Bitácora 4

Introducción

Visualización y Limpieza de la Nueva Base de Datos

Queremos señalar que fue necesario cambiar la base de datos utilizada, ya que la anterior parecía haber sido generada artificialmente, sin provenir de datos reales. Para la presente bitácora, hemos optado por una base de datos auténtica que incluye información sobre pagos por defecto, factores demográficos, datos de crédito, historial de pagos y estados de cuenta de clientes de tarjetas de crédito en Taiwán, correspondiente al periodo de abril a septiembre de 2005.

Empezamos el estudio de la base de datos, para ello lo primodiarla es cargarla.

```
Rows: 30000 Columns: 25
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (25): ID, LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3,...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# A tibble: 6 x 25
     ID LIMIT_BAL   SEX EDUCATION MARRIAGE   AGE PAY_0 PAY_2 PAY_3 PAY_4 PAY_5
  <dbl>     <dbl> <dbl>     <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     1     20000     2         2        1    24     2     2    -1    -1    -2
2     2    120000     2         2        2    26    -1     2     0     0     0
3     3     90000     2         2        2    34     0     0     0     0     0
4     4     50000     2         2        1    37     0     0     0     0     0
5     5     50000     1         2        1    57    -1     0    -1     0     0
6     6     50000     1         1        2    37     0     0     0     0     0
# i 14 more variables: PAY_6 <dbl>, BILL_AMT1 <dbl>, BILL_AMT2 <dbl>,
#   BILL_AMT3 <dbl>, BILL_AMT4 <dbl>, BILL_AMT5 <dbl>, BILL_AMT6 <dbl>,
#   PAY_AMT1 <dbl>, PAY_AMT2 <dbl>, PAY_AMT3 <dbl>, PAY_AMT4 <dbl>,
#   PAY_AMT5 <dbl>, PAY_AMT6 <dbl>, default.payment.next.month <dbl>
```

Una vez cargada la base de datos, es importante realizar algunas observaciones iniciales antes de comenzar a trabajar con ella. En primer lugar, queremos revisar los nombres de las variables y el tipo de datos que representan, ya que esto nos permite empezar a considerar qué técnicas estadísticas podríamos aplicar. Además, nos interesa verificar que la base de datos esté limpia; para ello, realizaremos un conteo de los valores faltantes en el dataset.

```r
names(data_credit)
```

```
 [1] "ID"                         "LIMIT_BAL"
 [3] "SEX"                        "EDUCATION"
 [5] "MARRIAGE"                   "AGE"
 [7] "PAY_0"                      "PAY_2"
 [9] "PAY_3"                      "PAY_4"
[11] "PAY_5"                      "PAY_6"
[13] "BILL_AMT1"                  "BILL_AMT2"
[15] "BILL_AMT3"                  "BILL_AMT4"
[17] "BILL_AMT5"                  "BILL_AMT6"
[19] "PAY_AMT1"                   "PAY_AMT2"
[21] "PAY_AMT3"                   "PAY_AMT4"
[23] "PAY_AMT5"                   "PAY_AMT6"
[25] "default.payment.next.month"
```

Con los nombres hacemos un pequeño resumen de qué significa cada uno:

- ID: ID de cada cliente.
- LIMIT_BAL: Monto de crédito otorgado en dólares taiwaneses (NT) (incluye crédito individual y familiar/suplementario).
- SEX: Género (1=hombre, 2=mujer).
- EDUCATION: Nivel educativo (1=posgrado, 2=universidad, 3=preparatoria, 4=otros, 5=desconocido, 6=desconocido).
- MARRIAGE: Estado civil (1=casado, 2=soltero, 3=otros).
- AGE: Edad en años.
- PAY_0: Estado de reembolso en septiembre de 2005 (-1=pago puntual, 1=atraso de un mes, 2=atraso de dos meses, …, 8=atraso de ocho meses, 9=atraso de nueve meses o más).
- PAY_2: Estado de reembolso en agosto de 2005 (escala igual a la anterior).
- PAY_3: Estado de reembolso en julio de 2005 (escala igual a la anterior).
- PAY_4: Estado de reembolso en junio de 2005 (escala igual a la anterior).
- PAY_5: Estado de reembolso en mayo de 2005 (escala igual a la anterior).
- PAY_6: Estado de reembolso en abril de 2005 (escala igual a la anterior).
- BILL_AMT1: Monto del estado de cuenta en septiembre de 2005 (dólares taiwaneses, NT).
- BILL_AMT2: Monto del estado de cuenta en agosto de 2005 (dólares taiwaneses, NT).
- BILL_AMT3: Monto del estado de cuenta en julio de 2005 (dólares taiwaneses, NT).
- BILL_AMT4: Monto del estado de cuenta en junio de 2005 (dólares taiwaneses, NT).
- BILL_AMT5: Monto del estado de cuenta en mayo de 2005 (dólares taiwaneses, NT).
- BILL_AMT6: Monto del estado de cuenta en abril de 2005 (dólares taiwaneses, NT).
- PAY_AMT1: Monto del pago anterior en septiembre de 2005 (dólares taiwaneses, NT).

- PAY_AMT2: Monto del pago anterior en agosto de 2005 (dólares taiwaneses, NT).
- PAY_AMT3: Monto del pago anterior en julio de 2005 (dólares taiwaneses, NT).
- PAY_AMT4: Monto del pago anterior en junio de 2005 (dólares taiwaneses, NT).
- PAY_AMT5: Monto del pago anterior en mayo de 2005 (dólares taiwaneses, NT).
- PAY_AMT6: Monto del pago anterior en abril de 2005 (dólares taiwaneses, NT).
- default.payment.next.month: Pago en mora (1=sí, 0=no).

Por otro lado, veamos con qué tipo de datos contamos.

```
# Verificar el tipo de datos del dataset.
str(data_credit)
```

```
spc_tbl_ [30,000 x 25] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ID                        : num [1:30000] 1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL                 : num [1:30000] 20000 120000 90000 50000 50000 50000 500000 1000
 $ SEX                       : num [1:30000] 2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : num [1:30000] 2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : num [1:30000] 1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : num [1:30000] 24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0                     : num [1:30000] 2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2                     : num [1:30000] 2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3                     : num [1:30000] -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4                     : num [1:30000] -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5                     : num [1:30000] -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6                     : num [1:30000] -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1                 : num [1:30000] 3913 2682 29239 46990 8617 ...
 $ BILL_AMT2                 : num [1:30000] 3102 1725 14027 48233 5670 ...
 $ BILL_AMT3                 : num [1:30000] 689 2682 13559 49291 35835 ...
 $ BILL_AMT4                 : num [1:30000] 0 3272 14331 28314 20940 ...
 $ BILL_AMT5                 : num [1:30000] 0 3455 14948 28959 19146 ...
 $ BILL_AMT6                 : num [1:30000] 0 3261 15549 29547 19131 ...
 $ PAY_AMT1                  : num [1:30000] 0 0 1518 2000 2000 ...
 $ PAY_AMT2                  : num [1:30000] 689 1000 1500 2019 36681 ...
 $ PAY_AMT3                  : num [1:30000] 0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4                  : num [1:30000] 0 1000 1000 1100 9000 ...
 $ PAY_AMT5                  : num [1:30000] 0 0 1000 1069 689 ...
 $ PAY_AMT6                  : num [1:30000] 0 2000 5000 1000 679 ...
 $ default.payment.next.month: num [1:30000] 1 1 0 0 0 0 0 0 0 0 ...
 - attr(*, "spec")=
  .. cols(
  ..   ID = col_double(),
  ..   LIMIT_BAL = col_double(),
```

```
  ..   SEX = col_double(),
  ..   EDUCATION = col_double(),
  ..   MARRIAGE = col_double(),
  ..   AGE = col_double(),
  ..   PAY_0 = col_double(),
  ..   PAY_2 = col_double(),
  ..   PAY_3 = col_double(),
  ..   PAY_4 = col_double(),
  ..   PAY_5 = col_double(),
  ..   PAY_6 = col_double(),
  ..   BILL_AMT1 = col_double(),
  ..   BILL_AMT2 = col_double(),
  ..   BILL_AMT3 = col_double(),
  ..   BILL_AMT4 = col_double(),
  ..   BILL_AMT5 = col_double(),
  ..   BILL_AMT6 = col_double(),
  ..   PAY_AMT1 = col_double(),
  ..   PAY_AMT2 = col_double(),
  ..   PAY_AMT3 = col_double(),
  ..   PAY_AMT4 = col_double(),
  ..   PAY_AMT5 = col_double(),
  ..   PAY_AMT6 = col_double(),
  ..   default.payment.next.month = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Antes de hacer un pequeño conteo de la cantidad de missing value, vamos a realizar un pequeño resumen de la base de datos.

```
library(skimr)
```

```
Warning: package 'skimr' was built under R version 4.4.2
```

```
# Resumen del dataset.
summary(data_credit)
```

```
       ID           LIMIT_BAL            SEX          EDUCATION     
 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
   MARRIAGE          AGE            PAY_0             PAY_2
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
    PAY_3             PAY_4             PAY_5             PAY_6
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
  BILL_AMT1         BILL_AMT2         BILL_AMT3         BILL_AMT4
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
  BILL_AMT5         BILL_AMT6          PAY_AMT1         PAY_AMT2
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
   PAY_AMT3          PAY_AMT4          PAY_AMT5           PAY_AMT6
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
default.payment.next.month
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.2212
3rd Qu.:0.0000
```

```
 Max.   :1.0000
```

```
# Generamos un cuadro resumen, con la información anterior.
skim(data_credit)
```

Table 1: Data summary

Name	data_credit
Number of rows	30000
Number of columns	25
_____	
Column type frequency:	
numeric	25
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1	15000.50	8660.40	1	7500.75	15000.5	22500.25	30000	
LIMIT_BAL	0	1	167484.32	129747.66	10000	50000.00	140000.0	240000.00	1000000	
SEX	0	1	1.60	0.49	1	1.00	2.0	2.00	2	
EDUCATION	0	1	1.85	0.79	0	1.00	2.0	2.00	6	
MARRIAGE	0	1	1.55	0.52	0	1.00	2.0	2.00	3	
AGE	0	1	35.49	9.22	21	28.00	34.0	41.00	79	
PAY_0	0	1	-0.02	1.12	-2	-1.00	0.0	0.00	8	
PAY_2	0	1	-0.13	1.20	-2	-1.00	0.0	0.00	8	
PAY_3	0	1	-0.17	1.20	-2	-1.00	0.0	0.00	8	
PAY_4	0	1	-0.22	1.17	-2	-1.00	0.0	0.00	8	
PAY_5	0	1	-0.27	1.13	-2	-1.00	0.0	0.00	8	
PAY_6	0	1	-0.29	1.15	-2	-1.00	0.0	0.00	8	
BILL_AMT1	0	1	51223.33	73635.86	-165580	3558.75	22381.5	67091.00	964511	
BILL_AMT2	0	1	49179.08	71173.77	-69777	2984.75	21200.0	64006.25	983931	
BILL_AMT3	0	1	47013.15	69349.39	-157264	2666.25	20088.5	60164.75	1664089	
BILL_AMT4	0	1	43262.95	64332.86	-170000	2326.75	19052.0	54506.00	891586	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
BILL_AMT5	0	1	40311.40	60797.16	-81334	1763.00	18104.50	50190.50	927171	
BILL_AMT6	0	1	38871.76	59554.11	-339603	1256.00	17071.00	49198.25	961664	
PAY_AMT1	0	1	5663.58	16563.28	0	1000.00	2100.0	5006.00	873552	
PAY_AMT2	0	1	5921.16	23040.87	0	833.00	2009.0	5000.00	1684259	
PAY_AMT3	0	1	5225.68	17606.96	0	390.00	1800.0	4505.00	896040	
PAY_AMT4	0	1	4826.08	15666.16	0	296.00	1500.0	4013.25	621000	
PAY_AMT5	0	1	4799.39	15278.31	0	252.50	1500.0	4031.50	426529	
PAY_AMT6	0	1	5215.50	17777.47	0	117.75	1500.0	4000.00	528666	
default.payment.next.month	0	1	0.22	0.42	0	0.00	0.0	0.00	1	

Apesar de qué el cuadro resumen anterior ya nos indica que las variables no tienen missing value, nos parece pertinente verificarlo de manera aislada, para ello.

```
#Verificamos la cantidad de datos nulos que hay nuestro dataset
sum(is.na(data_credit))
```

```
[1] 0
```

```
#Verificamos la cantidad de datos nulos que hay en cada columna
sapply(data_credit, function(x) sum(is.na(x)))
```

```
                ID                 LIMIT_BAL 
                 0                         0 
               SEX                 EDUCATION 
                 0                         0 
          MARRIAGE                       AGE 
                 0                         0 
             PAY_0                     PAY_2 
                 0                         0 
             PAY_3                     PAY_4 
                 0                         0 
             PAY_5                     PAY_6 
                 0                         0 
         BILL_AMT1                 BILL_AMT2 
                 0                         0 
         BILL_AMT3                 BILL_AMT4 
                 0                         0 
```

```
                  BILL_AMT5                   BILL_AMT6
                          0                           0
                   PAY_AMT1                    PAY_AMT2
                          0                           0
                   PAY_AMT3                    PAY_AMT4
                          0                           0
                   PAY_AMT5                    PAY_AMT6
                          0                           0
default.payment.next.month
                          0
```

Sin embargo, hay columnas que aunque no están vacías, contiene datos que no nos sirven, por eso hay que filtrar estos datos. Estos datos no sirven por el hecho de que no están definidos como parámetros significativos, es decir, si tenemos definidos la variable sexo como 1=hombre y 2=mujer, entonces aparecen números como el 3 y 0, por ello hay que filtrarlos, ya que afectan los análisis.

Nos damos cuenta de ello, gracias a ver el cuadro resumen, que aparecen valores que no deberían aparecer.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v purrr     1.0.2
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
# Filtramos los datos que están definidos.
data_credit <- data_credit %>%
  filter(MARRIAGE %in% c(1, 2, 3))

# Filtramos los datos que están definidos.
data_credit <- data_credit %>%
  filter(EDUCATION %in% c(1, 2, 3, 4, 5, 6))

data_credit <- data_credit %>%
  filter(PAY_0 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))
```

```r
data_credit <- data_credit %>%
  filter(PAY_2 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_3 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_4 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_5 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))

data_credit <- data_credit %>%
  filter(PAY_6 %in% c(-1, 1, 2, 3, 4, 5, 6, 7, 8, 9))
```

Realizamos de nuevo nuestro cuadro resumen después de esta filtración con el objetivo de observar las nuevas tendencias de la base.

```r
library(skimr)

# Resumen del dataset.
summary(data_credit)
```

```
       ID          LIMIT_BAL           SEX          EDUCATION
 Min.   :   12   Min.   : 10000   Min.   :1.000   Min.   :1.000
 1st Qu.: 6941   1st Qu.: 60000   1st Qu.:1.000   1st Qu.:1.000
 Median :13671   Median :150000   Median :2.000   Median :2.000
 Mean   :14278   Mean   :171695   Mean   :1.592   Mean   :1.757
 3rd Qu.:21713   3rd Qu.:240000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :29995   Max.   :740000   Max.   :2.000   Max.   :6.000
    MARRIAGE          AGE            PAY_0             PAY_2
 Min.   :1.000   Min.   :21.00   Min.   :-1.0000   Min.   :-1.0000
 1st Qu.:1.000   1st Qu.:29.00   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median :1.000   Median :35.00   Median :-1.0000   Median :-1.0000
 Mean   :1.493   Mean   :36.53   Mean   : 0.1819   Mean   : 0.2869
 3rd Qu.:2.000   3rd Qu.:43.00   3rd Qu.: 2.0000   3rd Qu.: 2.0000
 Max.   :3.000   Max.   :72.00   Max.   : 8.0000   Max.   : 8.0000
     PAY_3             PAY_4             PAY_5             PAY_6
 Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median :-1.0000   Median :-1.0000   Median :-1.0000   Median :-1.0000
```

```
 Mean   : 0.3188   Mean   : 0.2837   Mean   : 0.2424   Mean   : 0.2488
 3rd Qu.: 2.0000   3rd Qu.: 2.0000   3rd Qu.: 2.0000   3rd Qu.: 2.0000
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000
   BILL_AMT1         BILL_AMT2         BILL_AMT3         BILL_AMT4
 Min.   : -4316   Min.   :-24704   Min.   :-61506   Min.   : -3903
 1st Qu.:   931   1st Qu.:   856   1st Qu.:   835   1st Qu.:   828
 Median :  4394   Median :  4398   Median :  4192   Median :  4176
 Mean   : 22124   Mean   : 22296   Mean   : 22304   Mean   : 22641
 3rd Qu.: 22231   3rd Qu.: 22678   3rd Qu.: 22974   3rd Qu.: 22819
 Max.   :581775   Max.   :572677   Max.   :471175   Max.   :486776
   BILL_AMT5         BILL_AMT6          PAY_AMT1          PAY_AMT2
 Min.   : -3876   Min.   :-339603   Min.   :     0   Min.   :     0
 1st Qu.:   838   1st Qu.:    776   1st Qu.:   316   1st Qu.:   316
 Median :  4069   Median :   4120   Median :  1600   Median :  1595
 Mean   : 22589   Mean   :  22676   Mean   :  4669   Mean   :  4608
 3rd Qu.: 23341   3rd Qu.:  23710   3rd Qu.:  4427   3rd Qu.:  4398
 Max.   :503914   Max.   : 527711   Max.   :187206   Max.   :302961
    PAY_AMT3          PAY_AMT4          PAY_AMT5          PAY_AMT6
 Min.   :     0   Min.   :     0   Min.   :     0   Min.   :     0
 1st Qu.:   316   1st Qu.:   331   1st Qu.:   100   1st Qu.:     0
 Median :  1443   Median :  1443   Median :  1225   Median :  1044
 Mean   :  4756   Mean   :  4558   Mean   :  4594   Mean   :  4621
 3rd Qu.:  4200   3rd Qu.:  4100   3rd Qu.:  4000   3rd Qu.:  3710
 Max.   :417588   Max.   :193712   Max.   :303512   Max.   :345293
 default.payment.next.month
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.3548
 3rd Qu.:1.0000
 Max.   :1.0000
```

```
# Generamos un cuadro resumen, con la información anterior.
skim(data_credit)
```

Table 3: Data summary

Name	data_credit
Number of rows	4047
Number of columns	25
Column type frequency:	

numeric	25
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1	14277.57	8616.90	12	6941	13671	21712.5	29995	
LIMIT_BAL	0	1	171695.08	125912.17	10000	60000	150000	240000.0	740000	
SEX	0	1	1.59	0.49	1	1	2	2.0	2	
EDUCATION	0	1	1.76	0.75	1	1	2	2.0	6	
MARRIAGE	0	1	1.49	0.52	1	1	1	2.0	3	
AGE	0	1	36.53	9.19	21	29	35	43.0	72	
PAY_0	0	1	0.18	1.58	-1	-1	-1	2.0	8	
PAY_2	0	1	0.29	1.68	-1	-1	-1	2.0	8	
PAY_3	0	1	0.32	1.74	-1	-1	-1	2.0	8	
PAY_4	0	1	0.28	1.80	-1	-1	-1	2.0	8	
PAY_5	0	1	0.24	1.78	-1	-1	-1	2.0	8	
PAY_6	0	1	0.25	1.75	-1	-1	-1	2.0	8	
BILL_AMT1	0	1	22124.22	44383.68	-4316	931	4394	22231.0	581775	
BILL_AMT2	0	1	22296.39	44437.54	-24704	856	4398	22678.0	572677	
BILL_AMT3	0	1	22304.37	44520.48	-61506	835	4192	22974.0	471175	
BILL_AMT4	0	1	22640.65	45041.16	-3903	828	4176	22818.5	486776	
BILL_AMT5	0	1	22588.65	44576.24	-3876	838	4069	23341.0	503914	
BILL_AMT6	0	1	22675.85	45579.27	-339603	776	4120	23710.0	527711	
PAY_AMT1	0	1	4669.12	10921.61	0	316	1600	4427.0	187206	
PAY_AMT2	0	1	4608.06	11960.55	0	316	1595	4398.5	302961	
PAY_AMT3	0	1	4756.23	13590.04	0	316	1443	4200.0	417588	
PAY_AMT4	0	1	4557.61	11100.86	0	331	1443	4100.0	193712	
PAY_AMT5	0	1	4594.45	13511.57	0	100	1225	4000.0	303512	
PAY_AMT6	0	1	4620.82	15153.76	0	0	1044	3710.0	345293	
default.payment.next.month	0	1	0.35	0.48	0	0	0	1.0	1	

Gracias a todo lo anterior ya tenemos una base de datos limpia, lista para el análisis estadís-

tico.

Análisis Estadístico de la Base de Datos

Para esta sección, nuestro objetivo es explorar el comportamiento de nuestro conjunto de datos de manera más profunda. Para ello, aplicaremos algunas técnicas estadísticas que nos permitirán obtener información relevante.

En primer lugar, vamos a construir una matriz de correlación. Esta herramienta nos ayudará a identificar las relaciones más fuertes entre las variables de nuestro dataset. Es importante recordar que la matriz de correlación solo tiene sentido cuando se aplica a variables numéricas, por lo que es esencial seleccionar adecuadamente las variables antes de proceder con este análisis.

Matriz de Correlación

Vamos a seleccionar solo las variables que sean de interés, es decir, vamos a quitar variables como el id, y el sexo, ya que el id es único para cada persona en el data set, y la variable sexo es una variable de tipo categórico, lo mismo pasa con la variable de defaul.payment, la cual a pesar de representarse con números, estos significan que son categóricos.

```
library(tidyverse)
# Seleccionamos las variables.
data_credit_numerico <- data_credit %>% select(-ID, -SEX, -EDUCATION, -MARRIAGE, -PAY_0, -PA
```

Con los datos filtrados, podemos entonces realizar nuestra matriz de correlación.

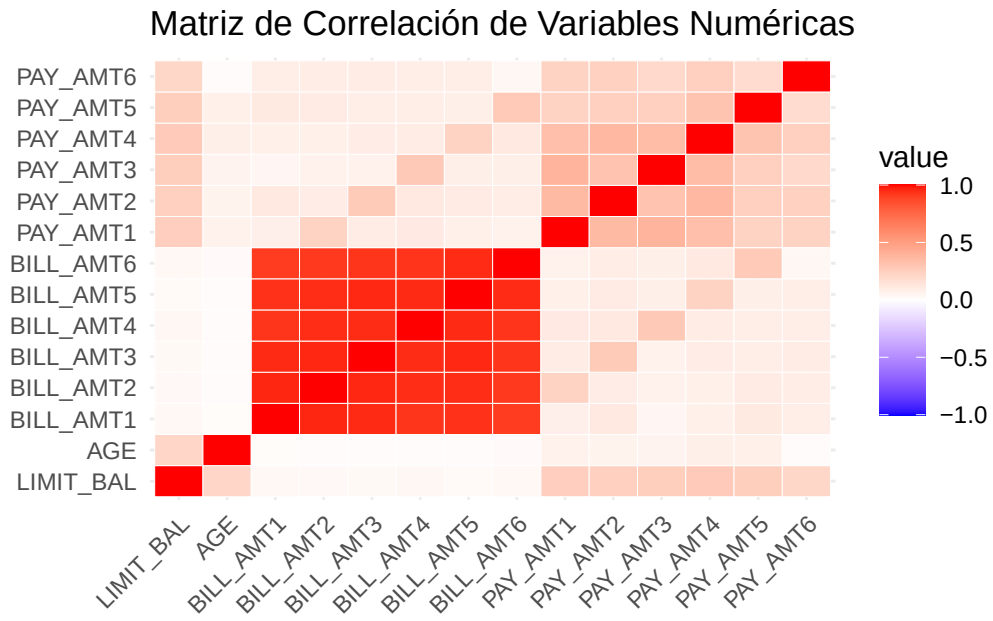
```
library(dplyr)
library(ggplot2)
library(reshape2)
```

```
Warning: package 'reshape2' was built under R version 4.4.2
```

```
Adjuntando el paquete: 'reshape2'
```

```
The following object is masked from 'package:tidyr':

    smiths
```

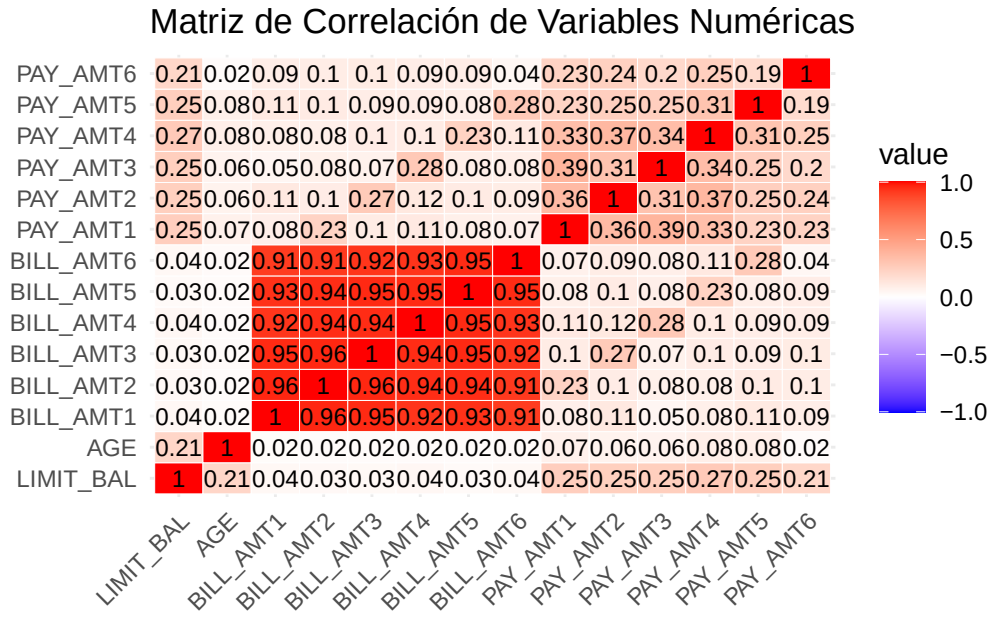
```r
# Crear la matriz de correlación
matriz_correlacion <- cor(data_credit_numerico, use = "complete.obs")

# Creamos la variable para ver la matriz de correlación como un gráfico de calor
base_matriz <- melt(matriz_correlacion)

# Hacemos el plot de la base
ggplot(base_matriz, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1
  theme_minimal() +
  labs(title = "Matriz de Correlación de Variables Numéricas",
       x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Agregamos la misma matriz, pero esta vez indicando los índices de correlación, para visualizar de manera gráfica y matemática la correlación existente.

```r
library(dplyr)
library(ggplot2)
library(reshape2)
```

```r
# Crear la matriz de correlación
matriz_correlacion <- cor(data_credit_numerico, use = "complete.obs")

# Creamos la variable para ver la matriz de correlación como un gráfico de calor
base_matriz <- melt(matriz_correlacion)

# Hacemos el plot de la base
ggplot(base_matriz, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1
  theme_minimal() + geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  labs(title = "Matriz de Correlación de Variables Numéricas",
       x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Gráficos relacionados a la Base de Datos

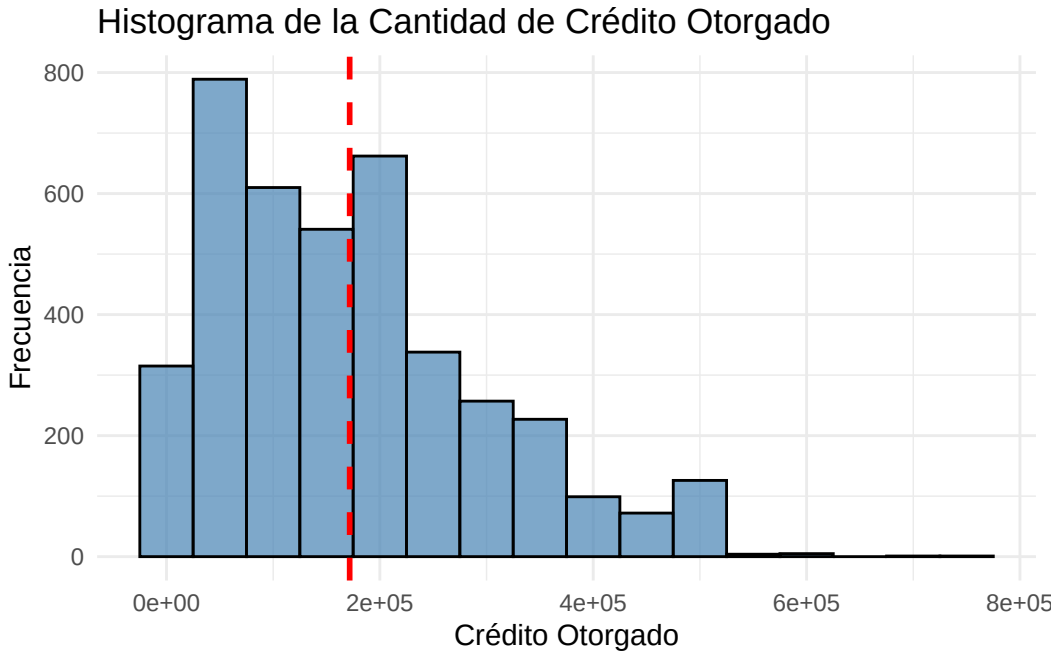
Gráficos de Variables Numéricas

Vamos a realizar algunos gráficos de la base, con el fin de observar cómo se comportan los datos, para ello, primero veamos algunos histogramas, recordemos que los histogramas están hechos para variables numéricas, por lo que trataremos de ir en orden a la hora de gráficarlas.

Damos inicio con la variable de “LIMIT_BAL”, la cual se refiere al crédito otorgado.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL)), linetype = "dashed", color = "red", size = 1
  labs(title = "Histograma de la Cantidad de Crédito Otorgado",
       x = "Crédito Otorgado",
       y = "Frecuencia") +
  theme_minimal()
```

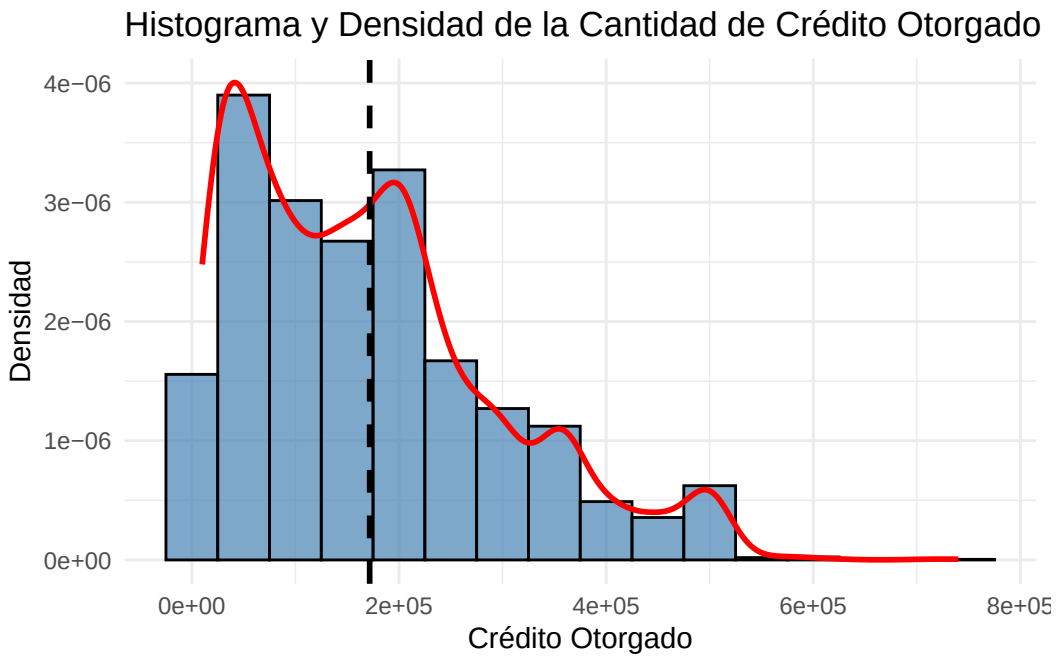
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Agregamos un gráfico adicional, donde podemos observar como se comporta la densidad de esta variable.

```
library(ggplot2)
library(dplyr)
```

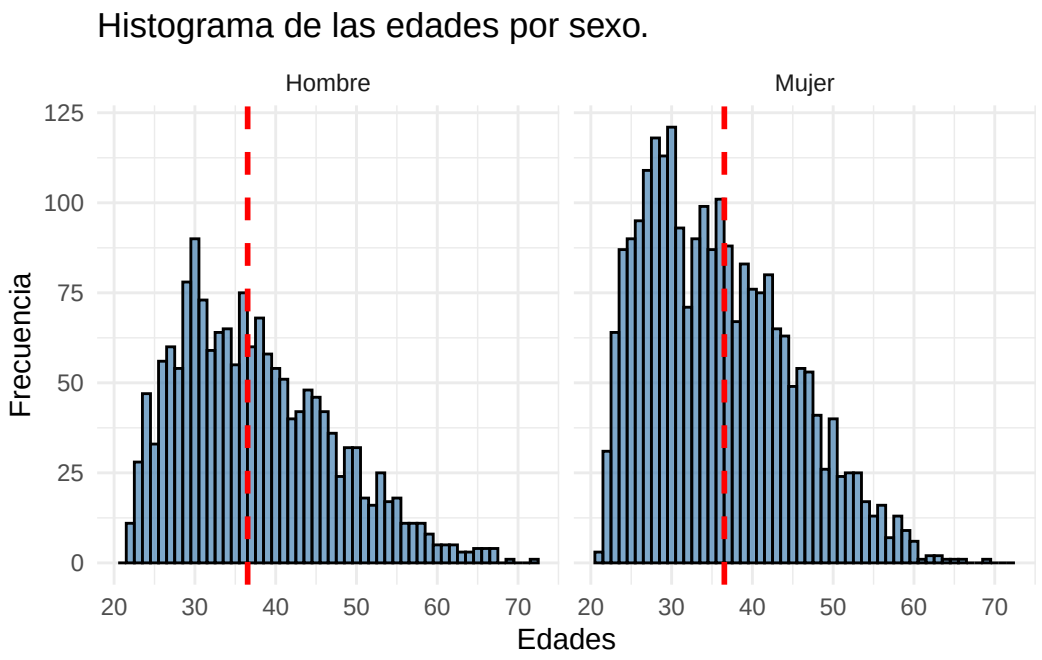
```r
data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 50000, fill = "steelblue", color =
  geom_density(color = "red", size = 1) + # Densidad en línea roja
  geom_vline(aes(xintercept = mean(LIMIT_BAL)), linetype = "dashed", color = "black", size =
  labs(title = "Histograma y Densidad de la Cantidad de Crédito Otorgado",
       x = "Crédito Otorgado",
       y = "Densidad") +
  theme_minimal()
```



Procedemos ahora con la variable de la edad, esto porque queremos visualizar la distribución de las edades de nuestro data set.

```r
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
```

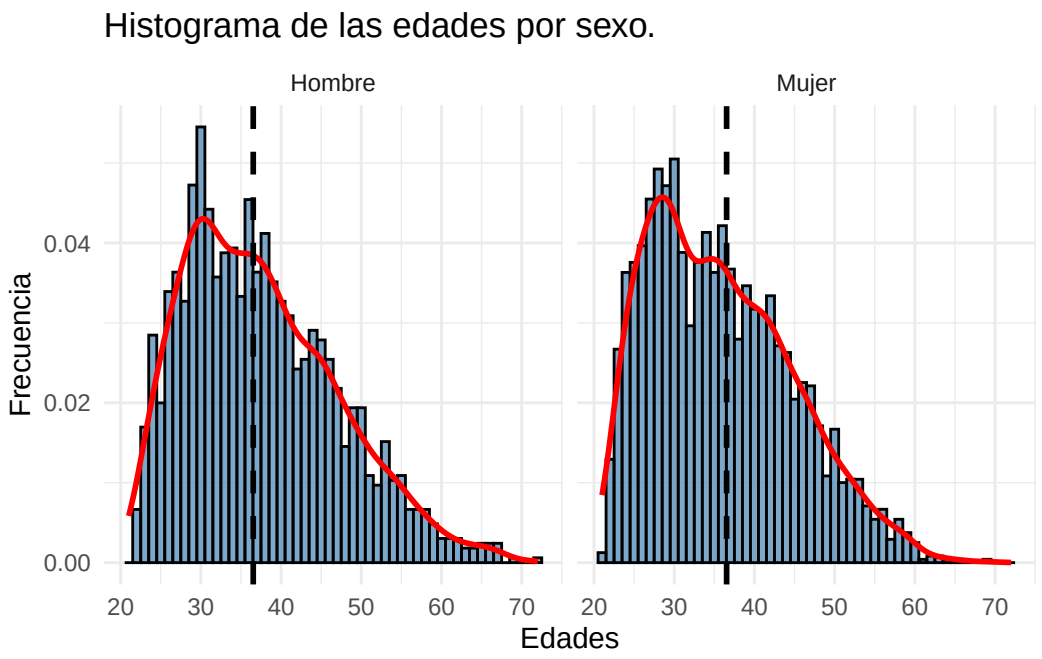
```
            size = 1) +
  labs(title = "Histograma de las edades por sexo.",
       x = "Edades",
       y = "Frecuencia") +
  facet_wrap(~SEX, labeller = as_labeller(c("1" = "Hombre", "2" = "Mujer"))) +
  theme_minimal()
```



Agregamos un gráfico adicional donde podemos observar la densidad de la variable.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "steelblue", color = "bla
  geom_density(aes(y = after_stat(density)), color = "red", size = 1) +  # Curva de densidad
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "black",
             linetype = "dashed",
             size = 1) +  # Línea roja punteada
  labs(title = "Histograma de las edades por sexo.",
       x = "Edades",
       y = "Frecuencia") +
```

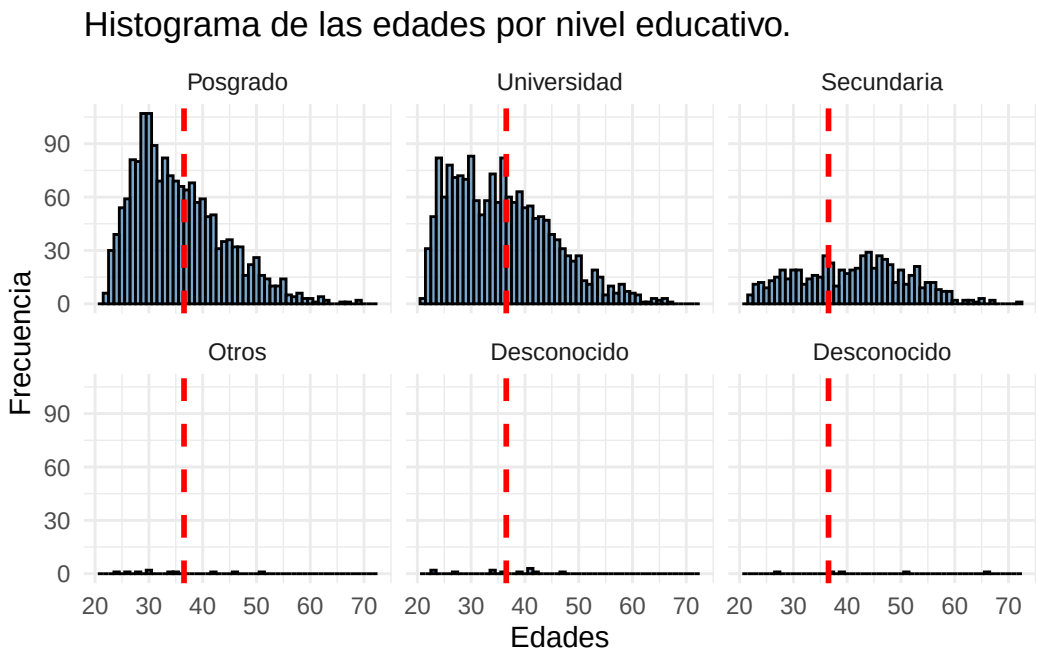
```
  facet_wrap(~SEX, labeller = as_labeller(c("1" = "Hombre", "2" = "Mujer"))) +
  theme_minimal()
```



Con estas dos variables es sufiente de gráficos aislados, debido a la naturaleza de las demás variables numéricas que tenemos en el dataset, en vez de eso, vamos a ver cómo se comportan los gráficos, cuando se plotean las relaciones entre ellos.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de las edades por nivel educativo.",
       x = "Edades",
       y = "Frecuencia") +
  facet_wrap(~EDUCATION, labeller = labeller(EDUCATION = c(
    "1" = "Posgrado",
    "2" = "Universidad",
```

```
    "3" = "Secundaria",
    "4" = "Otros",
    "5" = "Desconocido",
    "6" = "Desconocido"
  ))) + theme_minimal()
```

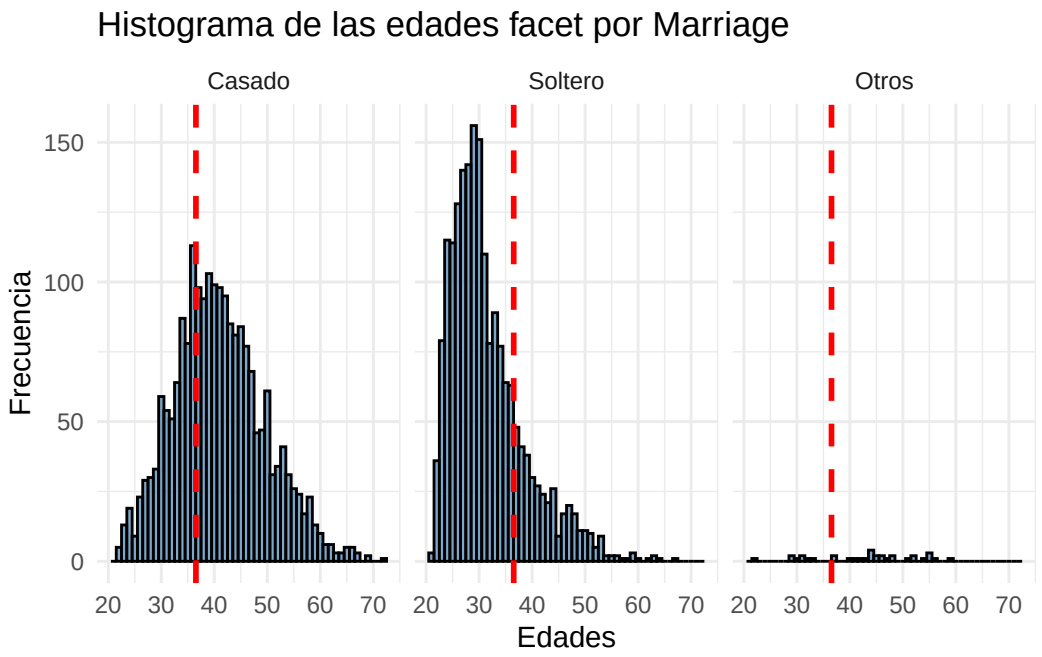


Gracias a este gráfico, podemos observar que la gran mayoría de los datos se encuentran en los subvariables de Universidad. Secundaria y Posgrado.

Ahora veamos como se distribuye la edad, con respecto a la variable de Estado Civil.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +  # Línea roja punteada para la media
  labs(title = "Histograma de las edades facet por Marriage",
       x = "Edades",
```

```
       y = "Frecuencia") +
  facet_wrap(~MARRIAGE, labeller = as_labeller(c("1" = "Casado", "2" = "Soltero", "3" = "Otr
  theme_minimal()
```

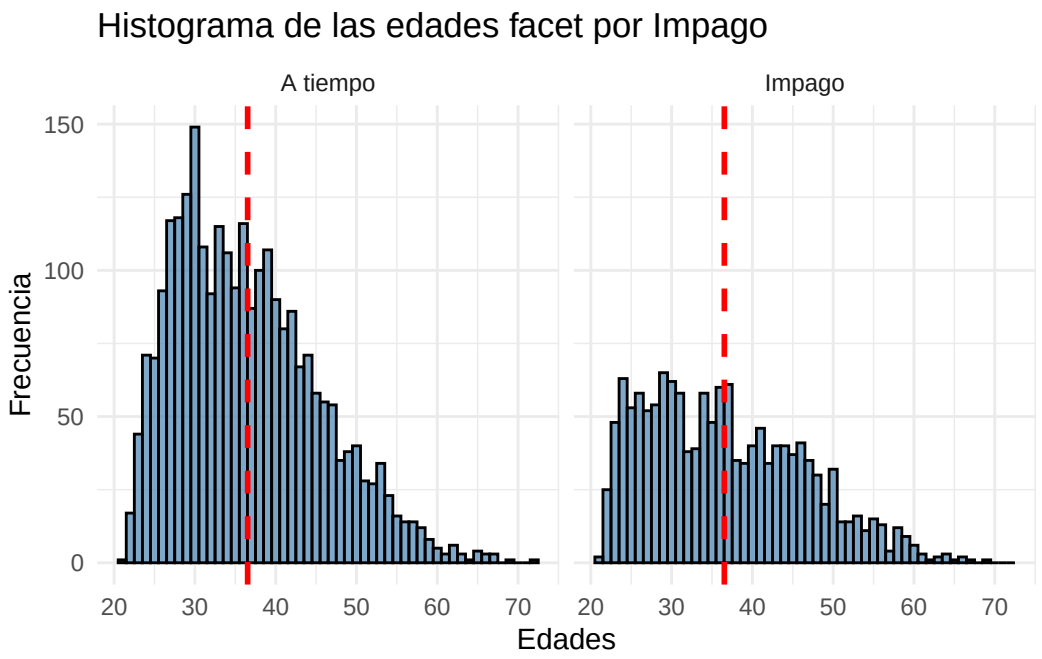


Conjeturando, podemos ver que la distribución de las edades de las personas que están casadas, sigue más o menos una distribución normal, mientras que las personas que están solteras lo hace como una distribución exponencial.

Por último nos interesa saber cómo se distribuyen las edades en relación a la variable de impago, que es la que tiene el principal peso de dicho estudio.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = AGE)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(AGE, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de las edades facet por Impago",
       x = "Edades",
```

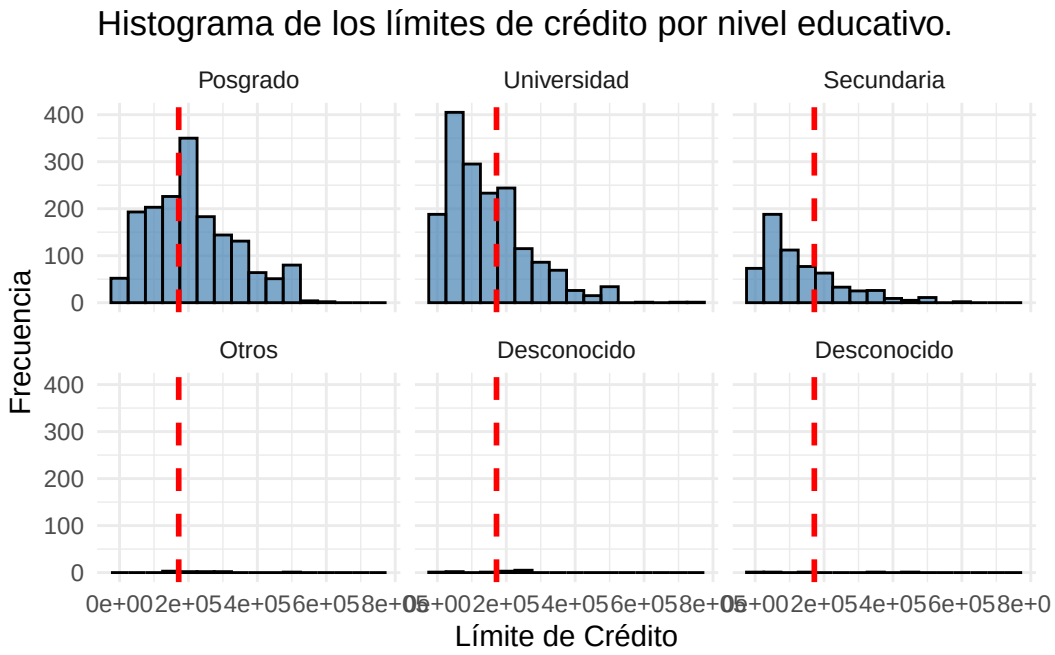
```
       y = "Frecuencia") +
  facet_wrap(~default.payment.next.month, labeller = as_labeller(c("0" = "A tiempo", "1" = "I
  theme_minimal()
```



Vamos a realizar las mismas gráficas, pero esta vez con la la variable de LIMIT_BAL, para observar su comportamiento.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito por nivel educativo.",
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~EDUCATION, labeller = labeller(EDUCATION = c(
    "1" = "Posgrado",
    "2" = "Universidad",
```

```
    "3" = "Secundaria",
    "4" = "Otros",
    "5" = "Desconocido",
    "6" = "Desconocido"
  ))) + theme_minimal()
```



Agregamos también un gráfico de densidades, esto con el objetivo de ver la distribución.

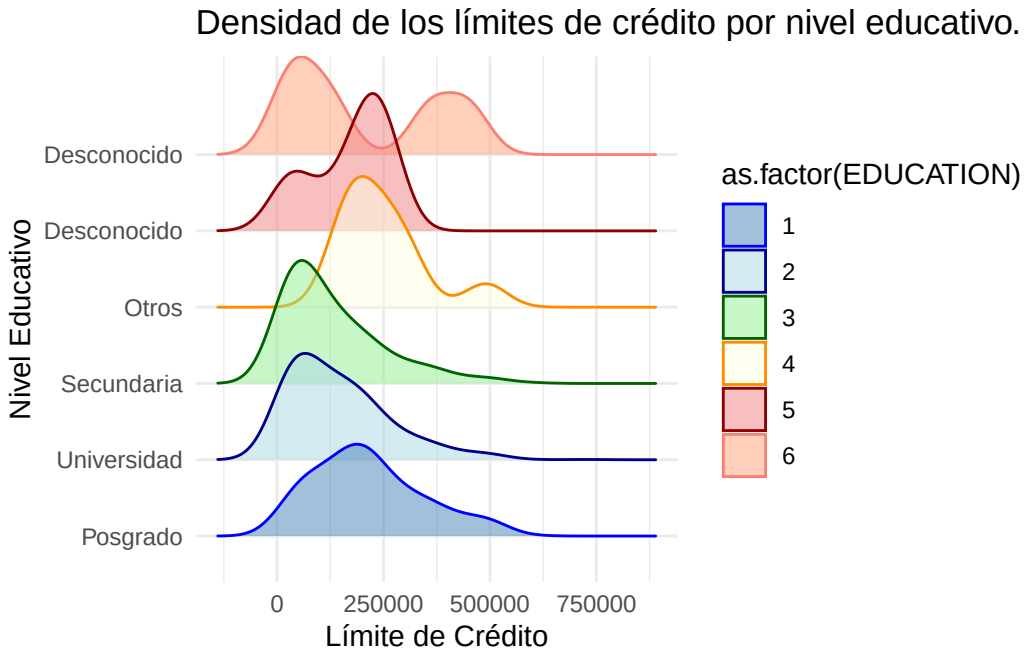
```
library(ggplot2)
library(dplyr)
library(ggridges)
```

```
Warning: package 'ggridges' was built under R version 4.4.2
```

```
data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(EDUCATION), fill = as.factor(EDUCATION), color = as
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por nivel educativo.",
       x = "Límite de Crédito",
       y = "Nivel Educativo") +
  scale_fill_manual(values = c("1" = "steelblue", "2" = "lightblue", "3" = "lightgreen",
                               "4" = "lightyellow", "5" = "lightcoral", "6" = "lightsalmon"))
```

```
  scale_color_manual(values = c("1" = "blue", "2" = "darkblue", "3" = "darkgreen",
                                "4" = "darkorange", "5" = "darkred", "6" = "salmon")) +
  scale_y_discrete(labels = c(
    "1" = "Posgrado",
    "2" = "Universidad",
    "3" = "Secundaria",
    "4" = "Otros",
    "5" = "Desconocido",
    "6" = "Desconocido"
  )) +
  theme_minimal()
```

```
Picking joint bandwidth of 49800
```

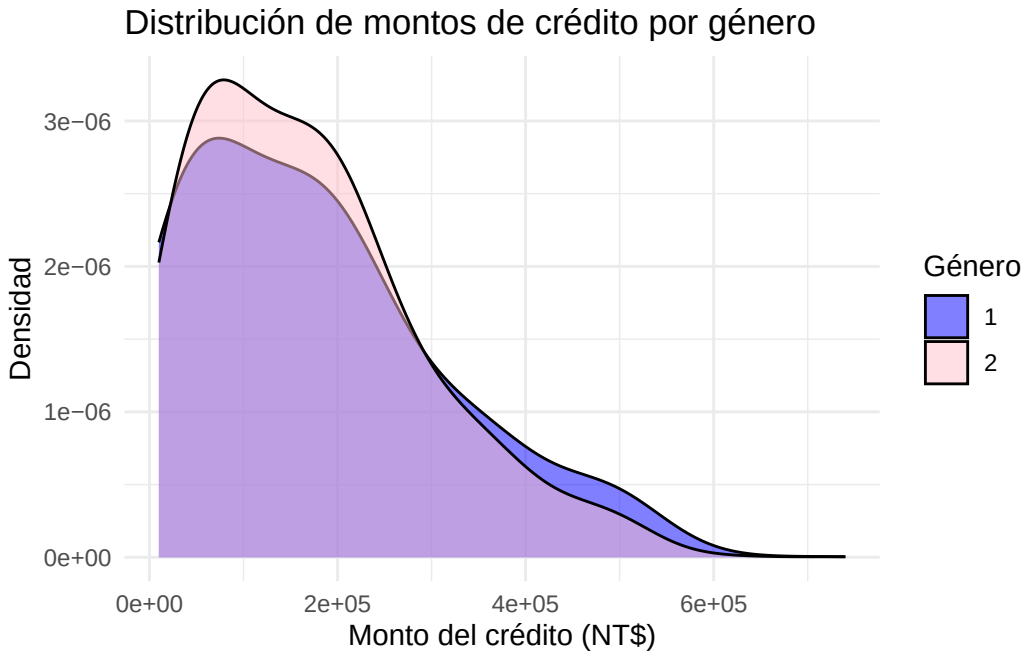


Las densidades que deben llamar nuestra atención son la de Secundaria, Universidad y Posgrado, esto por el hecho de que ellas son las que tienen la mayor concentración de datos.Además, veamos que las densidades tienen más o menos una distribución exponencial.

Por otro lado, observemos las densidades del límite de crédito con respecto al género.

```
library(ggplot2)
```

```r
ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(SEX))) +
  geom_density(adjust = 2, alpha = 0.5) +
  labs(
    x = "Monto del crédito (NT$)",
    y = "Densidad",
    title = "Distribución de montos de crédito por género",
    fill = "Género"
  ) +
  scale_fill_manual(values = c("1" = "blue", "2" = "pink")) +
  theme_minimal()
```

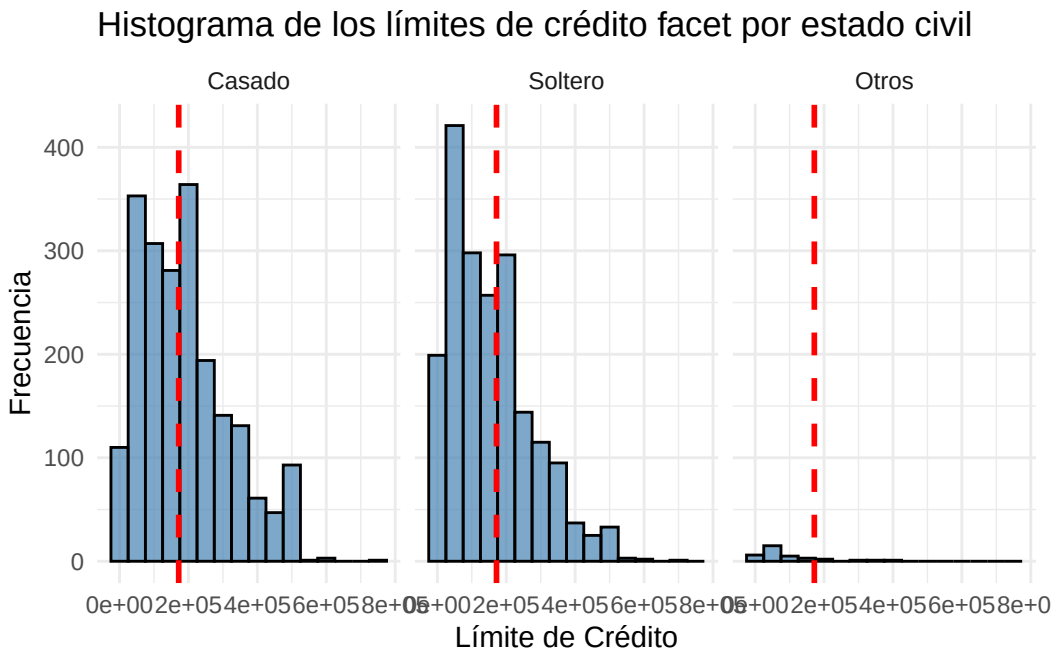


Del gráfico anterior, podemos determinar que a menores montos las mujeres tienen más crédito.

El siguiente gráfico es un histograma del límite de crédito con respecto al estado civil de las personas.

```r
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
```

```
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito facet por estado civil",
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~MARRIAGE, labeller = as_labeller(c("1" = "Casado", "2" = "Soltero", "3" = "Otro
  theme_minimal()
```

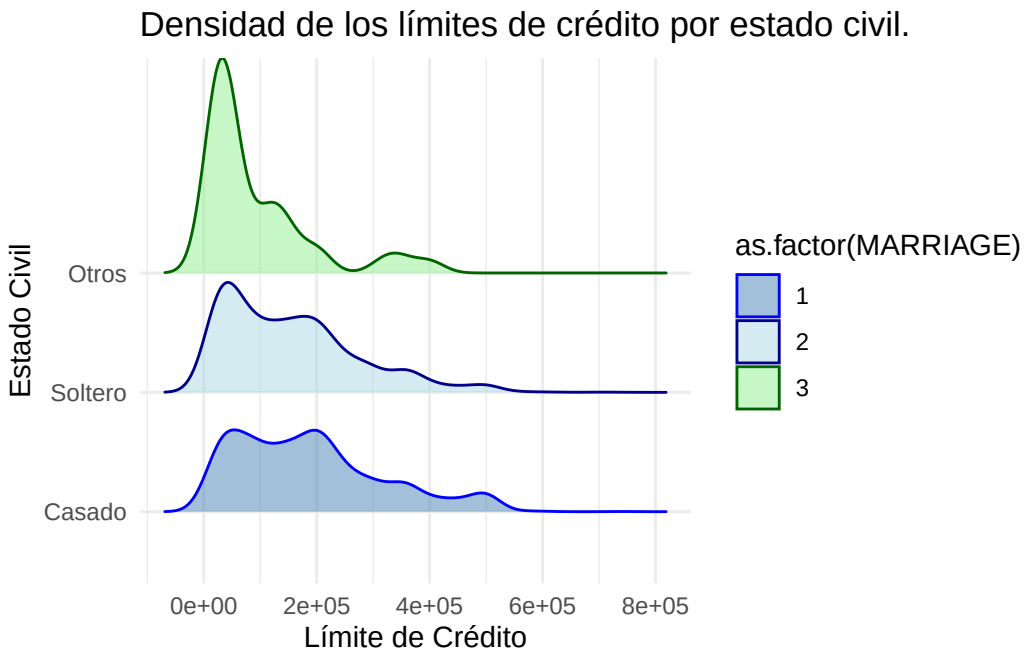


Agregamos densidades de las respectivas variables.

```
library(ggplot2)
library(dplyr)
library(ggridges)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(MARRIAGE), fill = as.factor(MARRIAGE), color = as.
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por estado civil.",
       x = "Límite de Crédito",
       y = "Estado Civil") +
  scale_fill_manual(values = c("1" = "steelblue", "2" = "lightblue", "3" = "lightgreen")) +
```

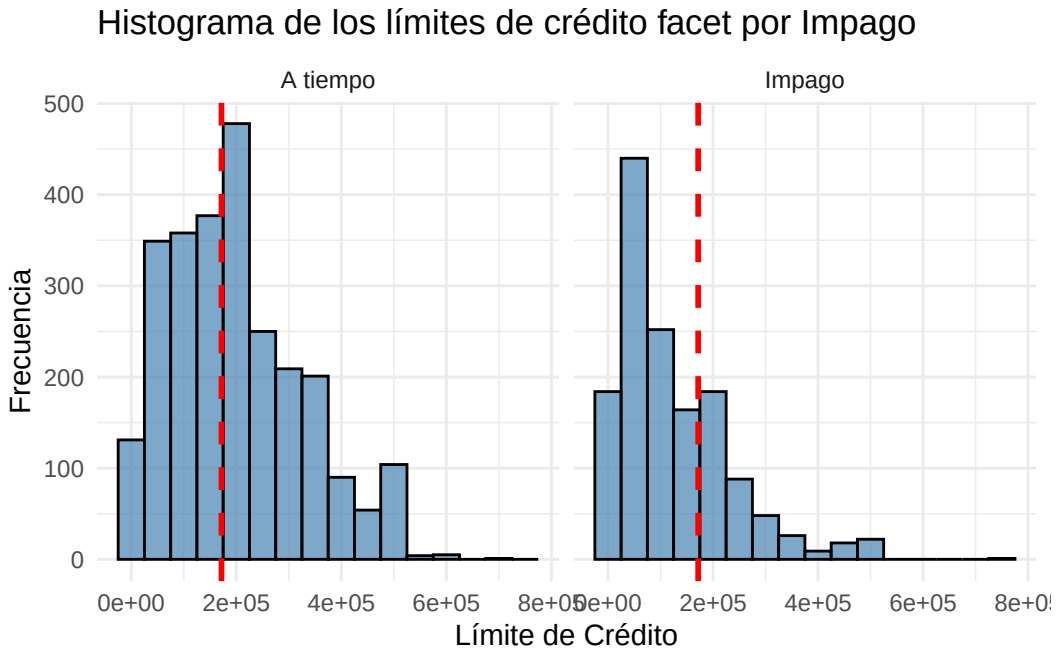
```
  scale_color_manual(values = c("1" = "blue", "2" = "darkblue", "3" = "darkgreen")) +
  scale_y_discrete(labels = c(
    "1" = "Casado",
    "2" = "Soltero",
    "3" = "Otros"
  )) +
  theme_minimal()
```

```
Picking joint bandwidth of 26300
```




```
library(ggplot2)
library(dplyr)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL)) +
  geom_histogram(binwidth = 50000, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_vline(aes(xintercept = mean(LIMIT_BAL, na.rm = TRUE)),
             color = "red",
             linetype = "dashed",
             size = 1) +
  labs(title = "Histograma de los límites de crédito facet por Impago",
```

```
       x = "Límite de Crédito",
       y = "Frecuencia") +
  facet_wrap(~default.payment.next.month, labeller = as_labeller(c("0" = "A tiempo", "1" = "
  theme_minimal()
```

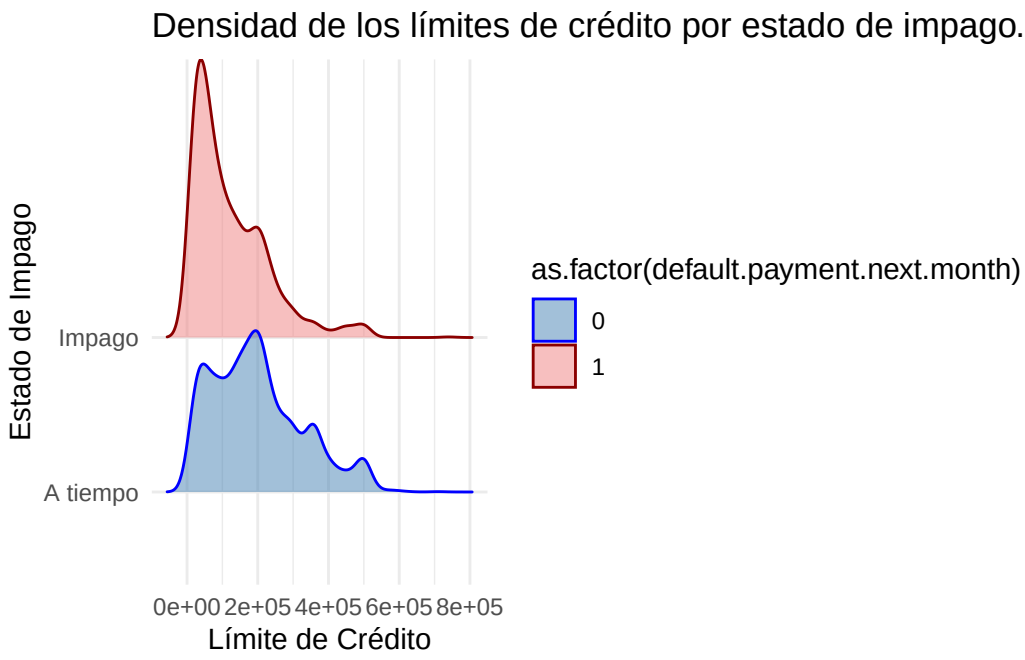


Agregamos dicho gráfico de densidades para las anteriores variables.

```
library(ggplot2)
library(dplyr)
library(ggridges)

data_credit %>%
  ggplot(aes(x = LIMIT_BAL, y = as.factor(default.payment.next.month), fill = as.factor(defau
  geom_density_ridges(alpha = 0.5) +
  labs(title = "Densidad de los límites de crédito por estado de impago.",
       x = "Límite de Crédito",
       y = "Estado de Impago") +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  scale_color_manual(values = c("0" = "blue", "1" = "darkred")) +
  scale_y_discrete(labels = c(
    "0" = "A tiempo",
    "1" = "Impago"
```

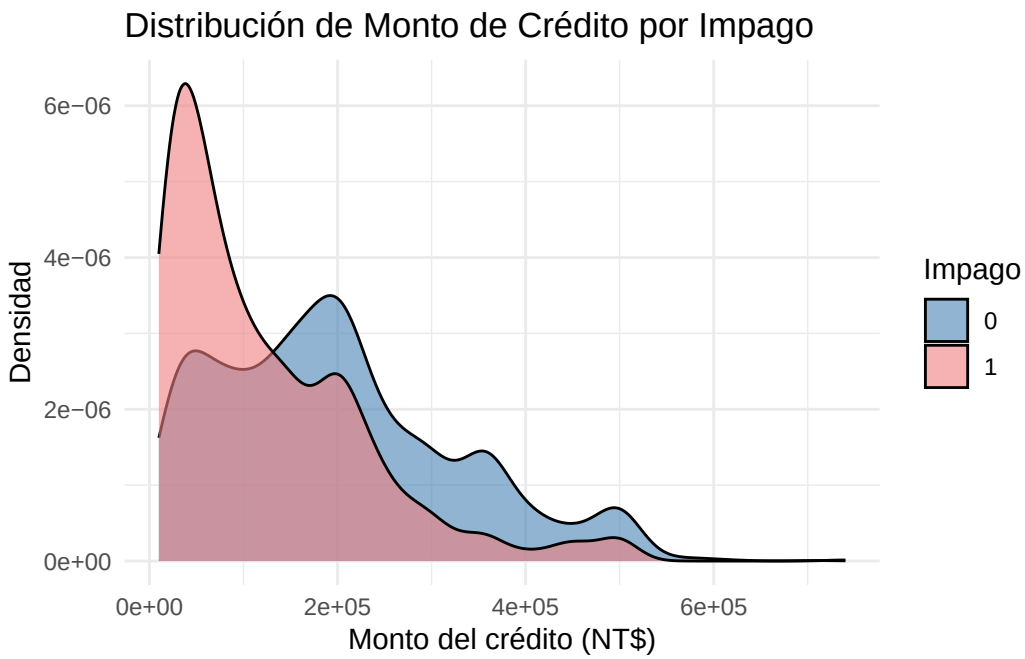
```
  )) +
  theme_minimal()
```

```
Picking joint bandwidth of 22100
```



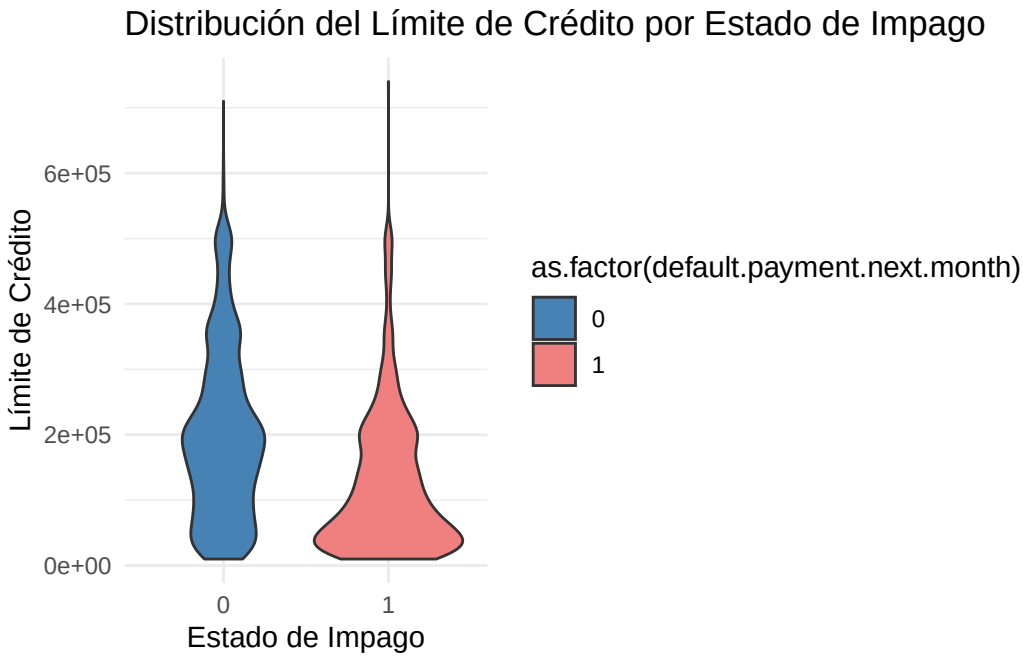
Ploteando los gráficos juntos, para una mejor visualización tenemos que:

```
ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(default.payment.next.month))) +
  geom_density(alpha = 0.6) +
  labs(
    x = "Monto del crédito (NT$)",
    y = "Densidad",
    title = "Distribución de Monto de Crédito por Impago",
    fill = "Impago"
  ) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  theme_minimal()
```



Utilizando una gráfica de violín, veamos la distribución para compararlas visualmente.

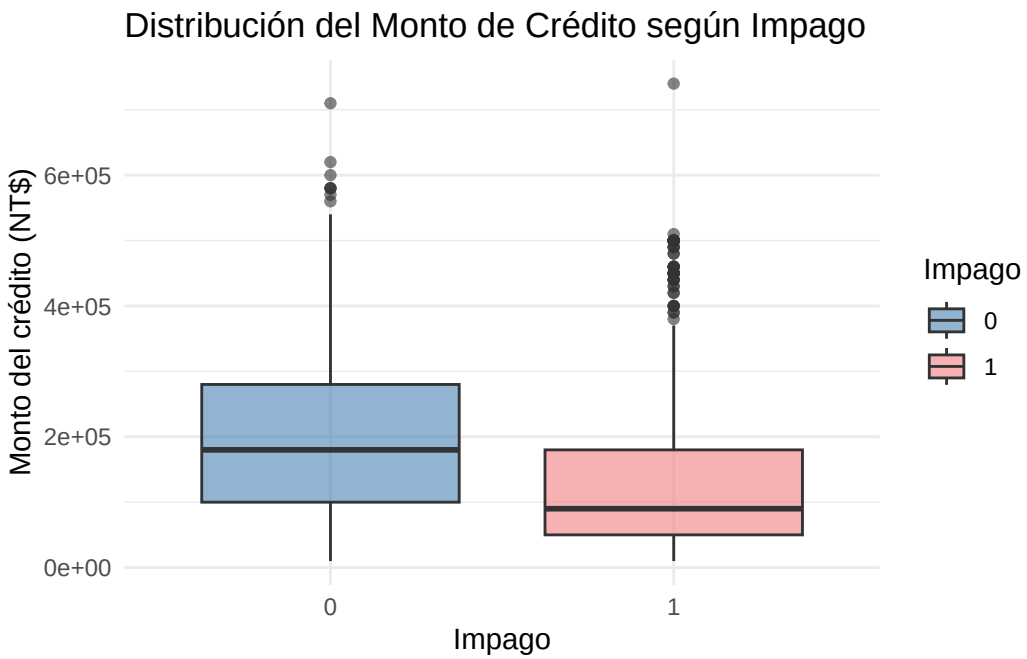
```
ggplot(data_credit, aes(x = as.factor(default.payment.next.month), y = LIMIT_BAL, fill = as.
  geom_violin() +
  labs(title = "Distribución del Límite de Crédito por Estado de Impago", x = "Estado de Impa
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  theme_minimal()
```



Por el gráfico de la izquierda, podemos observar como a menores límite de créditos hay una concentración de las personas que caen en impago, reduciendose conforme el límite de crédito es mayor. Note además que un gráfico de violín al final son las densidades reflejadas, podemos obtener la misma información de un gráfico de violín que de un gráfico de densidades.

luego utilizando un diagrama de cajas para hacer otra comparación.

```
library(ggplot2)

ggplot(data_credit, aes(x = as.factor(default.payment.next.month), y = LIMIT_BAL, fill = as.
  geom_boxplot(alpha = 0.6) +
  labs(
    x = "Impago",
    y = "Monto del crédito (NT$)",
    title = "Distribución del Monto de Crédito según Impago",
    fill = "Impago"
  ) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  theme_minimal()
```

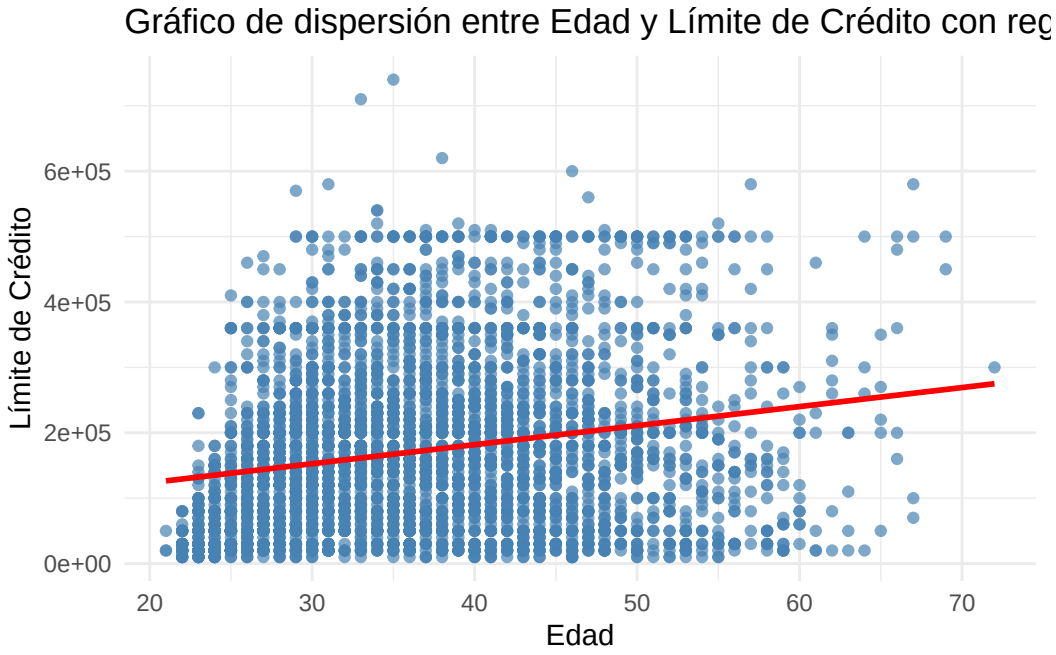


Del gráfico anterior entonces podemos observar que en promedio las personas que pagan a tiempo según el monto de crédito es mayor que el de las personas que cae en impago, lo cual nos ayudará a determinar más adelante si este factor es de importancia a la hora del riesgo de pago.

Por último veamos cómo de distribuyen las variables AGE y LIMIT_BAL. Usaremos un gráfico de dispersión

```
library(ggplot2)

ggplot(data_credit, aes(x = AGE, y = LIMIT_BAL)) +
  geom_point(color = "steelblue", alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Gráfico de dispersión entre Edad y Límite de Crédito con regresión lineal",
       x = "Edad",
       y = "Límite de Crédito") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

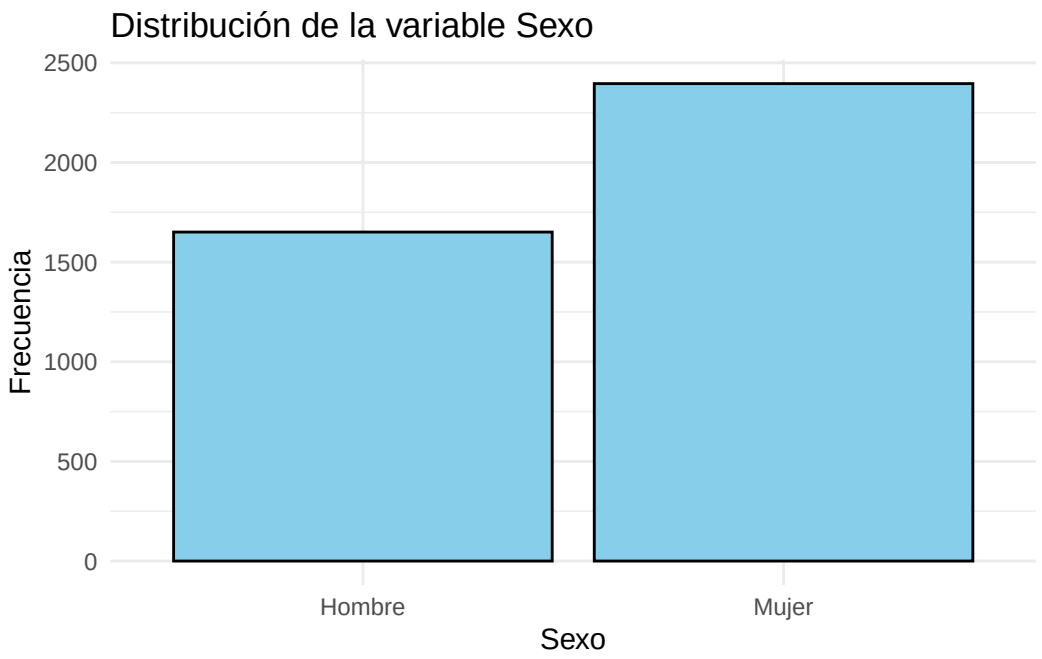


Cuando realizamos la matriz de correlación, vimos que la correlación entre estas dos variables es de 0.21, lo cual es algo bajo, podemos entonces ver esa tendencia en este gráfico, pues tenemos muchos puntos dispersos, al realizar una regresión lineal, podemos ver la linea que mejor se ajusta a estos puntos. Podemos inferir que hay una relación positiva débil entre las variables. Aunque un valor de 0.21 no ayuda a predecir qué pasaría cuando las variables aumenten.

Gráficos de Variables Categóricas.

Comenzaremos esta sección realizando gráficos de barras, con la intención de ver las frecuencias de las variables.

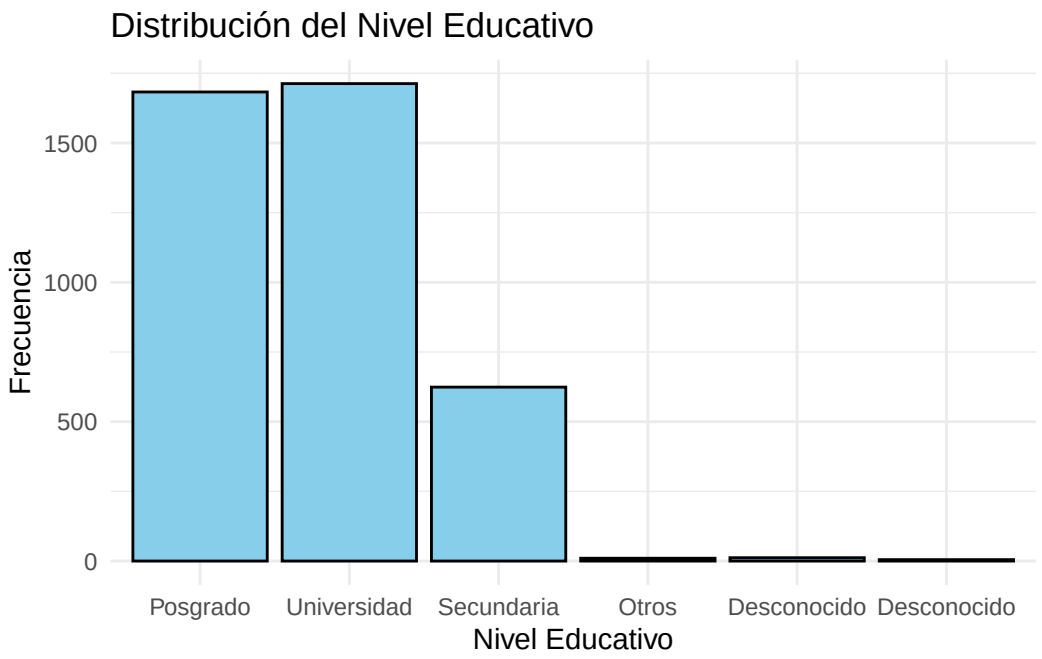
```
library(ggplot2)


ggplot(data = data_credit, aes(x = as.factor(SEX))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución de la variable Sexo", x = "Sexo", y = "Frecuencia") +
  scale_x_discrete(labels = c("1" = "Hombre", "2" = "Mujer")) +
  theme_minimal()
```



Con esto podemos ver que tenemos más datos de mujeres que de hombres.

Veamos como se comporta la variable de educación.

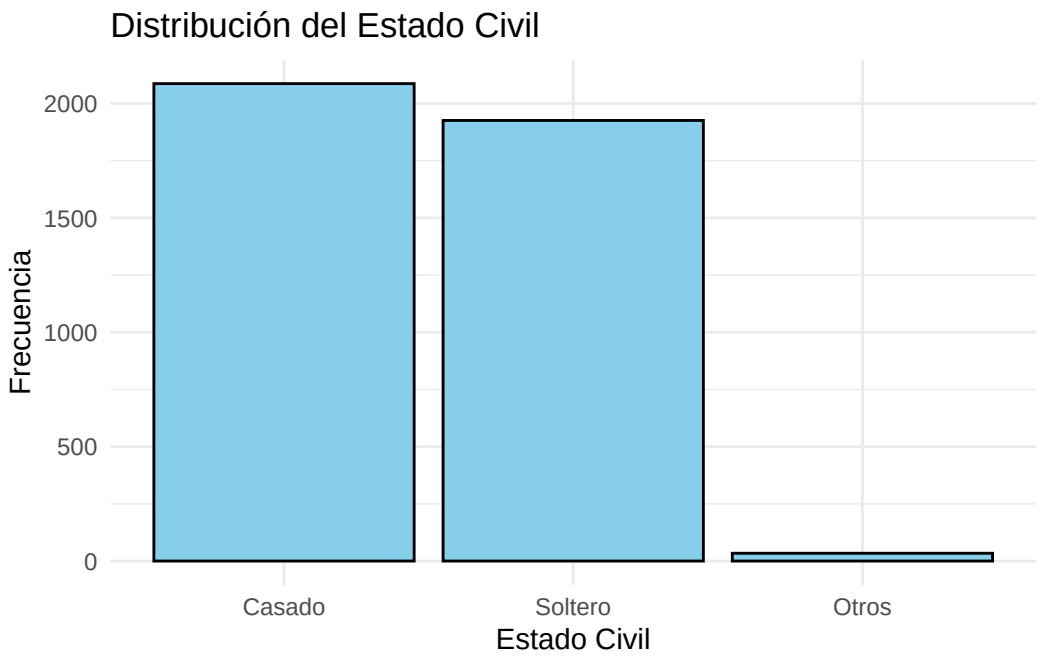
```
library(ggplot2)

ggplot(data = data_credit, aes(x = as.factor(EDUCATION))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución del Nivel Educativo", x = "Nivel Educativo", y = "Frecuencia") +
  scale_x_discrete(labels = c("1" = "Posgrado",
                              "2" = "Universidad",
                              "3" = "Secundaria",
                              "4" = "Otros",
                              "5" = "Desconocido",
                              "6" = "Desconocido")) +
  theme_minimal()
```



Observamos entonces que nuestra base de datos contiene más información de personas que están posgrados o que terminaron la universidad.

Luego para la variable de estado civil.

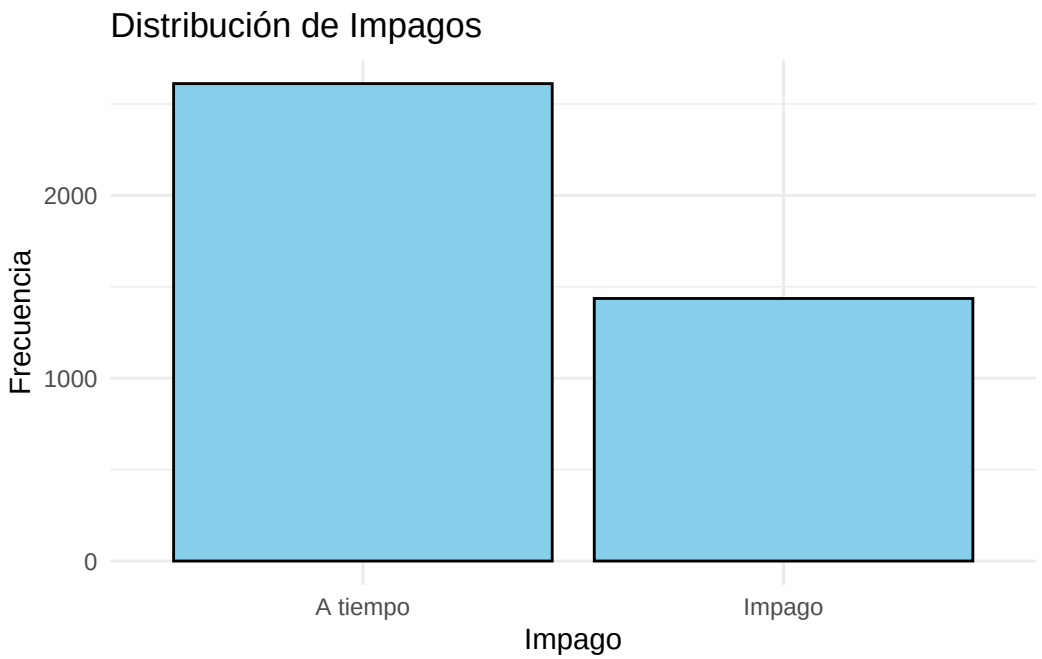
```
library(ggplot2)

ggplot(data = data_credit, aes(x = as.factor(MARRIAGE))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución del Estado Civil", x = "Estado Civil", y = "Frecuencia") +
  scale_x_discrete(labels = c("1" = "Casado",
                              "2" = "Soltero",
                              "3" = "Otros")) +
  theme_minimal()
```



Al igual que antes, tenemos más información de las personas que están casadas y de las que están solteras.

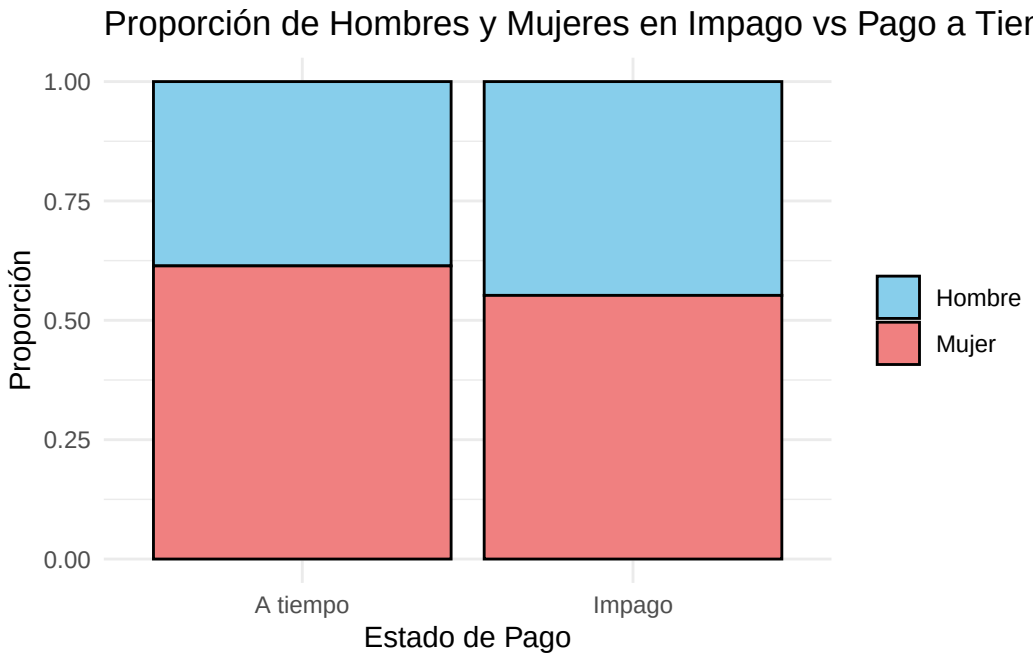
Por último vamos a ver la gráfica de barras de la variable de interés, la cual es si cayó en impago o no lo hizo.

```
library(ggplot2)

ggplot(data = data_credit, aes(x = as.factor(default.payment.next.month))) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribución de Impagos", x = "Impago", y = "Frecuencia") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  theme_minimal()
```



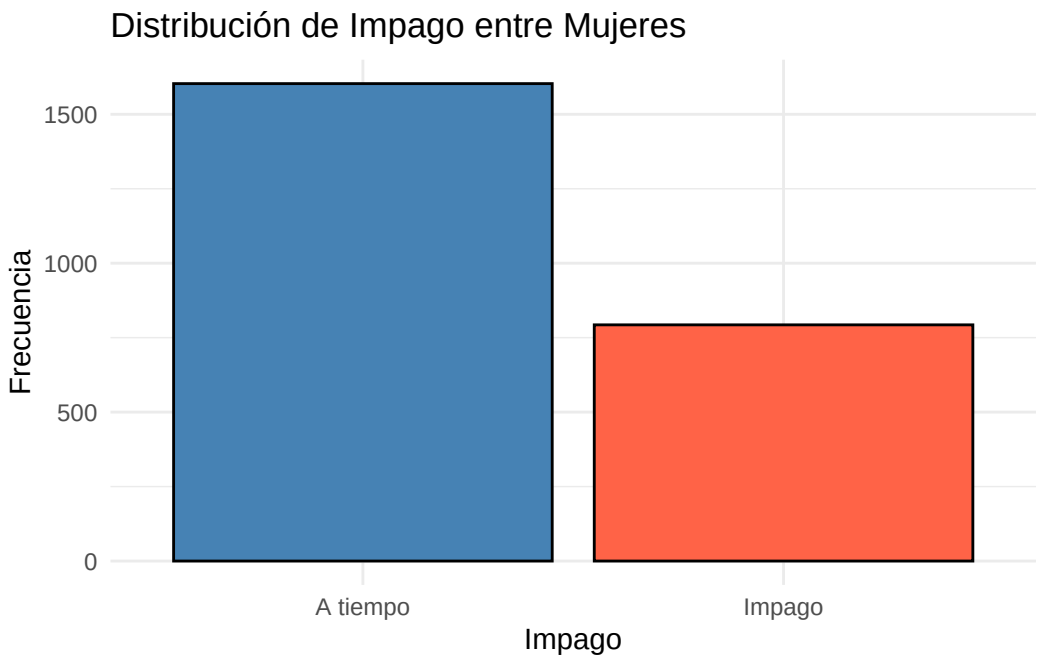
Con esto terminamos los gráficos aislados de las variables categóricas y damos inicio a ver cómo se distribuyen cuando las relacionamos.

```
library(ggplot2)
library(dplyr)

ggplot(data = data_credit, aes(x = as.factor(default.payment.next.month), fill = as.factor(SE
  geom_bar(position = "fill", color = "black") +
  labs(title = "Proporción de Hombres y Mujeres en Impago vs Pago a Tiempo",
       x = "Estado de Pago", y = "Proporción") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  scale_fill_manual(labels = c("1" = "Hombre", "2" = "Mujer"), values = c("skyblue", "lightco
  theme_minimal() +
  theme(legend.title = element_blank())
```



Del gráfico anterior, podemos observar entonces que de las personas que cayeron en impago, la mayoría son mujeres, al menos más del 50%, sin embargo, veamos de manera aislada esto.

```
library(ggplot2)
library(dplyr)

data_credit %>%
  filter(SEX == 2) %>%
  ggplot(aes(x = as.factor(default.payment.next.month), fill = as.factor(default.payment.next
  geom_bar(stat = "count", color = "black") +
  labs(title = "Distribución de Impago entre Mujeres",
       x = "Impago",
       y = "Frecuencia",
       fill = "Impago") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "tomato")) +
  theme_minimal() +
  theme(legend.position = "none")
```



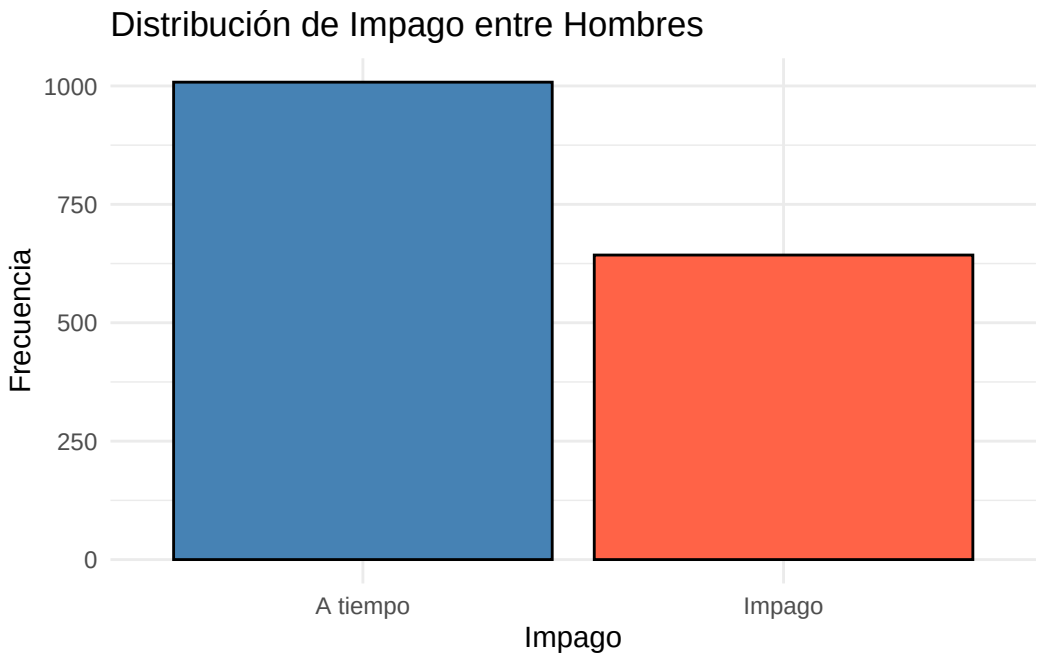
Numéricamente esto es:

```
library(dplyr)

mujeres_impago <- data_credit %>%
  filter(SEX == 2) %>%  # Filtramos los datos, porque nos interesan solo las mujeres
  summarise(
    total_mujeres = n(),
    mujeres_impago = sum(default.payment.next.month == 1)  # Número de mujeres en impago
  ) %>%
  mutate(porcentaje_impago = mujeres_impago / total_mujeres * 100)  # Calculamos el porcentaj

# Mostramos el resultado
mujeres_impago
```

```
# A tibble: 1 x 3
  total_mujeres mujeres_impago porcentaje_impago
          <int>          <int>             <dbl>
1          2396            793              33.1
```

Con esto podemos observar que de las mujeres totales, solo el 33% cayó en impago. Haremos un análisis similar con respecto a los hombre.

```r
library(ggplot2)
library(dplyr)

data_credit %>%
  filter(SEX == 1) %>%
  ggplot(aes(x = as.factor(default.payment.next.month), fill = as.factor(default.payment.next
  geom_bar(stat = "count", color = "black") +
  labs(title = "Distribución de Impago entre Hombres",
       x = "Impago",
       y = "Frecuencia",
       fill = "Impago") +
  scale_x_discrete(labels = c("0" = "A tiempo", "1" = "Impago")) +
  scale_fill_manual(values = c("0" = "steelblue", "1" = "tomato")) +
  theme_minimal() +
  theme(legend.position = "none")
```



Numéricamente podemos observar que:

```r
library(dplyr)

hombres_impago <- data_credit %>%
  filter(SEX == 1) %>%
```

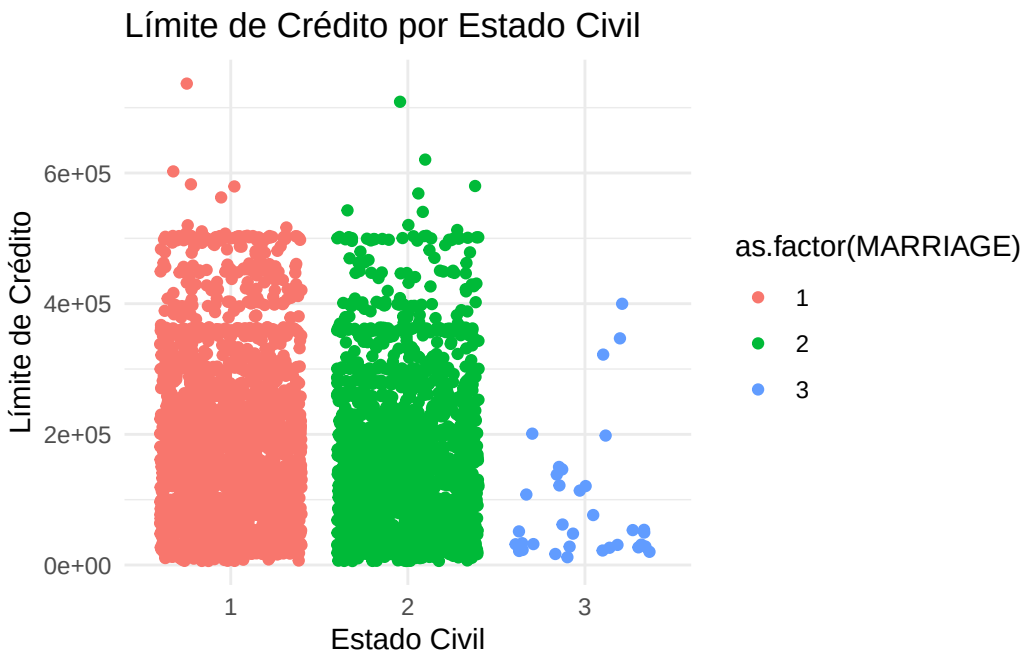
```r
  summarise(
    total_hombres = n(),
    hombres_impago = sum(default.payment.next.month == 1)
  ) %>%
  mutate(porcentaje_impago = hombres_impago / total_hombres * 100)

hombres_impago
```

```
# A tibble: 1 x 3
  total_hombres hombres_impago porcentaje_impago
          <int>          <int>             <dbl>
1          1651            643              38.9
```

Con esto observamos que el porcentaje de los hombres que cayeron en impago, aunque es por poco, es mayor que el de las mujeres que cayeron en impago. Esto lo hicimos porque anteriormente se estaban comparando magnitudes que no se podían comparar, con los porcentajes podemos determinar que relativamente, los hombres tienden a caer más en impago que las mujeres, al menos eso podemos inferrir gracias a la evidencia de los datos.

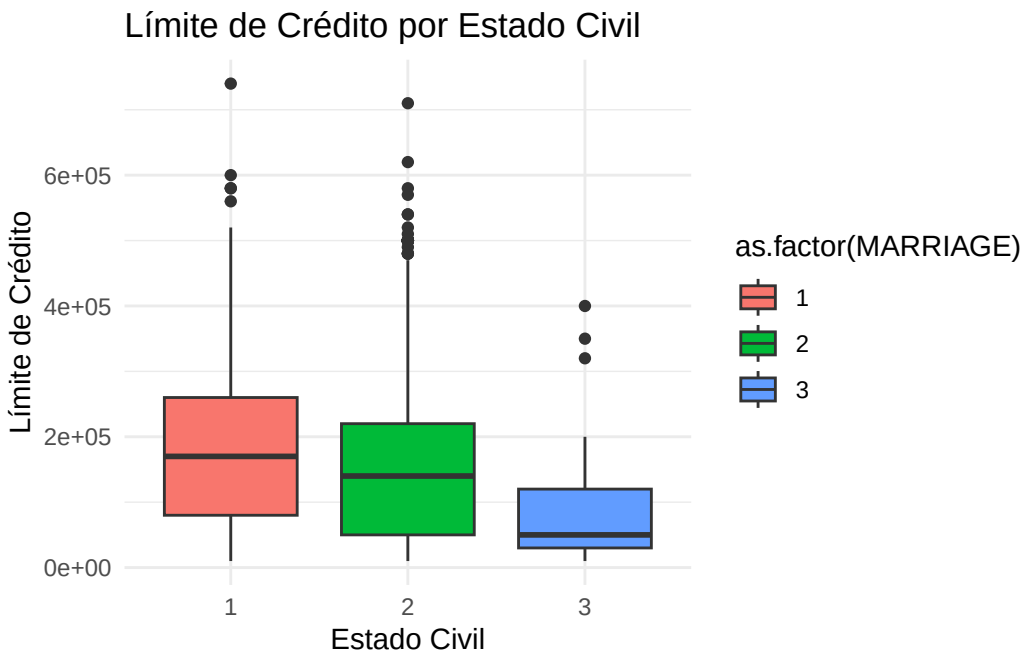
Ahora vamos a visualizar el crédito con respecto a la variable de MARRIAGE.

```r
library(ggplot2)

ggplot(data_credit, aes(x = as.factor(MARRIAGE), y = LIMIT_BAL, color = as.factor(MARRIAGE)))
  geom_jitter() +
  labs(title = "Límite de Crédito por Estado Civil", x = "Estado Civil", y = "Límite de Crédi
  theme_minimal()
```



Después analizaremos la relación entre las variables, ya que son variables de diferente naturaleza, es decir una categórica y una numérica, por lo que utilizaremos un análisis ANOVA para clarificar si las diferencias se deben al azar o si la evidencia estadística indican que están relacionados. Por el momento, haremos una comparación con gráficos de cajas, para observar de manera gráfica, como se siguen comportando.

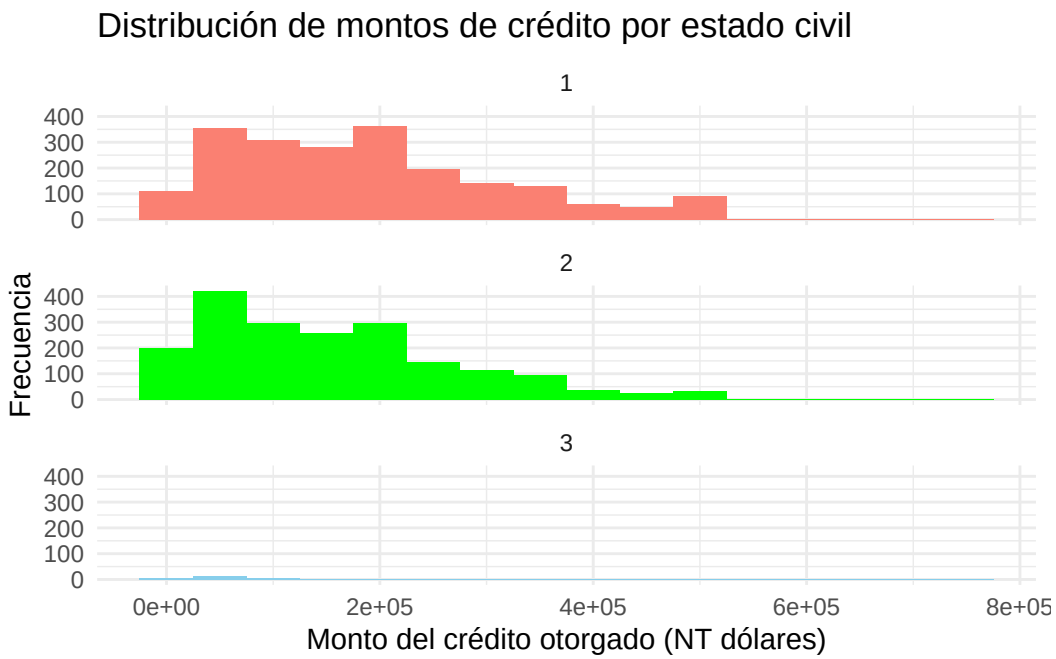
```
library(ggplot2)

ggplot(data_credit, aes(x = as.factor(MARRIAGE), y = LIMIT_BAL, fill = as.factor(MARRIAGE)))
  geom_boxplot() +
  labs(title = "Límite de Crédito por Estado Civil", x = "Estado Civil", y = "Límite de Crédi
  theme_minimal()
```



Podemos observar gráficamente que las diferencias no son tan significativas, sin embargo esto es una conjetura, nos ajustaremos a un análisis estadístico más riguroso en posteriores secciones. Por otro lado, podemos observar que el nivel 3 difiere de los otros niveles, esto se puede deber a las bajas observaciones que tenemos en este nivel.

Por último para tener la relación entre estas variables, observemos el siguiente gráfico.

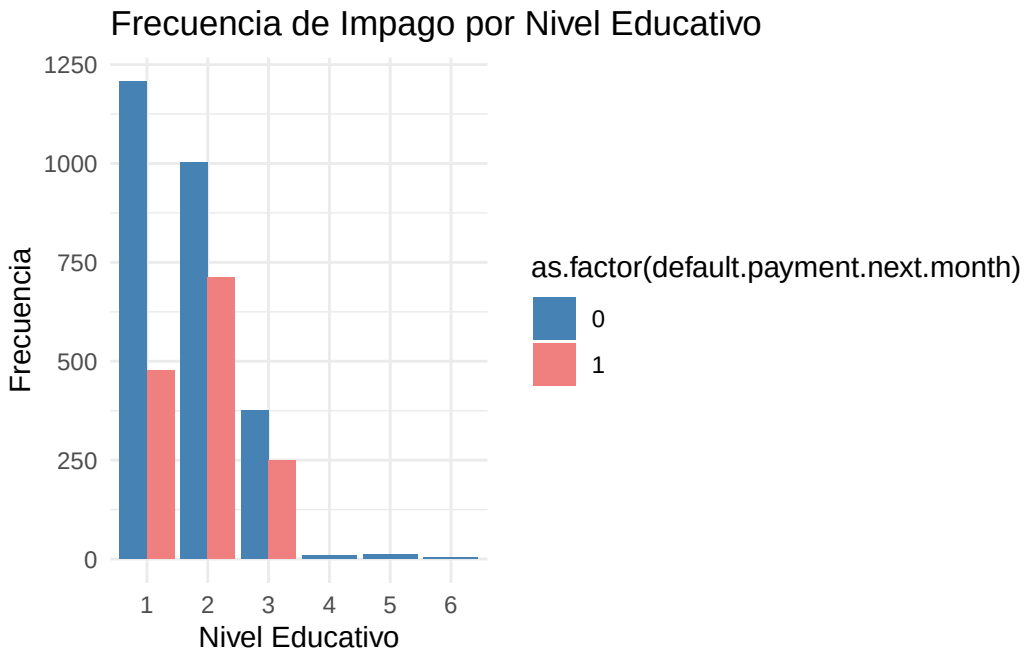
```
library(ggplot2)

ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(MARRIAGE))) +
  geom_histogram(binwidth = 50000) +
  facet_wrap(~MARRIAGE, nrow = 3) +
  labs(
    x = "Monto del crédito otorgado (NT dólares)",
    y = "Frecuencia",
    title = "Distribución de montos de crédito por estado civil",
    fill = "Estado civil"
  ) +
  scale_fill_manual(values = c("1" = "salmon", "2" = "green", "3" = "skyblue")) +
  theme_minimal() +
  theme(legend.position = "none")
```



Como hemos mencionado, la distribución de las variables casado y soltero se parecen mucho visualmente, no podemos decir más del estado "otro".

Analizamos ahora la frecuencia de impago en relación con el nivel educativo.

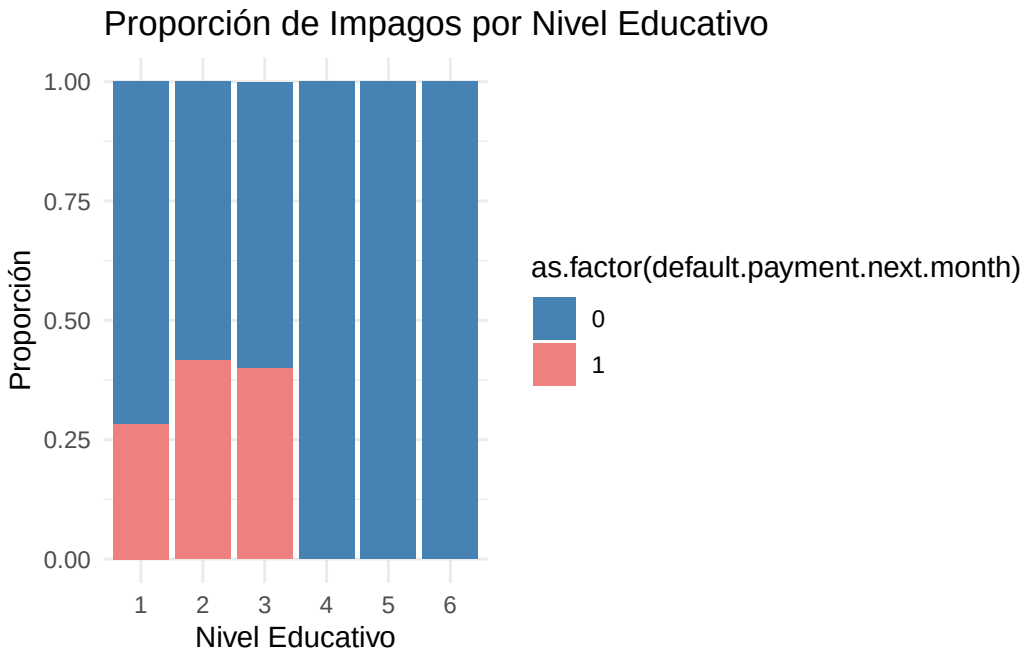
```
ggplot(data_credit, aes(x = as.factor(EDUCATION), fill = as.factor(default.payment.next.month
  geom_bar(position = "dodge") +
  labs(title = "Frecuencia de Impago por Nivel Educativo", x = "Nivel Educativo", y = "Frecue
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  theme_minimal()
```



Recordemos que la etiqueta 2 equivale a las personas que están en grado de haber terminado o concluido la universidad. Con este gráfico podemos interpretar que las personas que terminaron la universidad tienen una alta proporción de haber caido en impago, esto se puede deber a créditos estudiantiles y la dificultad de conseguir empleo, sin embargo esto es una conjetura y no vamos analizar esta consecuencia, ya que solo nos importa ver qué están diciendo nuestros datos.

Con el siguiente gráfico queremos observar la proporción de las personas que caen en impago, según el nievl educativo.

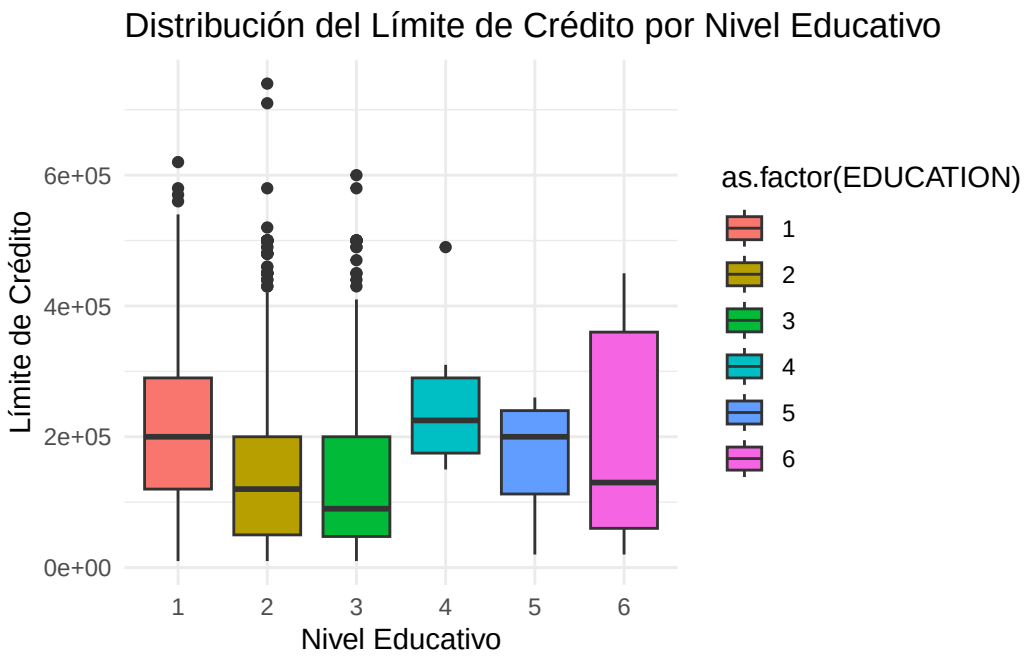
```
ggplot(data_credit, aes(x = as.factor(EDUCATION), fill = as.factor(default.payment.next.month
  geom_bar(position = "fill") +
  labs(title = "Proporción de Impagos por Nivel Educativo", x = "Nivel Educativo", y = "Propo
  scale_fill_manual(values = c("0" = "steelblue", "1" = "lightcoral")) +
  theme_minimal()
```



Justamente, este gráfico refleja que las personas que están en el nivel de universidad presentan una mayor proporción de impago.

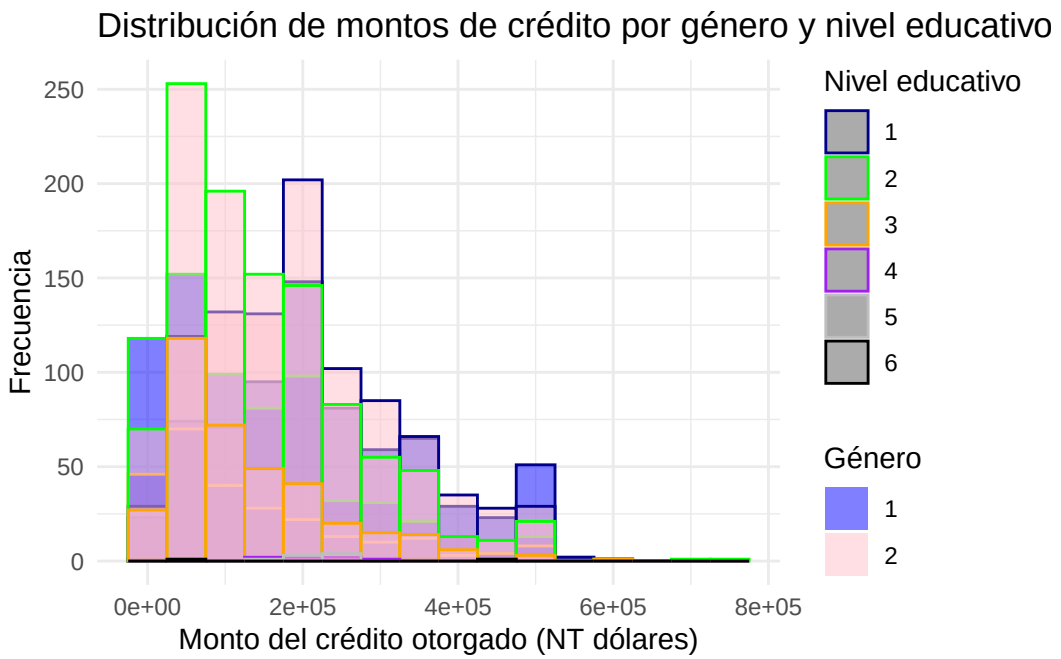
Por otro lado, vamos analizar gráficamente la relación de la variable educación con el límite de crédito.

```
ggplot(data_credit, aes(x = as.factor(EDUCATION), y = LIMIT_BAL, fill = as.factor(EDUCATION)
  geom_boxplot() +
  labs(title = "Distribución del Límite de Crédito por Nivel Educativo", x = "Nivel Educativ
  theme_minimal()
```



Observando las medias, inferimos que en promedio el límite de crédito otorgado a las personas que están en un nivel de posgrado es mayor que las personas que tiene solo universidad o secundaria. A su vez, la media de Universidad es mayor que la media de secundaria. Por la calidad de los datos de la base de datos, no podemos analizar con lujo de detalle los niveles 3,4 y 5.

```
library(ggplot2)

ggplot(data_credit, aes(x = LIMIT_BAL, fill = as.factor(SEX), color = as.factor(EDUCATION)))
  geom_histogram(binwidth = 50000, alpha = 0.5, position = "identity") +
  labs(
    x = "Monto del crédito otorgado (NT dólares)",
    y = "Frecuencia",
    title = "Distribución de montos de crédito por género y nivel educativo",
    fill = "Género",
    color = "Nivel educativo"
  ) +
  scale_fill_manual(values = c("1" = "blue", "2" = "pink")) +  # Colores para sexo
  scale_color_manual(values = c("1" = "darkblue", "2" = "green", "3" = "orange", "4" = "purpl
  theme_minimal()
```



Con el gráfico anterior podemos ver la relación existente entre el nivel educativo, el género y el monto del crédito otorgado, es decir, 3 variables graficadas.

Análisis Matemático de las correlaciones

Para esta sección, primero vamos analizar tablas de contingencias, esto con el fin de eencontrar relaciones entre las variables, utilizaremos las pruebas chi-cuadrado y la prueba ” — “, para determinar esta relaciones entre variables categóricas.

Posteriormente, utilizaremos los índices de correlación obtenidos en la matriz de correlación para el análisis de la base de datos y con ello apoyarnos en la evidencia teórica que existe.

Tablas de Contigencias y p-values

Para esta sección vamos hacer tablas de contigencias, esto con el objetivo de buscar las relaciones que tienen las variables categóricas.

```
library(gmodels)

# Realizamos la tabla de contingencia
CrossTable(data_credit$EDUCATION, data_credit$default.payment.next.month, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  4047


                     | data_credit$default.payment.next.month
data_credit$EDUCATION |         0 |         1 | Row Total |
----------------------|-----------|-----------|-----------|
                    1 |      1207 |       476 |      1683 |
                      |     0.717 |     0.283 |     0.416 |
                      |     0.462 |     0.331 |           |
                      |     0.298 |     0.118 |           |
----------------------|-----------|-----------|-----------|
                    2 |      1002 |       711 |      1713 |
                      |     0.585 |     0.415 |     0.423 |
                      |     0.384 |     0.495 |           |
                      |     0.248 |     0.176 |           |
----------------------|-----------|-----------|-----------|
                    3 |       375 |       249 |       624 |
                      |     0.601 |     0.399 |     0.154 |
                      |     0.144 |     0.173 |           |
                      |     0.093 |     0.062 |           |
----------------------|-----------|-----------|-----------|
                    4 |        10 |         0 |        10 |
                      |     1.000 |     0.000 |     0.002 |
                      |     0.004 |     0.000 |           |
                      |     0.002 |     0.000 |           |
----------------------|-----------|-----------|-----------|
                    5 |        12 |         0 |        12 |
                      |     1.000 |     0.000 |     0.003 |
                      |     0.005 |     0.000 |           |
                      |     0.003 |     0.000 |           |
----------------------|-----------|-----------|-----------|
                    6 |         5 |         0 |         5 |
```

```
                     |     1.000 |     0.000 |     0.001 |
                     |     0.002 |     0.000 |           |
                     |     0.001 |     0.000 |           |
---------------------|-----------|-----------|-----------|
         Column Total |      2611 |      1436 |      4047 |
                     |     0.645 |     0.355 |           |
---------------------|-----------|-----------|-----------|
```

De la tabla anterior podemos ver entonces las relaciones que hay entre las variables, por ejemplo, podemos ver que de las personas de educación, que pertencen al nivel de posgrado, un 71% de esas personas no imcuplieron su pago, es decir pagaron a tiempo. Y un 29% de esas personas si cayeron en impago.

Haremos la prueba Exacta de Fisher para realizar una Prueba de Hipótesis donde nuestra hipótesis nula es. H_0, no existe correlación entre las variables, y donde nuestra hipótesis alternativa, H_1, es que hay relación entre las variables.

```r
tabla_education <- table(data_credit$EDUCATION, data_credit$default.payment.next.month)

# Realizar la prueba exacta de Fisher con simulación de Monte Carlo
set.seed(2024)  # Fijamos la semilla
fisher_test_education <- fisher.test(tabla_education, simulate.p.value = TRUE, B = 10000)

print(fisher_test_education)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    10000 replicates)

data:  tabla_education
p-value = 9.999e-05
alternative hypothesis: two.sided
```

Gracias al test anterior podemos inferir entonces que existe una dependencia de las variables. Rechazamos la hipótesis nula, hay evidencia estadística suficiente para decir que las variables tienen una relación.

Más explicado aún, el p-value que obtuvimos fue de 0.0001, lo que es mucho más pequeño que el 5% del nivel de significancia, por lo que entonces podemos rechazar la hipótesis nula.

Por otro lado, ahora vamos a comparar las variables de sexo y de impago, esto con el objetivo de observar si el sexo influye o tiene relación con la probabilidad de caer en impago.

```r
library(gmodels)

# Realizamos la tabla de contingencia
CrossTable(data_credit$SEX, data_credit$default.payment.next.month, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  4047


                | data_credit$default.payment.next.month
data_credit$SEX |         0 |         1 | Row Total |
----------------|-----------|-----------|-----------|
              1 |      1008 |       643 |      1651 |
                |     0.611 |     0.389 |     0.408 |
                |     0.386 |     0.448 |           |
                |     0.249 |     0.159 |           |
----------------|-----------|-----------|-----------|
              2 |      1603 |       793 |      2396 |
                |     0.669 |     0.331 |     0.592 |
                |     0.614 |     0.552 |           |
                |     0.396 |     0.196 |           |
----------------|-----------|-----------|-----------|
   Column Total |      2611 |      1436 |      4047 |
                |     0.645 |     0.355 |           |
----------------|-----------|-----------|-----------|
```

Con la tabla vemos la distribución de las variables entre ellas. Aplicamos la prueba de Fisher con simulación al igual que en el caso anterior.

```r
tabla_sex <- table(data_credit$SEX, data_credit$default.payment.next.month)

# Realizar la prueba exacta de Fisher con simulación de Monte Carlo
set.seed(2024)
fisher_test_sex <- fisher.test(tabla_sex, simulate.p.value = TRUE, B = 10000)

print(fisher_test_sex)
```

```
    Fisher's Exact Test for Count Data

data:  tabla_sex
p-value = 0.0001387
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6792443 0.8856355
sample estimates:
odds ratio
 0.7755741
```

Gracias a la prueba anterior, obtenemos que el p-value es de 0.0001387, siendo bastante menor que el valor de significancia, el cual es de 5%, por lo que hay suficiente evidencia estadística para rechazar la hipótesis nula, así decimos entonces que existe una cierta dependencia entre estas variables.

Vamos a ver ahora como se comporta la variable MARRIAGE con el impago. Veamos primero su tabla de contingencia.

```r
library(gmodels)

# Realizamos la tabla de contingencia
CrossTable(data_credit$MARRIAGE, data_credit$default.payment.next.month, prop.chisq = FALSE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
```

```
|-------------------------|


Total Observations in Table:  4047


                     | data_credit$default.payment.next.month
data_credit$MARRIAGE |         0 |         1 | Row Total |
---------------------|-----------|-----------|-----------|
                   1 |      1359 |       728 |      2087 |
                     |     0.651 |     0.349 |     0.516 |
                     |     0.520 |     0.507 |           |
                     |     0.336 |     0.180 |           |
---------------------|-----------|-----------|-----------|
                   2 |      1235 |       691 |      1926 |
                     |     0.641 |     0.359 |     0.476 |
                     |     0.473 |     0.481 |           |
                     |     0.305 |     0.171 |           |
---------------------|-----------|-----------|-----------|
                   3 |        17 |        17 |        34 |
                     |     0.500 |     0.500 |     0.008 |
                     |     0.007 |     0.012 |           |
                     |     0.004 |     0.004 |           |
---------------------|-----------|-----------|-----------|
        Column Total |      2611 |      1436 |      4047 |
                     |     0.645 |     0.355 |           |
---------------------|-----------|-----------|-----------|
```

Hagamos la misma prueba de hipótesis para determinar la relación de las variables.

```r
tabla_marriage <- table(data_credit$MARRIAGE, data_credit$default.payment.next.month)

# Realizar la prueba exacta de Fisher con simulación de Monte Carlo
set.seed(2024)
fisher_test_marriage <- fisher.test(tabla_marriage, simulate.p.value = TRUE, B = 10000)

print(fisher_test_marriage)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
```

```
    10000 replicates)

data:  tabla_marriage
p-value = 0.1634
alternative hypothesis: two.sided
```

En este caso, el valor de significancia que hemos estado utilizando es del 5%, es decir, un 0,05, como el p-value nos dió un valor de 0.1634, el p-value es mayor que el nivel de significancia, por lo que no hay evidencia estadística suficientes para rechazar la hipótesis nula, es decir, las diferencias que hemos encontrado en las categorías se pueden deber al azar. En conclusión, no podemos afirmar que el estado civil afecte al riesdo de impago.

Con esto terminamos la sección de las variables categóricas, procedemos entonces con las variables numéricas y la variable de impago.

Variables Numéricas con la Variable de Riesgo de Pago

Vamos a realizar una prueba de hipótesis entonces para las variables de LIMIT_BAL e Impago, para ello vamos a utilizar una prueba t no pareada, ya que estamos comparando poblaciones diferentes.

```r
t_test_result <- t.test(LIMIT_BAL ~ default.payment.next.month, data = data_credit)

print(t_test_result)
```

```
    Welch Two Sample t-test

data:  LIMIT_BAL by default.payment.next.month
t = 19.866, df = 3418.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 67451.30 82223.61
sample estimates:
mean in group 0 mean in group 1
       198249.7        123412.3
```

Este resultado no debería ser una sorpresa, pues gráficamente habíamos visto que a mayores límites de crédito, las personas tendían a incumplir menos que las personas que si lo hacían. Analicemos, el p-value tomó un valor demasiado pequeño en comparación con 0.05, por lo que hay evidencia estadística suficiente para rechazar la hipótesis nula. En conclusión, el límite de

crédito tiene relación con la probabilidad de impago, en general, las personas con mayor límite de crédito tienen una menor porbabilidad de caer en impago.

Haremos un análisis similar, pero esta vez cambiando el límite de crédito por la variable de edad.

```
t_test_result <- t.test(AGE ~ default.payment.next.month, data = data_credit)

print(t_test_result)
```

```
    Welch Two Sample t-test

data:  AGE by default.payment.next.month
t = -0.10677, df = 2734.6, p-value = 0.915
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -0.6408875  0.5746956
sample estimates:
mean in group 0 mean in group 1
       36.51704        36.55014
```

De la prueba anterior podemos inferir que la edad no influye en el incumplimiento de los pagos, obtuvimos un p-value de 0.915 lo que es mucho mayor que nuestro nivel de significancia del 5%, así no hay evidencia estadística suficiente para rechazar la hipótesis nula, entonces no rechazamos la hipótesis nula. Concluímos que la edad no influye dado los datos en la probabilidad de caer en impago.

Con esto concluímos el análisis estadístico de la base de datos.

Parte de Planificación

Parte de Escritura

Escribir, escribir, escribir

Conclusion

Introducción

Este estudio se enfoca en desarrollar un estudio en el cual se avale si variables como lo son la edad, el género, el estado marital y el nivel de educación, tienen algún peso o relevancia a

la hora de medir el riesgo financiero de los prestatarios. De manera que lo que se busca es identificar patrones y con ello obtener una visión mas completa que no se limite únicamente a las características financieras de las personas, como lo es su nivel de ingresos, sino que también considera el contexto y características propias en las que se desenvuelve cada individuo. Por ello, es que este estudio comparar variables cuantitativas con variables cualitativas y tratar de ver la relación o correlación que pueda existir entre ellas, ya que al final todo ello se relaciona intrínsecamente con aspectos propios del prestatario, lo cual a su vez ayuda a generar una perspectiva mas robusta del riesgo que el mismo puede generar para alguna entidad financiera llamase banco u otra. Por lo cual, la presente investigación tiene como finalidad lograr determinar si variables cualitativas tienen una verdadera relevancia a la hora de estimar el riesgo de impago de las personas, lo cual ha su vez llega a ser pertinente en lo que es la carrera de ciencias actuariales, por lo que se pudieron usar herramientas que irán viendo a lo largo de los siguientes cursos, pero que llegan a ser muy útiles para el propósito de este trabajo. De manera que el objetivo principal de esta investigación es determinar si existe una correlación positiva entre algunas variables cualitativas y factores cuantitativos que se utilizan para determinar el riesgo de impago, esto a través del uso de una base de datos de un banco de Taiwán la cual brinda suficiente información para nuestro propósito. Por lo cual, esta investigación se fundamenta en un marco teórico y empírico con el que se busca lograr determinar a partir de varios métodos y pruebas la existencia o no de correlación entre nuestras variables. A su vez, el marco teórico que se maneja para esta investigación parece muy pertinente, esto debido a que los autores desean ir más allá de los modelos tradicionales y explorar factores adicionales al contexto económico de los prestatarios, lo cual da una mayor perspectiva que permite deslumbrar elementos de riesgo que podrían llegar a pasar desapercibidos en estudios un poco más convencionales.

Resumen

Ordenamiento Final

Titulo

Análisis de variables cualitativas en relación al riesgo crediticio

Resumen

Palabras Clave

- Educacion
- Edad
- Genero
- Monto de credito

- Limite de Credito
- Riesgo de impago

Introduccion

Este estudio se enfoca en desarrollar un estudio en el cual se avale si variables como lo son la edad, el género, el estado marital y el nivel de educación, tienen algún peso o relevancia a la hora de medir el riesgo financiero de los prestatarios. De manera que lo que se busca es identificar patrones y con ello obtener una visión mas completa que no se limite únicamente a las características financieras de las personas, como lo es su nivel de ingresos, sino que también considera el contexto y características propias en las que se desenvuelve cada individuo. Por ello, es que este estudio comparar variables cuantitativas con variables cualitativas y tratar de ver la relación o correlación que pueda existir entre ellas, ya que al final todo ello se relaciona intrínsecamente con aspectos propios del prestatario, lo cual a su vez ayuda a generar una perspectiva mas robusta del riesgo que el mismo puede generar para alguna entidad financiera llamase banco u otra. Por lo cual, la presente investigación tiene como finalidad lograr determinar si variables cualitativas tienen una verdadera relevancia a la hora de estimar el riesgo de impago de las personas, lo cual ha su vez llega a ser pertinente en lo que es la carrera de ciencias actuariales, por lo que se pudieron usar herramientas que irán viendo a lo largo de los siguientes cursos, pero que llegan a ser muy útiles para el propósito de este trabajo. De manera que el objetivo principal de esta investigación es determinar si existe una correlación positiva entre algunas variables cualitativas y factores cuantitativos que se utilizan para determinar el riesgo de impago, esto a través del uso de una base de datos de un banco de Taiwán la cual brinda suficiente información para nuestro propósito. Por lo cual, esta investigación se fundamenta en un marco teórico y empírico con el que se busca lograr determinar a partir de varios métodos y pruebas la existencia o no de correlación entre nuestras variables. A su vez, el marco teórico que se maneja para esta investigación parece muy pertinente, esto debido a que los autores desean ir más allá de los modelos tradicionales y explorar factores adicionales al contexto económico de los prestatarios, lo cual da una mayor perspectiva que permite deslumbrar elementos de riesgo que podrían llegar a pasar desapercibidos en estudios un poco más convencionales.

Metodología

Resultados

Conclusiones

Agradecimientos

Nos gustaría hacer un especial agradecimiento al profesor Maikol Solís por siempre encaminar y aconsejarnos a lo largo de este trabajo, además de siempre estar ahi para brindarnos su tutoria y recomendarnos la bibliografía más adecuada. También, agradecer a los compañeros que realizaron un trabajo similar donde la retroalimentación fue muy gratificante y de mucha ayuda para tener críticas constructivas y mejorar el estudio. De igual forma, nos parece oportuno agradecer a la asistente Ana Laura López, ya que se tomó el tiempo de darnos retroalimentación para mejorar considerablemente las bitácoras, a la vez que nos brindó material de referencia el cual nos ayudó mucho.

3. Revisiones Finales