Assignment 1

Mathew Zbitniff

2022-10-07

a)

First, we load the data. Notice that we use the "here" function. Using the "here" function is a good reproducible research practice.

test change

```
library("here")
```

```
## here() starts at C:/Users/Mathe/OneDrive/Desktop/Stat447_Homework_Mathew_Zbitniff
```

```
mydata<-read.csv(here("data","hyper.csv"))
```

Some overveiw:

```
head(mydata)
```

```
##   seqn age_mn gender hypertension bmi
## 1    1     29      2           NA   3
## 2    2    926      1            0   2
## 3    3    125      2            0   3
## 4    4     22      1           NA  NA
## 5    5    597      1            0   1
## 6    6    230      2            0   2
```

```
str(mydata)
```

```
## 'data.frame':    1000 obs. of  5 variables:
##  $ seqn        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age_mn      : int  29 926 125 22 597 230 712 159 133 518 ...
##  $ gender      : int  2 1 2 1 1 2 2 1 2 1 ...
##  $ hypertension: int  NA 0 0 NA 0 0 0 0 0 1 ...
##  $ bmi         : int  3 2 3 NA 1 2 1 3 3 0 ...
```

Since we have the data loaded properly, we can start cleaning it up. Notice above the NA's in our data, let's check how many we have and how big our data set is:

```
ncol(mydata)
```

```
## [1] 5
```

```
nrow(mydata)
```

```
## [1] 1000
```

```
sum(is.na(mydata))
```

```
## [1] 453
```

```
nrow(mydata[rowSums(is.na(mydata)) > 0,])
```

```
## [1] 307
```

```
colSums(is.na(mydata))
```

```
##         seqn       age_mn       gender hypertension          bmi
##            0           16            0          271          166
```

```
colSums(is.na(mydata))/(dim(mydata)[1])
```

```
##         seqn       age_mn       gender hypertension          bmi
##        0.000        0.016        0.000        0.271        0.166
```

The above results tell us that in our data we have 5 columns and 1000 rows, and there are 453 NA's within 307 rows. Most of the NA's are located in the hypertension column (it is almost 30% NA's). since we have not covered how to replace missing data, we simply remove The NA's. We will still have 693 observations, which should be plenty.

b)

Now, we check for potential errors:

```
c(min(mydata$age_mn),max(mydata$age_mn))
```

```
## [1]   96 1012
```

```
c(min(mydata$gender),max(mydata$gender))
```

```
## [1] 1 2
```

```
c(min(mydata$hypertension),max(mydata$hypertension))
```

```
## [1] 0 1
```

```
c(min(mydata$bmi),max(mydata$bmi))
```

```
## [1] 0 3
```

Based on our data, these are the expected max and min values for each variable. So we do not see any errors.

c)

Since there are no errors and no more NA's, we can assign our categorical variable's:

```
mydata$gender<-as.factor(mydata$gender)
mydata$bmi<-as.factor(mydata$bmi)
mydata$hypertension<-as.factor(mydata$hypertension)
```

Now that the variables are properly defined, we can look at the summary statistics:

```
summary(mydata)
```

```
##       seqn            age_mn       gender  hypertension bmi
##  Min.   :   2.0   Min.   :  96.0   1:351   0:598        0:141
##  1st Qu.: 260.0   1st Qu.: 187.0   2:342   1: 95        1:208
##  Median : 532.0   Median : 328.0                        2:271
##  Mean   : 516.5   Mean   : 412.2                        3: 73
##  3rd Qu.: 776.0   3rd Qu.: 627.0
##  Max.   :1000.0   Max.   :1012.0
```

Now, lets look at a few more things like Central tendency, Spread, proportion, normality, and occurrences of variables.

