LAB 1

SOLUTION

Install Latex Distribution

One of the goals of this lab is that you make sure you can create interactive document files (Rmd) and convert them to pdf, as well as html format. In order to be able to write mathematical text, you would need Latex, wich RStudio uses to produce such text in pdf files.

Latex is a free typesetting system/software that for quite some time has been a standard for writing books, scientific articles, theses and other documments in mathematics, statistics, physics, computer science and other disciplines.

If you've never used Latex on your machine, you will need some Latex distribution (MikTex, MacTex, TinyTex,…), which is a set of packages required for Latex. You can find various online tutorials how to install them. For example, MacTex.

You can read instructions here: https://www.latex-project.org/get/. Or else,

- for Windows
- for Mac
- for Linux/Ubuntu

Once you are done with Tex installation, you can (but don't have to) also install some Latex editor, if you want to create math/stats pdf files independently of RStudio, i.e. for some purposes outside of this class. There are various Latex editors. I use TexShop.

Hopefully, your MacTex installation went smoothly. Make sure you can create a pdf file using "Knit" button in RStudio. To do that, you need to create a new Rmarkdown file (`File --> New File --> R Markdown --> Document --> PDF`). Make sure this works or you fix problems ASAP. Let me know if you have any trouble ASAP. Don't wait couple of days before the lab deadline. You won't be able to finish the lab.

The above sequence of clicks should produce an Rmd file, which you can see it open in your RStudio. It is a template with some content that helps you figure out how basics of RMarkdown work. Save this file in order to run it, by clicking on `save` button (you would need to choose a name of the file). To run the file, click on `Knit`. Since you chose to create PDF (rather than HTML or Word), it will automatically generate pdf file.

Absenteeism at Work

The file `Absenteeism_at_work.csv` is posted on Canvas, as well as the documentation about the data. This is taken from UC Irvine Machine Learning Repository archive https://archive.ics.uci.edu/ml/datasets

Instructions

- Read the documentation to become familiar with the meanings of the variables/columns.
- Read in the data set using the command

```
df = read.csv("Absenteeism_at_work.csv",sep=";",header=TRUE)
```

- You will onle need to submit one PDF file, produced by your Rmd file. Include your code, plot and comments in your Rmarkdown file, so that they are shown in the pdf file.
- In each plot, include appropriate title and labels. Include the legend, if appropriate. Also, after each plot, write a short comment (one or two sentence) if you see something on the graph, i.e. if graph reveals or suggests something about the data. **Do not forget** to write these comments, even if you can't say much by looking at the graph (in that case, just say that the graph is not very useful, i.e. doesn't suggest anything).

- Use base plot (this time), not `ggplot2`.

1. Plot the scatter plot of height vs. weight (so, weight on x-axis) including all the (non-missing) data.
2. Plot the histogram of hours of absences. Do not group by ID, just treat each absence as one observation.
3. Plot the histogram of age of a person corresponding to each absence. Do not group by ID, just treat each absence as one observation.
4. Plot the bar plot of hours by month. So, each month is represented by one bar, whose height is the total number of absent hours of that month.
5. Plot the box plots of hours by social smoker variable. So, you will have two box plots in one figure. Use the legend, labels, title. Play with colors.
6. Plot the box plots of hours by social drinker variable. So, you will have two box plots in one figure. Use the legend, labels, title. Play with colors.

When you are done, before submitting, go over the instructions again and for each plot make sure you met all the requirements.

SOLUTION

```r
df = read.csv("Absenteeism_at_work.csv",sep=";",header=TRUE)
```

```r
head(df,3)
```

```
  ID Reason.for.absence Month.of.absence Day.of.the.week Seasons
1 11                 26                7               3       1
2 36                  0                7               3       1
3  3                 23                7               4       1
  Transportation.expense Distance.from.Residence.to.Work Service.time Age
1                    289                              36           13  33
2                    118                              13           18  50
3                    179                              51           18  38
  Work.load.Average.day Hit.target Disciplinary.failure Education Child
1               239.554         97                    0         1     2
2               239.554         97                    1         1     1
3               239.554         97                    0         1     0
  Social.drinker Social.smoker Pet Weight Height Body.mass.index
1              1             0   1     90    172              30
2              1             0   0     98    178              31
3              1             0   0     89    170              31
  Absenteeism.time.in.hours
1                         4
2                         0
3                         2
```

