Project Assignment 3

Elikem Asudo Tsatsu Gale-Zoyiku

2023-04-01

## **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

### 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

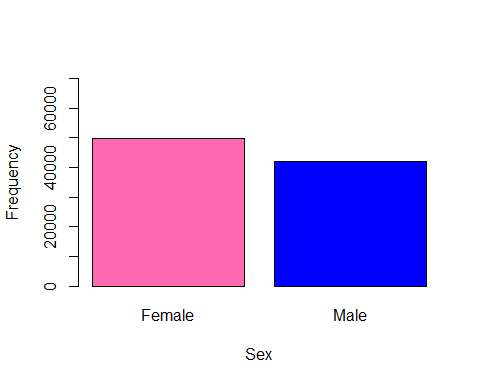
myHealthData$Gender <- rep(NA,nrow(myHealthData))  
myHealthData$Gender[myHealthData$SEX==1] <- 'Male'  
myHealthData$Gender[myHealthData$SEX==2] <- 'Female'  
  
  
myHealthData$WorkingAge <- rep(NA,nrow(myHealthData))  
myHealthData$WorkingAge[myHealthData$AGE<=18] <- 'Not working age'  
myHealthData$WorkingAge[myHealthData$AGE>18 & myHealthData$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] <- "Skipped"  
  
  
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] <- "Skipped"  
  
  
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", ]

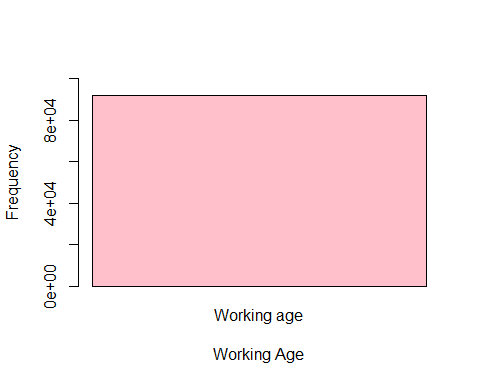
### 5. Descriptive statistics (sample means, standard deviations, proportions) and univariate displays

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



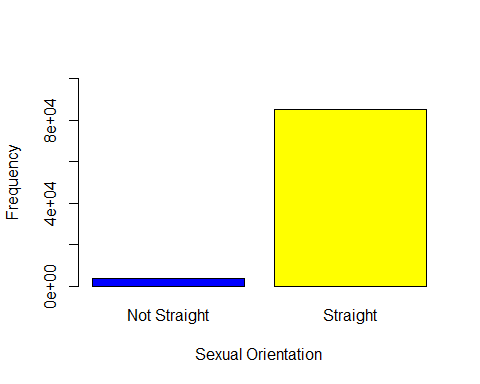
## myHealthData$Gender   
## Frequency Percent Valid Percent  
## Female 49932 54.0553 54.2  
## Male 42197 45.6816 45.8  
## NA's 243 0.2631   
## Total 92372 100.0000 100.0

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



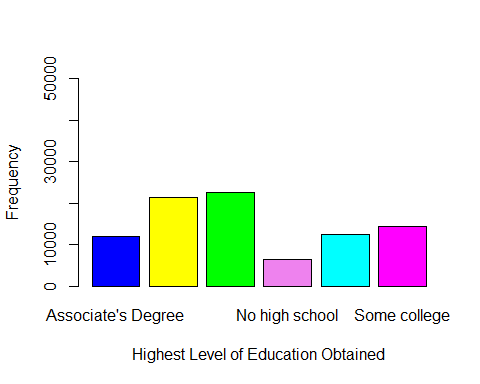
## myHealthData$WorkingAge   
## Frequency Percent Valid Percent  
## Working age 92135 99.7434 100  
## NA's 237 0.2566   
## Total 92372 100.0000 100

