Cleaning and Variable Analysis

Cleaning and Preparing the Data

```
dat = read.csv("train.csv", skip = 1)
validInd = sample(1:nrow(dat), nrow(dat)/4)
train = dat[-validInd, ]
valid = dat[validInd, ]
trainX = train[ ,1:(ncol(dat)-1)]
trainY = train[ ,ncol(dat)]
validX = valid[ ,1:(ncol(dat)-1)]
validY = valid[ ,ncol(dat)]
```

Univariate Analysis

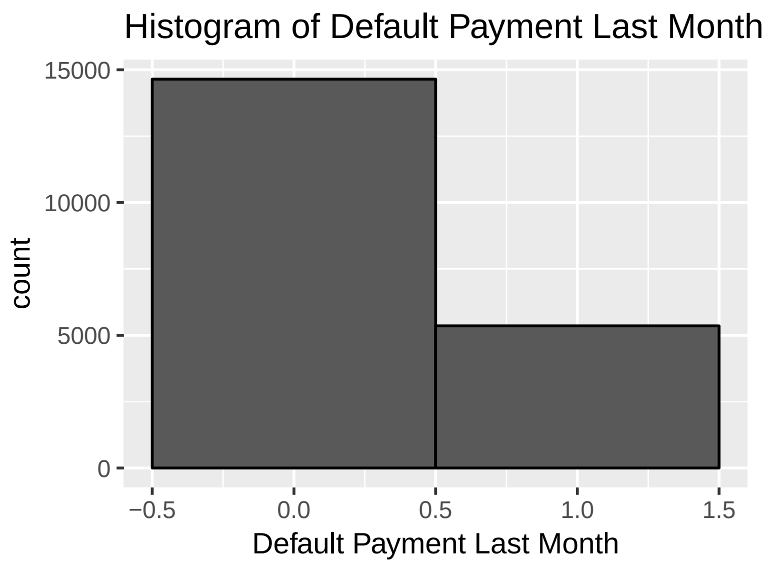
Variable Descriptions

Table 1: Description of variables in the data set.

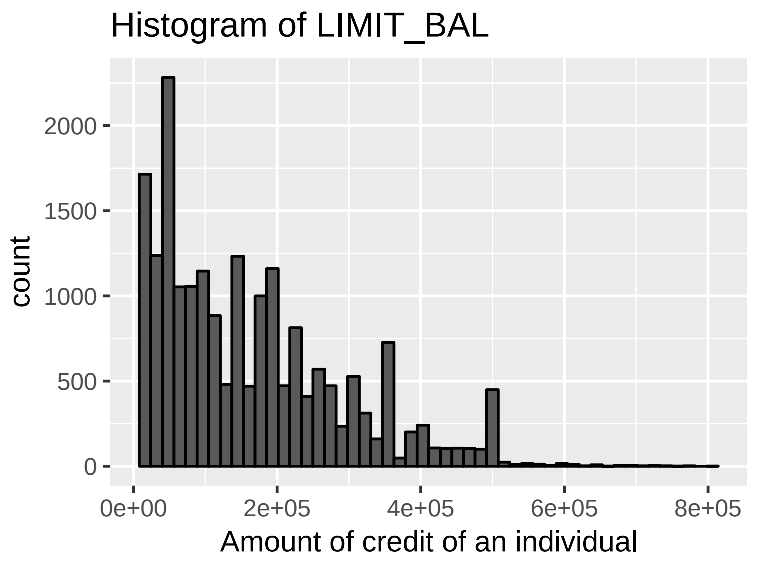
Variable Name	Data Type	Role in model	Description
default payment next month	Factor	Response	1 = a default payment, 0 = no default
LIMIT_BAL	Numeric	Predictor	Amount of credit of an individual, in NT dollars
SEX	Factor	Predictor	Sex of an individual; 1 = male, 2 = female
EDUCATION	Factor	Predictor	Education status of an individual; 1 = graduate school, 2 = university, 3 = high school, 4 = other education
MARRIAGE	Factor	Predictor	Martial status of an individual; 1 = married, 2 = single, 3 = other
AGE	Numeric	Predictor	Age of an individual
PAY_0 to PAY_6	Factor	Predictor	History of payment of an individual, from April (PAY_6) to September (PAY_0) 2015; -1 = on time, other values are months of delay in repayment
BILL_AMT1 to BILL_AMT6	Numeric	Predictor	Amount of bill statement, from April (BILL_AMT6) to September (BILL_AMT1) 2015, in NT dollars
PAY_AMT1 to PAY_AMT6	Numeric	Predictor	Amount of previous payment, from April (PAY_AMT6) to September (PAY_AMT1) 2015, in NT dollars

