Exploring the BRFSS data

Schwinn (Xuan) Chen

May 12, 2017

Part 1: Data

This Exploratory Data Analysis (EDA) discusses the following research questions, using data collected by the Behavioral Risk Factor Surveillance System (BRFSS) in 2013. BRFSS is an ongoing surveillance system designed to measure behavioral risk factors for the non-institutionalized adult population (18 years of age and older) residing in the US. The 2013 BRFSS data contains survey results from about 500000 respondents.

References BRFSS web site: http://www.cdc.gov/brfss/ BRFSS Questionnaire (Mandatory and Optional Modules): http://www.cdc.gov/brfss/questionnaires/pdf-ques/2013%20BRFSS_English.pdf BRFSS Codebook: http://www.cdc.gov/brfss/annual_data/2013/pdf/CODEBOOK13_LLCP.pdf BRFSS Guide to Calculated Variables: http://www.cdc.gov/brfss/annual_data/2013/pdf/2013_Calculated_Variables_Version15.pdf BRFSS Guide to Optional Modules Used, by State: http://apps.nccd.cdc.gov/BRFSSModules/ModByState.asp?Yr=2013

Based on how the data is collected, random sampling is not used in this report because when conducting the cellular telephone version of the BRFSS questionnaire, it is bias towards cellular phone users residing in private residences or college housing. All conclusion of this analysis are Exploratory Data Analysis of samples in BRFSS from 2013, which do not generalize the whole population. All Exploratory Data Analysis conducted in this report should be correlations rather than causalities because random assignment is not used. The surveys in this report are all observational, therefore, no causality will be drawn.

Load R packages

```
library(ggplot2)
library(dplyr)
```

Load data

```
load("brfss2013.RData")
```

Part 2: Research questions

Research question 1:

Does money buy happiness? Do money and happiness share a connection? It is said that money can't buy happiness, but numerous studies have found that there is a positive correlation between money and happiness. We will use our data to find out how respondents' income levels correlates to their mental health.

Research question 2:

Never ask a woman her weight! Are women more conscious about their body weight than men? There might be psychological reasons behind the fact that the majority of weight loss and diet advertisements are targeted towards women. We are curious to learn if the weight level of female respondents is different from that of males.

.

Research question 3:

Do tall people suffer more or less from heart disease? I have read controversial articles about how a person's height can affect his or her heart. Some articles claim that tall people are much more likely to suffer from heart attacks. Other say that being tall can give you a stronger heart. Let's find out from 2013's BRFSS.

