Data Formatting and Cleaning

Devan Morehouse

2023-10-03

This file uses the risk factors for cervival cancer data set found on this website.
https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
(https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29)

This code is designed for reading and analyzing the structure of the data that R has imported from this file.

```
stemp= "C:\\Users\\dev46\\OneDrive\\Desktop\\Fall 2023\\Data Stewardship\\Data\\risk_factors_cer
vical_cancer.csv"
risk_factors_cervical_cancer=read.csv(stemp)
str(risk_factors_cervical_cancer)
```

```
## 'data.frame':    858 obs. of  36 variables:
##  $ Age                                : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ Number.of.sexual.partners          : chr  "4.0" "1.0" "1.0" "5.0" ...
##  $ First.sexual.intercourse           : chr  "15.0" "14.0" "?" "16.0" ...
##  $ Num.of.pregnancies                 : chr  "1.0" "1.0" "1.0" "4.0" ...
##  $ Smokes                             : chr  "0.0" "0.0" "0.0" "1.0" ...
##  $ Smokes..years.                     : chr  "0.0" "0.0" "0.0" "37.0" ...
##  $ Smokes..packs.year.                : chr  "0.0" "0.0" "0.0" "37.0" ...
##  $ Hormonal.Contraceptives            : chr  "0.0" "0.0" "0.0" "1.0" ...
##  $ Hormonal.Contraceptives..years.    : chr  "0.0" "0.0" "0.0" "3.0" ...
##  $ IUD                                : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ IUD..years.                        : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs                               : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs..number.                      : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.condylomatosis                : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.cervical.condylomatosis       : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.vaginal.condylomatosis        : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.vulvo.perineal.condylomatosis: chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.syphilis                      : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.pelvic.inflammatory.disease   : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.genital.herpes                : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.molluscum.contagiosum         : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.AIDS                          : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.HIV                           : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.Hepatitis.B                   : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs.HPV                           : chr  "0.0" "0.0" "0.0" "0.0" ...
##  $ STDs..Number.of.diagnosis          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Time.since.first.diagnosis   : chr  "?" "?" "?" "?" ...
##  $ STDs..Time.since.last.diagnosis    : chr  "?" "?" "?" "?" ...
##  $ Dx.Cancer                          : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx.CIN                             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dx.HPV                             : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx                                 : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Hinselmann                         : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Schiller                           : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Citology                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Biopsy                             : int  0 0 0 0 0 0 1 0 0 0 ...
```

This code enables us to view the raw data from the initial rows.

```
rawString = readLines(stemp)
rawString[1:5]
```

```
## [1] "Age,Number of sexual partners,First sexual intercourse,Num of pregnancies,Smokes,Smokes
(years),Smokes (packs/year),Hormonal Contraceptives,Hormonal Contraceptives (years),IUD,IUD (yea
rs),STDs,STDs (number),STDs:condylomatosis,STDs:cervical condylomatosis,STDs:vaginal condylomato
sis,STDs:vulvo-perineal condylomatosis,STDs:syphilis,STDs:pelvic inflammatory disease,STDs:genit
al herpes,STDs:molluscum contagiosum,STDs:AIDS,STDs:HIV,STDs:Hepatitis B,STDs:HPV,STDs: Number o
f diagnosis,STDs: Time since first diagnosis,STDs: Time since last diagnosis,Dx:Cancer,Dx:CIN,D
x:HPV,Dx,Hinselmann,Schiller,Citology,Biopsy"
## [2] "18,4.0,15.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0,?,?,0,0,0,0,0,0,0,0"
## [3] "15,1.0,14.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0,?,?,0,0,0,0,0,0,0,0"
## [4] "34,1.0,?,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0,?,?,0,0,0,0,0,0,0,0"
## [5] "52,5.0,16.0,4.0,1.0,37.0,37.0,1.0,3.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0,?,?,1,0,1,0,0,0,0,0"
```

Noteworthy observations from this data include R's alteration of variable names, where spaces, colons, parentheses, and hyphens have been replaced with periods. This suggests that the data set is in a comma-separated values format, and it also indicates that NA values are represented as "?".

