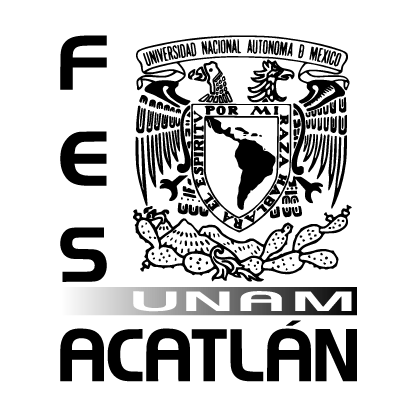
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FES ACATLÁN



CARDOSO OLVERA EDGAR DAVID

314551885

SEGUNDO EXAMEN MÓDULO I

DIPLOMADO

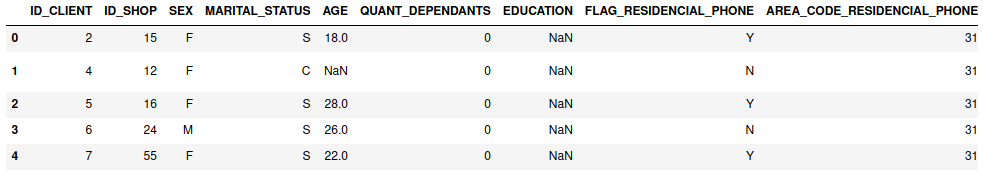
CIENCIA DE DATOS

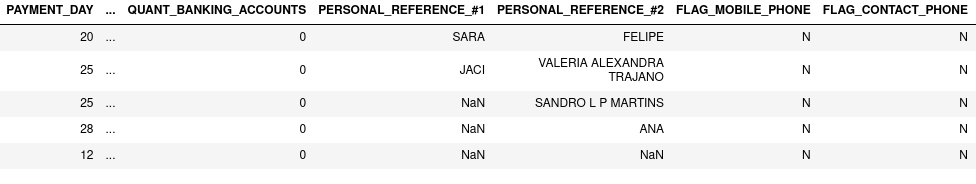
26/11/2020

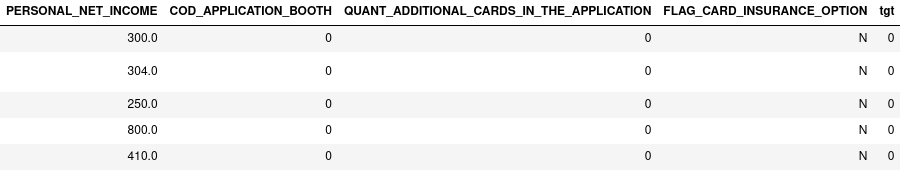
***Solicitud de Créditos***

**Conjunto de Datos**

Primera Tabla

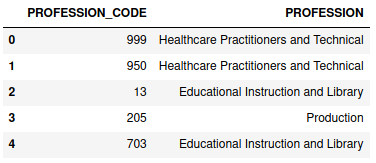
****

****

****

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



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.

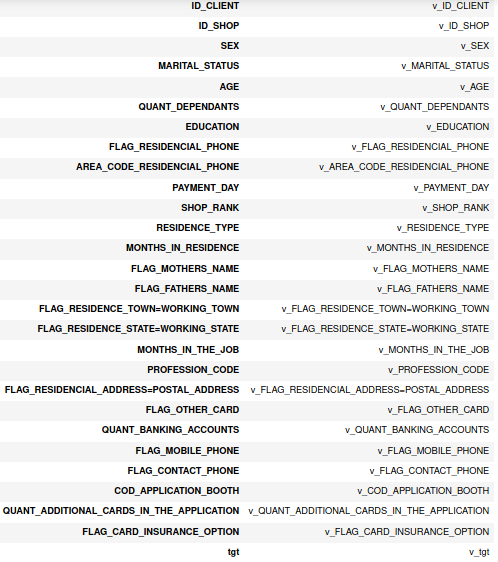
**Calidad de Datos**

**Realice el etiquetado de las variables de acuerdo a su tipo**

Se tomaron como variables continuas las siguientes características:



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.)

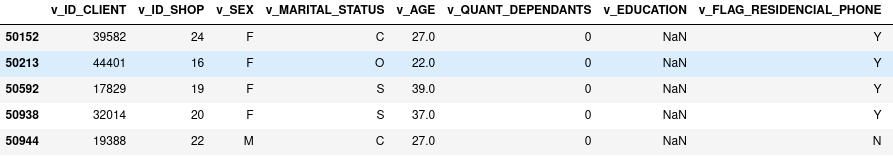


Se tomaron como variables de texto las siguientes características:



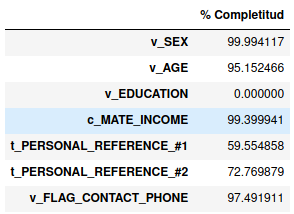
**Revisar y eliminar duplicados, mantenga el primer elemento de los duplicados**

Registros duplicados:



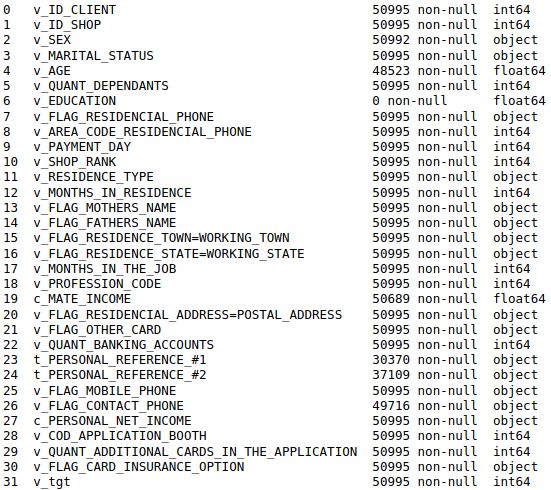
Sólo se eliminaron 5 registros repetidos en toda la tabla.

**Completitud**



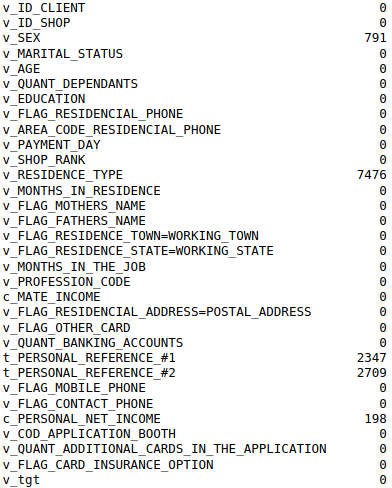
La gran mayoría de las variables tiene el 100% de completitud. De igual manera muchas características tienen más de un 90% de completitud, sólo las variables t\_PERSONAL\_REFERENCE\_#1 y t\_PERSONAL\_REFERENCE\_#2 lo que no indica que la tabla sea incorrecta, puede indicar que muchos clientes no tenían 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**



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

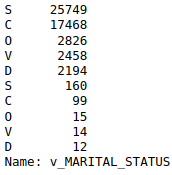
Cantidad de registros eliminados por ser de distinta naturaleza.



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.

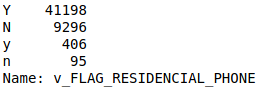
**Realice la limpieza de variables y haga transformaciones a tipo de dato int o float en continuas (haga normalización de categorías si es necesario)**

Revisión de contenido de variable, Estado Civil:



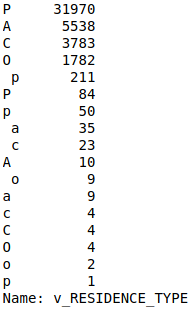
Observamos que en la variable v\_MARITAL\_STATUS se encuentran variables repetidas, pero con espacio, este se le eliminará para crear categorías únicas.

Revisión de contenido de variable, Teléfono Residencial:



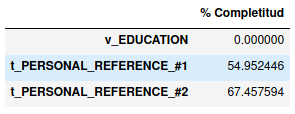
Los valores se encuentran en mayúsculas y minúsculas, estos se normalizarán a letras mayúsculas.

Revisión de contenido de variable, v\_RESIDENCE\_TYPE:



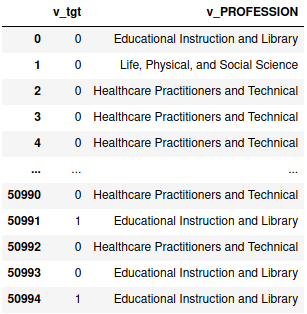
Observamos que existen varias categorías en minúsculas y con espacios que se normalizaran a categorías mayúsculas sin espacio.