Part 3: Exploratory data analysis

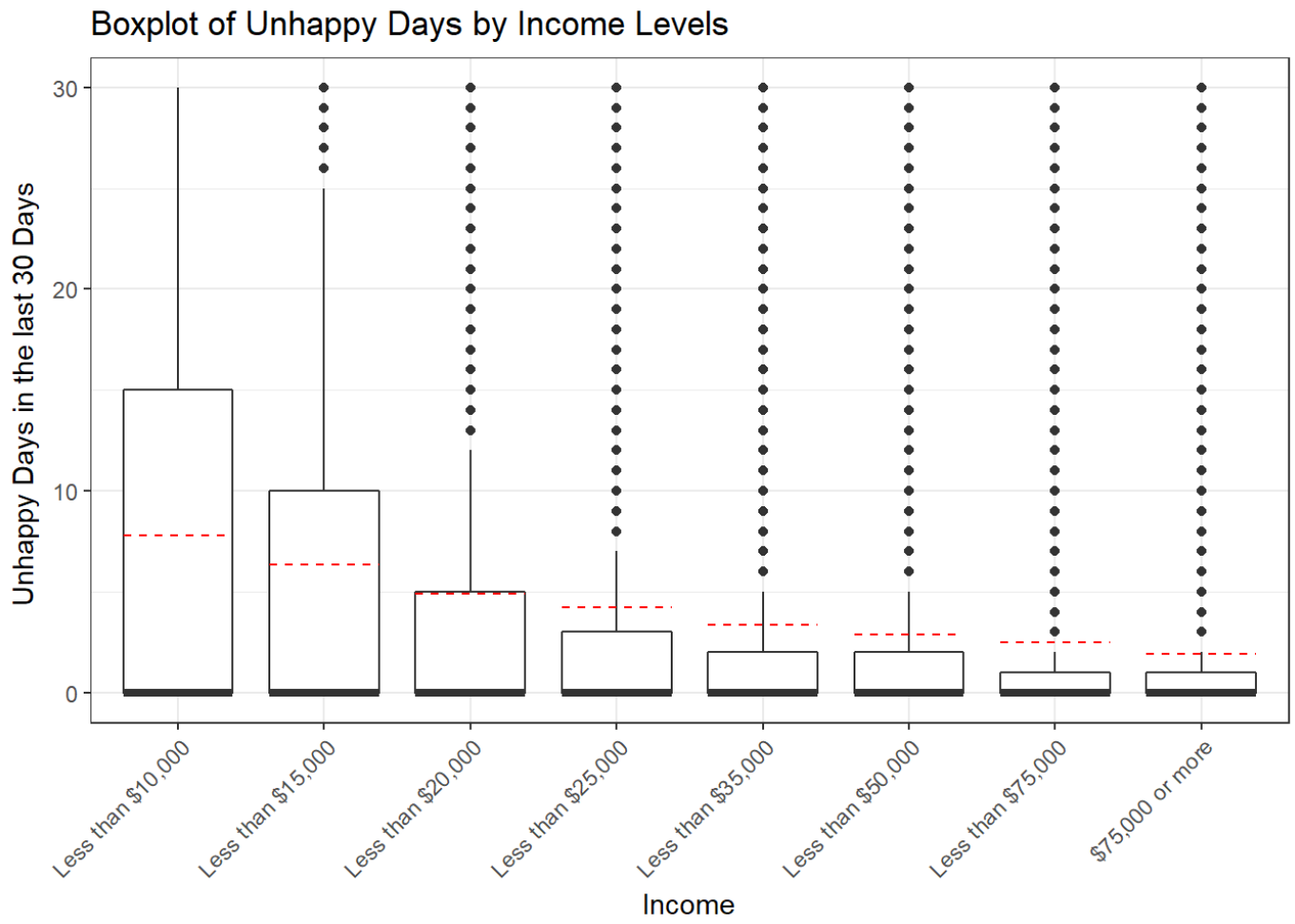
Research question 1:

** Does money buy happiness? Are money and happiness correlated?**

To answer this question, we found the following variables in the survey brfss2013 related to our research:

1. income2: Income Level This variable is directly relate to money.
2. menthlth: Number Of Days Mental Health Not Good during the past 30 days This variable shows the short-term happiness of a person (within the last 30 days)
3. addepev2: Ever Told You Had A Depressive Disorder This variable shows more of a long term happiness.

```
# data.frame titled inc_menth that has columns of income2, menthlth, and addepev2.
Remove all the NA data.

inc_menth<-brfss2013 %>%
filter(!is.na(income2),!is.na(menthlth), !is.na(addepev2)) %>%
select(income2,menthlth,addepev2)


#plot a series of boxplots that shows distribution of menthlth (number of unhappy d
ays in the past 30 days) in all income ranges.

mental_hplot<-ggplot(inc_menth, aes(x=income2,y=menthlth))+theme_bw()+theme(axis.te
xt.x = element_text(angle=45,hjust=1))+geom_boxplot(fatten=3)+stat_summary(fun.y=me
an,geom='errorbar',aes(ymax=..y..,ymin=..y..),width=0.75,linetype='dashed', color='
red')+labs(x='Income', y='Unhappy Days in the last 30 Days ', title='Boxplot of Unh
appy Days by Income Levels')
mental_hplot
```