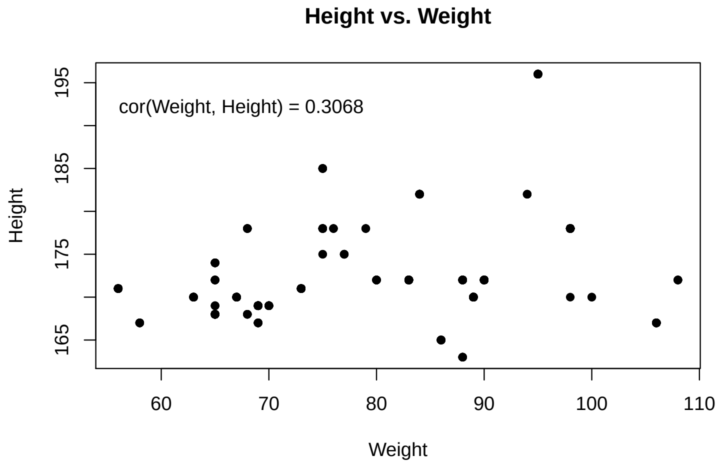
```r
dim(df)
```

```
[1] 740  21
```

```r
attach(df)
```

1.

```r
plot(Height ~ Weight, pch=16,
     main="Height vs. Weight")
text(55,192,paste("cor(Weight, Height) = ",round(cor(Weight,Height),4), sep=""),
     pos=4)
```



As weight increase, it seems so does the height. I included the correlation between weight and height, which is positive, as natural. Also, for larger weight the variance of height seems to be larger. This should not be surprising.

2.

Some of the names of columns are too long.

```
colnames(df)
```

```
 [1] "ID"                                "Reason.for.absence"
 [3] "Month.of.absence"                  "Day.of.the.week"
 [5] "Seasons"                           "Transportation.expense"
 [7] "Distance.from.Residence.to.Work" "Service.time"
 [9] "Age"                               "Work.load.Average.day"
[11] "Hit.target"                        "Disciplinary.failure"
[13] "Education"                         "Child"
[15] "Social.drinker"                    "Social.smoker"
[17] "Pet"                               "Weight"
[19] "Height"                            "Body.mass.index"
[21] "Absenteeism.time.in.hours"
```

Let us rename the last variable into `Hours`. For later use, we also change some other variable names as follows.

```
colnames(df)[21] = "Hours"
colnames(df)[3] = "Month"
colnames(df)[15] = "Drinker"
colnames(df)[16] = "Smoker"

colnames(df)
```

```
 [1] "ID"                                "Reason.for.absence"
 [3] "Month"                             "Day.of.the.week"
 [5] "Seasons"                           "Transportation.expense"
 [7] "Distance.from.Residence.to.Work" "Service.time"
 [9] "Age"                               "Work.load.Average.day"
[11] "Hit.target"                        "Disciplinary.failure"
[13] "Education"                         "Child"
[15] "Drinker"                           "Smoker"
[17] "Pet"                               "Weight"
[19] "Height"                            "Body.mass.index"
[21] "Hours"
```

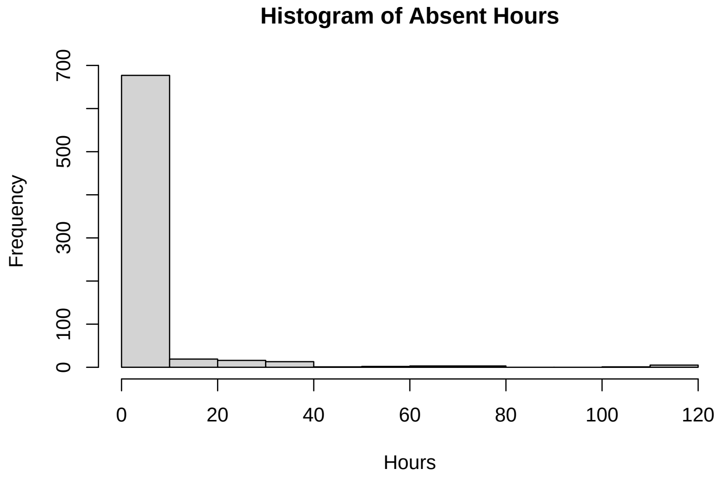
```
attach(df) ## need to update attached column names
```

```
The following objects are masked from df (pos = 3):

    Age, Body.mass.index, Child, Day.of.the.week, Disciplinary.failure,
    Distance.from.Residence.to.Work, Education, Height, Hit.target, ID,
    Pet, Reason.for.absence, Seasons, Service.time,
    Transportation.expense, Weight, Work.load.Average.day
```

