DATA607 WEEK One Assignment

John K. Hancock

September 1, 2018

DATA606 WEEK ONE ASSIGMENT

```
options(warn=-1)
setwd('C:/Users/jkhan/Documents/CUNY/Fall 2018/DATA606/Week One due 09-02-2018')
library(XML)
library(httr)
library(curl)
```

```
##
## Attaching package: 'curl'
```

```
## The following object is masked from 'package:httr':
##
##     handle_reset
```

```
library(stringr)
```

Use Curl to download the data to a csv file. Read the csv file into a dataframe.

```
curl_download('https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data', 'mushrooms.csv')
mushrooms <- read.csv('mushrooms.csv',header = FALSE)
unlink('mushrooms.csv')
```

```
head(mushrooms,5)
```

```
##   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
##   V21 V22 V23
## 1   k   s   u
## 2   n   n   g
## 3   n   n   m
## 4   k   s   u
## 5   n   a   g
```

Download the dictionary defining the columns.

```
dictionary <- curl_download('https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.names', 'dictionary.txt')
dictionary_text <- readLines('dictionary.txt')
```

Use grep to find the Attribute section of the page

```
grep('Attribute', dictionary_text)
```

```
## [1] 104 106 142
```

```
vec <- c(dictionary_text[107:139])
vec
```

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

Use a regular expression to find everything between the period and the colon. Remove extraneous values.

```
vec2 = gsub(".*\\.(.*)\\:.*", "\\1",vec)
vec2 <-vec2[-c(2,5,8,13,14,17,21,23,28,30,32)]
```

Trim the whitespaces

```
vec3 <- trimws(vec2)
vec4 <- c("edible_poisonous")
column_names <- c(vec4,vec3)
length(column_names)
```

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

Print out the first 7 rows of the data frame and take a look at its structure and names

```
colnames(mushrooms) <- c(column_names)

head(mushrooms,7)
```

```
##   edible_poisonous cap-shape cap-surface cap-color bruises? odor
## 1                p         x           s         n        t    p
## 2                e         x           s         y        t    a
## 3                e         b           s         w        t    l
## 4                p         x           y         w        t    p
## 5                e         x           s         g        f    n
## 6                e         x           y         y        t    a
## 7                e         b           s         w        t    a
##   gill-attachment gill-spacing gill-size gill-color stalk-shape stalk-root
## 1               f            c         n          k           e          e
## 2               f            c         b          k           e          c
## 3               f            c         b          n           e          c
## 4               f            c         n          n           e          e
## 5               f            w         b          k           t          e
## 6               f            c         b          n           e          c
## 7               f            c         b          g           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
## 7                        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
## 7                      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
## 7                 k          n       m
```

```
str(mushrooms)
```

```
## 'data.frame':    8124 obs. of  23 variables:
##  $ edible_poisonous        : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
##  $ cap-shape               : Factor w/ 6 levels "b","c","f","k",..: 6 6 1 6 6 6 1 1 6 1 ...
##  $ cap-surface             : Factor w/ 4 levels "f","g","s","y": 3 3 3 4 3 4 3 4 4 3 ...
##  $ cap-color               : Factor w/ 10 levels "b","c","e","g",..: 5 10 9 9 4 10 9 9 9 10
...
##  $ bruises?                : Factor w/ 2 levels "f","t": 2 2 2 2 1 2 2 2 2 2 ...
##  $ odor                    : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...
##  $ gill-attachment         : Factor w/ 2 levels "a","f": 2 2 2 2 2 2 2 2 2 2 ...
##  $ gill-spacing            : Factor w/ 2 levels "c","w": 1 1 1 1 2 1 1 1 1 1 ...
##  $ gill-size               : Factor w/ 2 levels "b","n": 2 1 1 2 1 1 1 1 2 1 ...
##  $ gill-color              : Factor w/ 12 levels "b","e","g","h",..: 5 5 6 6 5 6 3 6 8 3 ...
##  $ stalk-shape             : Factor w/ 2 levels "e","t": 1 1 1 1 2 1 1 1 1 1 ...
##  $ stalk-root              : Factor w/ 5 levels "?","b","c","e",..: 4 3 3 4 4 3 3 3 4 3 ...
##  $ stalk-surface-above-ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ stalk-surface-below-ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ stalk-color-above-ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 ...
##  $ stalk-color-below-ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 ...
##  $ veil-type               : Factor w/ 1 level "p": 1 1 1 1 1 1 1 1 1 ...
##  $ veil-color              : Factor w/ 4 levels "n","o","w","y": 3 3 3 3 3 3 3 3 3 ...
##  $ ring-number             : Factor w/ 3 levels "n","o","t": 2 2 2 2 2 2 2 2 2 ...
##  $ ring-type               : Factor w/ 5 levels "e","f","l","n",..: 5 5 5 5 1 5 5 5 5 ...
##  $ spore-print-color       : Factor w/ 9 levels "b","h","k","n",..: 3 4 4 3 4 3 3 4 3 3 ...
##  $ population              : Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...
##  $ habitat                 : Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...
```

names

```
names(mushrooms)
```

```
##  [1] "edible_poisonous"         "cap-shape"
##  [3] "cap-surface"              "cap-color"
##  [5] "bruises?"                 "odor"
##  [7] "gill-attachment"          "gill-spacing"
##  [9] "gill-size"                "gill-color"
## [11] "stalk-shape"              "stalk-root"
## [13] "stalk-surface-above-ring" "stalk-surface-below-ring"
## [15] "stalk-color-above-ring"   "stalk-color-below-ring"
## [17] "veil-type"                "veil-color"
## [19] "ring-number"              "ring-type"
## [21] "spore-print-color"        "population"
## [23] "habitat"
```

