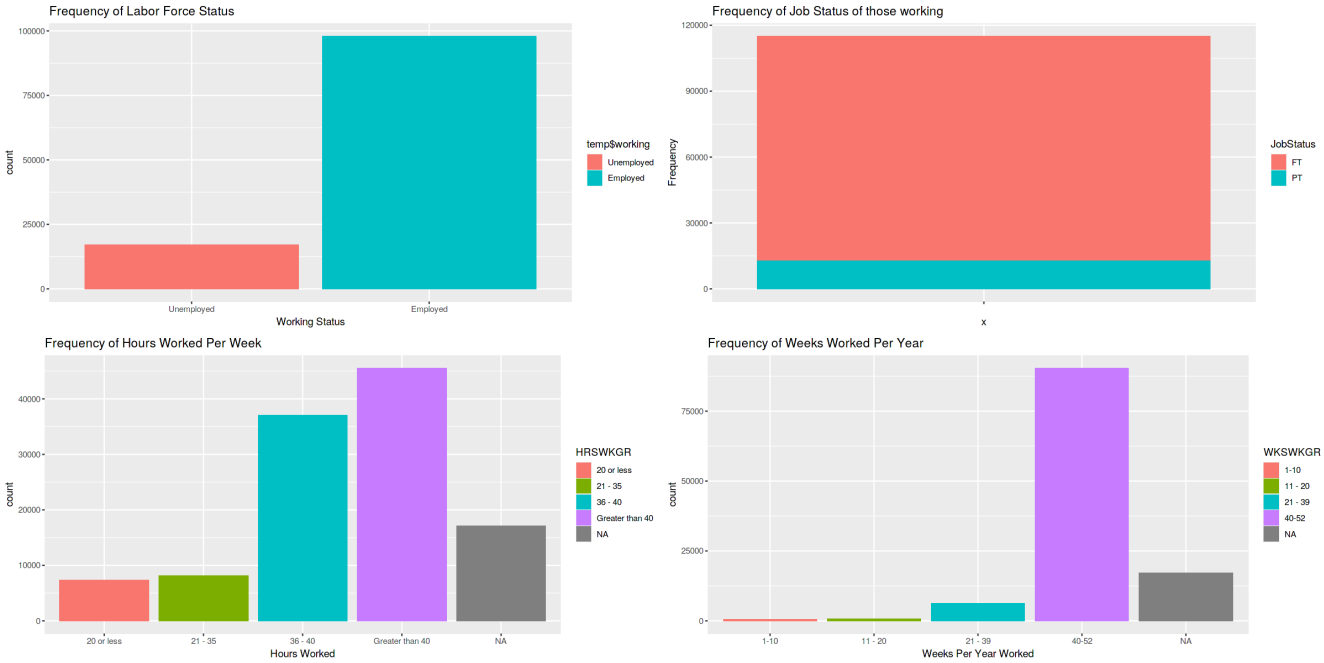
This section will show percent of people working, percent working part-time, number of hours per week and number of weeks per year. It will also cover statistical testing of relevance of degree with a variety of factors.

In [362]:

```
# % of people working
temp<-df
temp$LFSTAT <- as.factor(temp$LFSTAT)

temp$working <- temp$LFSTAT == "1"
temp$working <- as.factor(temp$working)
levels(temp$working) = c("Unemployed", "Employed")
p1 <- ggplot(temp, aes(x=working)) + geom_bar(aes(fill=temp$working))+xlab("Working Status") + ggti
tle("Frequency of Labor Force Status")

# count the occurences of rows for which all the part time responses are not yes.
# we must do this instead of counting the rows where the part time responses because a given perso
n can select more than one category and a response for "no" does not
# imply being full time
num_working_FT <- nrow(temp[temp$PTWTFT != 01 & temp$PTFAM != 01 & temp$PTNOND != 01 & temp$PTOCNA
!= 01 & temp$PTOTP !=01, ])
print(paste("Ratio of Full Time Workers:",num_working_FT/nrow(temp), ";Number of Part Time
Workers:", (nrow(temp)-num_working_FT)/nrow(temp)))

pt_df <- data.frame(JobStatus=c("PT","FT"),Frequency=c(nrow(df)-num_working_FT,num_working_FT))
p2 <- ggplot(pt_df, aes(x="",y=Frequency,fill=JobStatus)) + geom_bar(width=1,stat="identity")+ ggti
tle("Frequency of Job Status of those working ")


temp$HRSWKGR <- as.factor(temp$HRSWKGR)
print(paste("Ratio of Workers in NA category of Hours Worked Per Week: ", sum(is.na(temp$HRSWKGR))
/nrow(temp)))
levels(temp$HRSWKGR) = c("20 or less", "21 - 35", "36 - 40", "Greater than 40", "Logical Skip")
p3 <- ggplot(temp, aes(x=HRSWKGR)) + geom_bar(aes(fill=HRSWKGR))+xlab("Hours Worked")+ ggtitle("Fre
quency of Hours Worked Per Week")


temp$WKSWKGR <- as.factor(temp$WKSWKGR)
print(paste("Ratio of Workers in NA category of Number of Weeks Worked Per Year: ", sum(is.na(temp
$WKSWKGR))/nrow(temp)))
levels(temp$WKSWKGR) = c("1-10", "11 - 20", "21 - 39", "40-52", "Logical Skip")
p4 <- ggplot(temp, aes(x=WKSWKGR)) + geom_bar(aes(fill=WKSWKGR))+xlab("Weeks Per Year Worked")+ ggt
itle("Frequency of Weeks Worked Per Year")


prop.table(table(temp$working))
prop.table(table(temp$HRSWKGR))
prop.table(table(temp$WKSWKGR))
saved <- options(repr.plot.width=20, repr.plot.height=10)
```

```
grid.arrange(p1, p2, p3, p4, nrow = 2)
saved <- options(repr.plot.width=20, repr.plot.height=10)
```

```
[1] "Ratio of Full Time Workers: 0.887253369459497 ;Number of Part Time Workers:
0.112746630540503"
[1] "Ratio of Workers in NA category of Hours Worked Per Week:  0.148508058913436"
[1] "Ratio of Workers in NA category of Number of Weeks Worked Per Year:  0.148508058913436"


Unemployed   Employed
 0.1485081  0.8514919


     20 or less         21 - 35         36 - 40 Greater than 40    Logical Skip
     0.07495079      0.08271206      0.37754842      0.46478873      0.00000000


        1-10       11 - 20       21 - 39        40-52 Logical Skip
 0.006486420  0.007689876  0.063150809  0.922672895  0.000000000
```



- Percent of people working - There are approximately 14.85% workers unemployed and 85.14% employed workers
- Percent working part-time - There are ~88.73% full-time employees, and 11.27% part-time employees in the sample.
- Number of hours per week - The majority of workers work close to the standard amount of hours to be considered full-time employees(~40 hours) with approximately 84.22% in this category. This percentage does not include workers indicating NA(unemployed), but there is ~14.85% of the total sample in that category.
- Number of weeks per year - The majority of workers (~92.27%) work 40-52 weeks per year. The remaining participants fill the remaining categories with decreasing number of hours worked per year corresponding to decreasing percentages. These percentages does not include workers indicating NA(unemployed), but there is ~14.85% of the total sample in that category.

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

Assume a high number of hours per week to be greater than 40 (04) and a few weeks to be 1-10 weeks (01)

In [363]:

```
temp<-temp[temp$LFSTAT == "1",]
combos <- temp %>% count(HRSWKGR, WKSWKGR, sort = FALSE)
combo.df <- data.frame(combos)
combo.df
#equvialently nrow(temp[temp$HRSWKGR == "Greater than 40" & temp$WKSWKGR == "1-10",])-sum(is.na(te
mp$HRSWKGR))
```

A data.frame: 16 × 3

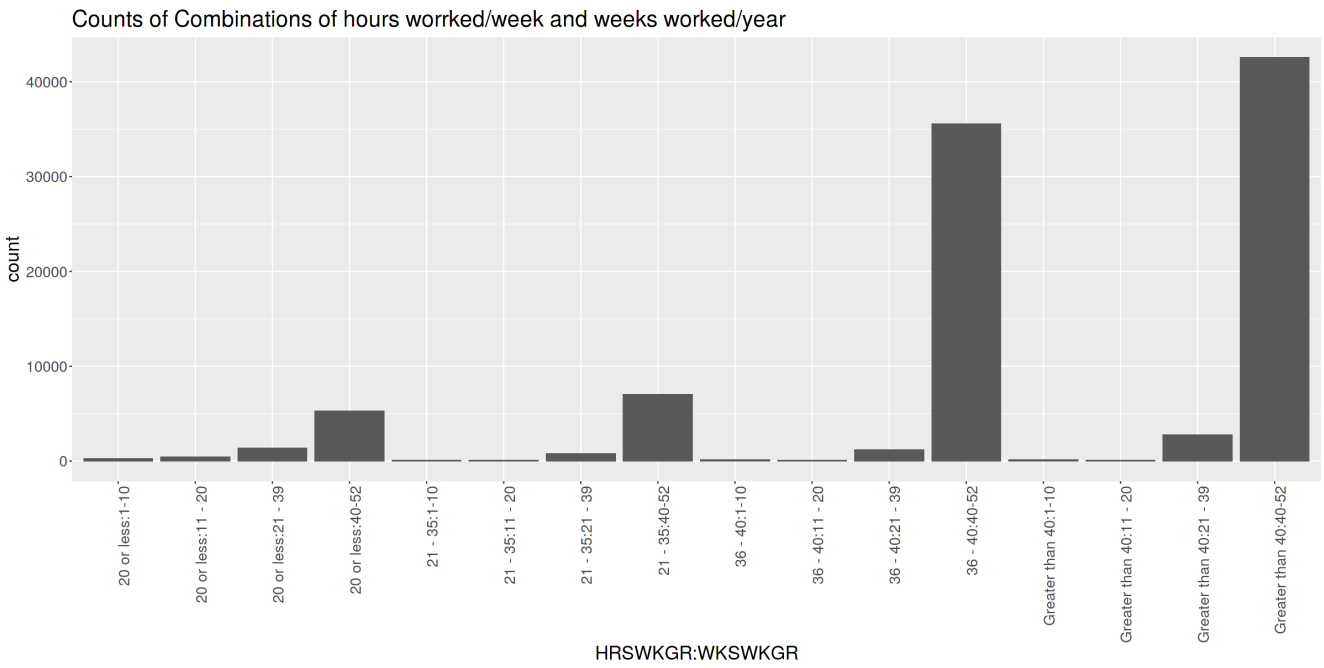
HRSWKGR	WKSWKGR	n

HRSWKGR <fct>	WKSWKGR <fct>	n <int>
20 or less	1-10	261
20 or less	11 - 20	452
20 or less	21 - 39	1363
20 or less	40-52	5273
21 - 35	1-10	109
21 - 35	11 - 20	127
21 - 35	21 - 39	806
21 - 35	40-52	7068
36 - 40	1-10	133
36 - 40	11 - 20	100
36 - 40	21 - 39	1229
36 - 40	40-52	35557
Greater than 40	1-10	133
Greater than 40	11 - 20	75
Greater than 40	21 - 39	2794
Greater than 40	40-52	42571

In [364]:

```
# levels(temp$HRSWKGR) = c("00", "01", "02", "03", "04")
# levels(temp$WKSWKGR) = c("00", "01", "02", "03", "04")
ggplot(temp,aes(x=HRSWKGR:WKSWKGR))+geom_bar()+theme(text = element_text(size=20),axis.text.x = ele
ment_text(angle=90, hjust=1))+ggtitle("Counts of Combinations of hours worrked/week and weeks
worked/year")
print((42571 + 35557 + 7068 + 5273)/sum(combo.df$n))
```

```
[1] 0.9226729
```



The following bar graph with the x-axis being the interaction of $HRSWKGR$ and $WKSWKGR$ shows that those employed work in the following variable combinations (listed in decreasing frequency) - (3,3), (2,3), (1,3), (0,3). This means very few work greater than 40 hours per week and 1-10 weeks per year(133 people), and in fact 78% of the sample work with regular hours all year round (variable combinations above)

We will show the major reasons that led people to not work at the time of survey. Since participants in the survey can choose more than one option, we have to look at the frequency of interactions to answer the question.

In [365]:

```
temp<-df
# to avoid NA warnings, remove those who are working to see reason for why people ARE
temp<-temp[is.na(temp$SALARY),]
temp$NWFAM <- as.factor(temp$NWFAM)
temp$NWLAY <- as.factor(temp$NWLAY)
temp$NWNOND <- as.factor(temp$NWNOND)
temp$NWOTP <- as.factor(temp$NWOTP)
temp$NWSTU <- as.factor(temp$NWSTU)

combos <- temp %>% count(NWFAM,NWLAY,NWNOND,NWOTP,NWSTU, sort = TRUE)
combo.df <- data.frame(combos)
combo.df
```

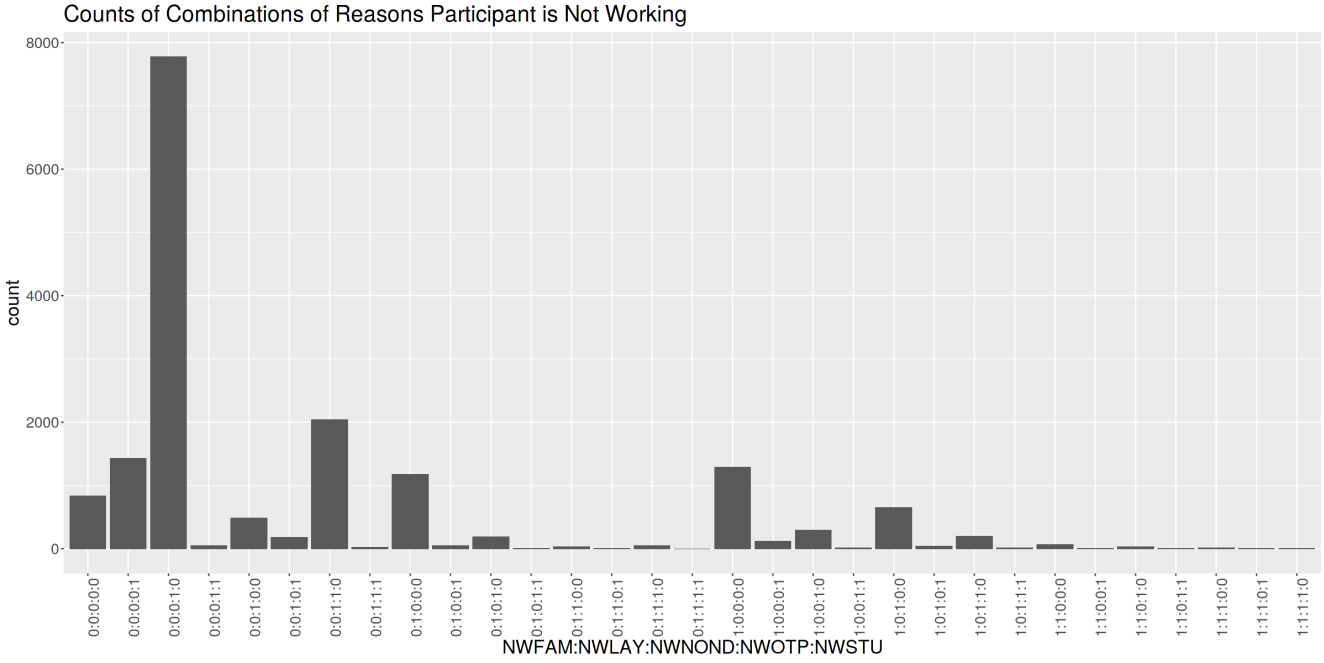
A data.frame: 31 × 6

NWFAM	NWLAY	NWNOND	NWOTP	NWSTU	n
<fct>	<fct>	<fct>	<fct>	<fct>	<int>
0	0	0	1	0	7776
0	0	1	1	0	2043
0	0	0	0	1	1429
1	0	0	0	0	1286
0	1	0	0	0	1178
0	0	0	0	0	834
1	0	1	0	0	656
0	0	1	0	0	487
1	0	0	1	0	297
1	0	1	1	0	199
0	1	0	1	0	193
0	0	1	0	1	181
1	0	0	0	1	124
1	1	0	0	0	67
0	0	0	1	1	53
0	1	0	0	1	49
0	1	1	1	0	49
1	0	1	0	1	43
0	1	1	0	0	32
1	1	0	1	0	31
0	0	1	1	1	22
1	0	1	1	1	17
1	1	1	0	0	16
1	0	0	1	1	12
1	1	1	1	0	9
0	1	0	1	1	5
1	1	1	0	1	4
0	1	1	0	1	3
1	1	0	1	1	3
1	1	0	0	1	2
0	1	1	1	1	1

In [366]:

```
saved <- options(repr.plot.width=20, repr.plot.height=10)
```

```
ggplot(temp,aes(x=NWFAM:NWLAY:NWNOND:NWOTP:NWSTU))+geom_bar()+theme(text = element_text(size=20),
        axis.text.x = element_text(angle=90, hjust=1)) + ggtitle("Counts of Combinations of Reasons
Participant is Not Working")
options(saved)
```



98051 particpants (those that are employed and obviously have no reason to have a reason not to work) put NA for all cateogories. However, there are those (834) who put 0 (No) for all reasons which means those individuals had another reason for not working that was not a option in the surveys. The most common reasons that led people not to work in decreasing frequency are

1. illness, retired, and other - 7776 people
2. did not need/want to work and illness, retired or other - 2043 people
3. student - 1429 people
4. family responsibilities - 1286 people
5. layoff -- 1178 people

The remaining reasons are combinations of each variable. The frequencies of those results are all below 1000 (insignificant compared to the top 5 reasons).

