DATA607-Week1-Mushrooms

Michael Y.

September 1, 2019

Assignment 1 - Loading Data into a Data Frame - Mushroom Dataset

```
knitr::opts_chunk$set(echo = TRUE)
directory = "C:/Users/Michael/Dropbox/priv/CUNY/MSDS/201909-Fall/DATA607_Tati_Andy/20190901_Week01/"
knitr::opts_knit$set(root.dir = directory)

### Make the output wide enough
options(scipen = 999, digits=6, width=120)

### Load some libraries
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
library(kableExtra)
```

Mushrooms Dataset

A famous-if slightly moldy-dataset about mushrooms can be found in the UCI repository here: https://archive.ics.uci.edu/ml/datasets/Mushroom The fact that this is such a well-known dataset in the data science community makes it a good dataset to use for comparative benchmarking. For example,

if someone was working to build a better decision tree algorithm (or other predictive classifier) to analyze categorical data, this dataset could be useful. A typical problem (which is beyond the scope of this assignment!) is to answer the question,

"Which other attribute or attributes are the best predictors of whether a particular mushroom is poisonous or edible?"

Your task is to study the dataset and the associated description of the data (i.e. "data dictionary"). You may need to look around a bit, but it's there! You should take the data,

```
### Avoid setwd when knitting -- instead use above knitr::opts_knit$set(root.dir = directory)
#setwd("C:/Users/Michael/Dropbox/priv/CUNY/MSDS/201909-Fall/DATA607_Tati_Andy/20190901_Week01/")
download.file('https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data', 'mushroom-database.csv')
```

Create a data frame with a subset of the columns in the dataset.

(First, I'll create a dataframe with all 23 columns:)

```
mushroom_df <- read.csv('mushroom-database.csv', header=FALSE, stringsAsFactors=TRUE)
```

According to the documentation, the size of the data frame should be 8124x3. Checking:

```
dim(mushroom_df)
```

```
## [1] 8124   23
```

Determine which lines represent poisonous mushrooms by converting to a 0/1 variable, usable in logistic regression

```
poisonous <- mushroom_df$V1=="p"
poisonous=as.integer(poisonous)
head(mushroom_df$V1)
```

```
## [1] p e e p e e
## Levels: e p
```

```
head(poisonous)
```

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

Get metadata – info about the data

```
download.file('https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.names', 'mushroom-info.txt')
# read the lines into R
mushroom_meta=readLines(con="mushroom-info.txt")
```

We only care about metalines 106-140

```
mushroom_meta2 = mushroom_meta[106:140]
# trim the whitespace on the overflow lines
mushroom_meta2 = trimws(mushroom_meta2)
kable(mushroom_meta2, caption="Mushroom Metadata text") %>%
  kable_styling(c("striped", "bordered"))
```

Attributes with many values are split onto two lines

Lines starting an attribute start with a number, while rollover lines start with a letter. Let's create a function which will join the rollover lines onto the starting lines:

```
pastearray = function(chararray) {
# take an array of character strings
# some lines start with a number, while other lines start with a letter
# the lines which start with a letter are to be pasted at the end of the preceding line,
# and the array is to be shortened

# first, determine which lines start with a number, and which with a letter
  # which lines in chararray start with a number, vs a letter?
  firstc <- substr(chararray,start = 1,stop = 1)          #Select the first character from each line
  firstc_numeric <- suppressWarnings(as.numeric(firstc))  # cast each character as numeric; characters return NA
                                                          # I don't want to see the warning messages, so suppress them
  firstc_not_numeric = sapply(firstc_numeric,is.na)       # true for each continuation line
  firstc_is_numeric = !firstc_not_numeric
  firstc_is_numeric
  chararray[firstc_is_numeric]

  tempoutputarray = NULL
```

```
  tempoutputline = ""
  j = 0
  for (i in 1:length(chararray))
    {
    if (firstc_is_numeric[i]) {
      # we are starting a new line, so print out the prior line assemblage --
      if (i>1) {
      # unless we are at the very beginning, in which case there is nothing to print
        j=j+1
        tempoutputarray[j]=tempoutputline
        #####print(paste(j,": ",  tempoutputarray[j]))
        # reset to blank line
        tempoutputline=""
      }
    # set this line as the newline
    tempoutputline = chararray[i]
    }
    else if (firstc_not_numeric[i])  # we are on a continuation !!!!!
      {
      tempoutputline = paste0(tempoutputline, chararray[i])   # paste this line onto the previous line
      #####print(paste("*** pasting ", i, j, tempoutputline))
    }
  }
  # when we reach the end, we still have to print out the prior line
  j=j+1
  tempoutputarray[j]=tempoutputline
  #####print(paste(j,": ",  tempoutputarray[j]))
  #####print ( paste(i,",",chararray[i]))
  return(tempoutputarray)
}
```

