Project Assignment 4

Elikem Asudo Tsatsu Gale-Zoyiku

2023-04-26

## Research Question

**Is there an association between the sexual orientation of cis-males, depression levels, and their income? Can education explain any association if there is?**

Explanatory Variables: Gender (Categorical), Sexual Orientation (Categorical), and Educational attainment (categorical)

Response Variables: Depression levels (Categorical), Anxiety levels (Categorical), Ratio of income to poverty level (Categorical)

### 1. Load data set(s) and libraries

load("C:/Users/egale/OneDrive - Ashesi University/Desktop/Statistics with Probability/IPUMS\_NHIS.RData")  
library(descr)

## Warning: package 'descr' was built under R version 4.2.3

library(stats)  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

### 2. Create variable subset

vars=c("AGE","SEX","SEXORIEN","EDUC","DEPFEELEVL","DEPFREQ","POVERTY","WORFEELEVL","WORFREQ")  
myHealthData = IPUMS\_NHIS[vars]  
str(myHealthData)

## 'data.frame': 116291 obs. of 9 variables:  
## $ AGE : int 47 61 77 37 8 34 68 46 36 5 ...  
## $ SEX : int 2 1 2 2 1 2 2 2 2 2 ...  
## $ SEXORIEN : int 2 2 2 2 0 2 2 2 2 0 ...  
## $ EDUC : int 400 201 301 303 0 400 201 400 301 0 ...  
## $ DEPFEELEVL: int 3 0 3 0 0 1 1 1 0 0 ...  
## $ DEPFREQ : int 4 5 4 5 5 3 4 4 5 5 ...  
## $ POVERTY : int 37 32 33 34 34 37 11 37 33 33 ...  
## $ WORFEELEVL: int 3 0 1 2 0 2 1 2 0 0 ...  
## $ WORFREQ : int 2 5 4 2 5 4 1 4 5 5 ...

myHealthData<-as.data.frame(myHealthData)  
str(myHealthData)

## 'data.frame': 116291 obs. of 9 variables:  
## $ AGE : int 47 61 77 37 8 34 68 46 36 5 ...  
## $ SEX : int 2 1 2 2 1 2 2 2 2 2 ...  
## $ SEXORIEN : int 2 2 2 2 0 2 2 2 2 0 ...  
## $ EDUC : int 400 201 301 303 0 400 201 400 301 0 ...  
## $ DEPFEELEVL: int 3 0 3 0 0 1 1 1 0 0 ...  
## $ DEPFREQ : int 4 5 4 5 5 3 4 4 5 5 ...  
## $ POVERTY : int 37 32 33 34 34 37 11 37 33 33 ...  
## $ WORFEELEVL: int 3 0 1 2 0 2 1 2 0 0 ...  
## $ WORFREQ : int 2 5 4 2 5 4 1 4 5 5 ...

### 3. Data management I: check for and recode errors and NAs

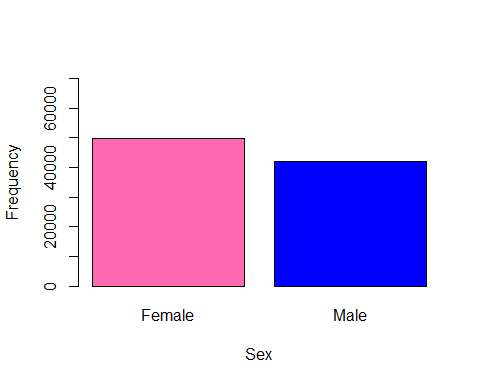
myHealthData$Gender <- rep(NA,nrow(myHealthData))  
myHealthData$Gender[myHealthData$SEX==1] <- 'Male'  
myHealthData$Gender[myHealthData$SEX==2] <- 'Female'  
  
  
myHealthData$WorkingAge <- rep(NA,nrow(myHealthData))  
myHealthData$WorkingAge[myHealthData$AGE<=18] <- 'Not working age'  
myHealthData$WorkingAge[myHealthData$AGE>18 & myHealthData$AGE<997] <- 'Working age'  
  
  
myHealthData$Orientation <- rep(NA,nrow(myHealthData))  
myHealthData$Orientation[myHealthData$SEXORIEN==1] <- 'Not Straight'  
myHealthData$Orientation[myHealthData$SEXORIEN==3] <- 'Not Straight'  
myHealthData$Orientation[myHealthData$SEXORIEN==4] <- 'Not Straight'  
myHealthData$Orientation[myHealthData$SEXORIEN==2] <- 'Straight'  
  
myHealthData$EDUC<-as.numeric(myHealthData$EDUC)  
myHealthData$Education <- rep(NA, nrow(myHealthData))  
myHealthData$Education[myHealthData$EDUC < 112] <- "No high school"  
myHealthData$Education[myHealthData$EDUC >= 112 & myHealthData$EDUC<116] <- "Some high school"  
myHealthData$Education[myHealthData$EDUC >= 200 & myHealthData$EDUC<=202] <- "High school diploma or equivalent"  
myHealthData$Education[myHealthData$EDUC >= 300 & myHealthData$EDUC<=301] <- "Some college"  
myHealthData$Education[myHealthData$EDUC >= 302 & myHealthData$EDUC<=303] <- "Associate's Degree"  
myHealthData$Education[myHealthData$EDUC == 400] <- "Bachelor's Degree"  
myHealthData$Education[myHealthData$EDUC == 504] <- "Other Degree"  
myHealthData$Education[myHealthData$EDUC >= 500 & myHealthData$EDUC<=503] <- "Postgraduate"  
  
  
myHealthData$DepressionFrequency <- rep(NA, nrow(myHealthData))  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 1] <- "Daily"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 2] <- "Weekly"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 3] <- "Monthly"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 4] <- "Rarely"  
myHealthData$DepressionFrequency[myHealthData$DEPFREQ == 5] <- "Never"  
  
  
myHealthData$DepressionLevel <- rep(NA, nrow(myHealthData))  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL == 1] <- "A lot"  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL== 2] <- "A little"  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL == 3] <- "Somewhere in between"  
myHealthData$DepressionLevel[myHealthData$DEPFEELEVL == 0] <- NA  
  
  
myHealthData$WorryFrequency <- rep(NA, nrow(myHealthData))  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 1] <- "Daily"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 2] <- "Weekly"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 3] <- "Monthly"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 4] <- "Rarely"  
myHealthData$WorryFrequency[myHealthData$WORFREQ == 5] <- "Never"  
  
  
myHealthData$WorryLevel <- rep(NA, nrow(myHealthData))  
myHealthData$WorryLevel[myHealthData$WORFEELEVL == 1] <- "A lot"  
myHealthData$WorryLevel[myHealthData$WORFEELEVL== 2] <- "A little"  
myHealthData$WorryLevel[myHealthData$WORFEELEVL == 3] <- "Somewhere in between"  
myHealthData$WorryLevel[myHealthData$WORFEELEVL == 0] <- NA  
  
  
myHealthData$PovertyRatio <- rep(NA, nrow(myHealthData))  
myHealthData$PovertyRatio[myHealthData$POVERTY<=14] <- "Below the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=25 & myHealthData$POVERTY>=20] <- "Between 1 to 1.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=32 & myHealthData$POVERTY>=31] <- "Between 2 to 2.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=34 & myHealthData$POVERTY>=33] <- "Between 3 to 3.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY<=36 & myHealthData$POVERTY>=35] <- "Between 4 to 4.99 units above the poverty line"  
myHealthData$PovertyRatio[myHealthData$POVERTY==37] <- "5 units and above the poverty line"

### 4. Data management II: further subset and create secondary variable

# Subset the data frame to take out observations below the working age  
myHealthData <- myHealthData[myHealthData$WorkingAge == "Working age", ]

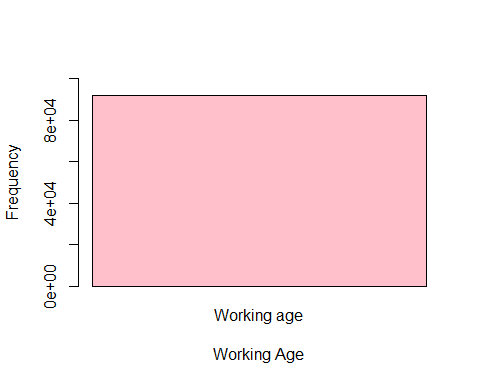
### 5. Descriptive statistics (sample means, standard deviations, proportions) and uni-variate displays

freq(myHealthData$Gender, ylab="Frequency",xlab="Sex",ylim=c(0,70000),col=c("hotpink","blue"))



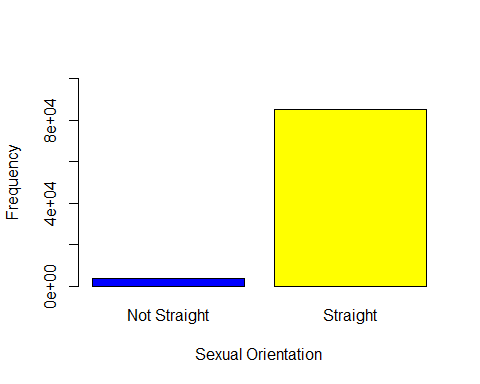
## myHealthData$Gender   
## Frequency Percent Valid Percent  
## Female 49932 54.0553 54.2  
## Male 42197 45.6816 45.8  
## NA's 243 0.2631   
## Total 92372 100.0000 100.0

freq(myHealthData$WorkingAge, ylab="Frequency",xlab="Working Age",ylim=c(0,100000),col=c("pink","blue"))



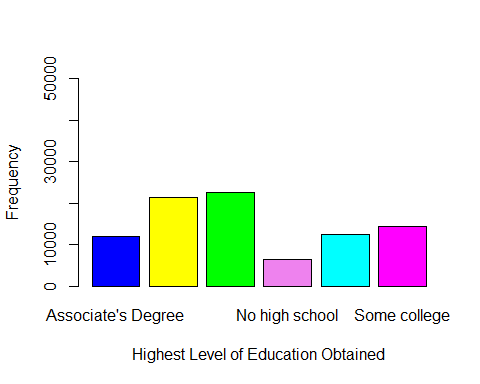
## myHealthData$WorkingAge   
## Frequency Percent Valid Percent  
## Working age 92135 99.7434 100  
## NA's 237 0.2566   
## Total 92372 100.0000 100

freq(myHealthData$Orientation,ylab="Frequency",xlab="Sexual Orientation",ylim=c(0,100000), col = c("blue","yellow"))



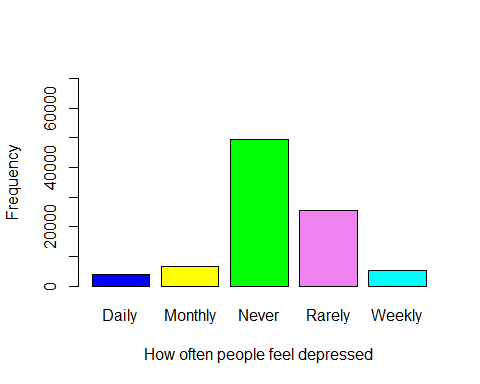
## myHealthData$Orientation   
## Frequency Percent Valid Percent  
## Not Straight 3418 3.700 3.86  
## Straight 85132 92.162 96.14  
## NA's 3822 4.138   
## Total 92372 100.000 100.00

freq(myHealthData$Education,ylab="Frequency",xlab="Highest Level of Education Obtained",ylim=c(0,50000),col=c("blue","yellow","green","violet","cyan","magenta"))



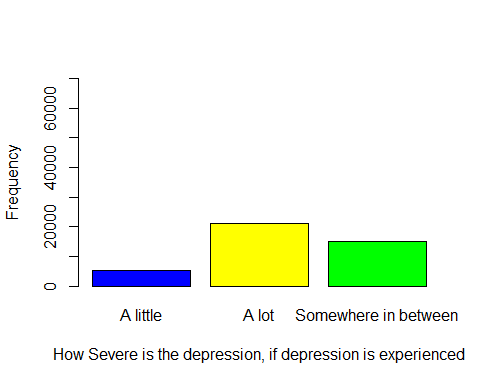
## myHealthData$Education   
## Frequency Percent Valid Percent  
## Associate's Degree 11966 12.954 13.403  
## Bachelor's Degree 21274 23.031 23.829  
## High school diploma or equivalent 22620 24.488 25.336  
## No high school 6430 6.961 7.202  
## Postgraduate 12525 13.559 14.029  
## Some college 14464 15.658 16.201  
## NA's 3093 3.348   
## Total 92372 100.000 100.000

freq(myHealthData$DepressionFrequency,ylab="Frequency",xlab="How often people feel depressed",ylim=c(0,70000),col=c("blue","yellow","green","violet","cyan","magenta"))



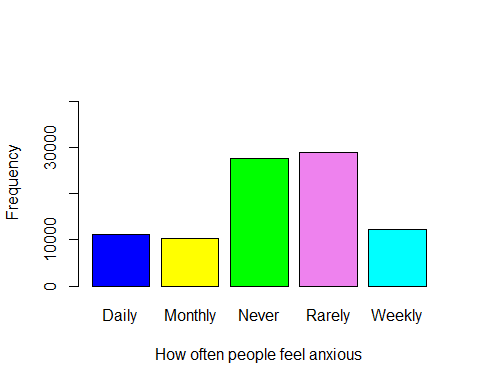
## myHealthData$DepressionFrequency   
## Frequency Percent Valid Percent  
## Daily 3750 4.060 4.151  
## Monthly 6494 7.030 7.188  
## Never 49367 53.444 54.642  
## Rarely 25423 27.522 28.140  
## Weekly 5312 5.751 5.880  
## NA's 2026 2.193   
## Total 92372 100.000 100.000

freq(myHealthData$DepressionLevel,ylab="Frequency",xlab="How Severe is the depression, if depression is experienced",ylim=c(0,70000),col=c("blue","yellow","green","violet","cyan","magenta"))