Next I subsetted a new data frame using columns, “edible_poisonous”, “odor”, “population”, and “habitat”

```
new_shrooms <- mushrooms[c(1,6,22:23)]
str(new_shrooms)
```

```
## 'data.frame':    8124 obs. of  4 variables:
##  $ edible_poisonous: Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
##  $ odor            : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...
##  $ population      : Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...
##  $ habitat         : Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...
```

```
names(new_shrooms)
```

```
## [1] "edible_poisonous" "odor"             "population"
## [4] "habitat"
```

To convert the abbreviations, I had to change the column to a character, convert the abbreviations, and then re-convert the column back to factor.

```
new_shrooms$edible_poisonous <- as.character(new_shrooms$edible_poisonous)
new_shrooms$edible_poisonous[new_shrooms$edible_poisonous == "p"] <- "poisonous"
new_shrooms$edible_poisonous[new_shrooms$edible_poisonous == "e"] <- "edible"
new_shrooms$edible_poisonous <- as.factor(new_shrooms$edible_poisonous)
```

```
str(new_shrooms)
```

```
## 'data.frame':    8124 obs. of  4 variables:
##  $ edible_poisonous: Factor w/ 2 levels "edible","poisonous": 2 1 1 2 1 1 1 1 2 1 ...
##  $ odor            : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...
##  $ population      : Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...
##  $ habitat         : Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...
```

Repeat the process for the odor observation, but use a switch statement

```
head(new_shrooms$odor,10)
```

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

```
new_shrooms$odor<- as.character(new_shrooms$odor)
new_shrooms$odor <- sapply(new_shrooms$odor, switch,
                    a='almond',
                    l='anise',
                    c='creosote',
                    y='fishy',
                    f='foul',
                    m='musty',
                    n='none',
                    p='pungent',
                    s='spicy')
new_shrooms$odor<- as.factor(new_shrooms$odor)
```

```
head(new_shrooms$odor,10)
```

```
##  [1] pungent almond  anise   pungent none    almond  almond  anise
##  [9] pungent almond
## Levels: almond anise creosote fishy foul musty none pungent spicy
```

```
str(new_shrooms)
```

```
## 'data.frame':    8124 obs. of  4 variables:
##  $ edible_poisonous: Factor w/ 2 levels "edible","poisonous": 2 1 1 2 1 1 1 1 2 1 ...
##  $ odor            : Factor w/ 9 levels "almond","anise",..: 8 1 2 8 7 1 1 2 8 1 ...
##  $ population      : Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...
##  $ habitat         : Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...
```

```
new_shrooms$population<- as.character(new_shrooms$population)
new_shrooms$population <- sapply(new_shrooms$population, switch,
                    a='abundant',
                    c='clustered',
                    n='numerous',
                    s='scattered',
                    v='several',
                    y='solitary')
new_shrooms$population<- as.factor(new_shrooms$population)
```

```
head(new_shrooms$population,10)
```

```
##  [1] scattered numerous  numerous  scattered abundant  numerous  numerous
##  [8] scattered several   scattered
## Levels: abundant clustered numerous scattered several solitary
```

Finally, the habitat observation

```
new_shrooms$habitat<- as.character(new_shrooms$habitat)
new_shrooms$habitat <- sapply(new_shrooms$habitat, switch,
                     g='grasses',
                     l='leaves',
                     m='meadows',
                     p='paths',
                     u='urban',
                     w='waste',
                     d='woods')
new_shrooms$habitat<- as.factor(new_shrooms$habitat)
```

```
head(new_shrooms$habitat,10)
```

```
##  [1] urban   grasses meadows urban   grasses grasses meadows meadows
##  [9] grasses meadows
## Levels: grasses leaves meadows paths urban waste woods
```

```
str(new_shrooms)
```

```
## 'data.frame':    8124 obs. of  4 variables:
##  $ edible_poisonous: Factor w/ 2 levels "edible","poisonous": 2 1 1 2 1 1 1 1 2 1 ...
##  $ odor            : Factor w/ 9 levels "almond","anise",..: 8 1 2 8 7 1 1 2 8 1 ...
##  $ population      : Factor w/ 6 levels "abundant","clustered",..: 4 3 3 4 1 3 3 4 5 4 ...
##  $ habitat         : Factor w/ 7 levels "grasses","leaves",..: 5 1 3 5 1 1 3 3 1 3 ...
```

```
head(new_shrooms,20)
```

```
##    edible_poisonous    odor population habitat
## 1         poisonous pungent  scattered   urban
## 2            edible  almond   numerous grasses
## 3            edible   anise   numerous meadows
## 4         poisonous pungent  scattered   urban
## 5            edible    none   abundant grasses
## 6            edible  almond   numerous grasses
## 7            edible  almond   numerous meadows
## 8            edible   anise  scattered meadows
## 9         poisonous pungent    several grasses
## 10           edible  almond  scattered meadows
## 11           edible   anise   numerous grasses
## 12           edible  almond  scattered meadows
## 13           edible  almond  scattered grasses
## 14        poisonous pungent    several   urban
## 15           edible    none   abundant grasses
## 16           edible    none   solitary   urban
## 17           edible    none   abundant grasses
## 18        poisonous pungent  scattered grasses
## 19        poisonous pungent  scattered   urban
## 20        poisonous pungent  scattered   urban
```