Apply this function to the above mushroom_meta

```
mushroom_meta2 = pastearray(mushroom_meta2)
mushroom_meta2
```

```
##  [1] "7. Attribute Information: (classes: edible=e, poisonous=p)"
##  [2] "1. cap-shape:                bell=b,conical=c,convex=x,flat=f,knobbed=k,sunken=s"
##  [3] "2. cap-surface:              fibrous=f,grooves=g,scaly=y,smooth=s"
##  [4] "3. cap-color:                brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y"
```

```
##  [5] "4. bruises?:                bruises=t,no=f"
##  [6] "5. odor:                    almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s"
##  [7] "6. gill-attachment:         attached=a,descending=d,free=f,notched=n"
##  [8] "7. gill-spacing:            close=c,crowded=w,distant=d"
##  [9] "8. gill-size:               broad=b,narrow=n"
## [10] "9. gill-color:              black=k,brown=n,buff=b,chocolate=h,gray=g,green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y"
## [11] "10. stalk-shape:             enlarging=e,tapering=t"
## [12] "11. stalk-root:              bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?"
## [13] "12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s"
## [14] "13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s"
## [15] "14. stalk-color-above-ring:   brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y"
## [16] "15. stalk-color-below-ring:   brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y"
## [17] "16. veil-type:               partial=p,universal=u"
## [18] "17. veil-color:              brown=n,orange=o,white=w,yellow=y"
## [19] "18. ring-number:             none=n,one=o,two=t"
## [20] "19. ring-type:               cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z"
## [21] "20. spore-print-color:       black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y"
## [22] "21. population:              abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y"
## [23] "22. habitat:                 grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d"
```

Success, now there are only 23 rows, each representing one column.

Modify the first line to resemble the other lines. Since ascertaining whether poisonous vs. edible is the TARGET, I'll relabel this line as such.

```
############
############
firstline = mushroom_meta2[1]
######firstline
# change the front part of the line
firstline = sub(pattern="7. Attribute Information: (classes:", replacement="0. TARGET:                  ", x=firstline, fixed=TRUE)
######firstline
# change the rear part of the line - remove the space and the final right parens
firstline = sub(pattern=", poisonous=p)", replacement=",poisonous=p", x=firstline, fixed=TRUE)
######firstline
################################
mushroom_meta2[1]=firstline
mushroom_meta2
```

```
##  [1] "0. TARGET:                  edible=e,poisonous=p"
##  [2] "1. cap-shape:               bell=b,conical=c,convex=x,flat=f,knobbed=k,sunken=s"
##  [3] "2. cap-surface:             fibrous=f,grooves=g,scaly=y,smooth=s"
```

```
##  [4] "3. cap-color:                brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y"
##  [5] "4. bruises?:                 bruises=t,no=f"
##  [6] "5. odor:                     almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s"
##  [7] "6. gill-attachment:          attached=a,descending=d,free=f,notched=n"
##  [8] "7. gill-spacing:             close=c,crowded=w,distant=d"
##  [9] "8. gill-size:                broad=b,narrow=n"
## [10] "9. gill-color:               black=k,brown=n,buff=b,chocolate=h,gray=g,green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y"
## [11] "10. stalk-shape:              enlarging=e,tapering=t"
## [12] "11. stalk-root:               bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?"
## [13] "12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s"
## [14] "13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s"
## [15] "14. stalk-color-above-ring:   brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y"
## [16] "15. stalk-color-below-ring:   brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y"
## [17] "16. veil-type:                partial=p,universal=u"
## [18] "17. veil-color:               brown=n,orange=o,white=w,yellow=y"
## [19] "18. ring-number:              none=n,one=o,two=t"
## [20] "19. ring-type:                cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z"
## [21] "20. spore-print-color:        black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y"
## [22] "21. population:               abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y"
## [23] "22. habitat:                  grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d"
```

```
###################################
```