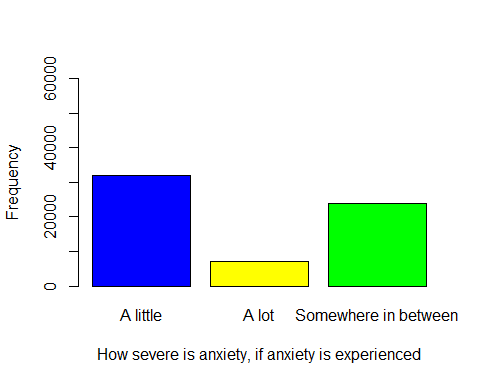
## myHealthData$DepressionLevel   
## Frequency Percent Valid Percent  
## A little 5338 5.779 12.80  
## A lot 21247 23.002 50.94  
## Somewhere in between 15126 16.375 36.26  
## NA's 50661 54.845   
## Total 92372 100.000 100.00

freq(myHealthData$WorryFrequency,ylab="Frequency",xlab="How often people feel anxious",ylim=c(0,45000),col=c("blue","yellow","green","violet","cyan","magenta"))



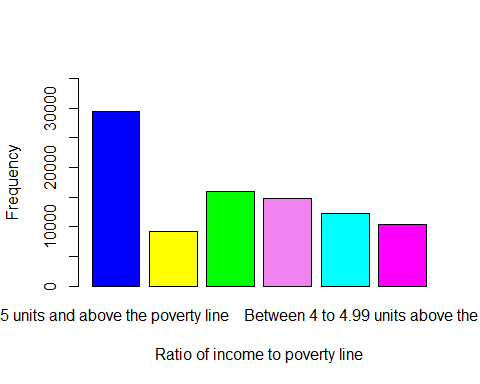
## myHealthData$WorryFrequency   
## Frequency Percent Valid Percent  
## Daily 11175 12.098 12.36  
## Monthly 10235 11.080 11.32  
## Never 27767 30.060 30.71  
## Rarely 29025 31.422 32.10  
## Weekly 12226 13.236 13.52  
## NA's 1944 2.105   
## Total 92372 100.000 100.00

freq(myHealthData$WorryLevel,ylab="Frequency",xlab="How severe is anxiety, if anxiety is experienced",ylim=c(0,60000),col=c("blue","yellow","green","violet","cyan","magenta"))



## myHealthData$WorryLevel   
## Frequency Percent Valid Percent  
## A little 31953 34.59 50.66  
## A lot 7187 7.78 11.39  
## Somewhere in between 23934 25.91 37.95  
## NA's 29298 31.72   
## Total 92372 100.00 100.00

freq(myHealthData$PovertyRatio,ylab="Frequency",xlab="Ratio of income to poverty line",ylim=c(0,35000),col=c("blue","yellow","green","violet","cyan","magenta"))



## myHealthData$PovertyRatio   
## Frequency Percent Valid Percent  
## 5 units and above the poverty line 29500 31.9361 32.018  
## Below the poverty line 9137 9.8915 9.917  
## Between 1 to 1.99 units above the poverty line 15952 17.2693 17.314  
## Between 2 to 2.99 units above the poverty line 14831 16.0557 16.097  
## Between 3 to 3.99 units above the poverty line 12312 13.3287 13.363  
## Between 4 to 4.99 units above the poverty line 10403 11.2621 11.291  
## NA's 237 0.2566   
## Total 92372 100.0000 100.000

t<-table(myHealthData$Gender)  
p<-prop.table(t)\*100;p

##   
## Female Male   
## 54.19792 45.80208

t<-table(myHealthData$WorkingAge)  
p<-prop.table(t)\*100;p

##   
## Working age   
## 100

t<-table(myHealthData$Orientation)  
p<-prop.table(t)\*100;p

##   
## Not Straight Straight   
## 3.859966 96.140034

t<-table(myHealthData$Education)  
p<-prop.table(t)\*100;p

##   
## Associate's Degree Bachelor's Degree   
## 13.402928 23.828672   
## High school diploma or equivalent No high school   
## 25.336305 7.202142   
## Postgraduate Some college   
## 14.029055 16.200898

t<-table(myHealthData$DepressionFrequency)  
p<-prop.table(t)\*100;p

##   
## Daily Monthly Never Rarely Weekly   
## 4.150709 7.187922 54.642153 28.139597 5.879618

t<-table(myHealthData$DepressionLevel)  
p<-prop.table(t)\*100;p

##   
## A little A lot Somewhere in between   
## 12.79758 50.93860 36.26382

t<-table(myHealthData$WorryFrequency)  
p<-prop.table(t)\*100;p

##   
## Daily Monthly Never Rarely Weekly   
## 12.35790 11.31840 30.70620 32.09736 13.52015

t<-table(myHealthData$WorryLevel)  
p<-prop.table(t)\*100;p

##   
## A little A lot Somewhere in between   
## 50.65954 11.39455 37.94590

t<-table(myHealthData$PovertyRatio)  
p<-prop.table(t)\*100;p

##   
## 5 units and above the poverty line   
## 32.01823   
## Below the poverty line   
## 9.91697   
## Between 1 to 1.99 units above the poverty line   
## 17.31372   
## Between 2 to 2.99 units above the poverty line   
## 16.09703   
## Between 3 to 3.99 units above the poverty line   
## 13.36300   
## Between 4 to 4.99 units above the poverty line   
## 11.29104

### 6. Bivariate tables and graphs

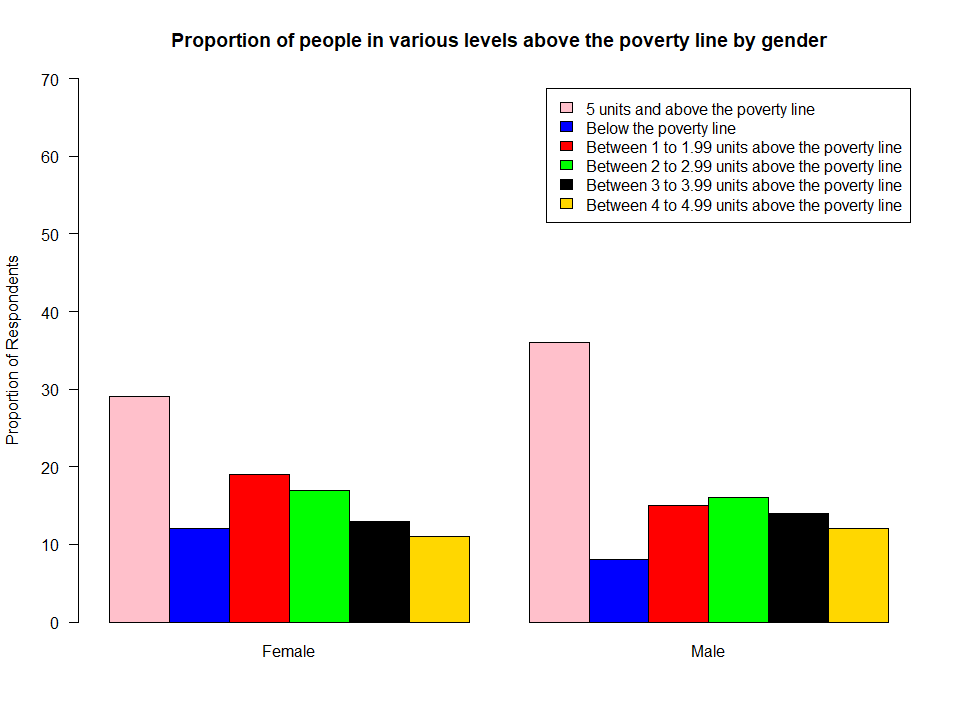
tab1<-table(myHealthData$PovertyRatio,myHealthData$Gender);tab1

##   
## Female Male  
## 5 units and above the poverty line 14400 15097  
## Below the poverty line 5768 3369  
## Between 1 to 1.99 units above the poverty line 9500 6450  
## Between 2 to 2.99 units above the poverty line 8286 6545  
## Between 3 to 3.99 units above the poverty line 6492 5820  
## Between 4 to 4.99 units above the poverty line 5486 4916

prob\_tab1<-round(prop.table(tab1,2),2)\*100; prob\_tab1

##   
## Female Male  
## 5 units and above the poverty line 29 36  
## Below the poverty line 12 8  
## Between 1 to 1.99 units above the poverty line 19 15  
## Between 2 to 2.99 units above the poverty line 17 16  
## Between 3 to 3.99 units above the poverty line 13 14  
## Between 4 to 4.99 units above the poverty line 11 12

barplot(prob\_tab1,main="Proportion of people in various levels above the poverty line by gender",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","black","gold"), beside = TRUE,legend=row.names(prob\_tab1),ylim=c(0,70))



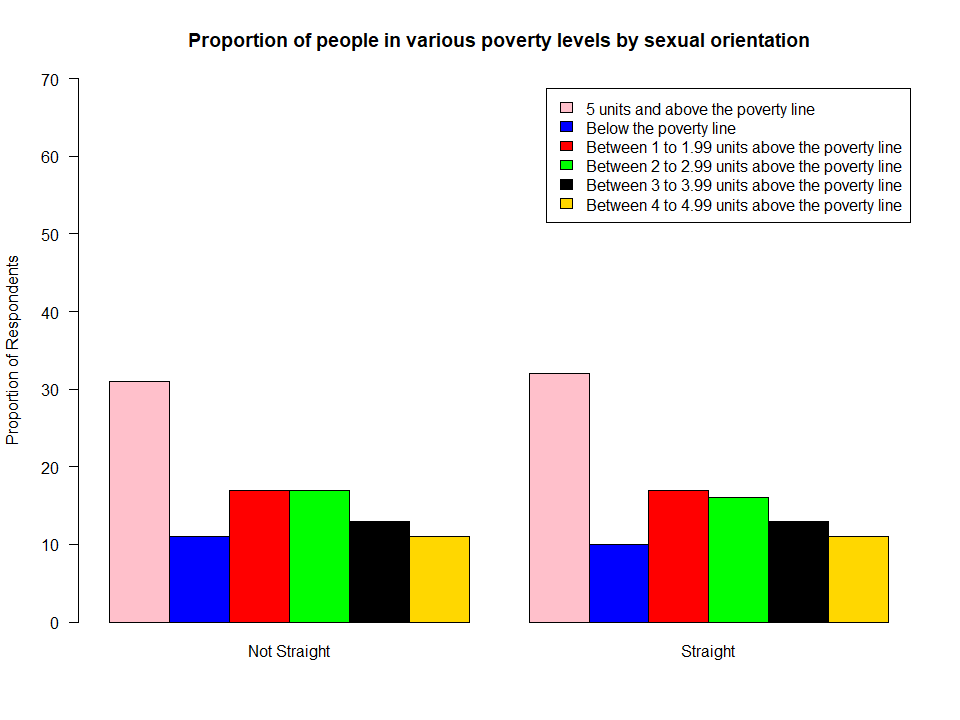
tab2<-table(myHealthData$PovertyRatio,myHealthData$Orientation);tab2

##   
## Not Straight Straight  
## 5 units and above the poverty line 1064 27477  
## Below the poverty line 392 8214  
## Between 1 to 1.99 units above the poverty line 590 14637  
## Between 2 to 2.99 units above the poverty line 571 13630  
## Between 3 to 3.99 units above the poverty line 440 11434  
## Between 4 to 4.99 units above the poverty line 361 9740

prob\_tab2<-round(prop.table(tab2,2),2)\*100;prob\_tab2

##   
## Not Straight Straight  
## 5 units and above the poverty line 31 32  
## Below the poverty line 11 10  
## Between 1 to 1.99 units above the poverty line 17 17  
## Between 2 to 2.99 units above the poverty line 17 16  
## Between 3 to 3.99 units above the poverty line 13 13  
## Between 4 to 4.99 units above the poverty line 11 11

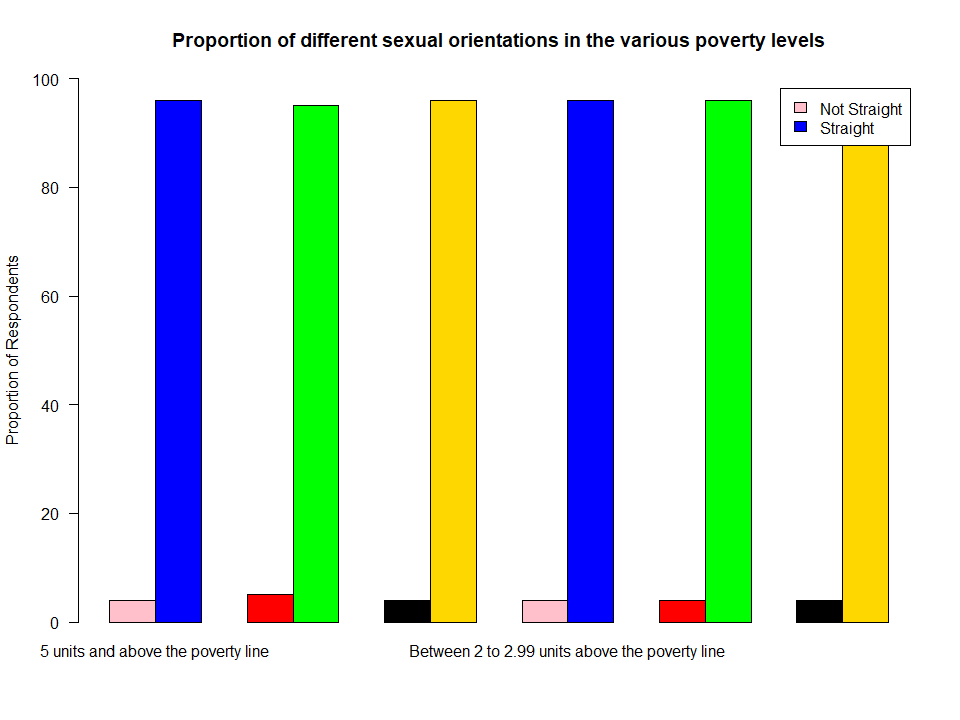
barplot(prob\_tab2,main="Proportion of people in various poverty levels by sexual orientation",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","black","gold"), beside = TRUE,legend=row.names(prob\_tab2),ylim=c(0,70))



tab2<-table(myHealthData$Orientation,myHealthData$PovertyRatio);tab2

##   
## 5 units and above the poverty line Below the poverty line  
## Not Straight 1064 392  
## Straight 27477 8214  
##   
## Between 1 to 1.99 units above the poverty line  
## Not Straight 590  
## Straight 14637  
##   
## Between 2 to 2.99 units above the poverty line  
## Not Straight 571  
## Straight 13630  
##   
## Between 3 to 3.99 units above the poverty line  
## Not Straight 440  
## Straight 11434  
##   
## Between 4 to 4.99 units above the poverty line  
## Not Straight 361  
## Straight 9740