**Eliminación de variables que posean una completitud inferior al 80%**



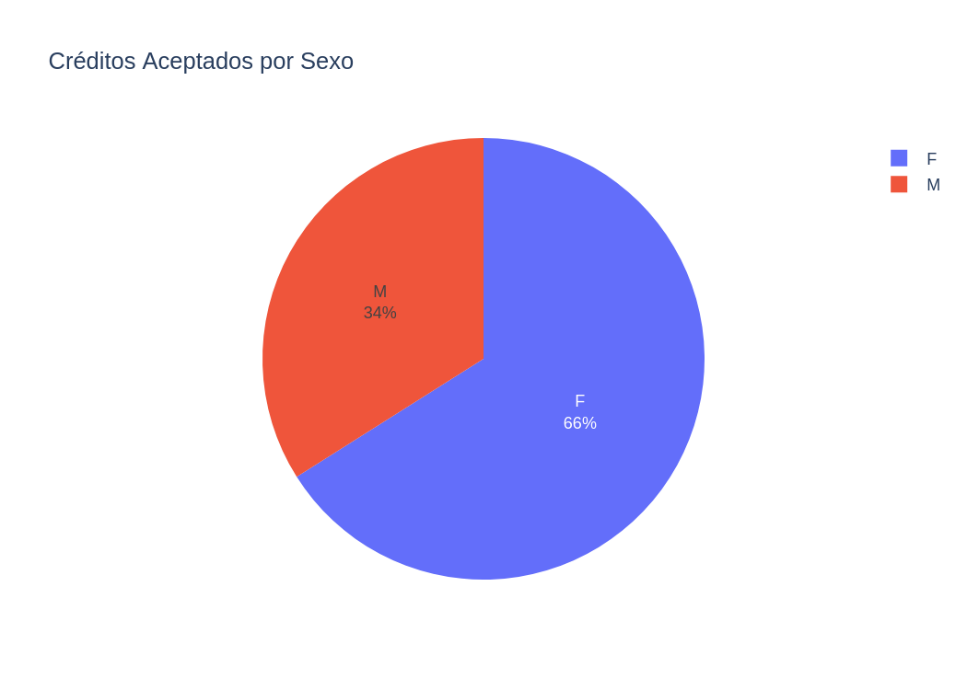
Observamos que las variables v\_EDUCATION, t\_PERSONAL\_REFERENCE\_#1, t\_PERSONAL\_REFERENCE\_#2, cuentan con menos del 80% de completitud, estas variables serán eliminadas de la tabla original.

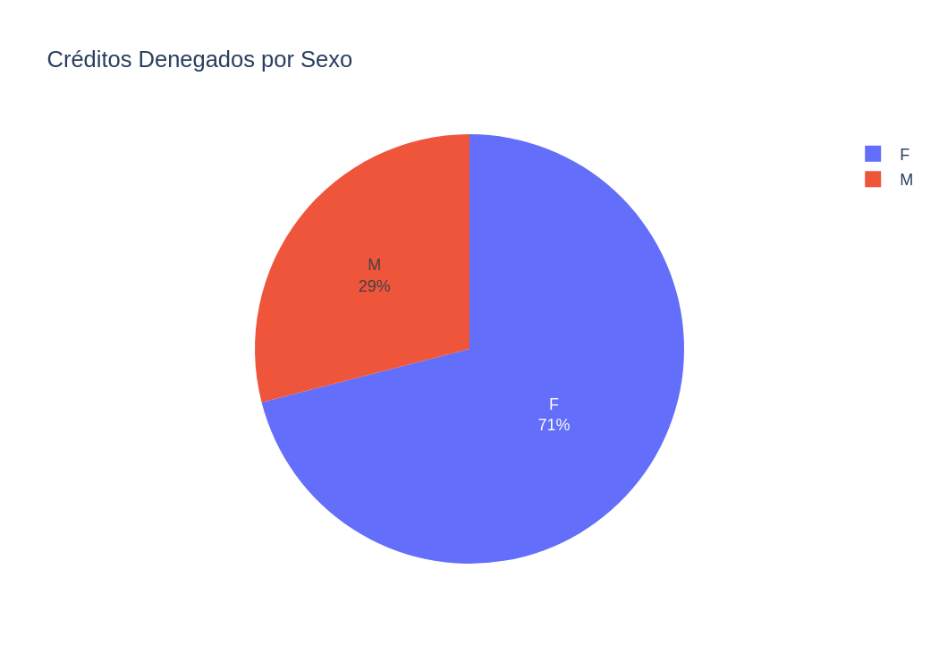
**Cruce con la tabla de ocupaciones**



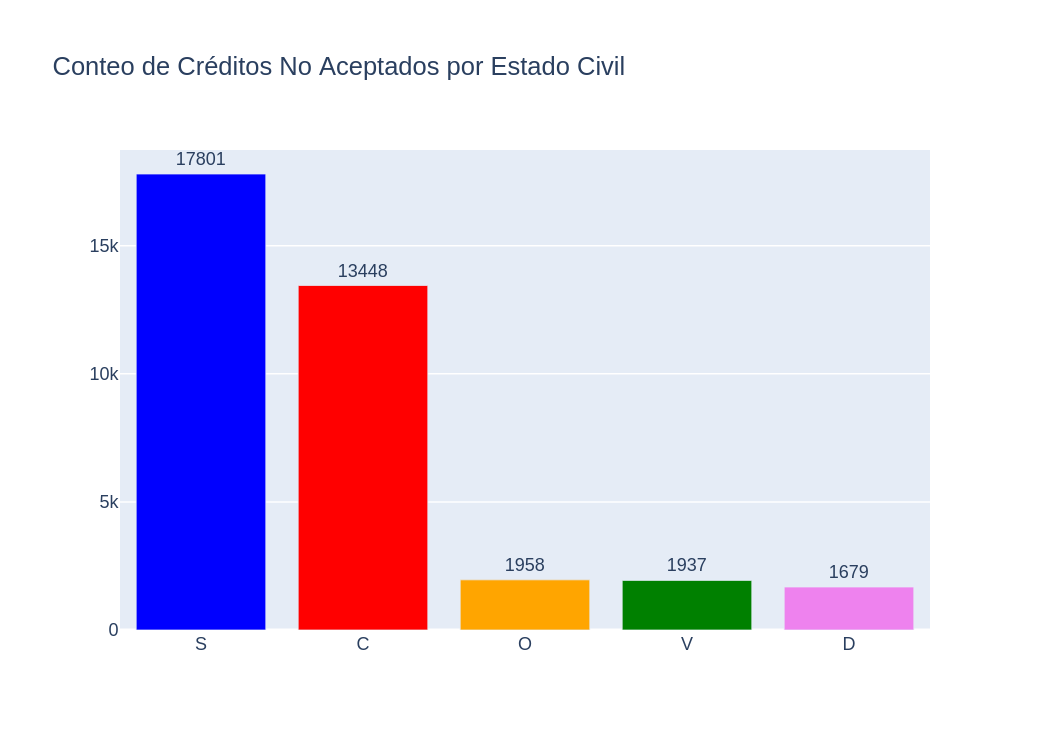
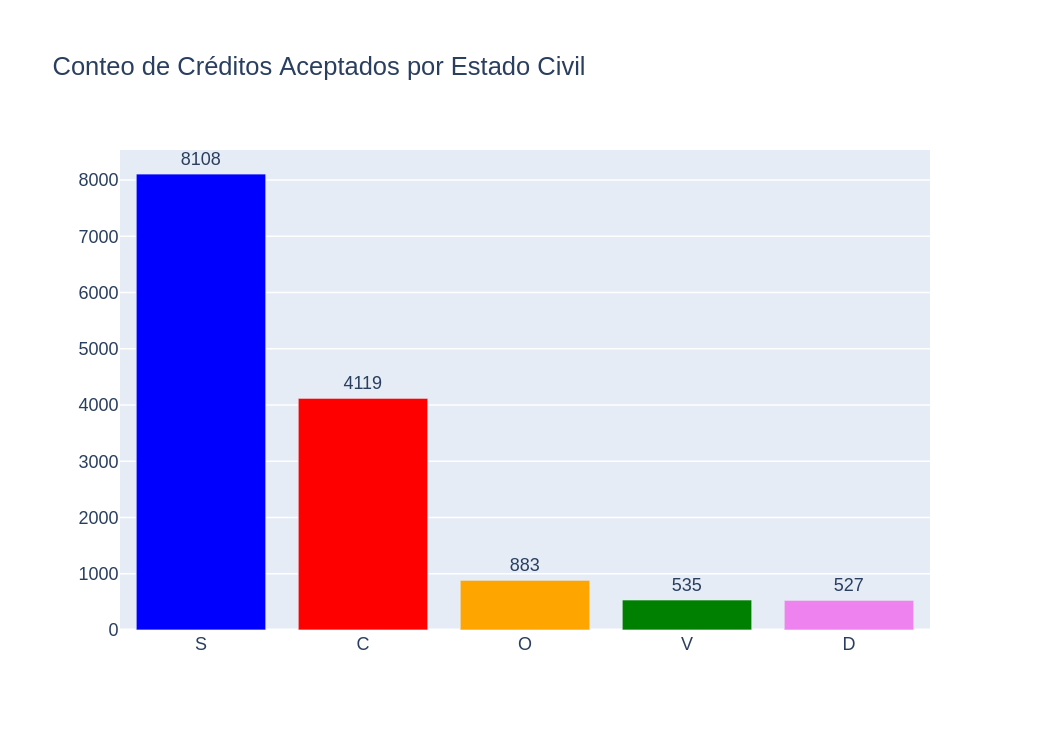
**Análisis Exploratorio de Datos**