Continous varible:

```
IQR(mydata$age_mn)
```

```
## [1] 440
```

Categorical varibles:

```
table(mydata$gender)
```

```
##
##   1   2
## 351 342
```

```r
table(mydata$hypertension)
```

```
##
##   0   1
## 598  95
```

```r
table(mydata$bmi)
```

```
##
##   0   1   2   3
## 141 208 271  73
```

```r
100*table(mydata$gender)/length(mydata$gender)
```

```
##
##        1        2
## 50.64935 49.35065
```

```r
100*table(mydata$hypertension)/length(mydata$hypertension)
```

```
##
##        0        1
## 86.29149 13.70851
```

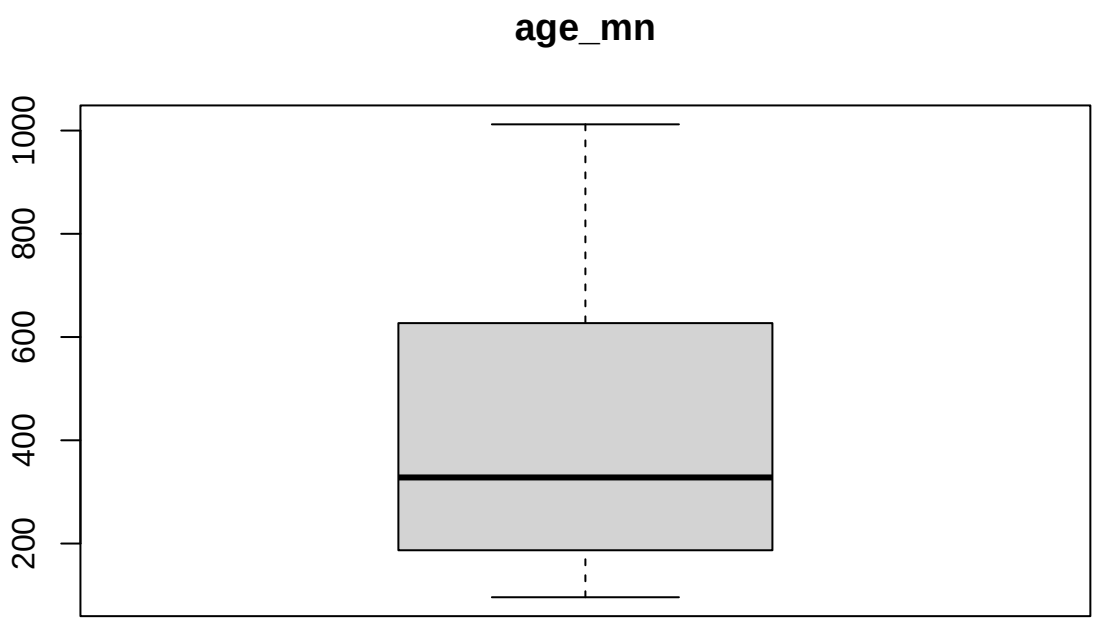
```r
100*table(mydata$bmi)/length(mydata$bmi)
```

```
##
##        0        1        2        3
## 20.34632 30.01443 39.10534 10.53391
```

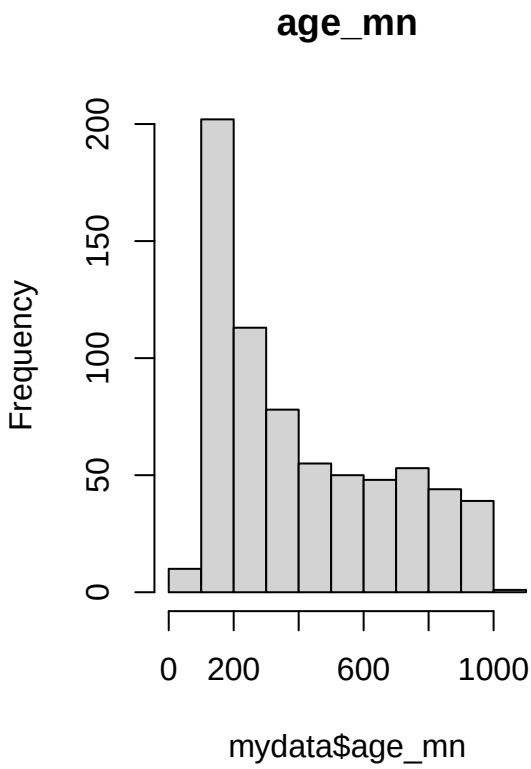
Note relative even spread from age_mn and bmi.

