Week 1 Assignment

Samriti Malhotra

January 30, 2019

Introduction

Week 1 assignment for DATA 607 is to subset the data provided by the UCI for Mushroom dataset located at below address:

https://archive.ics.uci.edu/ml/datasets/Mushroom

The actual dataset can be found here:

https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data

About the Data

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family.Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one

Data Dictionary

Attribute Information:

0. Type : 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

Problem Statement:-

Load data from given URL(https://archive.ics.uci.edu/ml/datasets/Mushroom) into R , subset and create a new Data frame selecting few columns from original including 1st column. Provide meangingfull name to columns Headers, and also update the values of each column based on the data dictionary values.

Loading Libraries

First step to load the necessary libraries required for this assignment.

```
library(stringr)
library(XML)
```

```
## Warning: package 'XML' was built under R version 3.5.2
```

```
library(maps)
library(httr)
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
```

R-code

```
mushroom_table <- read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricu

mushrooms <- as.data.frame(mushroom_table)


# subsetting the data to create a a new data frame with 5 columns
mush_subset <- subset(mushrooms, select=c(1,2,3,4,6))

#providing columns meaningfull names
colnames(mush_subset) <- c("Type","Shape","Surface","Color","Odor")

#Changing the column values with meaningfull values based on conditional statements
mush_subset$Type <- ifelse(str_detect(mush_subset$Type, "e") == TRUE, "Edible", "Poisonous")
ELSE <- TRUE
```

```
##Pipe operator %>% allows you to pipe the output of one function to the input of another functione. It
##mutate function will add new columns to dataframe
## with function hlps in constructing an environemnt from data, possible modifying(a copy of) the orign
#bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
mush_subset <- mush_subset %>% mutate(.,Shape1 = with(.,case_when(
  (mush_subset$Shape == "x") ~ "convex",
  (mush_subset$Shape == "b") ~ "bell",
  (mush_subset$Shape == "c") ~ "conical",
  (mush_subset$Shape == "k") ~ "knobbed",
  (mush_subset$Shape == "f") ~ "flat",
  ELSE ~ "sunken"
)))

#fibrous=f,grooves=g,scaly=y,smooth=s
mush_subset <- mush_subset %>% mutate(.,Surface = with(.,case_when(
  (mush_subset$Surface == "f") ~ "fibrous",
  (mush_subset$Surface == "g") ~ "grooves",
  (mush_subset$Surface == "y") ~ "scaly",
  ELSE ~ "smooth"
)))

#brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
mush_subset <- mush_subset %>% mutate(.,Color = with(.,case_when(
  (mush_subset$Color == "n") ~ "brown",
  (mush_subset$Color == "b") ~ "buff",
  (mush_subset$Color == "c") ~ "cinnamon",
  (mush_subset$Color == "g") ~ "gray",
  (mush_subset$Color == "r") ~ "green",
  (mush_subset$Color == "p") ~ "pink",
  (mush_subset$Color == "u") ~ "purple",
  (mush_subset$Color == "e") ~ "red",
  (mush_subset$Color == "w") ~ "white",
  ELSE ~ "yellow"
)))

#almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
mush_subset <- mush_subset %>% mutate(.,Odor = with(.,case_when(
  (mush_subset$Odor == "a") ~ "almond",
  (mush_subset$Odor == "l") ~ "anise",
  (mush_subset$Odor == "c") ~ "creosote",
  (mush_subset$Odor == "y") ~ "fishy",
  (mush_subset$Odor == "f") ~ "foul",
  (mush_subset$Odor == "m") ~ "musty",
  (mush_subset$Odor == "p") ~ "pungent",
  (mush_subset$Odor == "n") ~ "none",
  ELSE ~ "spicy"


)))
```

Summary

The data frame after subsetting, has been given meaningfull name and updating column values is as below.

```r
head(mush_subset,n=20)
```

```
##         Type Shape Surface  Color    Odor Shape1
## 1  Poisonous     x  smooth  brown pungent convex
## 2     Edible     x  smooth yellow  almond convex
## 3     Edible     b  smooth  white   anise   bell
## 4  Poisonous     x   scaly  white pungent convex
## 5     Edible     x  smooth   gray    none convex
## 6     Edible     x   scaly yellow  almond convex
## 7     Edible     b  smooth  white  almond   bell
## 8     Edible     b   scaly  white   anise   bell
## 9  Poisonous     x   scaly  white pungent convex
## 10    Edible     b  smooth yellow  almond   bell
## 11    Edible     x   scaly yellow   anise convex
## 12    Edible     x   scaly yellow  almond convex
## 13    Edible     b  smooth yellow  almond   bell
## 14 Poisonous     x   scaly  white pungent convex
## 15    Edible     x fibrous  brown    none convex
## 16    Edible     s fibrous   gray    none sunken
## 17    Edible     f fibrous  white    none   flat
## 18 Poisonous     x  smooth  brown pungent convex
## 19 Poisonous     x   scaly  white pungent convex
## 20 Poisonous     x  smooth  brown pungent convex
```