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

****

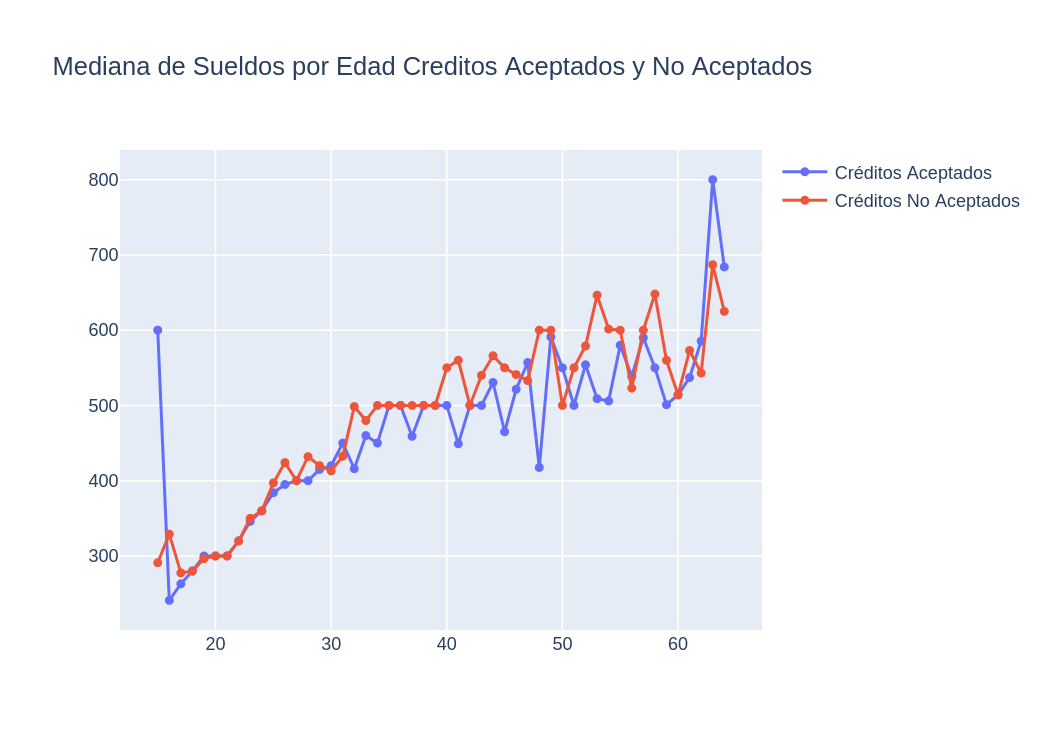
****

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.

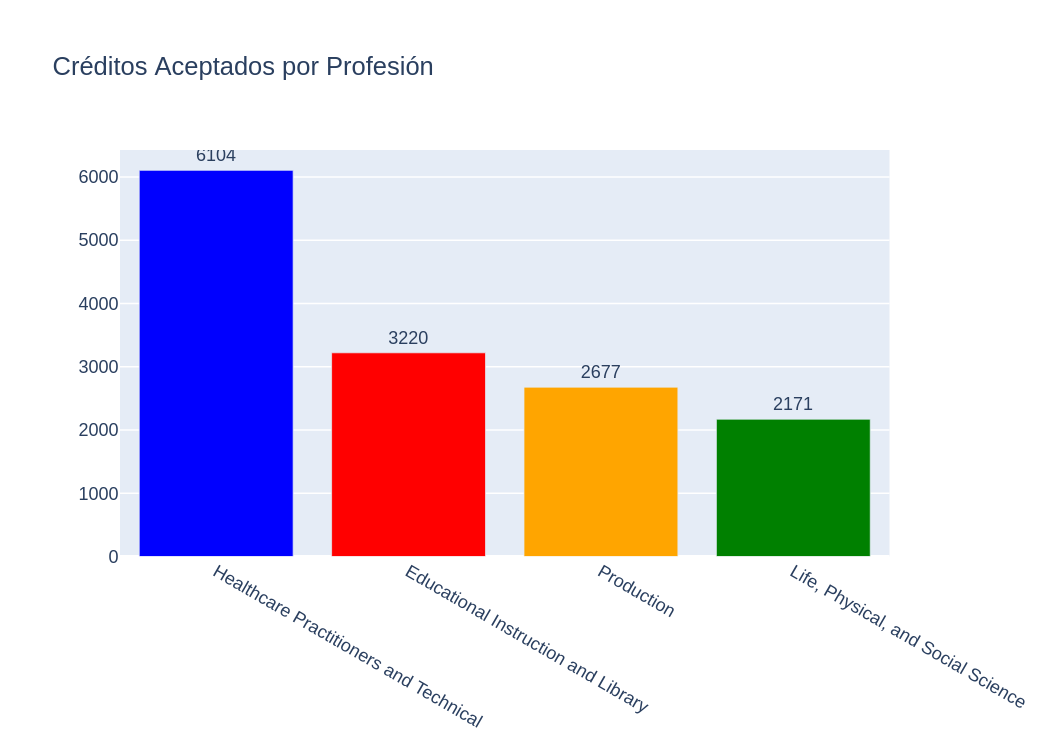
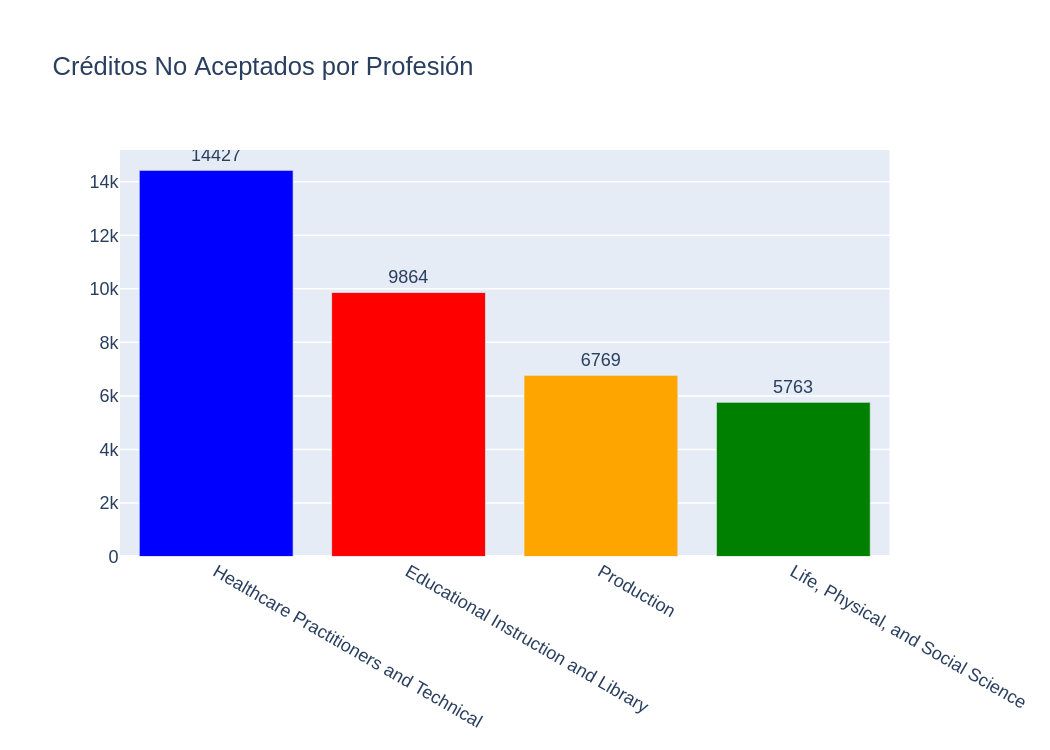
****

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.

