Machine Learning Assignment Phase 1

Chi Ting Low (s3611774) & Vidya Viswanathan (s3613547)

3/8/2018

Contents

Introduction

The aim of the current project is to replicate the research study by Yeh and Lien(2009). The purpose of Yeh and Lien's (2009) research is to compare the predictive accuracy of probability of default using four different machine learning methods(K-nearest neighbor classifiers (Altman, 1992), Logistic regression (Freedman, 2009), Naive Bayes Classifier (Russell & Norvig, 2016) and Classification Trees (Kelleher, Mac Namee and D'Arcy, 2015). The current project aims to examine all these methods using the *mlr* package in R.

The dataset is acquired from the UCI Machine Learning Repository. There are two phases in the current project. Phase I will focus on data cleaning, data exploration and data visualization. Phase II will focus on model building.

Data preprocessing

```r
library(readxl)     #reading excel files
library(tidyverse) #data manipulation packages
library(mlr)        #machine learning packages
library(psych)      #psych package for descriptive analysis
library(plyr)       #data manipulation packages
library(ggplot2)    #plotting
library(gridExtra) #arrange the plot
library(corrplot)  #correlation plot for the data

#loading data and looking at the structure of the data
default_crd <- read_excel("Default_risk/default of credit card clients.xls", skip = 1)
str(default_crd)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    30000 obs. of  25 variables:
##  $ ID                         : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ LIMIT_BAL                  : num  20000 120000 90000 50000 50000 50000 500000 100000 1400
##  $ SEX                        : num  2 2 2 2 1 1 1 2 2 1 ...
##  $ EDUCATION                  : num  2 2 2 2 2 1 1 2 3 3 ...
##  $ MARRIAGE                   : num  1 2 2 1 1 2 2 2 1 2 ...
##  $ AGE                        : num  24 26 34 37 57 37 29 23 28 35 ...
##  $ PAY_0                      : num  2 -1 0 0 -1 0 0 0 0 -2 ...
##  $ PAY_2                      : num  2 2 0 0 0 0 0 -1 0 -2 ...
##  $ PAY_3                      : num  -1 0 0 0 -1 0 0 -1 2 -2 ...
##  $ PAY_4                      : num  -1 0 0 0 0 0 0 0 0 -2 ...
##  $ PAY_5                      : num  -2 0 0 0 0 0 0 0 0 -1 ...
##  $ PAY_6                      : num  -2 2 0 0 0 0 0 -1 0 -1 ...
##  $ BILL_AMT1                  : num  3913 2682 29239 46990 8617 ...
##  $ BILL_AMT2                  : num  3102 1725 14027 48233 5670 ...
##  $ BILL_AMT3                  : num  689 2682 13559 49291 35835 ...
##  $ BILL_AMT4                  : num  0 3272 14331 28314 20940 ...
##  $ BILL_AMT5                  : num  0 3455 14948 28959 19146 ...
```

```
##  $ BILL_AMT6                 : num  0 3261 15549 29547 19131 ...
##  $ PAY_AMT1                  : num  0 0 1518 2000 2000 ...
##  $ PAY_AMT2                  : num  689 1000 1500 2019 36681 ...
##  $ PAY_AMT3                  : num  0 1000 1000 1200 10000 657 38000 0 432 0 ...
##  $ PAY_AMT4                  : num  0 1000 1000 1100 9000 ...
##  $ PAY_AMT5                  : num  0 0 1000 1069 689 ...
##  $ PAY_AMT6                  : num  0 2000 5000 1000 679 ...
##  $ default payment next month: num  1 1 0 0 0 0 0 0 0 0 ...
```

