Homework5

Lina Lee

March 11, 2019

# load libraries   
library(ggmosaic)

## Warning: package 'ggmosaic' was built under R version 3.5.2

## Loading required package: ggplot2

library(productplots)

## Warning: package 'productplots' was built under R version 3.5.2

##   
## Attaching package: 'productplots'

## The following objects are masked from 'package:ggmosaic':  
##   
## ddecker, happy, hbar, hspine, mosaic, prodcalc, spine, vbar,  
## vspine

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.2

## -- Attaching packages ------------------------------------------------------------- tidyverse 1.2.1 --

## v tibble 1.4.2 v purrr 0.2.5  
## v tidyr 0.8.2 v dplyr 0.7.7  
## v readr 1.1.1 v stringr 1.3.1  
## v tibble 1.4.2 v forcats 0.3.0

## Warning: package 'tidyr' was built under R version 3.5.2

## -- Conflicts ---------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(foreign) # package for reading "foreign" data formats (XPT/STATA/SPSS/dta/etc)

setwd(“C:/Users/linal/Desktop/2019 spring/visual/homwork/hw5”)

# reading in the data

full\_data <- foreign::read.xport(“LLCP2017.XPT”) head(full\_data) # YIKES!

# Subset the dataset with the following variables:

mydata <- full\_data %>% select(GENHLTH, IMONTH, SEX, PREGNANT, JOINPAI1, X\_BMI5CAT, X\_CHLDCNT, X\_EDUCAG)

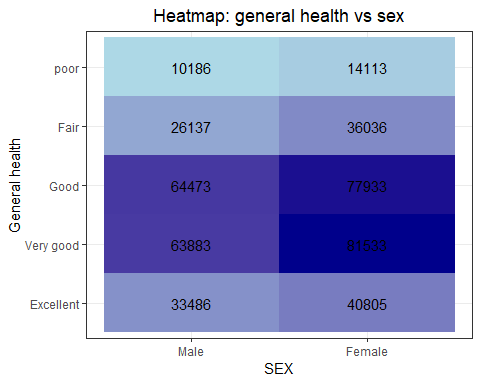
head(mydata)

remove(full\_data)

save (mydata, file = “mydata.RData”) #I saved mydata into “mydata.RData” and load it to avoid slow loading word document.

# 1) Heatmap: general health (GENHLTH) vs sex (SEX)

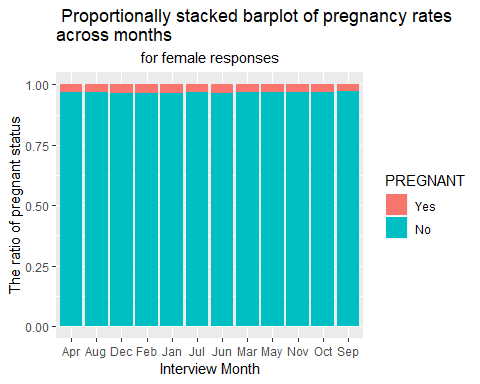
setwd("C:/Users/linal/Desktop/2019 spring/visual/homwork/hw5")  
load("mydata.RData")  
heatdata<-mydata %>%  
 select(GENHLTH,SEX) %>%  
 mutate(GENHLTH=factor(GENHLTH,levels =c(1,2,3,4,5),  
 labels=c("Excellent","Very good","Good","Fair","poor")),  
 SEX=factor(SEX,levels=c(1,2),labels=c("Male","Female"))) %>%  
 na.omit() %>%  
 group\_by(GENHLTH,SEX) %>%  
 summarise(Freq = n())  
  
  
ggplot()+  
 geom\_tile(aes(x=SEX,y=GENHLTH,fill=Freq)  
 ,heatdata) +  
 geom\_text(aes(x=SEX,y=GENHLTH,label=Freq),heatdata,color="black")+  
 scale\_fill\_gradient(low="lightblue",high="darkblue")+  
 theme\_bw()+  
 ggtitle("Heatmap: general health vs sex") +  
 xlab("SEX") + ylab("General health") +  
 theme(legend.position="none",plot.title = element\_text(hjust=0.5))



Description: For both sexes, more than 60000 persons think their general health is Good or Very good. Less than 15000 persons for each sex resported their general health is poor. Generally, General health of female is better than that of Male.

# 2) Proportionally stacked barplot of pregnancy (PREGNANT) rates across months (IMONTH) for female responses

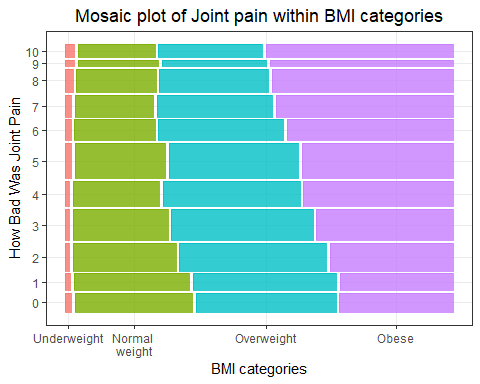
femaledata <-mydata %>%  
 select(PREGNANT,IMONTH,SEX) %>%  
 mutate(IMONTH =month.abb[IMONTH],  
 PREGNANT=factor(PREGNANT, levels=c(1,2),  
 labels=c("Yes","No")),  
 SEX=factor(SEX,levels=c(1,2),  
 labels=c("Male","Female"))) %>%  
 filter(SEX=="Female") %>%  
 select(PREGNANT,IMONTH) %>%  
 na.omit()   
  
  
ggplot() +  
 geom\_bar(aes(x=IMONTH, fill=PREGNANT), position="fill", data=femaledata) +  
 labs(x="Interview Month",y="The ratio of pregnant status",  
 title=" Proportionally stacked barplot of pregnancy rates \nacross months",   
 subtitle="for female responses") +  
 theme(plot.subtitle = element\_text(hjust=0.5))