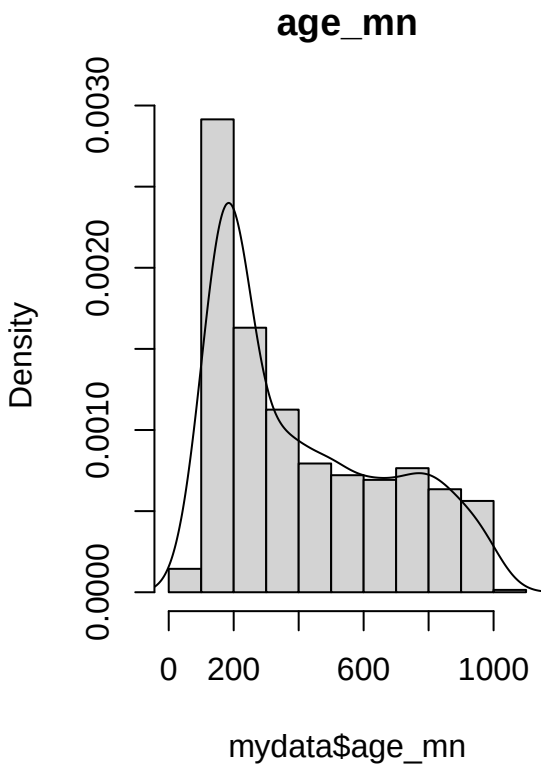
Graphical Display for the Continuous Variable:

```r
boxplot(mydata$age_mn, main = "age_mn")
```




```
par(mfrow = c(1, 2))
hist(mydata$age_mn, main = "age_mn")
hist(mydata$age_mn, main = "age_mn", prob = 1)
lines(density(mydata$age_mn))
```