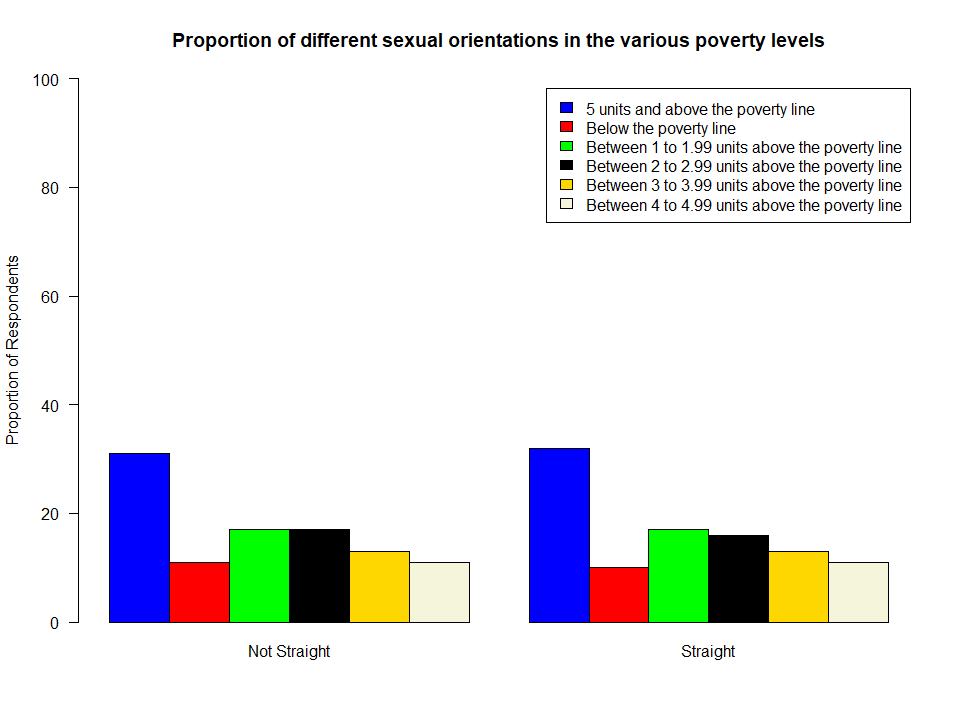
prob\_tab2<-round(prop.table(tab2,2),2)\*100  
barplot(prob\_tab2,main="Proportion of different sexual orientations in the various poverty levels",ylab="Proportion of Respondents", las=1,col=c("pink","blue","red","green","black","gold"), beside = TRUE,legend=row.names(prob\_tab2),ylim=c(0,100))



tab2<-table(myHealthData$PovertyRatio,myHealthData$Orientation);tab2

##   
## Not Straight Straight  
## 5 units and above the poverty line 1064 27477  
## Below the poverty line 392 8214  
## Between 1 to 1.99 units above the poverty line 590 14637  
## Between 2 to 2.99 units above the poverty line 571 13630  
## Between 3 to 3.99 units above the poverty line 440 11434  
## Between 4 to 4.99 units above the poverty line 361 9740

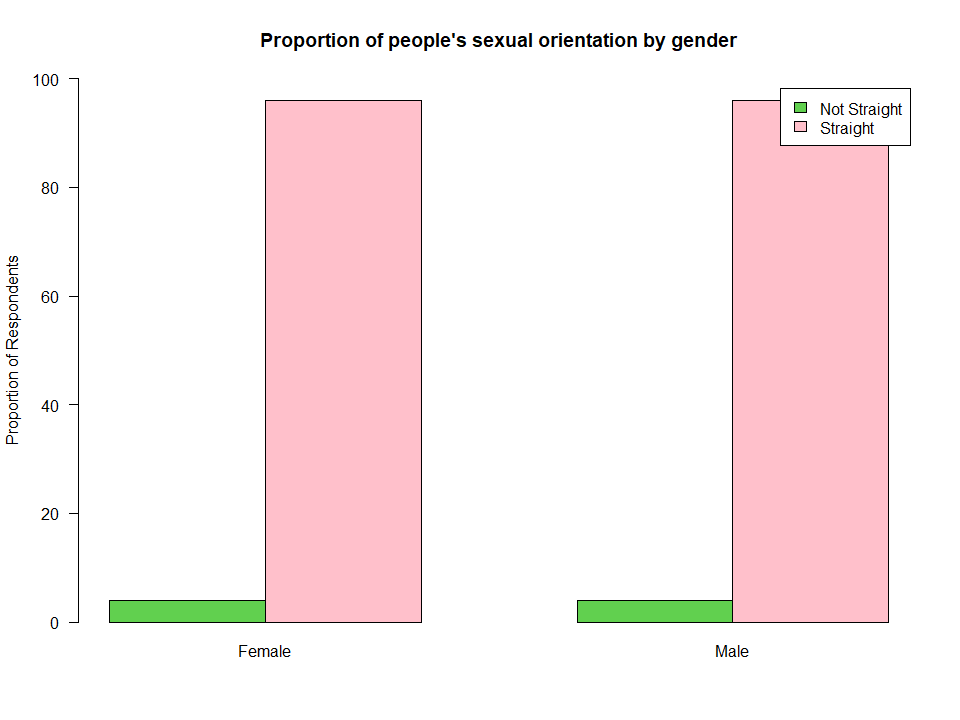
prob\_tab2<-round(prop.table(tab2,2),2)\*100  
barplot(prob\_tab2,main="Proportion of different sexual orientations in the various poverty levels",ylab="Proportion of Respondents", las=1, col=c("blue","red","green","black","gold","beige"), beside = TRUE,legend=row.names(prob\_tab2),ylim=c(0,100))



tab3<-table(myHealthData$Orientation,myHealthData$Gender);tab3

##   
## Female Male  
## Not Straight 1958 1456  
## Straight 45980 39151

prob\_tab3<-round(prop.table(tab3,2),2)\*100  
barplot(prob\_tab3,main="Proportion of people's sexual orientation by gender",ylab="Proportion of Respondents", las=1, col=c(123,"pink"), beside = TRUE,legend=row.names(prob\_tab3),ylim=c(0,100))



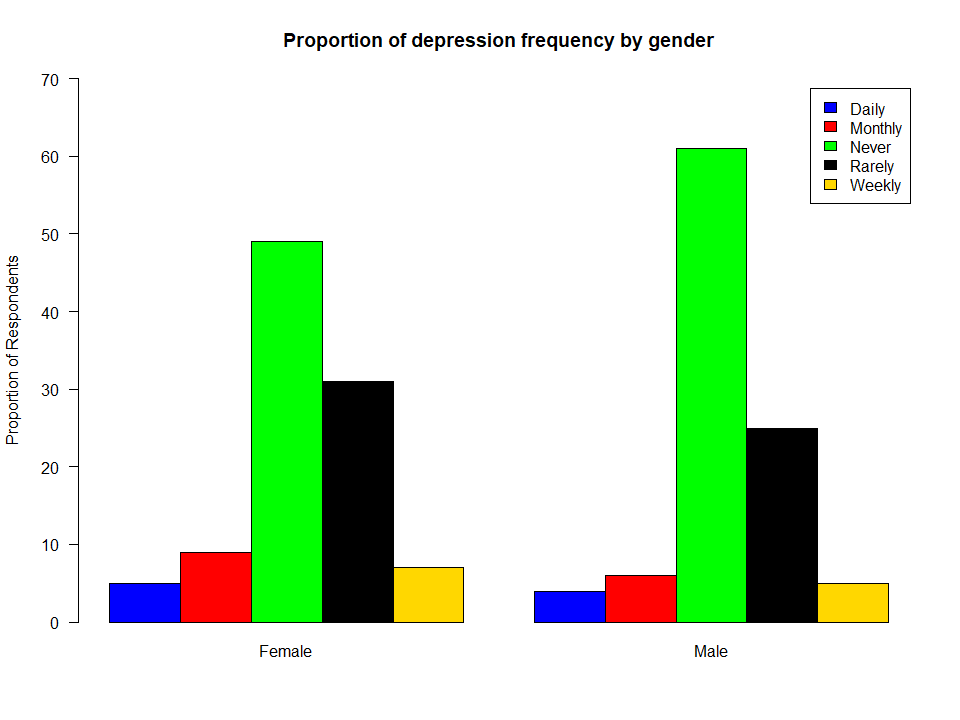
tab4 <- table(myHealthData$DepressionFrequency,myHealthData$Gender); tab4

##   
## Female Male  
## Daily 2287 1462  
## Monthly 4179 2313  
## Never 24215 25150  
## Rarely 14967 10456  
## Weekly 3274 2037

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Female Male  
## Daily 5 4  
## Monthly 9 6  
## Never 49 61  
## Rarely 31 25  
## Weekly 7 5

barplot(prop\_tab4,main="Proportion of depression frequency by gender",ylab="Proportion of Respondents", las=1, col=c("blue","red","green","black","gold"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



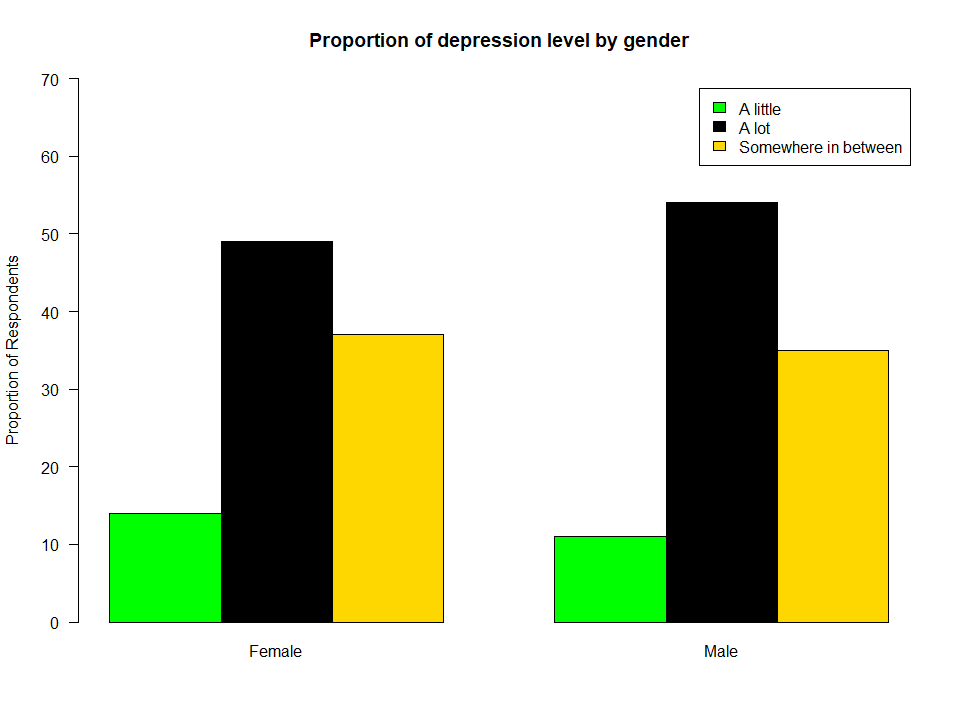
tab4 <- table(myHealthData$DepressionLevel,myHealthData$Gender); tab4

##   
## Female Male  
## A little 3498 1839  
## A lot 12381 8865  
## Somewhere in between 9359 5765

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Female Male  
## A little 14 11  
## A lot 49 54  
## Somewhere in between 37 35

barplot(prop\_tab4,main="Proportion of depression level by gender",ylab="Proportion of Respondents", las=1, col=c("green","black","gold"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



