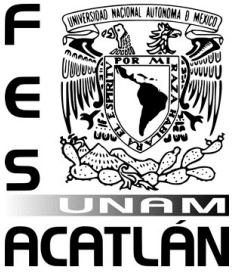

Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Diplomado Ciencia de Datos

Practica 2

Ambrocio Loreto Luis Manuel

Índice

1. Dataset

El conjunto de datos que tenemos a una tabla que contiene características de personas que aplicaron para solicitar un crédito, adicional a esa información contamos con una variable "tgt" que representa a nuestra variable objetivo que toma el valor de1 si se otorgó el crédito y 0 si se rechazó la solicitud, a continuación se muestra el diccionario de datos de las columnas que tiene el dataset

Cuadro 1: Descripción de Variables (Parte 1)

Variable	**Descripción**
ID_CLIENT	Número secuencial para el solicitante (utilizado como clave)
ID_SHOP	Código de la tienda donde se realizó la solicitud
SEX	M=Masculino, F=Femenino
MARITAL_STATUS	S=Soltero, C=Casado, D=Divorciado, V=Viudo, O=Otro
AGE	Edad del solicitante
FLAG_RESIDENCIAL_PHONE	Y=Sí, N=No; Si el solicitante posee un teléfono residencial
AREA_CODE_RESIDENCIAL_PHONE	Código de área modificado del teléfono residencial
PAYMENT_DAY	Día fijo del mes seleccionado para el eventual pago mensual
SHOP_RANK	Calificación de la empresa para la tienda en términos comerciales
RESIDENCE_TYPE	P=Propia, A=Alquilada, C=Casa de los padres, O=Otro
MONTHS_IN_RESIDENCE	Tiempo en la residencia actual en meses
FLAG_FATHERS_NAME	Y=Sí, N=No; Si el solicitante había completado el nombre del padre en el formulario
FLAG_MOTHERS_NAME	Y=Sí, N=No; Si el solicitante había completado el nombre de la madre en el formulario
FLAG_RESIDENCE_TOWN=WORKING_TOWN	Y=Sí, N=No; Si el solicitante trabaja en la misma ciudad donde vive
FLAG_RESIDENCE_STATE=WORKING_STATE	Y=Sí, N=No; Si el solicitante trabaja en el mismo estado donde vive
MONTHS_IN_THE_JOB	Tiempo en el trabajo actual en meses
CÓDIGO_PROFESIÓN	Código de profesión del solicitante
MATE_INCOME	Ingreso mensual neto del cónyuge del solicitante en moneda brasileña (R$)

Cuadro 2: Descripción de Variables (Parte 2)

Variable	**Descripción**
FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS	Y=Sí, N=No; Si el solicitante recibe el correo en la misma dirección donde vive
FLAG_OTHER_CARD	Y=Sí, N=No; Si el solicitante posee otra tarjeta de crédito o de marca privada
QUANT_BANKING_ACCOUNTS	Cantidad de cuentas bancarias del solicitante
FLAG_MOBILE_PHONE	Y=Sí, N=No; Si el solicitante posee un teléfono móvil
FLAG_CONTACT_PHONE	Y=Sí, N=No; Si el solicitante posee un teléfono de contacto
PERSONAL_NET_INCOME	Ingreso mensual neto personal del solicitante en moneda brasileña (R$)
COD_APPLICATION_BOOTH	Código de la caseta donde se entregó la solicitud
QUANT_ADDITIONAL_CARDS_ IN_THE_APPLICATION	Cantidad de tarjetas adicionales solicitadas en el mismo formulario de solicitud
FLAG_CARD_INSURANCE_OPTION	Y=Sí, N=No; Si el solicitante solicitó el servicio de seguro de tarjeta
TARGET_LABEL_BAD=1	Variable Objetivo: MALA=1, BUENA=0

Ahora en la siguiente tabla vemos una muestra de el conjunto de datos y el tipo de dato de cada columna

Cuadro 3: Descripción de Variables y Ejemplo de Datos

Variable	**Tipo de Dato**	**Valor**
ID_CLIENT	int64	1
ID_SHOP	int64	22
SEX	object	F
MARITAL_STATUS	object	O
AGE	int64	44
FLAG_RESIDENCIAL_PHONE	object	N
AREA_CODE_RESIDENCIAL_PHONE	int64	31
PAYMENT_DAY	int64	12
SHOP_RANK	int64	0
RESIDENCE_TYPE	object	P
MONTHS_IN_RESIDENCE	int64	12
FLAG_MOTHERS_NAME	object	Y
FLAG_FATHERS_NAME	object	Y
FLAG_RESIDENCE_TOWN_eq_WORKING_TOWN	object	N
FLAG_RESIDENCE_STATE_eq_WORKING_STATE	object	Y
MONTHS_IN_THE_JOB	int64	48
PROFESSION_CODE	int64	731
MATE_INCOME	float64	0.0
FLAG_RESIDENCIAL_ADDRESS_eq_POSTAL_ADDRESS	object	Y
FLAG_OTHER_CARD	object	N
QUANT_BANKING_ACCOUNTS	int64	0
FLAG_MOBILE_PHONE	object	N
FLAG_CONTACT_PHONE	object	N
PERSONAL_NET_INCOME	object	300
COD_APPLICATION_BOOTH	int64	0
QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	int64	N
FLAG_CARD_INSURANCE_OPTION	object	0
TARGET_LABEL_BAD=1	object	0

En total el conjunto de datos esta compuesto por 29 columnas y 40000 registros

2. Calidad de Datos

2.1. Etiquetado de variables

Renombramos las columnas con la siguiente nomenclatura:

- c_variable aquellas variables que son numéricas.
- v_variable aquellas variables que son categóricas.

El resultado es el siguiente:

Cuadro 4: Nombres de Columnas Originales y Modificados

Nombre Original	**Nombre Modificado**
ID_CLIENT	c_id_client
ID_SHOP	c_id_shop
SEX	v_sex
MARITAL_STATUS	v_marital_status
AGE	c_age
FLAG_RESIDENCIAL_PHONE	v_flag_residencial_phone
AREA_CODE_RESIDENCIAL_PHONE	v_area_code_residencial_phone
PAYMENT_DAY	c_payment_day
SHOP_RANK	c_shop_rank
RESIDENCE_TYPE	v_residence_type
MONTHS_IN_RESIDENCE	c_months_in_residence
FLAG_MOTHERS_NAME	v_flag_mothers_name
FLAG_FATHERS_NAME	v_flag_fathers_name
FLAG_RESIDENCE_TOWN_eq_WORKING_TOWN	v_flag_residence_town_eq_working_town
FLAG_RESIDENCE_STATE_eq_WORKING_STATE	v_flag_residence_state_eq_working_state
MONTHS_IN_THE_JOB	c_months_in_the_job
PROFESSION_CODE	v_profession_code
MATE_INCOME	c_mate_income
FLAG_RESIDENCIAL_ADDRESS_eq_POSTAL_ADDRESS	v_flag_residencial_address_eq_ postal_address
FLAG_OTHER_CARD	v_flag_other_card
QUANT_BANKING_ACCOUNTS	c_quant_banking_accounts
FLAG_MOBILE_PHONE	v_flag_mobile_phone
FLAG_CONTACT_PHONE	v_flag_contact_phone
PERSONAL_NET_INCOME	c_personal_net_income
COD_APPLICATION_BOOTH	v_cod_application_booth
QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	c_quant_additional_cards_in_the_application
FLAG_CARD_INSURANCE_OPTION	v_flag_card_insurance_option
TARGET_LABEL_BAD=1	v_target_label_bad=1

