Homework Week 1

CUNY MSDA DATA 607

Duubar Villalobos Jimenez mydvtech@gmail.com

February 5, 2017

R Assignment - Basic Data Loading and Transformations

Loading Data into a Data Frame

Very often, we're tasked with taking data in one form and transforming it for easier downstream analysis. We will spend several weeks in this course on tidying and transformation operations. Some of this work could be done in SQL or R (or Python or. . .). Here, you are asked to use R -you may use base functions or packages as you like.

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, and create a data frame with a subset of the columns in the dataset. 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." Your deliverable is the R code to perform these transformation tasks.

If you are working in a group, you also have the option of replacing the mushroom dataset in the assignment with a different data set that your group members might find more interesting.

Please place your solution in to a single R Markdown (.Rmd) file and publish your solution out to rpubs.com. You should post the .Rmd file in your GitHub repository, and provide the appropriate URLs to your GitHub repository and your rpubs.com file in your assignment link. You should also have the original data file accessible through your code-for example, stored in a GitHub repository and referenced in your code. We'll look together at some of the most interesting student solutions in next week's meetup.

Solution

Read data from url

```
url <- 'https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data'
mushrooms <- read.table(url, sep=",", header=FALSE, stringsAsFactors = FALSE)
head(mushrooms)
```

```
##   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20
## 1  p  x  s  n  t  p  f  c  n   k   e   e   s   s   w   w   p   w   o   p
## 2  e  x  s  y  t  a  f  c  b   k   e   c   s   s   w   w   p   w   o   p
## 3  e  b  s  w  t  l  f  c  b   n   e   c   s   s   w   w   p   w   o   p
## 4  p  x  y  w  t  p  f  c  n   n   e   e   s   s   w   w   p   w   o   p
## 5  e  x  s  g  f  n  f  w  b   k   t   e   s   s   w   w   p   w   o   e
```

```
## 6  e  x  y  y  t  a  f  c  b  n  e  c  s  s  w  w  p  w  o  p
##    V21 V22 V23
## 1   k   s   u
## 2   n   n   g
## 3   n   n   m
## 4   k   s   u
## 5   n   a   g
## 6   k   n   g
```

Provide Column Names

I have saved the dictionary from https://archive.ics.uci.edu/ml/machine-learning-databases/bridges/bridges.names in a Dictionary.txt file

The file I have included into GitHub with the following link: https://raw.githubusercontent.com/dvillalobos/MSDA/master/607/Homework/Villalobos-Homework1-dictionary.txt

```
file <- 'https://raw.githubusercontent.com/dvillalobos/MSDA/master/607/Homework/Villalobos-Homework1-di
mushroomsdict <- read.table(file, sep="|", header=TRUE, stringsAsFactors = FALSE)
mushroomsdict
```

```
##    Index               Attribute
## 1      0                   class
## 2      1               cap-shape
## 3      2             cap-surface
## 4      3               cap-color
## 5      4                bruises?
## 6      5                    odor
## 7      6          gill-attachment
## 8      7             gill-spacing
## 9      8                gill-size
## 10     9               gill-color
## 11    10              stalk-shape
## 12    11               stalk-root
## 13    12 stalk-surface-above-ring
## 14    13 stalk-surface-below-ring
## 15    14   stalk-color-above-ring
## 16    15   stalk-color-below-ring
## 17    16                veil-type
## 18    17               veil-color
## 19    18             ring-number
## 20    19               ring-type
## 21    20        spore-print-color
## 22    21              population
## 23    22                 habitat
##                                                                                  Information
## 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 Column Names for the Bridge data frame.

```
colnames(mushrooms) <- mushroomsdict$Attribute
head(mushrooms)
```

```
##   class cap-shape cap-surface cap-color bruises? odor gill-attachment
## 1     p         x           s         n        t    p               f
## 2     e         x           s         y        t    a               f
## 3     e         b           s         w        t    l               f
## 4     p         x           y         w        t    p               f
## 5     e         x           s         g        f    n               f
## 6     e         x           y         y        t    a               f
##   gill-spacing gill-size gill-color stalk-shape stalk-root
## 1            c         n          k           e          e
## 2            c         b          k           e          c
## 3            c         b          n           e          c
## 4            c         n          n           e          e
## 5            w         b          k           t          e
## 6            c         b          n           e          c
##   stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring
## 1                        s                        s                       w
## 2                        s                        s                       w
## 3                        s                        s                       w
## 4                        s                        s                       w
## 5                        s                        s                       w
## 6                        s                        s                       w
##   stalk-color-below-ring veil-type veil-color ring-number ring-type
## 1                      w         p          w           o         p
## 2                      w         p          w           o         p
## 3                      w         p          w           o         p
## 4                      w         p          w           o         p
## 5                      w         p          w           o         e
## 6                      w         p          w           o         p
##   spore-print-color population habitat
## 1                 k          s       u
## 2                 n          n       g
## 3                 n          n       m
## 4                 k          s       u
```

