TP Visualization

Panda is an open source library with powerful data exploratory and preparation tools.

Download Credit Approval data set for the UCI machine learning website. The attributes have been changed to protect the confidentiality of the data.

1) Provide basic statistics from the data set : number of attributes, number of objects, number of classes. For each attributes, gives

- for numerical features the minimum, the maximum, the mean, the standard deviation,
- for categorical features the mode and the domain,
- for both the percentage of missing values.

2) From this first analysis, some preprocessing step has to be performed

- assign the attributes to Gender, Age, debt, married, bankCustomer, EducationLevel, Ethnicity, YearsEmployed, PriorDefault, Employed, CreditScore, DriversLicense, Citizen, ZipCode, Income.
- Values 't' means 'true' and 'f' means false. Convert what is possible to booleans.
- Replace values 'a' and 'b' by 'male' and 'female' respectively.
- The ZipCode contains a lot of 00000 values. Replace it by NaN.

3) Clean missing values and drop column ‘ZipCode’.

4) Plot an histogram for the Age attribute.

5) Plot a cumulative histogram of approval based on employed. Plot the same information, but with a mosaic plot.

6) Present boxplots associated to continuous data. Is there some outliers ?

7) Create a scatter plot matrix for attributes Age, Debts and YearsEmployed. Can you observe some groups and/or some correlations ?

Corrections (supplementary to the code)

http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.html
https://www.kaggle.com/hafidhfikri/loan-approval-prediction

1) n=690, p=15, c=2. Missing values makes some calculation impossible. Replace '?' by 'nan'.

att.	type	min	max	mean	std	mode	domain	%MV
att0	categorical					b	a,b	1.74
att1	numerical	13.75	80.25	31.57	11.96			1.74
att2	numerical	0	28	4.76	4.98			0
att3	categorical					u	l,u,y	0.87
att4	categorical					g	g,gg,p	0.87
att5	categorical					c	aa,c,cc,d,e,ff,i,j,k,m,q,r,w,x	1.30
att6	categorical					v	bb,dd,ff,h,j,n,o,v,z	1.30
att7	numerical	0	28.5	2.22	3.35			0
att8	categorical					t	f,t	0
att9	categorical					f	f,t	0
att10	numerical	0	67	2.40	4.86			0
att11	categorical					f	f,t	0
att12	categorical					g	g,p,s	0
att13	categorical					00000	postale code	1.88
att14	numerical	0	100000	1017.39	5210.10			0

3) The percentage of MV per columns is low and the maximum of MV in a line is 4. Thus, we decide not to remove columns and lines. We are replacing with the median. Indeed, the last attribute has its min and its mean very close compare to its max. We suspect outliers.

4) The histogram is skewed to right (tail is longer). Note that some persons are less that 18 years old. (Note 2019/01 : mosaic plot for 3 variables does not exists yet)

5) As expected, employed persons have more luck to get a credit than unemployed persons.

6) There are a lot of outliers...

7) Yes, we can see correlations and 2 groups with age and debt.

Code

```
# python correction.py
# exec(open("correction.py").read())
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from statsmodels.graphics.mosaicplot import mosaic

######## Question 1 ########
data = pd.read_csv("crx.data",header=None)
#data.info()
x=data.loc[:,0:data.shape[1]-2]
x=x.replace('?', np.nan)
y=data[data.shape[1]-1]

n,p=x.shape
c=len(np.unique(y))

idxNum=[1,2,7,10,14]
idxCat=[0,3,4,5,6,8,9,11,12,13]

x[1]=x[1].astype(float) # because of nan, object type of column 1 is 'object' and not numerical

x[idxNum].min()
x[idxNum].max()
x[idxNum].mean()
x[idxNum].std()

x[idxCat].mode() # 0 at the begining means the first attributes values mode. If there are more, a second line is used with a 1, etc.
for i,idxi in enumerate(idxCat):
    #print(i,idxi)
    np.unique(x[idxi])

x.isnull().sum() # number of MV
x.isnull().sum()*100/len(x) # percentage of MV


######## Question 2 ########
x.columns=['Gender','Age','debt','married','bankCustomer','EducationLevel','Ethnicity','YearsEmployed','PriorDefault','Employed','CreditScore','DriversLicense','Citizen','ZipCode','Income']
#x.info()

x=x.replace('a','m')
x=x.replace('b','f')

x=x.replace('t',1)
x=x.replace('f',0)
x[['PriorDefault','Employed','DriversLicense']]=x[['PriorDefault','Employed','DriversLicense']].astype(bool) # convert 0,1 to boolean
```

```
# ou x.iloc[:,[8,9,11]]

x['ZipCode']=x['ZipCode'].replace('00000', np.nan)

######## Question 3 ########
np.where(x.isnull().sum(axis=1)>5) # index of lines with more than 5 MV.
# x.dropna(thresh=5,inplace=True) # drop line with at least 5 MV

# median for numeral data, mode for categorical data
values = x.mode()
median = x.median()
values[median.index]=median.values
x['Gender'].fillna(value=values['Gender'][0], inplace=True)
x['Age'].fillna(value=values['Age'][0], inplace=True)
x['married'].fillna(value=values['married'][0], inplace=True)
x['bankCustomer'].fillna(value=values['bankCustomer'][0], inplace=True)
x['EducationLevel'].fillna(value=values['EducationLevel'][0], inplace=True)
x['Ethnicity'].fillna(value=values['Ethnicity'][0], inplace=True)

x=x.drop(['ZipCode'],axis=1)

######## Question 4 ########
plt.hist(x['Age'],bins=10)
plt.show(block=False)


######## Question 5 ########
xsel=pd.concat([x['Employed'],y],axis=1)
xsel.columns=['Employed','Class']

df2=xsel.groupby(['Employed','Class'])['Employed'].count().unstack()
df2.plot(kind='bar', stacked=True,label='Employed')
plt.show(block=False)

mosaic(xsel, ['Employed','Class'])
plt.show(block=False)


######## Question 6 ########
x.plot.box()
plt.show(block=False)

x[['Age','debt','YearsEmployed','CreditScore']].plot.box()
plt.show(block=False)


######## Question 7 ########
spm = pd.tools.plotting.scatter_matrix(x[['Age','debt','YearsEmployed']], alpha=0.2, figsize=(6, 6), diagonal='hist')
plt.show(block=False)
```