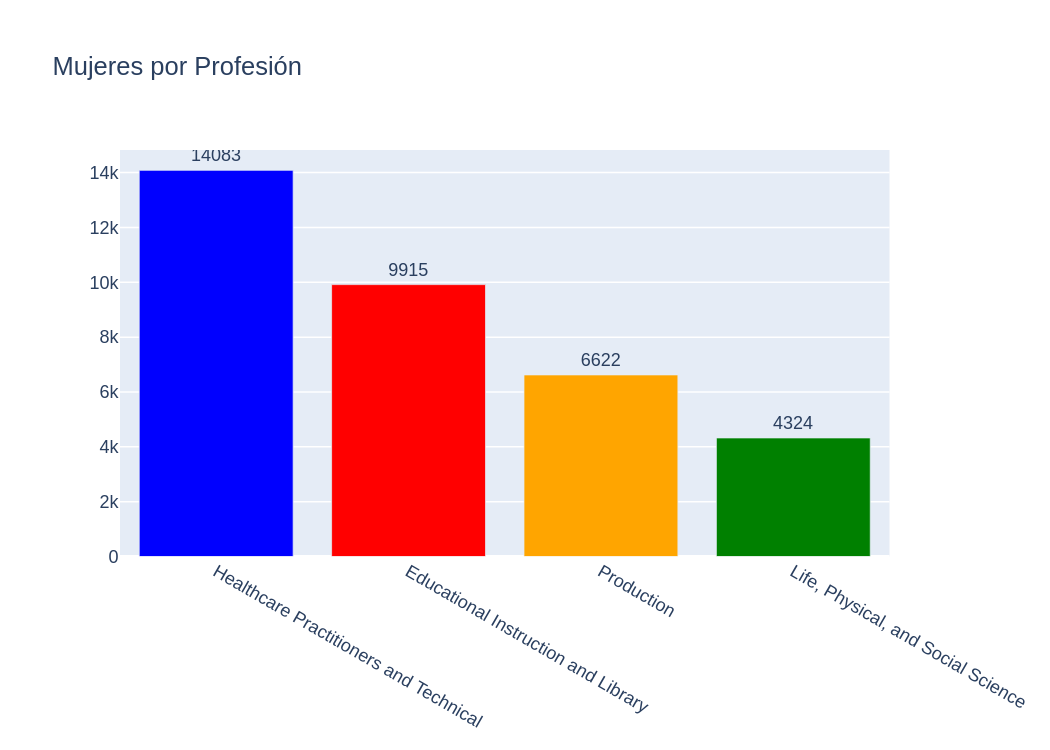
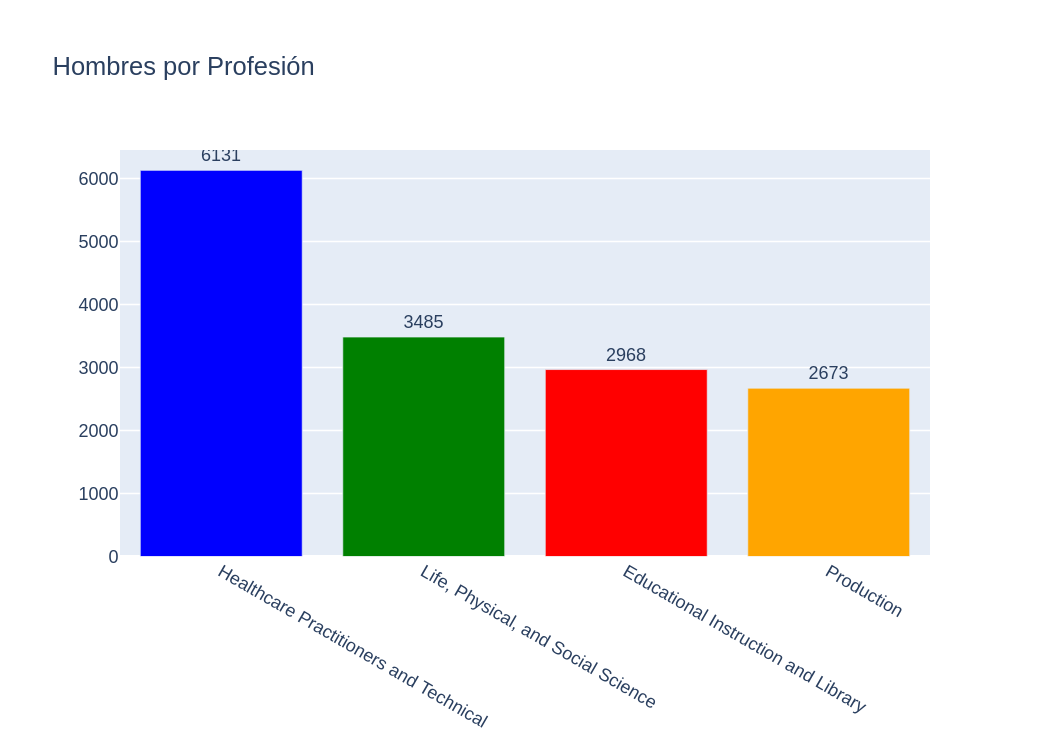
Viendo los rangos de edades observamos que existen valores extremos, para las siguientes gráficas se limitará a un rango entre 15 y 65 años de edad.



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.

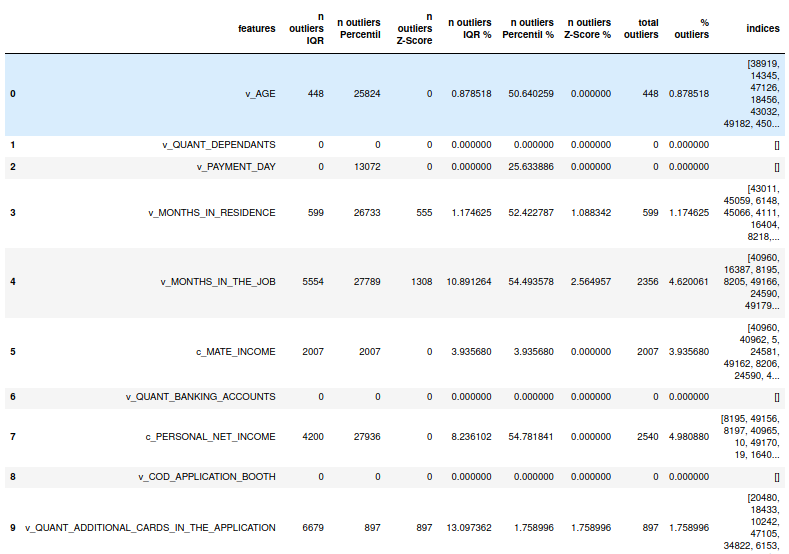


Los profesionistas de la salud tienen más aceptación de los créditos a comparación de las otras profesiones, vemos una baja no proporcional en la aceptación de créditos en las otras profesiones. Esto puede indicar que los profesionales de la salud podrán obtener un crédito con mayor facilidad.



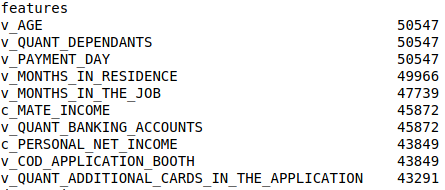
Observamos que los profesionales de la salud en ambos sexos solicitan más créditos, pero en las demás profesiones observamos un cambio en la solicitud de crédito por profesión. Los hombres que trabajan en las ciencias sociales solicitan más créditos proporcionalmente que las mujeres en esta área.

**Datos anómalos**



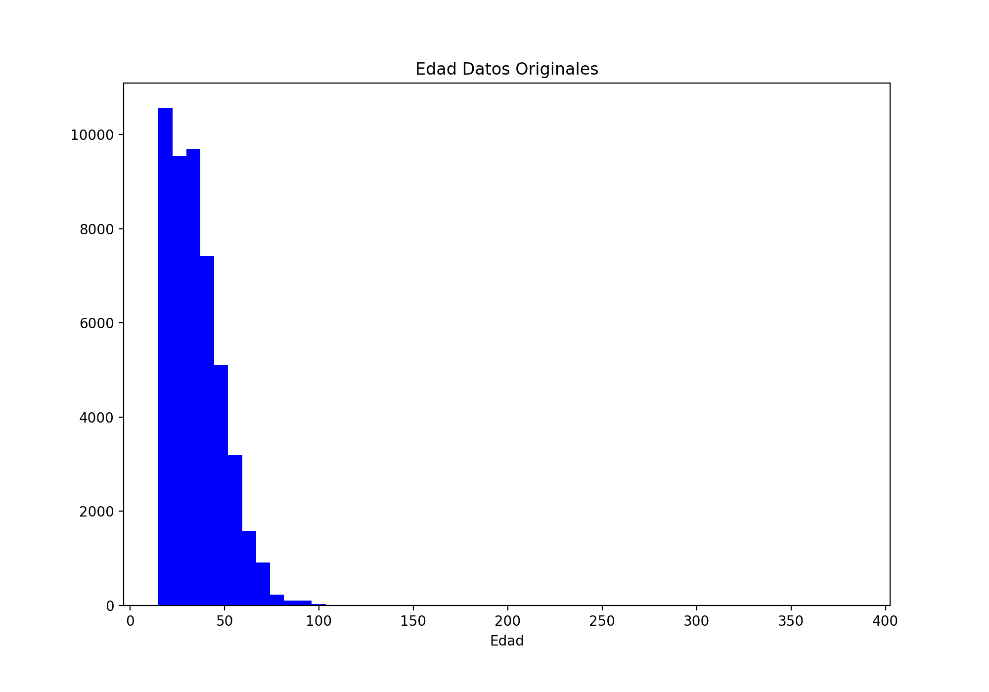
Las variables v\_AGE, v\_MONTHS\_IN\_RESIDENCE tuvieron menos de 600 valores extremos, un porcentaje menor al 1.20% de la tabla total, seguido de v\_QUANT\_ADDITIONAL\_CARDS\_IN\_THE\_APPLICATION con menos de 900 valores extremos lo que representa un 1.75% de la tabla. Finalmente, c\_MATE\_INCOME, c\_PERSONAL\_NET\_INCOME, 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.

