Data Simulation Project

Headache in School Children: Prevalence and Risk Factors

The Story

Recurrent headache is a common problem in school children. This data set is linked to the question whether number of headaches in school children is becoming more common and, if so, what risk factors are associated with the rise in the number of headaches. A survey of 300 children was done and the following factors were recorded:

- Age
- Consumption of alcohol per week
- Caffeine ingestion per day
- Number of smokes per week
- Number of video games owned
- Number of television sets in their bedroom
- Number of physical games played
- Physical or emotional abuse
- leisure time per week
- Number of chocolates eaten per week (caf)
- Physical abuse

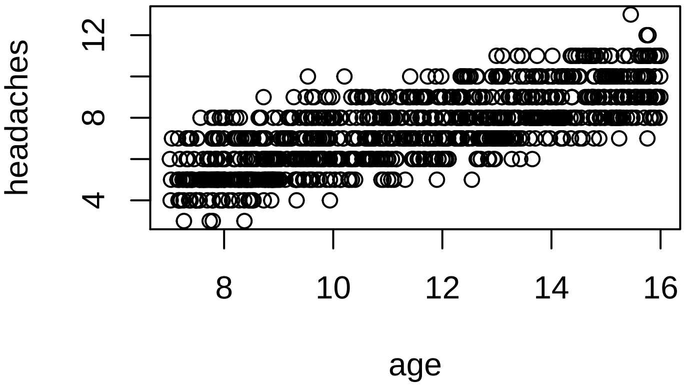
An analysis of the data needs to be made to determine the number of headaches and involved risk factors.

Data generation

This is a function to generate a simulated data set.

The `runif` function is used to get inputs for age as we need to have a mix of children from different ages. `rnorm` function is used to generate values where variables are continuous in nature for example alcohol intake can be floating point number.`rbinom` function is used to generate data for variables where a integer value is required and a definate probability for lamba value. The Caffeine_alcohol_index and Physical_abuse_index are calculated using various variables listed in the data frame. The car make is added as a distraction. Finally manupulate function is sued to determine coefficients for variables involved

```
generate_dataset <- function(N=1000){

    age <- runif(N, min=10-3, max=10+6)
    alcohol <- rnorm(N, mean=20, sd=5)
    caffeine <- rnorm(N, mean=200, sd=80)
    smokes <-  rbinom(N, 2, .20)
    chocolates <- rnorm(N, mean=2, sd=.5)
    Caffeine_alcohol_index <- ( 0.1 * (alcohol) + .001 * (caffeine) +
                                  .03 * (smokes) + 0.02 * (chocolates) )

    video_games <- rbinom(N, 4, .70)
    television <- rbinom(N, 1, .40)
    games_played <- rbinom(N, 2, .53)
    leisure_time <- rnorm (N, mean = 7, sd = 2)

    Physical_activity_index <-  (0.10 * games_played + 0.25* leisure_time
                                 - .10 * video_games - .20 * television)

    Physical_abuse <- sample(c("Y", "N"), N, replace=TRUE, prob=c(.34, .66))

    make <- c("Audi", "Camry", "Nissan", "Subaru",
              "Honda", "BMW", "Acura", "Infinity")
    car <- sample(make, N, replace=TRUE)

    bad_family <- ifelse((Physical_abuse == 'Y'), 2,0)

    headaches = floor(0.5 * age + 1.14 * Caffeine_alcohol_index
                      - 0.7 * Physical_activity_index + bad_family)


    headaches[which(headaches <= 0)] = 0

    data.frame(age, alcohol, caffeine, smokes, chocolates,video_games,
               television, games_played, leisure_time, Caffeine_alcohol_index,
               Physical_activity_index,Physical_abuse, bad_family, car,
               headaches)

}

df <- generate_dataset(1000)

with(df, plot(age, headaches))
```



The manipulate function is used to calculate appropriate intercepts for the number of headaches prediction equation

```
library(manipulate)
manipulate({
  df <- transform(df, headaches = floor(a * age + b * Caffeine_alcohol_index
  - c * Physical_activity_index + bad_family))

  plot(df$age , df$headaches)
}, a=slider(0, 10, step=0.1, initial = 0),b=slider(0, 10, step=0.01, initial = 0),
c=slider(0, 10, step=0.1, initial = 0))
```