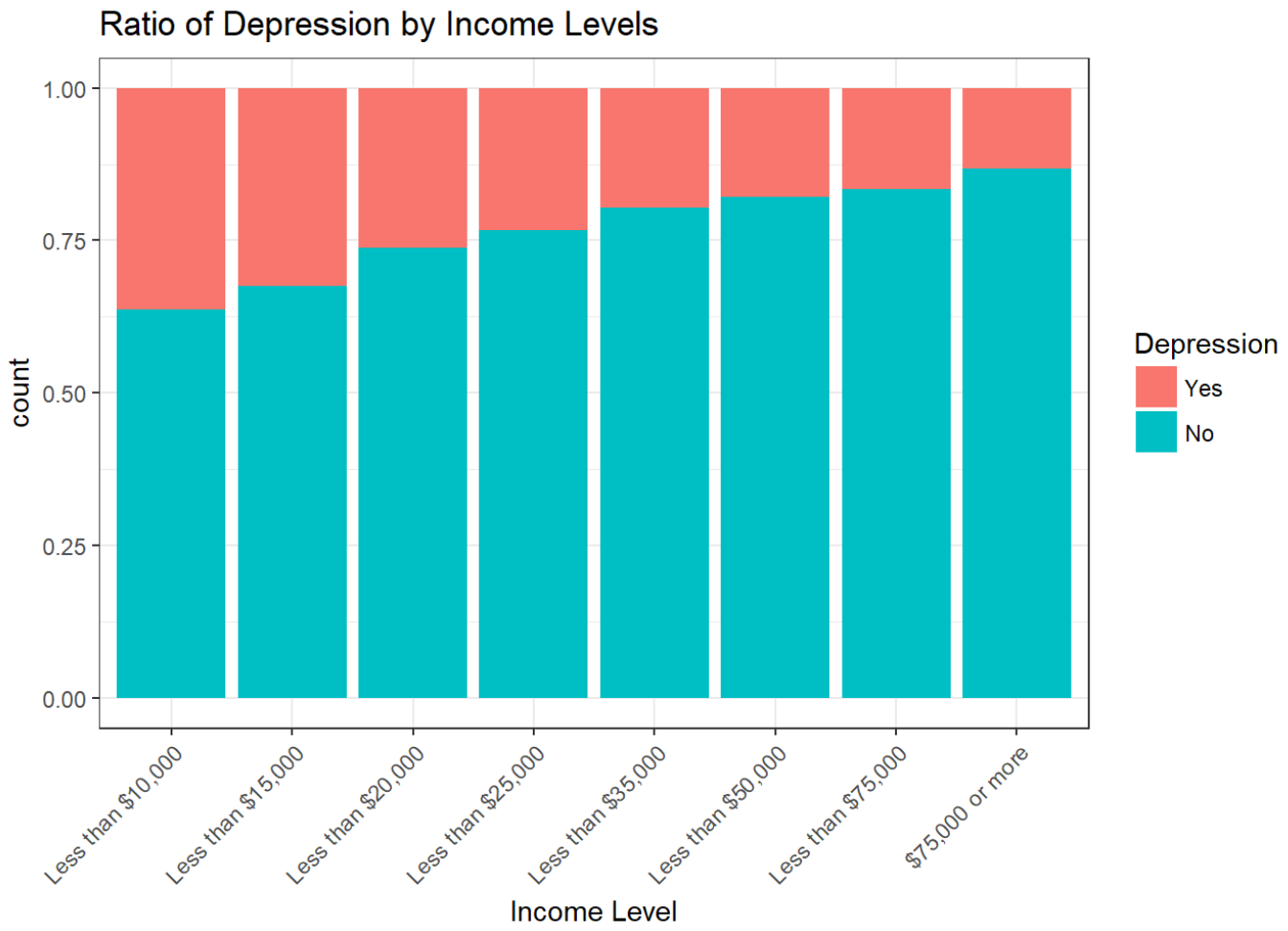
From the side-by-side boxplots, we can find the following conclusions:

1. The median of unhappy days (indicated by the thick black lines at the bottom of each box plot) is zero at all income levels.
2. The mean of unhappy days (indicated by the red dash lines) decreases as the income increases, indicating the higher the income, the less the average unhappy days in a month. In all income ranges, the mean is larger than the median, indicating that the distribution of unhappy days is right skewed.
3. The interquartile range (IQR) of unhappy days decreases as income increases, indicating that respondents with lower incomes have a wider range of unhappy days than those with higher incomes.

Based on this graph, we can say that there is a positive correlation between happiness and income levels in short-term (past 30 days).

Next we will find out the long term correlation between money and happiness. We will create a filled bar graph to as we are more interested in analyze the ratio of respondents with depression in each income range, therefore, a filled bar chart is the best choice in this situation.

```
# create a masaic chart that shows the ratio of depression with respect to income l
evels.
#Depression is marked red and without depression marked in blue

ggplot(inc_menth, aes(x=income2,fill=addepev2))+geom_bar(position = 'fill')+theme_b
w()+theme(axis.text.x = element_text(angle=45,hjust=1))+scale_fill_discrete(name="D
epression")+labs(x='Income Level', title='Ratio of Depression by Income Levels')
```



From this filled bar chart, we can tell that the proportion of depression to non-depression decreases as income increases, indicating the higher the income, the less respondents have been diagnosed with depression, thus, income levels and happiness shows a positive correlation in the long term as well.

