TP

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 ?