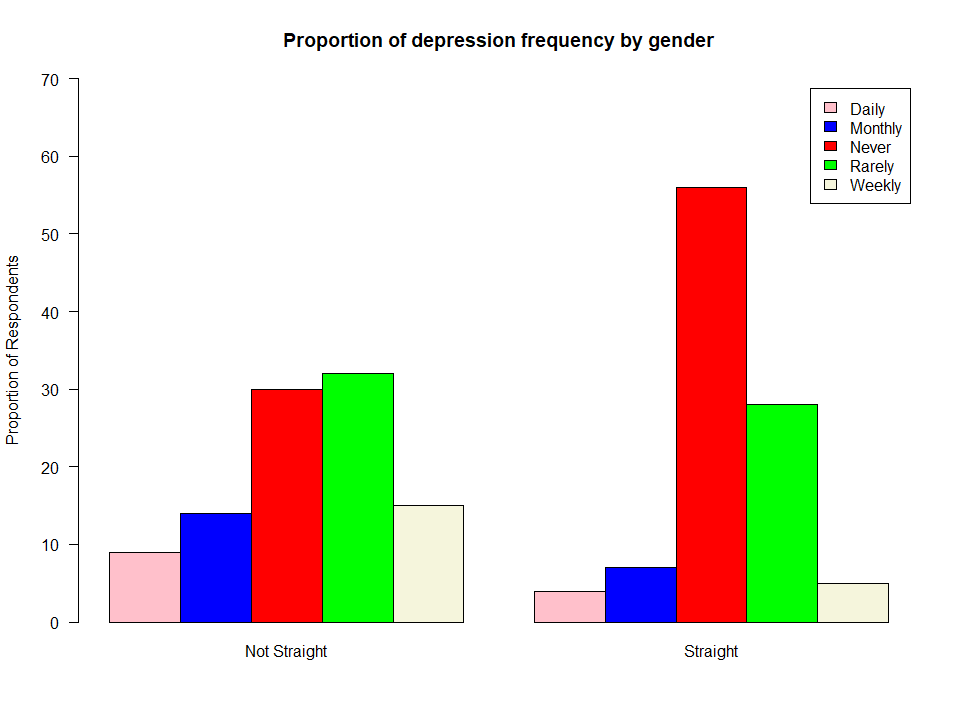
tab4 <- table(myHealthData$DepressionFrequency,myHealthData$Orientation); tab4

##   
## Not Straight Straight  
## Daily 313 3281  
## Monthly 493 5831  
## Never 1006 47099  
## Rarely 1079 23772  
## Weekly 513 4634

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Not Straight Straight  
## Daily 9 4  
## Monthly 14 7  
## Never 30 56  
## Rarely 32 28  
## Weekly 15 5

barplot(prop\_tab4,main="Proportion of depression frequency by gender",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","beige"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



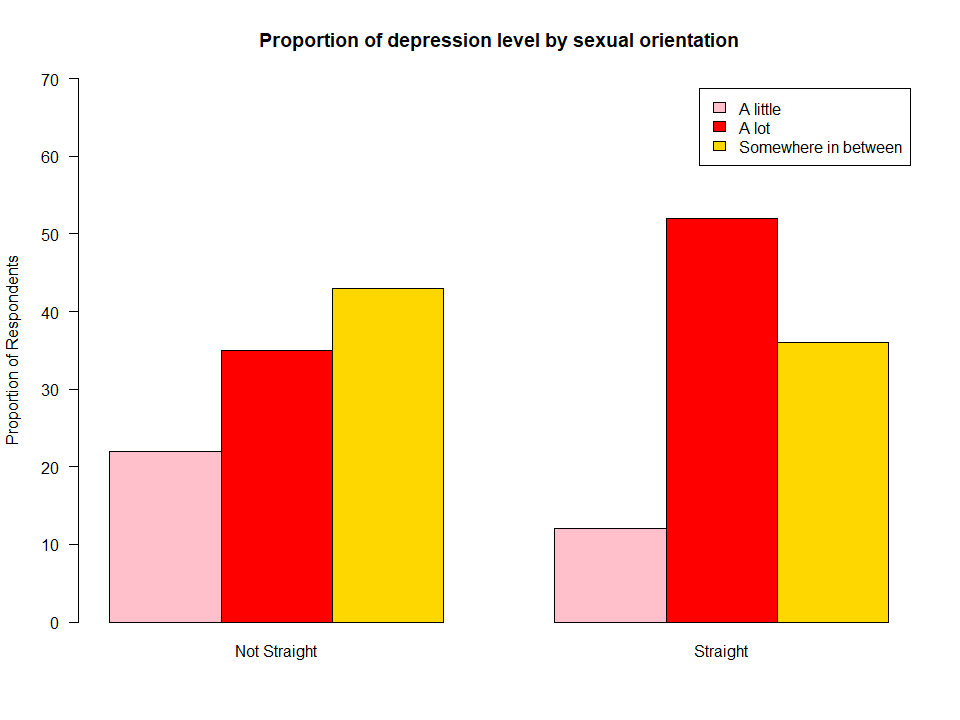
tab4 <- table(myHealthData$DepressionLevel,myHealthData$Orientation); tab4

##   
## Not Straight Straight  
## A little 530 4632  
## A lot 845 19929  
## Somewhere in between 1056 13653

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Not Straight Straight  
## A little 22 12  
## A lot 35 52  
## Somewhere in between 43 36

barplot(prop\_tab4,main="Proportion of depression level by sexual orientation",ylab="Proportion of Respondents", las=1, col=c("pink",'red',"gold"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



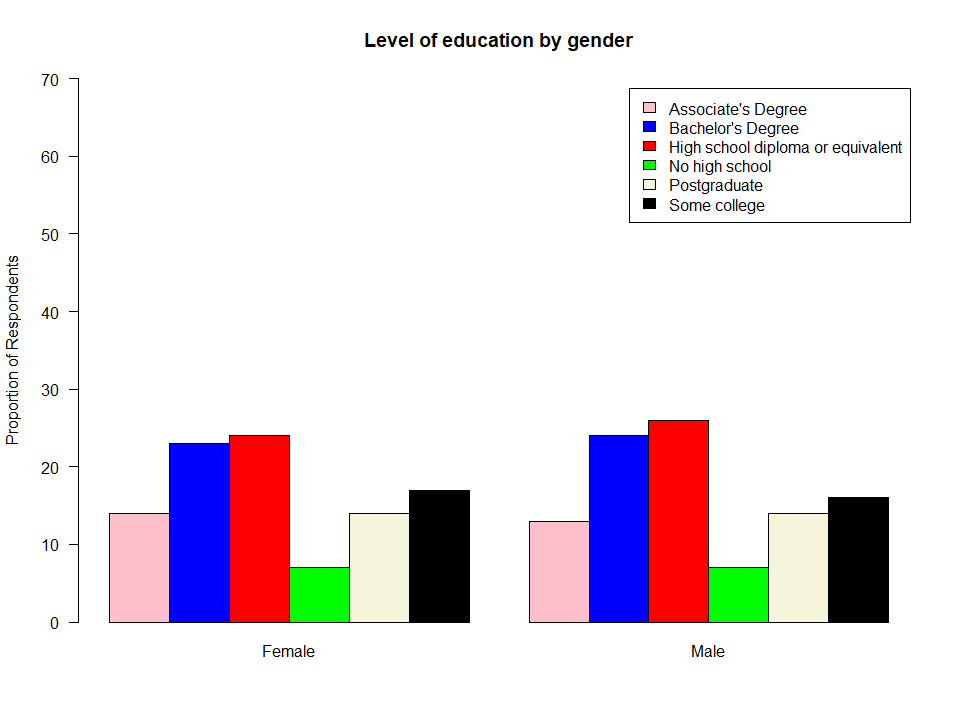
tab4 <- table(myHealthData$Education ,myHealthData$Gender); tab4

##   
## Female Male  
## Associate's Degree 6819 5147  
## Bachelor's Degree 11397 9874  
## High school diploma or equivalent 11855 10763  
## No high school 3481 2949  
## Postgraduate 6933 5591  
## Some college 8088 6376

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Female Male  
## Associate's Degree 14 13  
## Bachelor's Degree 23 24  
## High school diploma or equivalent 24 26  
## No high school 7 7  
## Postgraduate 14 14  
## Some college 17 16

barplot(prop\_tab4,main="Level of education by gender",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","beige","black"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



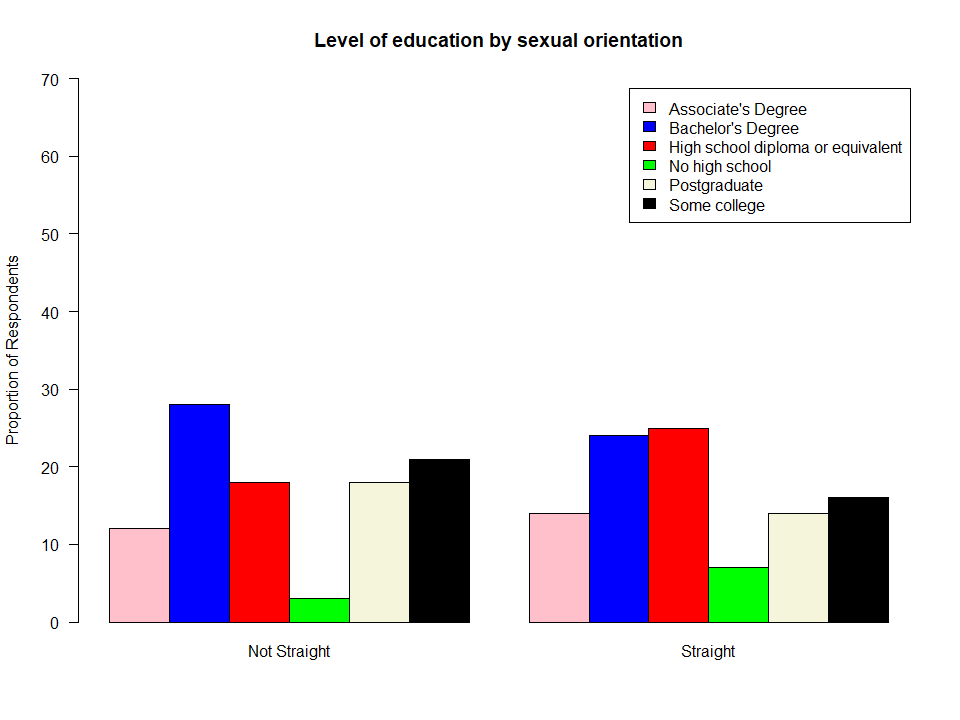
tab4 <- table(myHealthData$Education ,myHealthData$Orientation); tab4

##   
## Not Straight Straight  
## Associate's Degree 400 11149  
## Bachelor's Degree 934 19628  
## High school diploma or equivalent 600 21006  
## No high school 101 5940  
## Postgraduate 583 11577  
## Some college 690 13254

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Not Straight Straight  
## Associate's Degree 12 14  
## Bachelor's Degree 28 24  
## High school diploma or equivalent 18 25  
## No high school 3 7  
## Postgraduate 18 14  
## Some college 21 16

barplot(prop\_tab4,main="Level of education by sexual orientation",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","beige","black"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



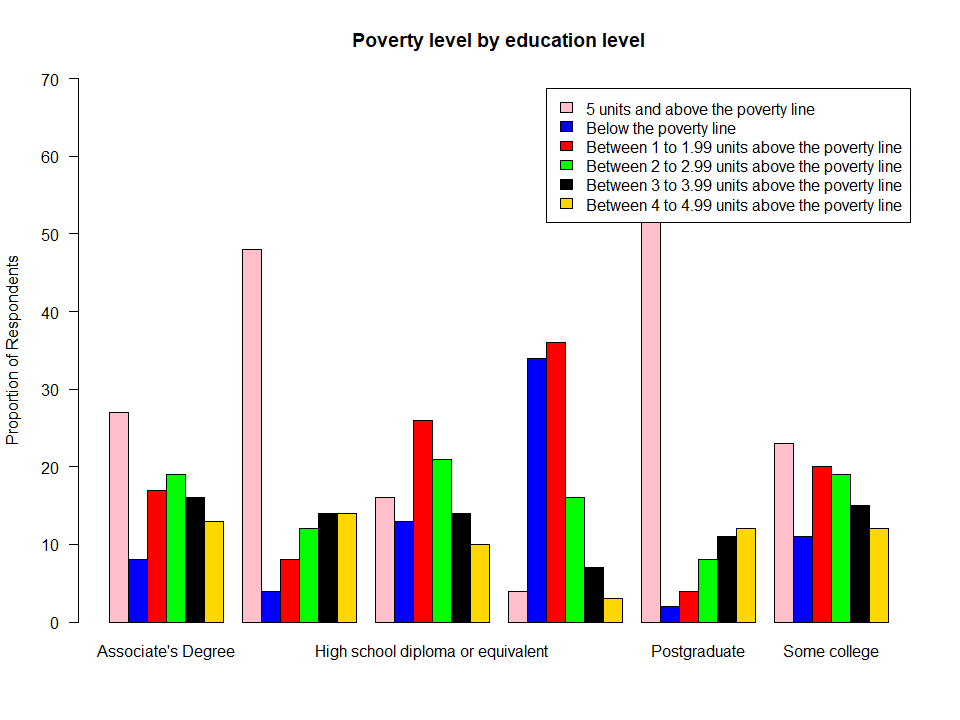
tab4 <- table(myHealthData$PovertyRatio ,myHealthData$Education); tab4

