2103 Project

2022-11-13

```
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
data <- read.csv("card.csv", sep = ",", skip = 2, header = FALSE)
header <- scan("card.csv",sep=",",nlines=2,what=character())
head(data)
```

```
##   V1     V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12   V13   V14   V15   V16   V17
## 1  1  20000  2  2  1 24  2  2 -1  -1  -2  -2  3913  3102   689     0     0
## 2  2 120000  2  2  2 26 -1  2  0   0   0   2  2682  1725  2682  3272  3455
## 3  3  90000  2  2  2 34  0  0  0   0   0   0 29239 14027 13559 14331 14948
## 4  4  50000  2  2  1 37  0  0  0   0   0   0 46990 48233 49291 28314 28959
## 5  5  50000  1  2  1 57 -1  0 -1   0   0   0  8617  5670 35835 20940 19146
## 6  6  50000  1  1  2 37  0  0  0   0   0   0 64400 57069 57608 19394 19619
##     V18  V19   V20   V21  V22  V23  V24 V25
## 1     0    0   689     0    0    0    0   1
## 2  3261    0  1000  1000 1000    0 2000   1
## 3 15549 1518  1500  1000 1000 1000 5000   0
## 4 29547 2000  2019  1200 1100 1069 1000   0
## 5 19131 2000 36681 10000 9000  689  679   0
## 6 20024 2500  1815   657 1000 1000  800   0
```

```
head(data)
```

```
##   V1     V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12   V13   V14   V15   V16   V17
## 1  1  20000  2  2  1 24  2  2 -1  -1  -2  -2  3913  3102   689     0     0
## 2  2 120000  2  2  2 26 -1  2  0   0   0   2  2682  1725  2682  3272  3455
## 3  3  90000  2  2  2 34  0  0  0   0   0   0 29239 14027 13559 14331 14948
## 4  4  50000  2  2  1 37  0  0  0   0   0   0 46990 48233 49291 28314 28959
## 5  5  50000  1  2  1 57 -1  0 -1   0   0   0  8617  5670 35835 20940 19146
## 6  6  50000  1  1  2 37  0  0  0   0   0   0 64400 57069 57608 19394 19619
##     V18  V19   V20   V21  V22  V23  V24 V25
## 1     0    0   689     0    0    0    0   1
## 2  3261    0  1000  1000 1000    0 2000   1
## 3 15549 1518  1500  1000 1000 1000 5000   0
## 4 29547 2000  2019  1200 1100 1069 1000   0
## 5 19131 2000 36681 10000 9000  689  679   0
## 6 20024 2500  1815   657 1000 1000  800   0
```

```
glimpse(data)
```

```
## Rows: 30,000
## Columns: 25
## $ V1  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,~
## $ V2  <int> 20000, 120000, 90000, 50000, 50000, 50000, 500000, 100000, 140000,~
## $ V3  <int> 2, 2, 2, 2, 1, 1, 1, 2, 2, 1, 2, 2, 2, 1, 1, 2, 1, 1, 2, 2, 2, 2, ~
## $ V4  <int> 2, 2, 2, 2, 2, 1, 1, 2, 3, 3, 3, 1, 2, 2, 1, 3, 1, 1, 1, 1, 3, 2, ~
## $ V5  <int> 1, 2, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 3, 2, 1, 1, 2, 2, 1, ~
## $ V6  <int> 24, 26, 34, 37, 57, 37, 29, 23, 28, 35, 34, 51, 41, 30, 29, 23, 24~
## $ V7  <int> 2, -1, 0, 0, -1, 0, 0, 0, 0, -2, 0, -1, -1, 1, 0, 1, 0, 0, 1, 1, 0~
## $ V8  <int> 2, 2, 0, 0, 0, 0, 0, -1, 0, -2, 0, -1, 0, 2, 0, 2, 0, 0, -2, -2, 0~
## $ V9  <int> -1, 0, 0, 0, -1, 0, 0, -1, 2, -2, 2, -1, -1, 2, 0, 0, 2, 0, -2, -2~
## $ V10 <int> -1, 0, 0, 0, 0, 0, 0, 0, 0, -2, 0, -1, -1, 0, 0, 0, 2, -1, -2, -2,~
## $ V11 <int> -2, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, -1, -1, 0, 0, 0, 2, -1, -2, -2,~
## $ V12 <int> -2, 2, 0, 0, 0, 0, 0, -1, 0, -1, -1, 2, -1, 2, 0, 0, 2, -1, -2, -2~
## $ V13 <int> 3913, 2682, 29239, 46990, 8617, 64400, 367965, 11876, 11285, 0, 11~
## $ V14 <int> 3102, 1725, 14027, 48233, 5670, 57069, 412023, 380, 14096, 0, 9787~
## $ V15 <int> 689, 2682, 13559, 49291, 35835, 57608, 445007, 601, 12108, 0, 5535~
## $ V16 <int> 0, 3272, 14331, 28314, 20940, 19394, 542653, 221, 12211, 0, 2513, ~
## $ V17 <int> 0, 3455, 14948, 28959, 19146, 19619, 483003, -159, 11793, 13007, 1~
## $ V18 <int> 0, 3261, 15549, 29547, 19131, 20024, 473944, 567, 3719, 13912, 373~
## $ V19 <int> 0, 0, 1518, 2000, 2000, 2500, 55000, 380, 3329, 0, 2306, 21818, 10~
## $ V20 <int> 689, 1000, 1500, 2019, 36681, 1815, 40000, 601, 0, 0, 12, 9966, 65~
## $ V21 <int> 0, 1000, 1000, 1200, 10000, 657, 38000, 0, 432, 0, 50, 8583, 6500,~
## $ V22 <int> 0, 1000, 1000, 1100, 9000, 1000, 20239, 581, 1000, 13007, 300, 223~
## $ V23 <int> 0, 0, 1000, 1069, 689, 1000, 13750, 1687, 1000, 1122, 3738, 0, 287~
## $ V24 <int> 0, 2000, 5000, 1000, 679, 800, 13770, 1542, 1000, 0, 66, 3640, 0, ~
## $ V25 <int> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, ~
```

```
str(data)
```

```
## ’data.frame':    30000 obs. of  25 variables:
##  $ V1 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ V2 : int  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
##  $ V3 : int  2 2 2 2 1 1 1 2 2 1 ...
##  $ V4 : int  2 2 2 2 2 1 1 2 3 3 ...
##  $ V5 : int  1 2 2 1 1 2 2 2 1 2 ...
##  $ V6 : int  24 26 34 37 57 37 29 23 28 35 ...
##  $ V7 : int  2 -1 0 0 -1 0 0 0 0 -2 ...
##  $ V8 : int  2 2 0 0 0 0 0 -1 0 -2 ...
##  $ V9 : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
##  $ V10: int  -1 0 0 0 0 0 0 0 0 -2 ...
##  $ V11: int  -2 0 0 0 0 0 0 0 0 -1 ...
##  $ V12: int  -2 2 0 0 0 0 0 -1 0 -1 ...
##  $ V13: int  3913 2682 29239 46990 8617 64400 367965 11876 11285 0 ...
##  $ V14: int  3102 1725 14027 48233 5670 57069 412023 380 14096 0 ...
##  $ V15: int  689 2682 13559 49291 35835 57608 445007 601 12108 0 ...
##  $ V16: int  0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
##  $ V17: int  0 3455 14948 28959 19146 19619 483003 -159 11793 13007 ...
##  $ V18: int  0 3261 15549 29547 19131 20024 473944 567 3719 13912 ...
##  $ V19: int  0 0 1518 2000 2000 2500 55000 380 3329 0 ...
##  $ V20: int  689 1000 1500 2019 36681 1815 40000 601 0 0 ...
##  $ V21: int  0 1000 1000 1200 10000 657 38000 0 432 0 ...
##  $ V22: int  0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
##  $ V23: int  0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
##  $ V24: int  0 2000 5000 1000 679 800 13770 1542 1000 0 ...
##  $ V25: int  1 1 0 0 0 0 0 0 0 0 ...
```

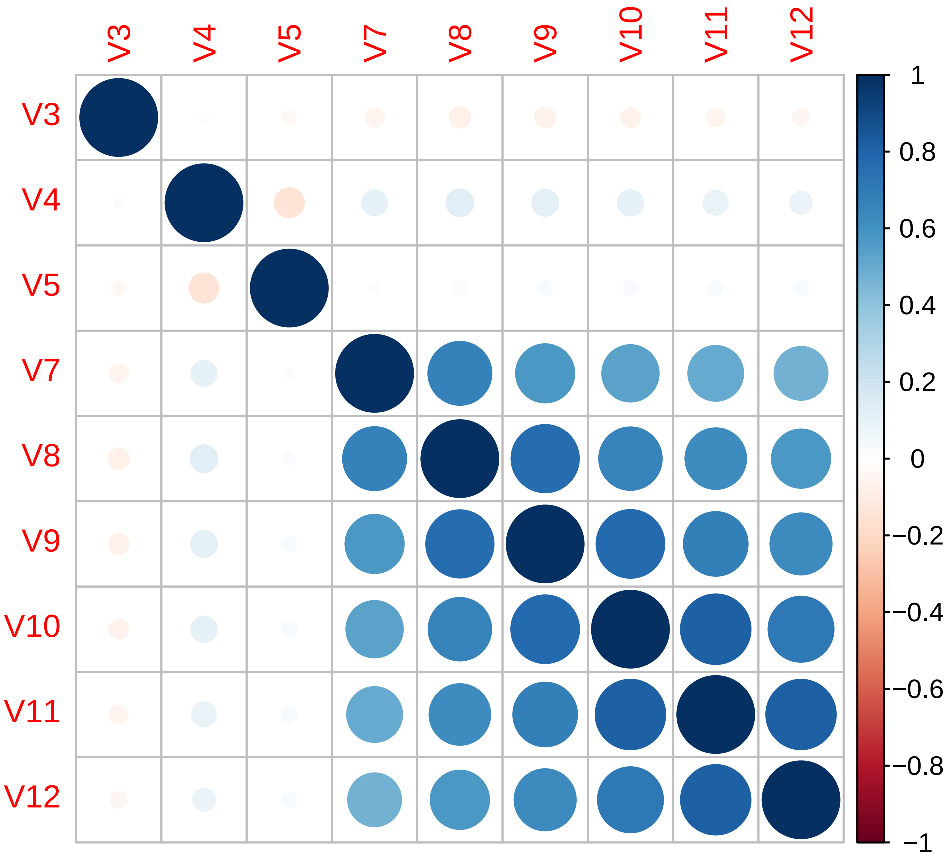
```
#Checking for NA values
any(is.na(data))
```

```
## [1] FALSE
```

```
summary(data)
```

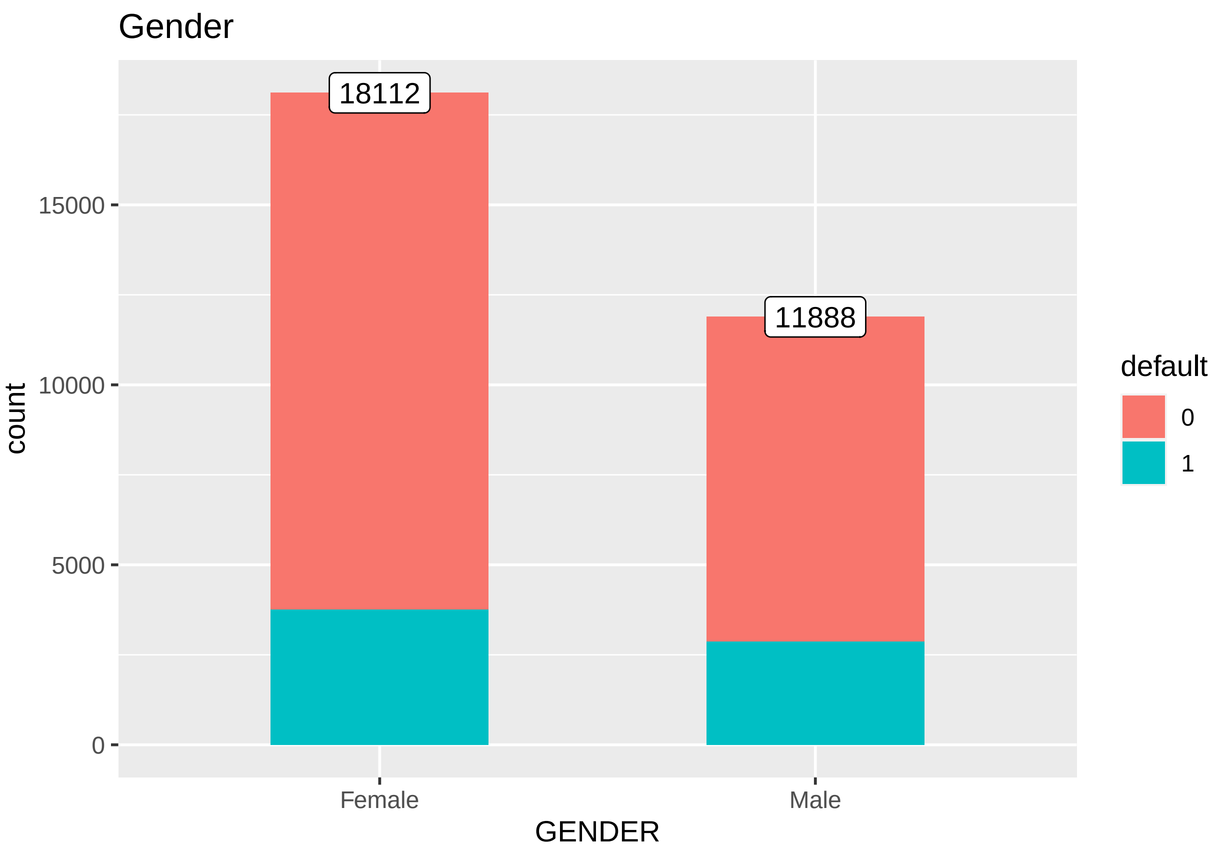
```
##        V1             V2                V3              V4
##  Min.   :    1   Min.   : 10000   Min.   :1.000   Min.   :0.000
##  1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000
##  Median :15000   Median : 140000   Median :2.000   Median :2.000
##  Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853
##  3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000
##        V5              V6              V7                V8
##  Min.   :0.000   Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000
##  1st Qu.:1.000   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000
##  Median :2.000   Median :34.00   Median : 0.0000   Median : 0.0000
##  Mean   :1.552   Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338
##  3rd Qu.:2.000   3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :3.000   Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000
##        V9               V10               V11               V12
##  Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000
##  1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000
##       V13               V14              V15               V16
##  Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000
##  1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327
##  Median :  22382   Median : 21200   Median :  20089   Median :  19052
##  Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263
##  3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506
##  Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586
##       V17              V18               V19              V20
##  Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :      0
##  1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:    833
##  Median : 18105   Median :  17071   Median :  2100   Median :   2009
##  Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :   5921
##  3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:   5000
##  Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259
##       V21              V22              V23                V24
##  Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0.0
##  1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8
##  Median :  1800   Median :  1500   Median :  1500.0   Median :  1500.0
##  Mean   :  5226   Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5
##  3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0
##  Max.   :896040   Max.   :621000   Max.   :426529.0   Max.   :528666.0
##       V25
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2212
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

```
#Correlation matrix
data_onlycat <- subset(data, select = c(c(V3,V4,V5,V7,V8,V9,V10,V11,V12)))
corrplot(cor(data_onlycat))
```

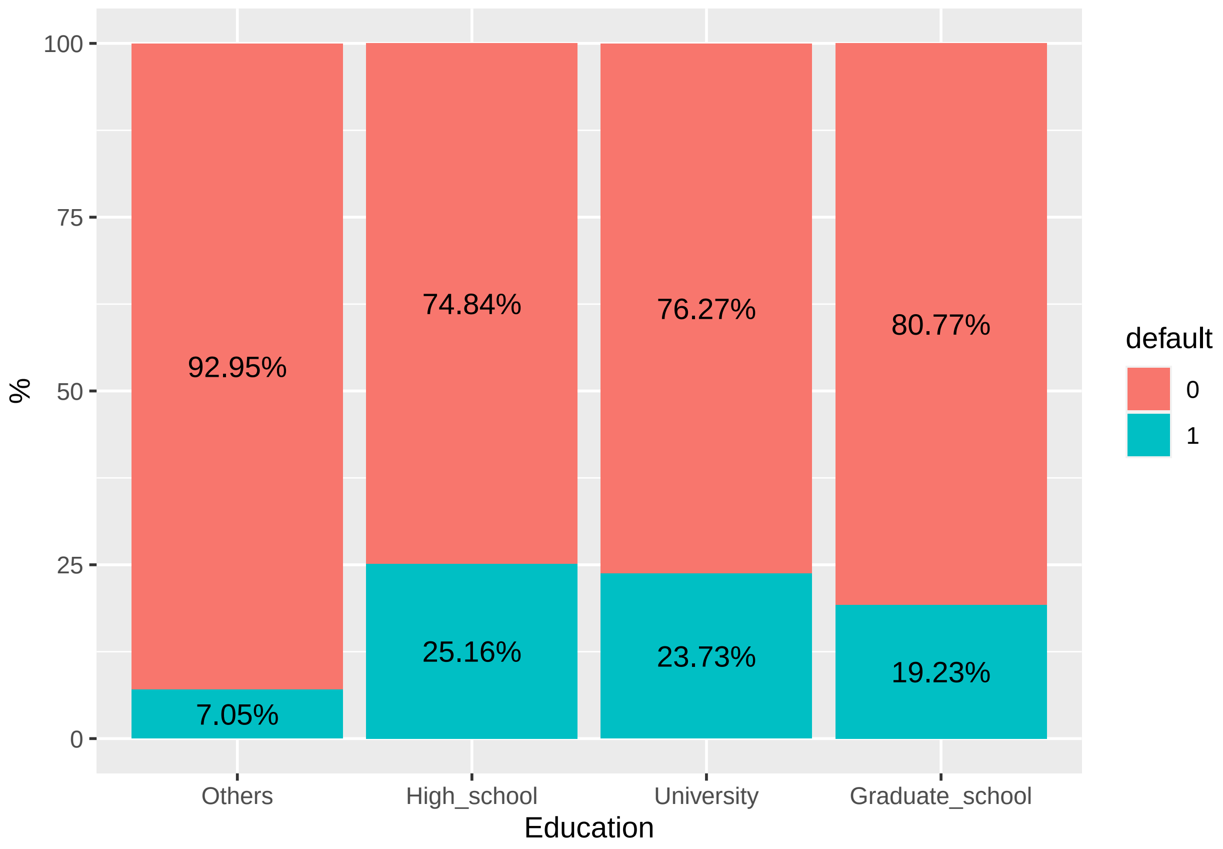