Extract the list of attribute names from the above

```
names1=gsub(pattern=":.*$", replacement = "", x=mushroom_meta2)
#names1
names2=gsub(pattern="^[0-9]*. ", replacement = "", x=names1, perl=FALSE)
#names2
# replace hyphens with underscores or it will cause problems later
names3=gsub(pattern="-", replacement="_", x=names2)
kable(names3, caption="Mushroom Attribute Names") %>%
  kable_styling(c("striped", "bordered"))
```

extract the list of factors (their descriptive names, and their single character abbreviations) from each line

```
factors1 = gsub(pattern="^.*:  *",replacement="",x=mushroom_meta2)
factors1
```

```
##  [1] "edible=e,poisonous=p"
##  [2] "bell=b,conical=c,convex=x,flat=f,knobbed=k,sunken=s"
```

```
##  [3] "fibrous=f,grooves=g,scaly=y,smooth=s"
##  [4] "brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y"
##  [5] "bruises=t,no=f"
##  [6] "almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s"
##  [7] "attached=a,descending=d,free=f,notched=n"
##  [8] "close=c,crowded=w,distant=d"
##  [9] "broad=b,narrow=n"
## [10] "black=k,brown=n,buff=b,chocolate=h,gray=g,green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y"
## [11] "enlarging=e,tapering=t"
## [12] "bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?"
## [13] "fibrous=f,scaly=y,silky=k,smooth=s"
## [14] "fibrous=f,scaly=y,silky=k,smooth=s"
## [15] "brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y"
## [16] "brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y"
## [17] "partial=p,universal=u"
## [18] "brown=n,orange=o,white=w,yellow=y"
## [19] "none=n,one=o,two=t"
## [20] "cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z"
## [21] "black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y"
## [22] "abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y"
## [23] "grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d"
```