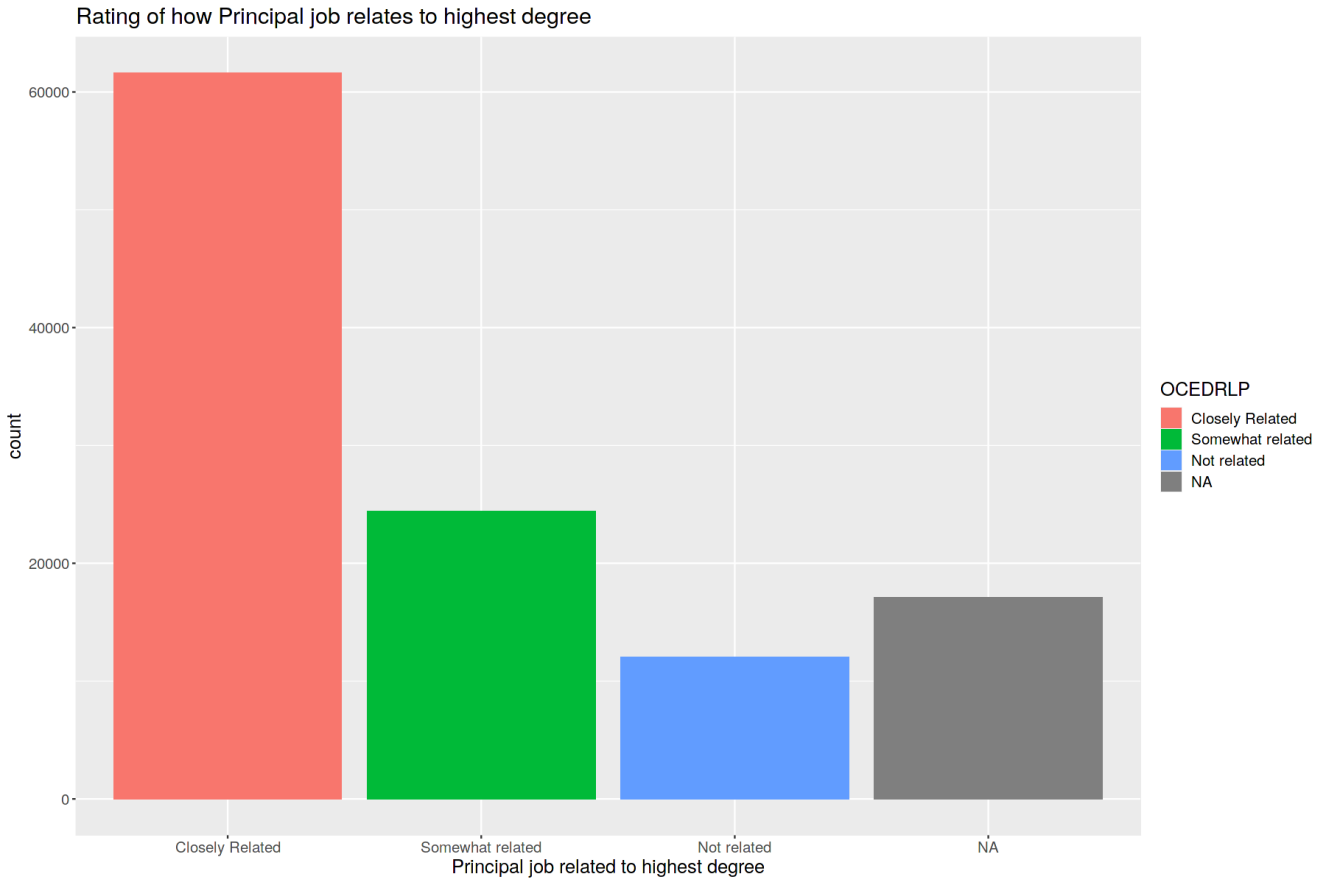
Degree Relevance

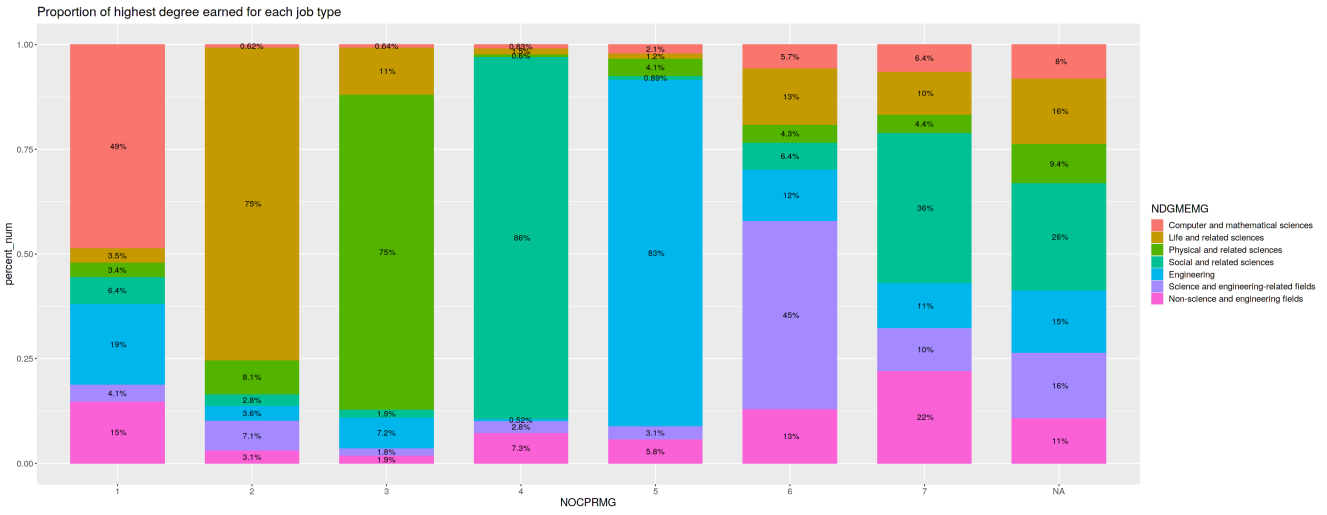
This subsection will show how relevant are people's degree to their principle job using $OCEDRLP$, $NDGMEMG,$ and $NOCPRMG$.

In [367]:

```
temp<-df
temp$OCEDRLP <- as.factor(temp$OCEDRLP)

saved <- options(repr.plot.width=15, repr.plot.height=10)
levels(temp$OCEDRLP) = c("Closely Related", "Somewhat related", "Not related", "Logical Skip")
levels(temp$OCEDRLP) = c("Closely Related", "Somewhat related", "Not related", "Logical Skip")
ggplot(temp, aes(x=OCEDRLP)) + geom_bar(aes(fill=OCEDRLP))+theme(text = element_text(size=15)) + xl
ab("Principal job related to highest degree")+ggtitle("Rating of how Principal job relates to high
est degree")
p<- prop.table(table(temp$OCEDRLP))

temp$NDGMEMG <- as.factor(temp$NDGMEMG)
temp$NOCPRMG <- as.factor(temp$NOCPRMG)

#copied from STACK OVERFLOW
temp.1 <- temp
levels(temp.1$NDGMEMG) <- c("Computer and mathematical sciences","Life and related sciences",
"Physical and related sciences", "Social and related sciences", "Engineering", "Science and engine
ering-related fields", "Non-science and engineering fields","Missing")
dt <- setDT(temp.1)[,list(count = .N), by = .(NOCPRMG,NDGMEMG)][,list(NDGMEMG = NDGMEMG, count = co
unt,
                percent_fmt = paste0(formatC(count*100/sum(count), digits = 2), "%"),
```

```
                percent_num = count/sum(count),
                cum_pct = cumsum(count/sum(count)),
                label_y = (cumsum(count/sum(count)) + cumsum(ifelse(is.na(shift(count/sum(count))),0
shift(count/sum(count))))) / 2
), by = NOCPRMG]

saved <- options(repr.plot.width=26, repr.plot.height=10)

ggplot(dt, aes (x=NOCPRMG,y=percent_num,fill = NDGMEMG))+geom_bar(position=position_fill(reverse=FA
LSE), stat = "identity",width=0.7) +geom_text(aes(label = percent_fmt),position = position_stack(vj
ust = 0.5))+theme(text = element_text(size=15))+ggtitle("Proportion of highest degree earned for e
ach job type")

# ggplot(temp, aes(x=NOCPRMG,fill=NDGMEMG)) + geom_bar(aes(fill=NDGMEMG))+theme(text =
element_text(size=15))
options(saved)
```





Since OCEDRLP is a categorical variable binning how close the principle job of the participant relates to their highest degree, we can use this as a measure of relevance. Not including those unemployed(picking NA), we see that ~62.83% of the sample states their principle job closely relates to their highest degree, followed by ~24.88% stating somewhat relates, and ~12.29% stating not related.

Using the second graph, we can see the distribution of majors for each occupation. Then, it is easy to examine if people work in the field that they were trained for. If this holds, we would expect the ith NOCPRMG category to have the ith NDGEMEMG as the dominant major.

1. 49% of Computer and mathematical scientists who have obtained higher degrees in the surveys have majored in Computer and mathematical sciences fields.
2. 75% of Biological, agricultural and other life scientists who have obtained higher degrees in the surveys have majored in Biological, agricultural and environmental life sciences fields.
3. 75% of Physical and related scientists who have obtained higher degrees in the surveys have majored in Physical and related sciences fields.
4. 86% of Social and related scientists who have obtained higher degrees in the surveys have majored in Social and related sciences fields.
5. 45% of Engineers who have obtained higher degrees in the surveys have majored in engineering fields.
6. 45% of those in the surveys in Science and engineering related occupations and have obtained higher degrees have majored in Science and engineering-related fields.
7. 36% of those in the surveys in Non-science and engineering occupations and have obtained higher degrees have majored in Non-science and engineering fields. This category has a much more uniform distribution that categories 1-6.
8. Those in the NA category must not have higher degrees than their bachelors.

We see that fields 2-4 are mainly dominated ($\geq 50\%$) by those who obtained degrees in preparation for the occupation. Fields 1,5,and 6 are quite close (sub 50% to being dominated by those who obtained degrees in preparation for the occupation. Field 7 is much more uniform in terms of what majors comprise it.

Statistical Testing of Relevance of Degree

Is there a statistically significant difference in relevance of degree vs

1. job type
2. the degree that they are trained for
3. principal activity in people's job

Let's look at 1

We can perform a permutation test

$ H_O: $ There is no difference in relevance of degree and job type.

$ H_A: $ There is a difference in relevance of degree and job type.

Let d_i be the observed $\sum_{i=1}^{7} |r_i - \bar{r}| $ where r is the rate of people with the ith job code of NOCPRMG that is relevant(OCEDRLP = 1) where i ranges from 1 to 7.

Form a distribution D from the permuation test such that we can find $P(E|H_0)$ where $E=D \geq d$

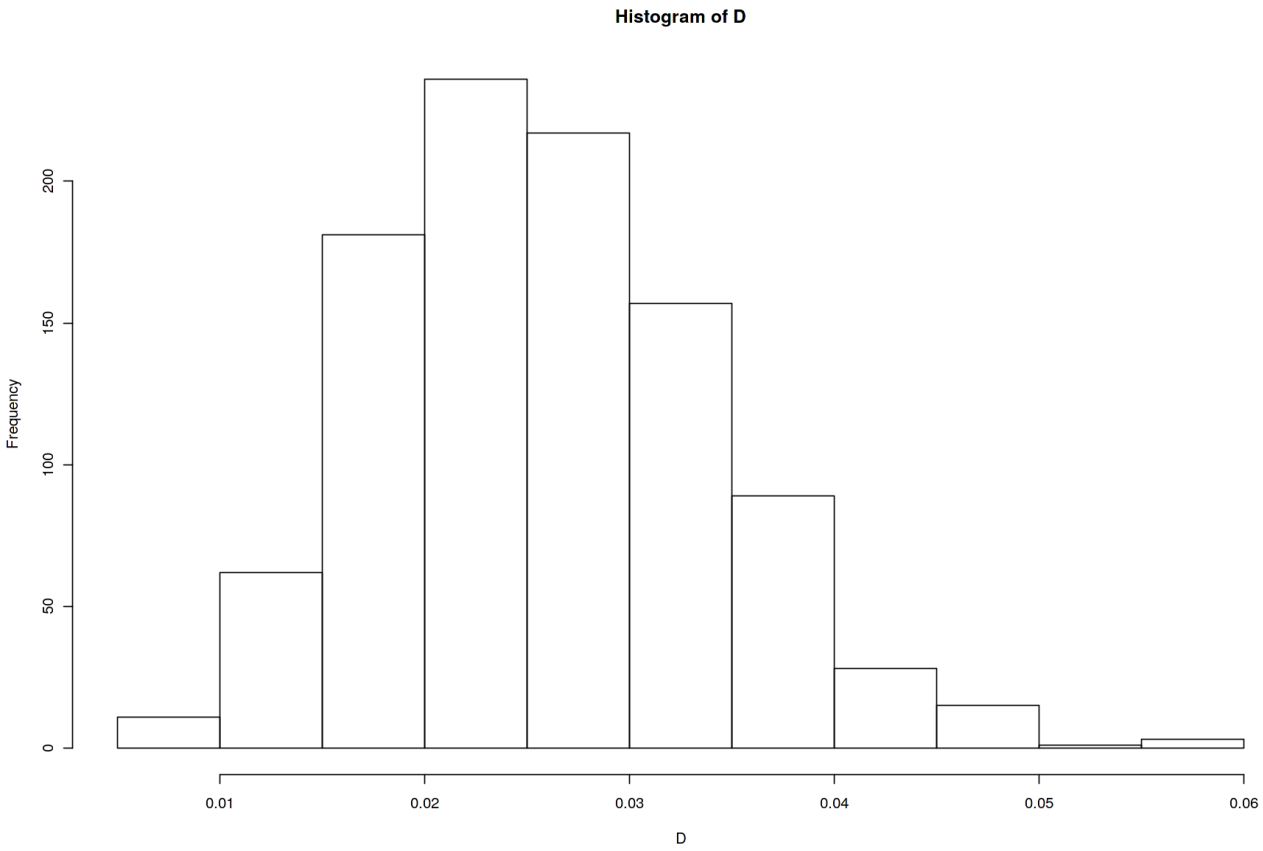
This means we are looking for the probability that an event is as extreme or more extreme than what we saw in the observed data.

In [368]:

```
temp$OCEDRLP <- as.factor(temp$OCEDRLP)
temp$NOCPRMG <- as.factor(temp$NOCPRMG)
# temp$WAPRSM <- as.factor(temp$WAPRSM)
temp <- temp[temp$OCEDRLP != "98",]
# temp <- temp[temp$WAPRSM != "98",]

#compute the data deviation
data.deviation <- function(rates){
    r.bar <- mean(rates)
    s <- sum(abs(rates-r.bar))
    return (s)
}

tb <- prop.table(table(temp$NOCPRMG,temp$OCEDRLP),margin=1)
i<-1
rates<-c()
while(i<=7){
    rates<-append(rates,tb[as.character(i),1])
    i <- i+1
}
d<- data.deviation(rates)

shuffle <- function(){
```

```
    OCEDRLP.shuffle <- sample(temp$OCEDRLP)
    tb <- prop.table(table(temp$NOCPRMG,OCEDRLP.shuffle),margin=1)
    rates<-c()
    i<-1
    while(i<=7){
        rates<-append(rates,tb[as.character(i),1])
        i <- i+1
    }
    test_statistic <- (data.deviation(rates))
    return (test_statistic)
}

num.exp <- 10**3
D <- replicate(num.exp,shuffle())

p.hat <-sum(D>=d)/num.exp
hist(D)
print(paste("Observed Test Statistic",d,"; p-value: ", p.hat))
```

```
[1] "Observed Test Statistic 0.708054688121245 ; p-value:  0"
```



phat(~0) is less than any α > 0, so we reject H_O and the results are significant. There is sufficient evidence to believe that for those who obtained more than a bachelor degree, there is a difference in the job type(job code) and relevance of degree.

Let's look at 2

We can perform a Chi Square Test of Independence

Assumptions: Random sample performed All expected counts are above 10

Our test statistic is $\chi^2 = \frac{(O_i - E_i)^2}{E_i}$

$ H_O: $ Relevance of degree and the degree trained for are independent.

$ H_A: $ Relevance of degree and the degree trained for are not independent.