```
## 5                 n          a       g
## 6                 k          n       g
```

Data transformation for the class column

```
# Data transformation for the class column
mushroomsdict$Information[1]
```

```
## [1] "edible=e,poisonous=p"
```

```
mushrooms$class[mushrooms$class == 'e'] <- 'edible'
mushrooms$class[mushrooms$class == 'p'] <- 'poisonous'
head(mushrooms)
```

```
##       class cap-shape cap-surface cap-color bruises? odor gill-attachment
## 1 poisonous         x           s         n        t    p               f
## 2    edible         x           s         y        t    a               f
## 3    edible         b           s         w        t    l               f
## 4 poisonous         x           y         w        t    p               f
## 5    edible         x           s         g        f    n               f
## 6    edible         x           y         y        t    a               f
##   gill-spacing gill-size gill-color stalk-shape stalk-root
## 1            c         n          k           e          e
## 2            c         b          k           e          c
## 3            c         b          n           e          c
## 4            c         n          n           e          e
## 5            w         b          k           t          e
## 6            c         b          n           e          c
##   stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring
## 1                        s                        s                       w
## 2                        s                        s                       w
## 3                        s                        s                       w
## 4                        s                        s                       w
## 5                        s                        s                       w
## 6                        s                        s                       w
##   stalk-color-below-ring veil-type veil-color ring-number ring-type
## 1                      w         p          w           o         p
## 2                      w         p          w           o         p
## 3                      w         p          w           o         p
## 4                      w         p          w           o         p
## 5                      w         p          w           o         e
## 6                      w         p          w           o         p
##   spore-print-color population habitat
## 1                 k          s       u
## 2                 n          n       g
## 3                 n          n       m
## 4                 k          s       u
## 5                 n          a       g
## 6                 k          n       g
```

```
table(mushrooms$class)
```

```
##
##    edible poisonous
##      4208      3916
```

Data transformation for the cap-surface column

```
# Data transformation for the cap-surface column

mushroomsdict$Information[3]
```

```
## [1] "fibrous=f,grooves=g,scaly=y,smooth=s"
```

```
mushrooms$`cap-surface`[mushrooms$`cap-surface` == 'f'] <- 'fibrous'
mushrooms$`cap-surface`[mushrooms$`cap-surface` == 'g'] <- 'grooves'
mushrooms$`cap-surface`[mushrooms$`cap-surface` == 'y'] <- 'scaly'
mushrooms$`cap-surface`[mushrooms$`cap-surface` == 's'] <- 'smooth'
head(mushrooms)
```

```
##       class cap-shape cap-surface cap-color bruises? odor gill-attachment
## 1 poisonous         x      smooth         n        t    p               f
## 2    edible         x      smooth         y        t    a               f
## 3    edible         b      smooth         w        t    l               f
## 4 poisonous         x       scaly         w        t    p               f
## 5    edible         x      smooth         g        f    n               f
## 6    edible         x       scaly         y        t    a               f
##   gill-spacing gill-size gill-color stalk-shape stalk-root
## 1            c         n          k           e          e
## 2            c         b          k           e          c
## 3            c         b          n           e          c
## 4            c         n          n           e          e
## 5            w         b          k           t          e
## 6            c         b          n           e          c
##   stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring
## 1                        s                        s                       w
## 2                        s                        s                       w
## 3                        s                        s                       w
## 4                        s                        s                       w
## 5                        s                        s                       w
## 6                        s                        s                       w
##   stalk-color-below-ring veil-type veil-color ring-number ring-type
## 1                      w         p          w           o         p
## 2                      w         p          w           o         p
## 3                      w         p          w           o         p
## 4                      w         p          w           o         p
## 5                      w         p          w           o         e
## 6                      w         p          w           o         p
##   spore-print-color population habitat
## 1                 k          s       u
## 2                 n          n       g
## 3                 n          n       m
## 4                 k          s       u
## 5                 n          a       g
## 6                 k          n       g
```