Now, let's plot the histogram of absent hours.

```
hist(Hours, main = "Histogram of Absent Hours")
```



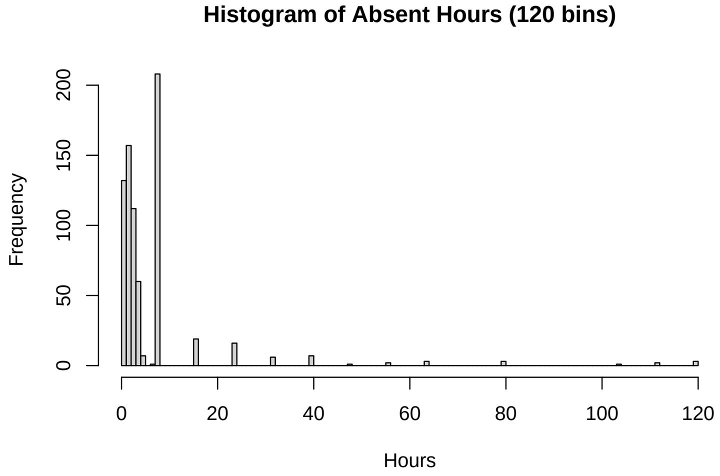
```
max(Hours)
```

```
[1] 120
```

This is not a surprise - people are more often absent for a day or two (or less than a day if they come later to work or leave earlier) than they are for a several days.

Also, the maximum number of absent hours is 120. For more sophisticated plot, we can require more breaks (this was not necessary for you to do)

```
hist(Hours, breaks=120,
     main = "Histogram of Absent Hours (120 bins)")
```



Now we see from the histogram there is another reason why data are right-skewed: there a some observations with zero-hours. This may either mean the hours were not recorded for these observations, or employees were just late for less than an hour. Let's look at it more closely.

```
sum(Hours==0)
```

```
[1] 44
```

```r
mean(Hours==0)
```

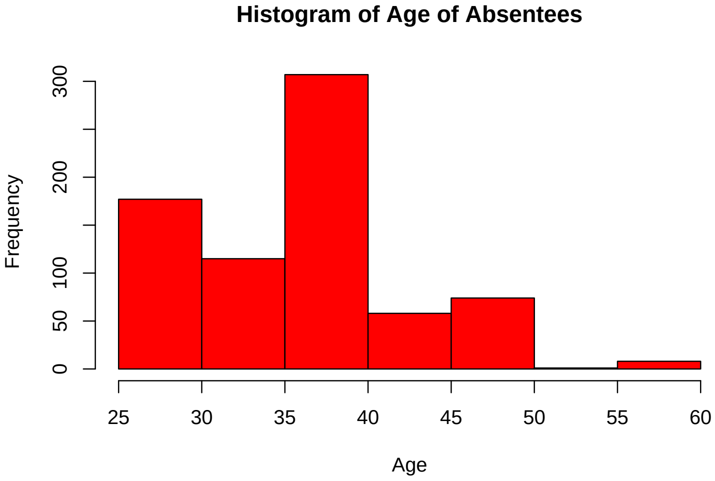
```
[1] 0.05945946
```

So, we see there are 44 zero-hours, and they make 5.95 % of the data. If we were to analyze distribution of hours further, it would be important to know what exactly zero-hour mean.

Either plot tells us the vast majority of absence times last not more than 10 hours (or in fact 8 hours, which is one work day). Note the spikes at 8, 16, 24,… This suggests people take whole days off. For example, no one was absent 2.5 days i.e. 20 hours - it's either 16 or 24. This is an expected behavior.