If the $P(>|\chi|) < \alpha$, reject H_O

In [369]:

```
combos <- df %>% count(OCEDRLP, NDGMEMG, sort = FALSE)
```

```
ctable <- xtabs(n ~ OCEDRLP+NDGMEMG, data=combos)
ctable
chisq <- chisq.test(ctable)
chisq$expected
chisq
```

```
       NDGMEMG
OCEDRLP     1     2     3     4     5     6     7
      1  6051  7948  4829 10820 13491 10794  7669
      2  2262  3658  2267  5050  5804  2055  3301
      3   796  1754   945  4049  1603  1100  1805
```

A matrix: 3 × 7 of type dbl

	1	2	3	4	5	6	7
1	5722.865	8393.619	5051.8779	12514.408	13129.480	8763.667	8026.084
2	2266.497	3324.228	2000.7575	4956.235	5199.830	3470.783	3178.669
3	1119.638	1642.153	988.3645	2448.356	2568.691	1714.550	1570.247

```
 Pearson's Chi-squared test

data:  ctable
X-squared = 3269, df = 12, p-value < 2.2e-16
```

phat(~0) is less than any α > 0, so we reject H_O and the results are significant. There is sufficient evidence to believe that for those who obtained more than a bachelor degree that relevance of degree and the degree trained for are not independent..

Let's look at 3

We can perform a Chi Square Test of Independence

Assumptions: Random sample performed

All expected counts are above 10

Our test statistic is $\chi^2 = \frac{(O_i - E_i)^2}{E_i}$

$ H_O: $ Relevance of degree and the principal activity in people's job are independent.

$ H_A: $ Relevance of degree and the principal activity in people's job are not independent.

In [370]:

```
combos <- df %>% count(OCEDRLP, WAPRSM, sort = FALSE)

ctable <- xtabs(n ~ OCEDRLP+WAPRSM, data=combos)
ctable
chisq <- chisq.test(ctable)
chisq$expected
chisq
```

```
       WAPRSM
OCEDRLP     1     2     3     4     5
      1 19620 10087 12357  4179 15359
      2  6096  2038  9606  1931  4726
      3  1237   713  5908   796  3398
```

A matrix: 3 × 5 of type dbl

	1	2	3	4	5
1	16933.623	8065.665	17510.371	4338.7973	14753.544
2	6706.432	3194.345	6934.848	1718.3474	5843.028
3	3312.945	1577.991	3425.781	848.8553	2886.428

```
 Pearson's Chi-squared test
```

```
data:  ctable
X-squared = 7890.5, df = 8, p-value < 2.2e-16
```

phat(~0) is less than any α > 0, so we reject H_O and the results are significant. There is sufficient evidence to believe that for those who obtained more than a bachelor degree that relevance of degree and the principal activity in people's job are not independent.

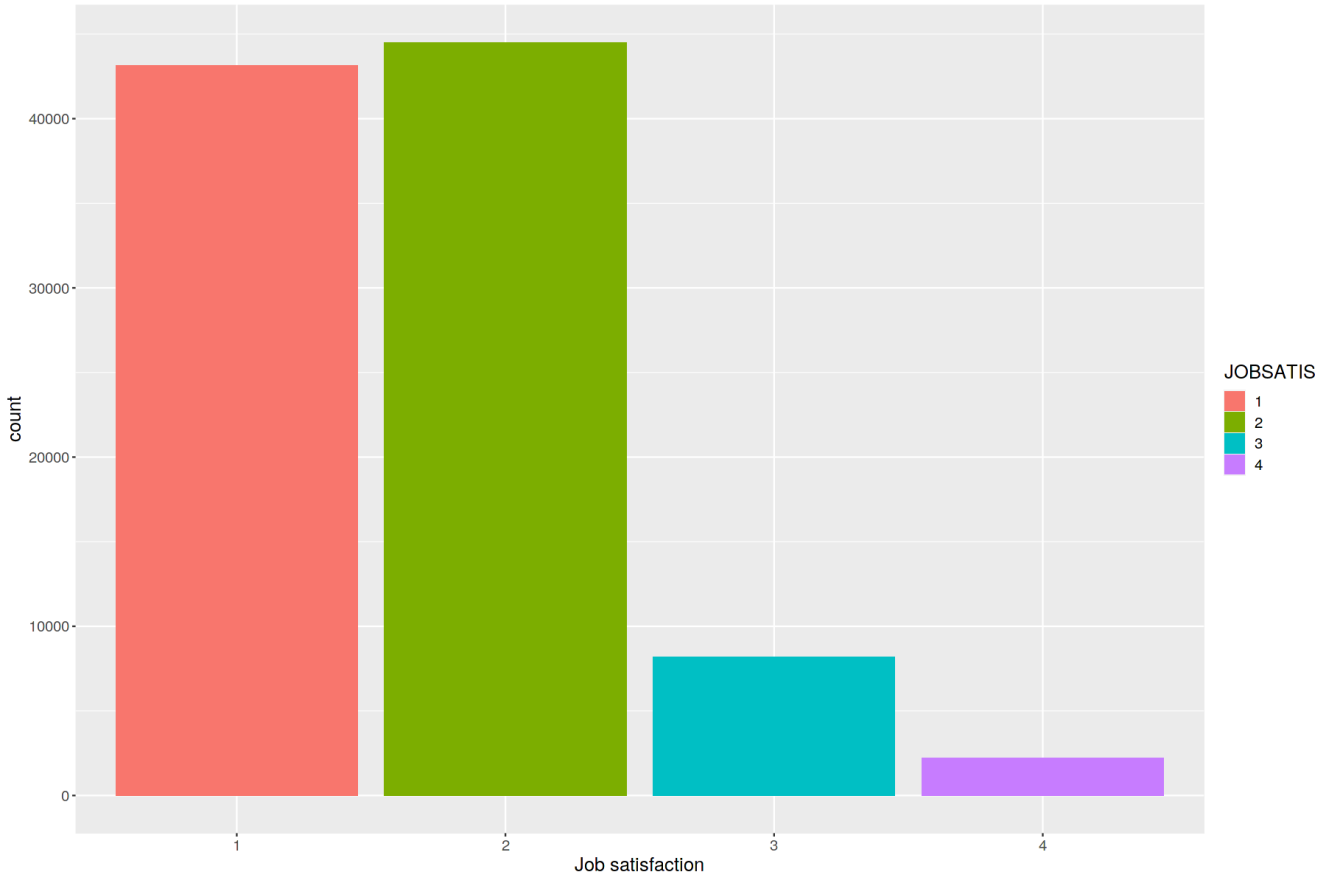
Job Satisfaction

We summarize job satisfaction by observing the proportions of satisfication for those with a job(drop those whose salary is NA).

In [371]:

```
temp <-df
temp<-temp[!is.na(temp$SALARY) & !is.na(temp$JOBSATIS),]
temp$JOBSATIS <- as.factor(temp$JOBSATIS)
ggplot(temp, aes(x=JOBSATIS)) + geom_bar(aes(fill=JOBSATIS))+theme(text = element_text(size=15)) +
xlab("Job satisfaction")
prop.table(table(temp$JOBSATIS))
# levels(temp$SATBEN)<-list("1"=c("1","2"),"2"=c("3","4"))
# levels(temp$SATCHAL)<-list("1"=c("1","2"),"2"=c("3","4"))
# levels(temp$SATIND)<-list("1"=c("1","2"),"2"=c("3","4"))
# levels(temp$SATLOC)<-list("1"=c("1","2"),"2"=c("3","4"))
# levels(temp$SATRESP)<-list("1"=c("1","2"),"2"=c("3","4"))
# levels(temp$SATSAL)<-list("1"=c("1","2"),"2"=c("3","4"))
# levels(temp$SATSEC)<-list("1"=c("1","2"),"2"=c("3","4"))
# levels(temp$SATSOC)<-list("1"=c("1","2"),"2"=c("3","4"))

levels(temp$JOBSATIS)<-list("1"=c("1","2"),"2"=c("3","4"))
```

```
          1           2           3           4
0.44004651 0.45392704 0.08340557 0.02262088
```



Out of those employed, ~99 % are very satisfied or somewhat satisified with their job. However, there is a significant minority (~.08 or ~10000 people) are somewhat dissatisfied with their job. There are .02 out of the employed sample that are very dissatisfied 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?

In [372]:

```
cols <- c("SATADV","SATBEN","SATCHAL","SATIND","SATLOC","SATRESP","SATSAL","SATSEC","SATSOC")
temp[,cols] <- lapply(temp[,cols], factor)

combos <- data.frame(temp[temp$JOBSATIS=="1",] %>% count(JOBSATIS,SATADV,SATBEN,SATCHAL,SATIND,SATL
OC,SATRESP,SATSAL,SATSEC,SATSOC,sort = FALSE))
head(combos)
counts.1<-c()
i<-1
while(i<length(cols)){
    counts.1<-append(counts.1,nrow(temp[temp$JOBSATIS == "1" & temp[, cols[i]] == "1",]) + nrow(tem
p[temp$JOBSATIS == "1" & temp[, cols[i]] == "2",]))
    i<-i+1
}
counts.1/nrow(temp[temp$JOBSATIS == "1",])

counts.2<-c()
i<-1
while(i<length(cols)){
    counts.2<-append(counts.2,nrow(temp[temp$JOBSATIS == "2" & temp[, cols[i]] == "3",]) + nrow(tem
p[temp$JOBSATIS == "2" & temp[, cols[i]] == "4",]))
    i<-i+1
}
counts.2/nrow(temp[temp$JOBSATIS == "2",])

counts.1/nrow(temp[temp$JOBSATIS == "1",])-(counts.2/nrow(temp[temp$JOBSATIS == "2",]))
```

A data.frame: 6 × 11

	JOBSATIS	SATADV	SATBEN	SATCHAL	SATIND	SATLOC	SATRESP	SATSAL	SATSEC	SATSOC	n
	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<int>
1	1	1	1	1	1	1	1	1	1	1	5937
2	1	1	1	1	1	1	1	1	1	2	584
3	1	1	1	1	1	1	1	1	1	3	48
4	1	1	1	1	1	1	1	1	1	4	8
5	1	1	1	1	1	1	1	1	2	1	513
6	1	1	1	1	1	1	1	1	2	2	176

0.739216245507957 0.833666077234613 0.900564713935315 0.948206035023672 0.895955735554161
0.933215446922594 0.825109805487422 0.86721807084593

0.82291265871489 0.463447479799923 0.59176606387072 0.372258560984994 0.298287803001154
0.505290496344748 0.643228164678723 0.476048480184686

-0.0836964132069331 0.37021859743469 0.308798650064595 0.575947474038678 0.597667932553007
0.427924950577846 0.1818816408087 0.391169590661244

For those who reported their primary employment as very satisfied or somewhat satisfied, the top 4 categories they were most satsified were in decreasing order are SATIND at 94.81%,SATRESP at 93.30%,SATBEN at 90.04%, and lastly SATLOC at 89.58%.

For those who reported their primary employment as somewhat/very dissatisfied, the top 4 categories they were most unsatsified were in decreasing order are SATADV at 82.40%,SATSAL at 64.08%,SATCHAL at 59.30%, and lastly SATRESP at 50.60%.

Which factors are most important to job satisfaction?

The factors most important to job satisfaction are those that are indicators of job satisfaction in both groups - those that are satisfied and unsatisfied. It seems that SATRESP is a strong predictor for which group one would belong to. Good predictors of positive job satisfaction are those with the highest positive difference between the proportions of predictors for each group. This would be SATRESP, SATLOC, and SATSAL.

Regression 1: SALARY vs other variables

Dropping Observations

First, numerics that were not relevant to SALARY such as PERSONID, YEAR, SAMPLE, SURID, and CHTOT were dropped. Those that are unemployed,not in the labor force and looking for work, and employed people that had no salary(615 people) were dropped.

The last type of observation is of particular interest to justify removal because the proportion of EMSEC shows that they mainly work for buisness or industry but are unpaid. It seems to only weaken the ability of predictor used for employees who do have a salary. To combine the concerns of a minority who take on unpaid roles(e.g those in retirement) with the vast majority taking on paid roles for vastly different reasons does not seem appropriate. There is no reason to create another model for the 3 categories deleted because the dropped subsets have salaries of 0 dollars, so the model could simply predict 0 for 100% accuracy. The figures discussed are seen in the table and matrix.

Dropping any salaries that are outliers among the full time group or part time group is not appropriate without knowing if its an influential point. Most likely, such observations are completely valid data points and are checked later by plotting Cook's distance.

Variable Selection

My variable selection is determined by building a correlation matrix. I choose the variables whose correlation with SALARY was at least 0.1. However, a linear model simply summing these numerical predictors performs with an R^2 of 0.3592. Next, I choose a subset of these predictors based on my intution of what would be a useful predictor. This means that subset, called cols, would be recoded as categoricals. I created a new cateogorical variable called years_since which represents how many years from 2013 did a person obtain their higher degree to be more clear in interpreting the model.

My second model involved the use of interaction terms. It was as follows $ log(SALARY) = (GENDER + DGRDG + HRSWKGR + WKSWKGR + JOBINS + JOBPROFT + EMSEC + FTPRET + WASCSM + MGRNAT+ SATSAL + NRREA):years_since:CH19IN$

It had a R^2 of .65 and failed the normality assumption. However, it was terribly complicated with 705 coefficients. I decided not to use this model.

My last model is below and is much simpler to interpret because it forgoes interaction terms.

Variable Transformation

Since the Residuals vs Fitted Graph had a curve to it that violated the independece assumption of linear regression, I took the log of the SALARY to remedy it. However, while that transformation increased the model's R^2 by ~.03 and met the independence assumption, it then failed the normality condition of the QQ plot.

Piecewise Fit

Lastly, I tried to create two models, one for part time employees and one for full time employees, that would predict SALARY. This was because I seemed to see a difference in salary among the full time group and part time group. However, I could not manage a model with a R^2 that was greater than .55 for both. I imagine that the amount of NAs present in both datasets made it such that the predictors conflicted each other.

In [400]:

```
temp<-df
head(temp[temp$SALARY==0 &temp$LFSTAT == 1,c("SALARY","NDGMEMG","AGE","FTPRET","EMSEC")])
table(temp[temp$SALARY==0 &temp$LFSTAT == 1,"EMSEC"])
temp<-temp[!is.na(temp$SALARY) & temp$SALARY != 0,]
temp <- temp[, !(names(temp) %in% c("LOOKWK","LFSTAT","YEAR","NWFAM","NWLAY","NWNOND","NWOCNA","NWO
TP","NWSTU","PERSONID","SAMPLE","SURID","CHTOT"))]
```

A data.frame: 6 × 5

	SALARY	NDGMEMG	AGE	FTPRET	EMSEC
	<dbl>	<int>	<int>	<int>	<int>
35	0	2	70	1	4
247	0	5	42	0	4
307	0	4	59	1	4
375	0	4	55	0	4
546	0	4	53	0	4

583 SALARY 0 NDGMEMG 3 AGE 60 FTPRET 1 EMSEC 4

1	2	3	4
23	50	14	528

In [401]:

```
res<-cor(temp,use="pairwise.complete.obs")

cor_df <- data.frame(res)

sig_cols <- data.frame((abs(cor_df["SALARY"])>.1 & abs(cor_df["SALARY"])<1)[,1])
names(sig_cols) <- c("Significant")
# sig_cols

# The first model
# model <- lm(SALARY ~ AGE + GENDER + MINRTY + RACETH + CHU2IN + DGRDG + HD03Y5 + HRSWKGR + WKSWKG
R + JOBINS + JOBPENS + JOBPROFT + JOBVAC + FTPRET + PTWTFT + PTNOND + PTOCNA + PTOTP + OCEDRLP + E
MSEC + WASCSM + NRREA + JOBSATIS + SATADV + SATBEN + SATCHAL + SATRESP + SATSAL + SATSAL + SATSEC
+ MGRNAT, temp)
# summary(model)
```

```
Warning message in cor(temp, use = "pairwise.complete.obs"):
"the standard deviation is zero"
```

In [402]:

```
predictors <- names(temp)[sig_cols$Significant]
print("Predictors that could be useful:")
predictors


temp$CHU2IN <- as.factor(temp$CHU2IN)
temp$CH25IN <- as.factor(temp$CH25IN)
temp$CH611N <- as.factor(temp$CH611IN)
temp$CH1218IN <- as.factor(temp$CH1218IN)
temp$CH19IN <- as.factor(temp$CH19IN)
# df <- df[sample(nrow(df),.1 * nrow(df)),]

cols <-
c("GENDER","DGRDG","HRSWKGR","WKSWKGR","JOBINS","JOBPROFT","EMSEC","FTPRET","WASCSM","MGRNAT","NOCP
RMG","NDGMEMG","NBAMEMG","SATSAL","NRREA", "HDDGRUS","JOBVAC","RACETH")
temp[,cols] <- lapply(temp[,cols], factor)

temp$HD03Y5 <- as.numeric(temp$HD03Y5)
temp$BA03Y5 <- as.numeric(temp$BA03Y5)
temp$years_since <- (2003) - temp$HD03Y5
# temp$years_since.2 <- (2003) - temp$BA03Y5
temp$years_since <- as.factor(temp$years_since)
# temp$years_since.2 <- as.factor(temp$years_since.2)
```

```
[1] "Predictors that could be useful:"
```