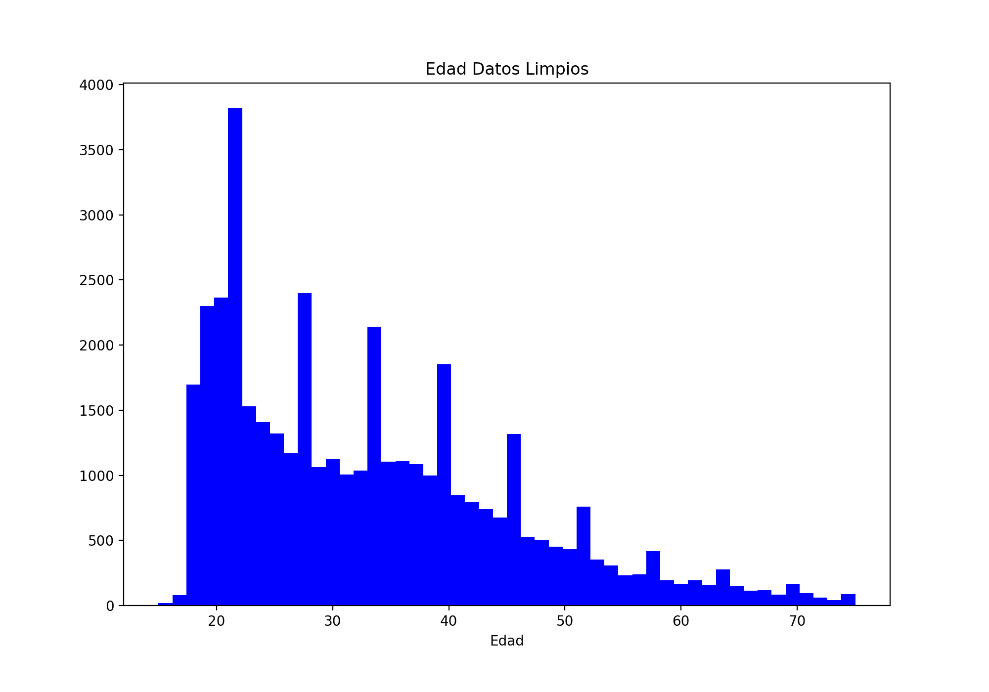
**Número de elementos al eliminar outliers**



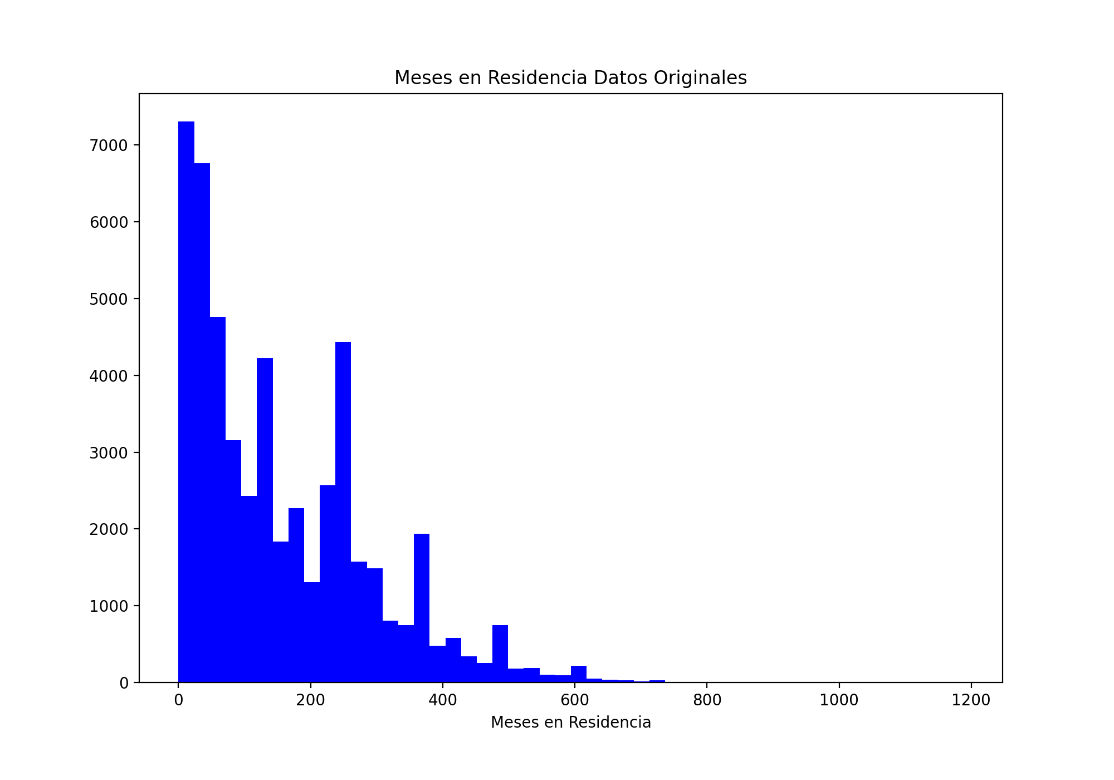
Vemos una reducción de 50,995 registros que se tenían en la tabla original a 43,291, lo que representa un 15.10% de variables eliminadas con respecto a la tabla inicial.

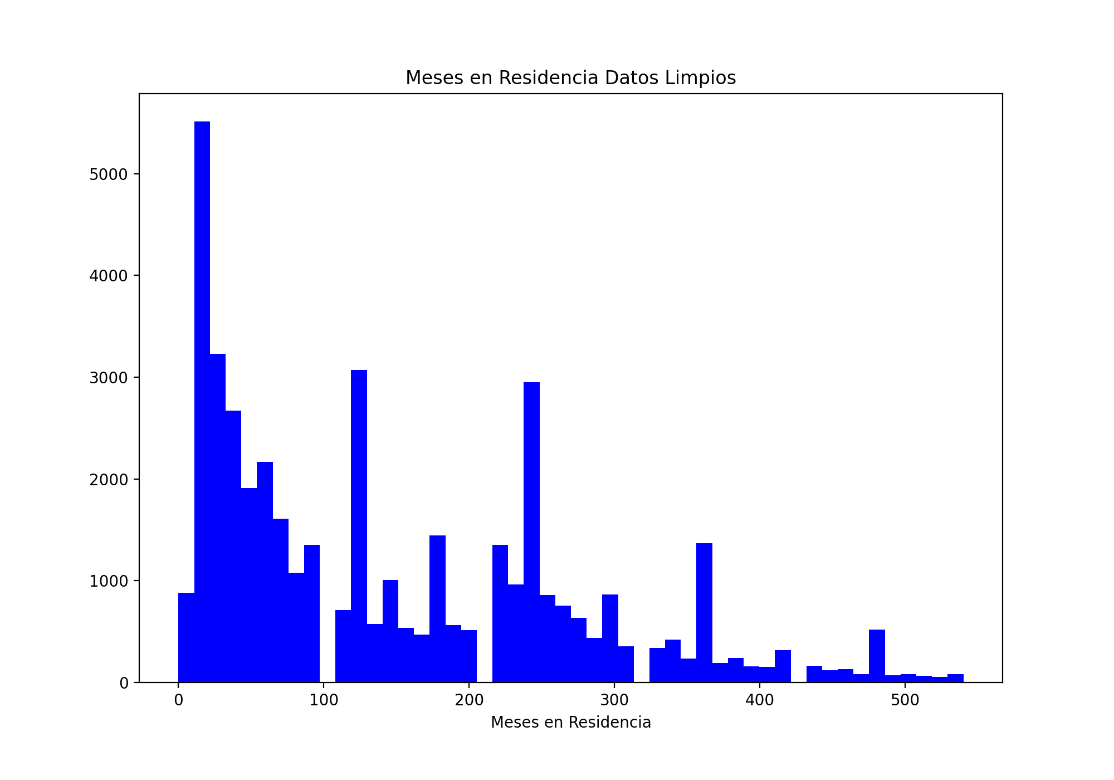
**Además, se debe añadir los gráficos del histograma antes de la remoción y después de la remoción de outliers de todas las variables continuas con datos anómalos.**