Variable Information

- ID: Identification information for each record
- Limit_balance : The amount of given credit includes both individuals and his/her family credit (*numeric*)
- Sex: 1 = male, 2 = Female (*categorical*)
- Education: 1 = graduate school, 2 = university, 3 = high school, 4 = others (*categorical*)
- Marital status: 1 = married, 2 = single, 3 = others (*categorical*)
- Age = age in year (*numeric*)
- History of past payments (Sept_repay - Apr_repay): These are the tracked past monthly payment records from April to September. The measurment scale for the payment status is: (-1 = pay duly, 1 = payment delay of one month, 2 = payment delay of two months … 9 = payment delay of nine months and above) (*categorical*)
- Number of bill statement(Sept_statement - Apr_statement): These variables represent the number of bill statement from April to September in 2005. (*numeric*)
- Amount of previous payment(Sept_amtpay - Apr_amtpay): These variables represent the amount of the prvious payments paid from April to September in 2005.(*numeric*)
- Default: binary variable represent default payment (1 = Yes, 0 = No)(*categorical*)

```r
#rename columns and recode the variables
colnames(default_crd) = c("ID", "Limit_balance", "Sex", "Education", "Marriage",
"Age", "Sept_repay", "Aug_repay","July_repay","Jun_repay","May_repay","Apr_repay",
"Sept_statement","Aug_statement","July_statement","Jun_statement","May_statement",
"Apr_statement","Sept_amtpay","Aug_amtpay","July_amtpay","Jun_amtpay","May_amtpay",
"Apr_amtpay","default")

default_nullID <- default_crd
default_nullID$ID <- NULL

#create copy of the data and remove id columns
recode_default <- default_crd[,-1]

#recode sex
recode_default$Sex[recode_default$Sex == 1] <- "Male"
recode_default$Sex[recode_default$Sex == 2] <- "Female"

#recode marriage
recode_default$Marriage[recode_default$Marriage == 1] <- "Married"
```

```r
recode_default$Marriage[recode_default$Marriage == 2] <- "Single"
recode_default$Marriage[recode_default$Marriage == 3] <- "others"

#recode education
recode_default$Education[recode_default$Education == 1] <- "Graduate_school"
recode_default$Education[recode_default$Education == 2] <- "University"
recode_default$Education[recode_default$Education == 3] <- "High_school"
recode_default$Education[recode_default$Education == 4] <- "Others"

#recode default
recode_default$default[recode_default$default == 1] <- "Yes"
recode_default$default[recode_default$default == 0] <- "No"
```

```r
#looking at data dimension and basic descriptive stats of variables
dim(recode_default)
```

```
## [1] 30000    24
```

```r
count(recode_default, 'Sex')
```

```
##      Sex  freq
## 1 Female 18112
## 2   Male 11888
```

```r
count(recode_default, 'Education')
```

```
##         Education  freq
## 1               0    14
## 2               5   280
## 3               6    51
## 4 Graduate_school 10585
## 5     High_school  4917
## 6          Others   123
## 7      University 14030
```

```r
count(recode_default, 'Marriage')
```

```
##   Marriage  freq
## 1        0    54
## 2  Married 13659
## 3   others   323
## 4   Single 15964
```

```r
count(recode_default,'default')
```

```
##   default  freq
## 1      No 23364
## 2     Yes  6636
```

```r
min(recode_default$Age); max(recode_default$Age)
```

```
## [1] 21
```

```
## [1] 79
```

```r
mean(recode_default$Age)
```

```
## [1] 35.4855
```

Prior building the models, the dataset is loaded for preprocessing. 30000 observations and 25 variables were observed in the dataset. To gain a better understanding of the data, variables(Sex, Marriage, Education, Default and the history of past payment) are recorded from numeric to categorical, as in the variable information, which represent the best for the variables. In addition, the values which are outside of the context from the variable information such as 0, 5, 6 in Education are removed. These errors may occur due to human error when inputting the data.

The aim of this dataset is to build a model using exisiting variables to predict whether a person will default on their credit payment. The target variable is identified as default and presented below:

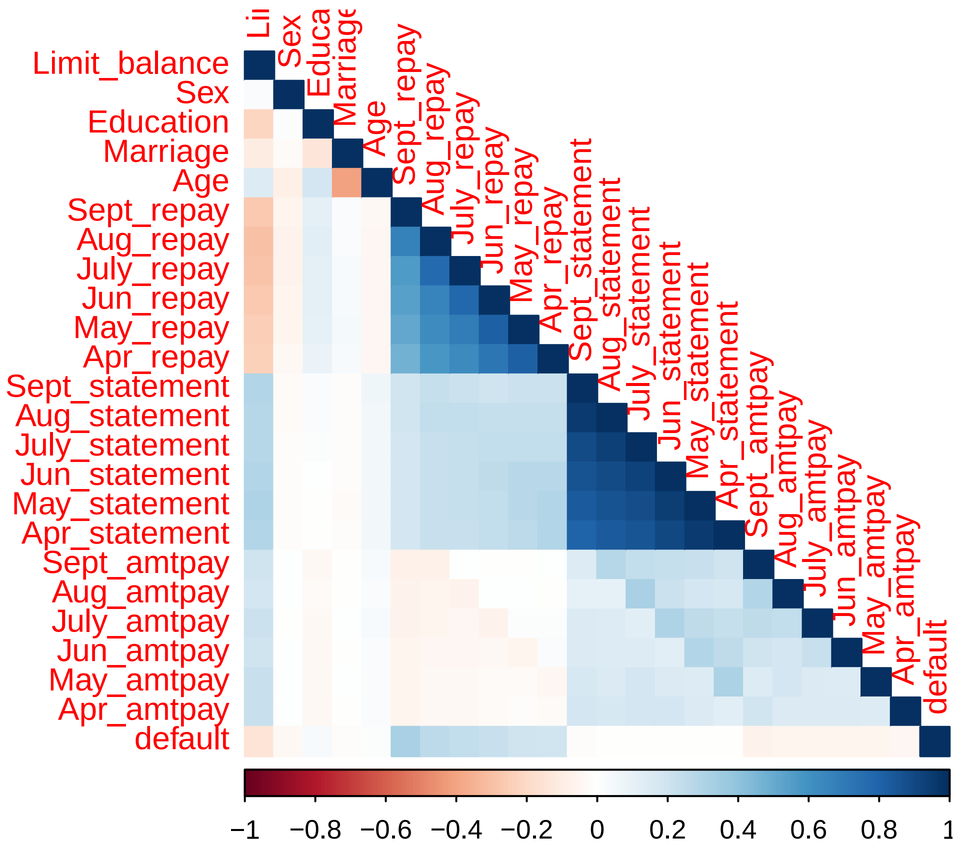
$$default = \Big\{ 1 = Yes, 0 = No \tag{1}$$