2.2. Duplicidad

Se reviso los datos suplicados por id_cliente y se encontró que había 9 clientes con más de un registro, a continuación se muestra algunas columnas de los registros duplicados, no se muestran todas por practicidad ya que son muchas columnas

Cuadro 5: Datos repetidos por id_cliente

id_cliente	id_shop	v_sex	c_age	v_target_label_bad=1
12	12	F	32	0
12	23	F	30	1
34	3	F	40	0
34	19	F	52	0
67	1	F	38	0
67	15	F	20	0
89	22	F	27	0
89	3	F	59	0
89	25	F	41	0
90	24	F	19	1
90	16	F	20	0
678	12	F	26	0
678	16	F	35	0
1345	55	F	21	0
1345	22	F	35	0
39819	14	F	43	0
39819	3	M	43	1
39995	15	F	40	0
39995	25	F	35	1
39995	1	F	53	0
39995	25	F	19	1

Primero se reviso si había nulos en los registros mostrados, se pensaba conservar para cada cliente aquellos registros que no tuvieran nulos, no se encontraron nulos por lo que se decidió conservar aquellos registros con c_age más alta, esto debido a que cada cliente puedo solicitar la tarjeta en diferentes años y nos es de más interés conservar los datos más actuales, para el cliente 39819 se eliminaron sus registros ya que es inconsistente que haya guardad diferente sexo en cada registro, es lógico que las otras variables cambien pero el sexo siempre será el mismo, a continuación se muestran los registros que se mantuvieron

Cuadro 6: Datos que se mantuvieron para clientes duplicados

id_client	id_shop	v_sex	c_age	v_target_label_bad=1
12	11	F	32	0
34	255	F	52	0
67	66	F	38	0
89	502	F	59	0
90	89	F	20	0
678	821	F	35	0
1345	176	F	35	0
39995	39926	F	53	0

2.3. Precisión/Orden

2.3.1. Variables unitarias

Se hizo una revisión de variables unitarias, es decir de aquellas variables que el mayor porcentaje de sus datos corresponden a un solo valor, usamos como umbral 91 %, es decir aquellas variables que tengan un valor cuyo porcentaje del total sea mayor a 91 % se consideraron como unitarias, las variables identificadas como unitarias son las siguientes:

- c_shop_rank
- v_flag_mothers_name
- v_flag_fathers_name
- v_flag_residence_state_eq_working_state
- c_mate_income
- v_flag_residencial_address_eq_postal_address
- v_flag_other_card
- c_quant_banking_accounts
- v_flag_mobile_phone
- v_flag_contact_phone
- v_cod_application_booth
- v_flag_card_insurance_option

A continuación se muestra un gráfico de barras para cada variable unitaria donde se observa que cada variable tiene el mayor porcentaje de sus datos en solo una categoría

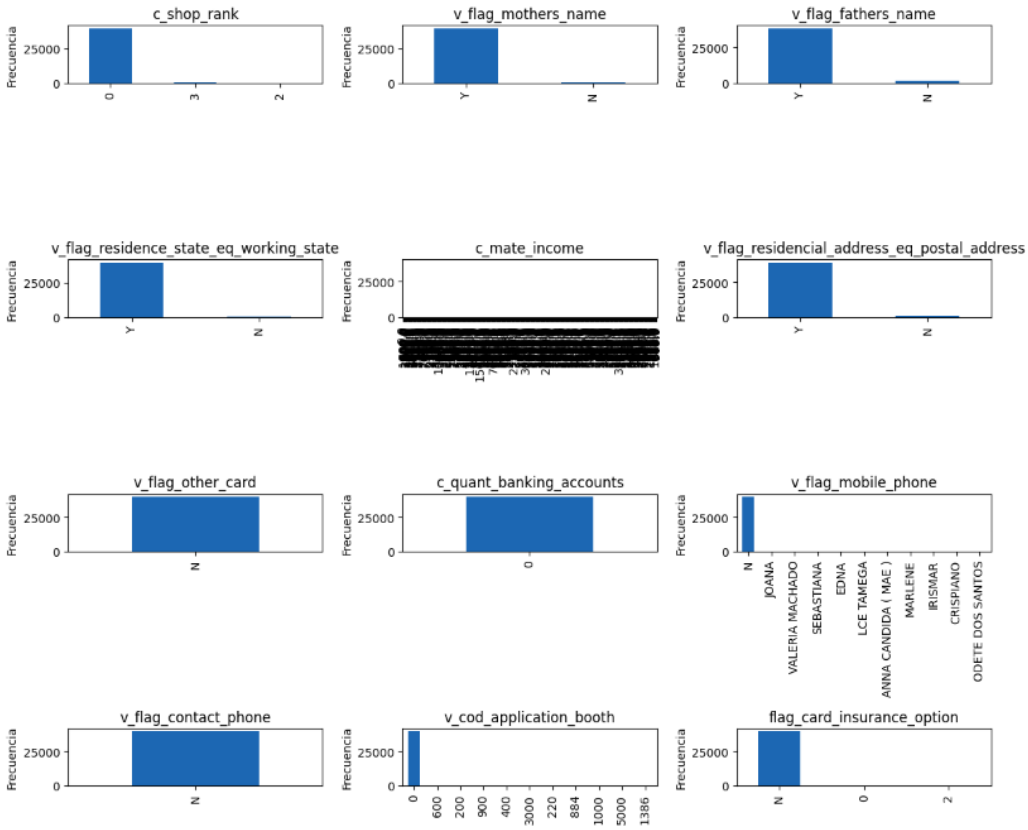

Figura 1: Gráfica de barras de variables unitarias

La variable c_mate_income no se observa bien ya que tiene un rango de valores más amplio, pero a continuación se muestra una parte de la tabla de frecuencias para esta variable donde se observa que el mayor porcentaje de datos cae dentro de un solo valor

Cuadro 7: Tabla de frecuencias de c_mate_income

Valor	**% de Aparición**
0.0	96.11 %
1000.0	0.19 %
600.0	0.16 %
800.0	0.15 %
500.0	0.15 %
...	...

Estas variable se eliminaron del Dataset por ser variables unitarias y proporcionar poca información.