The following table shows the income level, mean, median, standard division(SD) and IQR of Unhappiness During the Last 30 Days, as well as the percentage (yes_p) of diagnosed depression in the samples.

```
#summarise a table inc_ment_sum that shows the income level,  mean, median, standar
d division(SD) and IQR of Unhappiness During the Last 30 Days, as well as the perce
ntage (yes_p) of diagnosed depression in the samples.
 inc_ment_sum<-brfss2013 %>%
filter(!is.na(income2),!is.na(menthlth), !is.na(addepev2)) %>%
group_by(income2) %>%
 summarise(mean=mean(menthlth),median=median(menthlth),sd=sd(menthlth),iqr=IQR(ment
hlth),yes_p=100*sum(addepev2=='Yes')/n())
inc_ment_sum
```

```
## # A tibble: 8 x 6
##   income2             mean median    sd   iqr yes_p
##   <fct>              <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Less than $10,000  7.80       0 11.2     15  36.4
## 2 Less than $15,000  6.32       0 10.3     10  32.5
## 3 Less than $20,000  4.90       0  9.17     5  26.2
## 4 Less than $25,000  4.22       0  8.59     3  23.2
## 5 Less than $35,000  3.34       0  7.61     2  19.6
## 6 Less than $50,000  2.85       0  6.95     2  17.9
## 7 Less than $75,000  2.48       0  6.32     1  16.6
## 8 $75,000 or more    1.93       0  5.42     1  13.3
```

As we can see from this table, those making less than $10,000 have an average of 7.8 unhappy days in the last month, the highest in all income ranges, compared to those earning more than $75,000, which have the lowest average of 1.93 unhappy days.

In a similar fashion, the lowest income group also have the highest IQR and standard division of unhappy days. The lowest income groups (making less than $15,000) have a much larger IQR than the rest of the group, indicating a wider range of unhappy days.

More than 30% of samples have been diagnosed with depression when the income is less than $15,000 and less than 20% are diagnosed with depression when the income is more than $35,000.

From our analysis of data brfss2013, we can not say that money causes happiness, but our analysis does indicate that there is a positive correlation between money and happiness among our respondents.

Possible bias of samples in my analysis We used variable “addepev2” to analyze long term happiness. This variable indicates whether the interviewee has been told to have a depression disorder by a doctor, nurse or other health professionals. People with low income are very likely to not have health insurance and would very likely to answer this question as NA or No, which might cause bias in our analysis.

Research question 2:

Never ask a woman her weight! We are curious to learn if the weight level of female respondents is different from that of males.

The following variables are related to our question:

1. sex: Respondents Sex
2. X_bmi5cat: Computed Body Mass Index Categories (classified as ‘Underweight’, ‘Normal’, ‘Overweight’)

The table below shows the percentages of each body weight distribution in male and female. The graph visualizes the table. From the graph and the table, we observe that the weight level of female resondents are indeed different from that of males. of following:

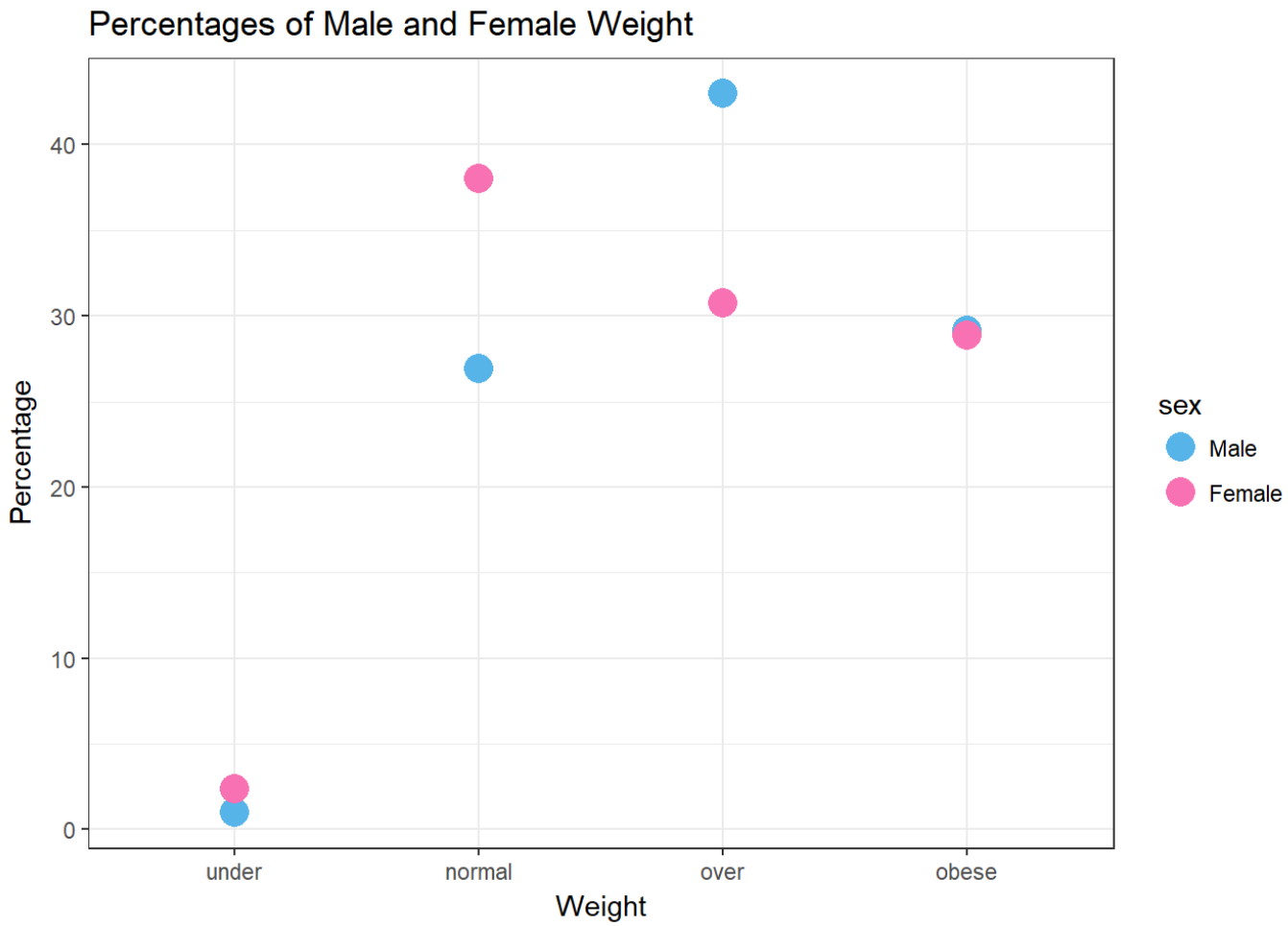
1. 2.12% of females are underweight, higher than 0.935% of the male counterpart.
2. 36.8% of females are normal weight, much higher than 26.7% of the male counterpart.
3. 31.0% of females are overweight, much lower than 42.8% of the male counterpart.
4. The percentages of obesities are very close in female and male, 29.6% and 30% respectively.

If females are more weight conscious, they are more likely to take actions to lose weight, such as working out, getting on a diet, or undergoing plastic surgeries. More surveys or research should be done to discover the psychological reasons behind our analysis.

```
#new data.frame titled weight_couscious which contains variables sex, X_bmi5cat. Remove all NA data.

weight_conscious<-brfss2013 %>%
filter(!is.na(sex),!is.na(X_bmi5cat)) %>%
select(sex,X_bmi5cat)

#calculate the percentage of Underweight, Normal Weight, Overweight, and Obesity in male and female.

weight_percentage<-weight_conscious %>%
group_by(sex) %>%
summarise(under=100*sum(X_bmi5cat=="Underweight")/n(), normal=100*sum(X_bmi5cat=="Normal weight")/n(),over=100*sum(X_bmi5cat=="Overweight")/n(), obese=100*sum(X_bmi5cat=="Obese")/n())
weight_percentage
```

```
## # A tibble: 2 x 5
##   sex    under normal  over obese
##   <fct>  <dbl>  <dbl> <dbl> <dbl>
## 1 Male   0.967   26.9  43.0  29.2
## 2 Female 2.37    38.0  30.7  28.9
```

```
#plot the percentages of weight types in male and female in a scattered plot. Pink represents female and blue represents male.
library(tidyr)
weight<-gather(weight_percentage,weight,percentage,-sex)
weight$weight<- as.character(weight$weight)
weight$weight <- factor(weight$weight, levels=unique(weight$weight))
plot<-ggplot(weight,aes(x=weight,y=percentage,color=sex))+theme_bw()+geom_point(size=5)+labs(x='Weight',y='Percentage',title='Percentages of Male and Female Weight')
plot+scale_color_manual(values=c("#56B4E9", "#f871b3"))
```



Research question 3:

Do tall people suffer more or less from heart disease?