##   
## Associate's Degree  
## 5 units and above the poverty line 3198  
## Below the poverty line 941  
## Between 1 to 1.99 units above the poverty line 2066  
## Between 2 to 2.99 units above the poverty line 2234  
## Between 3 to 3.99 units above the poverty line 1957  
## Between 4 to 4.99 units above the poverty line 1570  
##   
## Bachelor's Degree  
## 5 units and above the poverty line 10175  
## Below the poverty line 817  
## Between 1 to 1.99 units above the poverty line 1775  
## Between 2 to 2.99 units above the poverty line 2643  
## Between 3 to 3.99 units above the poverty line 2962  
## Between 4 to 4.99 units above the poverty line 2902  
##   
## High school diploma or equivalent  
## 5 units and above the poverty line 3587  
## Below the poverty line 2993  
## Between 1 to 1.99 units above the poverty line 5844  
## Between 2 to 2.99 units above the poverty line 4743  
## Between 3 to 3.99 units above the poverty line 3189  
## Between 4 to 4.99 units above the poverty line 2264  
##   
## No high school Postgraduate  
## 5 units and above the poverty line 231 7917  
## Below the poverty line 2180 245  
## Between 1 to 1.99 units above the poverty line 2315 555  
## Between 2 to 2.99 units above the poverty line 1058 968  
## Between 3 to 3.99 units above the poverty line 429 1350  
## Between 4 to 4.99 units above the poverty line 217 1490  
##   
## Some college  
## 5 units and above the poverty line 3390  
## Below the poverty line 1553  
## Between 1 to 1.99 units above the poverty line 2829  
## Between 2 to 2.99 units above the poverty line 2792  
## Between 3 to 3.99 units above the poverty line 2163  
## Between 4 to 4.99 units above the poverty line 1737

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Associate's Degree  
## 5 units and above the poverty line 27  
## Below the poverty line 8  
## Between 1 to 1.99 units above the poverty line 17  
## Between 2 to 2.99 units above the poverty line 19  
## Between 3 to 3.99 units above the poverty line 16  
## Between 4 to 4.99 units above the poverty line 13  
##   
## Bachelor's Degree  
## 5 units and above the poverty line 48  
## Below the poverty line 4  
## Between 1 to 1.99 units above the poverty line 8  
## Between 2 to 2.99 units above the poverty line 12  
## Between 3 to 3.99 units above the poverty line 14  
## Between 4 to 4.99 units above the poverty line 14  
##   
## High school diploma or equivalent  
## 5 units and above the poverty line 16  
## Below the poverty line 13  
## Between 1 to 1.99 units above the poverty line 26  
## Between 2 to 2.99 units above the poverty line 21  
## Between 3 to 3.99 units above the poverty line 14  
## Between 4 to 4.99 units above the poverty line 10  
##   
## No high school Postgraduate  
## 5 units and above the poverty line 4 63  
## Below the poverty line 34 2  
## Between 1 to 1.99 units above the poverty line 36 4  
## Between 2 to 2.99 units above the poverty line 16 8  
## Between 3 to 3.99 units above the poverty line 7 11  
## Between 4 to 4.99 units above the poverty line 3 12  
##   
## Some college  
## 5 units and above the poverty line 23  
## Below the poverty line 11  
## Between 1 to 1.99 units above the poverty line 20  
## Between 2 to 2.99 units above the poverty line 19  
## Between 3 to 3.99 units above the poverty line 15  
## Between 4 to 4.99 units above the poverty line 12

barplot(prop\_tab4,main="Poverty level by education level",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","black","gold"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



I have learnt that the percentage of the population above working age that identifies as gay, bisexual. or lesbian is under 4%. This figure was unexpected. I also observed that more women than men identify as not straight. I also found that the proportion of people who identify as gay, lesbian or bisexual is fairly constant across all poverty levels. I observed that non-straight people seem to be have proportional or greater representation than straight people in the higher levels of education, such as postgraduate, bachelor’s degree and some college education. Also, it is very evident that the higher the level of education, the higher the person’s poverty level. These observations have interesting repercussions on my research questions.

### 7. Bivariate analysis (hypothesis tests and post-hoc tests)

chisq\_results <- chisq.test(myHealthData$DepressionFrequency, myHealthData$Gender)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$DepressionFrequency and myHealthData$Gender  
## X-squared = 1209.2, df = 4, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Gender  
## Female Male  
## Daily 2287 1462  
## Monthly 4179 2313  
## Never 24215 25150  
## Rarely 14967 10456  
## Weekly 3274 2037

chisq\_results$expected

## myHealthData$Gender  
## Female Male  
## Daily 2030.203 1718.797  
## Monthly 3515.626 2976.374  
## Never 26732.727 22632.273  
## Rarely 13767.368 11655.632  
## Weekly 2876.076 2434.924

chisq\_results$residuals

## myHealthData$Gender  
## Female Male  
## Daily 5.699273 -6.194082  
## Monthly 11.188126 -12.159477  
## Never -15.398803 16.735724  
## Rarely 10.224043 -11.111692  
## Weekly 7.419925 -8.064121

chisq\_results <- chisq.test(myHealthData$DepressionLevel, myHealthData$Gender)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$DepressionLevel and myHealthData$Gender  
## X-squared = 112.91, df = 2, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Gender  
## myHealthData$DepressionLevel Female Male  
## A little 3498 1839  
## A lot 12381 8865  
## Somewhere in between 9359 5765

chisq\_results$expected

## myHealthData$Gender  
## myHealthData$DepressionLevel Female Male  
## A little 3229.559 2107.441  
## A lot 12856.512 8389.488  
## Somewhere in between 9151.929 5972.071

chisq\_results$residuals

## myHealthData$Gender  
## myHealthData$DepressionLevel Female Male  
## A little 4.723650 -5.847518  
## A lot -4.193725 5.191511  
## Somewhere in between 2.164524 -2.679516

chisq\_results <- chisq.test(myHealthData$DepressionFrequency, myHealthData$Orientation)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$DepressionFrequency and myHealthData$Orientation  
## X-squared = 1427.4, df = 4, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Orientation  
## Not Straight Straight  
## Daily 313 3281  
## Monthly 493 5831  
## Never 1006 47099  
## Rarely 1079 23772  
## Weekly 513 4634

chisq\_results$expected

## myHealthData$Orientation  
## Not Straight Straight  
## Daily 138.9893 3455.011  
## Monthly 244.5655 6079.435  
## Never 1860.3449 46244.655  
## Rarely 961.0525 23889.947  
## Weekly 199.0478 4947.952

chisq\_results$residuals

## myHealthData$Orientation  
## Not Straight Straight  
## Daily 14.7599649 -2.9604061  
## Monthly 15.8859952 -3.1862540  
## Never -19.8078119 3.9728527  
## Rarely 3.8046534 -0.7630993  
## Weekly 22.2528067 -4.4632453

chisq\_results <- chisq.test(myHealthData$DepressionLevel, myHealthData$Orientation)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$DepressionLevel and myHealthData$Orientation  
## X-squared = 341.47, df = 2, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Orientation  
## myHealthData$DepressionLevel Not Straight Straight  
## A little 530 4632  
## A lot 845 19929  
## Somewhere in between 1056 13653

chisq\_results$expected

## myHealthData$Orientation  
## myHealthData$DepressionLevel Not Straight Straight  
## A little 308.7421 4853.258  
## A lot 1242.5045 19531.496  
## Somewhere in between 879.7535 13829.247

chisq\_results$residuals

## myHealthData$Orientation  
## myHealthData$DepressionLevel Not Straight Straight  
## A little 12.592179 -3.176012  
## A lot -11.276986 2.844293  
## Somewhere in between 5.942102 -1.498723

chisq\_results <- chisq.test(myHealthData$PovertyRatio, myHealthData$Gender)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$PovertyRatio and myHealthData$Gender  
## X-squared = 858.5, df = 5, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Gender  
## myHealthData$PovertyRatio Female Male  
## 5 units and above the poverty line 14400 15097  
## Below the poverty line 5768 3369  
## Between 1 to 1.99 units above the poverty line 9500 6450  
## Between 2 to 2.99 units above the poverty line 8286 6545  
## Between 3 to 3.99 units above the poverty line 6492 5820  
## Between 4 to 4.99 units above the poverty line 5486 4916

chisq\_results$expected

## myHealthData$Gender  
## myHealthData$PovertyRatio Female Male  
## 5 units and above the poverty line 15986.760 13510.240  
## Below the poverty line 4952.064 4184.936  
## Between 1 to 1.99 units above the poverty line 8644.568 7305.432  
## Between 2 to 2.99 units above the poverty line 8038.093 6792.907  
## Between 3 to 3.99 units above the poverty line 6672.848 5639.152  
## Between 4 to 4.99 units above the poverty line 5637.667 4764.333

chisq\_results$residuals

## myHealthData$Gender  
## myHealthData$PovertyRatio Female Male  
## 5 units and above the poverty line -12.549632 13.651478  
## Below the poverty line 11.594796 -12.612808  
## Between 1 to 1.99 units above the poverty line 9.200552 -10.008352  
## Between 2 to 2.99 units above the poverty line 2.765106 -3.007880  
## Between 3 to 3.99 units above the poverty line -2.213897 2.408275  
## Between 4 to 4.99 units above the poverty line -2.019959 2.197310

chisq\_results <- chisq.test(myHealthData$PovertyRatio, myHealthData$Orientation)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$PovertyRatio and myHealthData$Orientation  
## X-squared = 16.531, df = 5, p-value = 0.005481

chisq\_results$observed

## myHealthData$Orientation  
## myHealthData$PovertyRatio Not Straight Straight  
## 5 units and above the poverty line 1064 27477  
## Below the poverty line 392 8214  
## Between 1 to 1.99 units above the poverty line 590 14637  
## Between 2 to 2.99 units above the poverty line 571 13630  
## Between 3 to 3.99 units above the poverty line 440 11434  
## Between 4 to 4.99 units above the poverty line 361 9740

chisq\_results$expected

## myHealthData$Orientation  
## myHealthData$PovertyRatio Not Straight Straight  
## 5 units and above the poverty line 1101.6729 27439.327  
## Below the poverty line 332.1887 8273.811  
## Between 1 to 1.99 units above the poverty line 587.7570 14639.243  
## Between 2 to 2.99 units above the poverty line 548.1538 13652.846  
## Between 3 to 3.99 units above the poverty line 458.3324 11415.668  
## Between 4 to 4.99 units above the poverty line 389.8952 9711.105

chisq\_results$residuals

## myHealthData$Orientation  
## myHealthData$PovertyRatio Not Straight Straight  
## 5 units and above the poverty line -1.13501883 0.22742734  
## Below the poverty line 3.28164001 -0.65755268  
## Between 1 to 1.99 units above the poverty line 0.09251718 -0.01853796  
## Between 2 to 2.99 units above the poverty line 0.97580486 -0.19552513  
## Between 3 to 3.99 units above the poverty line -0.85630507 0.17158058  
## Between 4 to 4.99 units above the poverty line -1.46336088 0.29321829

chisq\_results <- chisq.test(myHealthData$PovertyRatio ,myHealthData$Education)  
chisq\_results

##   
## Pearson's Chi-squared test  
##   
## data: myHealthData$PovertyRatio and myHealthData$Education  
## X-squared = 21733, df = 25, p-value < 2.2e-16

chisq\_results$observed

## myHealthData$Education  
## myHealthData$PovertyRatio Associate's Degree  
## 5 units and above the poverty line 3198  
## Below the poverty line 941  
## Between 1 to 1.99 units above the poverty line 2066  
## Between 2 to 2.99 units above the poverty line 2234  
## Between 3 to 3.99 units above the poverty line 1957  
## Between 4 to 4.99 units above the poverty line 1570  
## myHealthData$Education  
## myHealthData$PovertyRatio Bachelor's Degree  
## 5 units and above the poverty line 10175  
## Below the poverty line 817  
## Between 1 to 1.99 units above the poverty line 1775  
## Between 2 to 2.99 units above the poverty line 2643  
## Between 3 to 3.99 units above the poverty line 2962  
## Between 4 to 4.99 units above the poverty line 2902  
## myHealthData$Education  
## myHealthData$PovertyRatio High school diploma or equivalent  
## 5 units and above the poverty line 3587  
## Below the poverty line 2993  
## Between 1 to 1.99 units above the poverty line 5844  
## Between 2 to 2.99 units above the poverty line 4743  
## Between 3 to 3.99 units above the poverty line 3189  
## Between 4 to 4.99 units above the poverty line 2264  
## myHealthData$Education  
## myHealthData$PovertyRatio No high school Postgraduate  
## 5 units and above the poverty line 231 7917  
## Below the poverty line 2180 245  
## Between 1 to 1.99 units above the poverty line 2315 555  
## Between 2 to 2.99 units above the poverty line 1058 968  
## Between 3 to 3.99 units above the poverty line 429 1350  
## Between 4 to 4.99 units above the poverty line 217 1490  
## myHealthData$Education  
## myHealthData$PovertyRatio Some college  
## 5 units and above the poverty line 3390  
## Below the poverty line 1553  
## Between 1 to 1.99 units above the poverty line 2829  
## Between 2 to 2.99 units above the poverty line 2792  
## Between 3 to 3.99 units above the poverty line 2163  
## Between 4 to 4.99 units above the poverty line 1737