2.3.2. Variables Categóricas

Solo se detectaron errores para la target, en la siguiente gráfica de frecuencias se puede observar el error:

Cuadro 8: Tabla de frecuencias de v_target_label_bad=1

v_target_label_bad=1_valores	**%_aparición**
0	80.18 %
1	19.79 %
N	0.03 %

La variable v_target_label_bad=1 solo debería de contener 0 y 1 y contiene en algunos registros 'N', se dedujo que esto es porque en la mayoría de variables categóricas manejan 'N' para no y 'S' para si, para la target en lugar de manejar 'N' y 'S' se manejan 0 y 1, entonces en lugar de colocar 0 los usuarios pudieron colocar 'N' por equivocación, por lo tanto la solución que se dio es sustituir 0 por 'N', el total de registros afectados fueron 12. Las otras variables categóricas estan correctas.

2.3.3. Variables numéricas

Para la variable c_personal_net_income se detecto que el tipo de dato era Object, cuando debería ser Float, se revisó y se encontraron 'N' al igual que en la variable target, a contención se muestras los registros que se encontraron en esas condiciones

Cuadro 9: Selección de Columnas: id_client, c_personal_net_income, v_target_label_bad=1

id_client	**c_personal_net_income**	**v_target_label_bad=1**
895	N	0
2725	N	0
16163	N	0
28275	N	0
29344	N	0
30811	N	0
31503	N	0
32049	N	0
34468	N	0
35659	N	0
36625	N	0
37349	N	0

Esto es incorrecto, para este caso no aplica un error de valor como en la Target, en este caso pudieron haberse equivocado de columna, sin embargo, no hay forma de saber con que columna se equivocaron,y , el valor correcto del Ingreso mensual , se dejo el valor en Nulo para después imputar, y se convirtió la columna a Float.

2.4. Consistencia

Se encontró una posible inconsistencia en la variable c_months_in_the_job, primero se convirtieron los meses a años para tener una mejor perspectiva del tiempo laborando en el actual trabajo, y luego se obtuvieron estadísticas de los años laborando, se encontró lo siguiente:

Cuadro 10: Estadísticas de años laborando

Estadística	**Valor**
Conteo	39,987
Media	4.22
Desviación Estándar	6.19
Mínimo	0
Cuartil 25 %	1
Cuartil 50 % (Mediana)	2
Cuartil 75 %	5
Máximo	98

El valor que causo ruido fue el valor máximo, es difícil que alguien dure 98 años trabajando, se hizo una comparación entre los años de vida y los años laborando, se encontró que había casos en los que los años laborando era mayor a los años de vida (entre ellos el dato de 98 años laborando), a continuación se observan los datos inconsistentes:

Cuadro 11: Valores inconsistentes en c_months_in_the_job

id_client	**c_age**	**c_months_in_the_job**	**c_age_in_the_job**	**v_target_label_bad=1**
3323	22	780	65.0	0
5465	23	276	23.0	0
5984	18	852	71.0	0
6101	51	612	51.0	0
9964	21	360	30.0	0
10641	53	1104	92.0	0
14986	32	432	36.0	0
15771	43	600	50.0	0
15861	57	684	57.0	0
17499	22	324	27.0	1
18914	19	240	20.0	1
22636	19	228	19.0	0
25702	24	1176	98.0	0
26926	18	216	18.0	1
27856	23	276	23.0	0
28500	33	420	35.0	0
28995	20	324	27.0	0
29352	30	360	30.0	0
31251	20	252	21.0	0
32734	19	396	33.0	0
35034	30	360	30.0	0
35855	24	324	27.0	0
39908	21	480	40.0	0

No hay forma de determinar el valor correcto de estos datos, por lo que en estos registros la columna c_months_in_the_job se dejó en Nulo para después imputar, así se ven las estadísticas después de dejar en nulo los registros:

Cuadro 12: Estadísticas de años laborando quitando datos inconsistentes

Variable	**Valor**
Conteo	39,964
Media	4.20190755
Desviación Estándar	6.10538908
Mínimo	0.0
Cuartil 25 %	1.0
Cuartil 50 % (Mediana)	2.0
Cuartil 75 %	5.0
Máximo	59.0

Para la variable c_months_in_residence encontramos algo igual, de igual forma se convirtió a años los meses en la residencia actual, las estadísticas de los años en la residencia actual se muestran a continuación

Cuadro 13: Estadísticas de años en la residencia actual

Estadística	**Valor**
Conteo	39,987
Media	12.78
Desviación Estándar	11.37
Mínimo	0
Cuartil 25 %	3
Cuartil 50 % (Mediana)	10
Cuartil 75 %	20
Máximo	99

De igual forma hay valores muy altos,y , se hizo una comparó con los años de vida y se encontraron 845 registros que no cumplían con que la edad fuera mayor o igual a los años de residencia, a continuación se muestran solo algunos de estos registros:

Cuadro 14: Datos inconsistentes en c_months_in_residence

id_client	**c_age**	**c_months_in_residence**	**c_age_in_residence**	**v_target_label_bad=1**
39661	18	228	19.0	0
39684	20	252	21.0	1
39721	24	300	25.0	0
39764	45	552	46.0	0
39768	26	324	27.0	1

Usando el mismo método , se convirtieron los datos de c_months_in_residence a nulo para después imputar, las estadísticas actualizadas se muestran a continuación

Cuadro 15: Estadísticas de años en la residencia actual quitando datos inconsistente

Estadística	**Valor**
Conteo	39,142
Media	12.437
Desviación Estándar	11.144
Mínimo	0
Cuartil 25 %	3
Cuartil 50 % (Mediana)	10
Cuartil 75 %	20
Máximo	76

2.5. Normalización

Se detectaron algunas variables categóricas que se podían normalizar, por ejemplo en el siguiente recuadro se observa la frecuencia de datos para cada estado civil, se observa que la mayor parte de los registros se encuentran entre usuarios Solteros y Casados, así se considero buena opción juntar Viudo y Divorciado con la categoría Otro.

Cuadro 16: Distribución de frecuencias relativas del Estado Civil

Estado Civil	**Proporción**
Soltero (S)	50.94 %
Casado (C)	34.30 %
Otro (O)	5.57 %
Viudo (V)	4.89 %
Divorciado (D)	4.30 %

La distribución de frecuencias relativas después de la normalización se ve de la siguiente forma:

Cuadro 17: Distribución de frecuencias relativas del Estado Civil después de normalizar

Estado Civil	**Proporción**
Soltero (S)	50.94 %
Casado (C)	34.30 %
Otro (O)	14.76 %

Para la variable v_area_code_residencial_phone se encontró algo similar, la mayor cantidad de datos se encuentra en dos categorías, la demás categorías tienen una proporción muy baja de y es mejor normalizar ya que no aportan información significativa, a continuación se muestran las frecuencias relativas de v_area_code_residencial_phone.

