Report of Exercise 1-1

Yichen Lin

**Question a:**

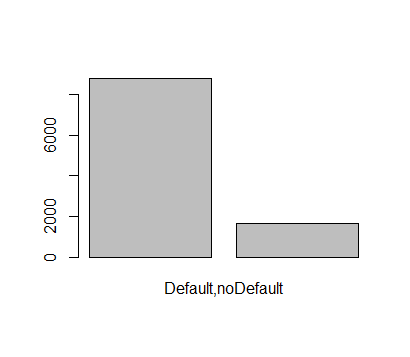
The dataset I choose is the public data from a P2P lending website. The link is as follows:

*https://www.bondora.com/en/public-reports#shared-legend*

Last semester, I use this dataset to train my model and predict whether the loan will be overdue and how likely it will be overdue. Honestly, the result of the model didn’t meet my expectation. And that’s why I want to visualize and have a better understanding of this dataset this semester.

**Question b:**

I use Rstudio to visualize an example bar chart for whether the loan was overdue on the train data. The plot is as follows:



It’s an essential process before we start to train model because we need to figure out the proportion of each value. When the proportion of data is unbalanced, using the default classification threshold may cause all outputs to be counterexamples, resulting in false high accuracy and classification failure.

The detailed R file will be uploaded also.