The ID variable is removed because it does not provide any meaningful measurement of the dataset. It was observed that a variable's name does not provide any meaningful meaning to the dataset. Therefore, variables are renamed to provide a better understanding of the dataset.

From the observation, it was observed that there were 30000 observations and 25 variables which included ID, Limited balance, Sex, Education, Marital status, Age, History of payments from April to September, Number of bill statements from April to September and a binary code from 0 (Not default) and 1 (Default payment) to determine whether a person is defaulting on their payment. Based on the observation, there are 18112 females and 11888 males in the current dataset. In addition, the observation also shows that there were 6636 observations defaulting on their payment. The minimum and maximum ages are 21 and 79 with mean of 35.4855.

Furthermore, it is shown that University(n = 14030) and Graduate school(n = 10585) are the highest education level achieved followed by high school (n = 4917) and other education(n = 123). Majority of the individuals are single (n = 15964) or married (n = 13659) followed by other relationship status(n = 323). The dataset also shown that majority of the individual are not likely to default (n = 23364) their credit.

```r
#correlation of the data
correlation = round(cor(default_nullID),2)
corrplot(correlation, type = "lower", method = 'color')
```



To examine the relationship between the variables, a correlation test using the Pearson product moment is performed. It was observed that the correlation between the variables was around -0.4 to 0.8. The observation shows that the correlation between variables is between medium and high correlation (Field, Miles & Field, 2012).

From the dataset and the papers, the variables of history of repayment show that there is no meaningful indicator for the values 0 and -2. Therefore, these values have been removed. In addition,the values in these variables are recoded to provide a better understanding of the data. After removing those values, there are 4030 observations left in the dataset.

According to Indira, Vasanthakumari and Sugumaran (2010) and Figueroa, Zeng-Treitler, Kandula and Ngo (2012), research on predicting the sample size required for classification performance shows that if the sample size is above 200, the performance of the mean absolute error and root mean squared error will be lower. This suggested that the performance of the machine learning method will have reliable and valid accuracy results compared to a dataset below 200.