Univariate Plots

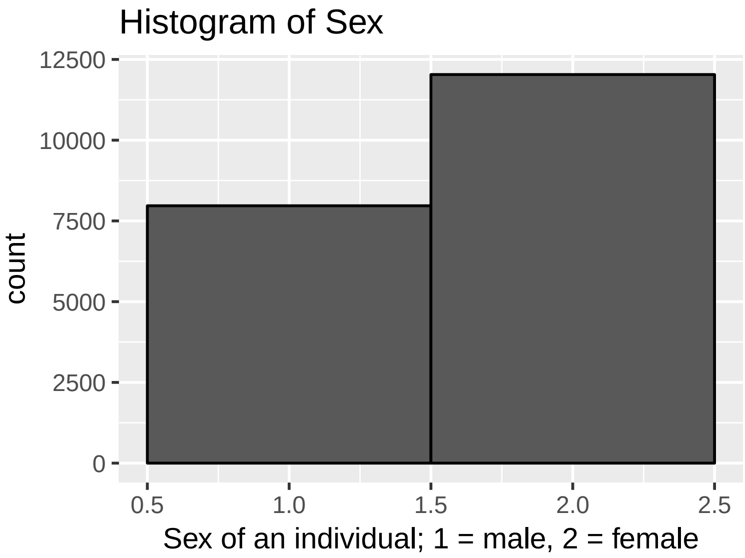
```
ggplot(dat, aes(x = default.payment.next.month)) +
  geom_histogram(bins = 2, col = "black") +
  ggtitle("Histogram of Default Payment Last Month (name?)") +
  xlab("Default Payment Last Month")
```



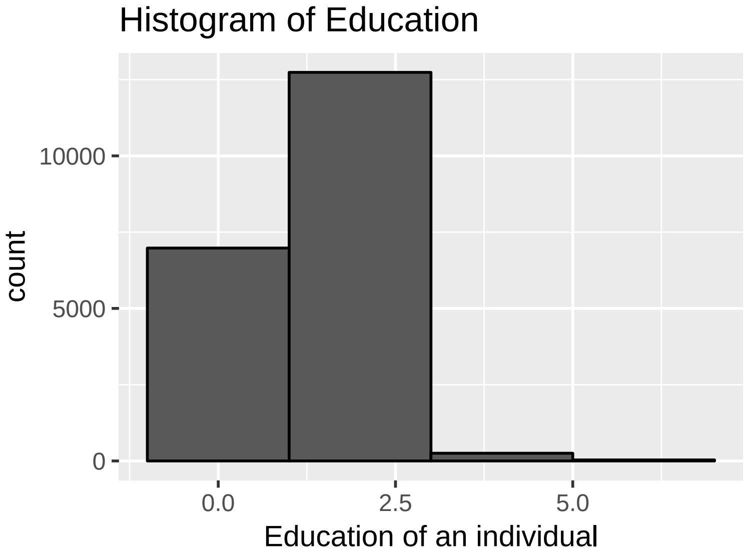
```r
ggplot(dat, aes(x = LIMIT_BAL)) +
  geom_histogram(bins = 50, col = "black") +
  ggtitle("Histogram of LIMIT_BAL") +
  xlab("Amount of credit of an individual")
```