'AGE' 'GENDER' 'MINRTY' 'RACETH' 'CHU2IN' 'DGRDG' 'HD03Y5' 'HRSWKGR' 'WKSWKGR' 'JOBINS' 'JOBPENS' 'JOBPROFT' 'JOBVAC' 'FTPRET' 'PTWTFT' 'PTNOND' 'PTOCNA' 'PTOTP' 'OCEDRLP' 'EMSEC' 'WASCSM' 'NRREA' 'JOBSATIS' 'SATADV' 'SATBEN' 'SATCHAL' 'SATRESP' 'SATSAL' 'SATSEC' 'MGRNAT'

Interpretation and Fitting

In [403]:

```
library(MASS)
fmla <- as.formula(paste("log(SALARY) ~  SATSAL + HDDGRUS + EMSEC + DGRDG + NRREA + CH19IN + years
_since + HRSWKGR + WKSWKGR + NDGMEMG + JOBINS + JOBPROFT + JOBVAC + JOBSATIS + MGRNAT + WASCSM + G
ENDER" ))
model <- lm(fmla,temp)
model.lm <- stepAIC(model)
summary(model.lm)
reg1<-model.lm
```

```
Start:  AIC=-5213.49
log(SALARY) ~ SATSAL + HDDGRUS + EMSEC + DGRDG + NRREA + CH19IN +
    years_since + HRSWKGR + WKSWKGR + NDGMEMG + JOBINS + JOBPROFT +
    JOBVAC + JOBSATIS + MGRNAT + WASCSM + GENDER

              Df Sum of Sq    RSS     AIC
- JOBPROFT     1      0.06 1639.9 -5215.3
<none>                     1639.9 -5213.5
- CH19IN       1      2.13 1642.0 -5209.2
- JOBSATIS     1      2.41 1642.3 -5208.3
- NDGMEMG      6      5.94 1645.8 -5207.9
- HDDGRUS      1      9.59 1649.5 -5187.0
- JOBVAC       1     13.61 1653.5 -5175.2
- EMSEC        3     17.12 1657.0 -5168.9
- GENDER       1     24.83 1664.7 -5142.2
- MGRNAT       1     29.94 1669.8 -5127.3
- WASCSM       5     34.61 1674.5 -5121.6
- NRREA        6     36.65 1676.5 -5117.7
- JOBINS       1     41.47 1681.3 -5093.7
- WKSWKGR      3     56.30 1696.2 -5054.9
- years_since 10     62.21 1702.1 -5051.9
- DGRDG        3     86.56 1726.4 -4968.6
- SATSAL       3    126.39 1766.3 -4857.4
- HRSWKGR      3    341.78 1981.7 -4296.2

Step:  AIC=-5215.32
log(SALARY) ~ SATSAL + HDDGRUS + EMSEC + DGRDG + NRREA + CH19IN +
    years_since + HRSWKGR + WKSWKGR + NDGMEMG + JOBINS + JOBVAC +
    JOBSATIS + MGRNAT + WASCSM + GENDER

              Df Sum of Sq    RSS     AIC
<none>                     1639.9 -5215.3
- CH19IN       1      2.13 1642.1 -5211.0
- JOBSATIS     1      2.37 1642.3 -5210.3
- NDGMEMG      6      5.95 1645.9 -5209.7
- HDDGRUS      1      9.63 1649.6 -5188.7
- JOBVAC       1     13.93 1653.9 -5176.1
- EMSEC        3     17.61 1657.5 -5169.2
- GENDER       1     24.82 1664.8 -5144.1
- MGRNAT       1     30.00 1669.9 -5128.9
- WASCSM       5     34.71 1674.6 -5123.2
- NRREA        6     36.70 1676.6 -5119.4
- JOBINS       1     42.25 1682.2 -5093.3
- WKSWKGR      3     56.27 1696.2 -5056.8
- years_since 10     62.19 1702.1 -5053.8
- DGRDG        3     86.60 1726.5 -4970.3
- SATSAL       3    126.96 1766.9 -4857.6
- HRSWKGR      3    341.75 1981.7 -4298.2


Call:
lm(formula = log(SALARY) ~ SATSAL + HDDGRUS + EMSEC + DGRDG +
    NRREA + CH19IN + years_since + HRSWKGR + WKSWKGR + NDGMEMG +
    JOBINS + JOBVAC + JOBSATIS + MGRNAT + WASCSM + GENDER, data = temp)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8659 -0.2593  0.0117  0.3157  2.4978

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.138158   0.133922  60.768  < 2e-16 ***
SATSAL2       -0.230882   0.021502 -10.738  < 2e-16 ***
SATSAL3       -0.474786   0.029510 -16.089  < 2e-16 ***
SATSAL4       -0.637414   0.037143 -17.161  < 2e-16 ***
HDDGRUS1       0.146359   0.027485   5.325 1.05e-07 ***
EMSEC2         0.091744   0.048988   1.873  0.06116 .
EMSEC3         0.243924   0.042405   5.752 9.35e-09 ***
EMSEC4         0.230178   0.035652   6.456 1.18e-10 ***
DGRDG2         0.132374   0.023638   5.600 2.26e-08 ***
DGRDG3         0.406961   0.025839  15.750  < 2e-16 ***
DGRDG4         0.179922   0.072423   2.484  0.01301 *
NRREA2        -0.162445   0.033892  -4.793 1.69e-06 ***
NRREA3        -0.180580   0.039995  -4.515 6.48e-06 ***
NRREA4        -0.068518   0.025486  -2.689  0.00720 **
NRREA5        -0.269188   0.028953  -9.297  < 2e-16 ***
```

```
NRREA6        -0.177190   0.027099  -6.539 6.85e-11 ***
NRREA7        -0.194843   0.038638  -5.043 4.76e-07 ***
CH19IN1       -0.060244   0.024069  -2.503  0.01235 *
years_since-3  0.024410   0.055496   0.440  0.66006
years_since2   0.236848   0.056781   4.171 3.08e-05 ***
years_since7   0.271889   0.056792   4.787 1.74e-06 ***
years_since12  0.310146   0.056757   5.464 4.88e-08 ***
years_since17  0.356867   0.058554   6.095 1.18e-09 ***
years_since22  0.375077   0.060827   6.166 7.56e-10 ***
years_since27  0.398172   0.064501   6.173 7.24e-10 ***
years_since32  0.292725   0.072944   4.013 6.09e-05 ***
years_since37  0.128117   0.102523   1.250  0.21149
years_since42  0.316040   0.194133   1.628  0.10360
HRSWKGR2       0.686406   0.035388  19.396  < 2e-16 ***
HRSWKGR3       0.952213   0.033933  28.062  < 2e-16 ***
HRSWKGR4       1.104111   0.035070  31.483  < 2e-16 ***
WKSWKGR2       0.323891   0.130347   2.485  0.01299 *
WKSWKGR3       0.602523   0.112221   5.369 8.28e-08 ***
WKSWKGR4       0.933071   0.102839   9.073  < 2e-16 ***
NDGMEMG2      -0.035806   0.038956  -0.919  0.35806
NDGMEMG3      -0.085405   0.044981  -1.899  0.05767 .
NDGMEMG4      -0.009212   0.035388  -0.260  0.79463
NDGMEMG5       0.031741   0.038951   0.815  0.41516
NDGMEMG6      -0.084300   0.041447  -2.034  0.04201 *
NDGMEMG7      -0.010821   0.039066  -0.277  0.78179
JOBINS1        0.321555   0.028834  11.152  < 2e-16 ***
JOBVAC1        0.188348   0.029419   6.402 1.68e-10 ***
JOBSATIS       0.033322   0.012611   2.642  0.00826 **
MGRNAT1        0.200070   0.021292   9.396  < 2e-16 ***
WASCSM2       -0.013992   0.049777  -0.281  0.77865
WASCSM3        0.039740   0.026557   1.496  0.13461
WASCSM4        0.021316   0.046950   0.454  0.64984
WASCSM5       -0.056247   0.035724  -1.574  0.11544
WASCSM6       -0.173773   0.028920  -6.009 2.01e-09 ***
GENDER2        0.162287   0.018987   8.547  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5829 on 4827 degrees of freedom
  (92559 observations deleted due to missingness)
Multiple R-squared:  0.6471, Adjusted R-squared:  0.6435
F-statistic: 180.6 on 49 and 4827 DF,  p-value: < 2.2e-16
```

$\widehat{SALARY} = 8.138158 + \text{-}0.230882 * SATSAL2 + \text{-}0.474786 * SATSAL3 + \text{-}0.637414 * SATSAL4 + 0.146359 * HDDGRUS1 +0.091744 * EMSEC2 + 0.243924 * EMSEC3 + 0.230178 * EMSEC4 + 0.132374 * DGRDG2 + 0.406961 * DGRDG3 + 0.179922 * DGRDG4 + \text{-}0.162445 * NRREA2 + \text{-}0.180580 * NRREA3 + \text{-}0.068518 * NRREA4 + \text{-}0.269188* NRREA5 + \text{-}0.177190 * NRREA6 + \text{-}0.194843 * NRREA7 + \text{-}0.060244 * CH19IN1 + 0.024410*years_since\text{-}3 + 0.236848*years_since2 + 0.271889*years_since7 + 0.310146*years_since12 + 0.356867*years_since17 + 0.375077*years_since22 + 0.398172*years_since27 + 0.292725*year_since32 + 0.128117*years_since37 + 0.316040*years_since42 + 0.686406 * HRSWKGR2 + 0.602523 * HRSWKGR3 + 1.104111 * HRSWKGR4 + 0.323891 * WKSWKGR2 + 0.602523 * WKSWKGR3 + 0.933071 * WKSWKGR4 + \text{-}0.035806* NDGMEMG2 + \text{-}0.085405 * NDGMEMG3 + \text{-}0.009212 * NDGMEMG4 + 0.031741 * NDGMEMG5 + \text{-}0.084300 * NDGMEMG6 + \text{-}0.010821 * NDGMEMG7 + 0.321555 * JOBINS1 + 0.188348 * JOBVAC1 + 0.033322 * JOBSATIS + 0.200070 * MGRNAT1 + \text{-}0.013992 * WASCSM2 + 0.039740 * WASCSM3 + 0.021316 * WASCSM4 + \text{-}0.056247 * WASCSM5 + \text{-}0.173773 * WASCSM6 + 0.162287*GENDER2 $

Interpretations:

For an index of 2 in SATSAL, there is a $\text{-}0.23088$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 3 in SATSAL, there is a $\text{-}0.474786$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 4 in SATSAL, there is a $\text{-}0.637414$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 1 in HDDGRUS, there is a 0.146359 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 2 in EMSEC, there is a 0.091744 dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 3 in EMSEC, there is a 0.243924 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 4 in EMSEC, there is a 0.230178 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 2 in DGRDG, there is a 0.132374 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 3 in DGRDG, there is a 0.406961 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 4 in DGRDG, there is a 0.179922 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 2 in NRREA, there is a $\text{-}0.162445$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 3 in NRREA, there is a $\text{-}0.180580$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 4 in NRREA, there is a $\text{-}0.068518$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 5 in NRREA, there is a $\text{-}0.269188$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 6 in NRREA, there is a $\text{-}0.177190$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 7 in NRREA, there is a $\text{-}0.194843$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 1 in CH19IN, there is a $\text{-}0.060244$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of -3 in years_since, there is a 0.024410 dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 2 in years_since, there is a 0.236848 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 7 in years_since, there is a 0.271889 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 12 in years_since, there is a 0.310146 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 17 in years_since, there is a 0.356867 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 22 in years_since, there is a 0.375077 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 27 in years_since, there is a 0.398172 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 32 in years_since, there is a 0.292725 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 37 in years_since, there is a 0.024410 dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 42 in years_since, there is a 0.316040 dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 2 in HRSWKGR, there is a 0.686406 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 3 in HRSWKGR, there is a 0.602523 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 4 in HRSWKGR, there is a 1.104111 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 2 in WKSWKGR, there is a 0.323891 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 3 in WKSWKGR, there is a 0.602523 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 4 in WKSWKGR, there is a 0.933071 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 2 in NDMGMEMG, there is a $\text{-}0.035806$ dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 3 in NDMGMEMG, there is a $\text{-}0.085405$ dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 4 in NDMGMEMG, there is a $\text{-}0.009212$ dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 5 in NDMGMEMG, there is a 0.031741 dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 6 in NDMGMEMG, there is a $\text{-}0.084300$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 7 in NDMGMEMG, there is a $\text{-}0.010821$ dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 1 in JOBINS, there is a 0.321555 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 1 in JOBVAC, there is a 0.188348 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For every one unit increase in JOBSATIS, there is a 0.033322 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 1 in MGRNAT, there is a 0.200070 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 2 in WASCSM, there is a $\text{-}0.013992$ dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 3 in WASCSM, there is a 0.039740 dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 4 in WASCSM, there is a 0.021316 dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 5 in WASCSM, there is a $\text{-}0.056247$ dollar increase in predicted log Salary. The pvalue is greater than .05, so I fail to reject H_O that the coefficient is 0.

For an index of 6 in WASCSM, there is a $\text{-}0.173773$ dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 2 in GENDER, there is a 0.162287 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

For an index of 1 in SATSAL, 0 in HDDGRUS, 1 in EMSEC,1 in DGRDG, 1 in NRREA, 0 in CH19IN, an index below -3 in years_since, 1 in WKSWKGR, 1 in HRSWKGR, 1 in NDGMEMG, 0 in JOBINS, 0 in JOBVAC, 0 in MGRNAT, 1 in WASCSM, or 1 in GENDER, there is a 8.138158 dollar increase in predicted log Salary. The pvalue is less than .05, so I reject H_O that the coefficient is 0, and there is sufficient evidence to believe that the coefficient is non-zero.

R^2 Interpretation, Diagnostic Plots, and Comments on Interpretability

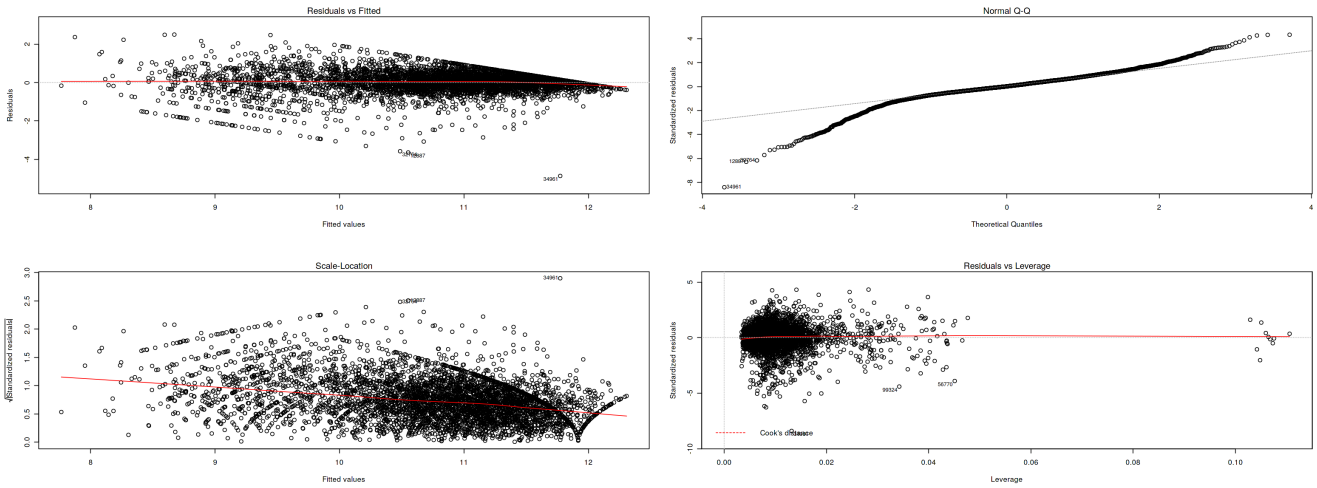
Since the model has 26 coefficients, the interpretability of the model is not as bad as my second attempt but is hard to visualize nevertheless.

The R^2 is 0.6471. That is 64.71% of the variation in the data can be explained by the model. This is $\frac{s_{response}^2-s_{residuals}^2}{s_{response}^2}*100$. The adjusted R^2 is 0.6435. R_{adj} is similar in interpretation but is a little smaller than

s_{residuals}^2}{s_{response}^2}*100.$ The adjusted R^2 is 0.6435. R_{adj} is similar in interpretation but is a little smaller than R^2 because it is calculated with a penalty for using more predictors. The R^2 value indicates the model is moderately good fit for the data. However, it could be the case that salary cannot be explained by linear regression, and more continuous,numerical data was needed. This model provides an acceptable balance between complexity and accuracy.