3.

```r
hist(Age, main="Histogram of Age of Absentees", col="red")
```



We see that the people who are most often absent are those in the age group between 35 - 40. This may be simple because of all the age groups this one is the largest - additional data would be needed to address whether this is the case.

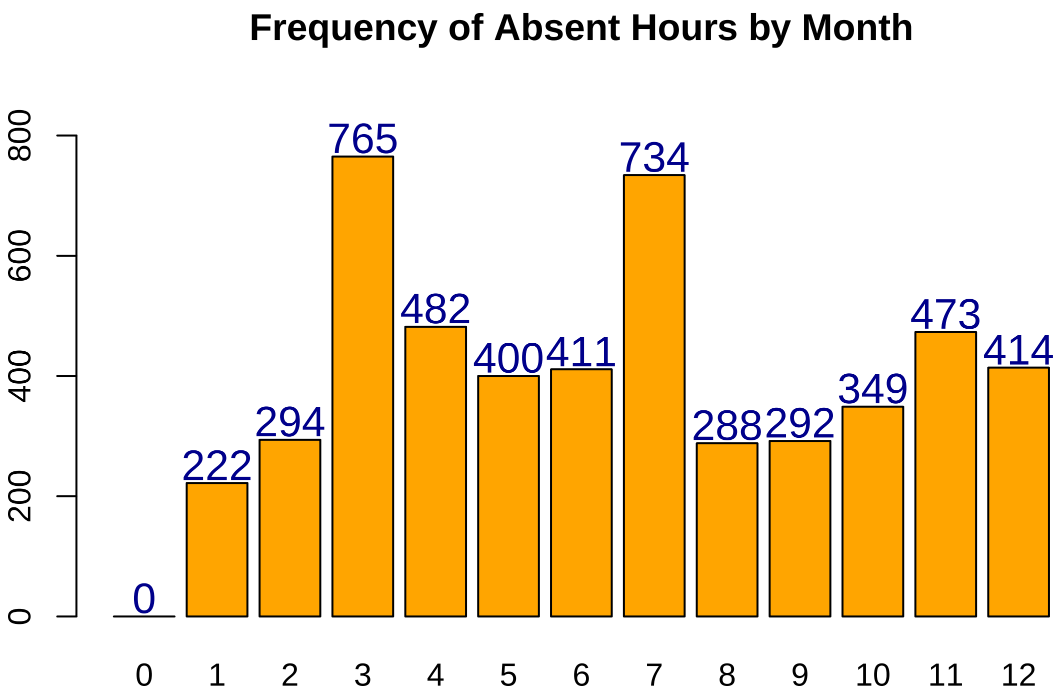
4.

```r
freq <- tapply(X=Hours, INDEX=Month, FUN=sum)
freq
```

```
  0   1   2   3   4   5   6   7   8   9  10  11  12 
  0 222 294 765 482 400 411 734 288 292 349 473 414 
```

```r
xx <-  barplot(freq, ylim=c(0,850), col="orange",
               main="Frequency of Absent Hours by Month")


text(x=xx, y=freq+30, label=as.character(freq),
     cex=1.3, col="darkblue")
```



We see there is appearance of month 0. Maybe the month is not recorded for that absence, but we cannot know that for sure. More over, we see that for month 0 the number of hours is 0, which could mean (although we cannot know for sure) that for these observations the number of absent hours was not recorded neither. To get more confidence the plot is right about that, we run

```
table(Month)
```

```
Month
 0  1  2  3  4  5  6  7  8  9 10 11 12 
 3 50 72 87 53 64 54 67 54 53 71 63 49 
```

```
df[df$Month==0, "Hours"]
```

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

```
which(df$Month==0)
```

```
[1] 738 739 740
```

```
df$Hours[which(df$Month==0)]
```

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

From this analysis we see there are 3 observations with `Month=0` (they are the last 3), and for all three of them the value of Hours is 0. It seems there is no reason to include `Month = 0` into our histogram. So, we can also plot