Cuadro 18: Distribución de frecuencias relativas de v_area_code_residencial_phone

cod	Proporción	cod	Proporción	cod	Proporción
31	70.1878 %	24	0.2451 %	49	0.1025 %
50	22.2497 %	32	0.0675 %	27	0.0650 %
5	4.8641 %	38	0.0275 %	42	0.0250 %
23	1.9431 %	52	0.0250 %	56	0.0225 %
68	0.0125 %	33	0.0100 %	11	0.0075 %
41	0.0075 %	43	0.0075 %	8	0.0075 %
26	0.0075 %	9	0.0050 %	48	0.0050 %
12	0.0050 %	7	0.0050 %	1	0.0050 %
34	0.0050 %	25	0.0025 %	61	0.0025 %
10	0.0025 %	58	0.0025 %	13	0.0025 %
3	0.0025 %	40	0.0025 %	19	0.0025 %
20	0.0025 %	67	0.0025 %	30	0.0025 %
35	0.0025 %	28	0.0025 %	60	0.0025 %
69	0.0025 %	39	0.0025 %	36	0.0025 %
37	0.0025 %	62	0.0025 %	29	0.0025 %
15	0.0025 %	46	0.0025 %	44	0.0025 %
45	0.0025 %	18	0.0025 %	6	0.0025 %
17	0.0025 %	2	0.0025 %	14	0.0025 %
54	0.0025 %	22	0.0025 %	59	0.0025 %
53	0.0025 %	21	0.0025 %		

La forma en que normalizamos fue conservar el código 31 y 50 y los demás códigos, al tener una proporción muy pequeña, juntarlos en una categoría llamada otros, la distribución de frecuencias relativas actualizada se muestran en la siguiente tabla.

Cuadro 19: Distribución de frecuencias relativas de v_area_code_residencial_phone después de normalizar

v_area_code_residencial_phone	Proporción
31	70.1878 %
50	22.2497 %
Otros	7.5625 %

Para la variable v_profession_code hay valores entre 0 y 999, para este caso no hay categorías dominantes, por lo que no se podría normalizar de la misma forma que los dos anteriores, lo que se hizo fue dividir en intervalos, se crearon 5 y son los siguientes:

- 0 - 200
- 201 - 400
- 401 - 600
- 601 - 800
- 801 - 999

La distribución de frecuencias relativas se ve de la siguiente forma:

Cuadro 20: Distribución de frecuencias relativas v_profession_code después de la normalización

v_profession_code	**Proporción**
0-200	35.9842 %
801-999	30.3073 %
601-800	15.6976 %
201-400	11.2687 %
401-600	6.7422 %

3. Análisis Exploratorio de Datos

Realizando un gráfico de barras para la target observamos que la mayoría de los créditos fueron otorgados (0), solo cerca del 20 % fueron rechazados.

Figura 2: Gráfica de barras para la target

También observamos que la mayoría de personas que aplicaron al crédito fueron mujeres, cerca del 70 % y solo cerca del 30 % fueron hombres

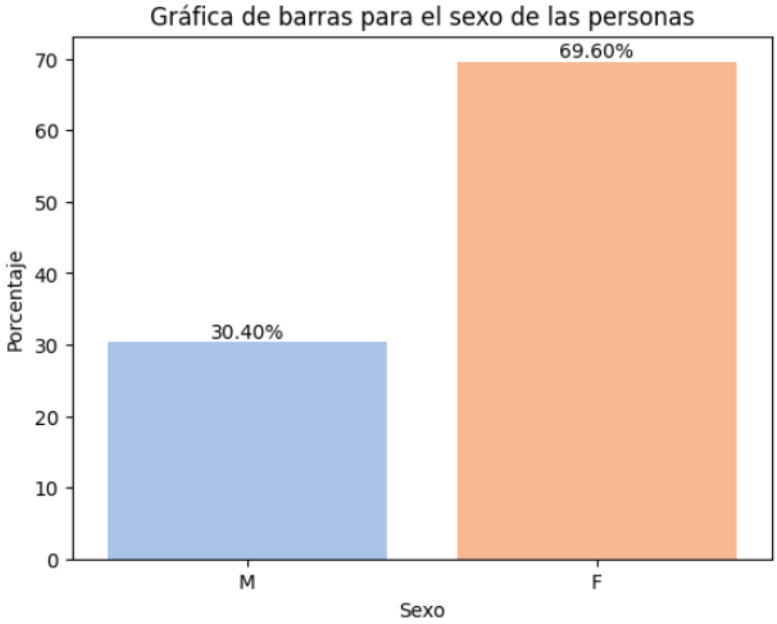

Figura 3: Gráfica de barras para el sexo

Realizando un mapa de calor vemos que más del 50 % de registros corresponden a mujeres que fueron aceptadas en el crédito, y solo el 23 % de los registros fueron de hombres a los que les aceptaron el crédito, con esto parecería que es más probable que a una mujer le acepten el

crédito, pero hay que recordar que la mayoría de solicitantes son mujeres, podemos recurrir a otra gráfica para tener una mejor visión.

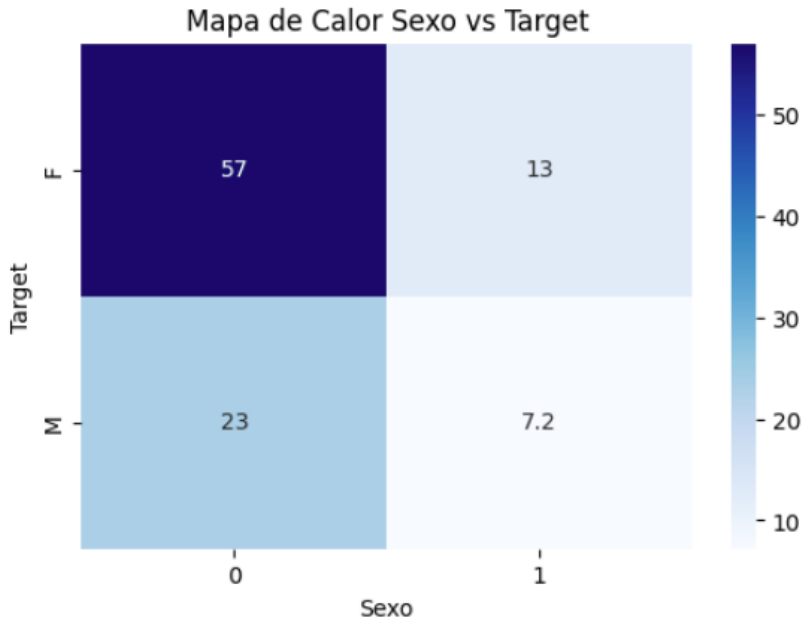

Figura 4: Mapa de calor sexo vs target

En esta gráfica de barras apiladas observamos que del total de mujeres a poco más del 81 % les aceptan el crédito y a cerca del 76 % les aceptan el crédito en el caso de los hombres, ahora si vemos que es un poco más probable que a una mujer le acepten el crédito, pero solo un poco, no hay una diferencia muy grande como se veía en el mapa de calor.

Figura 5: Gráfica de barras para sexo y target

Hacemos una comparación similar para el tipo de residencia, vemos que para aquellas personas con residencia propia les es un poco más probable adquirir el crédito en comparación con personas que viven rentando, esto suena lógico, pero de nuevo la diferencia no es grande.