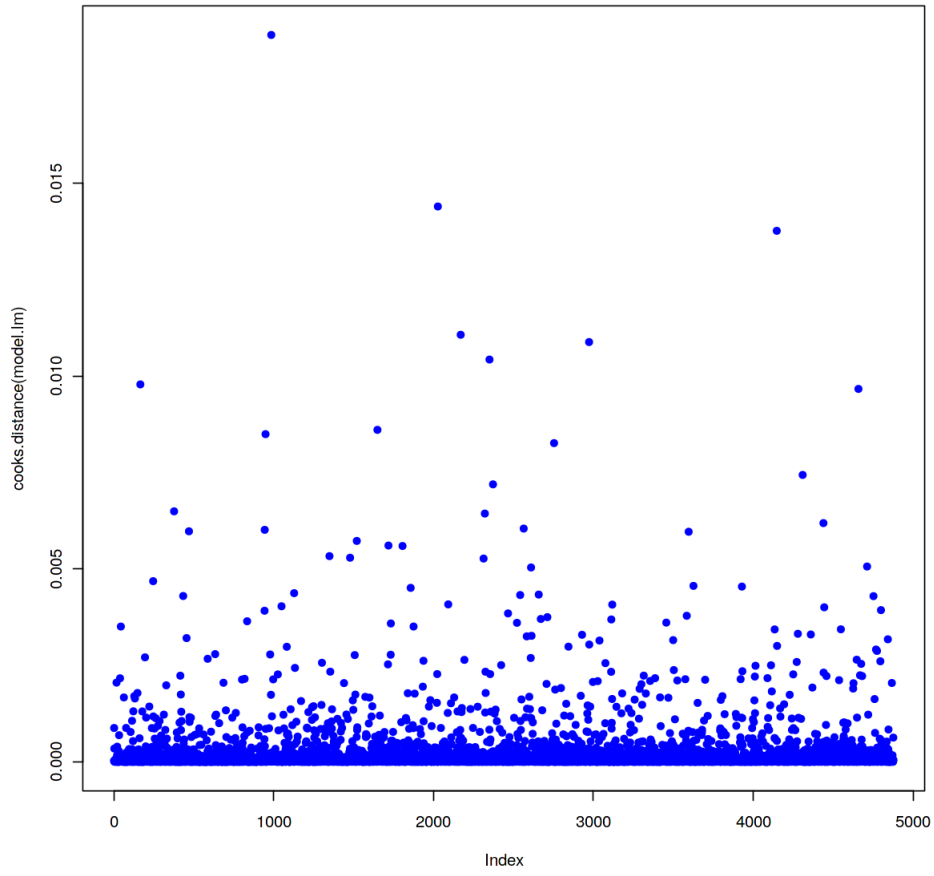
In [378]:

```
par(mfrow=c(2,2))
saved <- options(repr.plot.width=26, repr.plot.height=10)
plot(model.lm)
options(saved)
```



In [379]:

```
saved <- options(repr.plot.width=10, repr.plot.height=10)
plot(cooks.distance(model.lm), pch = 16, col = "blue") #Plot the Cooks Distances.
options(saved)
```



The model's assumption should be checked by the following:

Independence should be checked using a residual vs fitted plot and expect to see no pattern. This condition is met.

Normality should be checked using a Normal Proabability Plot and expect to see most of the points on the line. This condition is not met.

Constant Variance should be checked using a scaled location plot and expect to see a horizonatal(constant) variance. This condition is met.

From the plot of Cook's Distance, there does not seems to be any influential points that warrant removal from the dataset.

Which career path would you choose base on your model to maximize SALARY?

The categories under your control relevant when starting a career to maximize salary using the above model are NDGMEMG:EMSEC:HDDGRUS:WASCSM:JOBVAC:JOBINS:MGRNAT:DGRDG. If we look at the most common combinations(which implies the most realistic situations) of these variables for the top 20% of earners predicted by the model, this would be an indication of what could earn the most in a career.

In [380]:

```
#use only the highest 20% of SALARIES predicted by the model
top_20_percent <- length(model.lm$fitted.values)*.2
# get the fitted values along with the index of the observation it is predicting for
fitted <- data.frame(sort(model.lm$fitted.values, decreasing = TRUE)[1:top_20_percent])
names(fitted)<-c("fitted")
# get the observations that correspond to their fitted values
fitted_obs<-temp[row.names(fitted),]
fitted_obs$pred.SALARY <- exp(fitted(model.lm))[1:top_20_percent]
```

In [381]:

```
library(dplyr)
combos <- fitted_obs %>% count(NDGMEMG:EMSEC:HDDGRUS:WASCSM:JOBVAC:JOBINS:MGRNAT:DGRDG, sort = TRUE
)
combo.df <- data.frame(combos)
head(combo.df)

median(fitted_obs[fitted_obs$NDGMEMG =="3" & fitted_obs$EMSEC =="4" & fitted_obs$HDDGRUS =="1" & fi
tted_obs$WASCSM =="3" & fitted_obs$JOBVAC =="1" & fitted_obs$JOBINS =="1"  & fitted_obs$MGRNAT =="1
" & fitted_obs$DGRDG =="3","pred.SALARY"])
median(fitted_obs[fitted_obs$NDGMEMG =="5" & fitted_obs$EMSEC =="4" & fitted_obs$HDDGRUS =="1" & fi
tted_obs$WASCSM =="1" & fitted_obs$JOBVAC =="1" & fitted_obs$JOBINS =="1"  & fitted_obs$MGRNAT =="1
" & fitted_obs$DGRDG =="3","pred.SALARY"])
median(fitted_obs[fitted_obs$NDGMEMG =="5" & fitted_obs$EMSEC =="4" & fitted_obs$HDDGRUS =="1" & fi
tted_obs$WASCSM =="3" & fitted_obs$JOBVAC =="1" & fitted_obs$JOBINS =="1"  & fitted_obs$MGRNAT =="1
" & fitted_obs$DGRDG =="3","pred.SALARY"])
median(fitted_obs[fitted_obs$NDGMEMG =="2" & fitted_obs$EMSEC =="4" & fitted_obs$HDDGRUS =="1" & fi
tted_obs$WASCSM =="3" & fitted_obs$JOBVAC =="1" & fitted_obs$JOBINS =="1"  & fitted_obs$MGRNAT =="1
" & fitted_obs$DGRDG =="3","pred.SALARY"])
median(fitted_obs[fitted_obs$NDGMEMG =="4" & fitted_obs$EMSEC =="4" & fitted_obs$HDDGRUS =="1" & fi
tted_obs$WASCSM =="3" & fitted_obs$JOBVAC =="1" & fitted_obs$JOBINS =="1"  & fitted_obs$MGRNAT =="0
" & fitted_obs$DGRDG =="1","pred.SALARY"])
```

A data.frame: 6 × 2

	NDGMEMG.EMSEC.HDDGRUS.WASCSM.JOBVAC.JOBINS.MGRNAT.DGRDG	n
	<fct>	<int>
1	3:4:1:3:1:1:1:3	45
2	5:4:1:1:1:1:1:3	40
3	5:4:1:3:1:1:1:3	40
4	2:4:1:3:1:1:1:3	30
5	4:4:1:3:1:1:0:1	28
6	3:4:1:1:1:1:1:3	27

96223.0431235431

92304.60374766

103199.68383441

111932.900777806

85683.5549663911

Each career plan below is in the order of combinations listed above with their median pay.

1. \$96223.04 - Physcial and Related Sciences, Government, School awarding highest degree located in the US,Management and Administration,benefits available, health insurance available,Job required technical expertise: natural sciences, Doctorate Earned
2. \$92304.60 - Engineering, Government, School awarding highest degree located in the US,Research and Development,benefits available,health insurance available,Job required technical expertise: natural sciences,Doctorate Earned
3. \$103199.68 - Engineering, Government, School awarding highest degree located in the US,Management and Administration,benefits available,health insurance available,Job required technical expertise: natural sciences,Doctorate Earned
4. \$111932.90 - Biological, agricultural and environmental life sciences, Government, School awarding highest degree located in the US,Management and Administration,benefits available,health insurance available,Job required technical expertise: natural sciences,Doctorate Earned
5. \$85683.55 - Social and Related Sciences, Government,School awarding highest degree located in the US,Management and Administration,benefits available,health insurance available,Job does not require technical expertise: natural sciences,Doctorate Earned

The groups here are by no means the groups with the highest median, but investigating salary groups with 1 or 2 people who have a certain combination of factors that give them a higher salary is not useful in determining a feasible plan to maximize SALARY.

Alternatively, if we solely look at the coefficients, one would only want the largest and positive coefficients. Listing the factor corresponding to category in the order above that one would hypthetical fufill, this would be Engineering, Government, School awarding highest degree located in the US,Management and Administration, Available benefits: paid vacation/sick/personal days, Available benefits: health insurance,Job required technical expertise: natural sciences, and Doctorate.

Regression 2: job satisfaction vs other variables

Dropping Observations

First, I dropped the numerics that were not relevant to SALARY such as PERSONID, YEAR, SAMPLE, SURID, and CHTOT. In addition, Those that are unemployed and those not in the labor force and looking for work are dropped the dataset. It is not reasonable to predict job satisfaction for those with no jobs. Unlike Regression 1, I keep those that are employed and have a salary of 0 because they do have a JOBSATIS score.

While most people are satisfied or somewhat satisified (~99%) with their job according to the bar chart constructed in Question 6, it does not make sense to drop the remaining observations who are not satisfied or the model will have nothing to predict.

Variable Selection

Variable selection was determined by selecting all varaibles related to satisfaction. This achieved a baseline AUROC of ~0.93. Selecting other predictors like JOBxxx, NOCRPMG, NDGMEMG, NBAMEMG only served to increase the AUROC by ~.0002-.0005. However, removing even one of the satisfaction variables drops the AUROC consistently by .01 or more. Interaction terms increased complexity and did not increase AUROC. Accuracy ranges from (.88 to .92) depending on the cutoff chosen to round the probability fitted.

Determining Useful Fit

The model, while achieving an upper bound accuracy of .92 with a cutoff of .05 on the test data and .92 with a cutoff of .5 on the training data, this model performs worse than a far simpler model, simply outputting 1 to all training data, will achieve ~99% accuracy. This suggests the model, while initially impressive, is not useful.

Splitting Dataset

Since JOBSATIS in the dataset has been recoded such that 99% of the values are in the "satisfied" category, the accuracy of the model can be better tested when part of the data is used for the fitting(training) of the model, while the remaining data is used to analyze accuracy.

In [382]:

```
temp<-df
# employees who are employed but have no salary should be included
temp<-temp[!is.na(temp$SALARY) & !is.na(temp$JOBSATIS),]
temp <- temp[, !(names(temp) %in% c("LOOKWK","LFSTAT","YEAR","NWFAM","NWLAY","NWNOND","NWOCNA","NWO
TP","NWSTU","PERSONID","SAMPLE","SURID","CHTOT"))]
cols <-
c("JOBSATIS","GENDER","DGRDG","HRSWKGR","WKSWKGR","JOBVAC","JOBINS","JOBPROFT","EMSEC","FTPRET","WA
SCSM","MGRNAT","NOCPRMG","NDGMEMG","NBAMEMG","NRREA", "HDDGRUS","SATSAL","SATADV","SATBEN",
"SATCHAL","SATIND","SATLOC","SATRESP","SATSEC","SATSOC")
temp[,cols] <- lapply(temp[,cols], factor)
levels(temp$JOBSATIS)<-list("1"=c("1","2"),"0"=c("3","4"))
temp$JOBSATIS <- as.numeric(as.character(temp$JOBSATIS))
```

In [383]:

```
temp$JOBSATIS <- as.factor(temp$JOBSATIS)
table(temp$JOBSATIS)
```

```
    0     1
10396 87655
```

Split data into training and test data set

In [384]:

```
set.seed(101)
sample <- sample.int(n = nrow(temp), size = floor(.9*nrow(temp)), replace = F)
train <- temp[sample, ]
test  <- temp[-sample, ]
```

Fitting

In [385]:

```
# fit the Logistic Regression model
fmla <- as.formula(paste("JOBSATIS ~ SATSAL + SATADV + SATBEN +
SATCHAL+SATIND+SATLOC+SATRESP+SATSEC+SATSOC"))
model.lm <- glm(fmla,family="binomial",train)
summary(model.lm)
```

```
Call:
glm(formula = fmla, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3388   0.1042   0.1604   0.2876   3.2326

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.44339    0.07564  71.964  < 2e-16 ***
SATSAL2      0.03437    0.05117   0.672 0.501717
SATSAL3     -1.08814    0.05393 -20.178  < 2e-16 ***
SATSAL4     -1.93257    0.06030 -32.051  < 2e-16 ***
SATADV2     -0.32310    0.07532  -4.290 1.79e-05 ***
SATADV3     -1.06195    0.07511 -14.139  < 2e-16 ***
SATADV4     -1.73785    0.07750 -22.425  < 2e-16 ***
SATBEN2      0.09215    0.04162   2.214 0.026843 *
SATBEN3     -0.13990    0.04988  -2.805 0.005035 **
SATBEN4     -0.14294    0.05587  -2.559 0.010508 *
SATCHAL2    -0.17447    0.04756  -3.668 0.000244 ***
SATCHAL3    -0.89794    0.05263 -17.062  < 2e-16 ***
SATCHAL4    -1.35744    0.06708 -20.235  < 2e-16 ***
SATIND2     -0.44533    0.03710 -12.002  < 2e-16 ***
SATIND3     -1.14237    0.04788 -23.859  < 2e-16 ***
SATIND4     -1.35824    0.07615 -17.837  < 2e-16 ***
SATLOC2     -0.11140    0.03447  -3.231 0.001232 **
SATLOC3     -0.43588    0.04442  -9.812  < 2e-16 ***
SATLOC4     -0.73738    0.06331 -11.647  < 2e-16 ***
SATRESP2    -0.28109    0.04890  -5.748 9.05e-09 ***
```

```
SATRESP3    -0.93182    0.05618 -16.587  < 2e-16 ***
SATRESP4    -0.94884    0.08520 -11.136  < 2e-16 ***
SATSEC2     -0.04054    0.04031  -1.006 0.314549
SATSEC3     -0.62134    0.04661 -13.330  < 2e-16 ***
SATSEC4     -1.01061    0.05399 -18.718  < 2e-16 ***
SATSOC2     -0.40347    0.03870 -10.424  < 2e-16 ***
SATSOC3     -1.16241    0.04693 -24.770  < 2e-16 ***
SATSOC4     -1.43692    0.06467 -22.219  < 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: 59696  on 88244  degrees of freedom
Residual deviance: 32537  on 88217  degrees of freedom
AIC: 32593

Number of Fisher Scoring iterations: 7
```

Equation

In [386]:

```
eqn <- paste("$log(JOBSATIS odds) =", paste(round(model.lm$coefficients[1],2), paste(round(model.lm
$coefficients[-1],2), names(model.lm$coefficients[-1]), sep=" * ", collapse=" + "), sep=" + "), "$"
)
eqn
exp(model.lm$coefficients)
```

'$log(JOBSATIS odds) = 5.44 + 0.03 * SATSAL2 + -1.09 * SATSAL3 + -1.93 * SATSAL4 + -0.32 * SATADV2 + -1.06 * SATADV3 + -1.74 * SATADV4 + 0.09 * SATBEN2 + -0.14 * SATBEN3 + -0.14 * SATBEN4 + -0.17 * SATCHAL2 + -0.9 * SATCHAL3 + -1.36 * SATCHAL4 + -0.45 * SATIND2 + -1.14 * SATIND3 + -1.36 * SATIND4 + -0.11 * SATLOC2 + -0.44 * SATLOC3 + -0.74 * SATLOC4 + -0.28 * SATRESP2 + -0.93 * SATRESP3 + -0.95 * SATRESP4 + -0.04 * SATSEC2 + -0.62 * SATSEC3 + -1.01 * SATSEC4 + -0.4 * SATSOC2 + -1.16 * SATSOC3 + -1.44 * SATSOC4 $'

(Intercept)
231.224557770999
SATSAL2
1.03497193004585
SATSAL3
0.336843005630395
SATSAL4
0.144775335407839
SATADV2
0.723898110352018
SATADV3
0.345779791379875
SATADV4
0.175897667845828
SATBEN2
1.09652600193359
SATBEN3
0.869444242684256
SATBEN4
0.866805521611566
SATCHAL2
0.839904766173105
SATCHAL3
0.4074072535743
SATCHAL4
0.2573196346917
SATIND2
0.640613663874782
SATIND3
0.319062506823197
SATIND4
0.257112326947647

SATLOC2
0.89458161565542
SATLOC3
0.646694645073799
SATLOC4
0.478365971685723
SATRESP2
0.754962824718348
SATRESP3
0.393835463893604
SATRESP4
0.387188266700417
SATSEC2
0.960272305413907
SATSEC3
0.537221647024471
SATSEC4
0.363997718847742
SATSOC2
0.668000235883333
SATSOC3
0.312730270094951
SATSOC4
0.237658389816598

Interpretation of Coefficients and p-values

The baseline to which we compare coefficients as more likely than is someone who has SATSAL1,SATADV1, SATBEN1,SATCHAL1, SATIND1, SATLOC1, SATRESP1, SATSEC1, and SATSOC1.