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



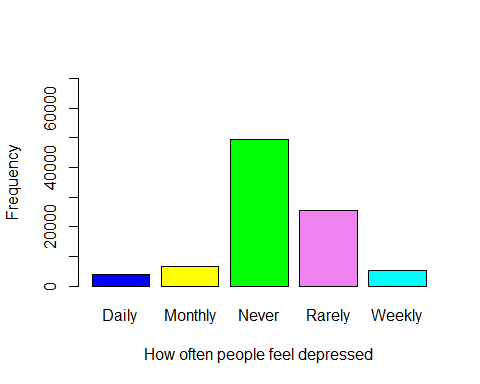
## myHealthData$Orientation   
## Frequency Percent Valid Percent  
## Not Straight 3418 3.700 3.86  
## Straight 85132 92.162 96.14  
## NA's 3822 4.138   
## Total 92372 100.000 100.00

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



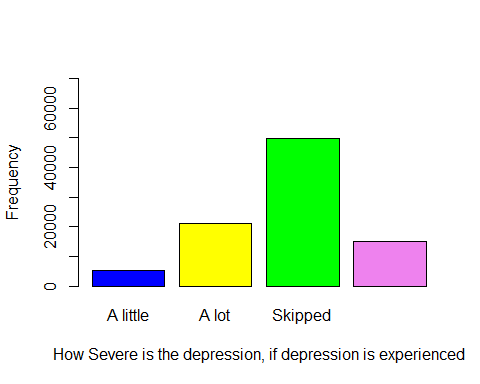
## myHealthData$Education   
## Frequency Percent Valid Percent  
## Associate's Degree 11966 12.954 13.403  
## Bachelor's Degree 21274 23.031 23.829  
## High school diploma or equivalent 22620 24.488 25.336  
## No high school 6430 6.961 7.202  
## Postgraduate 12525 13.559 14.029  
## Some college 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"))



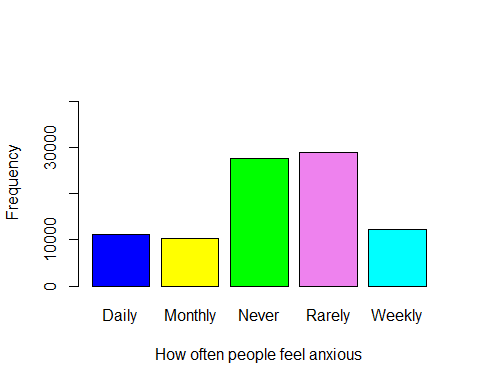
## myHealthData$DepressionFrequency   
## Frequency Percent Valid Percent  
## Daily 3750 4.060 4.151  
## Monthly 6494 7.030 7.188  
## Never 49367 53.444 54.642  
## Rarely 25423 27.522 28.140  
## Weekly 5312 5.751 5.880  
## NA's 2026 2.193   
## Total 92372 100.000 100.000

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"))



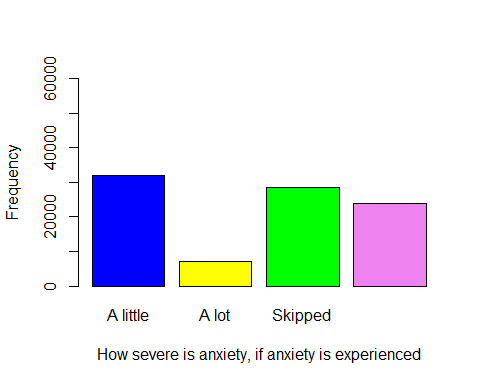
## myHealthData$DepressionLevel   
## Frequency Percent Valid Percent  
## A little 5338 5.7788 5.826  
## A lot 21247 23.0016 23.191  
## Skipped 49908 54.0294 54.473  
## Somewhere in between 15126 16.3751 16.510  
## NA's 753 0.8152   
## Total 92372 100.0000 100.000