chisq\_results$expected

## myHealthData$Education  
## myHealthData$PovertyRatio Associate's Degree  
## 5 units and above the poverty line 3819.566  
## Below the poverty line 1169.942  
## Between 1 to 1.99 units above the poverty line 2061.906  
## Between 2 to 2.99 units above the poverty line 1935.115  
## Between 3 to 3.99 units above the poverty line 1615.053  
## Between 4 to 4.99 units above the poverty line 1364.418  
## myHealthData$Education  
## myHealthData$PovertyRatio Bachelor's Degree  
## 5 units and above the poverty line 6790.695  
## Below the poverty line 2080.005  
## Between 1 to 1.99 units above the poverty line 3665.803  
## Between 2 to 2.99 units above the poverty line 3440.384  
## Between 3 to 3.99 units above the poverty line 2871.355  
## Between 4 to 4.99 units above the poverty line 2425.759  
## myHealthData$Education  
## myHealthData$PovertyRatio High school diploma or equivalent  
## 5 units and above the poverty line 7220.340  
## Below the poverty line 2211.606  
## Between 1 to 1.99 units above the poverty line 3897.737  
## Between 2 to 2.99 units above the poverty line 3658.056  
## Between 3 to 3.99 units above the poverty line 3053.025  
## Between 4 to 4.99 units above the poverty line 2579.236  
## myHealthData$Education  
## myHealthData$PovertyRatio No high school Postgraduate  
## 5 units and above the poverty line 2052.4663 3998.000  
## Below the poverty line 628.6749 1224.596  
## Between 1 to 1.99 units above the poverty line 1107.9775 2158.230  
## Between 2 to 2.99 units above the poverty line 1039.8452 2025.515  
## Between 3 to 3.99 units above the poverty line 867.8581 1690.501  
## Between 4 to 4.99 units above the poverty line 733.1780 1428.158  
## myHealthData$Education  
## myHealthData$PovertyRatio Some college  
## 5 units and above the poverty line 4616.932  
## Below the poverty line 1414.176  
## Between 1 to 1.99 units above the poverty line 2492.346  
## Between 2 to 2.99 units above the poverty line 2339.086  
## Between 3 to 3.99 units above the poverty line 1952.208  
## Between 4 to 4.99 units above the poverty line 1649.251

chisq\_results$residuals

## myHealthData$Education  
## myHealthData$PovertyRatio Associate's Degree  
## 5 units and above the poverty line -10.05727889  
## Below the poverty line -6.69333530  
## Between 1 to 1.99 units above the poverty line 0.09015046  
## Between 2 to 2.99 units above the poverty line 6.79440065  
## Between 3 to 3.99 units above the poverty line 8.50874822  
## Between 4 to 4.99 units above the poverty line 5.56559025  
## myHealthData$Education  
## myHealthData$PovertyRatio Bachelor's Degree  
## 5 units and above the poverty line 41.06883615  
## Below the poverty line -27.69317982  
## Between 1 to 1.99 units above the poverty line -31.22926064  
## Between 2 to 2.99 units above the poverty line -13.59452036  
## Between 3 to 3.99 units above the poverty line 1.69161129  
## Between 4 to 4.99 units above the poverty line 9.66948079  
## myHealthData$Education  
## myHealthData$PovertyRatio High school diploma or equivalent  
## 5 units and above the poverty line -42.75897057  
## Below the poverty line 16.61560405  
## Between 1 to 1.99 units above the poverty line 31.17419181  
## Between 2 to 2.99 units above the poverty line 17.93834031  
## Between 3 to 3.99 units above the poverty line 2.46090348  
## Between 4 to 4.99 units above the poverty line -6.20711955  
## myHealthData$Education  
## myHealthData$PovertyRatio No high school Postgraduate  
## 5 units and above the poverty line -40.20528313 61.98032566  
## Below the poverty line 61.87137011 -27.99307712  
## Between 1 to 1.99 units above the poverty line 36.26184651 -34.51015719  
## Between 2 to 2.99 units above the poverty line 0.56299822 -23.49734491  
## Between 3 to 3.99 units above the poverty line -14.89703131 -8.28153458  
## Between 4 to 4.99 units above the poverty line -19.06315532 1.63642781  
## myHealthData$Education  
## myHealthData$PovertyRatio Some college  
## 5 units and above the poverty line -18.05692335  
## Below the poverty line 3.69157304  
## Between 1 to 1.99 units above the poverty line 6.74340654  
## Between 2 to 2.99 units above the poverty line 9.36468058  
## Between 3 to 3.99 units above the poverty line 4.77079270  
## Between 4 to 4.99 units above the poverty line 2.16071020

To test the relationship between gender and depression levels which are both categorical variables, I conducted a Chi square test of independence. To conduct a Chi square test of independence, I followed these steps:

1. I stated the null and alternative hypotheses:

* Ho: There is no association between gender and depression levels.  
  H1: There is an association between gender and depression levels.

Summary of my samples:

Depression Frequency by Gender

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Daily (%) | Monthly (%) | Never (%) | Rarely (%) | Weekly (%) |
| Female | 61 | 64 | 49 | 59 | 62 |
| Male | 39 | 37 | 51 | 41 | 38 |

Depression Frequency by Gender (Actual Counts)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Daily** | **Monthly** | **Never** | **Rarely** | **Weekly** | **Total** |
| **Female** | 2287 | 4179 | 24215 | 14967 | 3274 | 48922 |
| **Male** | 1462 | 2313 | 25150 | 10456 | 2037 | 41418 |
| **Total** | 3749 | 6492 | 49365 | 25423 | 5311 | **90340** |

Depression Level of respondents by Gender

|  |  |  |  |
| --- | --- | --- | --- |
|  | A little (%) | A lot (%) | Somewhere in between (%) |
| Female | 66 | 58 | 62 |
| Male | 34 | 42 | 38 |

Depression Level of respondents by Gender (Actual Counts)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A little** | **A lot** | **Somewhere in between** | **Total** |
| **Female** | 3498 | 12381 | 9359 | 25238 |
| **Male** | 1839 | 8865 | 5765 | 16469 |
| **Total** | 5337 | 21246 | 15124 | **41707** |

//the counts in this table are less because I took out the respondents who said they had never felt depressed

The conditions for a Chi-Squared test are met because:

* Both gender and depression frequency / depression level are categorical variables.
* The cells in the contingency table are mutually exclusive, i.e. nobody reports that they experience two or more levels of depression, and nobody reports that they are both male and female.
* The expected value of all the cells are greater than 5, and no cell has an expected value less than 1.

Test Statistic (Chi-Square):

* Pearson's Chi-squared test  
    
  data: myHealthData$DepressionFrequency and myHealthData$Gender  
  X-squared = 1209.2, df = 4, p-value < 2.2e-16
* Pearson's Chi-squared test  
    
  data: myHealthData$DepressionLevel and myHealthData$Gender  
  X-squared = 112.91, df = 2, p-value < 2.2e-16
* The p-value of the test for the association between gender and depression frequency is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s gender and how frequently they experience depression.
* The p-value of the test for the association between gender and depression level is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s gender and how severe the person’s depression is.
* The type of error that could have been made in this test is a type I error, which occurs when the null hypothesis is rejected when it is actually true. This means that I could have falsely concluded that there is an association between gender and depression when there is none. The probability of making a type I error is equal to the significance level (alpha) for this test (0.05).

I repeated this test four (4) more times for different sets of hypothesis. These are:

**Set 2:**

To test the relationship between sexual orientation and depression levels which are both categorical variables, I conducted a Chi square test of independence. To conduct a Chi square test of independence, I followed these steps:

I stated the null and alternative hypotheses:

Ho: There is no association between sexual orientation and depression levels.

H1: There is an association between sexual orientation and depression levels.

Summary of my samples:

Frequency of depression by sexual orientation (actual count)  
  
 Not Straight Straight  
 Daily 313 3281  
 Monthly 493 5831  
 Never 1006 47099  
 Rarely 1079 23772  
 Weekly 513 4634

Level of depression by sexual orientation (actual count)  
  
 Not Straight Straight  
 A little 530 4632  
 A lot 845 19929  
 Somewhere in between 1056 13653

Frequency of depression by sexual orientation (proportions)   
   
 Not Straight Straight  
 Daily 9 4  
 Monthly 14 7  
 Never 30 56  
 Rarely 32 28  
 Weekly 15 5

Level of depression by sexual orientation (proportions)   
   
   
 Not Straight Straight  
 A little 22 12  
 A lot 35 52  
 Somewhere in between 43 36

The conditions for a Chi-Squared test are met because:

* Both sexual orientation and depression frequency / depression level are categorical variables.
* The cells in the contingency table are mutually exclusive, i.e. nobody reports that they experience two or more levels of depression, and nobody reports that they are both straight and not straight. Also, every person belongs to one category or the other.
* The expected value of all the cells are greater than 5, and no cell has an expected value less than 1.

Test Statistic (Chi-Square):

Pearson's Chi-squared test  
  
data: myHealthData$DepressionFrequency and myHealthData$Orientation  
X-squared = 1427.4, df = 4, p-value < 2.2e-16

* Pearson's Chi-squared test  
    
  data: myHealthData$DepressionLevel and myHealthData$Orientation  
  X-squared = 341.47, df = 2, p-value < 2.2e-16
* The p-value of the test of the association between sexual orientation and depression frequency is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s sexual orientation and how frequently they experience depression.
* The p-value of the test for the association between sexual orientation and depression level is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s sexual orientation and how severe the person’s depression is.
* The type of error that could have been made in this test is a type I error, which occurs when the null hypothesis is rejected when it is actually true. This means that I could have falsely concluded that there is an association between sexual orientation and depression when there is none. The probability of making a type I error is equal to the significance level (alpha) for this test (0.05).

**Set 3:**

To test the relationship between gender and poverty which are both categorical variables, I conducted a Chi square test of independence. To conduct a Chi square test of independence, I followed these steps:

I stated the null and alternative hypotheses:

Ho: There is no association between gender and poverty level.  
H1: There is an association between gender and poverty level.

Summary of my samples:

Distribution of different genders across different poverty levels (Actual counts)  
   
 Female Male  
 5 units and above the poverty line 14400 15097  
 Below the poverty line 5768 3369  
 Between 1 to 1.99 units above the poverty line 9500 6450  
 Between 2 to 2.99 units above the poverty line 8286 6545  
 Between 3 to 3.99 units above the poverty line 6492 5820  
 Between 4 to 4.99 units above the poverty line 5486 4916  
   
   
 Distribution of different genders across different poverty levels (proportions)  
   
 Female Male  
 5 units and above the poverty line 29 36  
 Below the poverty line 12 8  
 Between 1 to 1.99 units above the poverty line 19 15  
 Between 2 to 2.99 units above the poverty line 17 16  
 Between 3 to 3.99 units above the poverty line 13 14  
 Between 4 to 4.99 units above the poverty line 11 12

The conditions for a Chi-Squared test are met because:

* Both poverty level and gender are categorical variables.
* The cells in the contingency table are mutually exclusive, i.e. nobody is in more than one poverty level, and nobody reports that they are both male and female.
* The expected value of all the cells are greater than 5, and no cell has an expected value less than 1.

Test Statistic (Chi-Square):

Pearson's Chi-squared test  
  
data: myHealthData$PovertyRatio and myHealthData$Gender  
X-squared = 858.5, df = 5, p-value < 2.2e-16

The p-value of the test of the association between poverty level and gender is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s poverty level and their gender.

The type of error that could have been made in this test is a type I error, which occurs when the null hypothesis is rejected when it is actually true. This means that I could have falsely concluded that there is an association between gender and poverty ratio when there is none. The probability of making a type I error is equal to the significance level (alpha) for this test (0.05).

**Set 4:**

To test the relationship between sexual orientation and poverty which are both categorical variables, I conducted a Chi square test of independence. To conduct a Chi square test of independence, I followed these steps:

I stated the null and alternative hypotheses:

Ho: There is no association between sexual orientation and poverty level.  
H1: There is no association between sexual orientation and poverty level.

Summary of my samples:

:Distribution of different sexual orientations across different poverty levels (Actual counts)  
  
 Not Straight Straight  
 5 units and above the poverty line 1064 27477  
 Below the poverty line 392 8214  
 Between 1 to 1.99 units above the poverty line 590 14637  
 Between 2 to 2.99 units above the poverty line 571 13630  
 Between 3 to 3.99 units above the poverty line 440 11434  
 Between 4 to 4.99 units above the poverty line 361 9740

:Distribution of different sexual orientations across different poverty levels (proportions)   
 Not Straight Straight  
 5 units and above the poverty line 31 32  
 Below the poverty line 11 10  
 Between 1 to 1.99 units above the poverty line 17 17  
 Between 2 to 2.99 units above the poverty line 17 16  
 Between 3 to 3.99 units above the poverty line 13 13  
 Between 4 to 4.99 units above the poverty line 11 11

The conditions for a Chi-Squared test are met because:

* Both poverty level and sexual orientation are categorical variables.
* The cells in the contingency table are mutually exclusive, i.e. nobody is in more than one poverty level, and nobody reports that they are both male and female.
* The expected value of all the cells are greater than 5, and no cell has an expected value less than 1.

Test Statistic (Chi-Square):

Pearson's Chi-squared test  
  
data: myHealthData$PovertyRatio and myHealthData$Orientation  
X-squared = 16.531, df = 5, p-value = 0.005481

The p-value of the test of the association between poverty level and sexual orientation is significantly less than 0.05, so I have sufficient evidence to reject my null hypothesis. This implies that there is an association between a person’s poverty level and their sexual orientation.

The type of error that could have been made in this test is a type I error, which occurs when the null hypothesis is rejected when it is actually true. This means that I could have falsely concluded that there is an association between sexual orientation and poverty ratio when there is none. The probability of making a type I error is equal to the significance level (alpha) for this test (0.05).