```
table(mushrooms$`cap-surface`)
```

```
##
## fibrous grooves   scaly  smooth
##    2320       4    3244    2556
```

Data transformation for the bruises? column

```
# Data transformation for the bruises? column

mushroomsdict$Information[5]
```

```
## [1] "bruises=t,no=f"
```

```
mushrooms$`bruises?`[mushrooms$`bruises?` == 't'] <- 'bruises'
mushrooms$`bruises?`[mushrooms$`bruises?` == 'f'] <- 'no'
head(mushrooms)
```

```
##       class cap-shape cap-surface cap-color bruises? odor gill-attachment
## 1 poisonous         x      smooth         n  bruises    p               f
## 2    edible         x      smooth         y  bruises    a               f
## 3    edible         b      smooth         w  bruises    l               f
## 4 poisonous         x       scaly         w  bruises    p               f
## 5    edible         x      smooth         g       no    n               f
## 6    edible         x       scaly         y  bruises    a               f
##   gill-spacing gill-size gill-color stalk-shape stalk-root
## 1            c         n          k           e          e
## 2            c         b          k           e          c
## 3            c         b          n           e          c
## 4            c         n          n           e          e
## 5            w         b          k           t          e
## 6            c         b          n           e          c
##   stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring
## 1                        s                        s                       w
## 2                        s                        s                       w
## 3                        s                        s                       w
## 4                        s                        s                       w
## 5                        s                        s                       w
## 6                        s                        s                       w
##   stalk-color-below-ring veil-type veil-color ring-number ring-type
## 1                      w         p          w           o         p
## 2                      w         p          w           o         p
## 3                      w         p          w           o         p
## 4                      w         p          w           o         p
## 5                      w         p          w           o         e
## 6                      w         p          w           o         p
##   spore-print-color population habitat
## 1                 k          s       u
## 2                 n          n       g
## 3                 n          n       m
## 4                 k          s       u
## 5                 n          a       g
## 6                 k          n       g
```

```
table(mushrooms$`bruises?`)
```

```
##
## bruises      no
##    3376    4748
```

Data transformation for the gill-size column

```
# Data transformation for the gill-size column

mushroomsdict$Information[9]
```

```
## [1] "broad=b,narrow=n"
```

```
mushrooms$`gill-size`[mushrooms$`gill-size` == 'b'] <- 'broad'
mushrooms$`gill-size`[mushrooms$`gill-size` == 'n'] <- 'narrow'
head(mushrooms)
```

```
##       class cap-shape cap-surface cap-color bruises? odor gill-attachment
## 1 poisonous         x      smooth         n  bruises    p               f
## 2    edible         x      smooth         y  bruises    a               f
## 3    edible         b      smooth         w  bruises    l               f
## 4 poisonous         x       scaly         w  bruises    p               f
## 5    edible         x      smooth         g       no    n               f
## 6    edible         x       scaly         y  bruises    a               f
##   gill-spacing gill-size gill-color stalk-shape stalk-root
## 1            c    narrow          k           e          e
## 2            c     broad          k           e          c
## 3            c     broad          n           e          c
## 4            c    narrow          n           e          e
## 5            w     broad          k           t          e
## 6            c     broad          n           e          c
##   stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring
## 1                        s                        s                       w
## 2                        s                        s                       w
## 3                        s                        s                       w
## 4                        s                        s                       w
## 5                        s                        s                       w
## 6                        s                        s                       w
##   stalk-color-below-ring veil-type veil-color ring-number ring-type
## 1                      w         p          w           o         p
## 2                      w         p          w           o         p
## 3                      w         p          w           o         p
## 4                      w         p          w           o         p
## 5                      w         p          w           o         e
## 6                      w         p          w           o         p
##   spore-print-color population habitat
## 1                 k          s       u
## 2                 n          n       g
## 3                 n          n       m
## 4                 k          s       u
## 5                 n          a       g
## 6                 k          n       g
```

```
table(mushrooms$`gill-size`)
```

```
##
##  broad narrow
##   5612   2512
```

Presenting the four columns only

```
mush <- mushrooms[, c(1, 3, 5, 9)]
tail(mush)
```

```
##          class cap-surface bruises? gill-size
## 8119 poisonous       scaly       no    narrow
## 8120    edible      smooth       no     broad
## 8121    edible      smooth       no     broad
## 8122    edible      smooth       no     broad
## 8123 poisonous       scaly       no    narrow
## 8124    edible      smooth       no     broad
```