These variables are related to our question:

1. sex: Respondents Sex
2. htin4: Computed Height In Inches
3. cvdinfr4: Ever Diagnosed With Heart Attack
4. cvdcrhd4: Ever Diagnosed With Angina Or Coronary Heart Disease

```
#This R code subsets a new data.fame called f_height that re-groups the female heig
ht into the tallest 25%, middle 50%(Normal), and shortest 25%

f_height<-brfss2013 %>% filter(sex=='Female') %>%
filter(htin4<=100 &htin4>=30) %>%
filter(!is.na(cvdinfr4), !is.na(cvdcrhd4), !is.na(cvdstrk3)) %>%
select(htin4,cvdstrk3,cvdinfr4,cvdcrhd4) %>%
 arrange(desc(htin4))%>%
mutate(f_range=ifelse(htin4>=quantile(htin4,prob=0.75),'tallest25%',ifelse(htin4<=q
uantile(htin4,prob=0.25),'shortest25%','Normal')))
```

```
#summarise the percentages of strokes, heart attacks, and heart deseases for female
in each height subgroup.
#name the summary f_heart
f_heart<-f_height%>%
group_by(f_range) %>%
summarise(f_stroke_p=100*sum(cvdstrk3=='Yes')/n(),f_attack_p=100*sum(cvdinfr4=='Yes
')/n(),f_desease_p=100*sum(cvdcrhd4=='Yes')/n())
```

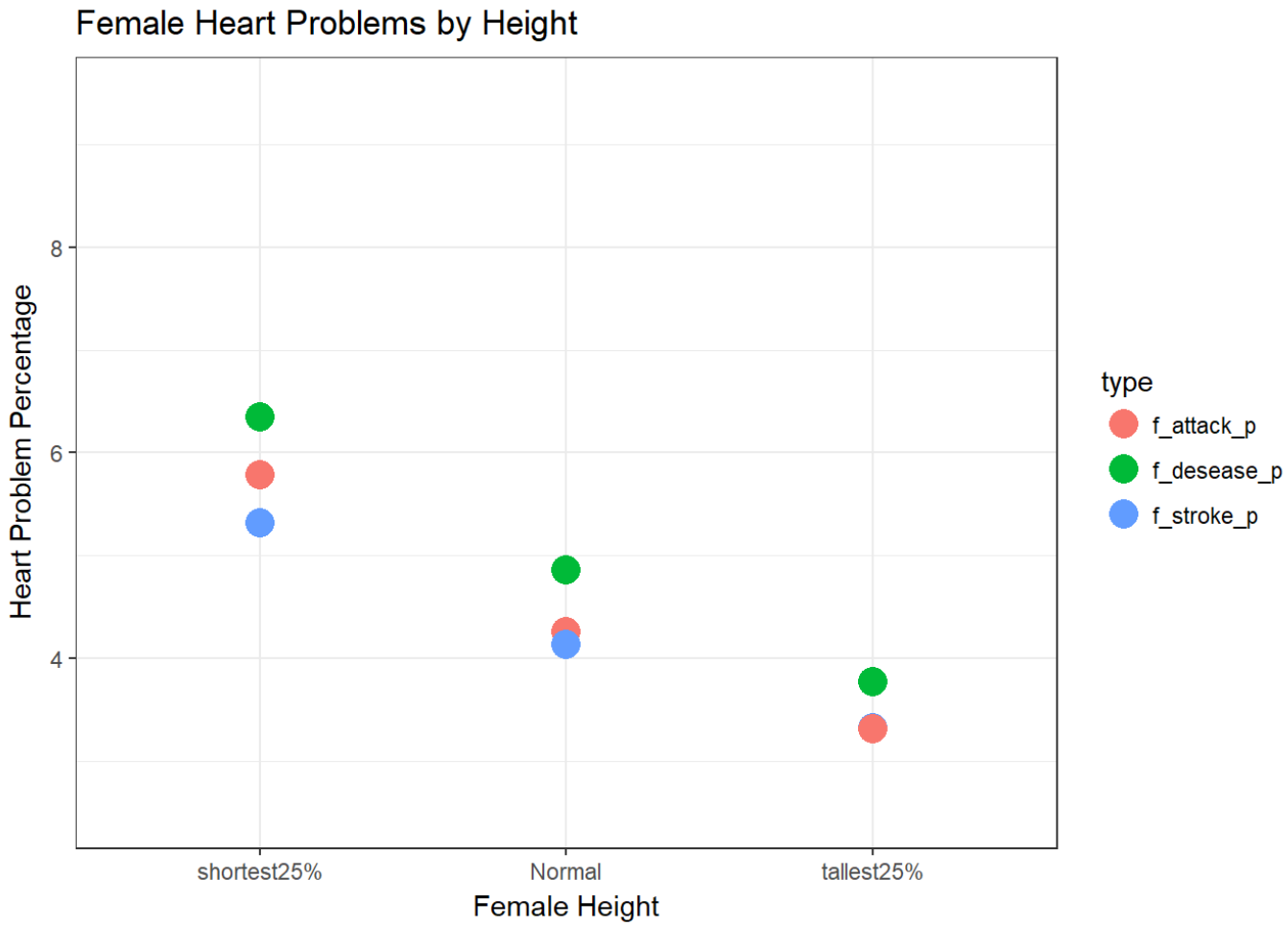
```
#This R code subsets a new data.fame called m_height that groups the male height in
to the tallest 25%, middle 50%(Normal), and shortest 25%

m_height<-brfss2013 %>%
filter(sex=='Male') %>%
filter(htin4<=100 &htin4>=30) %>%
filter(!is.na(cvdinfr4), !is.na(cvdcrhd4), !is.na(cvdstrk3)) %>%
select(htin4,cvdstrk3,cvdinfr4,cvdcrhd4) %>%
arrange(desc(htin4))%>%
mutate(m_range=ifelse(htin4>=quantile(htin4,prob=0.75),'tallest25%',ifelse(htin4<=q
uantile(htin4,prob=0.25),'shortest25%','Normal')))
```

```
#summarise the percentages of strokes, heart attacks, and heart deseases for male i
n each height subgroup.
#name the summary m_heart
m_heart<-m_height%>%
group_by(m_range) %>%
summarise(m_stroke_p=100*sum(cvdstrk3=='Yes')/n(),m_attack_p=100*sum(cvdinfr4=='Yes
')/n(),m_desease_p=100*sum(cvdcrhd4=='Yes')/n())
```

