German Credit Risk Basic EDA

Code

Kailash Baskar - W0790883

Project Proposal

In this project we are going to explore the German credit risk data and predict which customers has the best possibility to return back the credit i.e a good loan and which customers has the least possibility to return back the credit.

Section 1 - Introduction:

German credit risk data is a very interesting data which is obtained from a very confidential German financial data source to predict and to catch the customers who has the lease probability of returning back a credit to the bank. In this data set, each entry indicates a bank customer who takes a credit from a bank. From analyzing the given attributes in the data each customer's loan is identified as whether it is good or bad loan.

The project problem/research statement is to analyse the available variables and identify which attribute is important in predicting the credit risk and which customer with the right attributes should be given a credit from the bank. Also to create a model which will help predict whether a customer with various attributes will repay the loan i.e have a less or no credit risk associated with him.

The data set which is used in this project is selected from Kaggle and it is available in this link link_to_data_set (https://www.kaggle.com/kabure/german-credit-data-with-risk).

Code

The data set has 1000 observations and 11 variables in it. Below are the available variables in the data set and its type.

- Age (numeric)
- Sex (text: male, female)
- Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
- Housing (text: own, rent, or free)
- Saving accounts (text - little, moderate, quite rich, rich)
- Checking account (numeric, in DM - Deutsch Mark)
- Credit amount (numeric, in DM)
- Duration (numeric, in month)
- Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others

Code

```
## [1] 1000   11
```

Code

```
## cols(
##   ...1 = col_double(),
##   Age = col_double(),
##   Sex = col_character(),
##   Job = col_double(),
##   Housing = col_character(),
##   `Saving accounts` = col_character(),
##   `Checking account` = col_character(),
##   `Credit amount` = col_double(),
##   Duration = col_double(),
##   Purpose = col_character(),
##   Risk = col_character()
## )
```

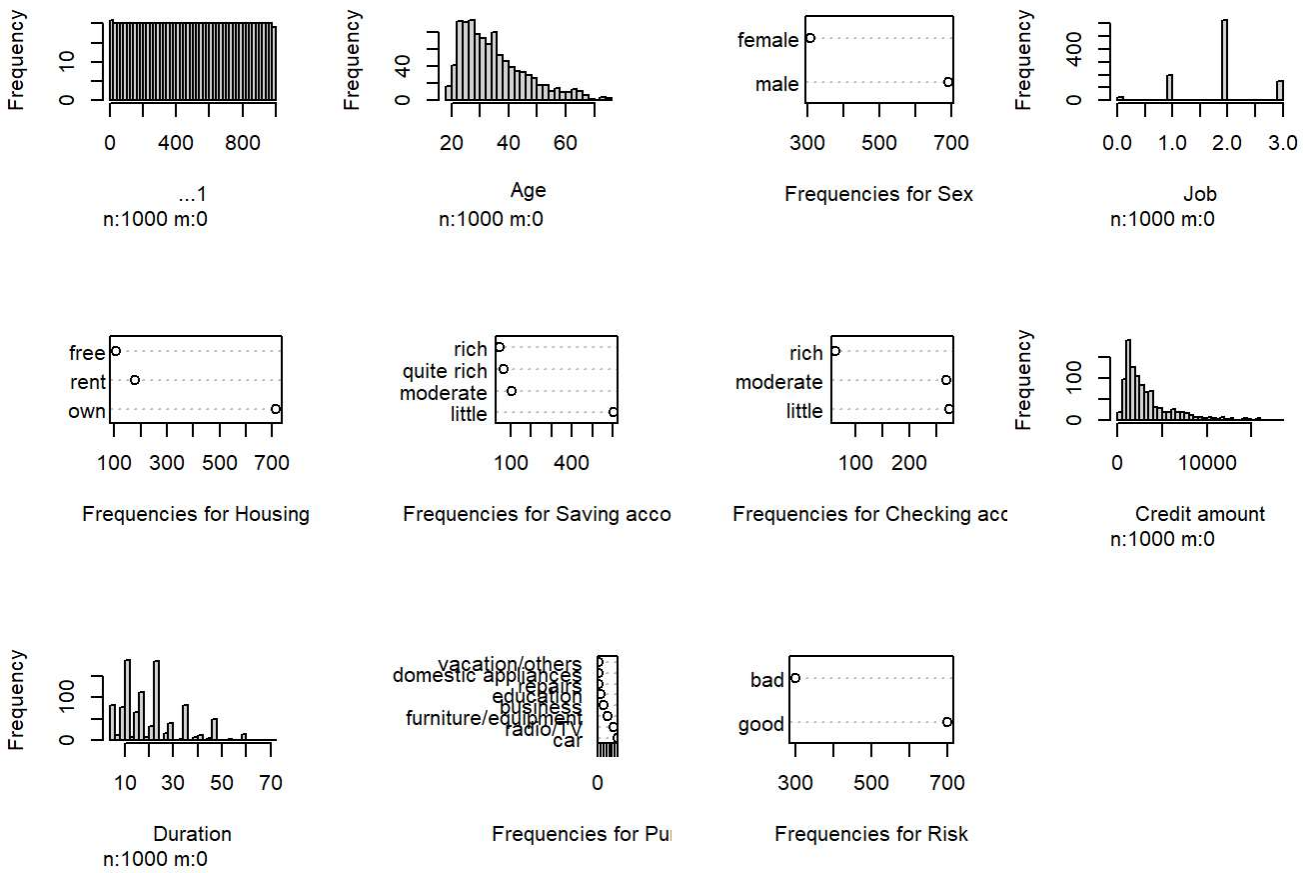
Section 2 - Data analysis plan:

- The main plan is to predict which and what customer attributes contribute in making a loan good or bad.
- The dependent variable “Risk” is to be compared with other predictor variables like Age, Duration and others for this project.
- The summary of this data set is mentioned below and also plotted a histogram graph to identify the distribution of the variables in the data set. The first variable is just the serial number and has to be removed, and there are some NA values in “Savings accounts” and “Checking account” variables will try to fill these observations with relevant values.

Code

```
##       ...1            Age            Sex                 Job        
##  Min.   :  0.0   Min.   :19.00   Length:1000        Min.   :0.000  
##  1st Qu.:249.8   1st Qu.:27.00   Class :character   1st Qu.:2.000  
##  Median :499.5   Median :33.00   Mode  :character   Median :2.000  
##  Mean   :499.5   Mean   :35.55                      Mean   :1.904  
##  3rd Qu.:749.2   3rd Qu.:42.00                      3rd Qu.:2.000  
##  Max.   :999.0   Max.   :75.00                      Max.   :3.000  
##    Housing          Saving accounts    Checking account   Credit amount  
##  Length:1000        Length:1000        Length:1000        Min.   :  250  
##  Class :character   Class :character   Class :character   1st Qu.: 1366  
##  Mode  :character   Mode  :character   Mode  :character   Median : 2320  
##                                                           Mean   : 3271  
##                                                           3rd Qu.: 3972  
##                                                           Max.   :18424  
##     Duration      Purpose             Risk          
##  Min.   : 4.0   Length:1000        Length:1000       
##  1st Qu.:12.0   Class :character   Class :character  
##  Median :18.0   Mode  :character   Mode  :character  
##  Mean   :20.9                                        
##  3rd Qu.:24.0                                        
##  Max.   :72.0
```

Code



Code

```
##            ...1             Age             Sex             Job
##               0               0               0               0
##         Housing  Saving accounts Checking account   Credit amount
##               0             183             394               0
##        Duration         Purpose            Risk
##               0               0               0
```

- The statistical methods used in this project are “IQR, correlation, lm, glm, summary, mean, and median”.
- Using the above mentioned statistical methods various statistical data are found from the variable in the dataset, also a couple of regression models are created with the available variables which will help in predicting if customer is a credit risky customer or not.

Section 3 - Data:

The data is placed in `/data` folder. Glimpse of the data

Code

```
## Rows: 1,000
## Columns: 11
## $ ...1               <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
## $ Age                <dbl> 67, 22, 49, 45, 53, 35, 53, 35, 61, 28, 25, 24, 22,~
## $ Sex                <chr> "male", "female", "male", "male", "male", "male", "~
## $ Job                <dbl> 2, 2, 1, 2, 2, 1, 2, 3, 1, 3, 2, 2, 2, 1, 2, 1, 2, ~
## $ Housing            <chr> "own", "own", "own", "free", "free", "free", "own",~
## $ `Saving accounts`  <chr> NA, "little", "little", "little", "little", NA, "qu~
## $ `Checking account` <chr> "little", "moderate", NA, "little", "little", NA, N~
## $ `Credit amount`    <dbl> 1169, 5951, 2096, 7882, 4870, 9055, 2835, 6948, 305~
## $ Duration           <dbl> 6, 48, 12, 42, 24, 36, 24, 36, 12, 30, 12, 48, 12, ~
## $ Purpose            <chr> "radio/TV", "radio/TV", "education", "furniture/equ~
## $ Risk               <chr> "good", "bad", "good", "good", "bad", "good", "good~
```

References:

https://www.tutorialspoint.com/how-to-create-histogram-of-all-columns-in-an-r-data-frame (https://www.tutorialspoint.com/how-to-create-histogram-of-all-columns-in-an-r-data-frame)

https://stackoverflow.com/questions/8317231/elegant-way-to-report-missing-values-in-a-data-frame (https://stackoverflow.com/questions/8317231/elegant-way-to-report-missing-values-in-a-data-frame)