Part I

Download the Mushroom Dataset from the UCI repository: <https://archive.ics.uci.edu/ml/datasets/mushroom> (Links to an external site.) The name of the dataset file is agaricus-lepiota.data Missing values in this dataset are represented using a questions mark: ?

#Prints the number of missing values in each row  
#Prints the number of missing values in each column  
  
  
import pandas as pd   
import numpy as np  
  
  
df = pd.read\_csv("agaricus-lepiota.data" ,header =None)  
df

*output:*

0 1 2 3 4 5 6 7 8 9 ... 13 14 15 16 17 18 19 20 21 22  
0 p x s n t p f c n k ... s w w p w o p k s u  
1 e x s y t a f c b k ... s w w p w o p n n g  
2 e b s w t l f c b n ... s w w p w o p n n m  
3 p x y w t p f c n n ... s w w p w o p k s u  
4 e x s g f n f w b k ... s w w p w o e n a g  
... .. .. .. .. .. .. .. .. .. .. ... .. .. .. .. .. .. .. .. .. ..  
8119 e k s n f n a c b y ... s o o p o o p b c l  
8120 e x s n f n a c b y ... s o o p n o p b v l  
8121 e f s n f n a c b n ... s o o p o o p b c l  
8122 p k y n f y f c n b ... k w w p w o e w v l  
8123 e x s n f n a c b y ... s o o p o o p o c l  
  
[8124 rows x 23 columns]

Text

Description automatically generated

for i in range(23):  
 print(df[i].value\_counts());  
 print("==end==")

*output:*

e 4208  
p 3916  
Name: 0, dtype: int64  
==end==  
x 3656  
f 3152  
k 828  
b 452  
s 32  
c 4  
Name: 1, dtype: int64  
==end==  
y 3244  
s 2556  
f 2320  
g 4  
Name: 2, dtype: int64  
==end==  
n 2284  
g 1840  
e 1500  
y 1072  
w 1040  
b 168  
p 144  
c 44  
r 16  
u 16  
Name: 3, dtype: int64  
==end==  
f 4748  
t 3376  
Name: 4, dtype: int64  
==end==  
n 3528  
f 2160  
y 576  
s 576  
a 400  
l 400  
p 256  
c 192  
m 36  
Name: 5, dtype: int64  
==end==  
f 7914  
a 210  
Name: 6, dtype: int64  
==end==  
c 6812  
w 1312  
Name: 7, dtype: int64  
==end==  
b 5612  
n 2512  
Name: 8, dtype: int64  
==end==  
b 1728  
p 1492  
w 1202  
n 1048  
g 752  
h 732  
u 492  
k 408  
e 96  
y 86  
o 64  
r 24  
Name: 9, dtype: int64  
==end==  
t 4608  
e 3516  
Name: 10, dtype: int64  
==end==  
**b 3776  
? 2480  
e 1120  
c 556  
r 192  
Name: 11, dtype: int64  
==end==**  
s 5176  
k 2372  
f 552  
y 24  
Name: 12, dtype: int64  
==end==  
s 4936  
k 2304  
f 600  
y 284  
Name: 13, dtype: int64  
==end==  
w 4464  
p 1872  
g 576  
n 448  
b 432  
o 192  
e 96  
c 36  
y 8  
Name: 14, dtype: int64  
==end==  
w 4384  
p 1872  
g 576  
n 512  
b 432  
o 192  
e 96  
c 36  
y 24  
Name: 15, dtype: int64  
==end==  
p 8124  
Name: 16, dtype: int64  
==end==  
w 7924  
n 96  
o 96  
y 8  
Name: 17, dtype: int64  
==end==  
o 7488  
t 600  
n 36  
Name: 18, dtype: int64  
==end==  
p 3968  
e 2776  
l 1296  
f 48  
n 36  
Name: 19, dtype: int64  
==end==  
w 2388  
n 1968  
k 1872  
h 1632  
r 72  
y 48  
b 48  
u 48  
o 48  
Name: 20, dtype: int64  
==end==  
v 4040  
y 1712  
s 1248  
n 400  
a 384  
c 340  
Name: 21, dtype: int64  
==end==  
d 3148  
g 2148  
p 1144  
l 832  
u 368  
m 292  
w 192  
Name: 22, dtype: int64  
==end==

df[df.columns[11]].value\_counts()

b 3776  
? 2480  
e 1120  
c 556  
r 192  
Name: 11, dtype: int64

Graphical user interface, application

Description automatically generated

#Removes the column that contains missing values.  
df =df.drop([11] ,axis=1)  
df

*output:*

0 1 2 3 4 5 6 7 8 9 ... 13 14 15 16 17 18 19 20 21 22  
0 p x s n t p f c n k ... s w w p w o p k s u  
1 e x s y t a f c b k ... s w w p w o p n n g  
2 e b s w t l f c b n ... s w w p w o p n n m  
3 p x y w t p f c n n ... s w w p w o p k s u  
4 e x s g f n f w b k ... s w w p w o e n a g  
... .. .. .. .. .. .. .. .. .. .. ... .. .. .. .. .. .. .. .. .. ..  
8119 e k s n f n a c b y ... s o o p o o p b c l  
8120 e x s n f n a c b y ... s o o p n o p b v l  
8121 e f s n f n a c b n ... s o o p o o p b c l  
8122 p k y n f y f c n b ... k w w p w o e w v l  
8123 e x s n f n a c b y ... s o o p o o p o c l  
  
[8124 rows x 22 columns]

Graphical user interface, text, application

Description automatically generated

**Part 2 :**

Download the Dataset file from Kaggle: <https://www.kaggle.com/lava18/google-play-store-apps?select=googleplaystore.csv> (Links to an external site.) The file name is googleplaystore.csv

#The first step is to decide whether the rows or the columns that contain missing values should be removed  
  
import pandas as pd   
import numpy as np   
  
  
df\_gps = pd.read\_csv("googleplaystore.csv")  
  
df\_gps

Graphical user interface, text, application, email

Description automatically generated

*Graphical user interface, application

Description automatically generated*

# check which columns contain missing values in the dataframe by calling the isnull() function:  
  
print(df\_gps.isnull().sum())

*output:*

App 0  
Category 0  
Rating 1474  
Reviews 0  
Size 0  
Installs 0  
Type 1  
Price 0  
Content Rating 1  
Genres 0  
Last Updated 0  
Current Ver 8  
Android Ver 3  
dtype: int64

A screenshot of a computer

Description automatically generated

#Find how many missing values are in each column.   
#List the number of missing values within each column.  
  
   
df\_gps = df\_gps.drop(['Rating'],axis=1)  
print(df\_gps.isnull().sum())  
   
df\_gps

*output:*

App 0  
Category 0  
Reviews 0  
Size 0  
Installs 0  
Type 1  
Price 0  
Content Rating 1  
Genres 0  
Last Updated 0  
Current Ver 8  
Android Ver 3  
dtype: int64

Graphical user interface, text, application

Description automatically generated

#Remove any other rows that contain missing values. You can do so using the dropna() function.  
  
df\_gps = df\_gps.dropna();  
print(df\_gps.isnull().sum())  
df\_gps

*output:*

App 0  
Category 0  
Reviews 0  
Size 0  
Installs 0  
Type 0  
Price 0  
Content Rating 0  
Genres 0  
Last Updated 0  
Current Ver 0  
Android Ver 0  
dtype: int64

Graphical user interface, application

Description automatically generated