Assignment1_607

Vishal Arora

Feburary 1, 2019

DATA 607 Week 1 Assignment - Loading Data into a Data Frame

Introduction

In week 1 assignment for DATA607. we will load data into dataframe that is provided by UCI Mushroom dataset located : https://archive.ics.uci.edu/ml/datasets/Mushroom

About The data

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended.

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.

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)
```

R-Code starting :-

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

mushrooms <- as.data.frame(mush_table)


# subsetting the data to create a a new data frame with 5 columns
mush_subset <- subset(mushrooms, select=c(1,2,3,4,6,22,23))

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

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

#bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
mush_subset$Shape <- ifelse(str_detect(mush_subset$Shape, "x") == TRUE, "convex", ifelse(str_detect(mus

#fibrous=f,grooves=g,scaly=y,smooth=s
mush_subset$Surface <- ifelse(str_detect(mush_subset$Surface, "f") == TRUE, "fibrous", ifelse(str_detec

#brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
mush_subset$Color <- ifelse(str_detect(mush_subset$Color, "n") == TRUE, "brown", ifelse(str_detect(mush

#almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
mush_subset$Odor <- ifelse(str_detect(mush_subset$Odor, "a") == TRUE, "almond", ifelse(str_detect(mush_

#population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
mush_subset$Population_Type <- ifelse(str_detect(mush_subset$Population_Type, "a") == TRUE, "abundant",
```

```
#habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d
mush_subset$Habitat <- ifelse(str_detect(mush_subset$Habitat, "g") == TRUE, "grasses", ifelse(str_detec
```

Summary:-

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

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

```
##          Type  Shape Surface  Color    Odor Population_Type Habitat
## 1  Poisonous convex  smooth  brown pungent       scattered   paths
## 2     Edible convex  smooth yellow  almond        numerous grasses
## 3     Edible   bell  smooth  white   anise        numerous meadows
## 4  Poisonous convex   scaly  white pungent       scattered   paths
## 5     Edible convex  smooth   gray    none        abundant grasses
## 6     Edible convex   scaly yellow  almond        numerous grasses
## 7     Edible   bell  smooth  white  almond        numerous meadows
## 8     Edible   bell   scaly  white   anise       scattered meadows
## 9  Poisonous convex   scaly  white pungent         several grasses
## 10    Edible   bell  smooth yellow  almond       scattered meadows
## 11    Edible convex   scaly yellow   anise        numerous grasses
## 12    Edible convex   scaly yellow  almond       scattered meadows
## 13    Edible   bell  smooth yellow  almond       scattered grasses
## 14 Poisonous convex   scaly  white pungent         several   paths
## 15    Edible convex fibrous  brown    none        abundant grasses
## 16    Edible sunken fibrous   gray    none        solitary   urban
## 17    Edible   flat fibrous  white    none        abundant grasses
## 18 Poisonous convex  smooth  brown pungent       scattered grasses
## 19 Poisonous convex   scaly  white pungent       scattered   paths
## 20 Poisonous convex  smooth  brown pungent       scattered   paths
```