Exploratory Data Analysis

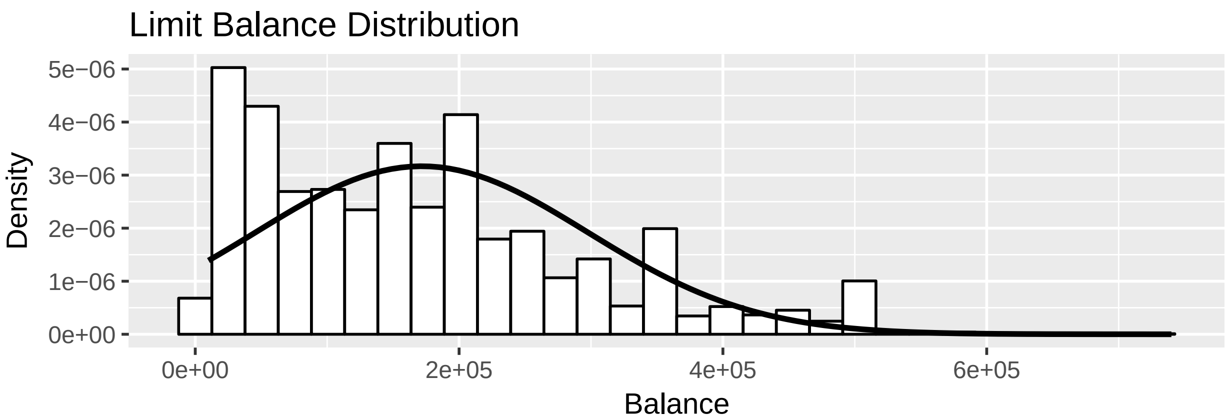
Assumption of Normality

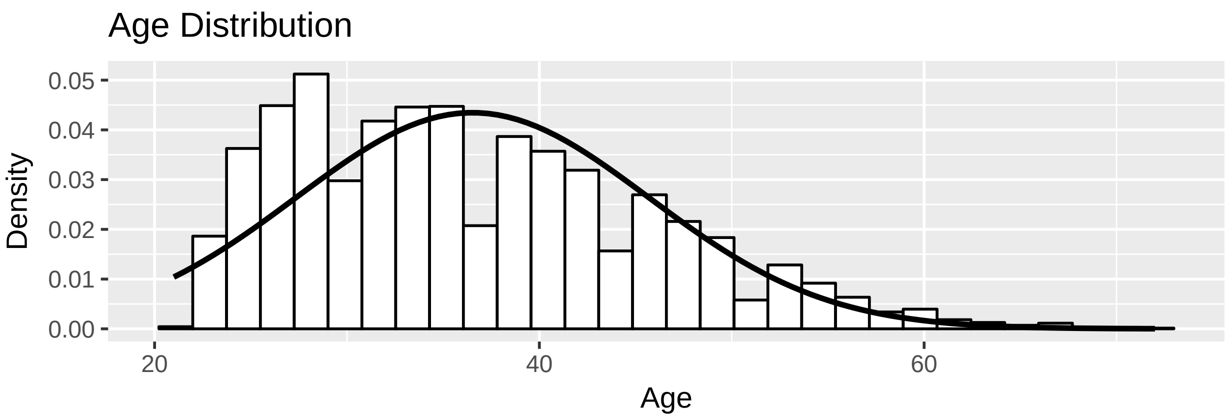
```
#descriptive statistics of the dataset
describe(recode_default)
```

```
##                 vars    n      mean        sd   median  trimmed       mad
## Limit_balance      1 4030 171657.57 125943.77 150000.0 157602.36 133434.00
## Sex*               2 4030      1.41      0.49      1.0      1.39      0.00
## Education*         3 4030      2.43      1.39      2.0      2.42      1.48
## Marriage*          4 4030      1.96      1.00      1.0      1.95      0.00
## Age                5 4030     36.52      9.18     35.0     35.84     10.38
## Sept_repay*        6 4030      8.06      3.58     11.0      8.33      0.00
## Aug_repay*         7 4030      8.35      3.33     11.0      8.57      0.00
## July_repay*        8 4030      8.35      3.32     11.0      8.56      0.00
## Jun_repay*         9 4030      8.53      3.26     11.0      8.79      0.00
## May_repay*        10 4030      7.61      3.23     10.0      7.89      0.00
## Apr_repay*        11 4030      7.53      3.27     10.0      7.78      0.00
## Sept_statement    12 4030  22090.76  44288.46   4405.0  11437.64   6047.53
## Aug_statement     13 4030  22256.29  44328.66   4414.0  11641.83   6075.69
## July_statement    14 4030  22321.21  44478.71   4238.5  11556.85   5815.50
## Jun_statement     15 4030  22618.34  44947.09   4180.0  11766.21   5765.83
## May_statement     16 4030  22590.35  44523.32   4087.0  11822.63   5614.61
## Apr_statement     17 4030  22701.87  45530.50   4162.0  11837.04   5874.06
## Sept_amtpay       18 4030   4654.06  10882.90   1600.0   2390.43   2372.16
## Aug_amtpay        19 4030   4609.70  11979.73   1595.0   2387.53   2364.75
## July_amtpay       20 4030   4719.07  13410.75   1443.0   2287.89   2139.39
## Jun_amtpay        21 4030   4547.38  11093.97   1443.5   2274.48   2140.13
## May_amtpay        22 4030   4605.92  13538.10   1228.0   2112.73   1820.63
## Apr_amtpay        23 4030   4590.07  14952.88   1048.0   1949.13   1553.76
## default*          24 4030      1.36      0.48      1.0      1.32      0.00
##                     min    max  range  skew kurtosis      se
## Limit_balance     10000 740000 730000  0.88     0.26 1983.92
## Sex*                  1      2      1  0.37    -1.86    0.01
## Education*            1      4      3  0.14    -1.84    0.02
## Marriage*             1      3      2  0.08    -1.99    0.02
## Age                  21     72     51  0.62    -0.15    0.14
## Sept_repay*           3     11      8 -0.44    -1.75    0.06
## Aug_repay*            3     11      8 -0.48    -1.74    0.05
## July_repay*           3     11      8 -0.49    -1.73    0.05
## Jun_repay*            3     11      8 -0.60    -1.59    0.05
## May_repay*            3     10      7 -0.65    -1.53    0.05
## Apr_repay*            3     10      7 -0.60    -1.61    0.05
## Sept_statement    -4316 581775 586091  4.16    25.03  697.65
## Aug_statement    -24704 572677 597381  4.13    24.71  698.28
## July_statement   -61506 471175 532681  3.91    20.49  700.65
## Jun_statement     -3903 486776 490679  3.90    20.56  708.03
## May_statement     -3876 503914 507790  3.83    19.86  701.35
## Apr_statement   -339603 527711 867314  3.59    19.82  717.22
## Sept_amtpay           0 187206 187206  6.92    71.12  171.43
## Aug_amtpay            0 302961 302961 10.66   179.28  188.71
## July_amtpay           0 417588 417588 12.47   274.22  211.25
## Jun_amtpay            0 193712 193712  7.37    81.26  174.76
## May_amtpay            0 303512 303512 11.30   194.37  213.26
```

```
## Apr_amtpay           0 345293 345293 10.58    155.16   235.54
## default*             1      2      1  0.60     -1.64     0.01
```