Rename the factors with the descriptive names

(I had to set this up manually because I couldn't get the proper processing of the quotation marks in R)

```
levels(mushroom_df$V1)
```

```
## [1] "e" "p"
```

```
levels(mushroom_df$V1) <-  list(edible="e",poisonous="p")
levels(mushroom_df$V1)
```

```
## [1] "edible"    "poisonous"
```

```
levels(mushroom_df$V2) <-  list(bell="b",conical="c",convex="x",flat="f",knobbed="k",sunken="s")
levels(mushroom_df$V3) <-  list(fibrous="f",grooves="g",scaly="y",smooth="s")
levels(mushroom_df$V4) <-  list(brown="n",buff="b",cinnamon="c",gray="g",green="r",pink="p",purple="u",red="e",white="w",yellow="y")
levels(mushroom_df$V5) <-  list(bruises="t",no="f")
```

```
levels(mushroom_df$V6)
```

```
## [1] "a" "c" "f" "l" "m" "n" "p" "s" "y"
```

```
levels(mushroom_df$V6) <-  list(almond="a",anise="l",creosote="c",fishy="y",foul="f",musty="m",none="n",pungent="p",spicy="s")
levels(mushroom_df$V6)
```

```
## [1] "almond"   "anise"    "creosote" "fishy"    "foul"     "musty"    "none"     "pungent"  "spicy"
```

```
levels(mushroom_df$V7) <-  list(attached="a",descending="d",free="f",notched="n")
levels(mushroom_df$V8) <-  list(close="c",crowded="w",distant="d")
levels(mushroom_df$V9) <-  list(broad="b",narrow="n")
levels(mushroom_df$V10) <-  list(black="k",brown="n",buff="b",chocolate="h",gray="g",green="r",orange="o",pink="p",purple="u",red="e",whit
levels(mushroom_df$V11) <-  list(enlarging="e",tapering="t")
levels(mushroom_df$V12) <-  list(bulbous="b",club="c",cup="u",equal="e",rhizomorphs="z",rooted="r",missing="?")
levels(mushroom_df$V13) <-  list(fibrous="f",scaly="y",silky="k",smooth="s")
levels(mushroom_df$V14) <-  list(fibrous="f",scaly="y",silky="k",smooth="s")
levels(mushroom_df$V15) <-  list(brown="n",buff="b",cinnamon="c",gray="g",orange="o",pink="p",red="e",white="w",yellow="y")
levels(mushroom_df$V16) <-  list(brown="n",buff="b",cinnamon="c",gray="g",orange="o",pink="p",red="e",white="w",yellow="y")
levels(mushroom_df$V17) <-  list(partial="p",universal="u")
levels(mushroom_df$V18) <-  list(brown="n",orange="o",white="w",yellow="y")
levels(mushroom_df$V19) <-  list(none="n",one="o",two="t")
levels(mushroom_df$V20) <-  list(cobwebby="c",evanescent="e",flaring="f",large="l",none="n",pendant="p",sheathing="s",zone="z")
levels(mushroom_df$V21) <-  list(black="k",brown="n",buff="b",chocolate="h",green="r",orange="o",purple="u",white="w",yellow="y")
levels(mushroom_df$V22) <-  list(abundant="a",clustered="c",numerous="n",scattered="s",several="v",solitary="y")
levels(mushroom_df$V23) <-  list(grasses="g",leaves="l",meadows="m",paths="p",urban="u",waste="w",woods="d")
```

Display mushroom df summary

```
summary(mushroom_df)
```

```
##          V1             V2             V3             V4             V5             V6                V7
##  edible   :4208   bell    : 452   fibrous:2320   brown  :2284   bruises:3376   none   :3528   attached  : 210
##  poisonous:3916   conical:   4   grooves:   4   gray   :1840   no     :4748   foul   :2160   descending:   0
##                   convex :3656   scaly  :3244   red    :1500                  fishy  : 576   free      :7914
##                   flat   :3152   smooth :2556   yellow :1072                  spicy  : 576   notched   :   0
##                   knobbed: 828                  white  :1040                  almond : 400
##                   sunken :  32                  buff   : 168                  anise  : 400
##                                                 (Other): 220                  (Other): 484
```

```
##        V8            V9             V10             V11               V12           V13           V14
##  close  :6812   broad :5612   buff     :1728   enlarging:3516   bulbous    :3776   fibrous: 552   fibrous: 600
##  crowded:1312   narrow:2512   pink     :1492   tapering :4608   club       : 556   scaly  :  24   scaly  : 284
##  distant:   0                 white    :1202                    cup        :   0   silky  :2372   silky  :2304
##                               brown    :1048                    equal      :1120   smooth :5176   smooth :4936
##                               gray     : 752                    rhizomorphs:   0
##                               chocolate: 732                    rooted     : 192
##                               (Other)  :1170                    missing    :2480
##       V15            V16             V17           V18         V19              V20             V21
##  white  :4464   white  :4384   partial  :8124   brown :  96   none:  36   pendant   :3968   white    :2388
##  pink   :1872   pink   :1872   universal:   0   orange:  96   one :7488   evanescent:2776   brown    :1968
##  gray   : 576   gray   : 576                    white :7924   two : 600   large     :1296   black    :1872
##  brown  : 448   brown  : 512                    yellow:   8               flaring   :  48   chocolate:1632
##  buff   : 432   buff   : 432                                             none      :  36   green    :  72
##  orange : 192   orange : 192                                             cobwebby  :   0   buff     :  48
##  (Other): 140   (Other): 156                                             (Other)   :   0   (Other)  : 144
##         V22            V23
##  abundant : 384   grasses:2148
##  clustered: 340   leaves : 832
##  numerous : 400   meadows: 292
##  scattered:1248   paths  :1144
##  several  :4040   urban  : 368
##  solitary :1712   waste  : 192
##                   woods  :3148
```