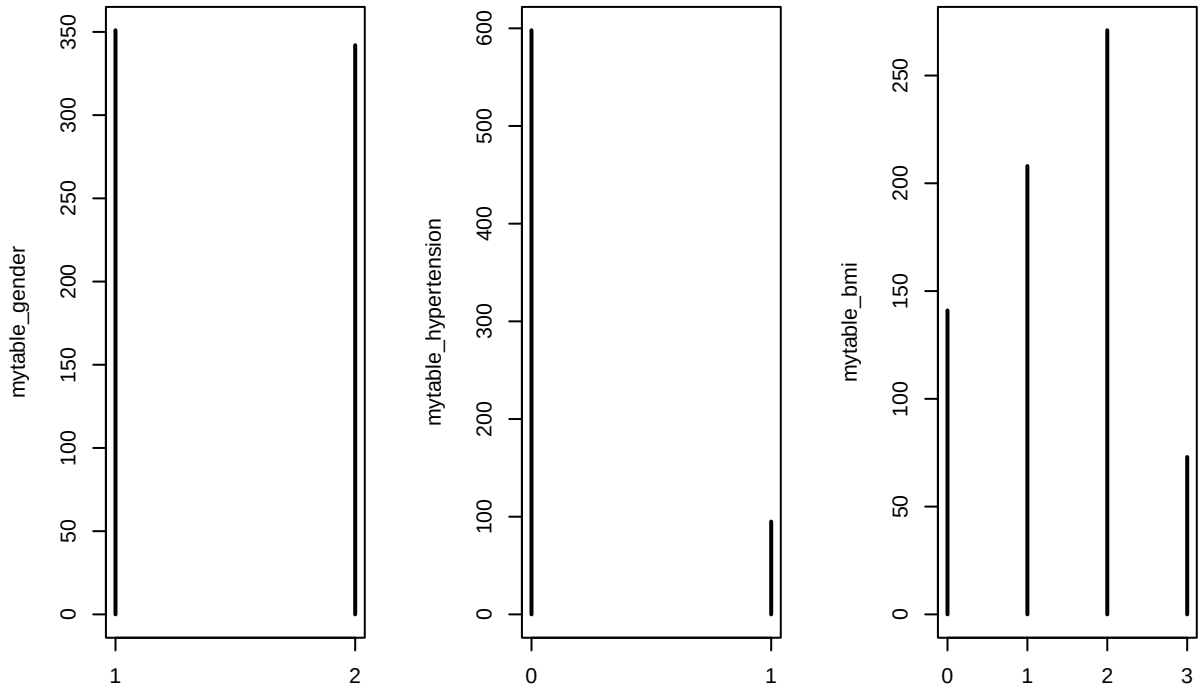




Note the higher frequency of younger ages.

Graphical Display for the Categorical Variables:

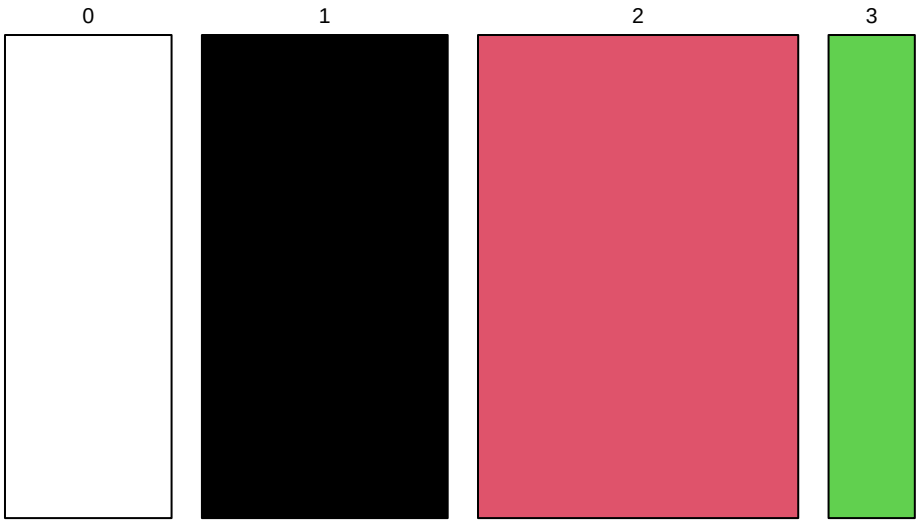
```
par(mfrow = c(1, 3))
mytable_gender <- table(mydata$gender)
plot(mytable_gender)
mytable_hypertension <- table(mydata$hypertension)
plot(mytable_hypertension)
mytable_bmi <- table(mydata$bmi)
plot(mytable_bmi)
```



Note that our previous tables told us the same info!

Here is another visualization for the bmi variable

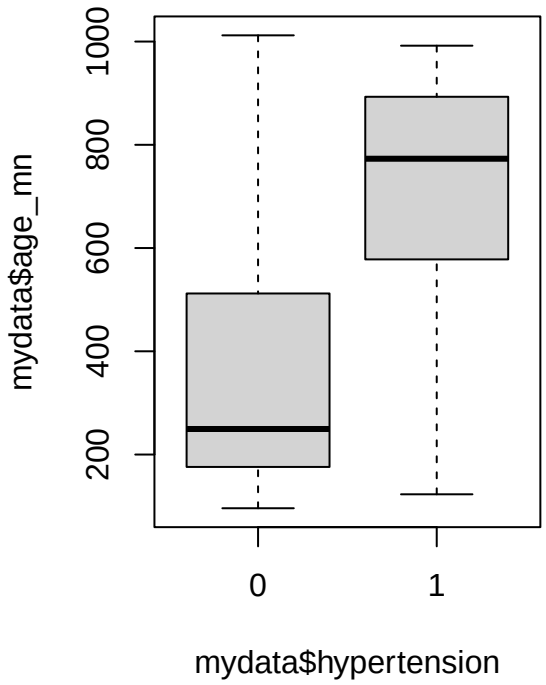
```
mosaicplot(mytable_bmi, col= c(names(mytable_bmi)), margin = T, main = "")
```