Analysis

For the analysis part a data frame containing the following factors is provided:

1. age : Recorded in number of years
2. alcohol: Amount of alcohol ingested in any form per week (in mg)
3. caffeine: Amount of caffeine intake per day (in mg)
4. smokes: Number of cigarettes smoked in a week
5. chocolates: Number of chocolates consumed per week
6. video_games: Number of video games owned
7. television: Number of televisions in the child's bedroom
8. games_played: Number of physical games played per week
9. leisure_time: Amount of free time per week (in hours)
10. Caffeine_alcohol_index : This is the amount of caffeine and alcohol ingested by the child in all forms per week. It is calculated on the basis of the alcohol ingested in all forms per week, amount of caffeine intake per day, number of smokes per week and number of chocolates per week
11. Physical_activity_index : This is a factor that evaluates the amount of physical activity performed by the child based on number of video games owned, number of televicions present in the child's bedroom, number of games played and the amount of leisure time available per week.
12. Physical_abuse: If the child has suffered from any sort of physical abuse ("Yes" or "No")
13. bad_family : It is an intercept value determined by the fact if the child has suffered physical abuse
14. car: The make of the car owned by the family
15. headaches: No. of headaches a child gets per week

We try to analyse the data by using lm.

```r
fit <- lm(headaches ~ age + Caffeine_alcohol_index + Physical_activity_index
          + Physical_abuse -1, data=df)

summary(fit)
```

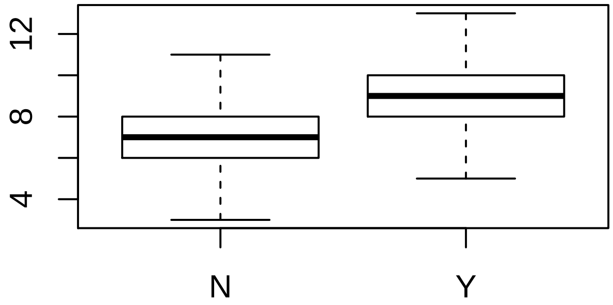
```
##
## Call:
## lm(formula = headaches ~ age + Caffeine_alcohol_index + Physical_activity_index +
##     Physical_abuse - 1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51362 -0.24138 -0.01369  0.24562  0.50936
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## age                      0.502981   0.003465 145.153   <2e-16 ***
## Caffeine_alcohol_index   1.147923   0.017360  66.125   <2e-16 ***
## Physical_activity_index -0.709384   0.017267 -41.082   <2e-16 ***
## Physical_abuseN         -0.534883   0.060409  -8.854   <2e-16 ***
## Physical_abuseY          1.434702   0.062192  23.069   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2821 on 995 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.465e+05 on 5 and 995 DF,  p-value: < 2.2e-16
```

The values obtained are similar to original values.

```
fit2 <- lm(headaches ~ age + Physical_abuse + alcohol+ caffeine + smokes + chocolates
           + video_games + television + games_played+ leisure_time + car -1, data=df)

summary(fit2)
```

```
##
## Call:
## lm(formula = headaches ~ age + Physical_abuse + alcohol + caffeine +
##     smokes + chocolates + video_games + television + games_played +
##     leisure_time + car - 1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53697 -0.23406 -0.01615  0.23898  0.53256
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## age               0.5024987  0.0034717 144.740  < 2e-16 ***
## Physical_abuseN -0.5440450  0.0809119  -6.724 3.00e-11 ***
## Physical_abuseY  1.4256110  0.0825955  17.260  < 2e-16 ***
## alcohol          0.1151190  0.0017740  64.891  < 2e-16 ***
## caffeine         0.0011175  0.0001124   9.940  < 2e-16 ***
## smokes           0.0381292  0.0163710   2.329   0.0201 *
## chocolates       0.0281738  0.0182463   1.544   0.1229
## video_games      0.0612643  0.0097697   6.271 5.37e-10 ***
## television       0.1772134  0.0183617   9.651  < 2e-16 ***
## games_played    -0.0715033  0.0123516  -5.789 9.52e-09 ***
## leisure_time    -0.1757196  0.0045548 -38.579  < 2e-16 ***
## carAudi          0.0304460  0.0343548   0.886   0.3757
## carBMW          -0.0462763  0.0348701  -1.327   0.1848
## carCamry        -0.0236486  0.0349159  -0.677   0.4984
## carHonda         0.0413005  0.0348982   1.183   0.2369
## carInfinity      0.0208342  0.0342431   0.608   0.5431
## carNissan        0.0110828  0.0354996   0.312   0.7550
## carSubaru        0.0012675  0.0349978   0.036   0.9711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.282 on 982 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 4.071e+04 on 18 and 982 DF,  p-value: < 2.2e-16
```

```
with(df, boxplot(headaches ~ Physical_abuse ))
```



```
plot (age~headaches, border = Physical_abuse, data = df)
```

```
## Warning in plot.window(...): "border" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "border" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "border" is
## not a graphical parameter
```