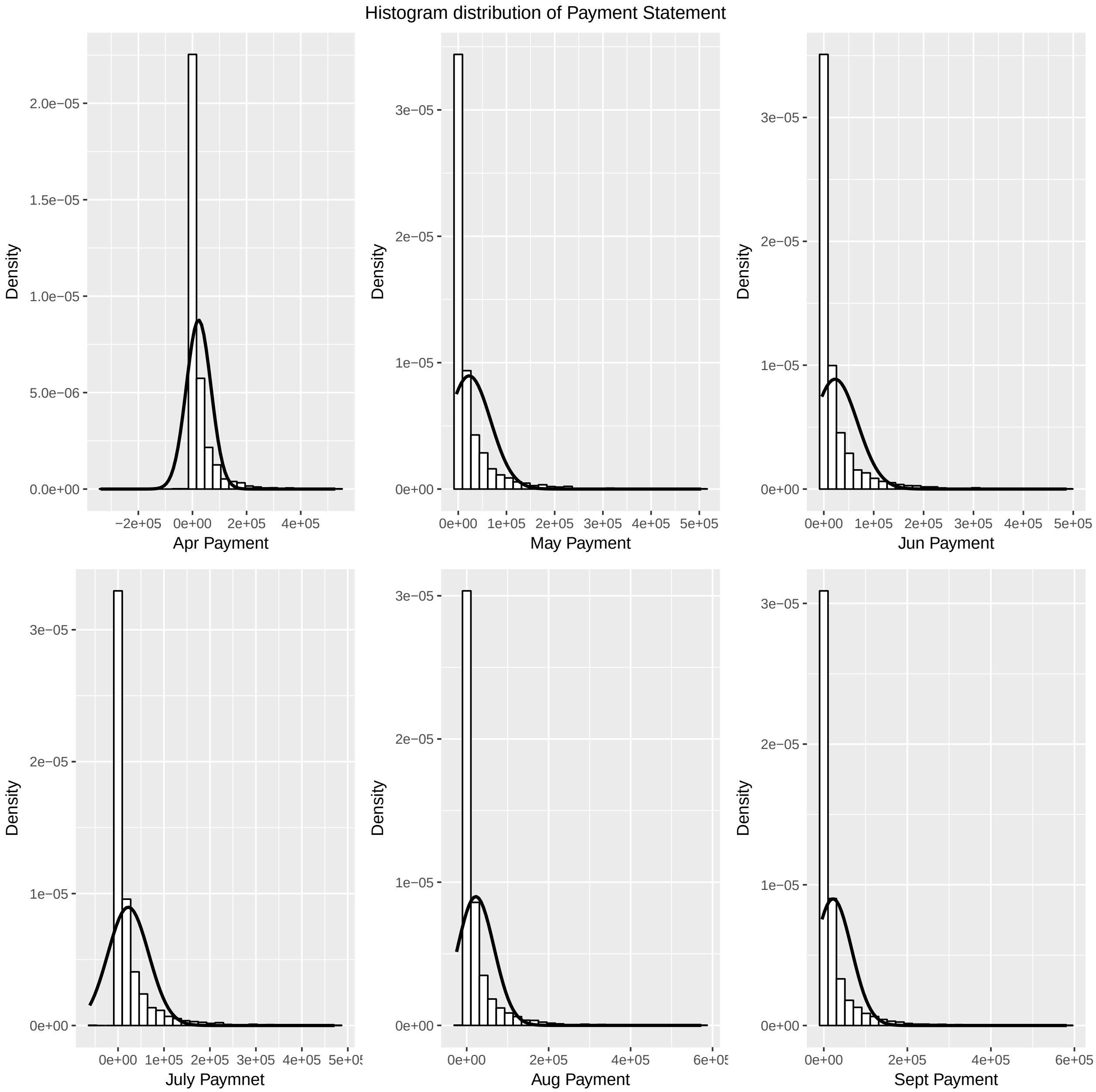
The *describe* function shows that the skew and kurtosis of the variables do not deviate further away from 0, which suggested that the data is normal or close to normal distribution.





