

Aim Shams University Faculty of Engineering ICHEP

Computer Engineering and Software Systems

# **CSE:382 Data Mining & Business Intelligence**

# **Course Project**

**Prepared by:**

Sohayla Ihab Hamed [19P7343]

Youssef Mahmoud Massoud [18P8814]

Salma Ihab Abdelmawgoud [19P8794]

Yasmin Haitham Abdelmoaty [18P3102]

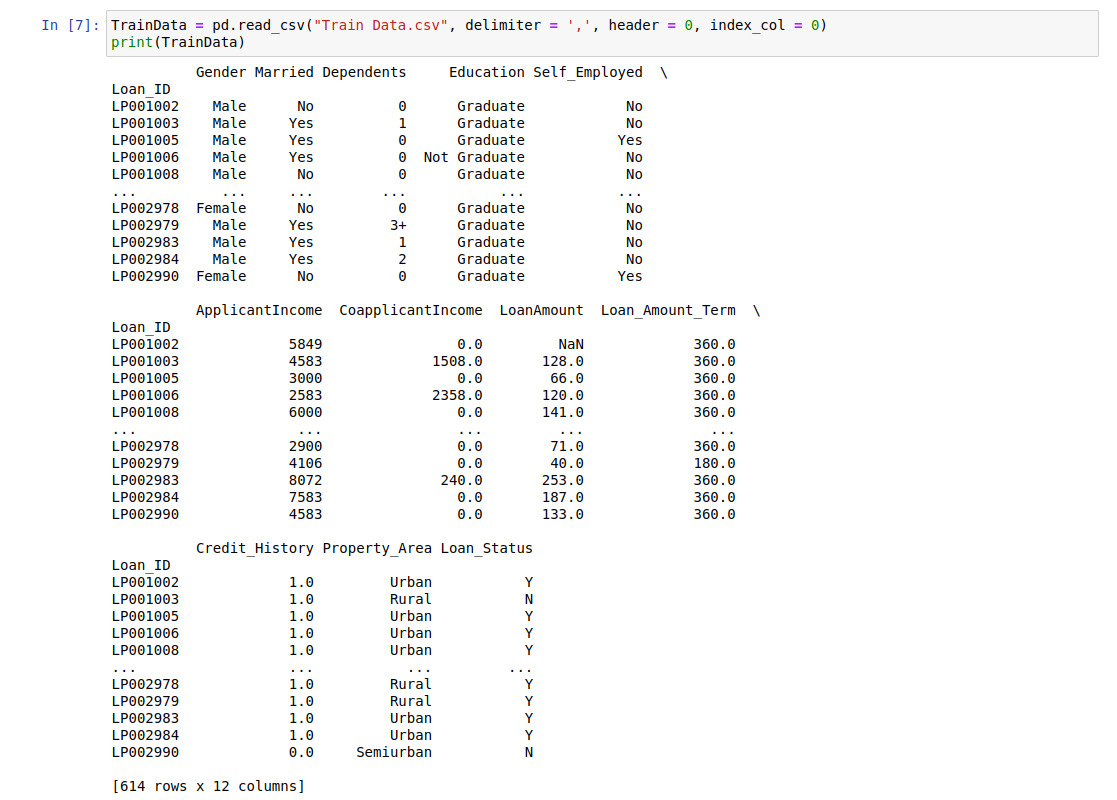
Youssef Hany Onsy [18P1789]

Loan prediction is a common data mining problem which most retail banks solve on a daily basis. It’s the process of predicting who deserves to receive a given loan and at what rate based on certain characteristics of the borrower, be it an individual or a company. Those characteristics are mined for their use in a risk assessment process to determine the amount of risks the lender (the bank) will be incurring when loaning the particular individual/company. For example, some banks can model their interest rate for lending based on how much risk their model assumes a certain individual pose, and thereby require higher interest rates for those who pose higher risks of default and vice versa.

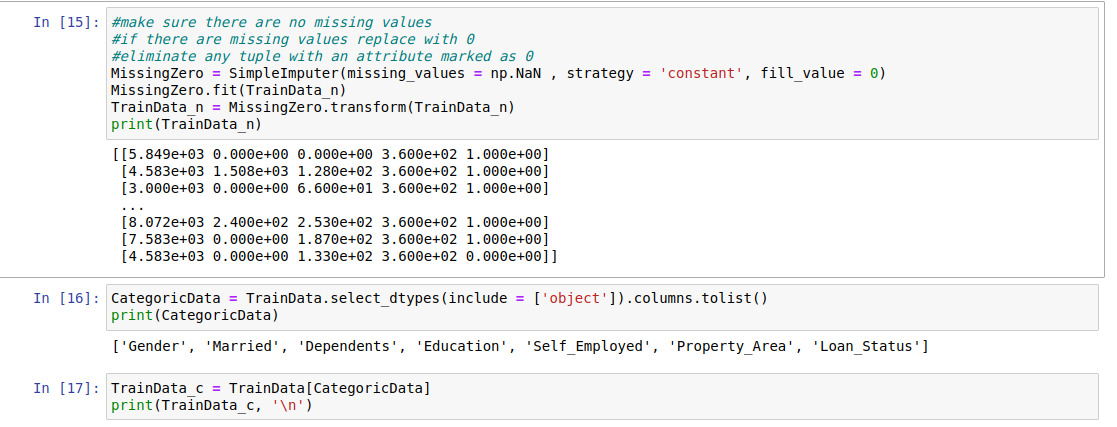
In our project, we use data mining techniques to analyze and predict whether a certain individual can be allowed to take a loan from our bank or whether said individual shall be denied the loan. This will be based on a set of features like marital status, education, employments and other features within the restrictions of the dataset we were provided. We will apply various data mining techniques to achieve our required goal where we end up with a classification model that we will train and later test.

