UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FES ACATLÁN

CARDOSO OLVERA EDGAR DAVID

314551885

SEGUNDO EXAMEN MÓDULO I

DIPLOMADO

CIENCIA DE DATOS

26/11/2020

Solicitud de Créditos

Conjunto de Datos

Primera Tabla

	ID_CLIENT	ID_SHOP	SEX	MARITAL_STATUS	AGE	QUANT_DEPENDANTS	EDUCATION	FLAG_RESIDENCIAL_PHONE	AREA_CODE_RESIDENCIAL_PHONE
0	2	15	F	S	18.0	0	NaN	Y	31
1	4	12	F	C	NaN	0	NaN	N	31
2	5	16	F	S	28.0	0	NaN	Y	31
3	6	24	M	S	26.0	0	NaN	N	31
4	7	55	F	S	22.0	0	NaN	Y	31

PAYMENT_DAY	...	QUANT_BANKING_ACCOUNTS	PERSONAL_REFERENCE_#1	PERSONAL_REFERENCE_#2	FLAG_MOBILE_PHONE	FLAG_CONTACT_PHONE
20	...	0	SARA	FELIPE	N	N
25	...	0	JACI	VALERIA ALEXANDRA TRAJANO	N	N
25	...	0	NaN	SANDRO L P MARTINS	N	N
28	...	0	NaN	ANA	N	N
12	...	0	NaN	NaN	N	N

PERSONAL_NET_INCOME	COD_APPLICATION_BOOTH	QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	FLAG_CARD_INSURANCE_OPTION	tgt
300.0	0	0	N	0
304.0	0	0	N	0
250.0	0	0	N	0
800.0	0	0	N	0
410.0	0	0	N	0

En la primera tabla observamos información de clientes que solicitan un crédito, esta tabla cuenta con 51,000 registros y 32 variables. Cada registro cuenta con características como tienda, sexo, estado civil, dependientes, ingresos, referencias, entre otros.

Segunda Tabla

	PROFESSION_CODE	PROFESSION
0	999	Healthcare Practitioners and Technical
1	950	Healthcare Practitioners and Technical
2	13	Educational Instruction and Library
3	205	Production
4	703	Educational Instruction and Library

En la segunda tabla encontramos el código de ocupaciones por cada profesión, contamos con 295 ocupaciones, de los cuales solo se cuentan con 4 ramas de profesiones, Healthcare Practitioners and Technical, Educational Instruction and Library, Production y Life, Physical, and Social Science.

Calidad de Datos

Realice el etiquetado de las variables de acuerdo a su tipo

Se tomaron como variables continuas las siguientes características:

MATE_INCOME	c_MATE_INCOME
PERSONAL_NET_INCOME	c_PERSONAL_NET_INCOME

Se tomaron como variables discretas las siguientes características: (La mayoría de las variables son discretas a excepción de los ingresos y el nombre de las personas de referencia.)

ID_CLIENT	v_ID_CLIENT
ID_SHOP	v_ID_SHOP
SEX	v_SEX
MARITAL_STATUS	v_MARITAL_STATUS
AGE	v_AGE
QUANT_DEPENDANTS	v_QUANT_DEPENDANTS
EDUCATION	v_EDUCATION
FLAG_RESIDENCIAL_PHONE	v_FLAG_RESIDENCIAL_PHONE
AREA_CODE_RESIDENCIAL_PHONE	v_AREA_CODE_RESIDENCIAL_PHONE
PAYMENT_DAY	v_PAYMENT_DAY
SHOP_RANK	v_SHOP_RANK
RESIDENCE_TYPE	v_RESIDENCE_TYPE
MONTHS_IN_RESIDENCE	v_MONTHS_IN_RESIDENCE
FLAG_MOTHERS_NAME	v_FLAG_MOTHERS_NAME
FLAG_FATHERS_NAME	v_FLAG_FATHERS_NAME
FLAG_RESIDENCE_TOWN=WORKING_TOWN	v_FLAG_RESIDENCE_TOWN=WORKING_TOWN
FLAG_RESIDENCE_STATE=WORKING_STATE	v_FLAG_RESIDENCE_STATE=WORKING_STATE
MONTHS_IN_THE_JOB	v_MONTHS_IN_THE_JOB
PROFESSION_CODE	v_PROFESSION_CODE
FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS	v_FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS
FLAG_OTHER_CARD	v_FLAG_OTHER_CARD
QUANT_BANKING_ACCOUNTS	v_QUANT_BANKING_ACCOUNTS
FLAG_MOBILE_PHONE	v_FLAG_MOBILE_PHONE
FLAG_CONTACT_PHONE	v_FLAG_CONTACT_PHONE
COD_APPLICATION_BOOTH	v_COD_APPLICATION_BOOTH
QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION
FLAG_CARD_INSURANCE_OPTION	v_FLAG_CARD_INSURANCE_OPTION
tgt	v_tgt

Se tomaron como variables de texto las siguientes características:

PERSONAL_REFERENCE_#1	t_PERSONAL_REFERENCE_#1
PERSONAL_REFERENCE_#2	t_PERSONAL_REFERENCE_#2

Revisar y eliminar duplicados, mantenga el primer elemento de los duplicados

Registros duplicados:

	v_ID_CLIENT	v_ID_SHOP	v_SEX	v_MARITAL_STATUS	v_AGE	v_QUANT_DEPENDANTS	v_EDUCATION	v_FLAG_RESIDENCIAL_PHONE
50152	39582	24	F	C	27.0	0	NaN	Y
50213	44401	16	F	O	22.0	0	NaN	Y
50592	17829	19	F	S	39.0	0	NaN	Y
50938	32014	20	F	S	37.0	0	NaN	Y
50944	19388	22	M	C	27.0	0	NaN	N

Sólo se eliminaron 5 registros repetidos en toda la tabla.

Completitud

	% Completitud
v_SEX	99.994117
v_AGE	95.152466
v_EDUCATION	0.000000
c_MATE_INCOME	99.399941
t_PERSONAL_REFERENCE_#1	59.554858
t_PERSONAL_REFERENCE_#2	72.769879
v_FLAG_CONTACT_PHONE	97.491911

La gran mayoría de las variables tiene el 100% de completitud. De igual manera muchas características tienen más de un 90% de completitud, sólo las variables t_PERSONAL_REFERENCE_#1 y t_PERSONAL_REFERENCE_#2 lo que no indica que la tabla sea incorrecta, puede indicar que muchos clientes no tenían una referencia que dar. La variable v_EDUCATION no contiene información alguna, esta se eliminará proximamente.

Revisión de valores fuera de la naturaleza de las variables (no validos) y conversión a NaN

```
0   v_ID_CLIENT                                  50995 non-null  int64
1   v_ID_SHOP                                    50995 non-null  int64
2   v_SEX                                        50992 non-null  object
3   v_MARITAL_STATUS                             50995 non-null  object
4   v_AGE                                        48523 non-null  float64
5   v_QUANT_DEPENDANTS                           50995 non-null  int64
6   v_EDUCATION                                  0 non-null      float64
7   v_FLAG_RESIDENCIAL_PHONE                     50995 non-null  object
8   v_AREA_CODE_RESIDENCIAL_PHONE                50995 non-null  int64
9   v_PAYMENT_DAY                                50995 non-null  int64
10  v_SHOP_RANK                                  50995 non-null  int64
11  v_RESIDENCE_TYPE                             50995 non-null  object
12  v_MONTHS_IN_RESIDENCE                        50995 non-null  int64
13  v_FLAG_MOTHERS_NAME                          50995 non-null  object
14  v_FLAG_FATHERS_NAME                          50995 non-null  object
15  v_FLAG_RESIDENCE_TOWN=WORKING_TOWN           50995 non-null  object
16  v_FLAG_RESIDENCE_STATE=WORKING_STATE         50995 non-null  object
17  v_MONTHS_IN_THE_JOB                          50995 non-null  int64
18  v_PROFESSION_CODE                            50995 non-null  int64
19  c_MATE_INCOME                                50689 non-null  float64
20  v_FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS    50995 non-null  object
21  v_FLAG_OTHER_CARD                            50995 non-null  object
22  v_QUANT_BANKING_ACCOUNTS                     50995 non-null  int64
23  t_PERSONAL_REFERENCE_#1                      30370 non-null  object
24  t_PERSONAL_REFERENCE_#2                      37109 non-null  object
25  v_FLAG_MOBILE_PHONE                          50995 non-null  object
26  v_FLAG_CONTACT_PHONE                         49716 non-null  object
27  c_PERSONAL_NET_INCOME                        50995 non-null  object
28  v_COD_APPLICATION_BOOTH                      50995 non-null  int64
29  v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION  50995 non-null  int64
30  v_FLAG_CARD_INSURANCE_OPTION                 50995 non-null  object
31  v_tgt                                        50995 non-null  int64
```

Se realizará la limpieza de estas variables que sean de tipo object ya que estas podrán contener información alfanumérica.

Cantidad de registros eliminados por ser de distinta naturaleza.

v_ID_CLIENT	0
v_ID_SHOP	0
v_SEX	791
v_MARITAL_STATUS	0
v_AGE	0
v_QUANT_DEPENDANTS	0
v_EDUCATION	0
v_FLAG_RESIDENCIAL_PHONE	0
v_AREA_CODE_RESIDENCIAL_PHONE	0
v_PAYMENT_DAY	0
v_SHOP_RANK	0
v_RESIDENCE_TYPE	7476
v_MONTHS_IN_RESIDENCE	0
v_FLAG_MOTHERS_NAME	0
v_FLAG_FATHERS_NAME	0
v_FLAG_RESIDENCE_TOWN=WORKING_TOWN	0
v_FLAG_RESIDENCE_STATE=WORKING_STATE	0
v_MONTHS_IN_THE_JOB	0
v_PROFESSION_CODE	0
c_MATE_INCOME	0
v_FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS	0
v_FLAG_OTHER_CARD	0
v_QUANT_BANKING_ACCOUNTS	0
t_PERSONAL_REFERENCE_#1	2347
t_PERSONAL_REFERENCE_#2	2709
v_FLAG_MOBILE_PHONE	0
v_FLAG_CONTACT_PHONE	0
c_PERSONAL_NET_INCOME	198
v_COD_APPLICATION_BOOTH	0
v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	0
v_FLAG_CARD_INSURANCE_OPTION	0
v_tgt	0

Observamos que las variables con mayor número de datos fuera de su rango son v_RESIDENCE_TYPE, t_PERSONAL_REFERENCE_#2 y t_PERSONAL_REFERENCE_#1, esto puede ser ya que no todas las personas cuenta con una residencia a su nombre o una referencia a quien dar, de igual manera se eliminaron las referencia que indicaran el parentesco, es decir si era su hermano/a, primo/a, cuñado/a, etc.

Realice la limpieza de variables y haga transformaciones a tipo de dato int o float en continuas (haga normalización de categorías si es necesario)

Revisión de contenido de variable, Estado Civil:

```
S     25749
C     17468
O      2826
V      2458
D      2194
S       160
C        99
O        15
V        14
D        12
Name: v_MARITAL_STATUS
```

Observamos que en la variable v_MARITAL_STATUS se encuentran variables repetidas, pero con espacio, este se le eliminará para crear categorías únicas.

Revisión de contenido de variable, Teléfono Residencial:

```
Y    41198
N     9296
y      406
n       95
Name: v_FLAG_RESIDENCIAL_PHONE
```

Los valores se encuentran en mayúsculas y minúsculas, estos se normalizarán a letras mayúsculas.

Revisión de contenido de variable, v_RESIDENCE_TYPE:

```
P     31970
A      5538
C      3783
O      1782
 p      211
P        84
p        50
 a       35
 c       23
A        10
 o        9
a         9
c         4
C         4
O         4
o         2
p         1
Name: v_RESIDENCE_TYPE
```

Observamos que existen varias categorías en minúsculas y con espacios que se normalizaran a categorías mayúsculas sin espacio.

Eliminación de variables que posean una completitud inferior al 80%

	% Completitud
v_EDUCATION	0.000000
t_PERSONAL_REFERENCE_#1	54.952446
t_PERSONAL_REFERENCE_#2	67.457594

Observamos que las variables v_EDUCATION, t_PERSONAL_REFERENCE_#1, t_PERSONAL_REFERENCE_#2, cuentan con menos del 80% de completitud, estas variables serán eliminadas de la tabla original.

Cruce con la tabla de ocupaciones

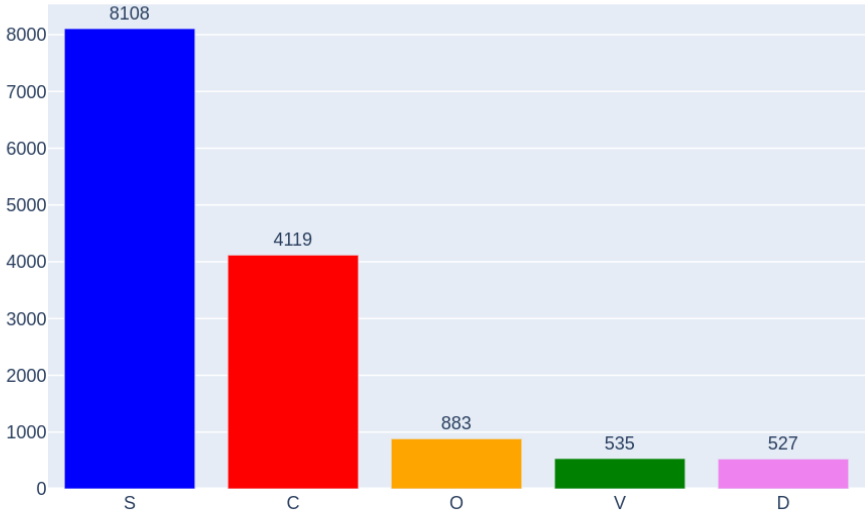
	v_tgt	v_PROFESSION
0	0	Educational Instruction and Library
1	0	Life, Physical, and Social Science
2	0	Healthcare Practitioners and Technical
3	0	Healthcare Practitioners and Technical
4	0	Healthcare Practitioners and Technical
...	...	...
50990	0	Healthcare Practitioners and Technical
50991	1	Educational Instruction and Library
50992	0	Healthcare Practitioners and Technical
50993	0	Educational Instruction and Library
50994	1	Educational Instruction and Library

Análisis Exploratorio de Datos

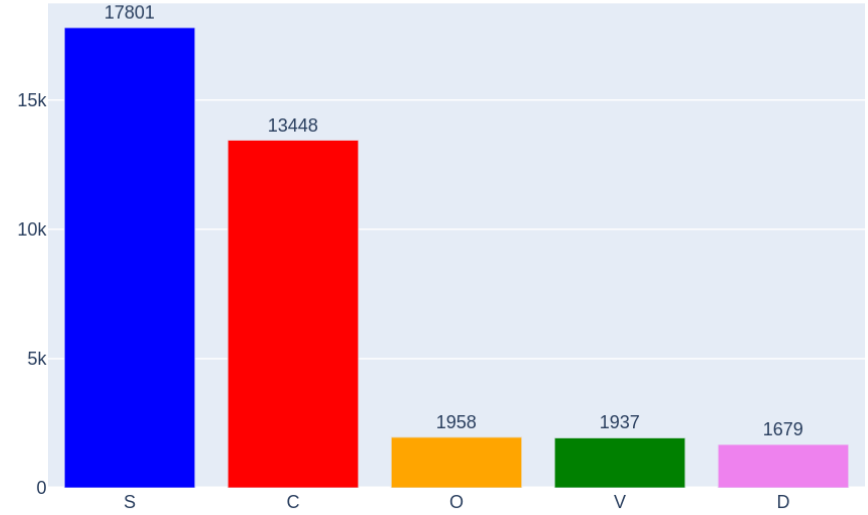
Realice análisis interesantes sobre los datos proporcionados, genere gráficas representativas.

Observamos que hay una mayoría de solicitudes de crédito en el sexo femenino y vemos que se aceptan más los créditos a mujeres que a hombres. Es decir que hay mayor probabilidad de aceptación de crédito si eres mujer.

Conteo de Créditos Aceptados por Estado Civil

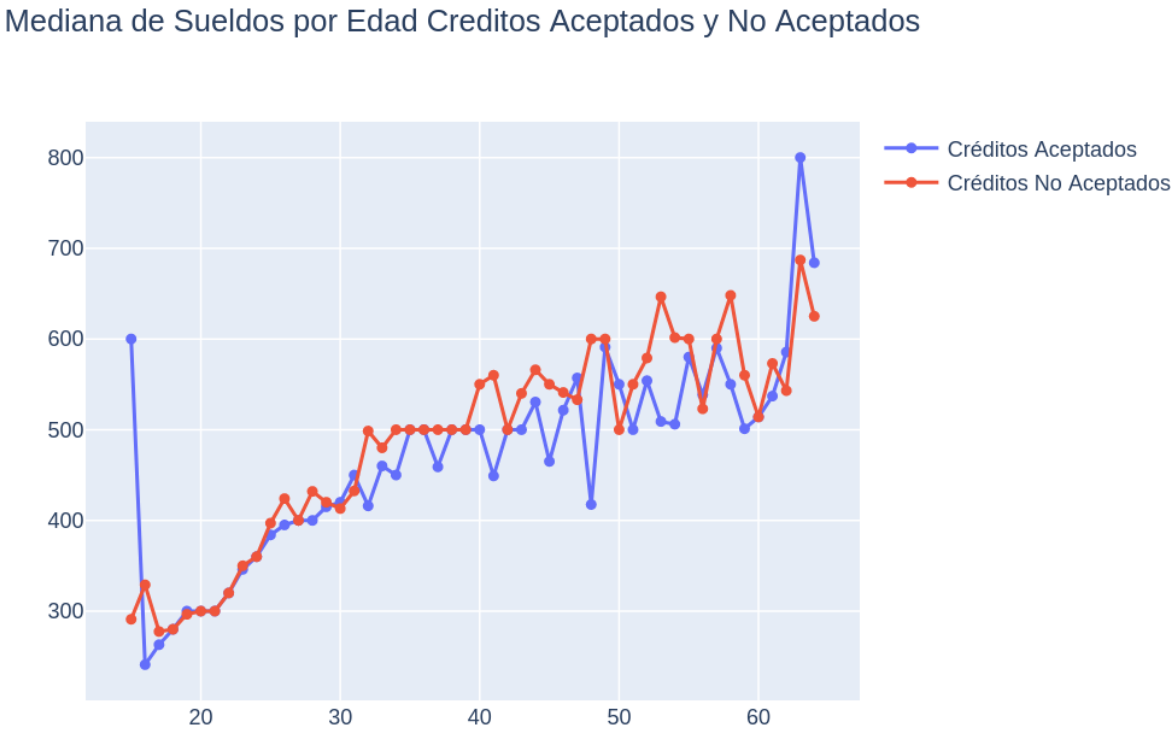


Conteo de Créditos No Aceptados por Estado Civil



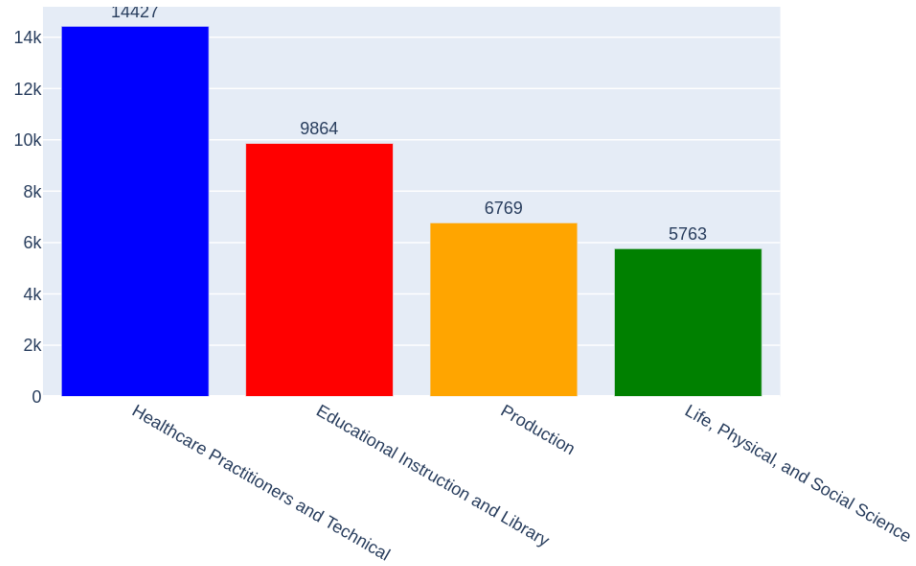
Observamos que las personas solteras y casadas son los que piden más créditos, pero se aceptan más créditos a las personas solteras a comparación de otro estado civil.

Viendo los rangos de edades observamos que existen valores extremos, para las siguientes gráficas se limitará a un rango entre 15 y 65 años de edad.

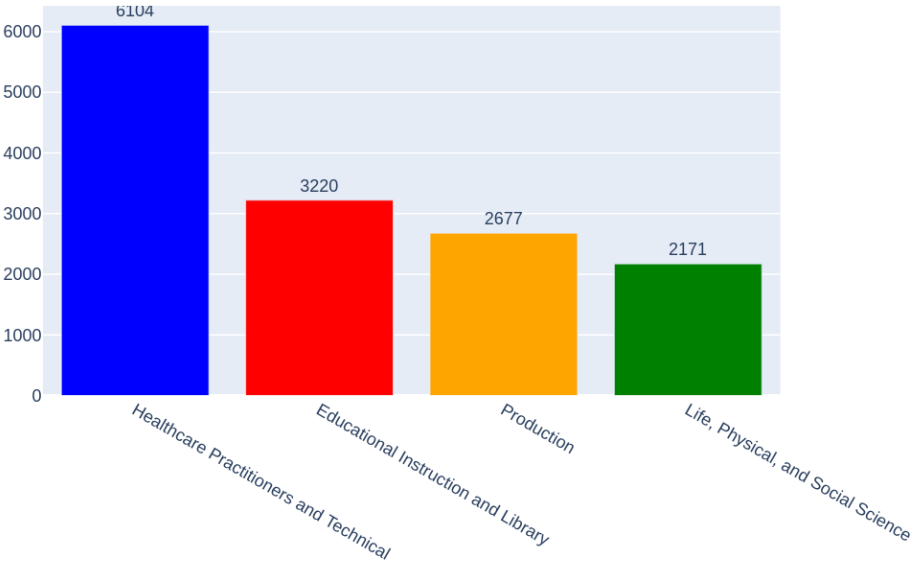


Podemos observar que los créditos aceptados fueron a personas que ganan en mediana menos dinero que otras personas que solicitaron créditos con su misma edad. Esto podría indicar que el sueldo no es importante para recibir un crédito.

Créditos No Aceptados por Profesión

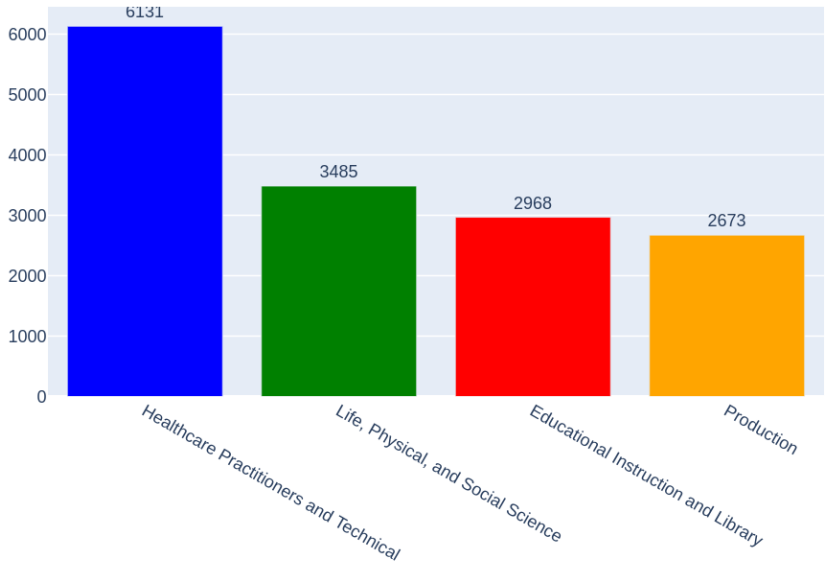


Créditos Aceptados por Profesión

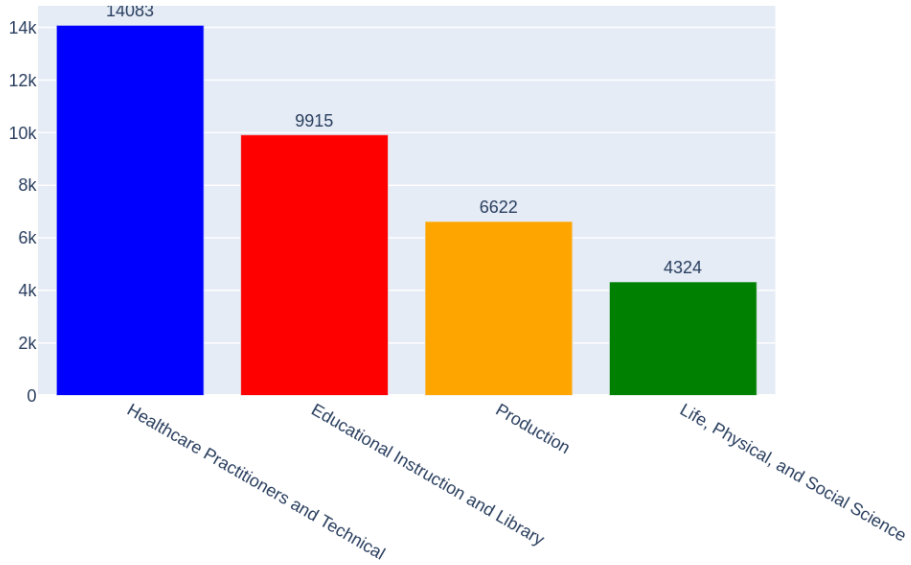


Los profesionistas de la salud tienen más aceptación de los créditos a comparación de las otras profesiones, vemos una baja no proporcional en la aceptación de créditos en las otras profesiones. Esto puede indicar que los profesionales de la salud podrán obtener un crédito con mayor facilidad.

Hombres por Profesión



Mujeres por Profesión



Observamos que los profesionales de la salud en ambos sexos solicitan más créditos, pero en las demás profesiones observamos un cambio en la solicitud de crédito por profesión. Los hombres que trabajan en las ciencias sociales solicitan más créditos proporcionalmente que las mujeres en esta área.

Datos anómalos

	features	n outliers IQR	n outliers Percentil	n outliers Z-Score	n outliers IQR %	n outliers Percentil %	n outliers Z-Score %	total outliers	% outliers	indices
0	v_AGE	448	25824	0	0.878518	50.640259	0.000000	448	0.878518	[38919, 14345, 47126, 18456, 43032, 49182, 450...
1	v_QUANT_DEPENDANTS	0	0	0	0.000000	0.000000	0.000000	0	0.000000	[]
2	v_PAYMENT_DAY	0	13072	0	0.000000	25.633886	0.000000	0	0.000000	[]
3	v_MONTHS_IN_RESIDENCE	599	26733	555	1.174625	52.422787	1.088342	599	1.174625	[43011, 45059, 6148, 45066, 4111, 16404, 8218,...
4	v_MONTHS_IN_THE_JOB	5554	27789	1308	10.891264	54.493578	2.564957	2356	4.620061	[40960, 16387, 8195, 8205, 49166, 24590, 49179...
5	c_MATE_INCOME	2007	2007	0	3.935680	3.935680	0.000000	2007	3.935680	[40960, 40962, 5, 24581, 49162, 8206, 24590, 4...
6	v_QUANT_BANKING_ACCOUNTS	0	0	0	0.000000	0.000000	0.000000	0	0.000000	[]
7	c_PERSONAL_NET_INCOME	4200	27936	0	8.236102	54.781841	0.000000	2540	4.980880	[8195, 49156, 8197, 40965, 10, 49170, 19, 1640...
8	v_COD_APPLICATION_BOOTH	0	0	0	0.000000	0.000000	0.000000	0	0.000000	[]
9	v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	6679	897	897	13.097362	1.758996	1.758996	897	1.758996	[20480, 18433, 10242, 47105, 34822, 6153,

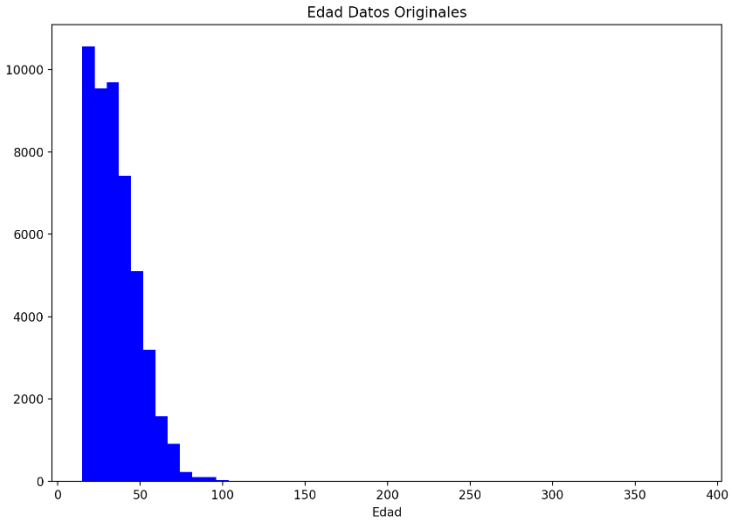
Las variables v_AGE, v_MONTHS_IN_RESIDENCE tuvieron menos de 600 valores extremos, un porcentaje menor al 1.20% de la tabla total, seguido de v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION con menos de 900 valores extremos lo que representa un 1.75% de la tabla. Finalmente, c_MATE_INCOME, c_PERSONAL_NET_INCOME, v_MONTHS_IN_THE_JOB tuvieron la mayor cantidad de outliers con un poco menos de 5% de los registros totales. Esto nos puede indicar que muchas personas pueden mentir al registrar sus ingresos y los meses en su trabajo, es decir no hay una relativa consistencia con las variables que tienen que ver con lo ingresos y el trabajo.

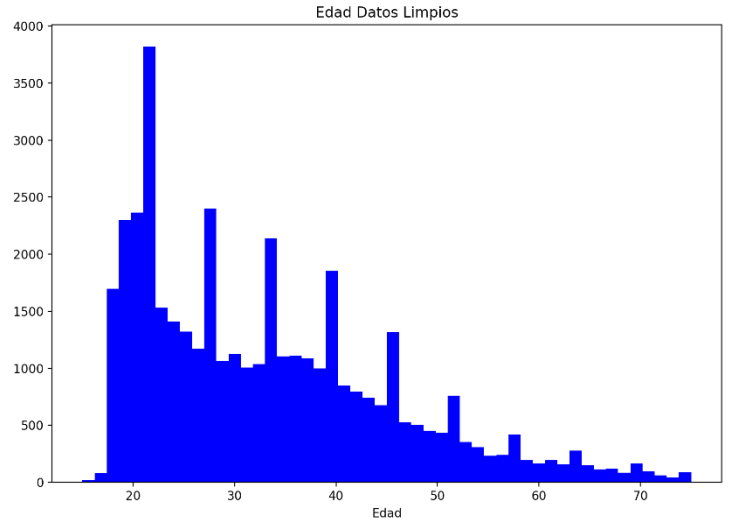
Número de elementos al eliminar outliers

```
features
v_AGE                                    50547
v_QUANT_DEPENDANTS                       50547
v_PAYMENT_DAY                            50547
v_MONTHS_IN_RESIDENCE                    49966
v_MONTHS_IN_THE_JOB                      47739
c_MATE_INCOME                            45872
v_QUANT_BANKING_ACCOUNTS                 45872
c_PERSONAL_NET_INCOME                    43849
v_COD_APPLICATION_BOOTH                  43849
v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION    43291
```

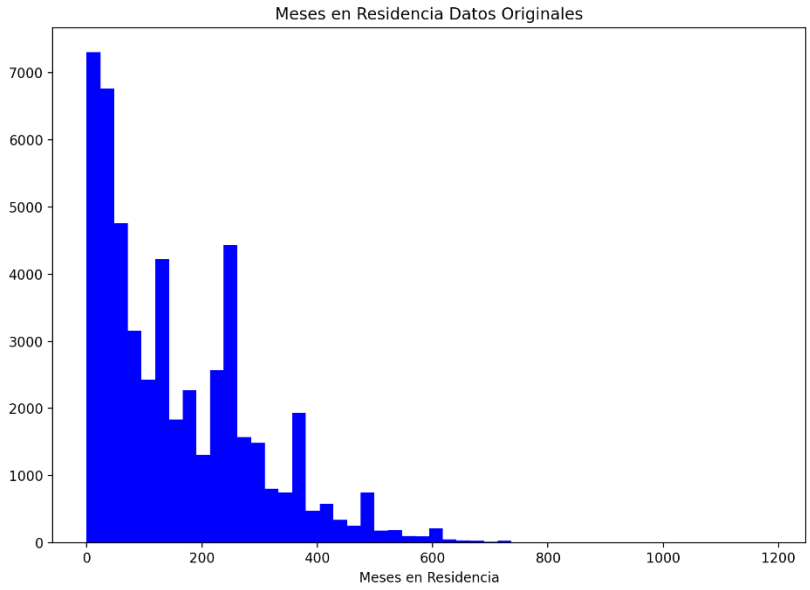
Vemos una reducción de 50,995 registros que se tenían en la tabla original a 43,291, lo que representa un 15.10% de variables eliminadas con respecto a la tabla inicial.

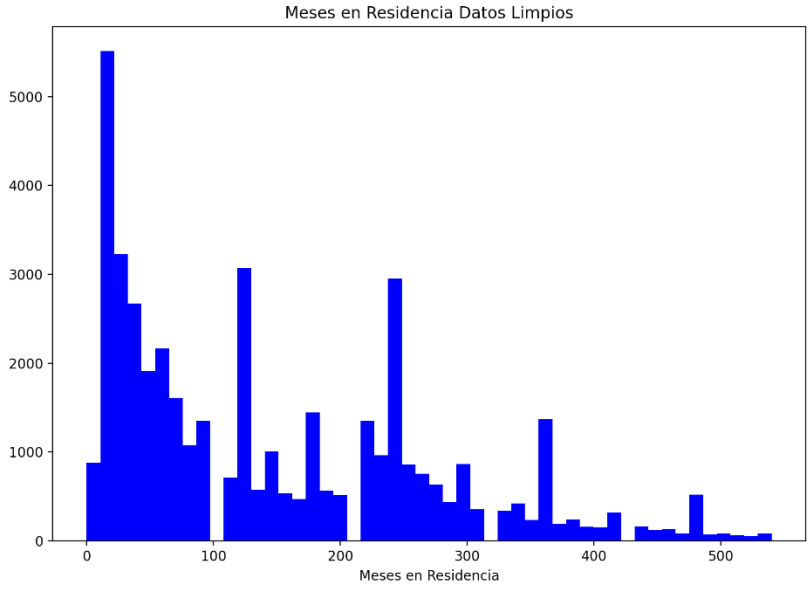
Además, se debe añadir los gráficos del histograma antes de la remoción y después de la remoción de outliers de todas las variables continuas con datos anómalos.



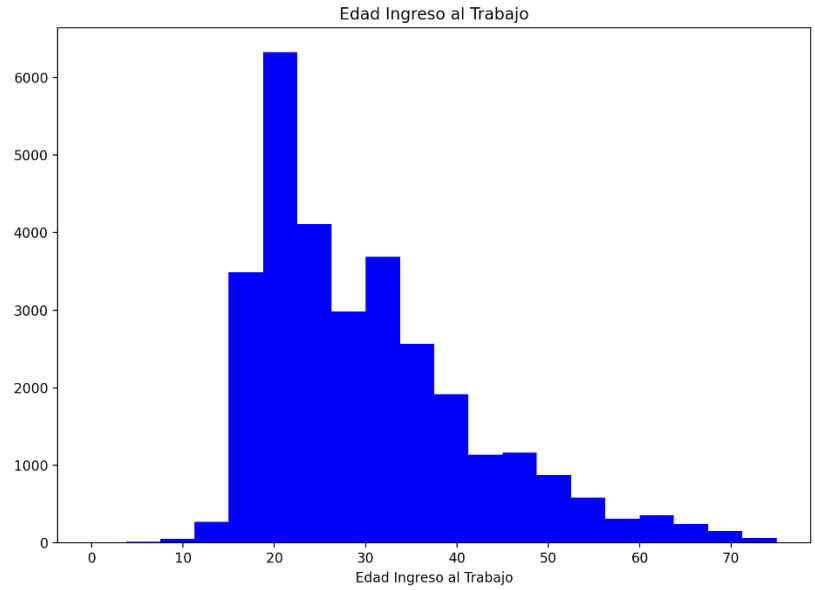


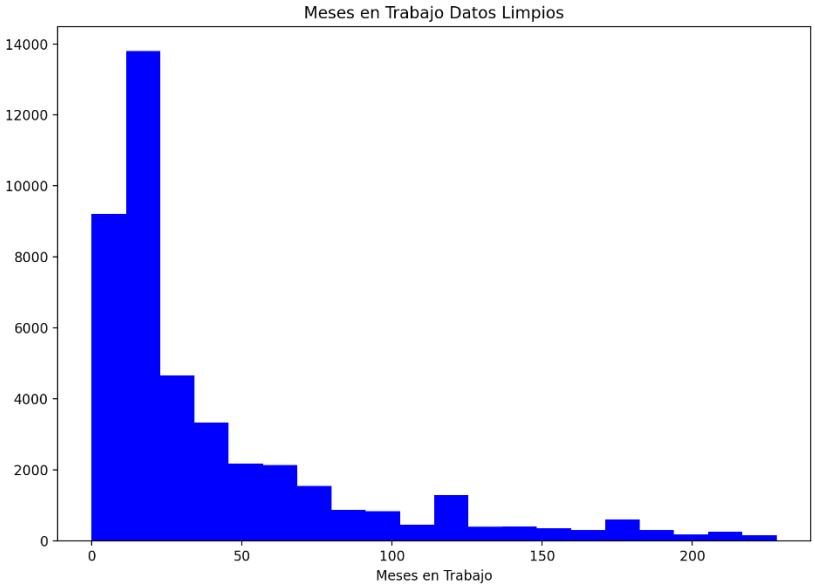
Observamos una reducción en el rango de variables y una concentración de edades entre los 18 a 25 años, esto puede ser ya que las personas de dichas edades se empiezan a independizar de sus familias y necesitan un apoyo económico para lograr sus objetivos.



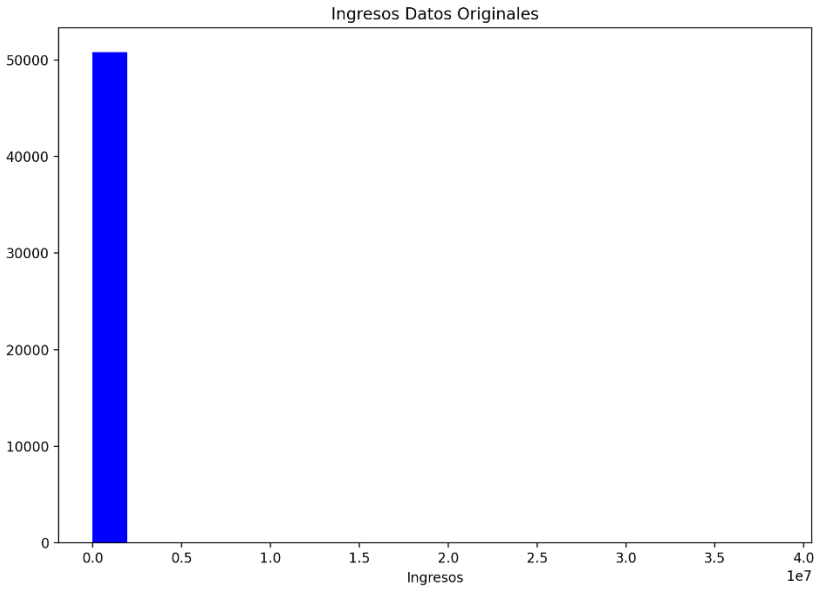


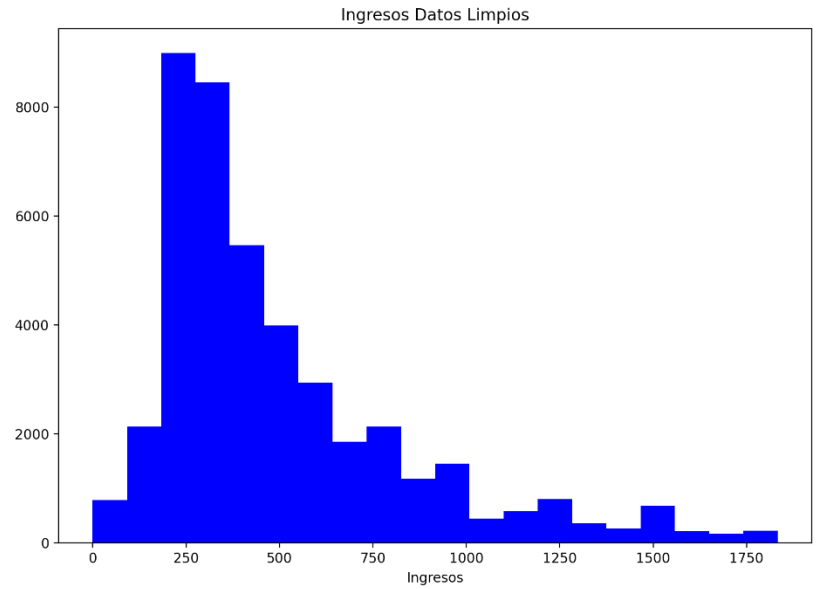
Vemos una reducción en el rango de Meses en residencia, y notamos una división en bloque en cada rango. Se puede observar que la concentración se encuentra en el primer bloque, es decir hay más solicitudes de crédito en personas que llevan menos de 100 meses o menos de una década en su residencia.





Observamos una disminución de rango en los meses de trabajo, de igual manera vemos una concentración de solicitud de créditos en un rango entre 0 y 24 meses, es decir que las personas que van empezando un trabajo son los mayores solicitantes de créditos.





En los datos originales existe un rango muy grande de ingresos, este se disminuye en los datos ya tratados, notamos que las personas que ganan entre 200 a 400 son lo que solicitan más crédito, esto es porque son los individuos que ganan menos en nuestro conjunto de datos.

Datos faltantes

Genere su conjunto de entrenamiento y prueba, donde el conjunto de prueba tenga el 30% de la información

Variables con valores ausentes:

v_AGE	1518
c_MATE_INCOME	198
c_PERSONAL_NET_INCOME	125

Se utilizará la mediana ya que no conocemos el comportamiento real de las variables continuas, si estas estuvieran divididas en forma normal se utilizaría la media.

Valores a Imputar:

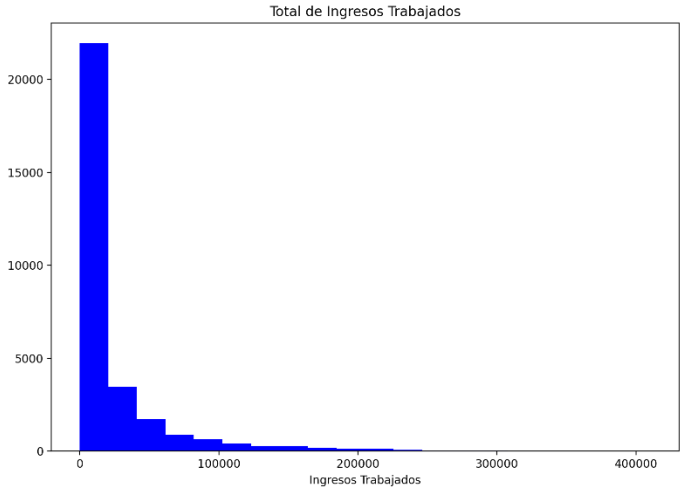
c_MATE_INCOME: 0.0

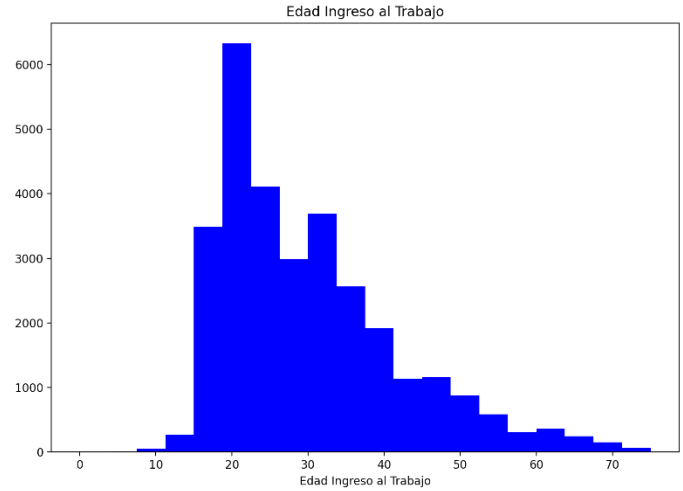
c_PERSONAL_NET_INCOME: 397.0

v_AGE: 31.0

Ingeniería de datos

Genere variables a partir de las variables continuas, al menos se deben crear dos





Se crearon estas dos nuevas variables para corroborar que las personas digan información realista, la primera variable es c_WORKED_INCOME e indica cual es el salario obtenido por los meses trabajados, podemos observar en el histograma que la mayoría de las personas tienen ingresos 0, por lo que no sería correcto asignar un crédito a una persona que no ha trabajado ni un mes dentro de una empresa. La segunda variable creada es edad de inicio en el trabajo, observamos en el histograma que la gran mayoría inicio un trabajo a partir de los 20 años, pero existen otros datos como las personas que empezaron a trabajar a edad 5 o más de 70 años de edad.

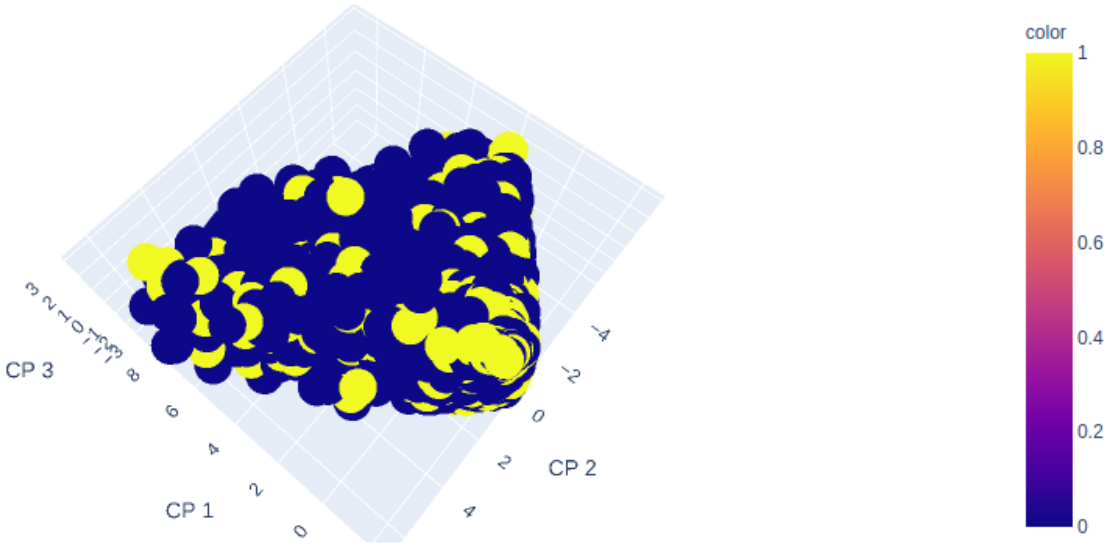
Reducción de dimensiones

Elimine variables con varianza baja

```
v_QUANT_DEPENDANTS                          0.000000
c_MATE_INCOME                               0.000000
v_QUANT_BANKING_ACCOUNTS                    0.000000
v_COD_APPLICATION_BOOTH                     0.000000
v_FLAG_MOTHERS_NAME_Y                       0.003912
v_FLAG_RESIDENCE_STATE=WORKING_STATE_Y      0.008429
```

Observamos que las variables v_QUANT_DEPENDANTS, c_MATE_INCOME, v_QUANT_BANKING_ACCOUNTS, v_COD_APPLICATION_BOOTH, v_FLAG_MOTHERS_NAME_Y, v_FLAG_RESIDENCE_STATE=WORKING_STATE_Y tienen una varianza cercana a 0 por lo que se eliminarán, ya que estas no aportarán nada al modelo.

Tome solo las variables continuas y utilice PCA para generar una visualización en 3D de las variables continuas con la distinción de los valores del target.



Varianza Explicada: 74.64%

Tablas Entrenamiento y Test Limpias

	v_ID_CLIENT	v_ID_SHOP	v_AGE	v_AREA_CODE_RESIDENCIAL_PHONE	v_PAYMENT_DAY	v_SHOP_RANK	v_MONTHS_IN_RESIDENCE
22212	27731	55	35.0	31	18	0	120
11543	14411	18	20.0	23	12	0	12
16591	20790	11	57.0	31	12	0	312
18200	22826	11	49.0	31	12	0	60
35637	44528	19	61.0	50	12	0	468

v_PROFESSION_Healthcare Practitioners and Technical	v_PROFESSION_Life, Physical, and Social Science	v_PROFESSION_Production	c_WORKED_INCOME	v_START_WORKING
0	0	0	20016.0	29.0
0	0	1	1800.0	19.0
0	0	1	0.0	57.0
1	0	0	7200.0	48.0
0	0	0	0.0	61.0

Tratamiento 2

Transformación entrópica

Varianza Cero

```
v_QUANT_DEPENDANTS          0.0
v_QUANT_BANKING_ACCOUNTS    0.0
v_COD_APPLICATION_BOOTH     0.0
```

Conjunto con transformación entrópica

v_AGE	v_AREA_CODE_RESIDENCIAL_PHONE	v_PAYMENT_DAY	v_SHOP_RANK	v_MONTHS_IN_RESIDENCE	v_MONTHS_IN_THE_JOB
(14.643, 372.0]	50	(19.0, 23.5]	0	(-1.188, 1188.0]	(-1.176, 1176.0]
(14.643, 372.0]	23	(23.5, 28.0]	0	(-1.188, 1188.0]	(-1.176, 1176.0]
(14.643, 372.0]	31	(23.5, 28.0]	0	(-1.188, 1188.0]	(-1.176, 1176.0]
(14.643, 372.0]	31	(10.0, 14.5]	0	(-1.188, 1188.0]	(-1.176, 1176.0]
(14.643, 372.0]	31	(19.0, 23.5]	0	(-1.188, 1188.0]	(-1.176, 1176.0]

v_PROFESSION_CODE	c_MATE_INCOME	c_PERSONAL_NET_INCOME	v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION
26	(-150.0, 150000.0]	(-38529.098, 38529098.0]	0
717	(-150.0, 150000.0]	(-38529.098, 38529098.0]	0
717	(-150.0, 150000.0]	(-38529.098, 38529098.0]	0
13	(-150.0, 150000.0]	(-38529.098, 38529098.0]	0
999	(-150.0, 150000.0]	(-38529.098, 38529098.0]	0

(W_v_RESIDENCE_TYPE_O,)	(W_v_RESIDENCE_TYPE_P,)	(W_v_FLAG_MOTHERS_NAME_Y,)	(W_v_FLAG_FATHERS_NAME_Y,)
0.001549	-0.032498	-0.000702	-0.010616
0.001549	0.054721	-0.000702	-0.010616
0.001549	-0.032498	-0.000702	-0.010616
0.001549	-0.032498	-0.000702	-0.010616
0.001549	-0.032498	-0.000702	-0.010616

(W_v_FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS_Y,)	(W_v_PROFESSION_Healthcare Practitioners and Technical,)	(W_v_PROFESSION_Life, Physical, and Social Science,)	(W_v_PROFESSION_Production,)
-0.000962	0.094454	0.005589	-0.011995
-0.000962	-0.065788	-0.030365	-0.011995
-0.000962	-0.065788	-0.030365	-0.011995
-0.000962	-0.065788	0.005589	-0.011995
-0.000962	0.094454	0.005589	-0.011995

Conjunto Final Transformaciones Entrópicas

(W_v_AGE,)	(W_v_AREA_CODE_RESIDENCIAL_PHONE,)	(W_v_PAYMENT_DAY,)	(W_v_SHOP_RANK,)	(W_v_MONTHS_IN_RESIDENCE,)
0.000947	-0.050532	-0.087403	-0.000312	0.0
0.000947	-0.050532	0.041668	-0.000312	0.0
0.000947	-0.050532	0.041668	-0.000312	0.0
0.000947	-0.050532	0.041668	-0.000312	0.0
0.000947	0.215390	0.011208	-0.000312	0.0

(W_v_MONTHS_IN_THE_JOB,)	(W_v_PROFESSION_CODE,)	(W_c_MATE_INCOME,)	(W_c_PERSONAL_NET_INCOME,)
0.0	-0.123190	-0.000564	-0.000367
0.0	-0.606728	-0.000564	-0.000367
0.0	-0.142246	-0.000564	-0.000367
0.0	-inf	-0.000564	-0.000367
0.0	0.277012	-0.000564	-0.000367

(W_v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION,)	...	(W_v_RESIDENCE_TYPE_P,)	(W_v_FLAG_MOTHERS_NAME_Y,)	(W_v_FLAG_FATHERS_NAME_Y,)
0.033735	...	-0.032498	-0.000702	-0.010616
0.033735	...	-0.032498	-0.000702	-0.010616
0.033735	...	0.054721	-0.000702	-0.010616
-0.500931	...	-0.032498	-0.000702	-0.010616
0.033735	...	-0.032498	-0.000702	-0.010616

(W_v_FLAG_FATHERS_NAME_Y,)	(W_v_FLAG_RESIDENCE_TOWN=WORKING_TOWN_Y,)	(W_v_FLAG_RESIDENCE_STATE=WORKING_STATE_Y,)
-0.010616	-0.033746	-0.000531
-0.010616	0.027883	-0.000531
-0.010616	0.027883	-0.000531
-0.010616	0.027883	-0.000531
-0.010616	0.027883	-0.000531

(W_v_FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS_Y,)	(W_v_PROFESSION_Healthcare Practitioners and Technical,)	(W_v_PROFESSION_Life, Physical, and Social Science,)	(W_v_PROFESSION_Production,)	v_tgt
-0.000962	-0.065788	0.005589	-0.011995	0
-0.000962	-0.065788	0.005589	-0.011995	0
-0.000962	-0.065788	-0.030365	-0.011995	1
-0.000962	-0.065788	-0.030365	-0.011995	1
-0.000962	-0.065788	0.005589	-0.011995	0

Observamos que todas las variables tienen una transformación entrópica y en la columna final se encuentra el target, esta tabla puede ser utilizada para el entrenamiento de algún modelo de clasificación binaria.

IV por Variable

v_AGE	4.575220e-07
v_AREA_CODE_RESIDENCIAL_PHONE	inf
v_PAYMENT_DAY	1.064961e-02
v_SHOP_RANK	3.303141e-04
v_MONTHS_IN_RESIDENCE	0.000000e+00
v_MONTHS_IN_THE_JOB	0.000000e+00
v_PROFESSION_CODE	inf
c_MATE_INCOME	2.110806e-05
c_PERSONAL_NET_INCOME	5.267641e-05
v_QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	inf
v_SEX_M	1.054177e-02
v_MARITAL_STATUS_D	2.019769e-03
v_MARITAL_STATUS_O	1.778334e-03
v_MARITAL_STATUS_S	2.992344e-02
v_MARITAL_STATUS_V	5.136864e-03
v_FLAG_RESIDENCIAL_PHONE_Y	3.491809e-02
v_RESIDENCE_TYPE_C	4.632509e-04
v_RESIDENCE_TYPE_O	3.456349e-05
v_RESIDENCE_TYPE_P	1.296049e-03
v_FLAG_MOTHERS_NAME_Y	1.282159e-05
v_FLAG_FATHERS_NAME_Y	2.317606e-03
v_FLAG_RESIDENCE_TOWN=WORKING_TOWN_Y	1.379746e-03
v_FLAG_RESIDENCE_STATE=WORKING_STATE_Y	1.223861e-05
v_FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS_Y	7.758093e-05
v_PROFESSION_Healthcare Practitioners and Technical	6.338747e-03
v_PROFESSION_Life, Physical, and Social Science	3.533714e-05
v_PROFESSION_Production	1.687287e-04

Cuestionario