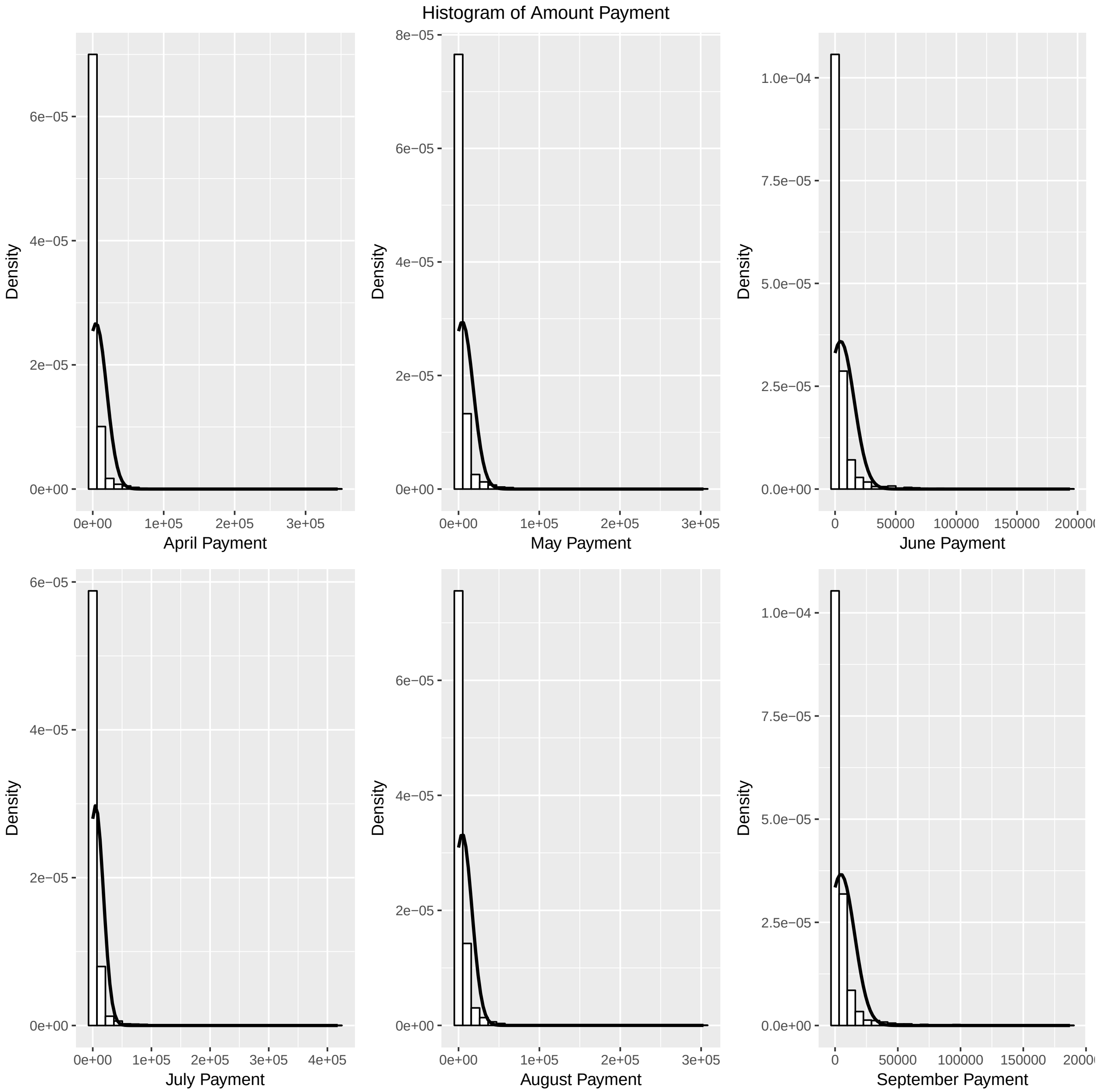

```
sept_statement <- ggplot(recode_default, aes(Sept_statement)) + theme(legend.position = "none"
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "Sept Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Sept_statement), sd = sd(re

aug_statement <- ggplot(recode_default, aes(Aug_statement)) + theme(legend.position = "none")
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "Aug Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Aug_statement), sd = sd(rec

july_statement <- ggplot(recode_default, aes(July_statement)) + theme(legend.position = "none"
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "July Paymnet", y = "Density") +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$July_statement), sd = sd(re

jun_statement <- ggplot(recode_default, aes(Jun_statement)) +
  theme(legend.position = "none") + geom_histogram(aes(y = ..density..), colour = "black", fil
  labs(x = "Jun Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Jun_statement), sd = sd(rec
```

```
may_statement <- ggplot(recode_default, aes(May_statement)) + theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "May Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$May_statement), sd = sd(re

apr_statement <- ggplot(recode_default, aes(Apr_statement)) +
  theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "Apr Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Apr_statement), sd = sd(re

grid.arrange(apr_statement, may_statement, jun_statement,july_statement,aug_statement,sept_sta
  layout_matrix = rbind(c(1, 2, 3),
                        c(4, 5, 6)),
  top = 'Histogram distribution of Payment Statement')
```




```
sept_amtpay <- ggplot(recode_default, aes(Sept_amtpay)) + theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "September Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Sept_amtpay), sd = sd(recod

aug_amtpay <- ggplot(recode_default, aes(Aug_amtpay)) + theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "August Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Aug_amtpay), sd = sd(recode

july_amtpay <- ggplot(recode_default, aes(July_amtpay)) + theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
```