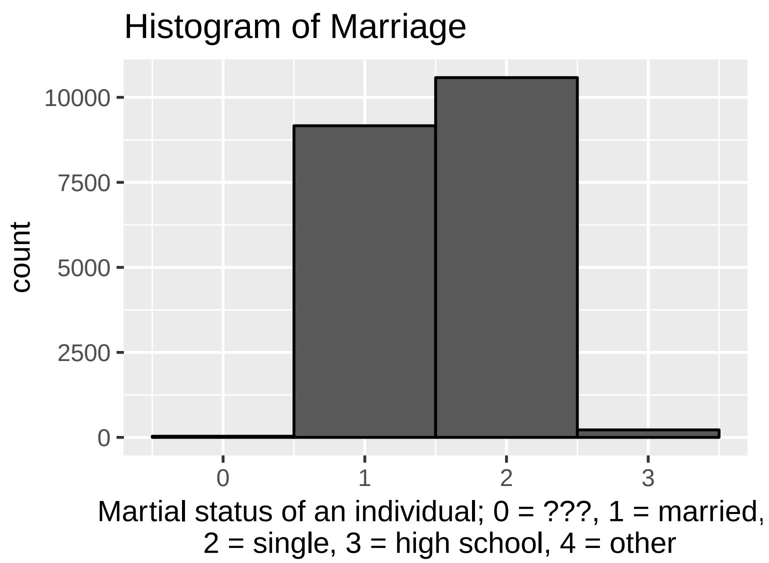
```r
ggplot(dat, aes(x = SEX)) +
  geom_histogram(bins = 2, col = "black") +
  ggtitle("Histogram of Sex") +
  xlab("Sex of an individual; 1 = male, 2 = female")
```



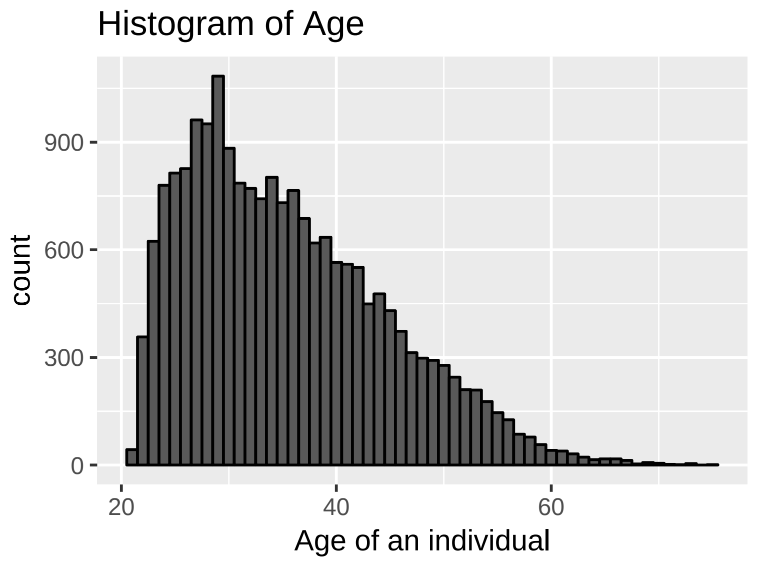
```r
ggplot(dat, aes(x = EDUCATION)) +
  geom_histogram(bins = 4, col = "black") +
  ggtitle("Histogram of Education") +
  xlab("Education of an individual")
```



```r
ggplot(dat, aes(x = MARRIAGE)) +
  geom_histogram(bins = 4, col = "black") +
  ggtitle("Histogram of Marriage") +
  xlab("Martial status of an individual; 0 = ???, 1 = married, \n 2 = single, 3 = high school, 4 = othe
```



```r
ggplot(dat, aes(x = AGE)) +
  geom_histogram(binwidth = 1, col = "black") +
  ggtitle("Histogram of Age") +
  xlab("Age of an individual")
```


Histogram of Age
count
Age of an individual