Now, replace the names of the columns in the data set with descriptive attribute names (with underscores replacing hyphens:

```
names(mushroom_df)=names3
head(mushroom_df)
```

```
##      TARGET cap_shape cap_surface cap_color bruises?    odor gill_attachment gill_spacing gill_size gill_color
## 1 poisonous    convex      smooth     brown  bruises pungent            free        close    narrow      black
## 2    edible    convex      smooth    yellow  bruises  almond            free        close     broad      black
## 3    edible      bell      smooth     white  bruises   anise            free        close     broad      brown
## 4 poisonous    convex       scaly     white  bruises pungent            free        close    narrow      brown
## 5    edible    convex      smooth      gray       no    none            free      crowded     broad      black
## 6    edible    convex       scaly    yellow  bruises  almond            free        close     broad      brown
##   stalk_shape stalk_root stalk_surface_above_ring stalk_surface_below_ring stalk_color_above_ring
```

```
## 1   enlarging      equal                  smooth                  smooth                  white
## 2   enlarging       club                  smooth                  smooth                  white
## 3   enlarging       club                  smooth                  smooth                  white
## 4   enlarging      equal                  smooth                  smooth                  white
## 5    tapering      equal                  smooth                  smooth                  white
## 6   enlarging       club                  smooth                  smooth                  white
##   stalk_color_below_ring veil_type veil_color ring_number  ring_type spore_print_color population habitat
## 1                  white   partial      white         one    pendant             black  scattered   urban
## 2                  white   partial      white         one    pendant             brown   numerous grasses
## 3                  white   partial      white         one    pendant             brown   numerous meadows
## 4                  white   partial      white         one    pendant             black  scattered   urban
## 5                  white   partial      white         one evanescent             brown   abundant grasses
## 6                  white   partial      white         one    pendant             black   numerous grasses
```

```
summary(mushroom_df)
```

```
##        TARGET       cap_shape     cap_surface     cap_color       bruises?         odor        gill_attachment
##  edible   :4208   bell   : 452   fibrous:2320   brown  :2284   bruises:3376   none   :3528   attached  : 210
##  poisonous:3916   conical:   4   grooves:   4   gray   :1840   no     :4748   foul   :2160   descending:   0
##                   convex :3656   scaly  :3244   red    :1500                  fishy  : 576   free      :7914
##                   flat   :3152   smooth :2556   yellow :1072                  spicy  : 576   notched   :   0
##                   knobbed: 828                  white  :1040                  almond : 400
##                   sunken :  32                  buff   : 168                  anise  : 400
##                                                 (Other): 220                  (Other): 484
##   gill_spacing   gill_size        gill_color      stalk_shape         stalk_root   stalk_surface_above_ring
##  close  :6812   broad :5612   buff     :1728   enlarging:3516   bulbous    :3776   fibrous: 552
##  crowded:1312   narrow:2512   pink     :1492   tapering :4608   club       : 556   scaly  :  24
##  distant:   0                 white    :1202                    cup        :   0   silky  :2372
##                               brown    :1048                    equal      :1120   smooth :5176
##                               gray     : 752                    rhizomorphs:   0
##                               chocolate: 732                    rooted     : 192
##                               (Other)  :1170                    missing    :2480
##  stalk_surface_below_ring stalk_color_above_ring stalk_color_below_ring     veil_type     veil_color   ring_number
##  fibrous: 600             white  :4464           white  :4384           partial  :8124   brown :  96   none:  36
##  scaly  : 284             pink   :1872           pink   :1872           universal:   0   orange:  96   one :7488
##  silky  :2304             gray   : 576           gray   : 576                            white :7924   two : 600
##  smooth :4936             brown  : 448           brown  : 512                            yellow:   8
##                           buff   : 432           buff   : 432
##                           orange : 192           orange : 192
##                           (Other): 140           (Other): 156
```

```
##       ring_type    spore_print_color     population      habitat
##  pendant   :3968   white    :2388    abundant : 384   grasses:2148
##  evanescent:2776   brown    :1968    clustered: 340   leaves : 832
##  large     :1296   black    :1872    numerous : 400   meadows: 292
##  flaring   :  48   chocolate:1632    scattered:1248   paths  :1144
##  none      :  36   green    :  72    several  :4040   urban  : 368
##  cobwebby  :   0   buff     :  48    solitary :1712   waste  : 192
##  (Other)   :   0   (Other)  : 144                     woods  :3148
```