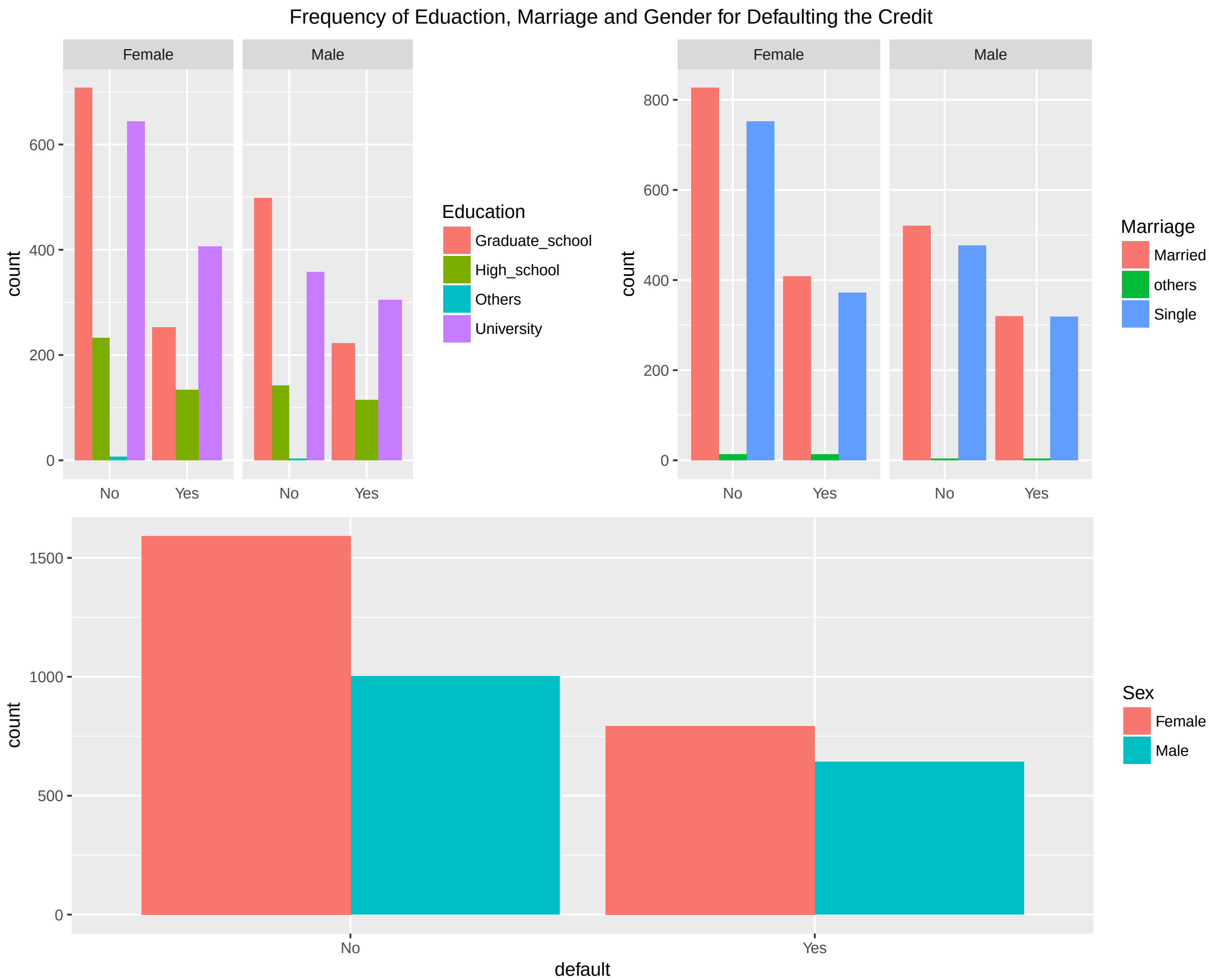
```
  labs(x = "July Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$July_amtpay), sd = sd(recod

jun_amtpay <- ggplot(recode_default, aes(Jun_amtpay)) + theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "June Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Jun_amtpay), sd = sd(recode

may_amtpay <- ggplot(recode_default, aes(May_amtpay)) + theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "May Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$May_amtpay), sd = sd(recode

apr_amtpay <- ggplot(recode_default, aes(Apr_amtpay)) + theme(legend.position = "none") +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(x = "April Payment", y = "Density")  +
  stat_function(fun = dnorm, args = list(mean = mean(recode_default$Apr_amtpay), sd = sd(recode

grid.arrange(apr_amtpay, may_amtpay, jun_amtpay, july_amtpay, aug_amtpay, sept_amtpay,
  layout_matrix = rbind(c(1, 2, 3),
                        c(4, 5, 6)), top = "Histogram of Amount Payment")
```