β_0 is the intercept, β_1 is SATSAL2 coefficient, and so on.

e^{β_0} is the odds ratio for a person with SATSAL1 for fixed SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are 231.22 times more likely to be satisified

$e^{\beta_{1}}$ is the odds ratio for a person with SATSAL2 versus one with SATSAL1 for fixed SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are 1.03 times more likely to be satisfied with their job than someone with SATSAL1 and fixed remaining variables.

e^{β_2} is the odds ratio for a person with SATSAL3 versus one with SATSAL1 for fixed SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .34 times more likely to be satisfied with their job than someone with SATSAL1 and fixed remaining variables.

e^{β_3} is the odds ratio for a person with SATSAL3 versus one with SATSAL1 for fixed SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .14 times more likely to be satisfied with their job than someone with SATSAL1 and fixed remaining variables.

e^{β_4} is the odds ratio for a person with SATADV2 versus one with SATADV1 for fixed SATSAL, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .72 times more likely to be satisfied with their job than someone with SATADV1 and fixed remaining variables.

e^{β_5} is the odds ratio for a person with SATADV3 versus one with SATADV1 for fixed SATSAL, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .35 times more likely to be satisfied with their job than someone with SATADV1 and fixed remaining variables.

e^{β_6} is the odds ratio for a person with SATADV4 versus one with SATADV1 for fixed SATSAL, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .18 times more likely to be satisfied with their job than someone with SATADV1 and fixed remaining variables.

e^{β_7} is the odds ratio for a person with SATBEN2 versus one with SATBEN1 for fixed SATSAL, SATADV, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are 1.10 times more likely to be satisfied with their job than someone with SATBEN1 and fixed remaining variables.

e^{β_8} is the odds ratio for a person with SATBEN3 versus one with SATBEN1 for fixed SATSAL, SATADV, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .87 times more likely to be satisfied with their job than someone with SATBEN1 and fixed remaining variables.

e^{β_9} is the odds ratio for a person with SATBEN4 versus one with SATBEN1 for fixed SATSAL, SATADV, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .87 times more likely to be satisfied with their job than someone with SATBEN1 and fixed remaining variables.

$e^{\beta_{10}}$ is the odds ratio for a person with SATCHAL2 versus one with SATCHAL1 for fixed SATSAL, SATADV, SATBEN, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .84 times more likely to be satisfied with their job than someone with SATCHAL1 and fixed remaining variables.

$e^{\beta_{11}}$ is the odds ratio for a person with SATCHAL3 versus one with SATCHAL1 for fixed SATSAL, SATADV, SATBEN, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .41 times more likely to be satisfied with their job than someone with SATCHAL1 and fixed remaining variables.

$e^{\beta_{12}}$ is the odds ratio for a person with SATCHAL4 versus one with SATCHAL1 for fixed SATSAL, SATADV, SATBEN, SATIND, SATLOC, SATRESP, SATSEC, SATSOC. They are .26 times more likely to be satisfied with their job than someone with SATCHAL1 and fixed remaining variables.

$e^{\beta_{13}}$ is the odds ratio for a person with SATIND2 versus one with SATIND1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATLOC, SATRESP, SATSEC, SATSOC. They are .64 times more likely to be satisfied with their job than someone with SATIND1 and fixed remaining variables.

$e^{\beta_{14}}$ is the odds ratio for a person with SATLOC2 versus one with SATLOC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATRESP, SATSEC, SATSOC. They are .32 times more likely to be satisfied with their job than someone with SATLOC1 and fixed remaining variables.

$e^{\beta_{15}}$ is the odds ratio for a person with SATLOC3 versus one with SATLOC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATRESP, SATSEC, SATSOC. They are .26 times more likely to be satisfied with their job than someone with SATLOC1 and fixed remaining variables.

$e^{\beta_{16}}$ is the odds ratio for a person with SATLOC4 versus one with SATLOC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATRESP, SATSEC, SATSOC. They are .48 times more likely to be satisfied with their job than someone with SATLOC1 and fixed remaining variables.

$e^{\beta_{17}}$ is the odds ratio for a person with SATRESP2 versus one with SATRESP1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATSEC, SATSOC. They are .75 times more likely to be satisfied with their job than someone with SATRESP1 and fixed remaining variables.

$e^{\beta_{18}}$ is the odds ratio for a person with SATRESP3 versus one with SATRESP1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATSEC, SATSOC. They are .39 times more likely to be satisfied with their job than someone with SATRESP1 and fixed remaining variables.

$e^{\beta_{19}}$ is the odds ratio for a person with SATRESP4 versus one with SATRESP1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATSEC, SATSOC. They are .39 times more likely to be satisfied with their job than someone with SATRESP1 and fixed remaining variables.

$e^{\beta_{20}}$ is the odds ratio for a person with SATSEC2 versus one with SATSEC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSOC. They are .96 times more likely to be satisfied with their job than someone with SATSEC1 and fixed remaining variables.

$e^{\beta_{21}}$ is the odds ratio for a person with SATSEC3 versus one with SATSEC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSOC. They are .54 times more likely to be satisfied with their job than someone with SATSEC1 and fixed remaining variables.

$e^{\beta_{22}}$ is the odds ratio for a person with SATSEC4 versus one with SATSEC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSOC. They are .36 times more likely to be satisfied with their job than someone with SATSEC1 and fixed remaining variables.

$e^{\beta_{23}}$ is the odds ratio for a person with SATSOC2 versus one with SATSOC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC. They are .67 times more likely to be satisfied with their job than someone with SATSOC1 and fixed remaining variables.

$e^{\beta_{24}}$ is the odds ratio for a person with SATSOC3 versus one with SATSOC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC. They are .31 times more likely to be satisfied with their job than someone with SATSOC1 and fixed remaining variables.

$e^{\beta_{25}}$ is the odds ratio for a person with SATSOC4 versus one with SATSOC1 for fixed SATSAL, SATADV, SATBEN, SATCHAL, SATIND, SATLOC, SATRESP, SATSEC. They are .23 times more likely to be satisfied with their job than someone with SATSOC1 and fixed remaining variables.

Since the pvalues for every coefficient,except SATSEC2, are less than $\alpha=.05$, we reject H_O, and there is sufficient evidence to beleive these variables are relevant. SATSEC2 versus SATSEC1 must not have a significant impact on JOBSATIS

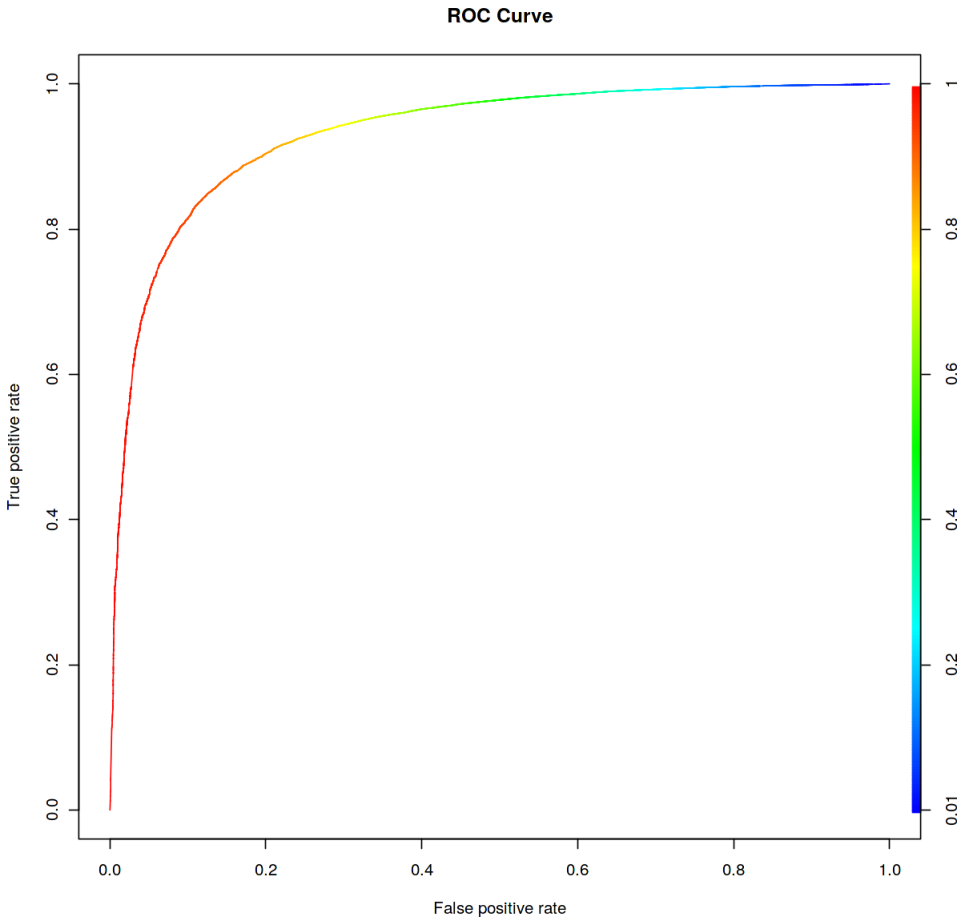
evidence to believe these variables are relevant. GATECL2 versus GATECL1 must not have a significant impact on JOBSATIS.

ROC curve and AUROC

In [387]:

```
library(ROCR)
pred <- prediction(model.lm$fitted.values, train$JOBSATIS)
perf <- performance(pred,"tpr","fpr")
AUROC <- performance(pred, measure = "auc")@y.values[[1]]
print(paste("AUROC:",AUROC))

saved <- options(repr.plot.width=10, repr.plot.height=10)
plot(perf,colorize=TRUE,main="ROC Curve")
options(saved)
```

```
[1] "AUROC: 0.932467611650645"
```



Determine training set accuracy and test set accuracy over a range of cutoffs.

In [388]:

```
get.Accuracy<-function(cutoff){
    model.probs <- model.lm$fitted.values
    model.pred = model.probs >= cutoff
    #create confusion matrix
    table=table(train$JOBSATIS,model.pred)
    accuracy = (table[2,2]+table[1,1])/sum(table)
    return (accuracy)
}
i<-.05
accuracy <- c()
while(i<=.9){
    accuracy <- append(accuracy,get.Accuracy(i))
    i<-i+.05
}
accuracy
```

```
accuracy
```

0.901694146977166 0.90724686951102 0.911757040058927 0.915281319054904 0.918261657884299
0.921241996713695 0.923236443991161 0.9245509660604 0.926307439514987 0.927168678112074 0.92739531984815
0.926851379681568 0.925355544223469 0.92254518669613 0.915847923395093 0.906895574820103
0.888401609156326

In [389]:

```
pred<-data.frame(predict(model.lm, newdata = test, type = "response"))
get.Accuracy<-function(cutoff){
    names(pred)=c("fit")
    model.pred =pred$fit >= cutoff
    # create confusion matrix
    table=table(test$JOBSATIS,model.pred)
    accuracy = (table[2,2]+table[1,1])/sum(table)
    return(accuracy)
}
i<-.05
accuracy <- c()
while(i<=.9){
    accuracy <- append(accuracy,get.Accuracy(i))
    i<-i+.05
}
sort(accuracy,decreasing = TRUE)
```

0.927187436263512 0.926473587599429 0.926065674077096 0.925759738935346 0.925147868651846
0.922088517234346 0.92188456047318 0.919946971242097 0.918723230675097 0.91729553334693
0.914440138690598 0.911584744034265 0.910666938609015 0.905975933102182 0.901794819498266
0.901488884356516 0.881501121762186

Which career path would you choose based on your model to maximize JOBSATIS?

If we observe from the coefficients above which combination of satisfaction variables that would result in the maximum odds ratio of JOBSATIS above 1, this would allow one to pick the remaining variables that determine career path. Some factors to look for when starting a career are NDGMEMG:DGRDG:EMSEC:HDDGRUS:JOBINS.

In [390]:

```
# get the fitted values along with the index of the observation it is predicting for
fitted <- data.frame(sort(model.lm$fitted.values, decreasing = TRUE))
names(fitted)<-c("fitted")
# get the observations that correspond to their fitted values
fitted_obs<-temp[row.names(fitted),]
fitted_obs$pred.JOBSATIS <- fitted(model.lm)>.6
```

Now that we have obtained the observations associated with the prediction, we can observe the most common combinations of factors that are relevant to starting a career out of the subset formed by our coefficients as well as a JOBSATIS of 1.

In [391]:

```
combos <- fitted_obs[fitted_obs$SATADV=="1" & fitted_obs$SATBEN=="1" & fitted_obs$SATCHAL=="1" & fi
tted_obs$SATIND=="1" & fitted_obs$SATLOC=="1" & fitted_obs$SATRESP=="1"
                     & fitted_obs$SATSAL=="1"& fitted_obs$SATSEC=="1" & fitted_obs$SATSOC=="1" & fi
ted_obs$JOBSATIS=="1",] %>% count(NOCPRMG:NDGMEMG:DGRDG:EMSEC:HDDGRUS:JOBINS, sort = TRUE)
combo.df <- data.frame(combos)
head(combo.df)
```

A data.frame: 6 × 2

	NOCPRMG.NDGMEMG.DGRDG.EMSEC.HDDGRUS.JOBINS	n
	<fct>	<int>
1	5:5:1:4:1:1	226
2	7:4:1:4:1:1	162
3	6:6:4:4:1:1	143

4	**NOCPRMG.NDGMEMG.DGRDG.EMSEC.HDDGRUS.JOBINS**	12n
5	4:4:3:2:1:1	123
6	5:5:2:4:1:1	114

The top 5 most common/attaintable career plans to maximize satisifaction,listed in the combination formed above and based on the chosen factor combinations, are as follows

1. Engineering, engineering, highest degree is bachelor degree, buisness or industry employer sector, highest degree awarded in US ,job has health insurance
2. Non-science and engineering, Social and related sciences, highest degree is bachelor degree, buisness or industry employer sector, highest degree awarded in US ,job has health insurance
3. Science and engineering-related, Science and engineering-related, highest degree is professional degree, buisness or industry employer sector, highest degree awarded in US ,job has health insurance
4. Non-science and engineering, Non-science and engineering, master degree, buisness or industry employer sector, highest degree awarded in US ,job has health insurance
5. Social and related sciences, Social and related sciences, doctorate degree, 4 year college or medical institution, highest degree awarded in US ,job has health insurance

However, these recommendations are not to say that these groups have the highest rates of JOBSATIS within their own factor combination because we excluded those that were unsatisfied in their combination pool. Instead, we can say that they occur the most frequently within the pool of those who are satisfied with their job, and thus, the best factors one could optimize to maximize JOBSATIS.

Fact-check news outlets

Gallup: Does Higher Learning = Higher Job Satisfaction?

- Claim #1: Education level has very little to do with job satisfaction, or satisfaction with income and time flexibility.

  The analysis of the claim will ignore the analysis of high school graduates with job satisfaction in the article because the surveys used in our dataset imply the participant has at least a bachelors(NSCG+SDR). There are three parts to this claim.

  i) Education level has very little to do with job satisfaction

  Since my recommendation to maximize job satisfaction involves 4/7 job types as well as 3/4 levels of education in the survey, it does not seem like any education level alone implies satisfaction or dissatisfaction. Just as the article states "educational achievement...seems to have very little to do with overall job satisfaction", the data provided in the NSCG and SDR surveys shows that each DGRDG distribution with respect to JOBSATIS reveals similar proportions of satisfaction in each each DGRDG category(Graph 1) and as a result, implies the same conclusion as the article.For example, one could include DGRDG in the logistic regression model, and while some levels are significant, most are not and do not improve AUROC or accuracy. However, while the claim itself is valid, it is important to note, like we did in the regression analysis, that the addition of other factors associated with education level could change the rates of those being satisfied with their job,and as a result, the probability of a given person being satisfied. This is evident in that some combinations of factors and being satisfied are rarer than others and an indication that with a larger context, education level does somewhat deal with job satisfaction.

  ii) Education level has very little to do with satisfaction and time flexibility

  To determine satisfaction with time flexibility for a given highest education level attained,the variables HRSWKGR and WKSWKGR were used. Much of HRSWKGR and WKSWKGR factors were not relevant(pvalue) to the odds ratio of satisfaction and did not improve model accuracy, so they were left out of the model. Furthermore, the data from the surveys shows that each DGRDG distribution with respect to HRSWKGR/WKSWKGR and JOBSATIS reveals similar proportions in each category on the x-axis(Graph 2/3) and as a result, implies the same conclusion as the article. However, we also claim the last point of the previous paragraph replaced with the appropriate variable.

  iii) Education level has very little to do with satisfaction and income

  To determine satisfaction with income for a given highest education level attained, the variable SATSAL were used. The data from the surveys shows that each DGRDG distribution with respect to SATSAL and JOBSATIS reveals similar proportions in each category (Graph 4) and as a result, implies the same conclusion as the article. However, the minimum salary threshold that the article states ,where the association between highest education level attained and SATSAL disappears, could be present for samples where that salary threshold for any DGRDG level is not met.