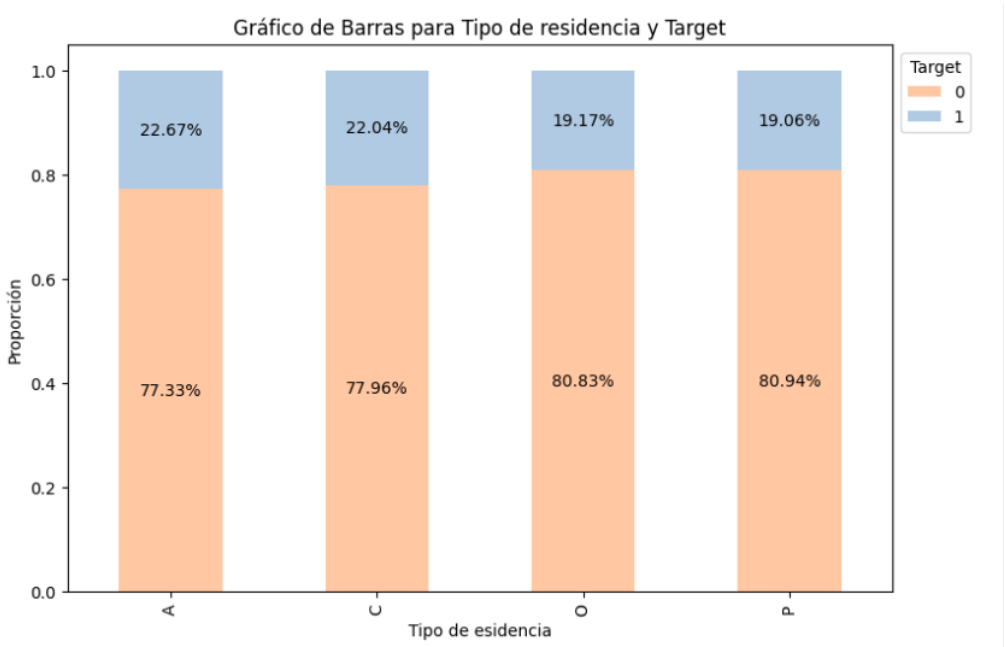

Figura 6: Gráfica de barras para tipo de residencia y target

Para el estado civil también hay diferencias (de casi el 10 %) en la probabilidad de adquirir un crédito, aparentemente es más probable adquirir un crédito estando casa que soltero, esto también puede estar relacionados a que la mayoría de las personas casadas son más grandes y tienen una mayor estabilidad económica (claro, no es una regla general).

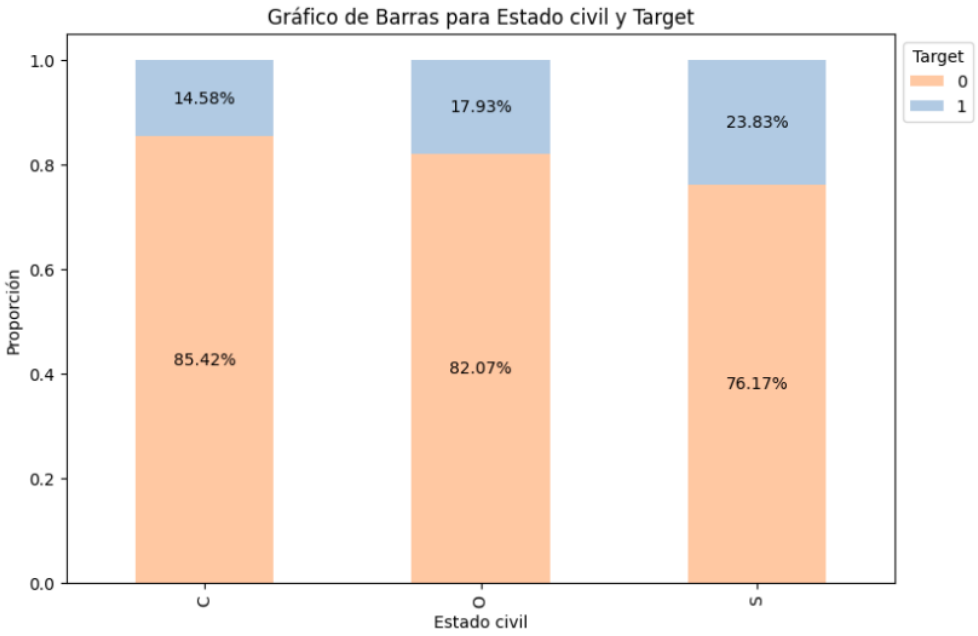

Figura 7: Gráfica de barras para estado civil y target

Ahora pasamos a analizar las variables continuas, primero vamos a hacer una comparación con la target, en el siguiente gráfico se observa un gráfico de caja y bigotes de la edad para el caso cuando fue aprobado el crédito y para el caso cuando fue rechazado el crédito, vemos que para el caso cuando fue aprobado el crédito la distribución esta ligeramente más hacia arriba, es decir más a edades grandes, esto lo vemos porque los cuartiles están ligeramente más arriba que los cuartiles de el gráfico para cuando fue rechazado el crédito, lo cual indica que la edad

es un factor que afecta a la hora de aprobar el crédito

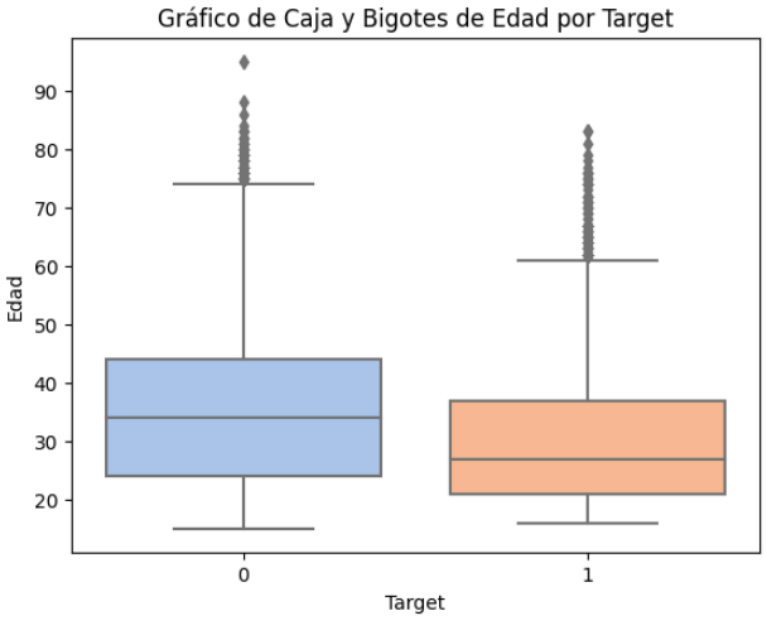

Figura 8: Gráfica de caja y bigotes de edad por target

Hacemos lo mismo para la variable de ingreso mensual, en este caso los gráficos no se observan bien ya que se nota la presencia de posibles datos atípicos, por el momento solo los ignoramos y hacemos otro gráfico.