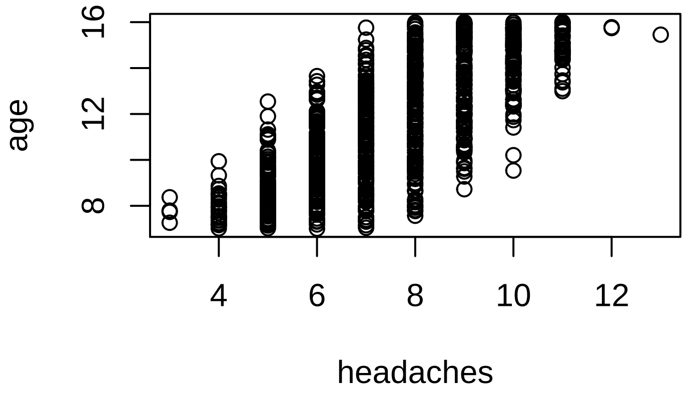
```
## Warning in axis(side = side, at = at, labels = labels, ...): "border" is
## not a graphical parameter
```

```
## Warning in box(...): "border" is not a graphical parameter
```

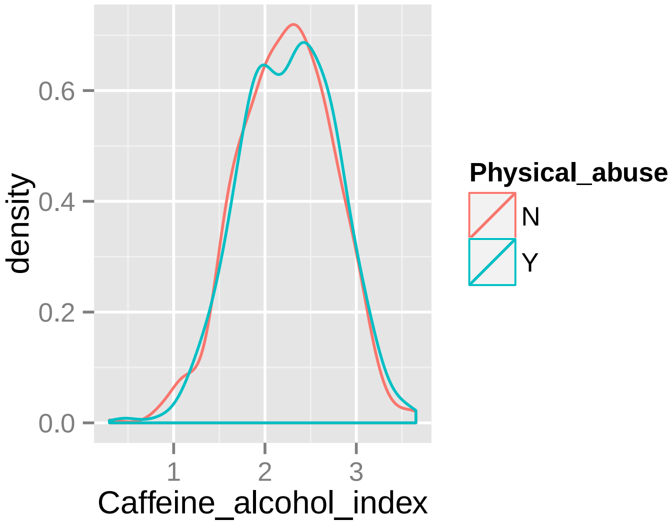
```
## Warning in title(...): "border" is not a graphical parameter
```

```
library(ggplot2)
```

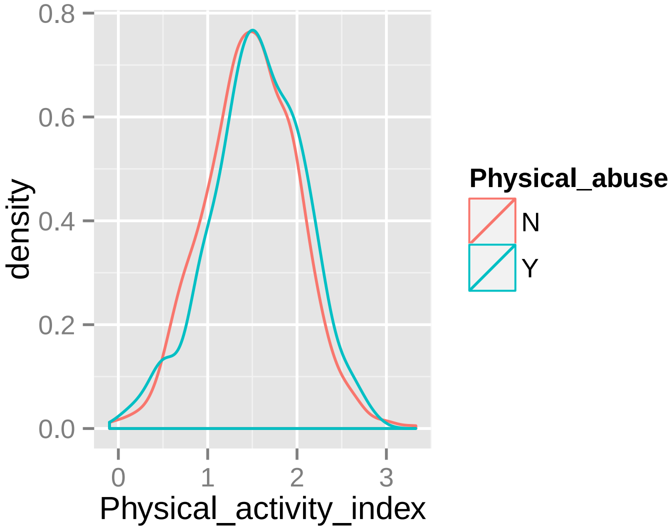
```
## Warning: package 'ggplot2' was built under R version 3.1.3
```