Description: Across the interview months, less than 10 percent of female respondents reported being pregnant. The ratio of the female responspondent wh0 reported being pregnant is quite small across all interview months.

# 3) Mosaic plot: of joint pain (JOINPAI1) within BMI categories (X\_BMI5CAT)

mosaicdata <- mydata %>%   
 select(JOINPAI1,X\_BMI5CAT) %>%  
 filter(JOINPAI1<=10, X\_BMI5CAT<=4) %>%  
 mutate(JOINPAI1=factor(JOINPAI1,levels=c(0,1,2,3,4,5,6,7,8,9,10)),  
 X\_BMI5CAT=factor(X\_BMI5CAT,levels=c(1,2,3,4),labels=c("Underweight","Normal \nweight","Overweight","Obese"))) %>%  
 na.omit() %>%  
 group\_by(JOINPAI1,X\_BMI5CAT) %>%   
 summarise(Freq = n())  
  
  
  
ggplot()+  
 geom\_mosaic(aes(x=product(JOINPAI1),fill=X\_BMI5CAT,weight=Freq),  
 data=mosaicdata)+  
 theme\_bw()+  
 ggtitle("Mosaic plot of Joint pain within BMI categories") +  
 xlab("How Bad Was Joint Pain") + ylab("BMI categories") +  
 theme(legend.position="none",plot.title = element\_text(hjust=0.5))+  
 coord\_flip()



Description: There are four categories of BMI, and respondent reported how bad their joint pain on average during past 30days on 10 scale, where 0 is no pain or aching, and 10 indicates most severe. The severity of joint pain is different by BMI categories. For the underweight group, people reported feeling no pain or severe pain most frequently, but the differences among the proportion of response across the pain scales are not significant.

In normal weight, more respondent said they feel less pain. As the pain scale increases, the proportion of responses decreases in the normal weight group. The overweight group has a similar pattern with the normal weight group although the proportion of resposes for middle pain scale is not that smaller than lower pain scale, unlike the normal weight group.

However, the obese group has a quite different pattern from other groups. The proportion of responses increases as pain scale becomes higher. In conclusion, as the persons have a larger body mass index, they have a tendency to have more severe joint pain. I can guess that body weight might affect the severeness of joint pains.

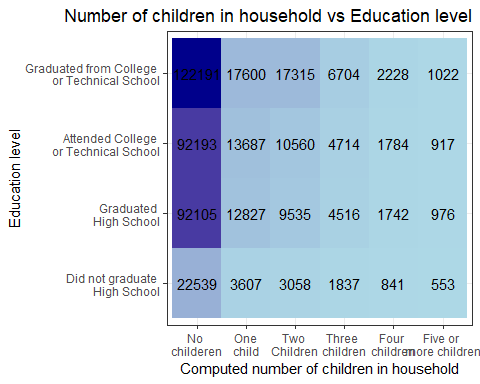
# 4) Your Choice of Plot: association between number of children (X\_CHLDCNT) and education (X\_EDUCAG)

choicedata<-mydata %>%  
filter(X\_CHLDCNT<=6, X\_EDUCAG<=4) %>%  
 mutate(X\_CHLDCNT=factor(X\_CHLDCNT,levels=c(1,2,3,4,5,6),  
 labels=c("No \nchilderen"," One \nchild","Two \nChildren",  
 "Three \nchildren","Four \nchildren",  
 "Five or \nmore children")),  
 X\_EDUCAG=factor(X\_EDUCAG,levels=c(1,2,3,4),  
 labels=c("Did not graduate \nHigh School",  
 "Graduated \nHigh School",  
 "Attended College \nor Technical School",  
 "Graduated from College \nor Technical School"))) %>%  
 group\_by(X\_CHLDCNT,X\_EDUCAG) %>%  
 summarize(Freq=n())

ggplot()+  
 geom\_tile(aes(x=X\_CHLDCNT,y=X\_EDUCAG,fill=Freq)  
 ,choicedata) +  
 scale\_fill\_gradient(low="lightblue",high="darkblue")+  
 theme\_bw()+  
 ggtitle("Number of children in household vs Education level") +  
 xlab("Computed number of children in household") + ylab("Education level")

+ geom\_text(aes(x=X\_CHLDCNT,y=X\_EDUCAG,label=Freq),choicedata,color="black")

+ theme(legend.position="none",plot.title = element\_text(hjust=0.5))



Description: For all groups by education levels, the frequency of the group who don’t have any child is the most among groups by the number of children in a household. The frequency of each group by the number of children decreases as the number of children in a household of which the group represents increases. The frequency of people of the group who has five or more than five children is the least.