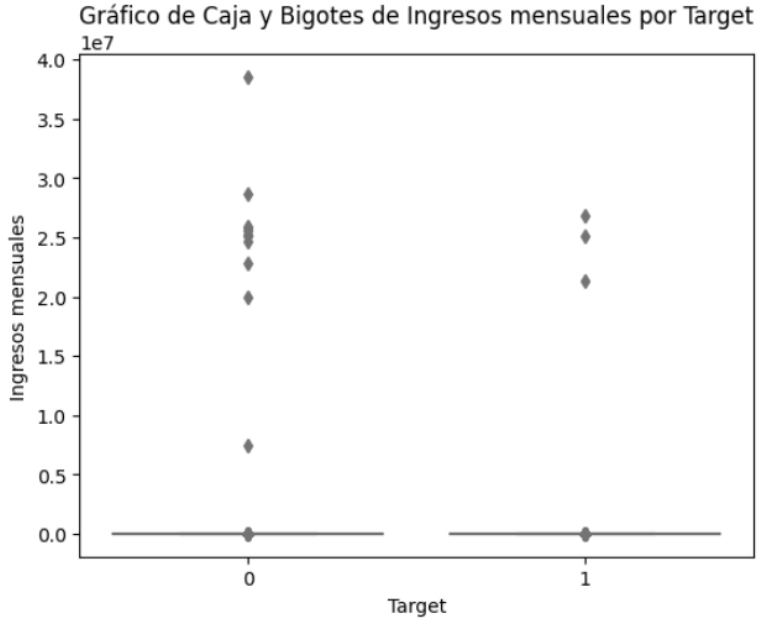

Figura 9: Gráfica de caja y bigotes de ingresos mensuales por target

En este nuevo gráfico de ingresos mensuales encontramos algo similar a la edad, la distribución del ingreso mensual para cuando fue aceptado el crédito esta ligeramente más hacia arriba, aunque en este caso no es tan evidente como en la edad

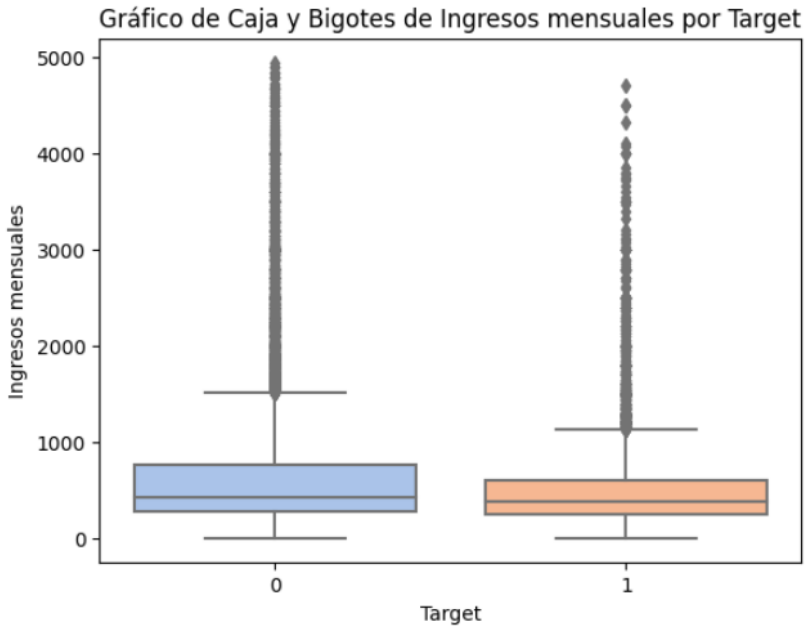

Figura 10: Gráfica 2 de caja y bigotes de ingresos mensuales por target

El siguiente gráfico muestra los mismos gráficos pero en lugar de comparar por la target, comparamos por el sexo, vemos que si hay una pequeña diferencia a favor de los hombres en la distribución de los ingresos mensuales, para este ultimo caso realicemos un agrupamiento y calculemos la media.

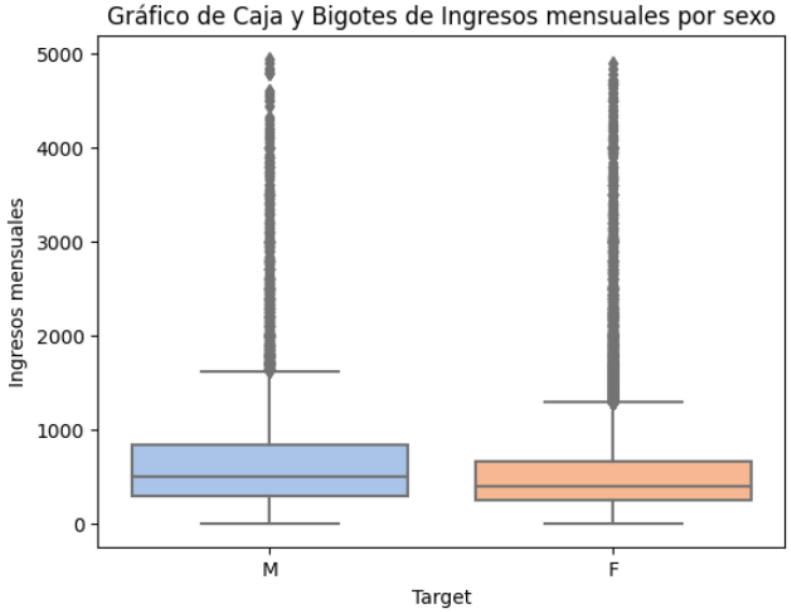

Figura 11: Gráfica de caja y bigotes de ingresos mensuales por sexo

La siguiente tabla muestra la media de ingresos mensuales para hombres y mujeres, confirmamos que si hay una diferencia para los ingresos mensuales entre hombres y mujeres, por lo menos en este conjunto de datos, es importante mencionar que se quitaron los posibles datos atípicos del calculo de la media ya que son demasiados extremos e iban a sesgar el resultado.