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



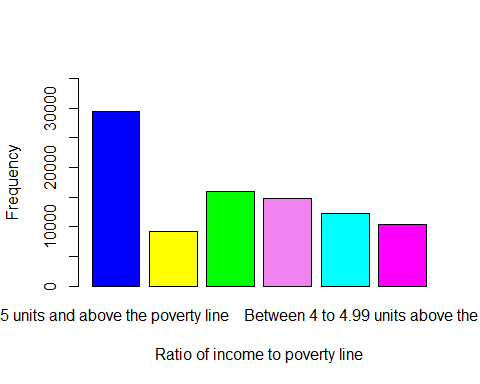
## myHealthData$WorryFrequency   
## Frequency Percent Valid Percent  
## Daily 11175 12.098 12.36  
## Monthly 10235 11.080 11.32  
## Never 27767 30.060 30.71  
## Rarely 29025 31.422 32.10  
## Weekly 12226 13.236 13.52  
## NA's 1944 2.105   
## Total 92372 100.000 100.00

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 34.5917 34.864  
## A lot 7187 7.7805 7.842  
## Skipped 28576 30.9358 31.179  
## Somewhere in between 23934 25.9104 26.115  
## NA's 722 0.7816   
## Total 92372 100.0000 100.000

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 Skipped   
## 5.826302 23.190605 54.473417   
## Somewhere in between   
## 16.509676

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

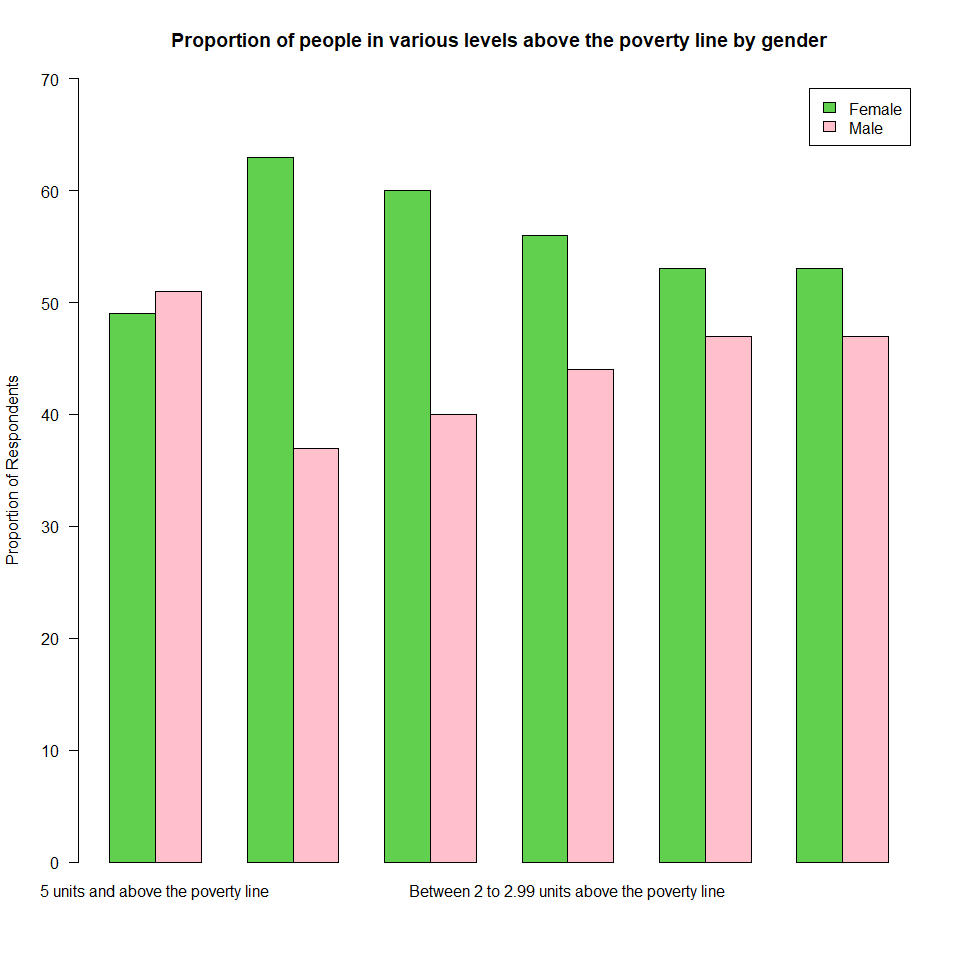
##   
## A little A lot Skipped   
## 34.864157 7.841789 31.179487   
## Somewhere in between   
## 26.114566