According to the documentation, a small number of columns provide an excellent prediction of which mushrooms are poisonous:

These columns are odor, spore-print-color, stalk-surface-below-ring, and stalk-color-above-ring

Rule #1: odor (V6) is not Almond ("a"), Anise ("l"), or None ("n")

```
# Rule1: odor (V6) is NOT (Almond , Anise , or None)
rule1 = (!(mushroom_df$odor == "almond" | mushroom_df$odor == "anise" | mushroom_df$odor == "none"))
sum(as.integer(rule1))
```

```
## [1] 3796
```

Rule #2: spore-print-color (V21) is Green ("r")

```
# Rule2: spore-print-color is Green
rule2 = (mushroom_df$spore_print_color== "green")
sum(as.integer(rule2))
```

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

Rule #3: odor (V6) is None ("n") AND stalk-surface-below-ring (V14) is Scaly ("y") AND stalk-color-above-ring (V15) is NOT brown (n)

```
# Rule3: odor is None AND
#       stalk-surface-below-ring is Scaly AND
#       stalk-color-above-ring is NOT brown
rule3 = ((mushroom_df$odor == "none") & (mushroom_df$stalk_surface_below_ring == "scaly" )&( mushroom_df$stalk_color_above_ring != "brown"
sum(as.integer(rule3))
```

```
## [1] 40
```

run logistic regression using just rule 1

```
mmodel1 <- glm(poisonous ~ rule1, data = mushroom_df, family="binomial")
summary(mmodel1)
```

```
##
## Call:
## glm(formula = poisonous ~ rule1, family = "binomial", data = mushroom_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.237  -0.237  -0.237   0.000   2.678
##
## Coefficients:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept)  -3.5573     0.0926  -38.42 <0.0000000000000002 ***
## rule1TRUE    25.1233   474.4626    0.05                0.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11251.8  on 8123  degrees of freedom
## Residual deviance:  1097.1  on 8122  degrees of freedom
## AIC: 1101
##
## Number of Fisher Scoring iterations: 20
```

```
model1predictor <- (as.integer(mmodel1$residuals)==1)
model1errors <- sum(as.integer(model1predictor != poisonous))
model1errors
```

```
## [1] 120
```

run logistic regression using rules 1 and 2

```
mmodel2 <- glm(poisonous ~ rule1 + rule2, data = mushroom_df, family="binomial")
summary(mmodel2)
```

```
##
## Call:
## glm(formula = poisonous ~ rule1 + rule2, family = "binomial",
##     data = mushroom_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.151  -0.151  -0.151   0.000   2.995
##
## Coefficients:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept)   -4.474      0.145  -30.82 <0.0000000000000002 ***
## rule1TRUE     27.040    782.257    0.03                0.97
## rule2TRUE     27.040   5679.970    0.00                1.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11251.76  on 8123  degrees of freedom
## Residual deviance:   526.01  on 8121  degrees of freedom
## AIC: 532
##
## Number of Fisher Scoring iterations: 21
```

```
model2predictor <- (as.integer(mmodel2$residuals)==1)
model2errors <- sum(as.integer(model2predictor != poisonous))
model2errors
```

```
## [1] 120
```

