Data\_607\_Week\_1\_Assignment

Jimmy Ng

1/29/2019

# load packages  
library(tidyverse)

## -- Attaching packages -------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts ----------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# download, read data  
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data"  
download.file(url, "mushroom.txt", method = "wininet", quiet = FALSE, mode = "w", cacheOK = TRUE)

# read, extract & transform data  
df <- read.table("mushroom.txt", header = F, sep = ",") %>%  
 # select 4 columns only  
 dplyr::select(V1, V2, V3, V4) %>%  
 dplyr::rename("classes" = V1, "cap-shape" = V2, "cap-surface" = V3, "cap-color" = V4) %>%  
 dplyr::mutate(classes = dplyr::case\_when(classes == "e" ~ "edible",  
 classes == "p" ~ "poisonous",  
 TRUE ~ "others"),  
 `cap-shape` = dplyr::case\_when(`cap-shape` == "b" ~ "bell",  
 `cap-shape` == "c" ~ "conical",  
 `cap-shape` == "x" ~ "convex",  
 `cap-shape` == "f" ~ "flat",  
 `cap-shape` == "k" ~ "knobbed",  
 `cap-shape` == "s" ~ "sunken",  
 TRUE ~ "others"),  
 `cap-surface` = dplyr::case\_when(`cap-surface` == "f" ~ "fibrous",  
 `cap-surface` == "g" ~ "grooves",   
 `cap-surface` == "y" ~ "scaly",  
 `cap-surface` == "s" ~ "smooth",  
 TRUE ~ "others"),  
 `cap-color` = dplyr::case\_when(`cap-color` == "n" ~ "brown",  
 `cap-color` == "b" ~ "buff",  
 `cap-color` == "c" ~ "cinnamon",  
 `cap-color` == "g" ~ "gray",  
 `cap-color` == "r" ~ "green",  
 `cap-color` == "p" ~ "pink",  
 `cap-color` == "u" ~ "purple",  
 `cap-color` == "e" ~ "red",  
 `cap-color` == "w" ~ "white",  
 `cap-color` == "y" ~ "yellow",  
 TRUE ~ "others"))

# check summary, missing value & unique values by column  
purrr::map(list(summary,   
 function(x) {colSums(is.na(x))},  
 function(x) {sapply(x, unique)}),   
 function(x){x(df)})  
## [[1]]  
## classes cap-shape cap-surface   
## Length:8124 Length:8124 Length:8124   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
## cap-color   
## Length:8124   
## Class :character   
## Mode :character   
##   
## [[2]]  
## classes cap-shape cap-surface cap-color   
## 0 0 0 0   
##   
## [[3]]  
## [[3]]$classes  
## [1] "poisonous" "edible"   
##   
## [[3]]$`cap-shape`  
## [1] "convex" "bell" "sunken" "flat" "knobbed" "conical"  
##   
## [[3]]$`cap-surface`  
## [1] "smooth" "scaly" "fibrous" "grooves"  
##   
## [[3]]$`cap-color`  
## [1] "brown" "yellow" "white" "gray" "red" "pink"   
## [7] "buff" "purple" "cinnamon" "green"

# print & plot distribution by column  
par(mfrow = c(2, 2))  
sapply(df, function(x) round(prop.table(ftable(x)), 2)) %>%  
 print %>%  
lapply(., barplot) %>%  
 invisible

## $classes  
## x edible poisonous  
##   
## 0.52 0.48  
##   
## $`cap-shape`  
## x bell conical convex flat knobbed sunken  
##   
## 0.06 0.00 0.45 0.39 0.10 0.00  
##   
## $`cap-surface`  
## x fibrous grooves scaly smooth  
##   
## 0.29 0.00 0.40 0.31  
##   
## $`cap-color`  
## x brown buff cinnamon gray green pink purple red white yellow  
##   
## 0.28 0.02 0.01 0.23 0.00 0.02 0.00 0.18 0.13 0.13

