Homework 3

Tauqeer Kasam Rumaney

November 2, 2022

Download this R Markdown file, save it on your computer, and perform all the below tasks by inserting your answer in text or by inserting R chunks below. After you are done, upload this file with your solutions on Moodle.

Exercise 1: Compute frequencies in the Pima diabetes dataset

Load the Pima diabetes dataset:

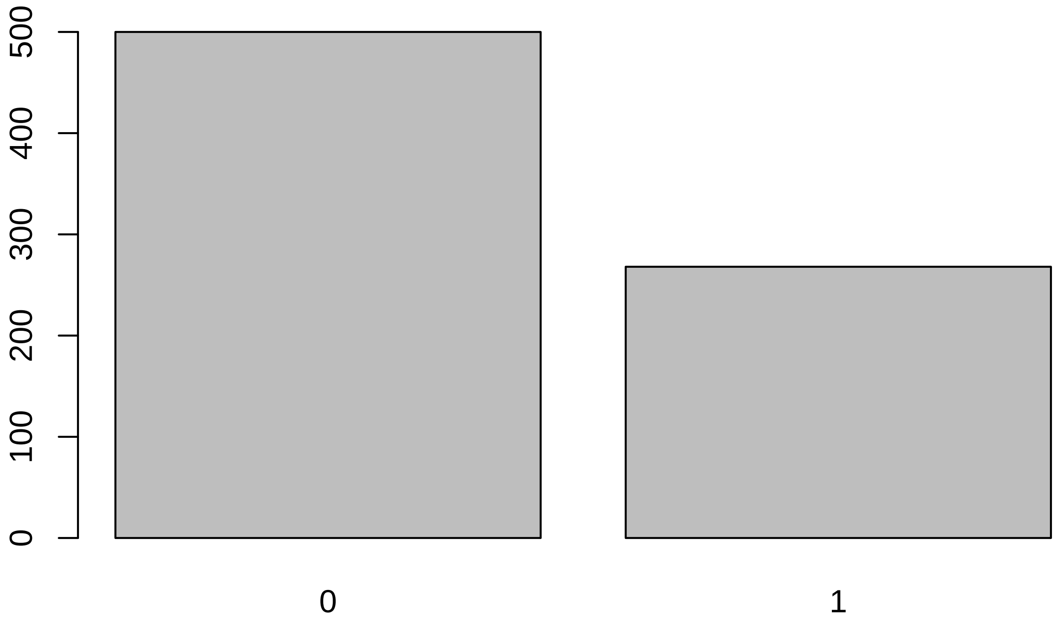
```
abc <- read.csv(file="/Users/tauqeerrumaney/BioStat/Pima_diabetes.csv")
```

Which variables are measured on a nominal level? # Outcome since other variables have a clear ordering Now compute frequency tables, barplots, and mosaic plots of all nominal variables in the dataset.

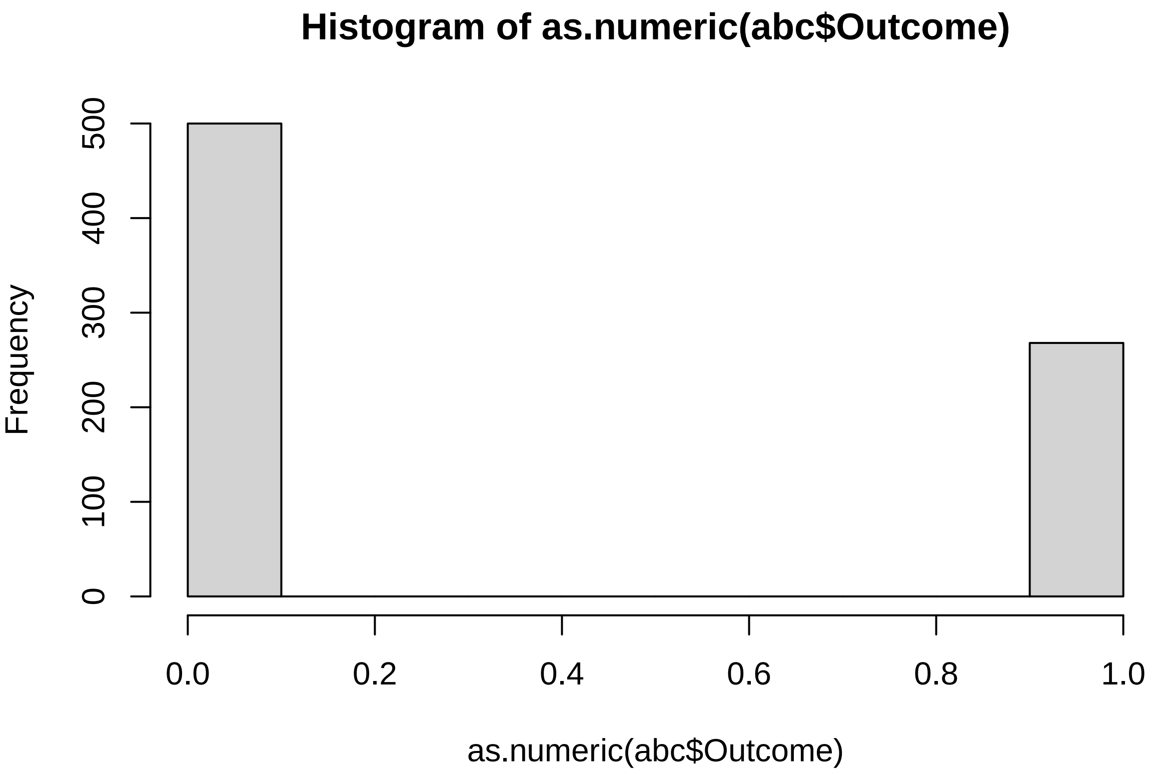
```
table(abc$Outcome)
```

```
##
##   0   1
## 500 268
```

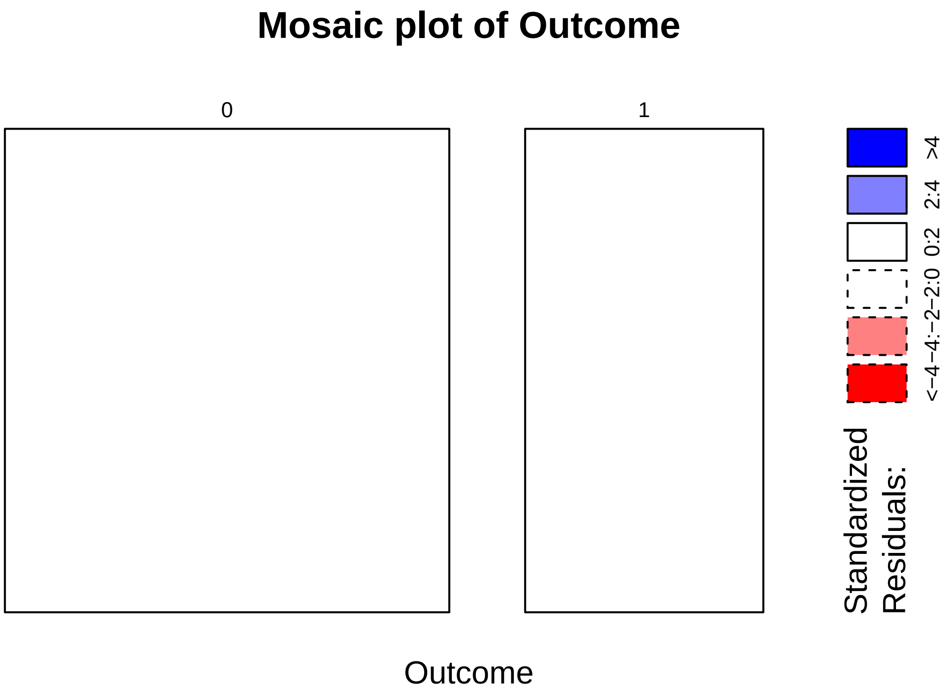
```
barplot(table(abc$Outcome))
```



```
hist(as.numeric(abc$Outcome))
```



```
table1 <- table(abc$Outcome)
mosaicplot(table1, shade = TRUE, main = "Mosaic plot of Outcome", xlab = "Outcome")
```



Next, create a variable which describes whether a woman had more or less than 4 pregnancies. Then, use this variable to create a 2x2 table with diabetes outcome. Do you see an indication of whether the number of pregnancies is associated with diabetes prevalence? Do you think your investigation is a good way to investigate this?

```
abc$P[(abc$Pregnancies <= 4)] <- 0
abc$P[(abc$Pregnancies > 4)] <- 1
table(abc$P)
```

```
##
##   0   1
## 492 276
```

```
table2 <- table(abc$P,abc$Outcome)
prop.table(table2, 1)
```

```
##
##             0         1
##   0 0.7235772 0.2764228
##   1 0.5217391 0.4782609
```

```
prop.table(table2, 2)
```

```
##
##             0         1
##   0 0.7120000 0.5074627
##   1 0.2880000 0.4925373
```

```
#install.packages("expss")
library(expss)
```

```
## Loading required package: maditr
```

```
##
## Use magrittr pipe '%>%' to chain several operations:
##              mtcars %>%
##                  let(mpg_hp = mpg/hp) %>%
##                  take(mean(mpg_hp), by = am)
##
```

```
expss::cro(abc$P,abc$Outcome)
```

abc$Outcome

0

1

abc$P

0

356

136

1

144

132

#Total cases

500

268

```
#helpdata <- abc[, c(9,10)]
#names(helpdata) <- c("Outcome", "Pregnancy")
#helpdata = apply_labels(helpdata,
                      #Outcome = "Is person Diabetic",
                      #Pregnancy = "Pregnancy greater than 4"
                      #)
#calculate(helpdata, cro(Outcome, Pregnancy)
#Error: Incomplete expression: calculate(helpdata, cro(Outcome, Pregnancy)
```

Exercise 2: Generate a table with descriptive statistics (optional, but recommended)

Use any dataset (a dataset that you have worked with in the past, or that you are currently working with, a dataset that is available on Blackboard, in R or that you have downloaded from the internet), and generate a table with descriptive statistics of the main variables of interest.

```
xyz <- read.csv(file="/Users/tauqeerrumaney/BioStat/Pima_diabetes.csv")
table(xyz$BloodPressure)
```

```
## 
##   0  24  30  38  40  44  46  48  50  52  54  55  56  58  60  61  62  64  65  66 
##  35   1   2   1   1   4   2   5  13  11  11   2  12  21  37   1  34  43   7  30 
##  68  70  72  74  75  76  78  80  82  84  85  86  88  90  92  94  95  96  98 100 
##  45  57  44  52   8  39  45  40  30  23   6  21  25  22   8   6   1   4   3   3 
## 102 104 106 108 110 114 122 
##   1   2   3   2   3   1   1 
```

```
n <- nrow(xyz)
table(xyz$BloodPressure)/n
```

```
## 
##           0          24          30          38          40          44 
## 0.045572917 0.001302083 0.002604167 0.001302083 0.001302083 0.005208333 
##          46          48          50          52          54          55 
## 0.002604167 0.006510417 0.016927083 0.014322917 0.014322917 0.002604167 
##          56          58          60          61          62          64 
## 0.015625000 0.027343750 0.048177083 0.001302083 0.044270833 0.055989583 
##          65          66          68          70          72          74 
## 0.009114583 0.039062500 0.058593750 0.074218750 0.057291667 0.067708333 
##          75          76          78          80          82          84 
## 0.010416667 0.050781250 0.058593750 0.052083333 0.039062500 0.029947917 
##          85          86          88          90          92          94 
## 0.007812500 0.027343750 0.032552083 0.028645833 0.010416667 0.007812500 
##          95          96          98         100         102         104 
## 0.001302083 0.005208333 0.003906250 0.003906250 0.001302083 0.002604167 
##         106         108         110         114         122 
## 0.003906250 0.002604167 0.003906250 0.001302083 0.001302083 
```

```
prop.table(table(xyz$BloodPressure))
```

```
## 
##           0          24          30          38          40          44 
## 0.045572917 0.001302083 0.002604167 0.001302083 0.001302083 0.005208333 
##          46          48          50          52          54          55 
## 0.002604167 0.006510417 0.016927083 0.014322917 0.014322917 0.002604167 
##          56          58          60          61          62          64
## 0.015625000 0.027343750 0.048177083 0.001302083 0.044270833 0.055989583
##          65          66          68          70          72          74
## 0.009114583 0.039062500 0.058593750 0.074218750 0.057291667 0.067708333
##          75          76          78          80          82          84
## 0.010416667 0.050781250 0.058593750 0.052083333 0.039062500 0.029947917
##          85          86          88          90          92          94
## 0.007812500 0.027343750 0.032552083 0.028645833 0.010416667 0.007812500
##          95          96          98         100         102         104
## 0.001302083 0.005208333 0.003906250 0.003906250 0.001302083 0.002604167
##         106         108         110         114         122
## 0.003906250 0.002604167 0.003906250 0.001302083 0.001302083
```

```
#table(xyz$BloodPressure, xyz$Glucose)
#expss::cro(xyz$BloodPressure, xyz$Glucose)
?table()
min(xyz$BloodPressure)
```

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

```
max(xyz$BloodPressure)
```

```
## [1] 122
```

```
range(xyz$BloodPressure)
```

```
## [1]   0 122
```

```
median(xyz$BloodPressure)
```

```
## [1] 72
```

```
mean(xyz$BloodPressure)
```

```
## [1] 69.10547
```

```
mode(xyz$BloodPressure)
```

```
## [1] "numeric"
```

```
mad(xyz$BloodPressure)
```

```
## [1] 11.8608
```

```
var(xyz$BloodPressure)
```

```
## [1] 374.6473
```

```
sd(xyz$BloodPressure)
```

```
## [1] 19.35581
```

```
quantile(xyz$BloodPressure, seq(0, 1, 0.25))
```

```
##   0%  25%  50%  75% 100% 
##    0   62   72   80  122 
```

```
summary(xyz)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness  
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
##     Insulin           BMI        DiabetesPedigreeFunction      Age       
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00  
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00  
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00  
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24  
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00  
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00  
##     Outcome     
##  Min.   :0.000  
##  1st Qu.:0.000  
##  Median :0.000  
##  Mean   :0.349  
##  3rd Qu.:1.000  
##  Max.   :1.000
```

```
library(fBasics)
skewness(xyz$BloodPressure)
```

```
## [1] -1.836413
## attr(,"method")
## [1] "moment"
```

```
kurtosis(xyz$BloodPressure)
```

```
## [1] 5.11751
## attr(,"method")
## [1] "excess"
```

```
tapply(xyz$BloodPressure, xyz$Glucose, mean, na.rm = TRUE)
```

```
##        0       44       56       57       61       62       65       67 
## 67.60000 62.00000 56.00000 70.00000 82.00000 78.00000 72.00000 76.00000 
##       68       71       72       73       74       75       76       77 
## 79.33333 64.50000 78.00000 36.66667 47.50000 73.00000 61.00000 69.00000 
##       78       79       80       81       82       83       84       85 
## 64.00000 71.66667 59.50000 73.66667 62.00000 70.16667 63.60000 67.14286 
##       86       87       88       89       90       91       92       93 
## 67.33333 57.71429 64.22222 65.33333 65.54545 60.88889 67.77778 65.71429 
##       94       95       96       97       98       99      100      101 
## 59.28571 69.46154 64.75000 68.22222 66.66667 58.58824 66.94118 66.11111 
##      102      103      104      105      106      107      108      109 
## 73.46154 69.11111 70.66667 69.15385 69.28571 69.63636 65.38462 67.41667 
##      110      111      112      113      114      115      116      117 
## 75.33333 71.07143 75.30769 62.80000 64.63636 60.40000 63.14286 71.27273 
##      118      119      120      121      122      123      124      125 
## 71.33333 48.36364 67.09091 69.33333 70.50000 74.00000 70.00000 73.28571 
##      126      127      128      129      130      131      132      133 
## 78.22222 75.60000 75.09091 67.57143 74.28571 57.20000 64.80000 82.80000 
##      134      135      136      137      138      139      140      141 
## 70.66667 54.00000 78.50000 75.62500 58.40000 65.37500 80.00000 43.20000 
##      142      143      144      145      146      147      148      149 
## 79.60000 80.33333 73.71429 66.40000 67.66667 80.28571 70.50000 68.00000
##      150      151      152      153      154      155      156      157
## 73.33333 75.00000 84.50000 85.00000 73.66667 69.60000 82.33333 73.00000
##      158      159      160      161      162      163      164      165
## 80.75000 65.00000 54.00000 68.00000 75.66667 71.33333 81.33333 80.50000
##      166      167      168      169      170      171      172      173
## 74.00000 60.00000 78.50000 74.00000 69.00000 84.66667 68.00000 77.66667
##      174      175      176      177      178      179      180      181
## 73.00000 75.00000 88.00000 60.00000 84.00000 75.40000 59.60000 76.40000
##      182      183      184      186      187      188      189      190
## 74.00000 52.66667 82.33333 90.00000 66.00000 80.00000 84.50000 92.00000
##      191      193      194      195      196      197      198      199
## 68.00000 60.00000 75.33333 70.00000 80.66667 71.00000 66.00000 76.00000
```

Exercise 3: Plots using ggplot2

Load the NoShow dataset:

```
load(file = url("https://www.dropbox.com/s/4oqg79cn1qfnhsh/NoShowdata.RData?dl=1"))
head(NoShowdata)
```

```
##      PatientId AppointmentID Gender        ScheduledDay AppointmentDay Age
## 1 2.987250e+13       5642903      F 2016-04-29 18:38:08     2016-04-29  62
## 2 5.589978e+14       5642503      M 2016-04-29 16:08:27     2016-04-29  56
## 3 4.262962e+12       5642549      F 2016-04-29 16:19:04     2016-04-29  62
## 4 8.679512e+11       5642828      F 2016-04-29 17:29:31     2016-04-29   8
## 5 8.841186e+12       5642494      F 2016-04-29 16:07:23     2016-04-29  56
## 6 9.598513e+13       5626772      F 2016-04-27 08:36:51     2016-04-29  76
##       Neighbourhood Scholarship Hipertension Diabetes Alcoholism Handcap
## 1   JARDIM DA PENHA           0            1        0          0       0
## 2   JARDIM DA PENHA           0            0        0          0       0
## 3     MATA DA PRAIA           0            0        0          0       0
## 4 PONTAL DE CAMBURI           0            0        0          0       0
## 5   JARDIM DA PENHA           0            1        1          0       0
## 6          REPÚBLICA          0            1        0          0       0
##   SMS_received No-show
## 1            0      No
## 2            0      No
## 3            0      No
## 4            0      No
## 5            0      No
## 6            0      No
```

Use ggplot2 to generate the following plots:

- Create a boxplots of Age (stratified) by neighborhood.
- Create a histogram of Age.
- Create a histogram of Age, stratified by whether the person showed up - in one panel using the or in multiple panels.
- Stratify this plot further by gender.

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:expss':
##
```

```
##     vars
```

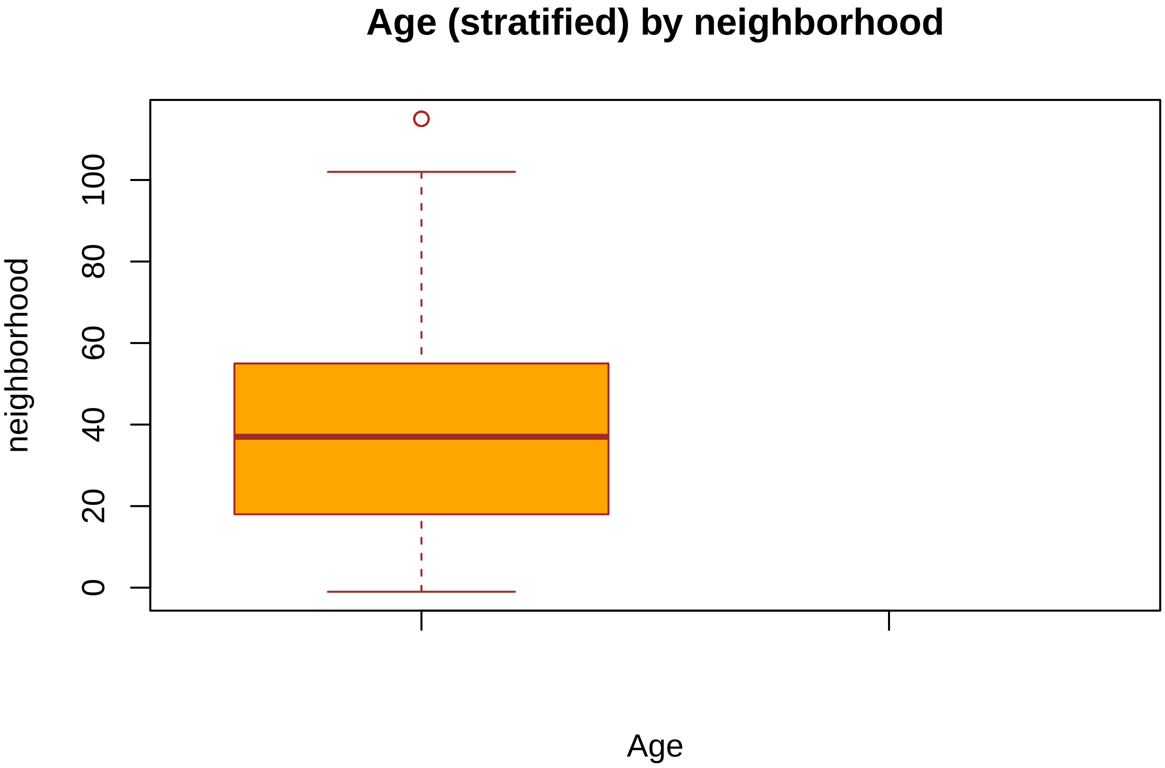
```
age <- NoShowdata$Age
neighborhood <- NoShowdata$Neighbourhood
neighborhood_norm <- rnorm(110527,mean=mean(neighborhood, na.rm=TRUE), sd=sd(neighborhood, na.rm=TRUE))
```

```
## Warning in mean.default(neighborhood, na.rm = TRUE): argument is not numeric or
## logical: returning NA
```

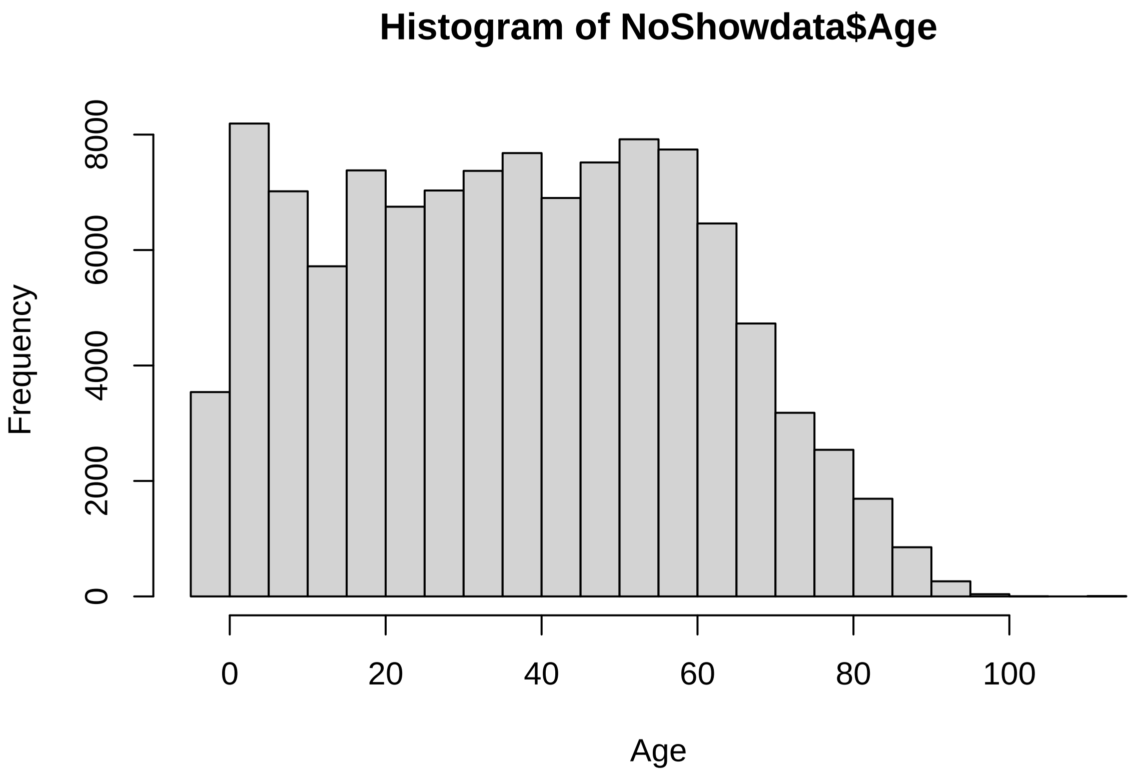
```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion
```

```
## Warning in rnorm(110527, mean = mean(neighborhood, na.rm = TRUE), sd =
## sd(neighborhood, : NAs produced
```

```
boxplot(age,neighborhood_norm, main = "Age (stratified) by neighborhood",xlab = "Age",ylab = "neighborh
```



```
hist(NoShowdata$Age, xlab ="Age")
```

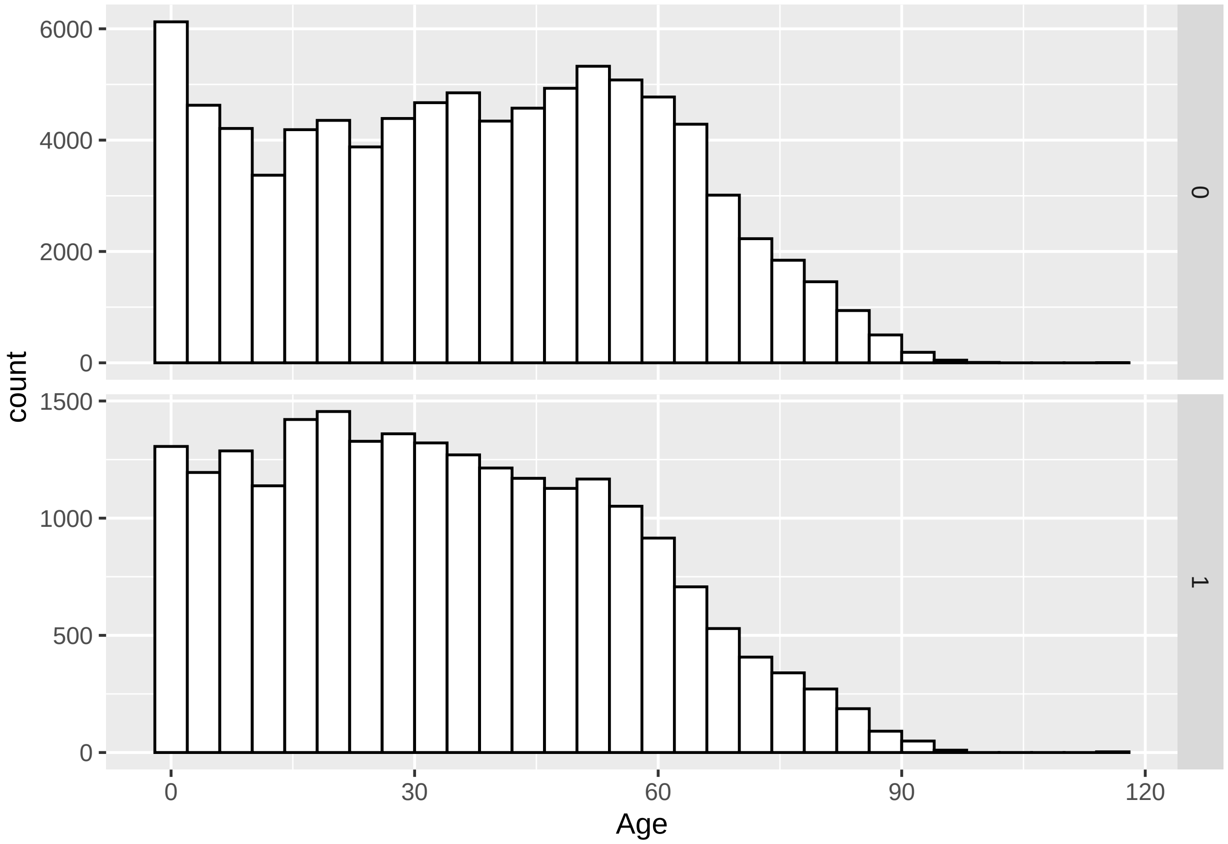


```
NoShowdata$NoShow[(NoShowdata$`No-show` == "No")] <- 0
```

```
## Warning: Unknown or uninitialised column: `NoShow`.
```

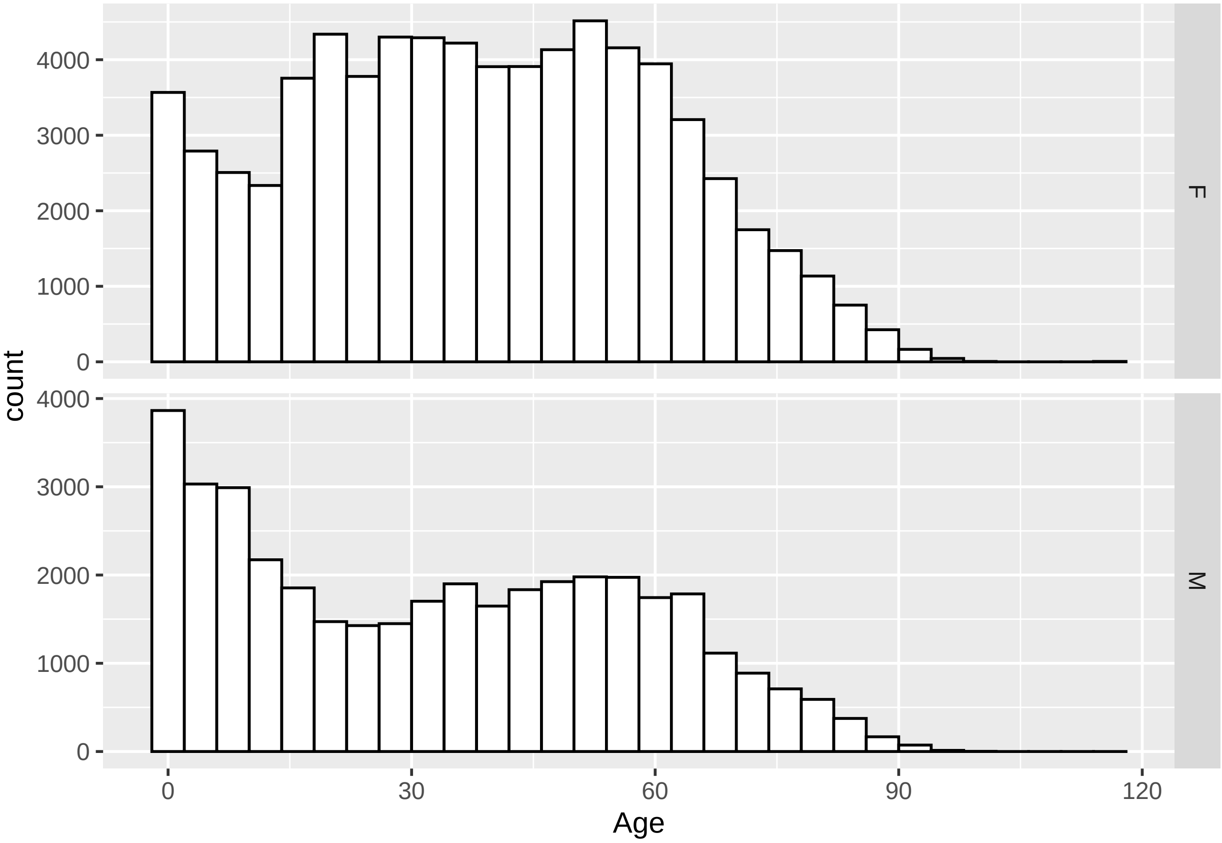
```
NoShowdata$NoShow[(NoShowdata$`No-show` == "Yes")] <- 1

ggplot(NoShowdata, aes(x = Age)) +
  geom_histogram(fill = "white", colour = "black") +
  facet_grid(NoShow ~ ., scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(NoShowdata, aes(x = Age)) +
  geom_histogram(fill = "white", colour = "black") + facet_grid(Gender ~ ., scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



What insights can you get from these plots? For which goal would you create these plots? #This plots allows to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.The goal is to ease our task of Statistics as well as to find relations between differnet variables and the factor which influence those variables.