Based on the figures above, it is shown that the limited balance and individual's age show a positive skew pattern. Overall, the variables are normally distributed. This can also be observed for other variables (April to September Statement and amount paid). These suggest that the data falls within the normal distribution.

```
gender <- ggplot(recode_default, aes(default, ..count..)) +
  geom_bar(aes(fill = Sex), position = 'dodge')

education <- ggplot(recode_default, aes(default, ..count..)) +
  geom_bar(aes(fill = Education), position = 'dodge') +
  facet_wrap(~ Sex)  + theme(axis.title.x = element_blank(), axis.ticks.x = element_blank())

marriage <- ggplot(recode_default, aes(default, ..count..)) +
```

```r
  geom_bar(aes(fill = Marriage), position = 'dodge') +
  facet_wrap(~ Sex) + theme(axis.title.x = element_blank(), axis.ticks.x = element_blank())

grid.arrange(education,marriage,gender,
  layout_matrix = rbind(c(1, 1, 2, 2),
                        c(3, 3, 3, 3)),
  top = 'Frequency of Eduaction, Marriage and Gender for Defaulting the Credit')
```



Basic exploratory data analysis shows that females with a higher education level or that are married are the most likely not to default their credit compared to males. Overall observation shows that females are more likely not to default their credit.

Summary

The final data checking shows that all the variables in the dataset perform well. There are no missing values, outliners or extreme values, and the values follow a normal distribution.

References

Altman, N. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175. http://dx.doi.org/10.2307/2685209

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid"Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G and Jones Z (2016). "mlr: Machine Learning in R."*Journal of Machine Learning Research*, *17*(170), pp. 1-5. <URL: http://jmlr.org/papers/v17/15-066.html>.

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. London: Sage.

Figueroa, R., Zeng-Treitler, Q., Kandula, S., & Ngo, L. (2012). Predicting sample size required for classification performance. BMC Medical Informatics And Decision Making, 12(1). http://dx.doi.org/10.1186/1472-6947-12-8

Freedman, D. A. (2009). Statistical models: theory and practice. cambridge university press.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Hadley Wickham and Jennifer Bryan (2017). readxl: Read Excel Files. R package version 1.0.0. https://CRAN.R-project.org/package=readxl

Indira, V., Vasanthakumari, R., & Sugumaran, V. (2010). Minimum sample size determination of vibration signals in machine learning approach to fault diagnosis using power analysis. Expert Systems With Applications, 37(12), 8650-8658. http://dx.doi.org/10.1016/j.eswa.2010.06.068

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

Revelle, W. (2017) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.7.8.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,.

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Yeh, I., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems With Applications, 36(2), 2473-2480. http://dx.doi.org/10.1016/j.eswa.2007.12.020