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

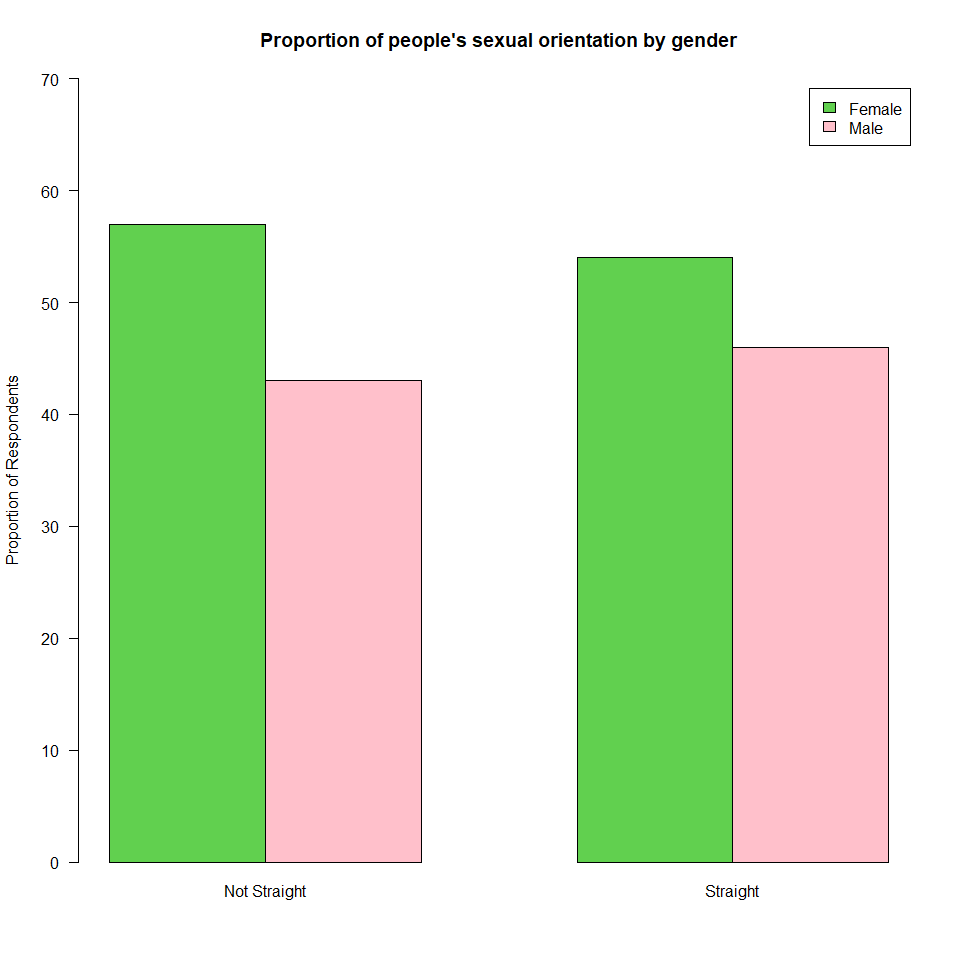
##   
## 5 units and above the poverty line   
## 32.01823   
## Below the poverty line   
## 9.91697   
## Between 1 to 1.99 units above the poverty line   
## 17.31372   
## Between 2 to 2.99 units above the poverty line   
## 16.09703   
## Between 3 to 3.99 units above the poverty line   
## 13.36300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