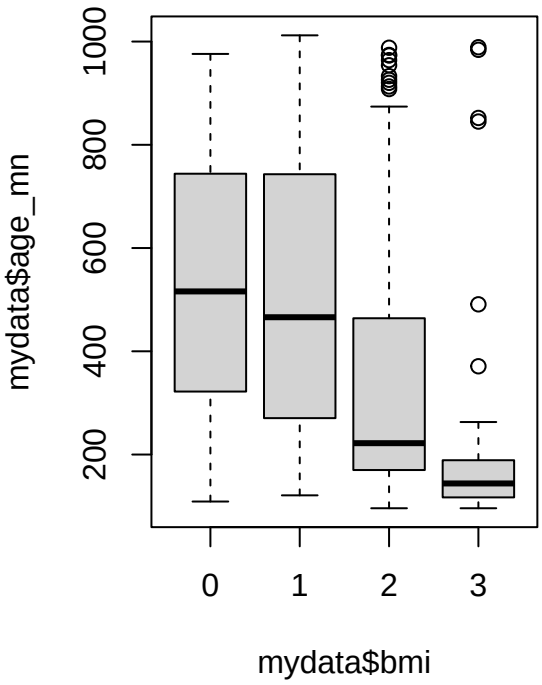


Note the most common bmi status is normal.

Relationships between variables

```
par(mfrow = c(1, 2))
boxplot(mydata$age_mn ~ mydata$hypertension)
boxplot(mydata$age_mn ~ mydata$bmi)
```



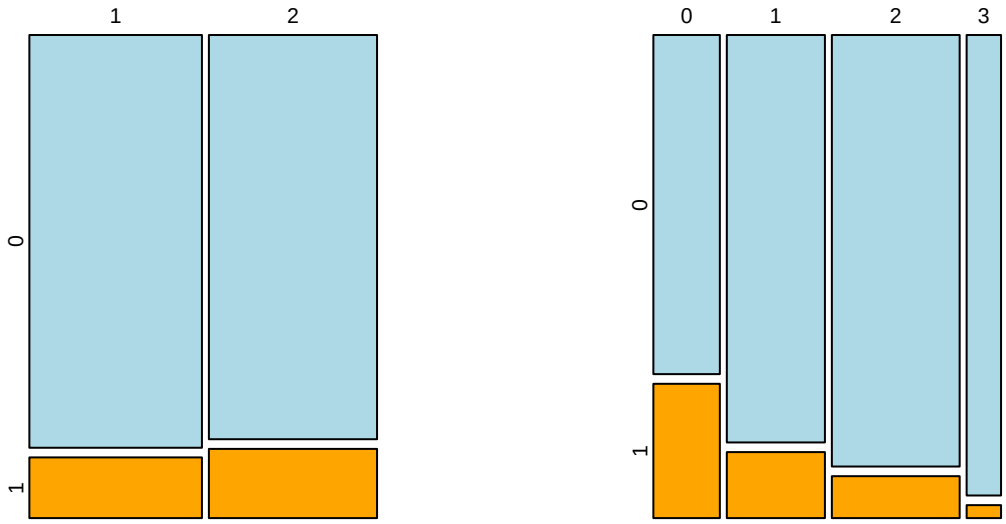


We see that the older one is, the more likely they are to have hypertension. Also, we see that lower age is correlated with higher bmi.

```
par(mfrow = c(1, 2))

mytable1 <- table(mydata$gender, mydata$hypertension)
mosaicplot(mytable1, col = c("light blue", "orange"), main = "")

mytable2 <- table(mydata$bmi, mydata$hypertension)
mosaicplot(mytable2, col = c("light blue", "orange"), main = "")
```



The first graphs tell us that males and females experience hypertension at a similar rate, the second graph tells us that the higher one bmi is, the higher the chance they'll have hypertension.

Conclusion

At this point, we have cleaned our data and checked for errors. We also have a good idea about the relationships between the different variables. We are now ready to do future analysis and modeling on this data.