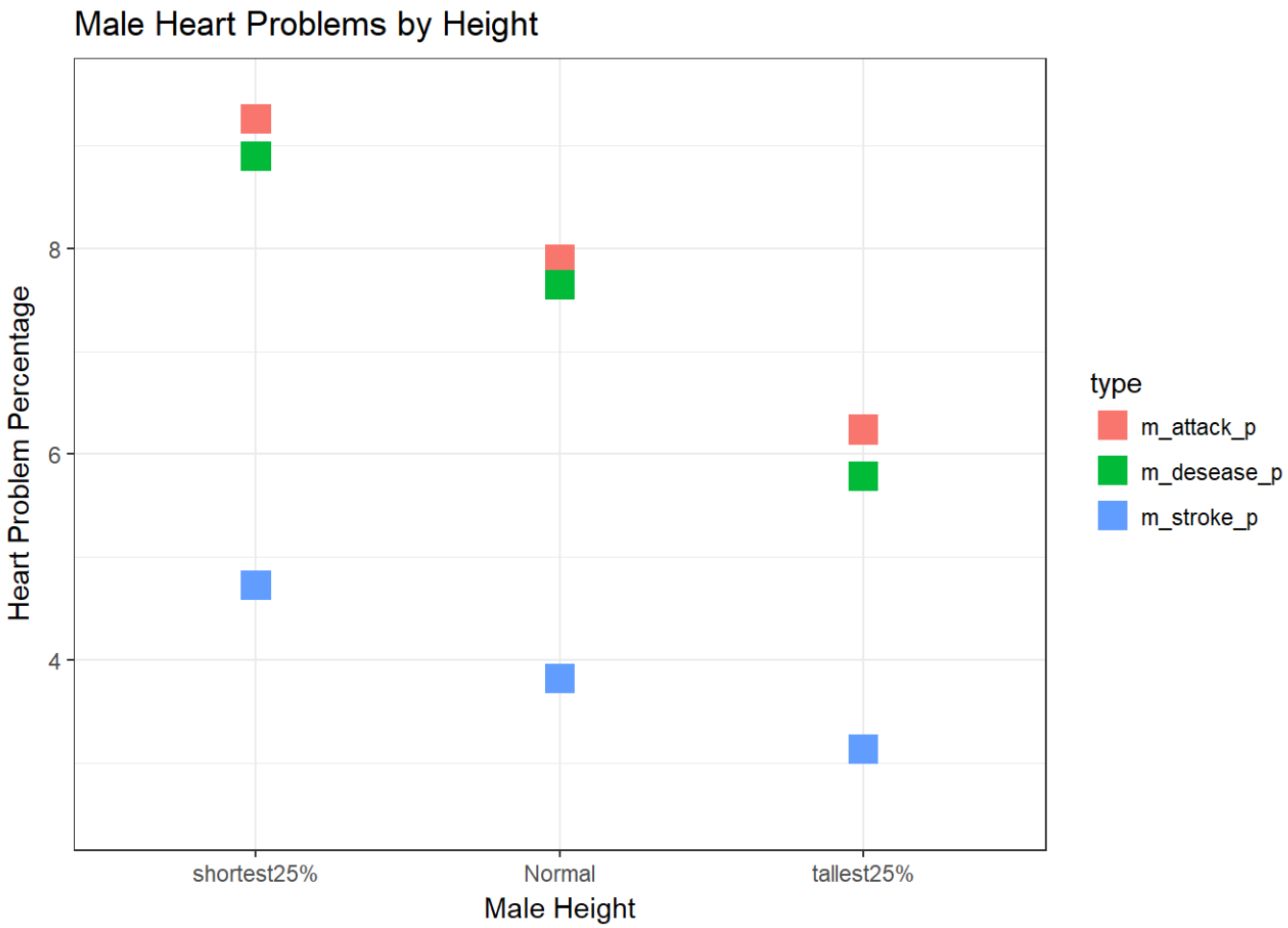
```
#reshape the vector of m_heart and f_heart into its transpose matrices in order to
plot properly.
m_h<-gather(m_heart,type,percent,m_stroke_p, m_attack_p, m_desease_p)%>%arrange(des
c(percent))

m_h$m_range<- as.character(m_h$m_range)
m_h$m_range <- factor(m_h$m_range, levels=unique(m_h$m_range))

f_h<-gather(f_heart,type,percent,f_stroke_p, f_attack_p, f_desease_p)%>%arrange(des
c(percent))
f_h$f_range<- as.character(f_h$f_range)
f_h$f_range <- factor(f_h$f_range, levels=unique(f_h$f_range))
```

```
#plot percentages of female with heart problems v.s. height range.

f_plot<-ggplot(f_h,aes(x=f_range,y=percent, color=type))+theme_bw()+geom_point(size
=5)+scale_y_continuous(limits=c(2.5,9.5))+labs(x='Female Height',y='Heart Problem P
ercentage',title='Female Heart Problems by Height')

#plot percentages of male with heart problems v.s. height range.

m_plot<-ggplot(m_h,aes(x=m_range,y=percent,color=type))+theme_bw()+geom_point(size=
5,shape=15)+scale_y_continuous(limits=c(2.5,9.5))+labs(x='Male Height',y='Heart Pro
blem Percentage',title='Male Heart Problems by Height')
```

```
f_plot
```