****

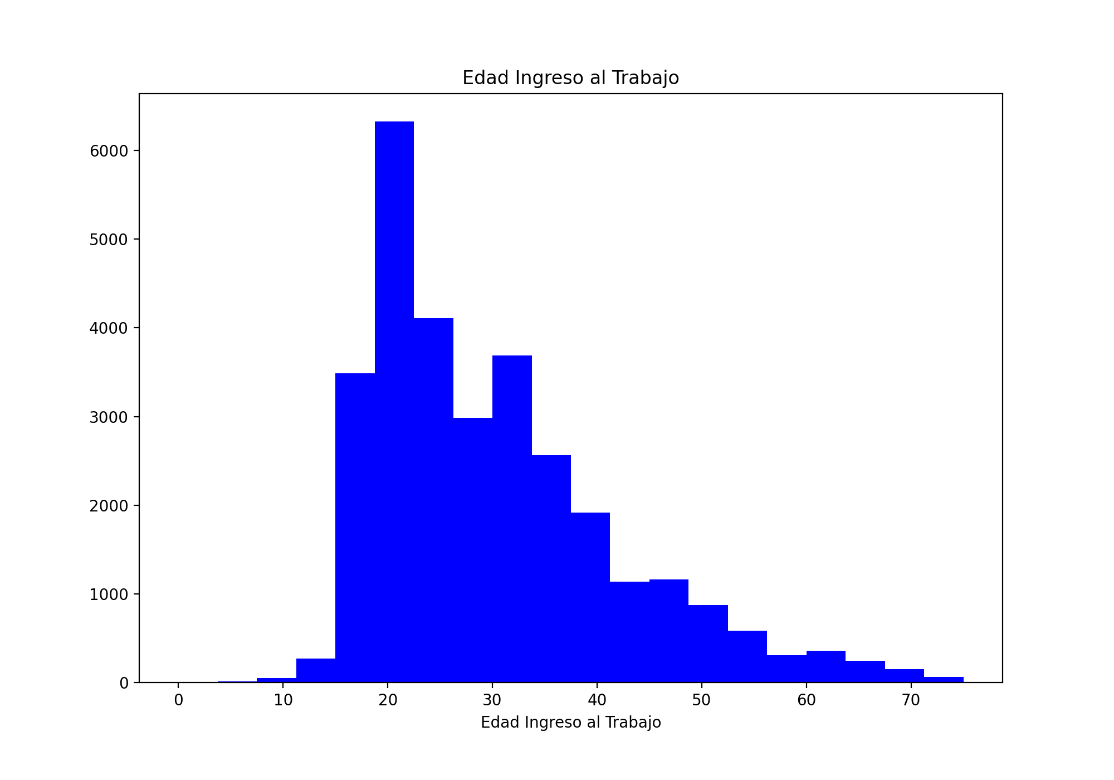
****

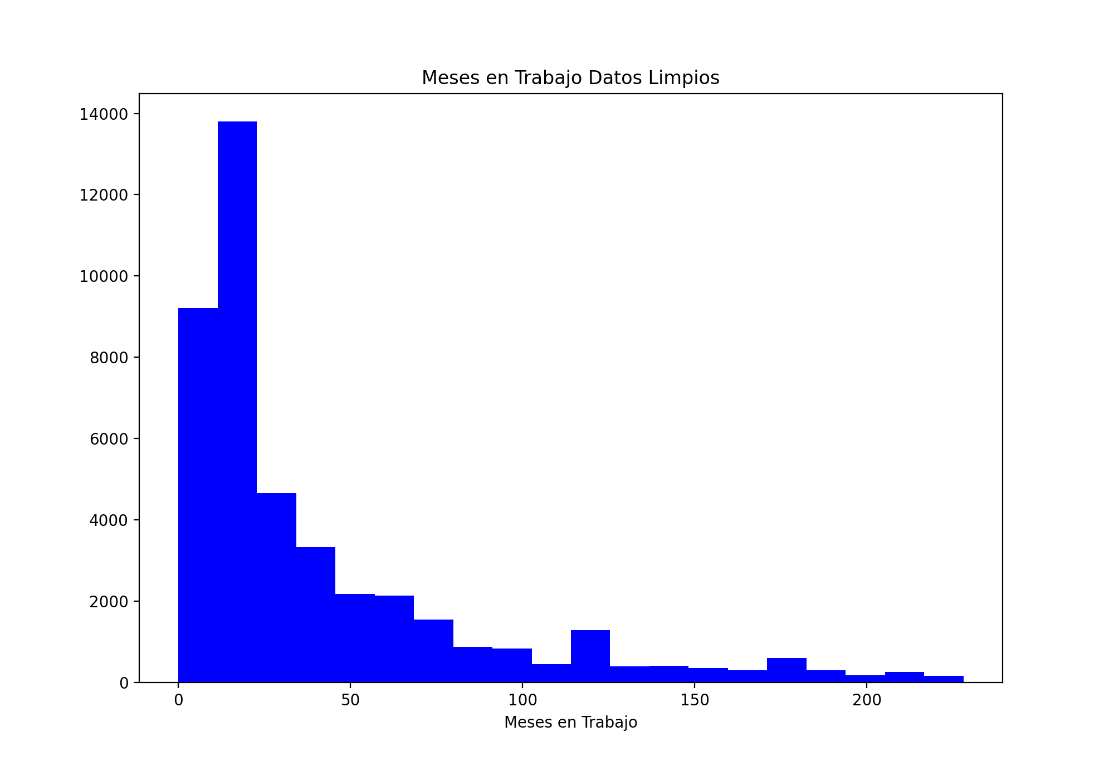
Observamos una reducción en el rango de variables y una concentración de edades entre los 18 a 25 años, esto puede ser ya que las personas de dichas edades se empiezan a independizar de sus familias y necesitan un apoyo económico para lograr sus objetivos.