Once we have examined the data set and made our observations, we can seamlessly format it to the "read.table()" function in R, where we can then review its structure once again.

```
risk_factors_cervical_cancer = read.table(stemp, header=TRUE, na.strings = "?", sep=",")
head(risk_factors_cervical_cancer)
```

```
##   Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 1  18                         4                       15                 1
## 2  15                         1                       14                 1
## 3  34                         1                       NA                 1
## 4  52                         5                       16                 4
## 5  46                         3                       21                 4
## 6  42                         3                       23                 2
##   Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 1      0              0                   0                       0
## 2      0              0                   0                       0
## 3      0              0                   0                       0
## 4      1             37                  37                       1
## 5      0              0                   0                       1
## 6      0              0                   0                       0
##   Hormonal.Contraceptives..years. IUD IUD..years. STDs STDs..number.
## 1                               0   0           0    0             0
## 2                               0   0           0    0             0
## 3                               0   0           0    0             0
## 4                               3   0           0    0             0
## 5                              15   0           0    0             0
## 6                               0   0           0    0             0
##   STDs.condylomatosis STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
## 1                   0                            0                          0
## 2                   0                            0                          0
## 3                   0                            0                          0
## 4                   0                            0                          0
## 5                   0                            0                          0
## 6                   0                            0                          0
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                                  0             0
## 2                                  0             0
## 3                                  0             0
## 4                                  0             0
## 5                                  0             0
## 6                                  0             0
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                                0                   0
## 2                                0                   0
## 3                                0                   0
## 4                                0                   0
## 5                                0                   0
## 6                                0                   0
##   STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 1                          0         0        0                0        0
## 2                          0         0        0                0        0
## 3                          0         0        0                0        0
## 4                          0         0        0                0        0
## 5                          0         0        0                0        0
## 6                          0         0        0                0        0
##   STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
## 1                         0                              NA
## 2                         0                              NA
```

```
## 3                         0                                NA
## 4                         0                                NA
## 5                         0                                NA
## 6                         0                                NA
##   STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV Dx Hinselmann
## 1                              NA         0      0      0  0          0
## 2                              NA         0      0      0  0          0
## 3                              NA         0      0      0  0          0
## 4                              NA         1      0      1  0          0
## 5                              NA         0      0      0  0          0
## 6                              NA         0      0      0  0          0
##   Schiller Citology Biopsy
## 1        0        0      0
## 2        0        0      0
## 3        0        0      0
## 4        0        0      0
## 5        0        0      0
## 6        0        0      0
```

```
str(risk_factors_cervical_cancer)
```

```
## 'data.frame':    858 obs. of  36 variables:
##  $ Age                                 : int  18 15 34 52 46 42 51 26 45 44 ...
##  $ Number.of.sexual.partners           : num  4 1 1 5 3 3 3 1 1 3 ...
##  $ First.sexual.intercourse            : num  15 14 NA 16 21 23 17 26 20 15 ...
##  $ Num.of.pregnancies                  : num  1 1 1 4 4 2 6 3 5 NA ...
##  $ Smokes                              : num  0 0 0 1 0 0 1 0 0 1 ...
##  $ Smokes..years.                      : num  0 0 0 37 0 ...
##  $ Smokes..packs.year.                 : num  0 0 0 37 0 0 3.4 0 0 2.8 ...
##  $ Hormonal.Contraceptives             : num  0 0 0 1 1 0 0 1 0 0 ...
##  $ Hormonal.Contraceptives..years.     : num  0 0 0 3 15 0 0 2 0 0 ...
##  $ IUD                                 : num  0 0 0 0 0 0 1 1 0 NA ...
##  $ IUD..years.                         : num  0 0 0 0 0 0 7 7 0 NA ...
##  $ STDs                                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..number.                       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.condylomatosis                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.cervical.condylomatosis        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.vaginal.condylomatosis         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.vulvo.perineal.condylomatosis: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.syphilis                       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.pelvic.inflammatory.disease    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.genital.herpes                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.molluscum.contagiosum          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.AIDS                           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.HIV                            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.Hepatitis.B                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.HPV                            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Number.of.diagnosis           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs..Time.since.first.diagnosis    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ STDs..Time.since.last.diagnosis     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Dx.Cancer                           : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx.CIN                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dx.HPV                              : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ Dx                                  : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Hinselmann                          : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Schiller                            : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Citology                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Biopsy                              : int  0 0 0 0 0 0 1 0 0 0 ...
```