### 8. Moderation

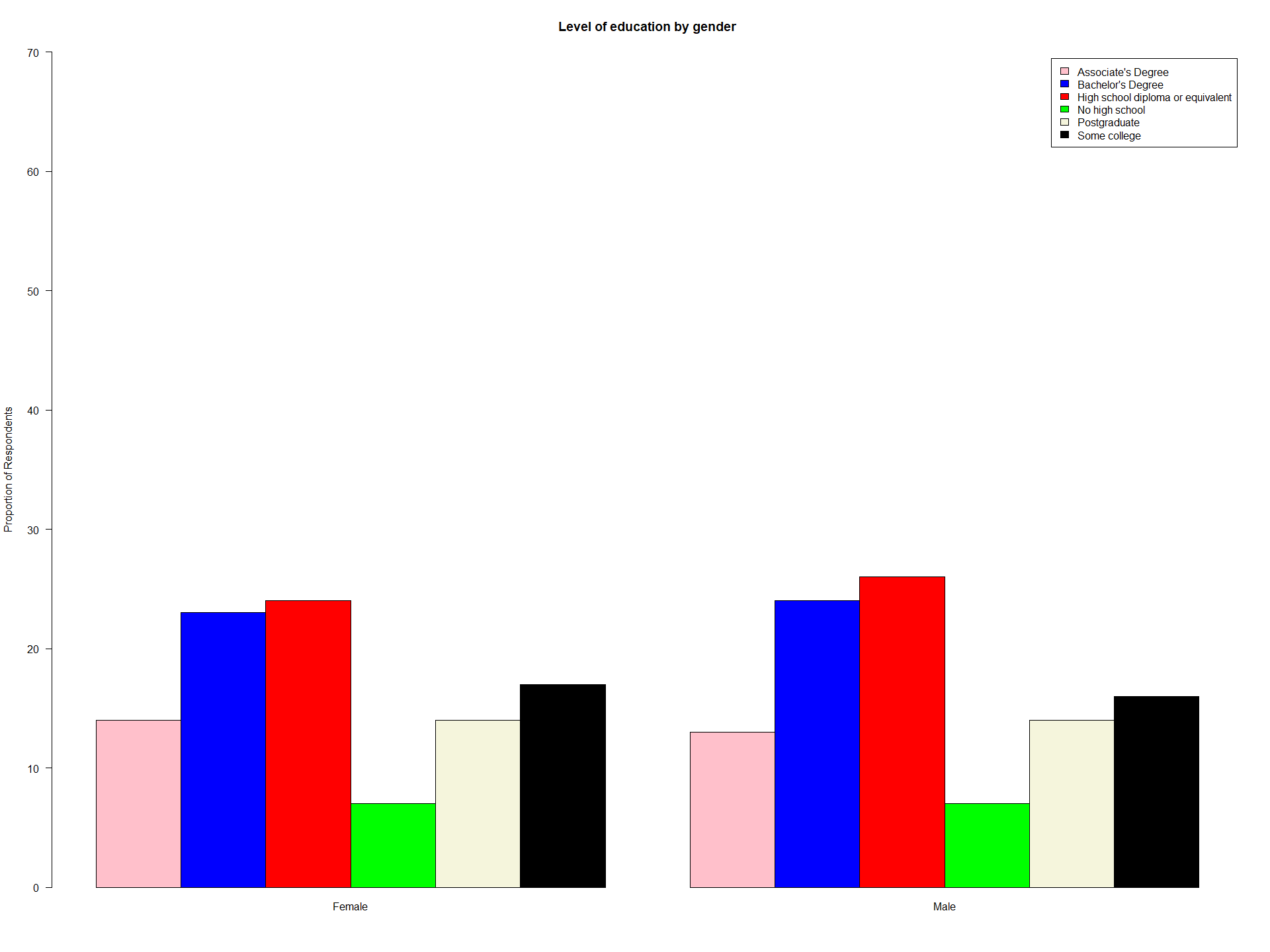
tab4 <- table(myHealthData$Education ,myHealthData$Gender); tab4

##   
## Female Male  
## Associate's Degree 6819 5147  
## Bachelor's Degree 11397 9874  
## High school diploma or equivalent 11855 10763  
## No high school 3481 2949  
## Postgraduate 6933 5591  
## Some college 8088 6376

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Female Male  
## Associate's Degree 14 13  
## Bachelor's Degree 23 24  
## High school diploma or equivalent 24 26  
## No high school 7 7  
## Postgraduate 14 14  
## Some college 17 16

barplot(prop\_tab4,main="Level of education by gender",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","beige","black"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



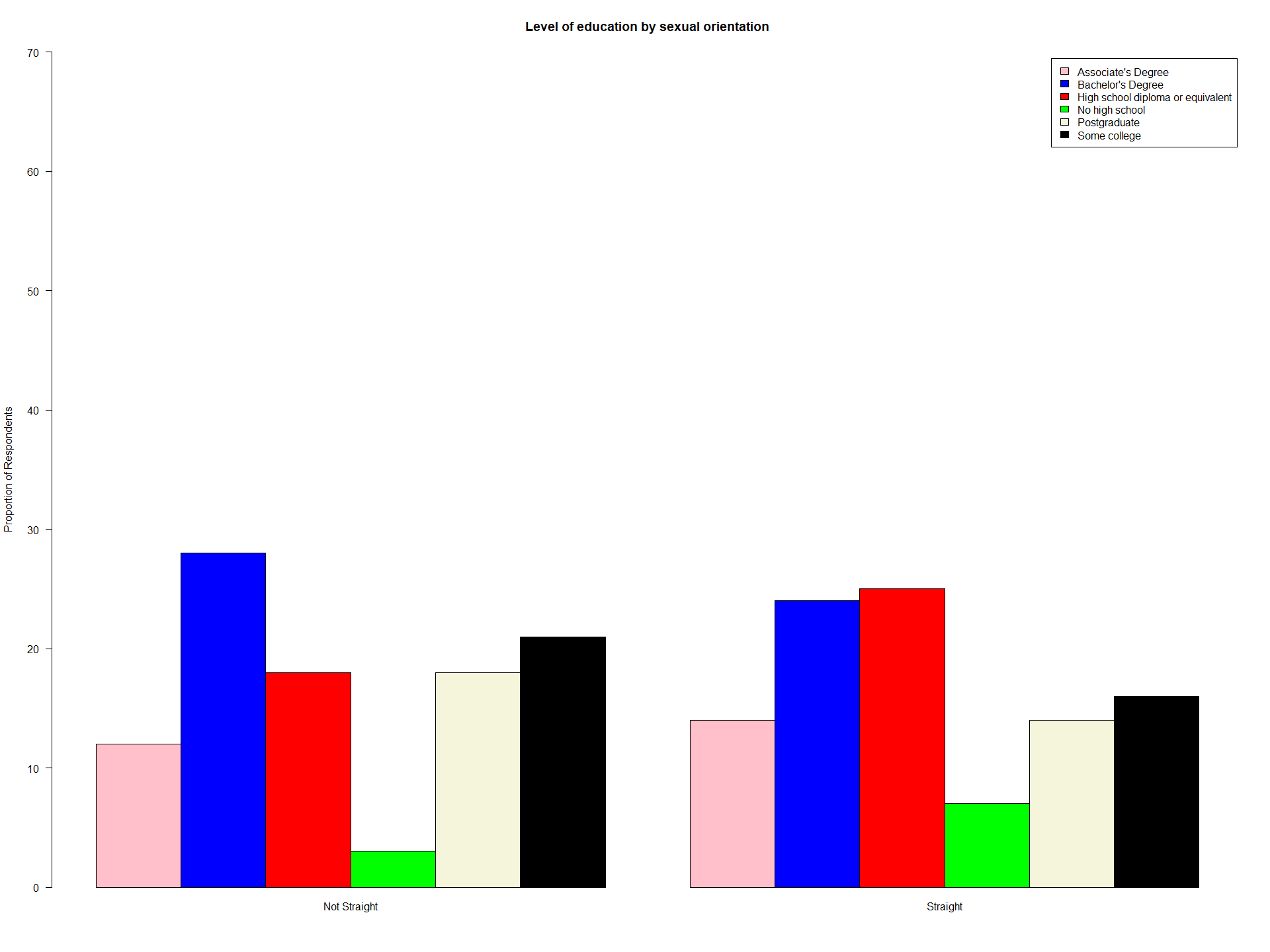
tab4 <- table(myHealthData$Education ,myHealthData$Orientation); tab4

##   
## Not Straight Straight  
## Associate's Degree 400 11149  
## Bachelor's Degree 934 19628  
## High school diploma or equivalent 600 21006  
## No high school 101 5940  
## Postgraduate 583 11577  
## Some college 690 13254

prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Not Straight Straight  
## Associate's Degree 12 14  
## Bachelor's Degree 28 24  
## High school diploma or equivalent 18 25  
## No high school 3 7  
## Postgraduate 18 14  
## Some college 21 16

barplot(prop\_tab4,main="Level of education by sexual orientation",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","beige","black"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



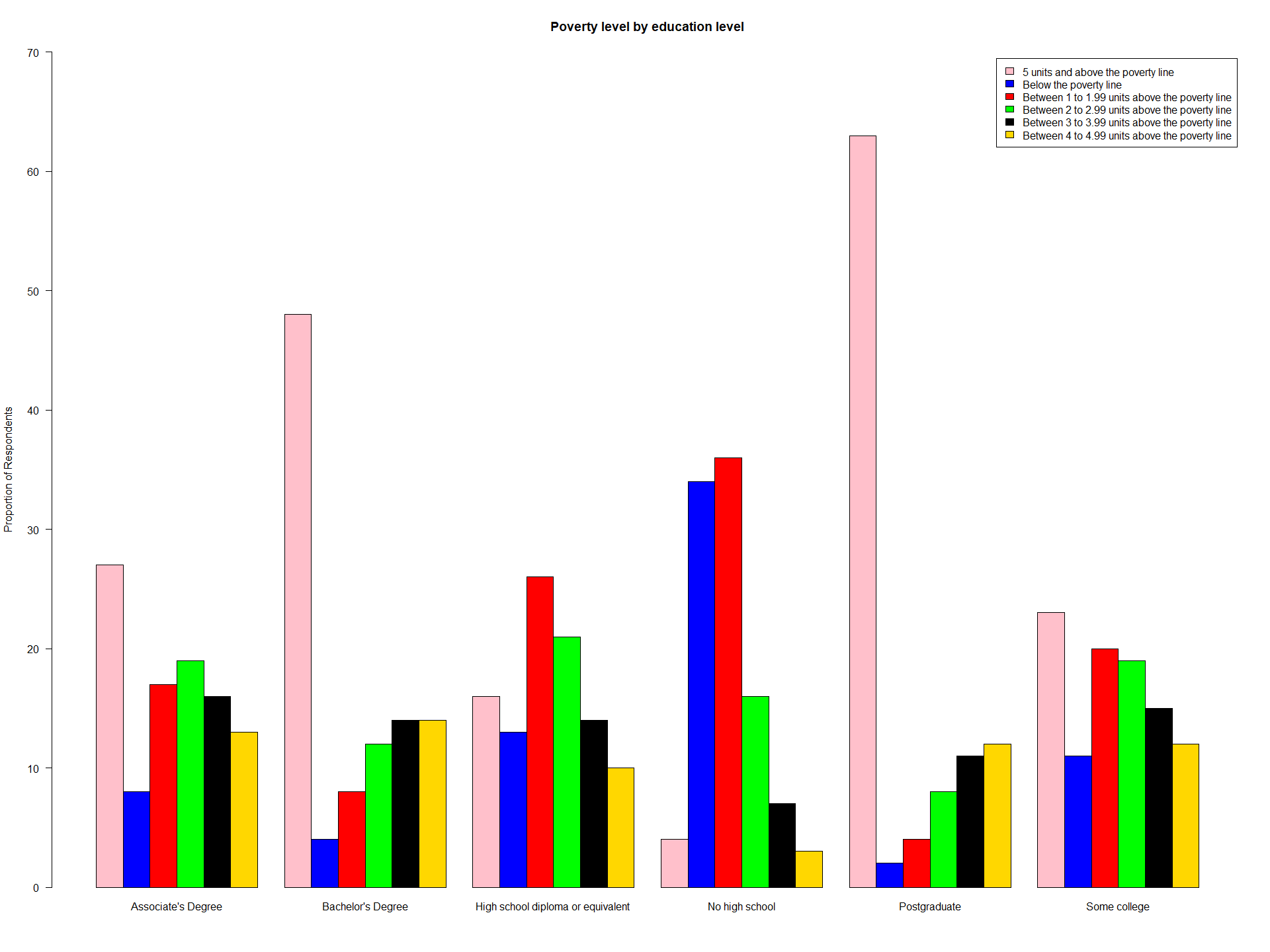
tab4 <- table(myHealthData$PovertyRatio ,myHealthData$Education); tab4

##   
## Associate's Degree  
## 5 units and above the poverty line 3198  
## Below the poverty line 941  
## Between 1 to 1.99 units above the poverty line 2066  
## Between 2 to 2.99 units above the poverty line 2234  
## Between 3 to 3.99 units above the poverty line 1957  
## Between 4 to 4.99 units above the poverty line 1570  
##   
## Bachelor's Degree  
## 5 units and above the poverty line 10175  
## Below the poverty line 817  
## Between 1 to 1.99 units above the poverty line 1775  
## Between 2 to 2.99 units above the poverty line 2643  
## Between 3 to 3.99 units above the poverty line 2962  
## Between 4 to 4.99 units above the poverty line 2902  
##   
## High school diploma or equivalent  
## 5 units and above the poverty line 3587  
## Below the poverty line 2993  
## Between 1 to 1.99 units above the poverty line 5844  
## Between 2 to 2.99 units above the poverty line 4743  
## Between 3 to 3.99 units above the poverty line 3189  
## Between 4 to 4.99 units above the poverty line 2264  
##   
## No high school Postgraduate  
## 5 units and above the poverty line 231 7917  
## Below the poverty line 2180 245  
## Between 1 to 1.99 units above the poverty line 2315 555  
## Between 2 to 2.99 units above the poverty line 1058 968  
## Between 3 to 3.99 units above the poverty line 429 1350  
## Between 4 to 4.99 units above the poverty line 217 1490  
##   
## Some college  
## 5 units and above the poverty line 3390  
## Below the poverty line 1553  
## Between 1 to 1.99 units above the poverty line 2829  
## Between 2 to 2.99 units above the poverty line 2792  
## Between 3 to 3.99 units above the poverty line 2163  
## Between 4 to 4.99 units above the poverty line 1737