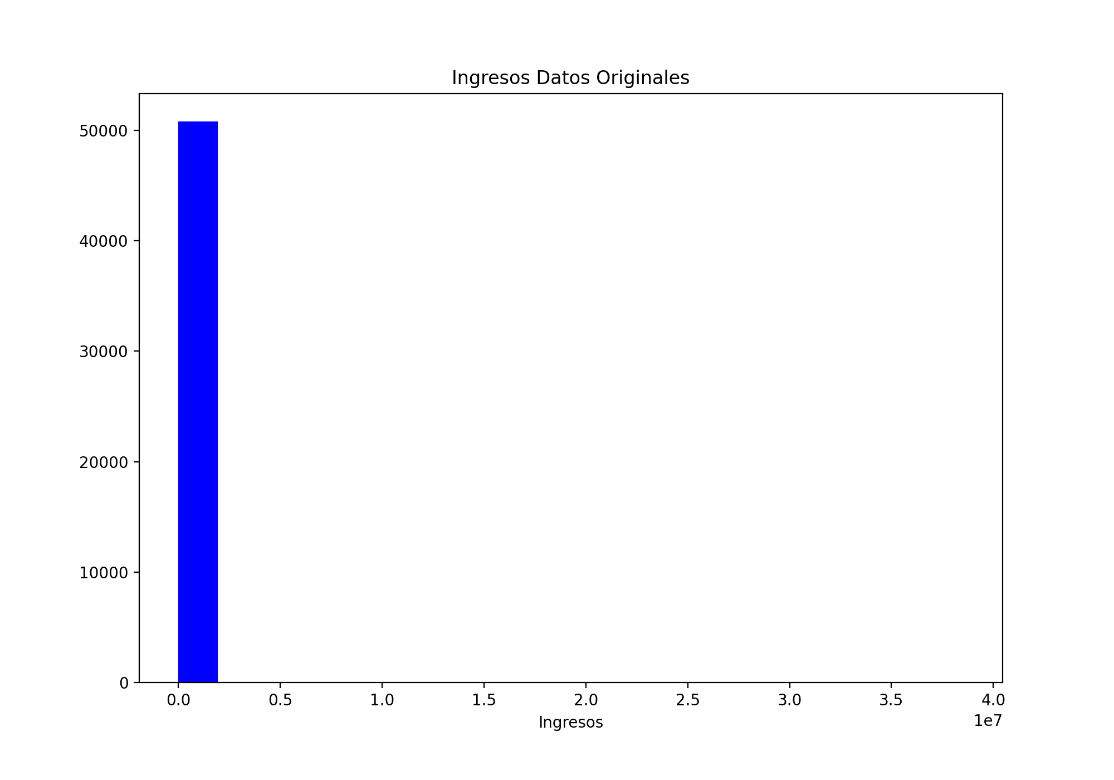


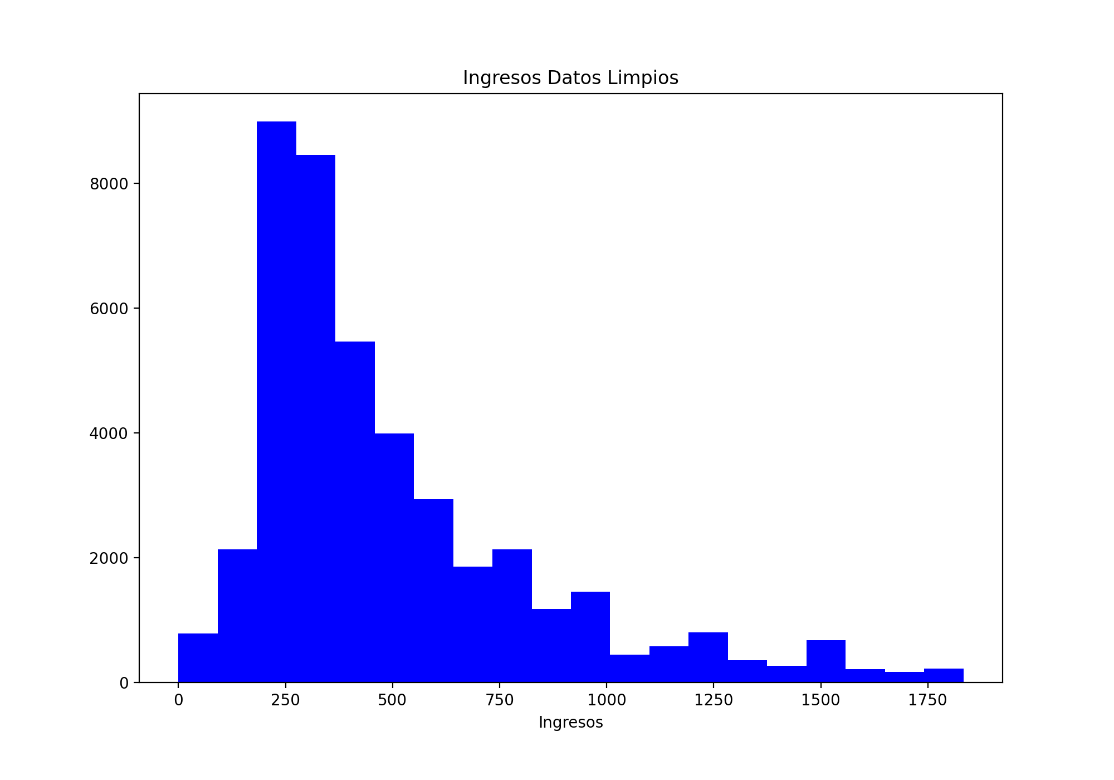
Vemos una reducción en el rango de Meses en residencia, y notamos una división en bloque en cada rango. Se puede observar que la concentración se encuentra en el primer bloque, es decir hay más solicitudes de crédito en personas que llevan menos de 100 meses o menos de una década en su residencia.





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.



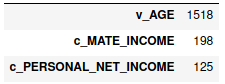


En los datos originales existe un rango muy grande de ingresos, este se disminuye en los datos ya tratados, notamos que las personas que ganan entre 200 a 400 son lo que solicitan más crédito, esto es porque son los individuos que ganan menos en nuestro conjunto de datos.

**Datos faltantes**

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

Variables con valores ausentes:



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:

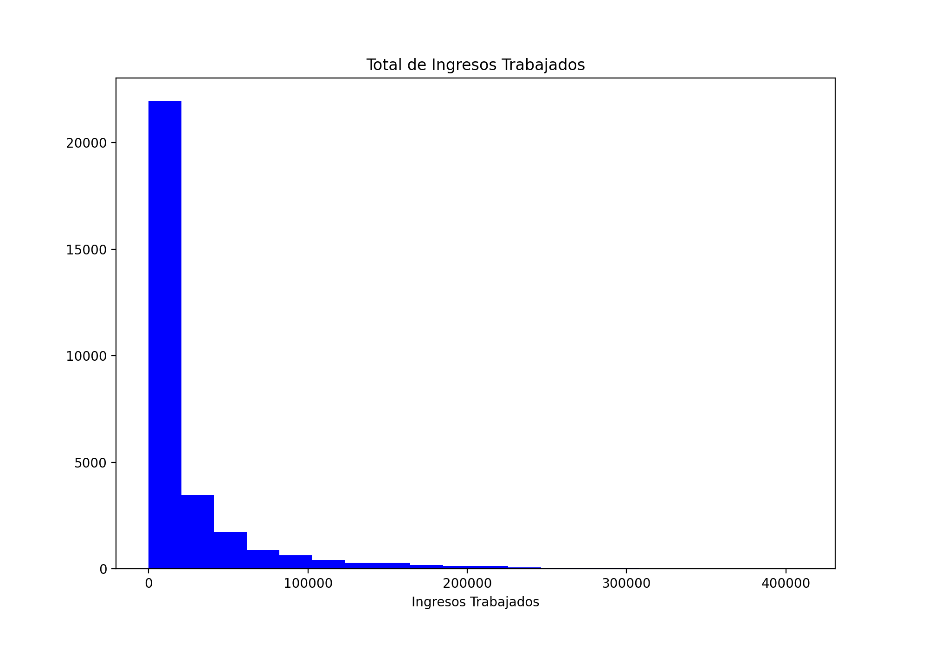
c\_MATE\_INCOME: 0.0

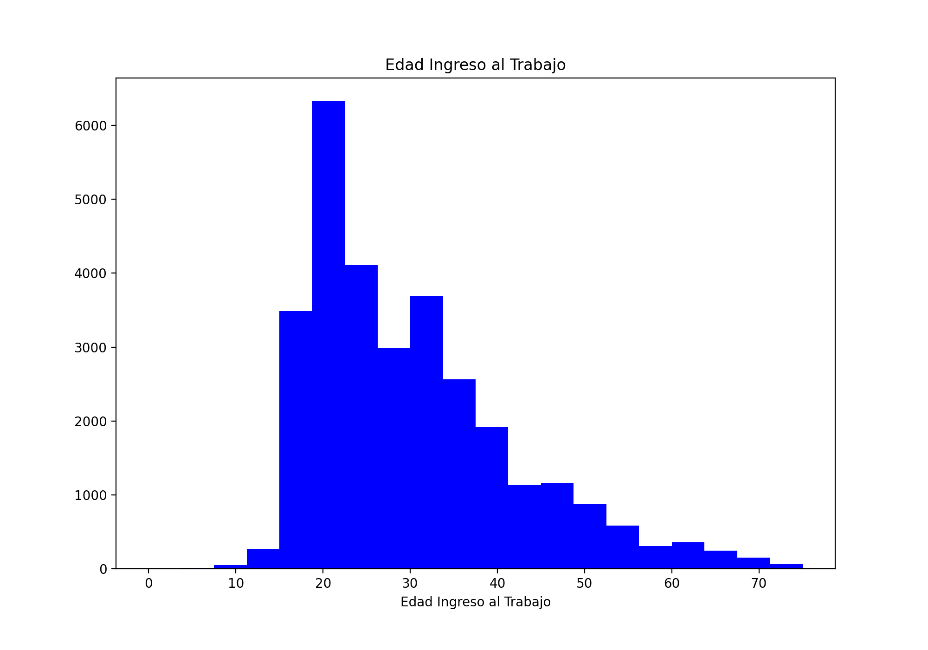
c\_PERSONAL\_NET\_INCOME: 397.0

v\_AGE: 31.0

**Ingeniería de datos**

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

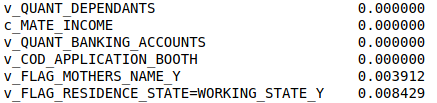
****

****

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.