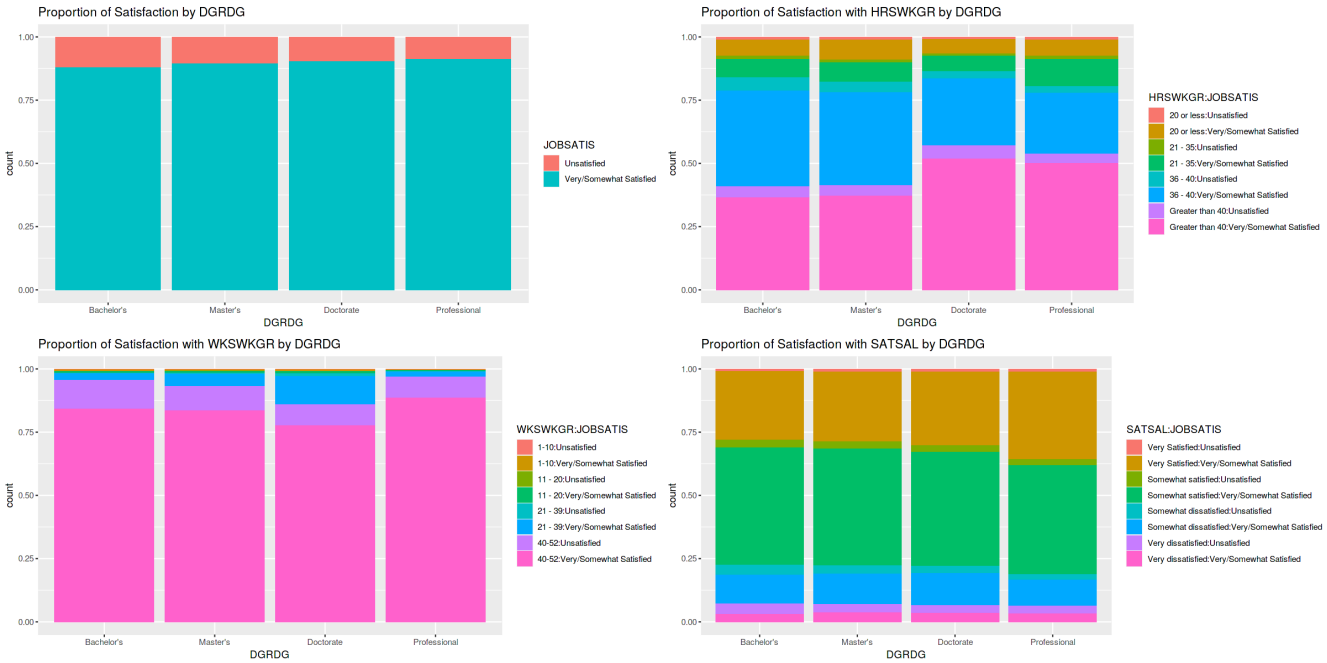
  Broken down by DGRDG, Graphs 2-4 corroborate the article's claim that "income and time flexibility doesn't seem to have much to do with your educational attainment".
- Claim #2: Having the opportunity to do what you do best is the one factor that correlates most highly with overall job satisfaction is

is.

According to Question 7 of the basic analysis section, the best predictors of job satisfaction are those of SATRESP,SATLOC, and SATSAL. Only one of which, SATRESP. can be attributed to doing what you do best because level of responsibility is tied into the job picked by the participant. Therfore, this claim is largely refuted.

In [392]:

```
require(gridExtra)
levels(temp$DGRDG)<-c("Bachelor's","Master's","Doctorate","Professional")
levels(temp$JOBSATIS)<-c("Unsatisfied","Very/Somewhat Satisfied")
levels(temp$HRSWKGR) = c("20 or less", "21 - 35", "36 - 40", "Greater than 40")
levels(temp$WKSWKGR) = c("1-10", "11 - 20", "21 - 39", "40-52")
levels(temp$SATSAL) = c("Very Satisfied","Somewhat satisfied","Somewhat dissatisfied","Very dissati
sfied")

saved <- options(repr.plot.width=20, repr.plot.height=10)
p1<-ggplot(temp,aes(x=DGRDG,fill=JOBSATIS))+geom_bar(position="fill")+xlab("DGRDG")+ ggtitle("Propo
rtion of Satisfaction by DGRDG")
p2<-ggplot(temp,aes(x=DGRDG,fill=HRSWKGR:JOBSATIS))+geom_bar(position="fill")+ ggtitle("Proportion
of Satisfaction with HRSWKGR by DGRDG")
p3<-ggplot(temp,aes(x=DGRDG,fill=WKSWKGR:JOBSATIS))+geom_bar(position="fill")+ ggtitle("Proportion
of Satisfaction with WKSWKGR by DGRDG")
p4<-ggplot(temp,aes(x=DGRDG,fill=SATSAL:JOBSATIS))+geom_bar(position="fill")+ ggtitle("Proportion o
f Satisfaction with SATSAL by DGRDG")
grid.arrange(p1,p2,p3,p4,ncol=2)
options(saved)
```



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

- Claim #1: Certain race groups earn less than others when they have the same education level.

  According to model1, MINRTY is not a significant coefficient and as a result, not relevant to predicting SALARY. There is a difference among races according to the boxplot below, just as the article states. However, the question of whether that there is a significant difference in salary for a given education level between races may be revealed using ANOVA testing for each level of DGRDG. Of course, there is the case where a person one race earns less than that of another race, but we assume the claim discusses that on average, certain race groups earn less than others when they have the same education level.

  Assumptions:

  1. Independence - the data is collected via a sampling method ensuring this (form of SRS)
  2. Approximately normal - we have no major outliers across all DGRDG groups(using boxplots) and the sample size is sufficiently large
  3. Constant Variance - For each DGRDG group, the boxplots reveal similar variances across groups.

Conducting 4 ANOVA tests, where each test uses the same hypotheses, it is possible to observe which groups have a significant difference using pairwise analysis.

H_O: The mean outcome is the same across all groups.

H_A: At least one mean is different.

For all tests: Since the p-value is smaller than $\alpha = 0.05$, we reject H_O and there is sufficient evidence to believe that at least one mean is different.

For the first three tests(bachelor,master,doctorate): There is evidence among Asians and Whites whose highest degree is the ith DGRDG category that the mean salary differs. There is evidence among URMs and Whiteswhose highest degree is the ith DGRDG category that the mean salary differs.There is evidence among Asians and URMs whose highest degree is the ith DGRDG category that the mean salary differs.

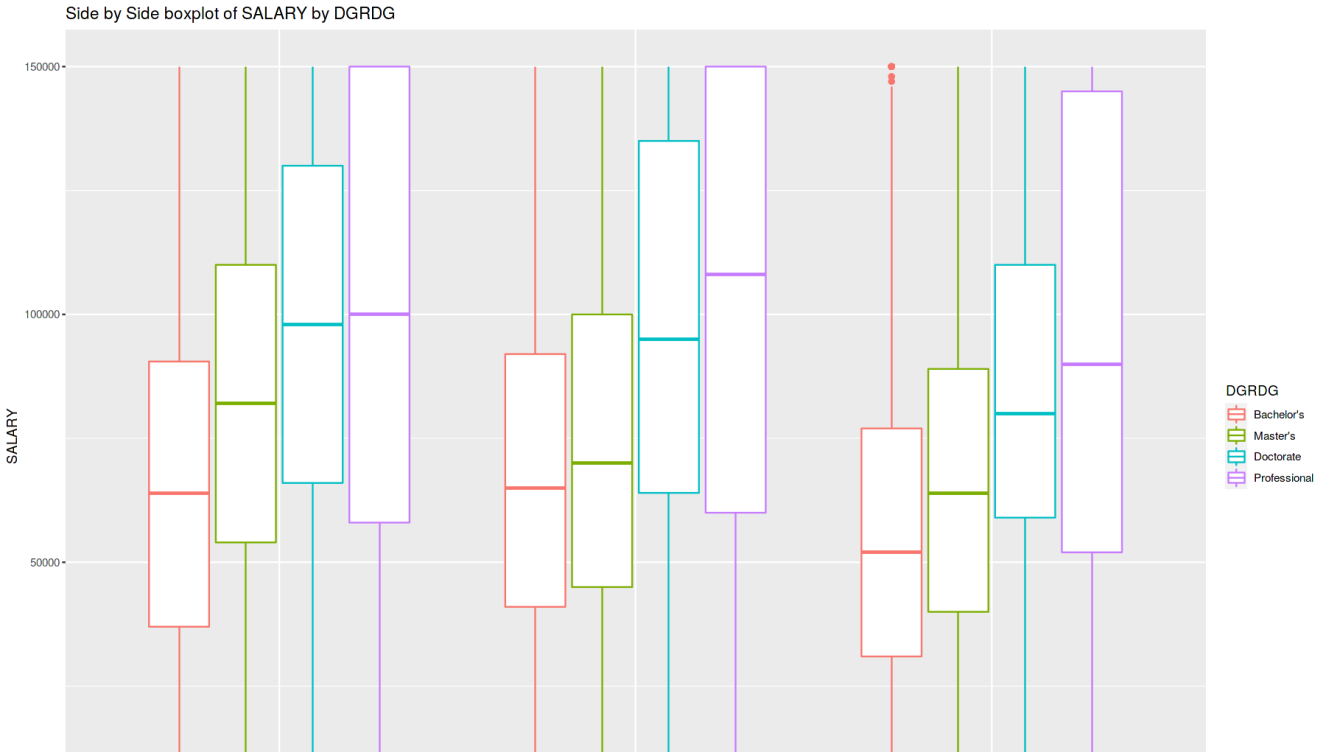
For those with a professional degree(last degree): There is evidence among URMs and Whites whose highest degree is a professional degree that the mean salary differs.There is evidence among Asians and URMs whose highest degree is a professional degree that the mean salary differs.

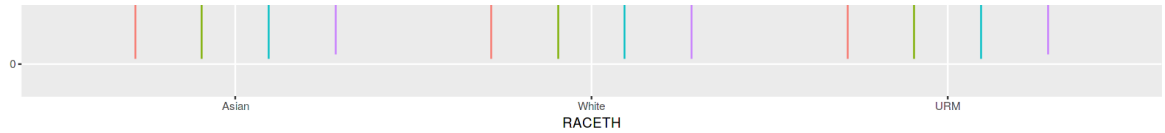
The claim is verified with a dataset that captures that changes in the distribution of SALARY over time, unlike that of the article which analyzes a single year(2008). Furthermore, it is verified that at certain education levels, the differences in salaries among certain race groups are significant.

In [393]:

```
# employees who are employed but have no salary should not be included
temp<-temp[!is.na(temp$SALARY) & temp$SALARY!=0,]
temp <- temp[, !(names(temp) %in% c("LOOKWK","LFSTAT","YEAR","NWFAM","NWLAY","NWNOND","NWOCNA","NWO
TP","NWSTU","PERSONID","SAMPLE","SURID","CHTOT"))]
cols <-
c("JOBSATIS","GENDER","DGRDG","HRSWKGR","WKSWKGR","JOBVAC","JOBINS","JOBPROFT","EMSEC","FTPRET","WA
SCSM","MGRNAT","NOCPRMG","NDGMEMG","NBAMEMG","NRREA", "HDDGRUS","SATSAL","SATADV","SATBEN",
"SATCHAL","SATIND","SATLOC","SATRESP","SATSEC","SATSOC","RACETH","OCEDRLP")
temp[,cols] <- lapply(temp[,cols], factor)
levels(temp$DGRDG)<-c("Bachelor's","Master's","Doctorate","Professional")
levels(temp$RACETH)<-c("Asian","White","URM","Other")
saved <- options(repr.plot.width=15, repr.plot.height=10)
ggplot(temp,aes(x=RACETH,y=SALARY,color=DGRDG))+geom_boxplot()+ggtitle("Side by Side boxplot of SAL
ARY by DGRDG")
options(saved)
table(temp$DGRDG)
```

```
Bachelor's     Master's    Doctorate Professional
     36334        28542        28923         3637
```





Perform the One-Way Anova test for each degree level

In [394]:

```
levels(temp$DGRDG)<-c("1","2","3","4")
levels(temp$RACETH)<-c("1","2","3","4")
i<-1
while(i<=4){
    res.aov.1 <- aov(SALARY ~ RACETH, data = temp[temp$DGRDG==as.character(i),])
    print(summary(res.aov.1))
    print(TukeyHSD(res.aov.1))
    i<-i+1
}
```

```
              Df    Sum Sq   Mean Sq F value Pr(>F)
RACETH         2 8.981e+11 4.490e+11   330.7 <2e-16 ***
Residuals   36331 4.934e+13 1.358e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SALARY ~ RACETH, data = temp[temp$DGRDG == as.character(i), ])

$RACETH
          diff         lwr       upr    p adj
2-1   2126.775    799.4309   3454.12 0.000508
3-1  -9934.674 -11458.3082  -8411.04 0.000000
3-2 -12061.449 -13164.1680 -10958.73 0.000000

              Df    Sum Sq   Mean Sq F value Pr(>F)
RACETH         2 6.738e+11 3.369e+11   228.6 <2e-16 ***
Residuals   28539 4.206e+13 1.474e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SALARY ~ RACETH, data = temp[temp$DGRDG == as.character(i), ])

$RACETH
          diff        lwr        upr p adj
2-1  -7844.256  -9266.398  -6422.115     0
3-1 -15361.775 -17047.922 -13675.629     0
3-2  -7517.519  -8847.801  -6187.237     0

              Df    Sum Sq   Mean Sq F value Pr(>F)
RACETH         2 4.860e+11 2.430e+11     152 <2e-16 ***
Residuals   28920 4.622e+13 1.598e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SALARY ~ RACETH, data = temp[temp$DGRDG == as.character(i), ])

$RACETH
          diff        lwr          upr     p adj
2-1  -1423.075  -2821.938    -24.21228 0.0450831
3-1 -12044.155 -13868.410 -10219.89941 0.0000000
3-2 -10621.080 -12148.864  -9093.29519 0.0000000

             Df    Sum Sq   Mean Sq F value  Pr(>F)
RACETH        2 5.687e+10 2.843e+10   13.77 1.1e-06 ***
Residuals  3634 7.503e+12 2.065e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
```

```
    95% family-wise confidence level

Fit: aov(formula = SALARY ~ RACETH, data = temp[temp$DGRDG == as.character(i), ])

$RACETH
         diff        lwr        upr     p adj
2-1  3683.269  -1264.841  8631.3786 0.1885739
3-1 -6110.889 -11926.152  -295.6272 0.0367233
3-2 -9794.158 -14194.662 -5393.6537 0.0000006
```

Claim #2: STEM (science, technology, engineering and mathematics) careers, in which minorities are underrepresented,tend to pay more than careers in social sciences.

There are two parts to this claim.

1. Are minorities underrepresented in STEM? What is the threshold percentage of the total STEM job force would be considered underepresented?

   Considering that Asians and URMs in STEM are at around the same proportion, they should both be seens as underrepresented because for total equity, combined they are still less in proportion to Whites, who are ~62% of the STEM work force. The threshold percentage is a number not determined by fairness but by politics, available labor,VISA restrictions,etc.

   This part of the claim is verified despite the ambigious definition of underrepresented provided - is it a threshold or is that for there should be equal rates of acceptance among whites and non-whites(makes getting a job for whites more compeititive),etc.

In [395]:

```
temp$STEM <- temp$NOCPRMG == "1" | temp$NOCPRMG == "2" | temp$NOCPRMG == "3" | temp$NOCPRMG == "5"
| temp$NOCPRMG == "6"
prop.table(table(temp$STEM,temp$RACETH),margin=1)
```

```
               1         2         3         4
  FALSE 0.1375269 0.6033113 0.2591618 0.0000000
  TRUE  0.1971676 0.6284214 0.1744110 0.0000000
```

1. STEM careers tend to pay more than careers in social sciences.

   Using the observations that Regression 1 used to predict SALARY, we can observe if it is the case that those with a STEM career(determined by NOCPRMG) have a difference in SALARY from those in social sciences. The boxplot below shows that the median salary of those in STEM careers is only marginally higher than that of those in STEM careers. Furthermore, the variability around the median in both boxplots shows that there is a right skew toward higher salaries but that STEM careers have the highest salaries. This implies that there is a large cluster of salaries between ~\$230000 to ~\$70000 among both groups and is confirmed by the histrograms. Lastly, the middle 50% of data shows that there is equal variablity around the median SALARY for the respective group.

   Considering all salaries, this claim holds. Interestingly, the middle 50% of both groups are very similar in their summary statisitics(variability,median) and are only distinguishable in the last quartile of salaries. Since the article relied on one year(2008) to derive that a "graduate's given occupation may explain the difference in income....", there was a possibility that the disparity in SALARY shown by the article for STEM and Non-STEM groups may only exist when analyzing a single year. However, such a disparity never disappeared when considering all years(using the NSCG+SDR) and the implications of a changing distribution of majors year over year.

   While the difference in SALARY of STEM careers versus that of social sciences is seen though Regression 1 and the corresponding boxplots, determining if the difference is significant is determined by a Two Sample T test. However, multiple outliers in both groups violate the normality condition of such a test.