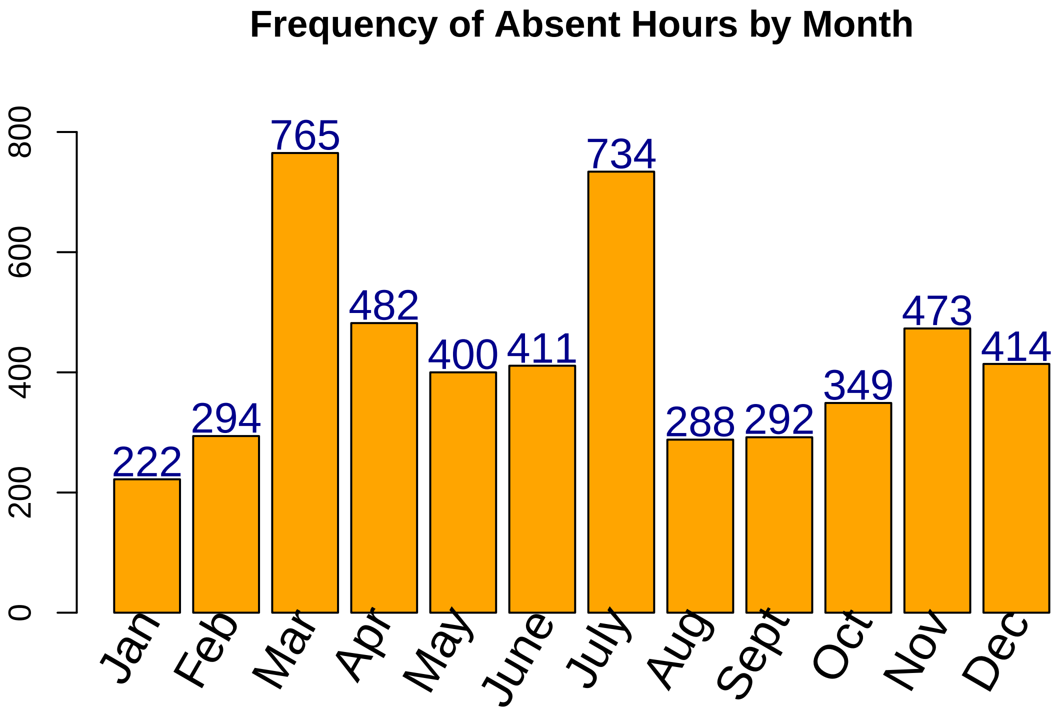
```
freqNo0 = freq[-1] ## take out the first component of vector freq
freqNo0
```

```
  1   2   3   4   5   6   7   8   9  10  11  12 
222 294 765 482 400 411 734 288 292 349 473 414 
```

```
xx <-  barplot(freqNo0, ylim=c(0,850), col="orange",
               xaxt = "n", ## don't include default ticks
               main="Frequency of Absent Hours by Month")

text(x=xx,
     srt = 60, adj= 1, par("usr")[3]-0.25, xpd = TRUE, cex=1.5,
     labels = c("Jan","Feb","Mar","Apr","May","June","July",
                "Aug","Sept","Oct","Nov","Dec"))
```

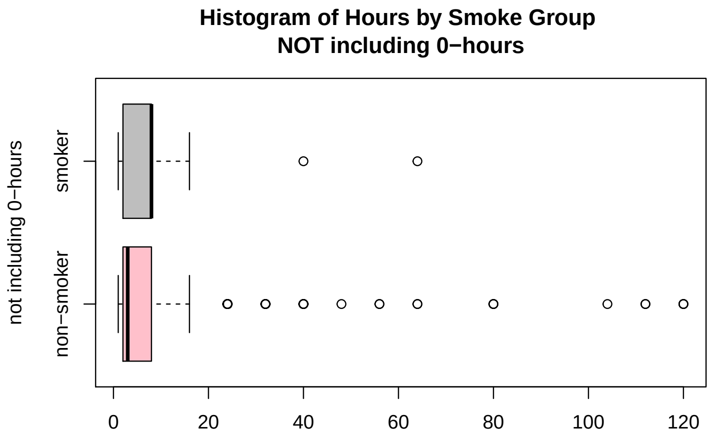
```r
text(x=xx, y=freqNo0+30, label=as.character(freqNo0),
     cex=1.3, col="darkblue")
```



5.