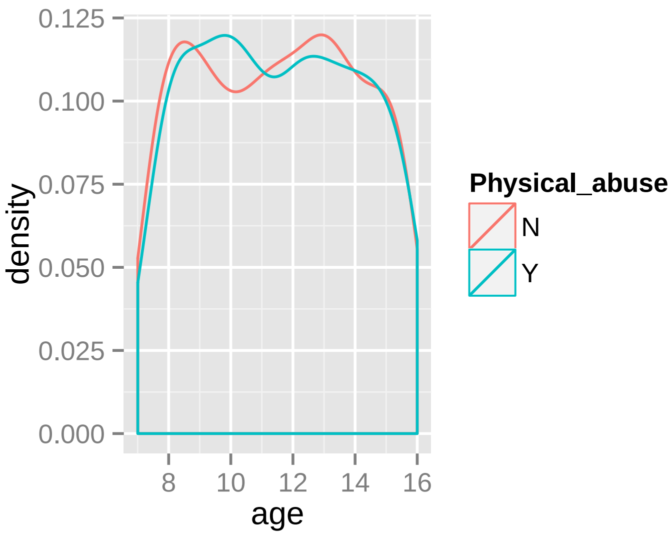
```
ggplot(df, aes(x = Caffeine_alcohol_index, col = Physical_abuse )) + geom_density()
```



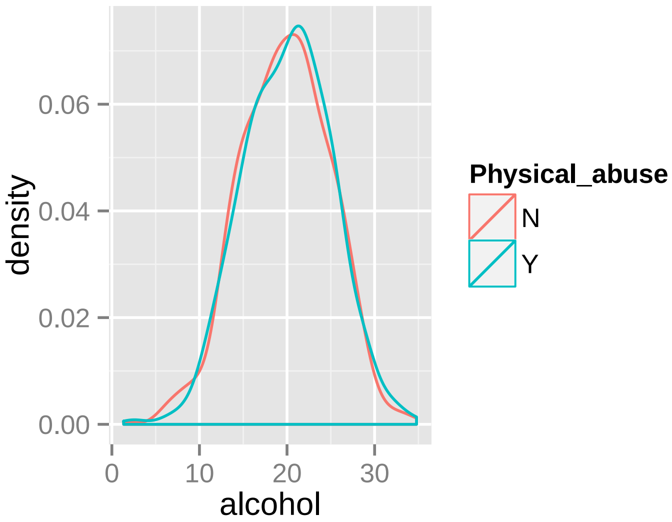
```
ggplot(df, aes(x = Physical_activity_index, col = Physical_abuse )) + geom_density()
```



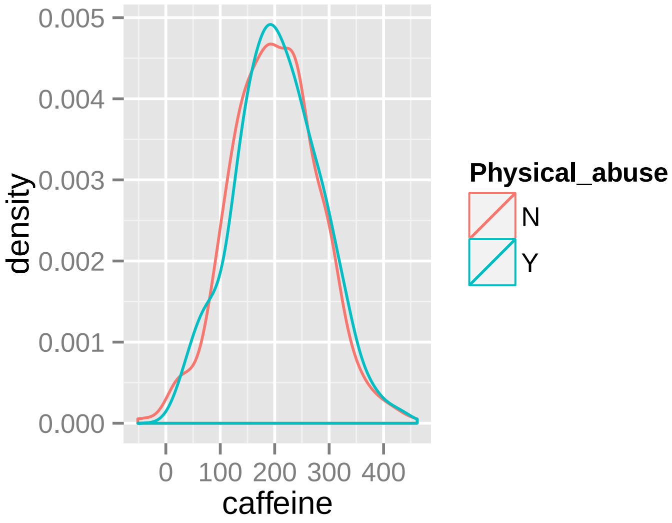
```
ggplot(df, aes(x = age, col = Physical_abuse  )) + geom_density()
```



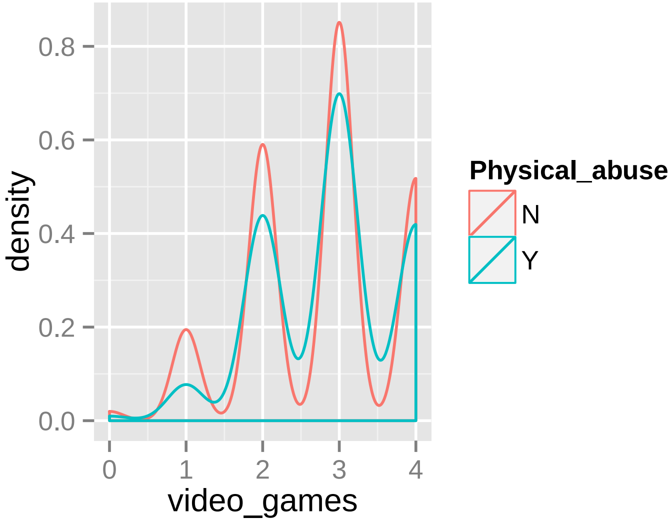
```
ggplot(df, aes(x = alcohol, col = Physical_abuse  )) + geom_density()
```



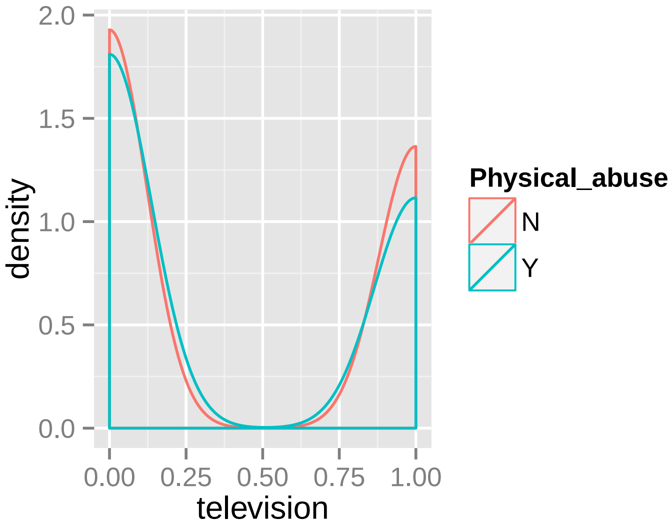
```
ggplot(df, aes(x = caffeine, col = Physical_abuse  )) + geom_density()
```



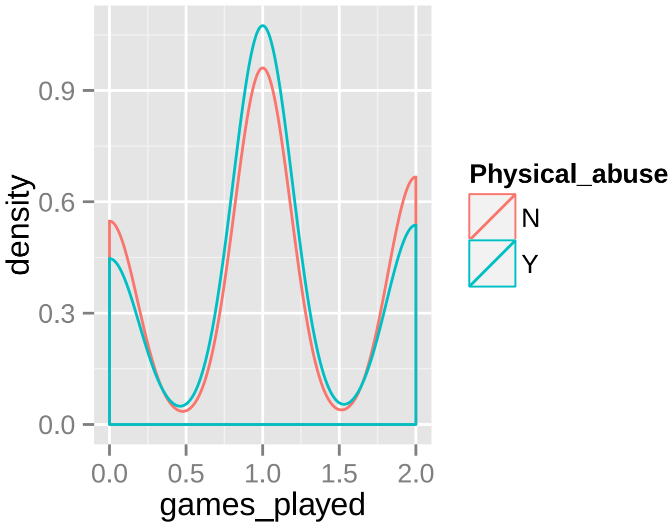
```
ggplot(df, aes(x = video_games, col = Physical_abuse  )) + geom_density()
```



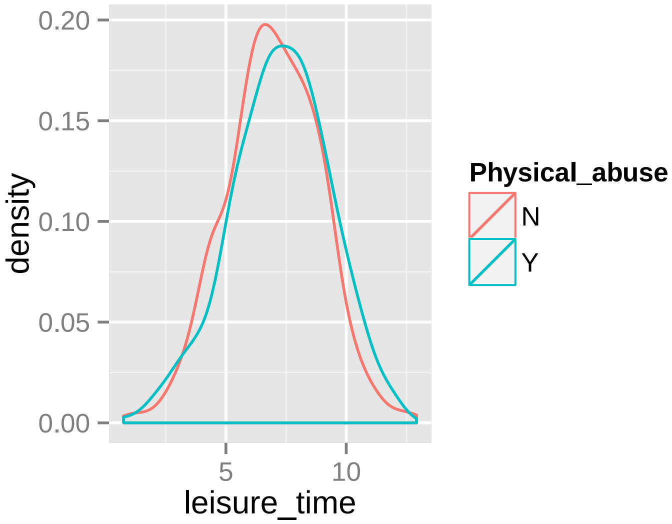
```
ggplot(df, aes(x = television, col = Physical_abuse  )) + geom_density()
```



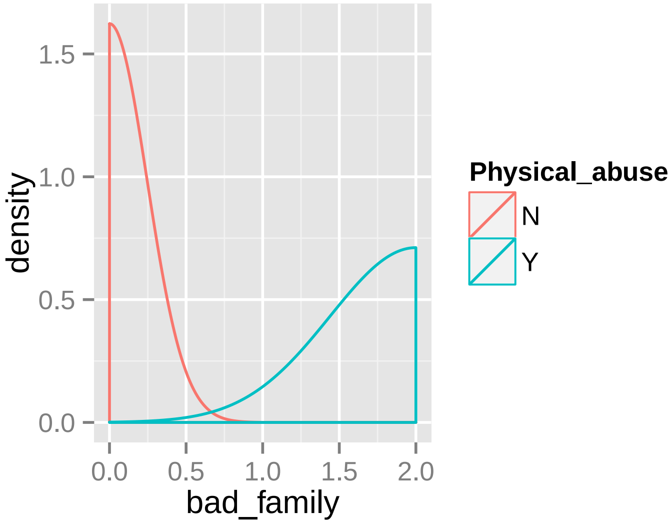
```
ggplot(df, aes(x = games_played, col = Physical_abuse  )) + geom_density()
```



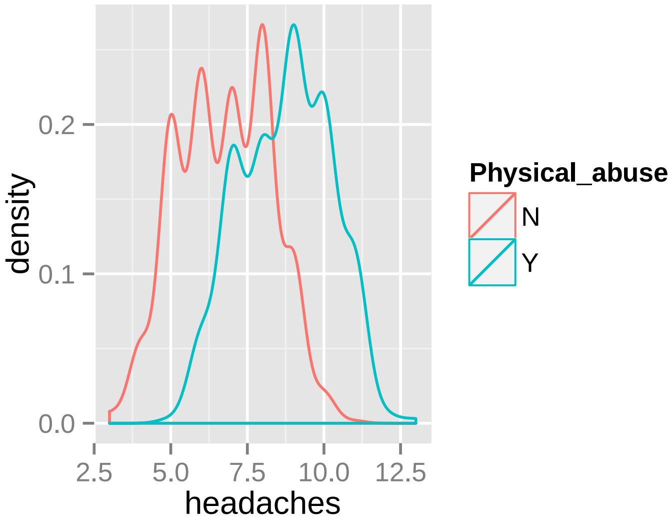
```
ggplot(df, aes(x = leisure_time, col = Physical_abuse  )) + geom_density()
```



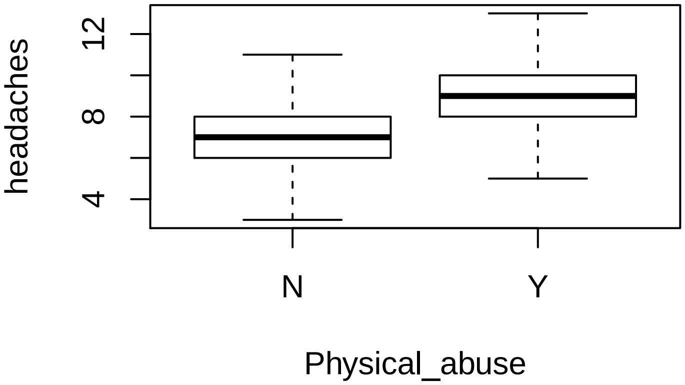
```
ggplot(df, aes(x = bad_family, col = Physical_abuse  )) + geom_density()
```



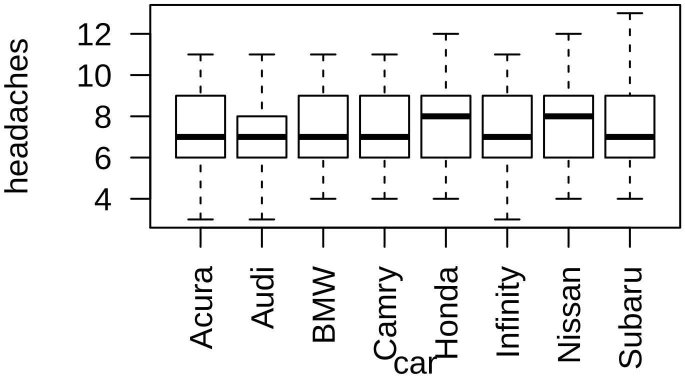
```
ggplot(df, aes(x = headaches, col = Physical_abuse  )) + geom_density()
```



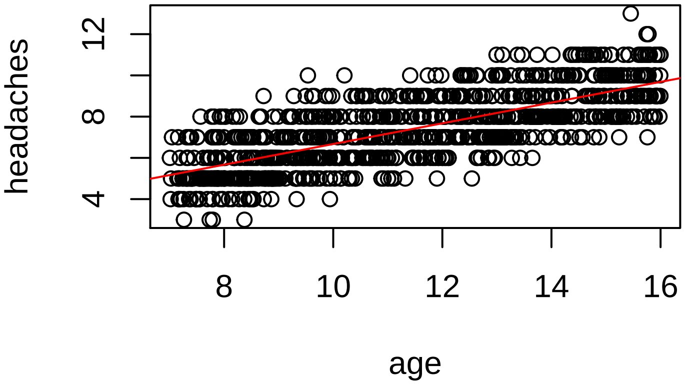
```
plot(headaches ~ Physical_abuse, data =df)
```



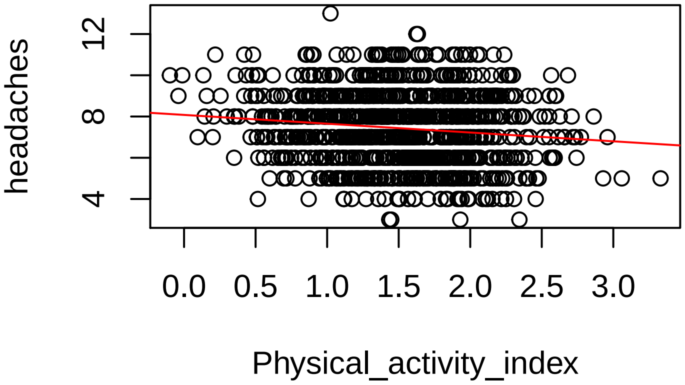
```
plot(headaches ~ car, data =df, las =2)
```



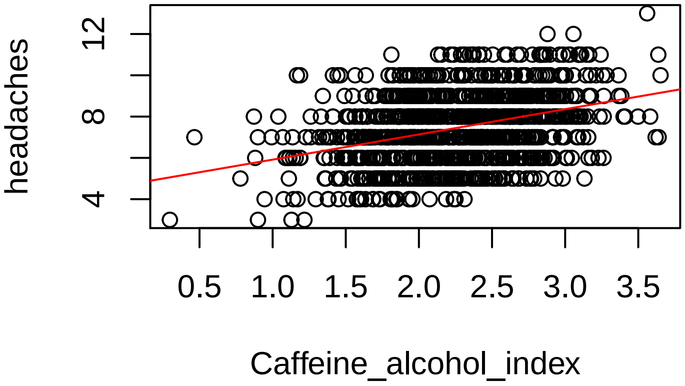
```
plot (headaches ~ age , data =df)
abline (lm(headaches ~ age , data =df), col ='red')
```



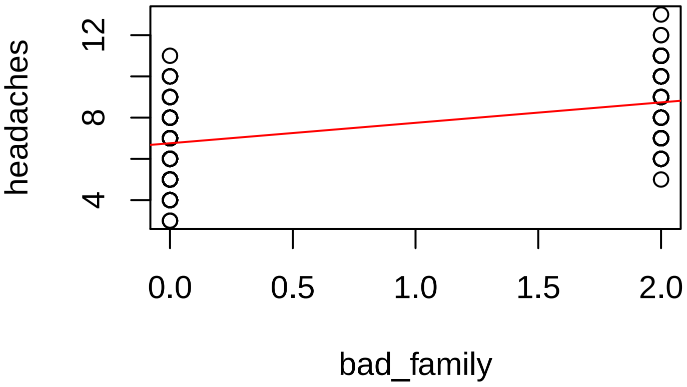
```
plot (headaches ~ Physical_activity_index , data =df)
abline (lm(headaches ~ Physical_activity_index , data =df), col ='red')
```



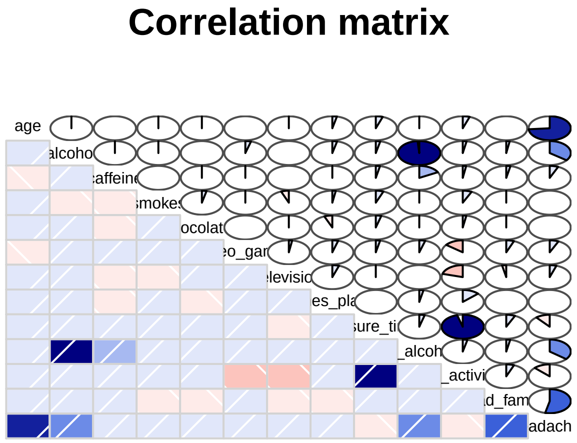
```
plot (headaches ~ Caffeine_alcohol_index , data =df)
abline (lm(headaches ~ Caffeine_alcohol_index , data =df), col ='red')
```



```
plot (headaches ~ bad_family , data =df)
abline (lm(headaches ~ bad_family , data =df), col ='red')
```

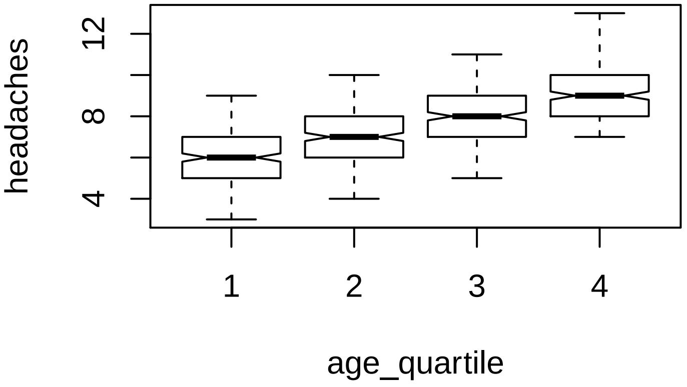


```
library(corrgram)
corrgram(df, order=FALSE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt,
  main="Correlation matrix")
```



Age effects can be analysed through the following graphs:

```
df$age_quartile <- with(df, cut(age, breaks=quantile(age, 0:4/4), labels=1:4))
plot(headaches ~ age_quartile, data=df, outline=F, notch=T)
```



Saving data in a csv

```r
dfnew <- df

colnames(dfnew) <- c("Age (in years)", "Alcohol intake per week (in mg)",
                     "Caffeine intake per day (in mg)", "Smokes per week",
                     "Chocolates per week","Video games owned", " No. tv sets in bedroom",
                     " No. of physical games played", "Leisure time per week (in hours)",
                     "Caffeine_alcohol_index", "Physical_activity_index",
                     "Physical abuse", "bad_family","Car make",
                     "No. of headaches per week" )

write.csv(dfnew,file="/Users/chamanpreetkaur/Desktop/data")
```

Story inspiration:

- http://www.ncbi.nlm.nih.gov/pubmed/24333367
- http://pediatrics.aappublications.org/content/133/3/386.full.pdf