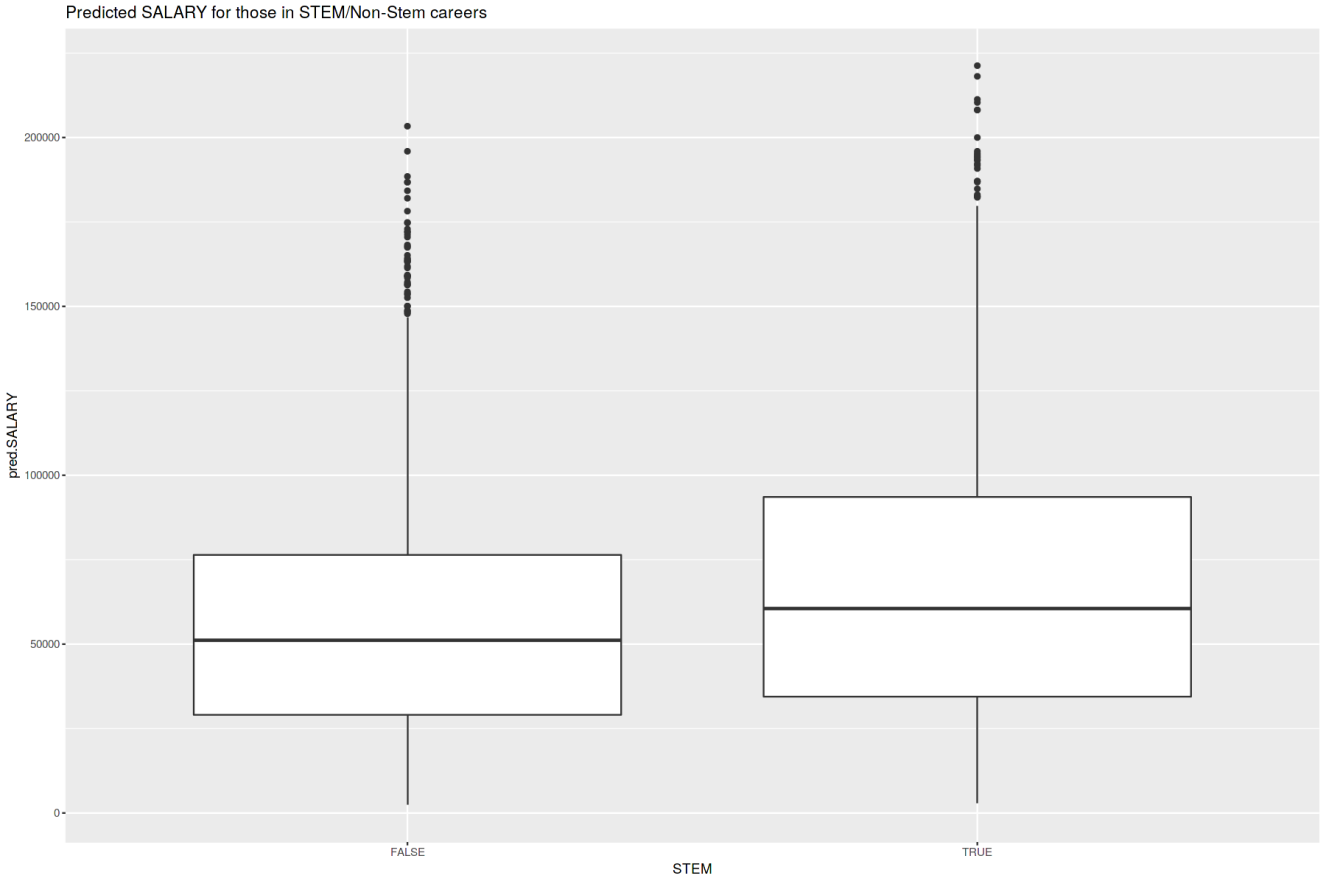
   The claim is valid but the difference in salary between the two groups may or may not be significant.

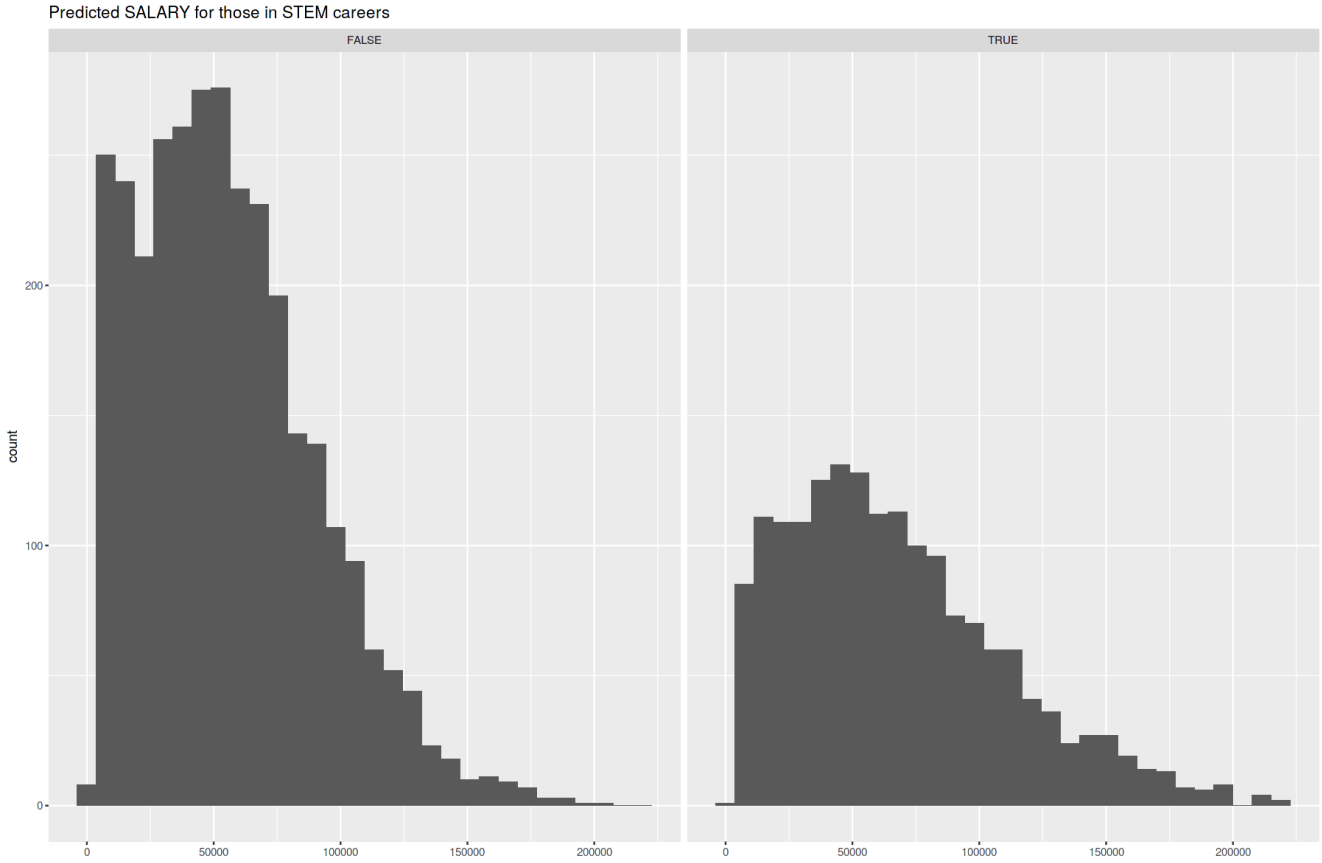
In [396]:

```
# get the fitted values along with the index of the observation it is predicting for
fitted <- data.frame(sort(reg1$fitted.values, decreasing = TRUE))
names(fitted)<-c("fitted")
# get the observations that correspond to their fitted values
fitted_obs<-temp[row.names(fitted),]
fitted_obs$pred.SALARY <- exp(fitted(reg1))
fitted_obs$STEM <- fitted_obs$NOCPRMG == "1" | fitted_obs$NOCPRMG == "2" | fitted_obs$NOCPRMG == "3
```

```
" | fitted_obs$NOCPRMG == "5" | fitted_obs$NOCPRMG == "6"
ggplot(fitted_obs,aes(x=STEM,y=pred.SALARY))+geom_boxplot()+ggtitle("Predicted SALARY for those in STEM/Non-Stem careers")

ggplot(fitted_obs,aes(x=pred.SALARY))+geom_histogram()+ggtitle("Predicted SALARY for those in STEM careers")+facet_grid(~STEM)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





PEW: the rising cost of not going to college

Claim #1: 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.

From the SDR and NSCG surveys, this claim fails to hold by the stacked bar charts below because there are around equal proportions of those who say their current job is “very closely” related to their college or graduate field of study among the STEM and non-STEM groups. However, to test whether the SDR and NSCG surveys have data due to chance(meaning the claim could be valid), we can perform a permutation test.

The permutation test is performed with the highest degree trained for, for those that had an OCEDRLP opinion, just as the article does.

STEM majors are defined as all categories in NDGMEMG except for 4 and 7.

$ H_O: $ There is no difference in relevance of degree and degree trained for (STEM vs Non-STEM) .

$ H_A: $ There is a difference in relevance of degree and degree trained for (STEM vs Non-STEM) .

Let d_i be the observed $\sum_{i=1}^{2} |r_i - \bar{r}| $ where r is the rate of people with the ith category of STEM that is relevant(OCEDRLP = 1,very closely) where i ranges from 1 to 2.

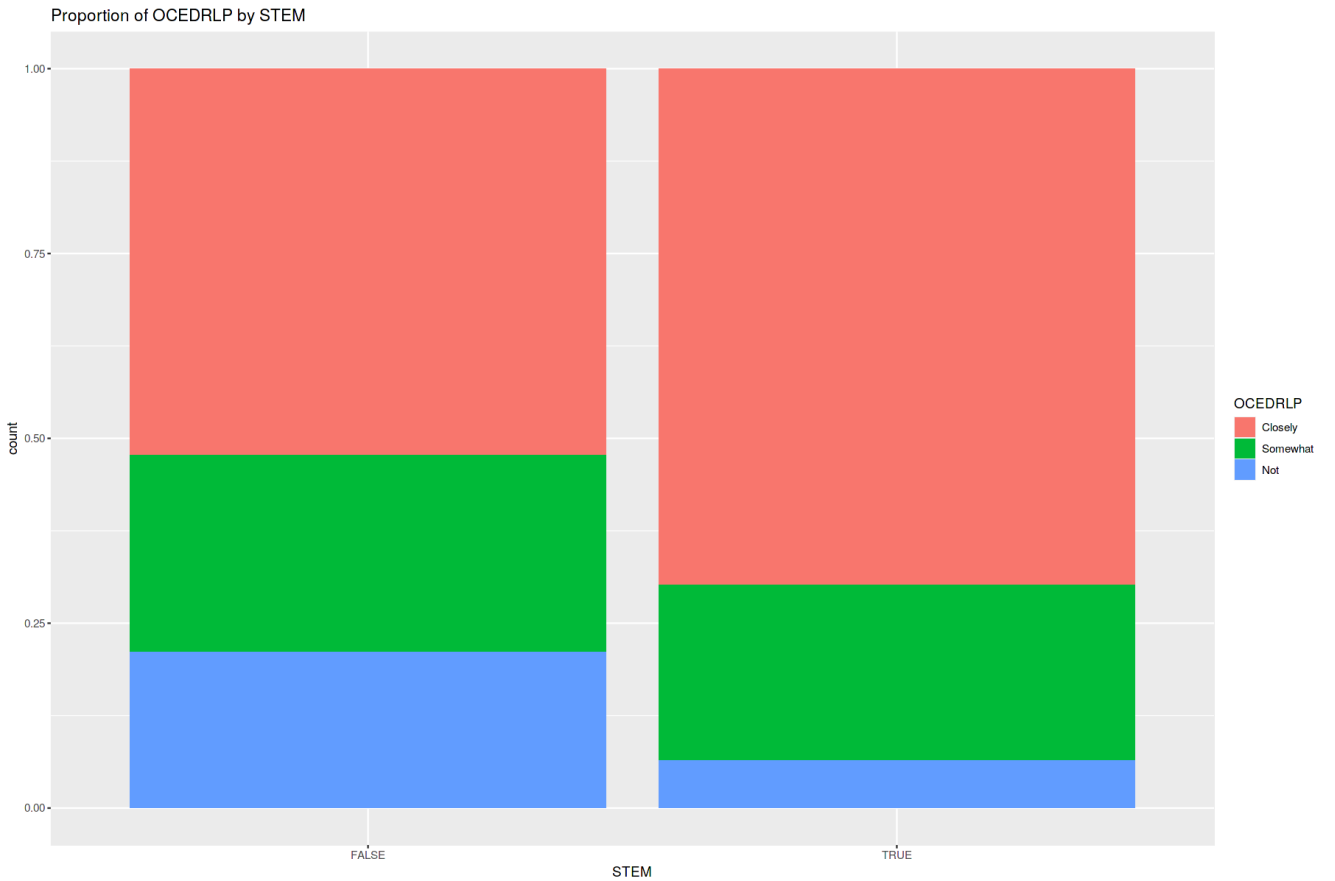
Forming a contingency table for every shuffled OCEDRLP and calculating rates of very close relationships among STEM and Non-STEM groups would aid in calculation of r_i

Form a distribution D from the permuation test such that we can find $P(E|H_0)$ where $E=D \geq d$

This means we are looking for the probability that an event is as extreme or more extreme than what we saw in the observed data.

In [397]:

```
levels(temp$OCEDRLP)<-c("Closely","Somewhat","Not","Logical Skip")
ggplot(temp,aes(x=STEM,fill=OCEDRLP))+geom_bar(position="fill")+xlab("STEM")+ ggtitle("Proportion o
f OCEDRLP by STEM")
levels(temp$OCEDRLP)<-c("1","2","3","98")
```



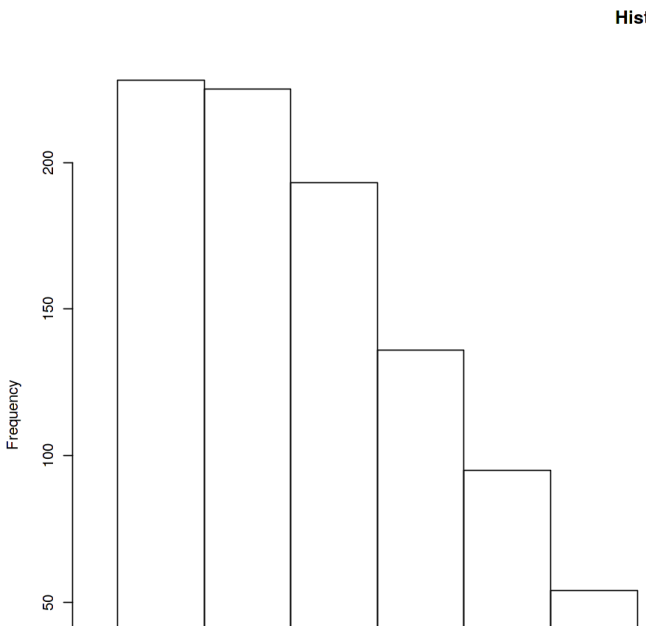
Permutation Test

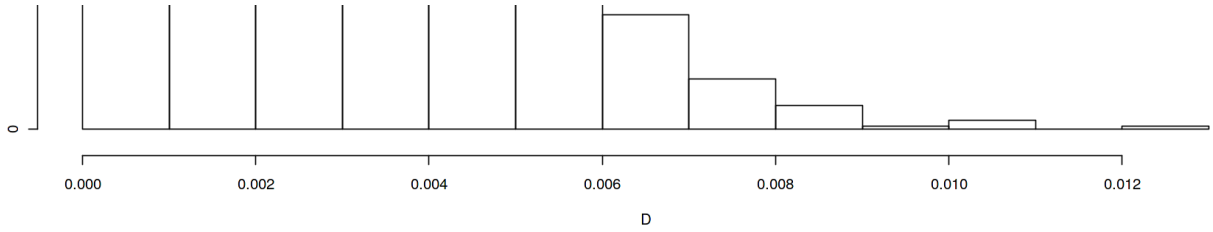
In [398]:

```
temp$STEM <- temp$NDGMEMG == "1" | temp$NDGMEMG == "2" | temp$NDGMEMG == "3" | temp$NDGMEMG == "5"
| temp$NDGMEMG == "6"
tb<-prop.table(table(temp$STEM,temp$OCEDRLP),margin=1)
tb

#compute the data deviation
data.deviation <- function(rates){
    r.bar <- mean(rates)
    s <- sum(abs(rates-r.bar))
    return (s)
}

i<-1
rates<-c()
while(i<=2){
    rates<-append(rates,tb[i,1])
    i <- i+1
}
d<- data.deviation(rates)

shuffle <- function(){
    OCEDRLP.shuffle <- sample(temp$OCEDRLP)
    tb <- prop.table(table(temp$STEM,OCEDRLP.shuffle),margin=1)
    rates<-c()
    i<-1
    while(i<=2){
        rates<-append(rates,tb[i,1])
        i <- i+1
    }
    test_statistic <- (data.deviation(rates))
    return (test_statistic)
}

num.exp <- 10**3
D <- replicate(num.exp,shuffle())

p.hat <-sum(D>=d)/num.exp
hist(D)
print(paste("Observed Test Statistic",d,"; p-value: ", p.hat))
```

```
                1          2          3         98
  FALSE 0.56665844 0.25516433 0.17817722 0.00000000
  TRUE  0.66074890 0.24536168 0.09388942 0.00000000
```

```
[1] "Observed Test Statistic 0.0940904551920124 ; p-value:  0"
```





phat(~0) is less than any α > 0, so we reject H_O and the results are significant. There is sufficient evidence to believe there is a difference in the degree trained for(STEM vs Non-STEM) and relevance of degree. As a result, the claim is verified because the probability of observing no difference in relevance of degree and data under the null is quite unlikely. The survey methodology of the article was able the opinions of relevance for all age groups due to weighting,so it goes to further corroborate Claim #1.