We begin with a set of preprocessing methods used to preprocess our given dataset, the first of which is performing a data cleaning step.

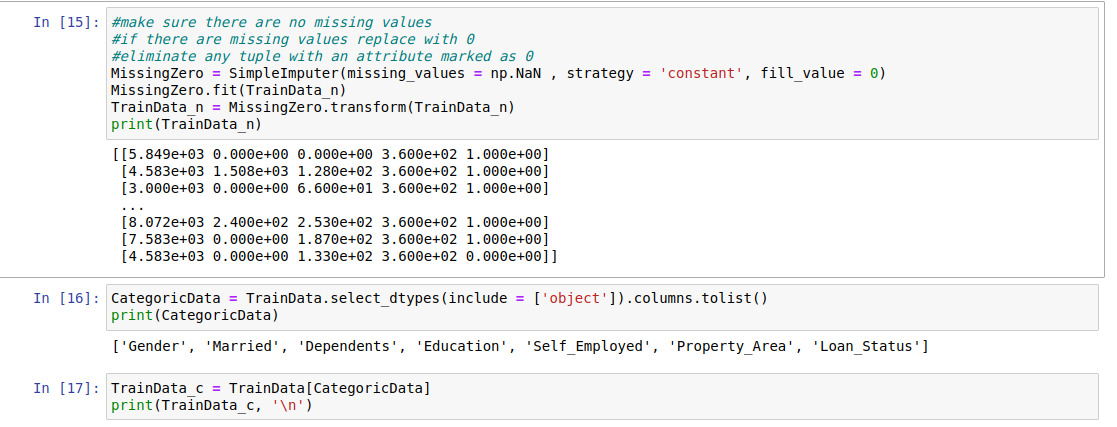
We noticed that some of the training datapoints in the dataset lack information, and so we begin our data cleaning phase by handling the missing Data.

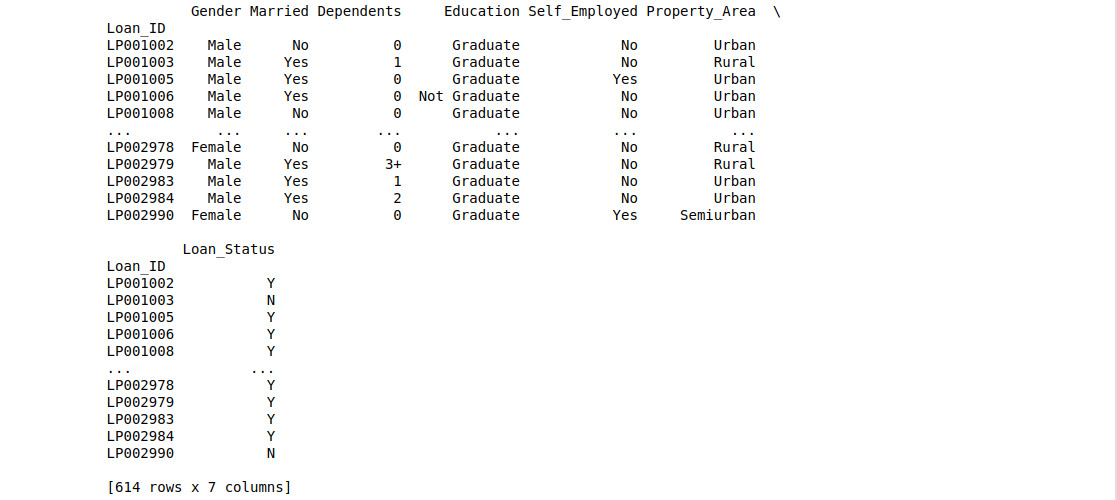


We identify missing data of both types, numerical data as well as categorical ones.



Then once we identified the missing values we replace all of those missing values with zeros, and eliminate any tuple with attributes marked as zero.





“Some information shall be added here it seems like youssef hany filled the missing data using some form of regression I will make sure of that before I add it into the documentation”

The next step in our data cleaning phase is handling the noise found in the dataset, we handle the noise also referred to as outliers using a common binning approach which is a pandas cut, and we group the values into 10 bins as shown below.