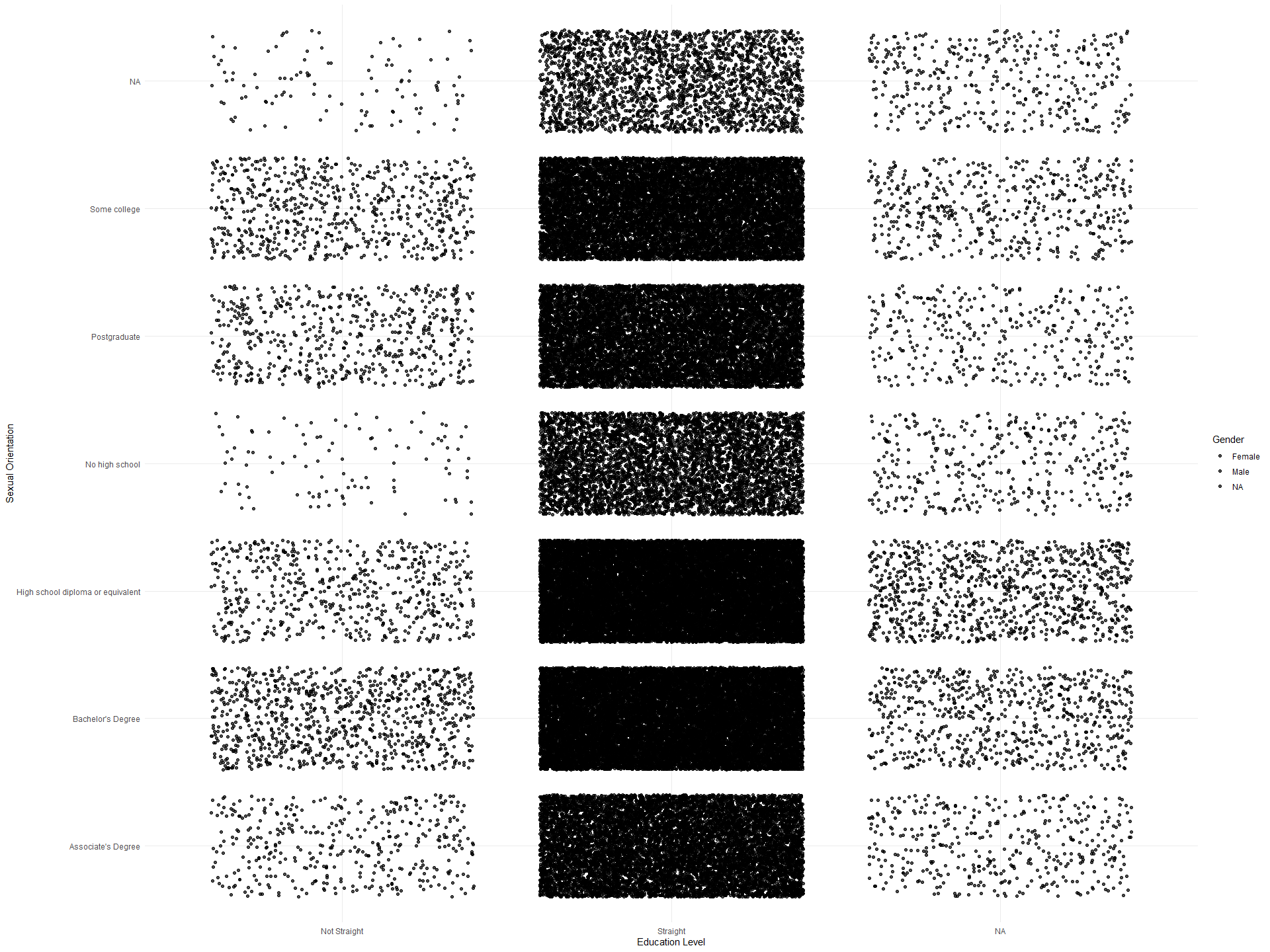
prop\_tab4 <- round(prop.table(tab4,2),2)\*100; prop\_tab4

##   
## Associate's Degree  
## 5 units and above the poverty line 27  
## Below the poverty line 8  
## Between 1 to 1.99 units above the poverty line 17  
## Between 2 to 2.99 units above the poverty line 19  
## Between 3 to 3.99 units above the poverty line 16  
## Between 4 to 4.99 units above the poverty line 13  
##   
## Bachelor's Degree  
## 5 units and above the poverty line 48  
## Below the poverty line 4  
## Between 1 to 1.99 units above the poverty line 8  
## Between 2 to 2.99 units above the poverty line 12  
## Between 3 to 3.99 units above the poverty line 14  
## Between 4 to 4.99 units above the poverty line 14  
##   
## High school diploma or equivalent  
## 5 units and above the poverty line 16  
## Below the poverty line 13  
## Between 1 to 1.99 units above the poverty line 26  
## Between 2 to 2.99 units above the poverty line 21  
## Between 3 to 3.99 units above the poverty line 14  
## Between 4 to 4.99 units above the poverty line 10  
##   
## No high school Postgraduate  
## 5 units and above the poverty line 4 63  
## Below the poverty line 34 2  
## Between 1 to 1.99 units above the poverty line 36 4  
## Between 2 to 2.99 units above the poverty line 16 8  
## Between 3 to 3.99 units above the poverty line 7 11  
## Between 4 to 4.99 units above the poverty line 3 12  
##   
## Some college  
## 5 units and above the poverty line 23  
## Below the poverty line 11  
## Between 1 to 1.99 units above the poverty line 20  
## Between 2 to 2.99 units above the poverty line 19  
## Between 3 to 3.99 units above the poverty line 15  
## Between 4 to 4.99 units above the poverty line 12

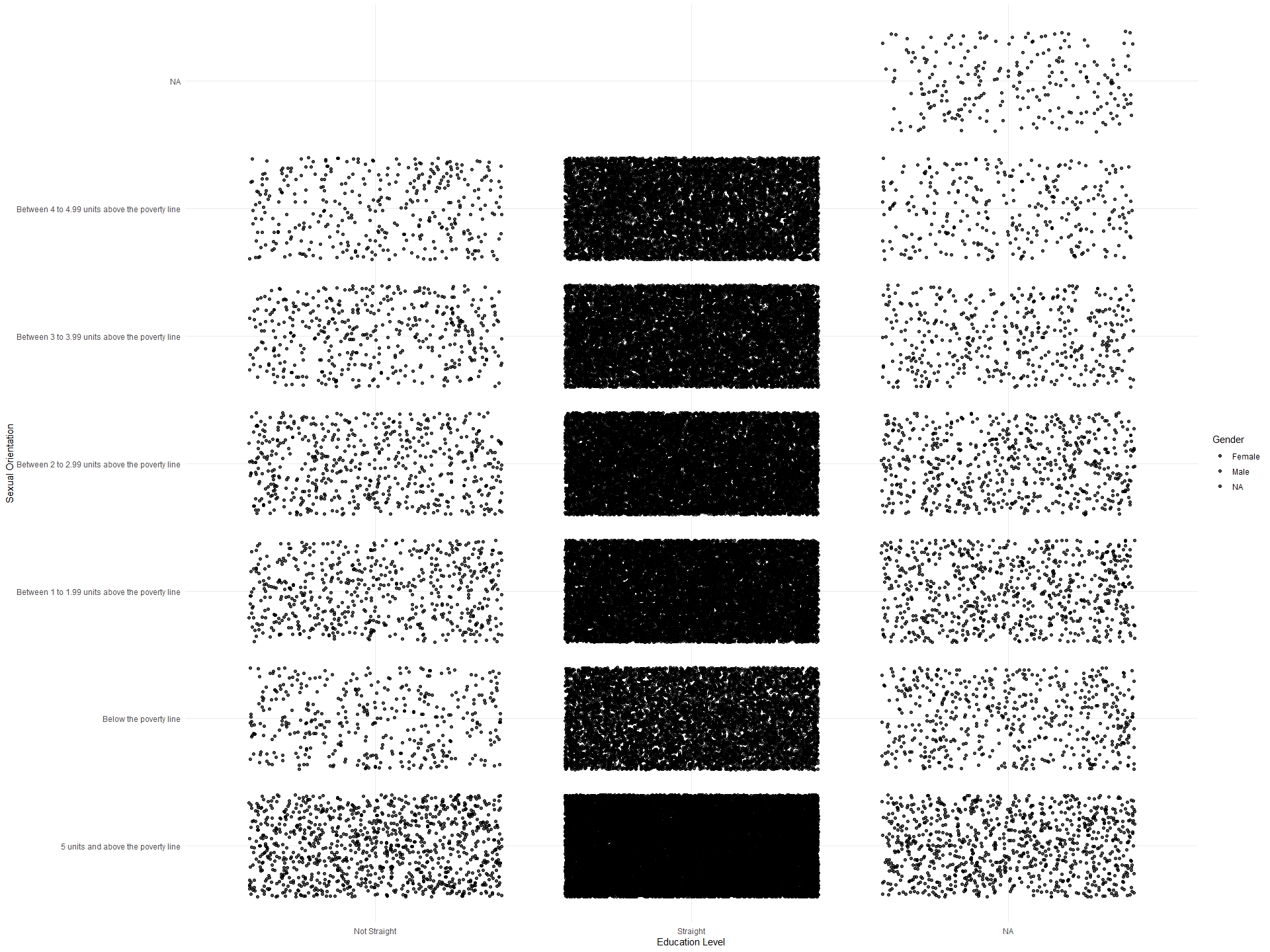
barplot(prop\_tab4,main="Poverty level by education level",ylab="Proportion of Respondents", las=1, col=c("pink","blue","red","green","black","gold"), beside = TRUE,legend=row.names(prop\_tab4),ylim=c(0,70))



#attempting to view the relationship between sexual orientation and education level, by gender (not working satisfactorily)  
  
myHealthData$Gender <- factor(myHealthData$Gender, levels = c("Female", "Male"))  
  
ggplot(myHealthData, aes(y = myHealthData$Education, x = myHealthData$Orientation, fill = myHealthData$Gender)) +  
 geom\_point(position = "jitter", alpha = 0.7) +  
 labs(x = "Education Level", y = "Sexual Orientation", fill = "Gender") +  
 scale\_fill\_manual(values = c("#E69F00", "#56B4E9","red")) +  
 theme\_minimal()



#attempting to view the relationship between sexual orientation and poverty level level, by gender (not working satisfactorily)  
  
ggplot(myHealthData, aes(y = myHealthData$PovertyRatio, x = myHealthData$Orientation, fill = myHealthData$Gender)) +  
 geom\_point(position = "jitter", alpha = 0.7) +  
 labs(x = "Education Level", y = "Sexual Orientation", fill = "Gender") +  
 scale\_fill\_manual(values = c("#E69F00", "#56B4E9","red")) +  
 theme\_minimal()



### 9. Save

save(myHealthData, file = "myHealthData.RDATA")

## Summary

The proportion of the population above working age that identifies as gay, bisexual. or lesbian is under 4%. More women than men identify as not straight. Non-straight people seem to be have proportional or greater representation than straight people in the higher levels of education, such as postgraduate, bachelor’s degree and some college education. Also, it is very evident that the higher the level of education, the better a person’s poverty level.

From the four Chi tests I carried out, I found that:

1. There is a relationship between gender and depression, as demonstrated by the tests of the association between gender and depression frequency, and gender and depression level. Men reported never feeling depressed more often than women. But among the men and women who reported feeling depressed, more men than women reported being depressed a lot.
2. There is a relationship between sexual orientation and depression, as demonstrated by the tests of the association between sexual orientation and depression frequency, and sexual orientation and depression level. A greater proportion of people who reported their depression level as “A little” or “Somewhere in between” were not straight.
3. There is an relationship between a person’s poverty level and their sexual orientation, as shown by the test. Fewer non-straight people are in the higher level above the poverty level.
4. There is an association between between a person’s sexual orientation and their educational attainment. A higher proportion of people with college education, some college, or postgraduate degree are not-straight.

There seems to be some credence to my research question, however, my findings now are generalized to non-straight people, instead of specific to men in this category. My findings need to incorporate this parameter to properly answer my research question. Education might have been a lurking variable, but that would be covered by implementing a 3-way categorical plot to examine the relationship between sexual orientation and education level, with gender as a control parameter. This is what I tried to carry out in my moderation chunk, but I need more guidance on how to do this.

The findings of my study so far could have important implications for policy, particularly in the areas of education and mental health. The study revealed that non-straight individuals are more represented in higher education, which suggests that policies aimed at promoting inclusive and supportive environments in education settings could have a positive impact on this group. Such policies might include the provision of safe spaces and support networks, as well as anti-bullying measures that are sensitive to issues around sexual orientation and gender identity.

Additionally, the study found that there is a relationship between sexual orientation and depression, with a greater proportion of non-straight individuals reporting experiencing depression. This suggests that policies aimed at addressing mental health issues among non-straight individuals are needed. Such policies might include targeted mental health support services, awareness campaigns that reduce stigma and promote positive mental health practices, and the development of appropriate screening tools that are sensitive to the unique needs of non-straight individuals.

Furthermore, the study suggests that non-straight individuals are more likely to experience poverty, which suggests that policies aimed at addressing poverty and income inequality could benefit this group. Such policies might include income support programs, access to affordable housing, and employment initiatives that promote equality of opportunity and support diversity in the workplace.

My findings of this study highlight the need for policy initiatives that are sensitive to the unique needs and experiences of non-straight individuals. By addressing issues related to education, mental health, and poverty, policymakers can work towards creating a more inclusive and equitable society for all individuals, regardless of their sexual orientation or gender identity.