After visiting the website (https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29 (https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29)), we have identified certain column variables that should be factored into True or False labels. Below this is R code that defines a function to select and process the relevant columns accordingly.

```
Factor_List = c('Smokes','Hormonal.Contraceptives','IUD','STDs','STDs.condylomatosis','STDs.cervical.condylomatosis','STDs.vaginal.condylomatosis','STDs.vulvo.perineal.condylomatosis','STDs.syphilis','STDs.pelvic.inflammatory.disease','STDs.genital.herpes','STDs.molluscum.contagiosum','STDs.AIDS','STDs.HIV','STDs.Hepatitis.B','STDs.HPV','Dx.Cancer','Dx.CIN','Dx.HPV','Dx','Hinselmann','Schiller','Citology','Biopsy')

for(x in Factor_List){
  risk_factors_cervical_cancer[,x] <- factor(risk_factors_cervical_cancer[,x], labels=c('False','True'), levels=c(0,1))
}
head(risk_factors_cervical_cancer)
```

```
##   Age Number.of.sexual.partners First.sexual.intercourse Num.of.pregnancies
## 1  18                         4                       15                 1
## 2  15                         1                       14                 1
## 3  34                         1                       NA                 1
## 4  52                         5                       16                 4
## 5  46                         3                       21                 4
## 6  42                         3                       23                 2
##   Smokes Smokes..years. Smokes..packs.year. Hormonal.Contraceptives
## 1  False              0                   0                  False
## 2  False              0                   0                  False
## 3  False              0                   0                  False
## 4   True             37                  37                   True
## 5  False              0                   0                   True
## 6  False              0                   0                  False
##   Hormonal.Contraceptives..years.   IUD IUD..years.  STDs STDs..number.
## 1                               0 False           0 False             0
## 2                               0 False           0 False             0
## 3                               0 False           0 False             0
## 4                               3 False           0 False             0
## 5                              15 False           0 False             0
## 6                               0 False           0 False             0
##   STDs.condylomatosis STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
## 1               False                        False                       False
## 2               False                        False                       False
## 3               False                        False                       False
## 4               False                        False                       False
## 5               False                        False                       False
## 6               False                        False                       False
##   STDs.vulvo.perineal.condylomatosis STDs.syphilis
## 1                              False         False
## 2                              False         False
## 3                              False         False
## 4                              False         False
## 5                              False         False
## 6                              False         False
##   STDs.pelvic.inflammatory.disease STDs.genital.herpes
## 1                            False               False
## 2                            False               False
## 3                            False               False
## 4                            False               False
## 5                            False               False
## 6                            False               False
##   STDs.molluscum.contagiosum STDs.AIDS STDs.HIV STDs.Hepatitis.B STDs.HPV
## 1                      False     False    False            False    False
## 2                      False     False    False            False    False
## 3                      False     False    False            False    False
## 4                      False     False    False            False    False
## 5                      False     False    False            False    False
## 6                      False     False    False            False    False
##   STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis
## 1                         0                              NA
## 2                         0                              NA
```

```
## 3                         0                                  NA
## 4                         0                                  NA
## 5                         0                                  NA
## 6                         0                                  NA
##   STDs..Time.since.last.diagnosis Dx.Cancer Dx.CIN Dx.HPV    Dx Hinselmann
## 1                              NA     False  False  False False      False
## 2                              NA     False  False  False False      False
## 3                              NA     False  False  False False      False
## 4                              NA      True  False   True False      False
## 5                              NA     False  False  False False      False
## 6                              NA     False  False  False False      False
##   Schiller Citology Biopsy
## 1    False    False  False
## 2    False    False  False
## 3    False    False  False
## 4    False    False  False
## 5    False    False  False
## 6    False    False  False
```

Now that the data set has been successfully formatted and cleaned, we can save this file to our desktop.

```
write.table(risk_factors_cervical_cancer,file="C:\\Users\\dev46\\OneDrive\\Desktop\\Fall 2023\\Data Stewardship\\HW\\10.03.2023 Week 5\\risk_factors_cervical_cancer.data",sep=",",row.names=TRUE,col.names=TRUE)
```