Cuadro 21: Media de ingresos mensuales por sexo

Sexo	**Media**
F	586.893367
M	716.835641

En las gráficas de caja y bigotes de meses de residencia por target no observamos una diferencia significativa entre la distribución de meses de residencia para cuando se aprobó el crédito y para cuando se rechazo

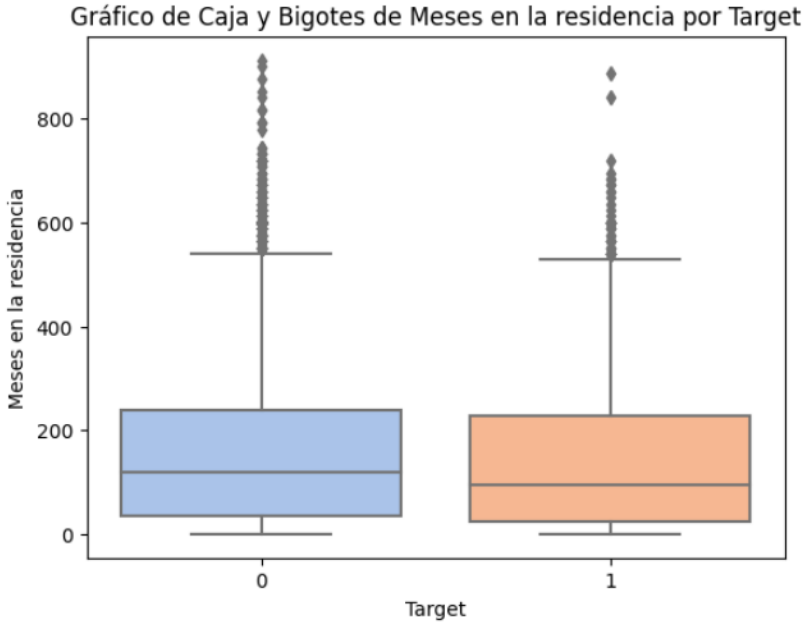

Figura 12: Gráfica de caja y bigotes de meses en la residencia por target

Para el caso de meses en el trabajo de nuevos notamos una diferencia, mostrando que la distribución de meses en el trabajo para cuando se acepto el crédito esta ligeramente hacia arriba, esto es lógico pues pues los bancos prefieren clientes con un empleo ya estable.

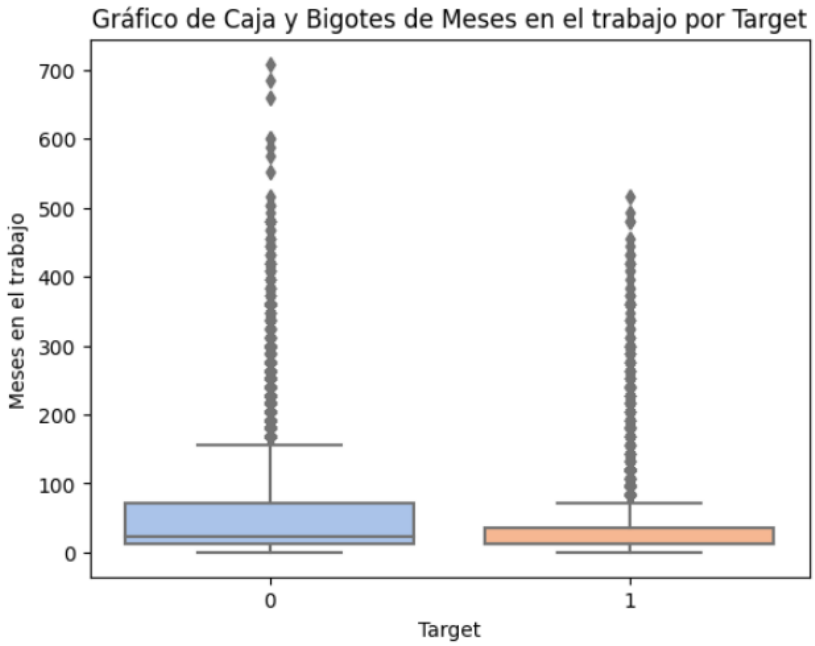

Figura 13: Gráfica de caja y bigotes de meses en el trabajo por target

Revisemos ahora gráficos de dispersión, primero observemos una gráfica de ingresos mensuales vs Edad, vemos de nuevo esos posibles datos atípicos que han estado causando ruido, además vemos que a la mayoría de personas con estos ingreso tan altos se les aprobó el crédito, de nuevo vamos a ignorar estos datos para tener una mejor gráfica.

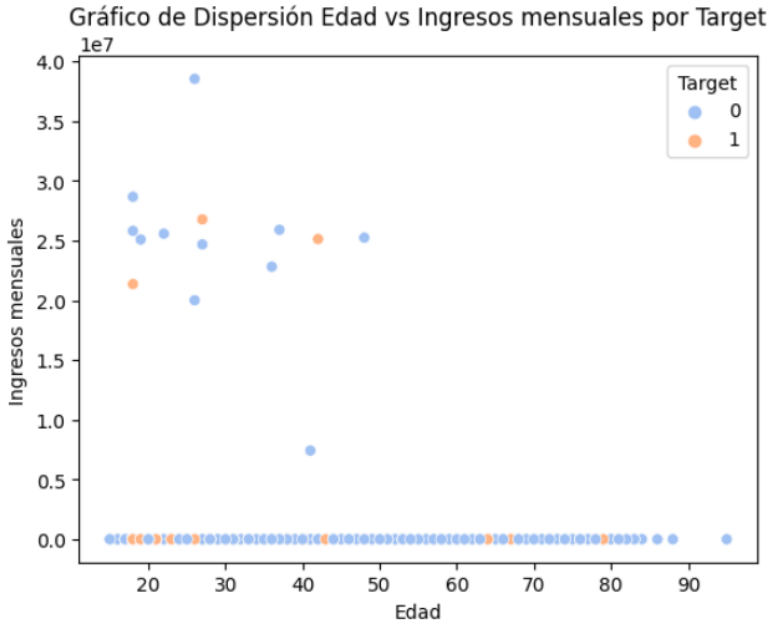

Figura 14: Gráfico de dispersión edad vs ingresos mensuales por target

Vemos algo curioso en esta gráfica, yo me imaginaría que a mayor edad mayor ingreso, pero no hay una relación aparente, incluso podríamos poner dos lineas horizontales donde entrarían la mayor cantidad de registros

En la siguiente gráfica de meses en el trabajo vs ingresos vemos de nuevo datos atípicos, además se observa que corresponden a personas con menos de 100 meses en el trabajo.

La siguiente gráfica igual es de meses en el trabajo vs ingresos pero omitiendo los posibles datos atípicos de los ingresos, se esperaría que a medida que aumentan los meses en el trabajo