Female Heart Problems by Height
Heart Problem Percentage
8
6
4
shortest25%
Normal
tallest25%
Female Height
type
f_attack_p
f_desease_p
f_stroke_p

```
m_plot
```



The two graphs above illustrate the percentage of male and female with heart problems with respect to their height. The x-axis is divided into three groups: the tallest 25%, middle 50% (Normal), and shortest 25% of female and male samples. In both male and female graphs, we see a negative correlation between the rate of heart problems and height. It appears that taller respondents have a less chance of heart problems than shorter ones. It is also interesting that males have a higher chance of heart problems than females. The most common heart problem is heart attack for male, and heart disease such as angina or coronary diseases for female.

The tables below show details of the data in the graphs. The chance of heart related problems decreases as height increases. In both male and female samples, the tallest 25% have the lowest chance of stroke, heart attack, and heart diseases. The shortest 25% of samples have the highest chances of heart problems.

```
#Display Female Heart Problem Percentages
f_heart%>%
  arrange(desc(f_stroke_p))
```

```
## # A tibble: 3 x 4
##   f_range     f_stroke_p f_attack_p f_desease_p
##   <chr>            <dbl>      <dbl>       <dbl>
## 1 shortest25%       5.32       5.78        6.35
## 2 Normal            4.13       4.26        4.86
## 3 tallest25%        3.32       3.32        3.77
```

```
#Display Male Heart Problem Percentages
m_heart%>%
    arrange(desc(m_stroke_p))
```

```
## # A tibble: 3 x 4
##   m_range     m_stroke_p m_attack_p m_desease_p
##   <chr>            <dbl>      <dbl>       <dbl>
## 1 shortest25%       4.73       9.25        8.89
## 2 Normal            3.82       7.89        7.64
## 3 tallest25%        3.13       6.24        5.78
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: