Project code

Jessica Hin (008515095)

2023-06-04

setwd('/Users/stephenreagin/Desktop/ADS503\_project/ADS503')  
data <- read.csv('risk\_factors\_cervical\_cancer.csv')  
  
library(MASS)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ggplot2)  
library(reshape2)

# checking for zero variance predictors  
  
# Hinselmann test shows up in the results, but the other three tests do not (Schiller,Cytology,Biopsy)  
data\_nearzero <- as.integer(nearZeroVar(data))  
  
# remove IUD column  
#data <- data[,IUD]   
  
# checking for missing values  
data\_miss <- sapply(data, function(x) sum(is.na(x)))  
  
# data has question marks in data, finding out which rows have columns with question marks in it  
q\_idx <- as.data.frame(unlist(which(data == "?", arr.ind = TRUE)))  
rows <- unique(unlist(q\_idx[1]))  
  
# subsetting the data set with the rows that have columns with question marks in it  
data\_test <- data[c(rows),]  
  
# data frame with stds included  
data\_wstd <- data  
  
# subbing in the values with question marks as NAs  
data[data == "?"] <- NA  
  
# checking for missing values again  
data\_miss <- sapply(data, function(x) sum(is.na(x)))  
  
# removing the std cols  
data <- subset(data, select = -c(STDs..Time.since.first.diagnosis, STDs..Time.since.last.diagnosis))  
sapply(data, function(x) sum(is.na(x)))

## Age Number.of.sexual.partners   
## 0 26   
## First.sexual.intercourse Num.of.pregnancies   
## 7 56   
## Smokes Smokes..years.   
## 13 13   
## Smokes..packs.year. Hormonal.Contraceptives   
## 13 108   
## Hormonal.Contraceptives..years. IUD   
## 108 117   
## IUD..years. STDs   
## 117 105   
## STDs..number. STDs.condylomatosis   
## 105 105   
## STDs.cervical.condylomatosis STDs.vaginal.condylomatosis   
## 105 105   
## STDs.vulvo.perineal.condylomatosis STDs.syphilis   
## 105 105   
## STDs.pelvic.inflammatory.disease STDs.genital.herpes   
## 105 105   
## STDs.molluscum.contagiosum STDs.AIDS   
## 105 105   
## STDs.HIV STDs.Hepatitis.B   
## 105 105   
## STDs.HPV STDs..Number.of.diagnosis   
## 105 0   
## Dx.Cancer Dx.CIN   
## 0 0   
## Dx.HPV Dx   
## 0 0   
## Hinselmann Schiller   
## 0 0   
## Citology Biopsy   
## 0 0

# changing character variables to integers  
data[,2:27] <- sapply(data[,2:27], as.integer)  
  
# Changing certain variables to factors  
#data[,c(5,8,10,12,14:25,27:34)] <- sapply(data[,c(5,8,10,12,14:25,27:34)], as.factor)

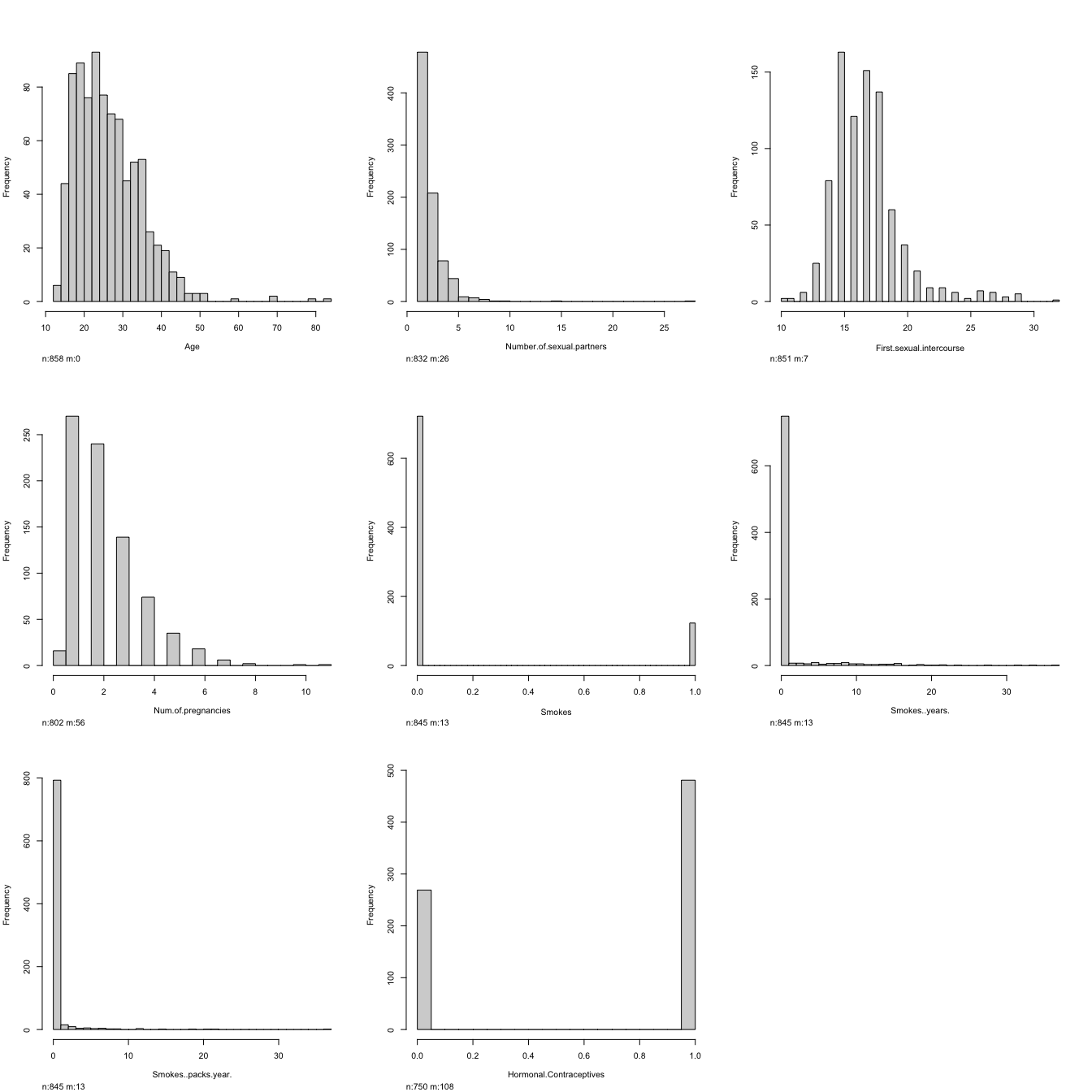
# Histograms of data  
library(Hmisc)

##   
## Attaching package: 'Hmisc'

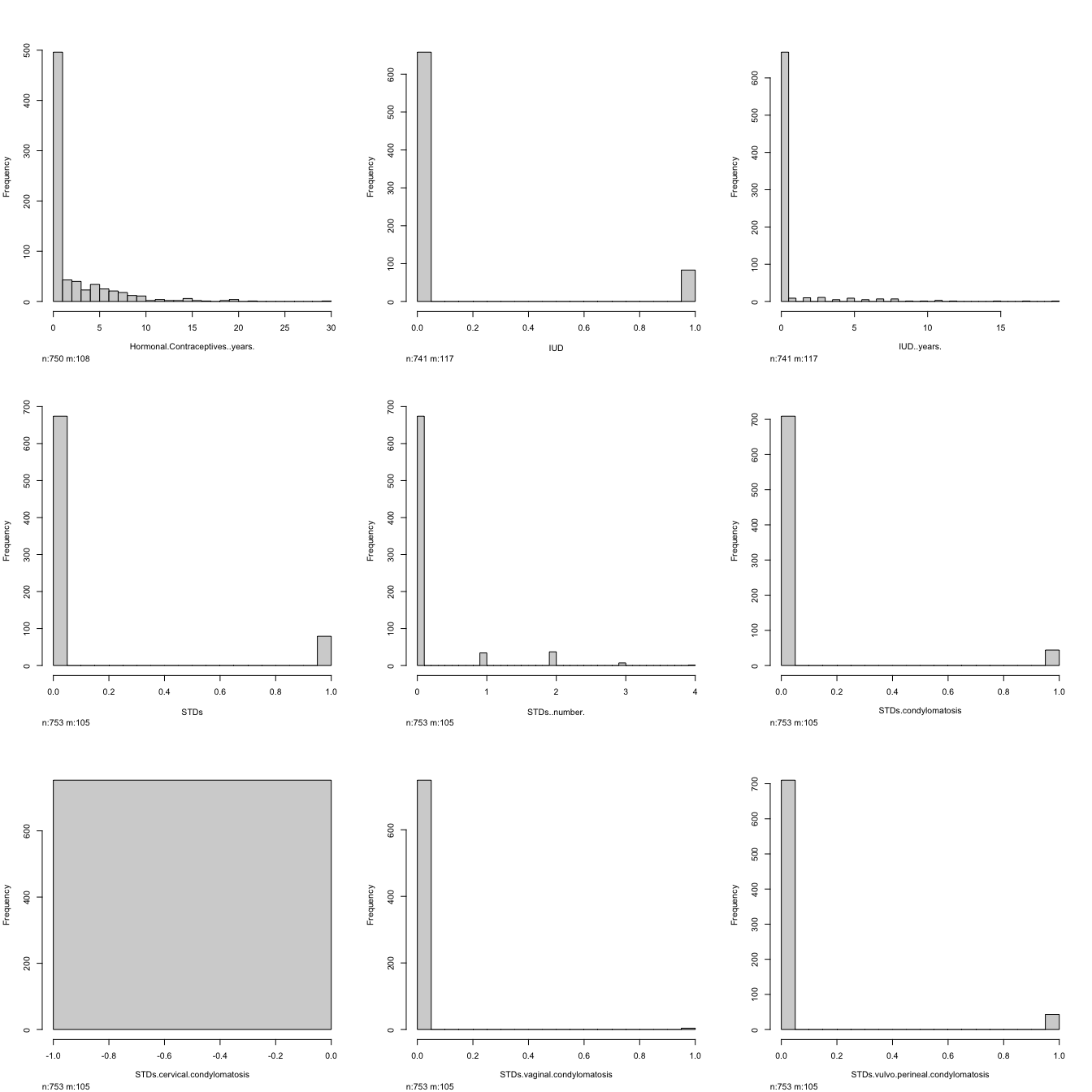
## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

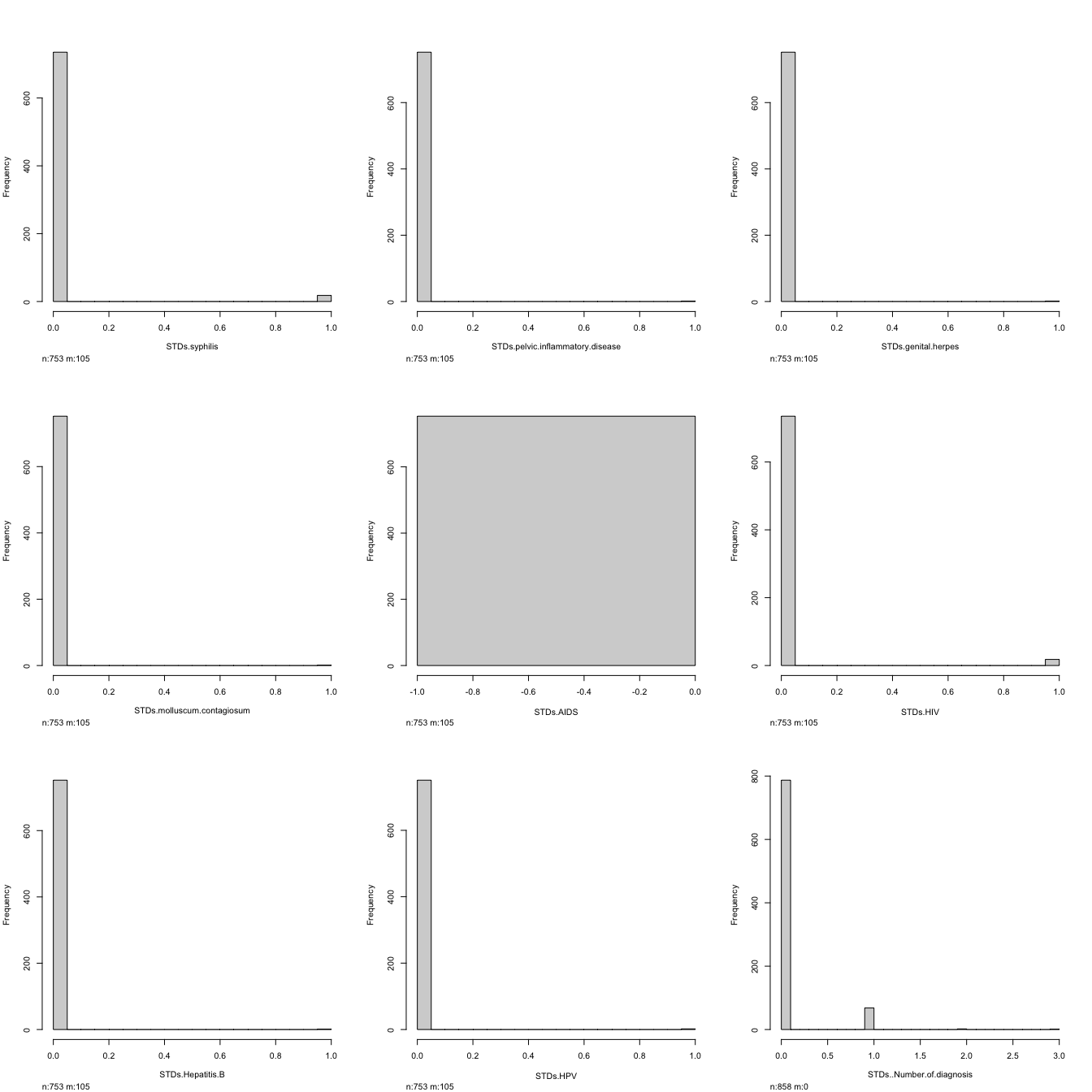
hist.data.frame(data[1:8], n.unique=1)



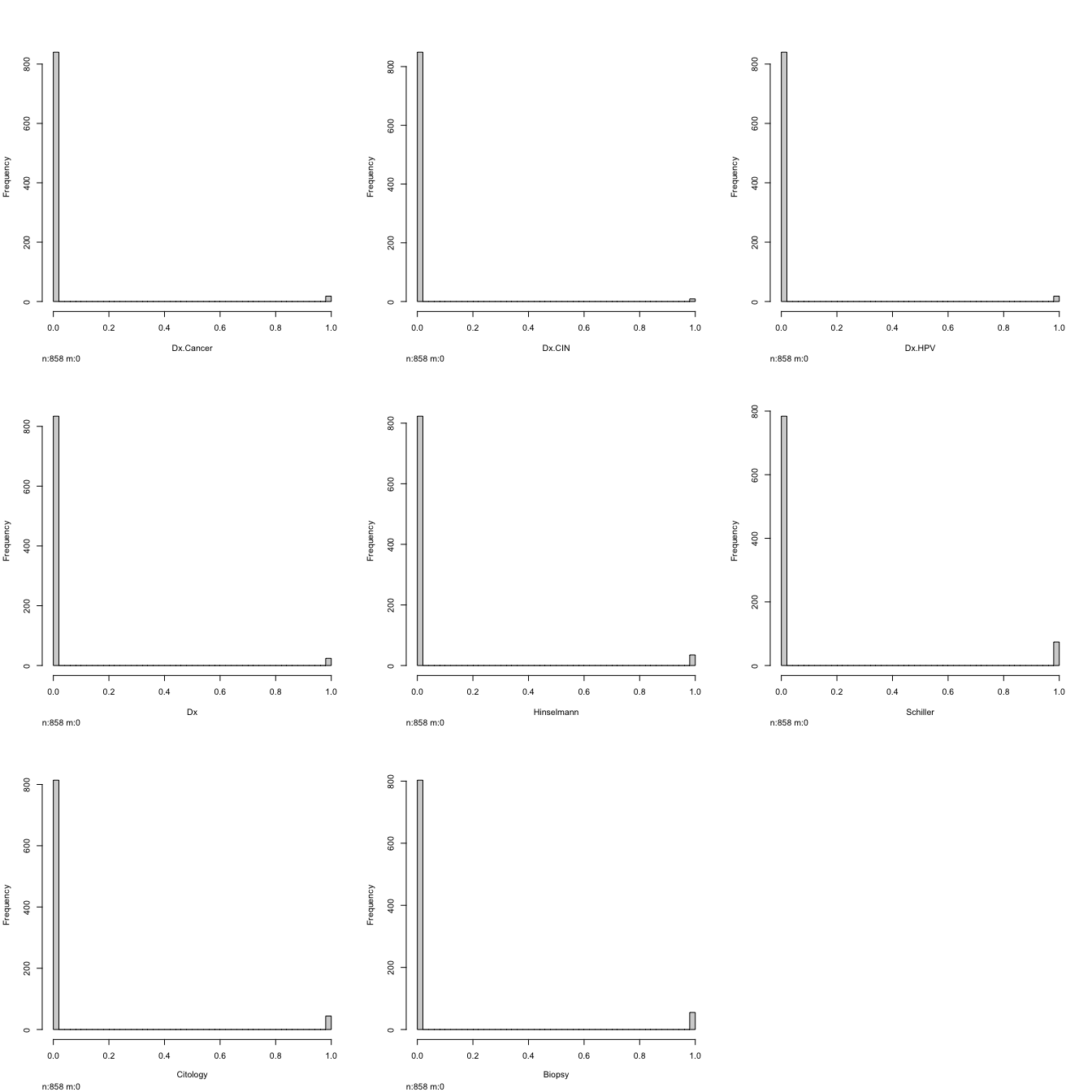
hist.data.frame(data[9:17], n.unique=1)



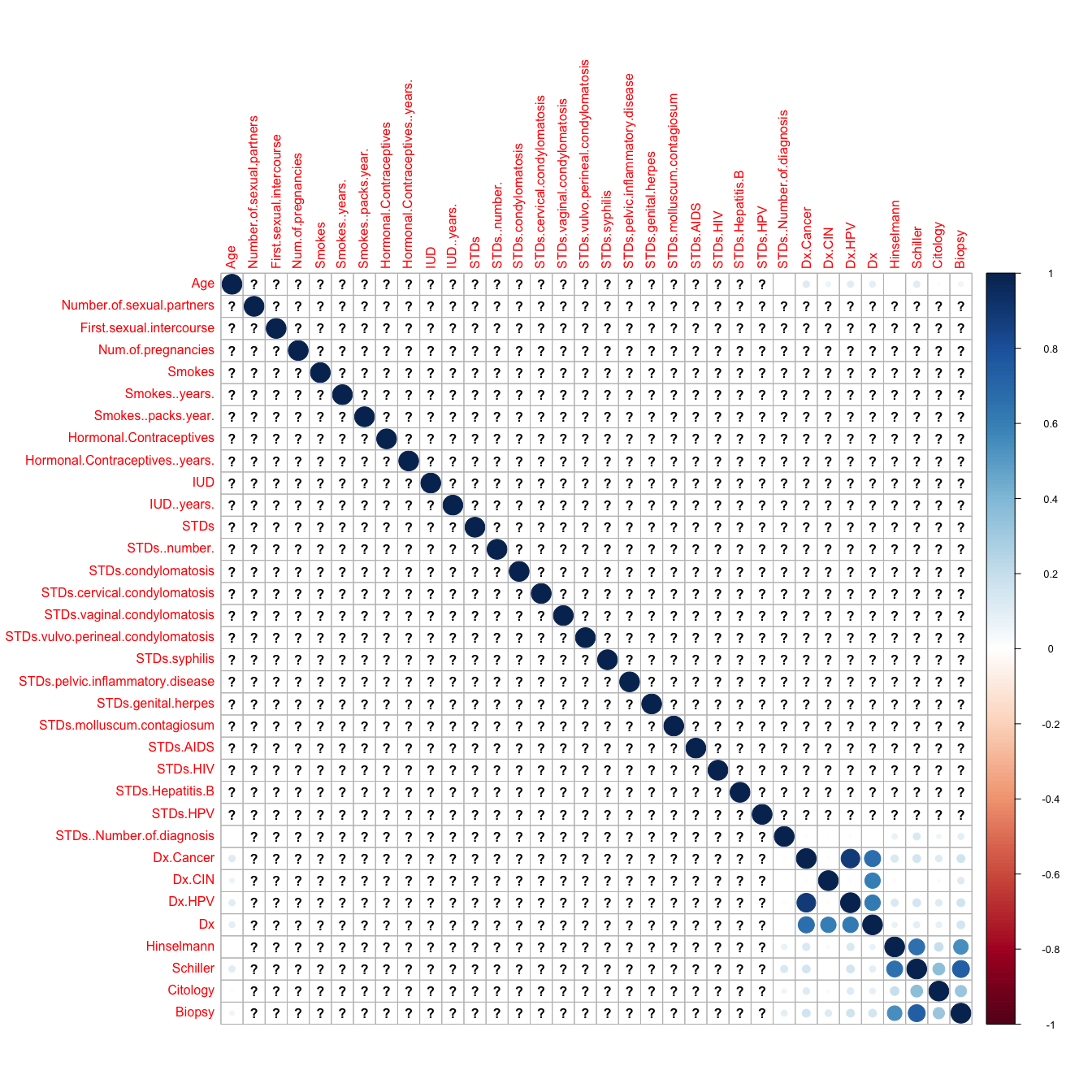
hist.data.frame(data[18:26], n.unique=1)



hist.data.frame(data[27:34], n.unique=1)



# Heatmap of the variables  
#heatmap(as.matrix(data))  
#melt(cor(data))  
#ggplot(melt(cor(data)),aes(x = Var1, y = Var2),  
# fill = value) + geom\_tile()  
  
# correlation plot for the variables  
corrplot::corrplot(cor(data))



library(psych)

##   
## Attaching package: 'psych'

## The following object is masked from 'package:Hmisc':  
##   
## describe

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

tetrachoric(data[,-c(1:4,6:7,9,11,13,26)])

## Warning in tetrachoric(data[, -c(1:4, 6:7, 9, 11, 13, 26)]): Item =  
## STDs.cervical.condylomatosis had no variance and was deleted

## Warning in tetrachoric(data[, -c(1:4, 6:7, 9, 11, 13, 26)]): Item = STDs.AIDS  
## had no variance and was deleted

## Warning in cor.smooth(mat): Matrix was not positive definite, smoothing was  
## done

## Call: tetrachoric(x = data[, -c(1:4, 6:7, 9, 11, 13, 26)])  
## tetrachoric correlation   
## Smoks Hrm.C IUD STDs STDs.c STDs.v.  
## Smokes 1.00   
## Hormonal.Contraceptives -0.01 1.00   
## IUD -0.17 0.07 1.00   
## STDs 0.26 -0.07 0.10 1.00   
## STDs.condylomatosis 0.19 -0.04 0.22 0.78 1.00   
## STDs.vaginal.condylomatosis 0.32 -0.30 0.19 0.59 0.73 1.00   
## STDs.vulvo.perineal.condylomatosis 0.19 -0.04 0.19 0.79 0.98 0.69   
## STDs.syphilis 0.27 -0.01 -0.10 0.63 0.18 0.28   
## STDs.pelvic.inflammatory.disease 0.19 0.02 0.24 0.46 0.34 0.63   
## STDs.genital.herpes 0.20 0.01 0.23 0.47 0.34 0.61   
## STDs.molluscum.contagiosum 0.19 -0.23 0.23 0.46 0.33 0.65   
## STDs.HIV 0.21 -0.26 0.06 0.65 0.37 0.28   
## STDs.Hepatitis.B 0.41 -0.22 0.23 0.48 0.31 0.65   
## STDs.HPV 0.31 0.15 0.13 0.47 0.23 0.58   
## Dx.Cancer -0.04 0.08 0.32 0.02 -0.12 0.24   
## Dx.CIN -0.17 -0.04 0.28 0.08 0.01 0.33   
## Dx.HPV 0.06 0.13 0.21 0.01 -0.12 0.24   
## Dx -0.32 0.00 0.39 -0.05 -0.16 0.20   
## Hinselmann 0.13 0.08 0.13 0.15 0.19 0.12   
## Schiller 0.11 0.00 0.19 0.24 0.24 0.05   
## Citology -0.01 -0.03 0.03 0.13 0.21 0.12   
## Biopsy 0.05 0.03 0.15 0.18 0.22 0.13   
## STDs.v.. STDs.s STDs.p.. STDs.g. STDs.m.  
## Smokes   
## Hormonal.Contraceptives   
## IUD   
## STDs   
## STDs.condylomatosis   
## STDs.vaginal.condylomatosis   
## STDs.vulvo.perineal.condylomatosis 1.00   
## STDs.syphilis 0.19 1.00   
## STDs.pelvic.inflammatory.disease 0.34 0.44 1.00   
## STDs.genital.herpes 0.34 0.42 0.80 1.00   
## STDs.molluscum.contagiosum 0.34 0.44 0.80 0.79 1.00   
## STDs.HIV 0.39 0.24 0.45 0.44 0.45   
## STDs.Hepatitis.B 0.31 0.41 0.80 0.77 0.80   
## STDs.HPV 0.24 0.39 0.72 0.70 0.71   
## Dx.Cancer -0.12 0.02 0.45 0.46 0.45   
## Dx.CIN 0.01 0.13 0.53 0.55 0.52   
## Dx.HPV -0.12 0.03 0.46 0.46 0.45   
## Dx -0.17 -0.03 0.44 0.43 0.44   
## Hinselmann 0.19 0.02 0.38 0.39 0.38   
## Schiller 0.26 0.00 0.26 0.27 0.27   
## Citology 0.22 -0.12 0.36 0.36 0.35   
## Biopsy 0.24 -0.12 0.31 0.50 0.32   
## STD.HI STD.H. STD.HP Dx.Cn D.CIN D.HPV Dx Hnslm Schll Ctlgy  
## STDs.HIV 1.00   
## STDs.Hepatitis.B 0.62 1.00   
## STDs.HPV 0.38 0.72 1.00   
## Dx.Cancer 0.03 0.43 0.72 1.00   
## Dx.CIN 0.29 0.51 0.40 0.22 1.00   
## Dx.HPV 0.03 0.44 0.75 0.98 0.22 1.00   
## Dx 0.10 0.42 0.57 0.78 0.71 0.76 1.00   
## Hinselmann 0.29 0.36 0.24 0.43 0.07 0.44 0.24 1.00   
## Schiller 0.33 0.28 0.16 0.41 0.03 0.42 0.28 0.89 1.00   
## Citology 0.28 0.34 0.23 0.39 0.03 0.39 0.26 0.49 0.65 1.00  
## Biopsy 0.38 0.33 0.24 0.40 0.32 0.42 0.42 0.80 0.87 0.62  
## Bipsy  
## STDs.HIV   
## STDs.Hepatitis.B   
## STDs.HPV   
## Dx.Cancer   
## Dx.CIN   
## Dx.HPV   
## Dx   
## Hinselmann   
## Schiller   
## Citology   
## Biopsy 1.00  
##   
## with tau of   
## Smokes Hormonal.Contraceptives   
## 1.06 -0.36   
## IUD STDs   
## 1.22 1.25   
## STDs.condylomatosis STDs.vaginal.condylomatosis   
## 1.57 2.55   
## STDs.vulvo.perineal.condylomatosis STDs.syphilis   
## 1.58 1.98   
## STDs.pelvic.inflammatory.disease STDs.genital.herpes   
## 3.00 3.00   
## STDs.molluscum.contagiosum STDs.HIV   
## 3.00 1.98   
## STDs.Hepatitis.B STDs.HPV   
## 3.00 2.79   
## Dx.Cancer Dx.CIN   
## 2.03 2.31   
## Dx.HPV Dx   
## 2.03 1.91   
## Hinselmann Schiller   
## 1.74 1.36   
## Citology Biopsy   
## 1.63 1.52