Cervical Cancer Cleaning

Mike Kozlowski

2023-02-27

```
ccDataFile="F:\\Chrome Downloads\\risk_factors_cervical_cancer.csv"
```

Checking the types that the attributes were set to in order to verify that the types are properly set.

```
cervCanc=read.table(ccDataFile,header=TRUE,na.strings="?",sep=",",fill=TRUE)
head(cervCanc)
```

```
##   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(cervCanc)
```

```
## '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 ...
```

It appears that several data types that should have been set to bool have been set to num. The data set doesn't really make sense in some ways according to the attribute information listed online though, as fields like "Smokes (years)" is set to bool, when it's telling you how many years that patient smoked. So I have no idea why it says it should be a bool on the website. I'm going to use some common sense and ignore the attributes that seem like they should remain integers/num and will fix the ones that make sense to be boolean, such as "Smokes". Fixing those:

```
cervCanc$Smokes <- as.logical(cervCanc$Smokes)
cervCanc$STDs <- as.logical(cervCanc$STDs)
cervCanc$Hormonal.Contraceptives <- as.logical(cervCanc$Hormonal.Contraceptives)
cervCanc$IUD <- as.logical(cervCanc$IUD)
cervCanc$STDs.condylomatosis <- as.logical(cervCanc$STDs.condylomatosis)
cervCanc$STDs.cervical.condylomatosis <- as.logical(cervCanc$STDs.cervical.condylomatosis)
cervCanc$STDs.vaginal.condylomatosis <- as.logical(cervCanc$STDs.vaginal.condylomatosis)
cervCanc$STDs.vulvo.perineal.condylomatosis <- as.logical(cervCanc$STDs.vulvo.perineal.condylomatosis)
cervCanc$STDs.syphilis <- as.logical(cervCanc$STDs.syphilis)
cervCanc$STDs.pelvic.inflammatory.disease <- as.logical(cervCanc$STDs.pelvic.inflammatory.disease)
cervCanc$STDs.genital.herpes <- as.logical(cervCanc$STDs.genital.herpes)
cervCanc$STDs.molluscum.contagiosum <- as.logical(cervCanc$STDs.molluscum.contagiosum)
cervCanc$STDs.AIDS <- as.logical(cervCanc$STDs.AIDS)
cervCanc$STDs.HIV <- as.logical(cervCanc$STDs.HIV)
cervCanc$STDs.Hepatitis.B <- as.logical(cervCanc$STDs.Hepatitis.B)
cervCanc$STDs.HPV <- as.logical(cervCanc$STDs.HPV)
cervCanc$Dx.Cancer <- as.logical(cervCanc$Dx.Cancer)
cervCanc$Dx.CIN <- as.logical(cervCanc$Dx.CIN)
cervCanc$Dx.HPV <- as.logical(cervCanc$Dx.HPV)
cervCanc$Dx <- as.logical(cervCanc$Dx)
cervCanc$Hinselmann <- as.logical(cervCanc$Hinselmann)
cervCanc$Schiller <- as.logical(cervCanc$Schiller)
cervCanc$Citology <- as.logical(cervCanc$Citology)
cervCanc$Biopsy <- as.logical(cervCanc$Biopsy)
```

Verifying the types.

```
str(cervCanc)
```

```
## '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                            : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ 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           : logi  FALSE FALSE FALSE TRUE TRUE FALSE ...
##  $ Hormonal.Contraceptives..years.   : num  0 0 0 3 15 0 0 2 0 0 ...
##  $ IUD                               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ IUD..years.                       : num  0 0 0 0 0 0 7 7 0 NA ...
##  $ STDs                              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs..number.                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ STDs.condylomatosis               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.cervical.condylomatosis      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.vaginal.condylomatosis       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.vulvo.perineal.condylomatosis: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.syphilis                     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.pelvic.inflammatory.disease  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.genital.herpes               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.molluscum.contagiosum        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.AIDS                         : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.HIV                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.Hepatitis.B                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ STDs.HPV                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ 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                         : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Dx.CIN                            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Dx.HPV                            : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Dx                                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Hinselmann                        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Schiller                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Citology                          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Biopsy                            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
head(cervCanc)
```

```
##   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
```

Checking to see if these columns have entries besides NA in them, as they appear to only have NA in the head of the data set

```
sum(!is.na(cervCanc[,28]))
```

```
## [1] 71
```

```
sum(!is.na(cervCanc[,27]))
```

```
## [1] 71
```

Viewing the table in excel to verify that everything looks correct.

```
View(cervCanc)
```

Data types appear correct, if a value has an entry with a real number in it at some point, the column is correctly set to num, and the ones with only integers are set to integers. If it has NA, it should be NA.

```
write.table(cervCanc,file="C:\\Users\\Mike\\Documents\\DAT511\\HW5\\cervical_cancer_formatted.csv",sep
=",",row.names=TRUE,col.names=TRUE)
```