run logistic regression using rules 1, 2 and 3

```
mmodel3 <- glm(poisonous ~ rule1 + rule2 + rule3, data = mushroom_df, family="binomial")
summary(mmodel3)
```

```
##
## Call:
## glm(formula = poisonous ~ rule1 + rule2 + rule3, family = "binomial",
```

```
##     data = mushroom_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.062  -0.062  -0.062   0.000   3.540
##
## Coefficients:
##              Estimate Std. Error z value            Pr(>|z|)
## (Intercept)    -6.265      0.354  -17.70 <0.0000000000000002 ***
## rule1TRUE      29.831   1289.723    0.02                0.98
## rule2TRUE      29.831   9364.689    0.00                1.00
## rule3TRUE      29.831  12564.045    0.00                1.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11251.76  on 8123  degrees of freedom
## Residual deviance:   116.26  on 8120  degrees of freedom
## AIC: 124.3
##
## Number of Fisher Scoring iterations: 22
```

```
model3predictor <- (as.integer(mmodel3$residuals)==1)
model3errors <- sum(as.integer(model3predictor != poisonous))
model3errors
```

```
## [1] 8
```

Combining the first three rules gives a result which misses only 8 poisonous mushrooms.

Subsetting the data

You should include the column that indicates edible or poisonous and three or four other columns.

You should also add meaningful column names and replace the abbreviations used in the data-for example, in the appropriate column, “e” might become “edible.”

```
mushroom_subsetdf = subset(mushroom_df, select=c(TARGET, odor, spore_print_color, stalk_surface_below_ring, stalk_color_above_ring))
head(mushroom_subsetdf)
```

```
##      TARGET    odor spore_print_color stalk_surface_below_ring stalk_color_above_ring
## 1 poisonous pungent             black                   smooth                  white
## 2    edible  almond             brown                   smooth                  white
## 3    edible   anise             brown                   smooth                  white
## 4 poisonous pungent             black                   smooth                  white
## 5    edible    none             brown                   smooth                  white
## 6    edible  almond             black                   smooth                  white
```

Summary of subsets

```
summary(mushroom_subsetdf)
```

```
##        TARGET          odor      spore_print_color stalk_surface_below_ring stalk_color_above_ring
##  edible   :4208   none   :3528   white    :2388    fibrous: 600             white  :4464
##  poisonous:3916   foul   :2160   brown    :1968    scaly  : 284             pink   :1872
##                   fishy  : 576   black    :1872    silky  :2304             gray   : 576
##                   spicy  : 576   chocolate:1632    smooth :4936             brown  : 448
##                   almond : 400   green    :  72                             buff   : 432
##                   anise  : 400   buff     :  48                             orange : 192
##                   (Other): 484   (Other)  : 144                             (Other): 140
```

Your deliverable is the R code to perform these transformation tasks.

Table 1: Mushroom Metadata text

x
7. Attribute Information: (classes: edible=e, poisonous=p)
1. cap-shape: bell=b,conical=c,convex=x,flat=f,
knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,
pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,
musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g,
green=r,orange=o,pink=p,purple=u,red=e,
white=w,yellow=y
10. stalk-shape: enlarging=e,tapering=t
11. stalk-root: bulbous=b,club=c,cup=u,equal=e,
rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,
pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,
pink=p,red=e,white=w,yellow=y
16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,
none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,
orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n,
scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p,
urban=u,waste=w,woods=d

Table 2: Mushroom Attribute Names

x
TARGET
cap_shape
cap_surface
cap_color
bruises?
odor
gill_attachment
gill_spacing
gill_size
gill_color
stalk_shape
stalk_root
stalk_surface_above_ring
stalk_surface_below_ring
stalk_color_above_ring
stalk_color_below_ring
veil_type
veil_color
ring_number
ring_type
spore_print_color
population
habitat