DATA 607 - Week 1 Assignment

Binish Kurian Chandy

February 2, 2018

Read mushroom data set

```
df <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data
str(df)
```

```
## 'data.frame':    8124 obs. of  23 variables:
##  $ V1 : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
##  $ V2 : Factor w/ 6 levels "b","c","f","k",..: 6 6 1 6 6 6 1 1 6 1 ...
##  $ V3 : Factor w/ 4 levels "f","g","s","y": 3 3 3 4 3 4 3 4 4 3 ...
##  $ V4 : Factor w/ 10 levels "b","c","e","g",..: 5 10 9 9 4 10 9 9 9 10 ...
##  $ V5 : Factor w/ 2 levels "f","t": 2 2 2 2 1 2 2 2 2 2 ...
##  $ V6 : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...
##  $ V7 : Factor w/ 2 levels "a","f": 2 2 2 2 2 2 2 2 2 2 ...
##  $ V8 : Factor w/ 2 levels "c","w": 1 1 1 1 2 1 1 1 1 1 ...
##  $ V9 : Factor w/ 2 levels "b","n": 2 1 1 2 1 1 1 1 2 1 ...
##  $ V10: Factor w/ 12 levels "b","e","g","h",..: 5 5 6 6 5 6 3 6 8 3 ...
##  $ V11: Factor w/ 2 levels "e","t": 1 1 1 1 2 1 1 1 1 1 ...
##  $ V12: Factor w/ 5 levels "?","b","c","e",..: 4 3 3 4 4 3 3 3 4 3 ...
##  $ V13: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ V14: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ V15: Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ V16: Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ V17: Factor w/ 1 level "p": 1 1 1 1 1 1 1 1 1 1 ...
##  $ V18: Factor w/ 4 levels "n","o","w","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ V19: Factor w/ 3 levels "n","o","t": 2 2 2 2 2 2 2 2 2 2 ...
##  $ V20: Factor w/ 5 levels "e","f","l","n",..: 5 5 5 5 1 5 5 5 5 5 ...
##  $ V21: Factor w/ 9 levels "b","h","k","n",..: 3 4 4 3 4 3 3 4 3 3 ...
##  $ V22: Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...
##  $ V23: Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...
```

```
dim(df)
```

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

Rename column names

```
colnames(df) <- c("class", "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")
head(df)
```

```
##   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
```

Create new dataframe using subset of columns

```
mushrooms <- subset(df, select=c("class", "odor", "gill-size", "population", "habitat"))
```

Check if new data frame has same number of observation as original

```
dim(df)
```

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

```
dim(mushrooms)
```

```
## [1] 8124    5
```

Replace abbreviations with actual values

```
levels(mushrooms$class)[levels(mushrooms$class) == "e"] <- "edible"
levels(mushrooms$class)[levels(mushrooms$class) == "p"] <- "poisonous"
```

```r
levels(mushrooms$odor)[levels(mushrooms$odor) == "a"] <- "almond"
levels(mushrooms$odor)[levels(mushrooms$odor) == "l"] <- "anise"
levels(mushrooms$odor)[levels(mushrooms$odor) == "c"] <- "creosote"
levels(mushrooms$odor)[levels(mushrooms$odor) == "y"] <- "fishy"
levels(mushrooms$odor)[levels(mushrooms$odor) == "f"] <- "foul"
levels(mushrooms$odor)[levels(mushrooms$odor) == "m"] <- "musty"
levels(mushrooms$odor)[levels(mushrooms$odor) == "n"] <- "none"
levels(mushrooms$odor)[levels(mushrooms$odor) == "p"] <- "pungent"
levels(mushrooms$odor)[levels(mushrooms$odor) == "s"] <- "spicy"

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

levels(mushrooms$population)[levels(mushrooms$population) == "a"] <- "abundant"
levels(mushrooms$population)[levels(mushrooms$population) == "c"] <- "clustered"
levels(mushrooms$population)[levels(mushrooms$population) == "n"] <- "numerous"
levels(mushrooms$population)[levels(mushrooms$population) == "s"] <- "scattered"
levels(mushrooms$population)[levels(mushrooms$population) == "v"] <- "several"
levels(mushrooms$population)[levels(mushrooms$population) == "y"] <- "solitary"

levels(mushrooms$habitat)[levels(mushrooms$habitat) == "g"] <- "grasses"
levels(mushrooms$habitat)[levels(mushrooms$habitat) == "l"] <- "leaves"
levels(mushrooms$habitat)[levels(mushrooms$habitat) == "m"] <- "meadows"
levels(mushrooms$habitat)[levels(mushrooms$habitat) == "p"] <- "paths"
levels(mushrooms$habitat)[levels(mushrooms$habitat) == "u"] <- "urban"
levels(mushrooms$habitat)[levels(mushrooms$habitat) == "w"] <- "waste"
levels(mushrooms$habitat)[levels(mushrooms$habitat) == "d"] <- "woods"

head(mushrooms)
```

```
##       class    odor gill-size population habitat
## 1 poisonous pungent    narrow  scattered   urban
## 2    edible  almond     broad   numerous grasses
## 3    edible   anise     broad   numerous meadows
## 4 poisonous pungent    narrow  scattered   urban
## 5    edible    none     broad   abundant grasses
## 6    edible  almond     broad   numerous grasses
```

```r
tail(mushrooms)
```

```
##          class  odor gill-size population habitat
## 8119 poisonous  foul    narrow    several   woods
## 8120    edible  none     broad  clustered  leaves
## 8121    edible  none     broad    several  leaves
## 8122    edible  none     broad  clustered  leaves
## 8123 poisonous fishy    narrow    several  leaves
## 8124    edible  none     broad  clustered  leaves
```