```
#Replacing values
data$V7[data$V7 >2 ]<- 2
data$V8[data$V8 >2 ]<- 2
data$V9[data$V9 >2 ]<- 2
data$V10[data$V10 >2 ]<- 2
data$V11[data$V11 >2 ]<- 2
data$V12[data$V12 >2 ]<- 2
```

```
data$GENDER = ifelse(data$V3 == 1, "Male", "Female")

#V4: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
data$EDUCATION <- ifelse(data$V4%in%c(0,4,5,6), 0, data$V4)
data$EDUCATION <- factor(data$EDUCATION,
                         labels = c("Others","Graduate_school", "University", "High_school"))
data$default <- as.factor(data$V25)

# Bar Graph for gender
gender_plot<- ggplot(data, aes(GENDER))+
  geom_bar(aes(fill=default), width = 0.5) +
  labs(title="Gender") +
  stat_count(aes(label = ..count..), geom = "label")

gender_plot
```



```
# Bar graph for Education
data %>%
  count(EDUCATION, default) %>%
    group_by(EDUCATION) %>%
    mutate(n = n/sum(n) * 100) %>%
    ggplot() +
    aes(factor(EDUCATION,
               levels = c("Others", "High_school", "University", "Graduate_school")), n,
        fill = default, label = paste0(round(n, 2), "%")) +
    geom_col() +
    geom_text(position=position_stack(0.5))+
    xlab("Education")+
    ylab("%")
```


%
100
75
50
25
0
92.95%
7.05%
74.84%
25.16%
76.27%
23.73%
80.77%
19.23%
Others
High_school
University
Graduate_school
Education
default
0
1