### 6. Bivariate tables and graphs

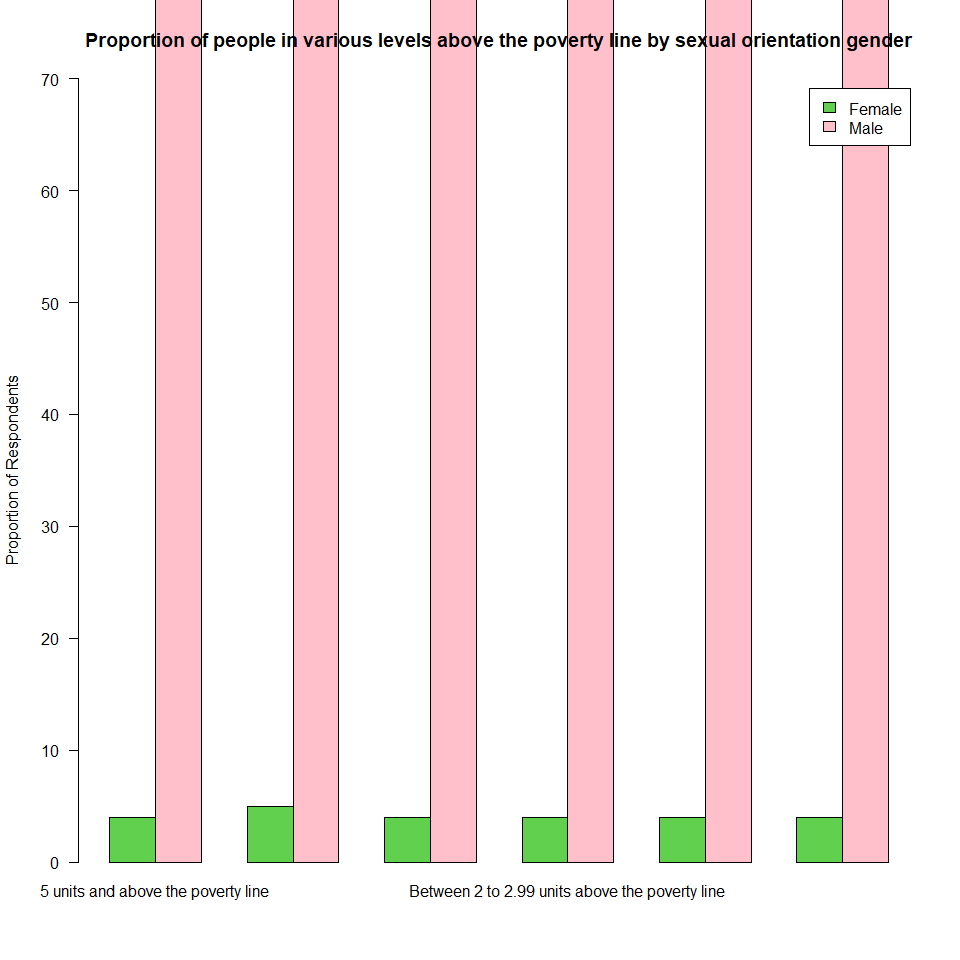
tab1<-table(myHealthData$Gender,myHealthData$PovertyRatio)  
prob\_tab1<-round(prop.table(tab1,2),2)\*100  
barplot(prob\_tab1,main="Proportion of people in various levels above the poverty line by gender",ylab="Proportion of Respondents", las=1, col=c(123,"pink"), beside = TRUE,legend=row.names(prob\_tab1),ylim=c(0,70))



tab2<-table(myHealthData$Orientation,myHealthData$PovertyRatio)  
prob\_tab2<-round(prop.table(tab2,2),2)\*100  
  
  
  
tab3<-table(myHealthData$Gender,myHealthData$Orientation)  
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(0,70))



barplot(prob\_tab2,main="Proportion of people in various levels above the poverty line by sexual orientation gender",ylab="Proportion of Respondents", las=1, col=c(123,"pink"), beside = TRUE,legend=row.names(prob\_tab3),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. These observations have interesting repercussions on my research questions. ### 7. Bivariate analysis (hypothesis tests and post-hoc tests)

### 8. Moderation

### 9. Save

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