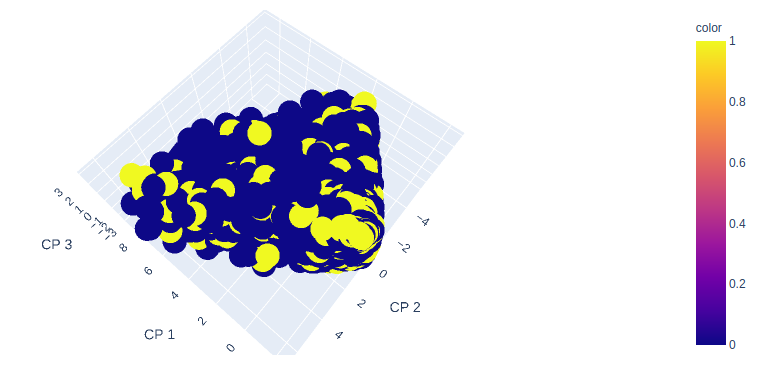
**Reducción de dimensiones**

**Elimine variables con varianza baja**



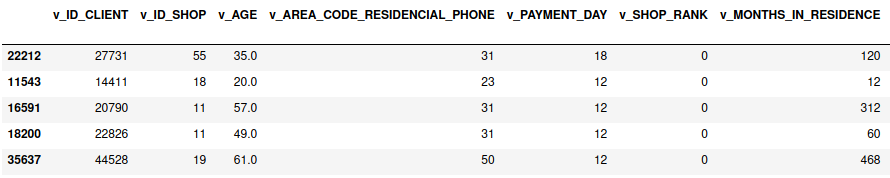
Observamos que las variables v\_QUANT\_DEPENDANTS, c\_MATE\_INCOME, v\_QUANT\_BANKING\_ACCOUNTS, v\_COD\_APPLICATION\_BOOTH, v\_FLAG\_MOTHERS\_NAME\_Y, v\_FLAG\_RESIDENCE\_STATE=WORKING\_STATE\_Y tienen una varianza cercana a 0 por lo que se eliminarán, ya que estas no aportarán nada al modelo.

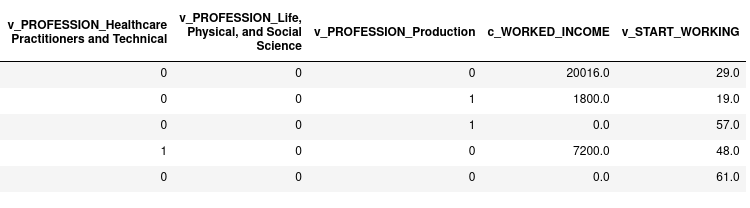
**Tome solo las variables continuas y utilice PCA para generar una visualización en 3D de las variables continuas con la distinción de los valores del target.**



Varianza Explicada: 74.64%

**Tablas Entrenamiento y Test Limpias**





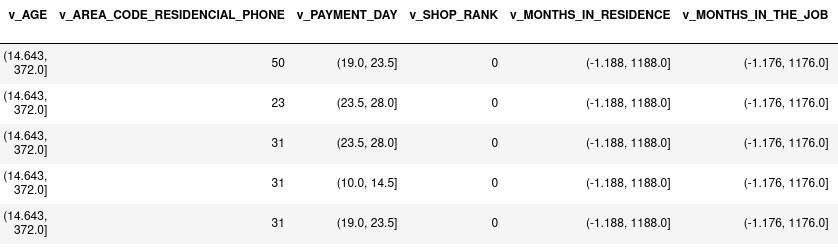
**Tratamiento 2**

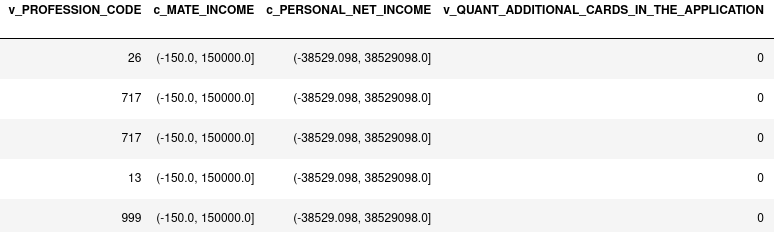
**Transformación entrópica**

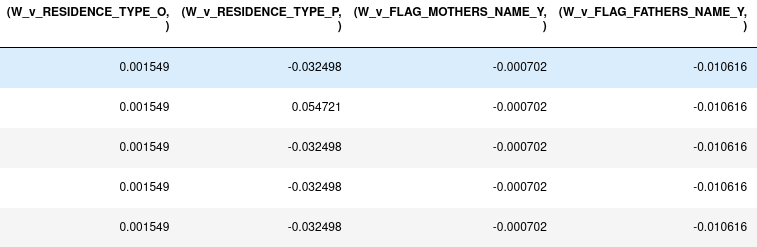
**Varianza Cero**

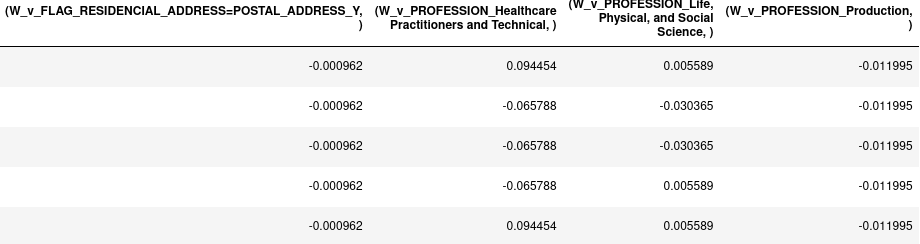


**Conjunto con transformación entrópica**

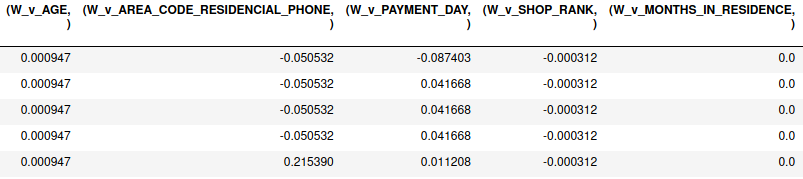


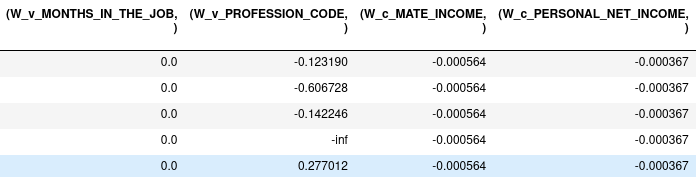


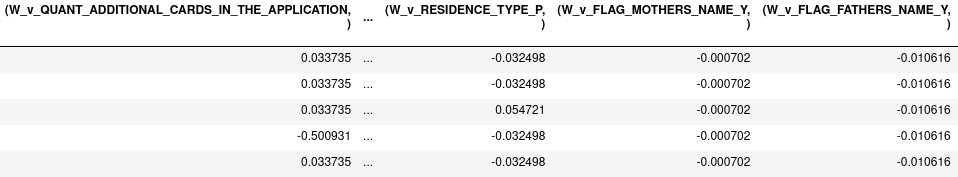


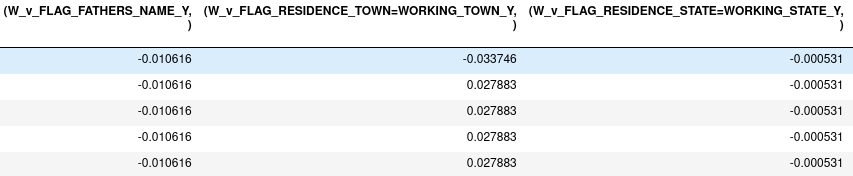


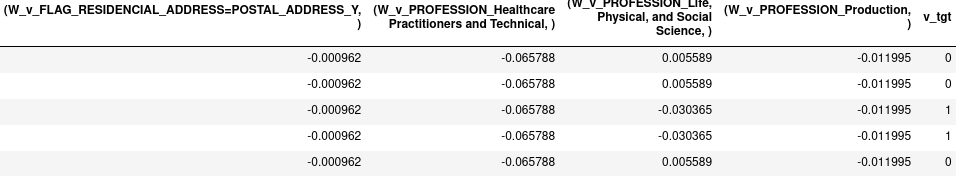
**Conjunto Final Transformaciones Entrópicas**





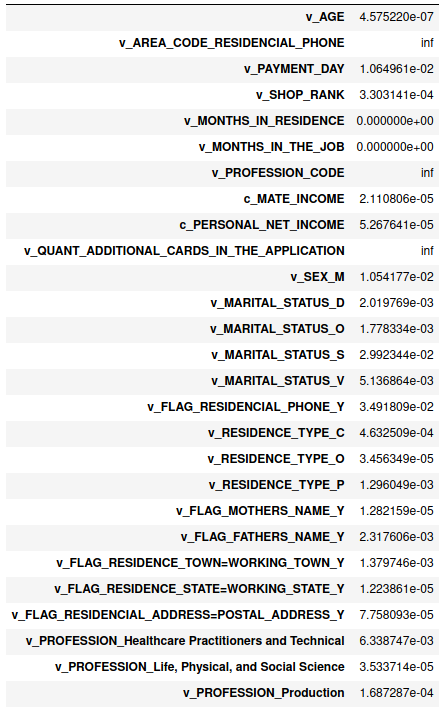






Observamos que todas las variables tienen una transformación entrópica y en la columna final se encuentra el target, esta tabla puede ser utilizada para el entrenamiento de algún modelo de clasificación binaria.

**IV por Variable**



**Cuestionario**