The reason why we would be interested in plotting absent hours by smoker group is to see possible affect of smoking on absent hours. If the meaining of 0-hours is that data are missing, or person was late, it would make sense to take out zero-hours:

```r
boxplot(Hours[Hours!=0] ~ Smoker[which(Hours!=0)], horizontal=TRUE,
        col=c("pink","gray"),
        main="Histogram of Hours by Smoke Group \nNOT including 0-hours",
        names=c("non-smoker", "smoker"),
        xlab = "", ylab="not including 0-hours")
```



```r
table(Hours==0)
```

```
FALSE  TRUE 
  696    44 
```

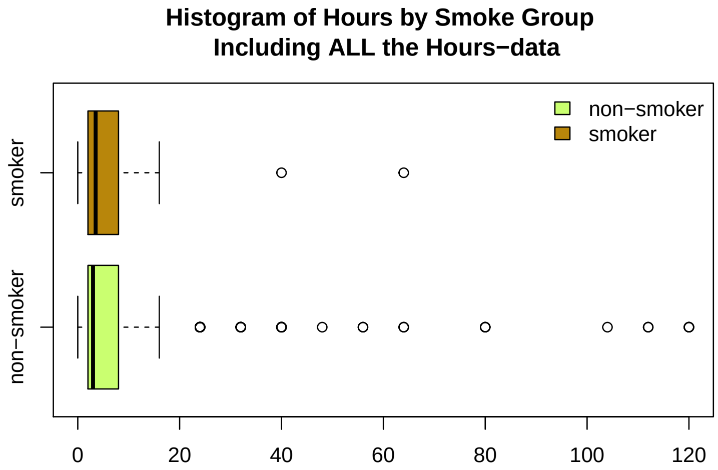
```r
mean(Hours==0)
```

```
[1] 0.05945946
```

Note there are 44 zero-hours in the data, which make 5.95% of the data. Without including them, the median of absent hours for non-smoker is smaller than that of smokers. However, without knowing what exactly 0-hours means, and since I didn't specify you should feel free to investigate whatever you think it might be useful, let us plot boxplot without taking out 0-hours:

```
boxplot(Hours ~ Smoker, horizontal=TRUE,
        col=c("darkolivegreen1","darkgoldenrod"),
        main="Histogram of Hours by Smoke Group \n Including ALL the Hours-data",
        names=c("non-smoker", "smoker"),
        xlab = "", ylab="")

legend("topright",legend=c('non-smoker','smoker'),
       fill=c("darkolivegreen1","darkgoldenrod"),
       bty="n")
```



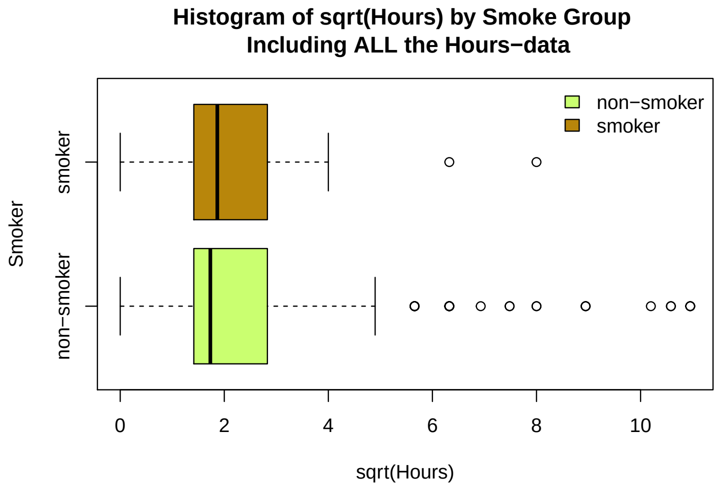
```
table(Smoker)
```

```
Smoker
  0   1 
686  54 
```

We already saw the hours data are highly right-skewed, and these boxplots also show that. In order to more easily compare the two groups visually, we can transform data. For example, let us transworm by taking the square root of the data.

```
boxplot(sqrt(Hours) ~ Smoker, horizontal=TRUE,
        col=c("darkolivegreen1","darkgoldenrod"),
        main="Histogram of sqrt(Hours) by Smoke Group \n Including ALL the Hours-data",
        names=c("non-smoker", "smoker"))

legend("topright",legend=c('non-smoker','smoker'),
       fill=c("darkolivegreen1","darkgoldenrod"),
       bty="n")
```



The transformed hours (as well as non-transformed) for non-smokers have larger range than that of smokers. This may be because there significantly more data (320) of non-smokers than smokers (420), so it is more likely to occasionally get extreme values). However, the median of non-smokers is a bit smaller. The question of what 0 hours means and whether we can ignore 0-hours data - is important, since on the answer to that question depends whether the two midians are considerably different or not.

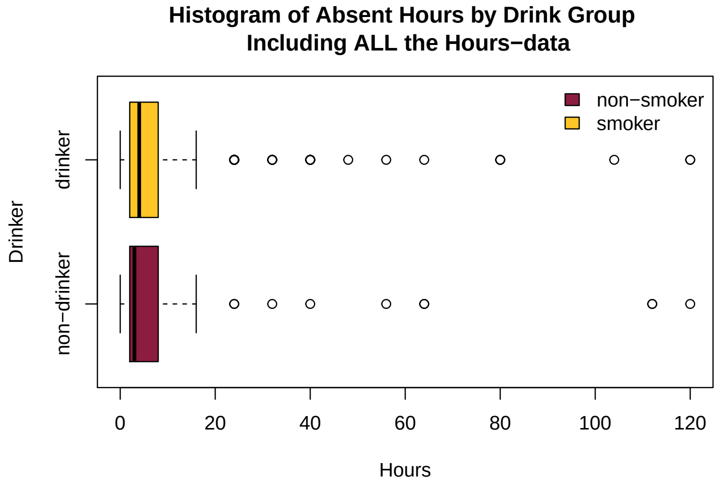
Nevertheless, this is only exploratory analysis. More reliable and scientifically valid analysis requires hypothesis testing.

6.

This part is similar to the previous one.

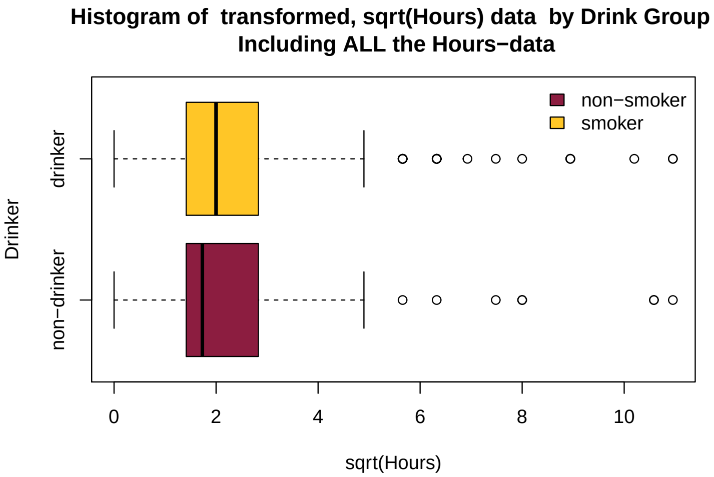
```
boxplot(Hours ~ Drinker, horizontal=TRUE,
        col=c("#8C1D40","#FFC627"),
        main="Histogram of Absent Hours by Drink Group \n Including ALL the Hours-data",
        names=c("non-drinker", "drinker"))

legend("topright",legend=c('non-smoker','smoker'),
       fill=c("#8C1D40","#FFC627"),
       bty="n")
```



```
boxplot(sqrt(Hours) ~ Drinker, horizontal=TRUE,
        col=c("#8C1D40","#FFC627"),
        main="Histogram of  transformed, sqrt(Hours) data  by Drink Group \n Including ALL the Hours-
        names=c("non-drinker", "drinker"))

legend("topright",legend=c('non-smoker','smoker'),
       fill=c("#8C1D40","#FFC627"),
       bty="n")
```



```
table(Drinker)
```

```
Drinker
  0   1 
320 420 
```

The situation is similar to that with Smoker data. However, the difference is that neither group has significantly smaller number of data in the sample, which may explain why both groups have somewhat similar number of extreme values i.e. outliers.