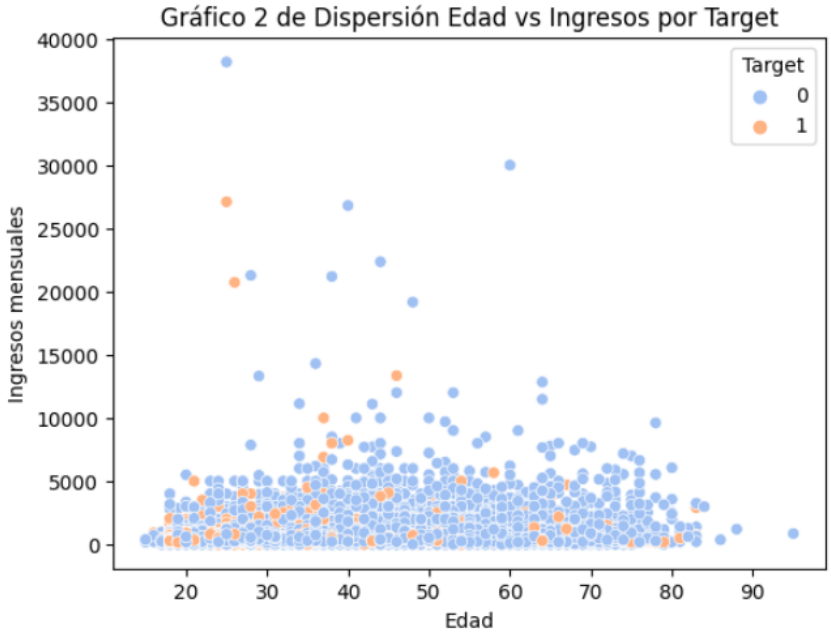

Figura 15: Gráfico 2 de dispersión edad vs ingresos mensuales por target

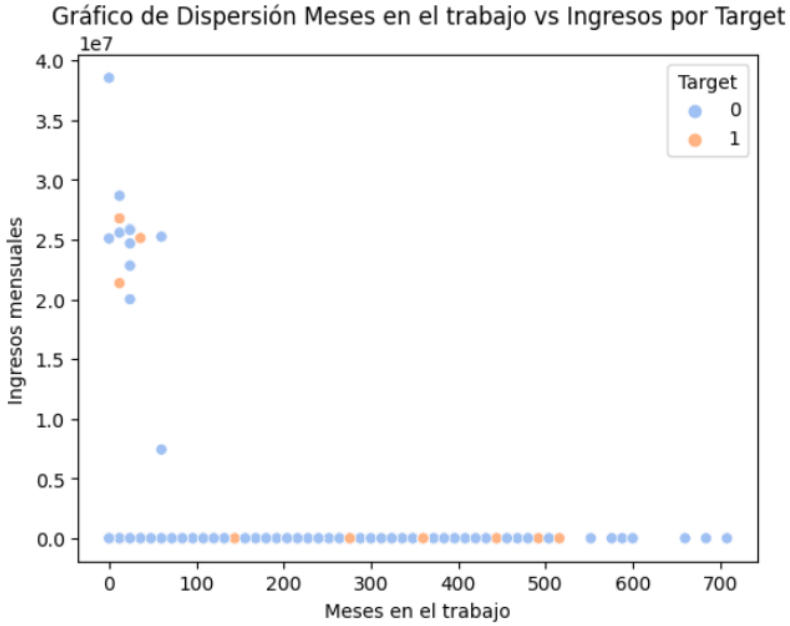

Figura 16: Gráfico de dispersión meses en el trabajo vs ingresos mensuales por target

también aumenta el ingreso, pero se observa que esto no es así, si bien hay algunos ingresos altos dentro en los primeros meses de trabajo, estos son pocos así que tampoco se podría decir que hay una relación negativa entre los meses de trabajo y los ingresos mensuales, en realidad parecería que no hay una relación.

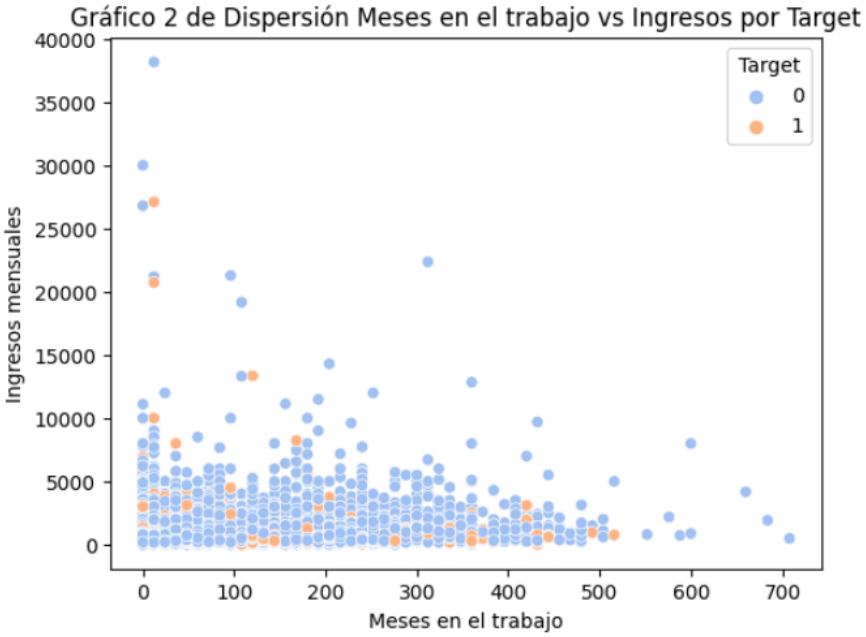

Figura 17: Gráfico 2 de dispersión meses en el trabajo vs ingresos mensuales por target

Por ultimo graficamos la edad vs los mese en el trabajo, para este caso si se ve una relación, los meses en el trabajo aumenta conforme aumente la edad, lo cual es bastante lógico.

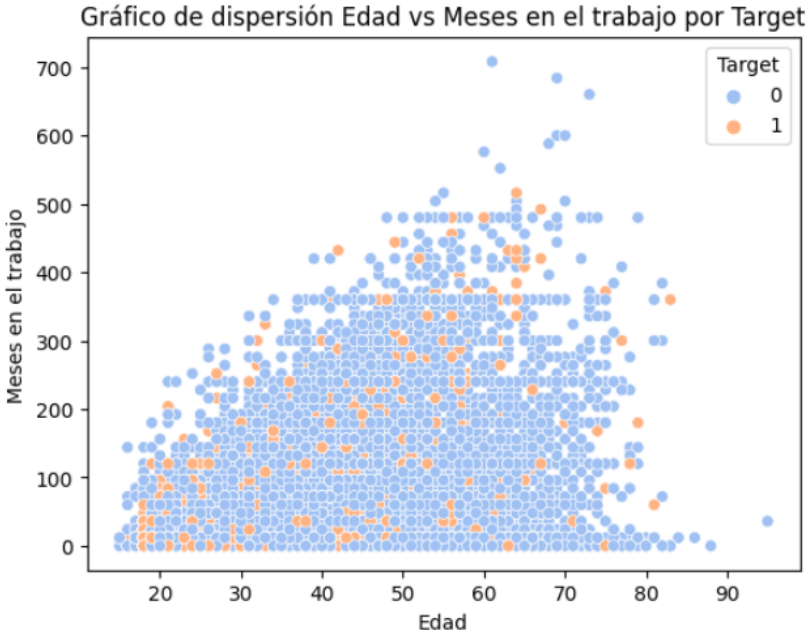

Figura 18: Gráfico de dispersión de edad vs meses en el trabajo por tarjet

Logramos identificar como se relacionan algunas variables con la target, muchos de los descubrimientos son bastante lógicos, algunos no, y también hay variables que uno pensaría que se relacionan pero no se muestra una relación aparente.

4. Datos anómalos

Empecemos revisando los histogramas de nuestras variables continuas, para la edad se observa que hay edades arriba de 80 que parecen atípicas, si bien es coherente una edad es 80, no es algo muy común y pudieran sesgar los datos.

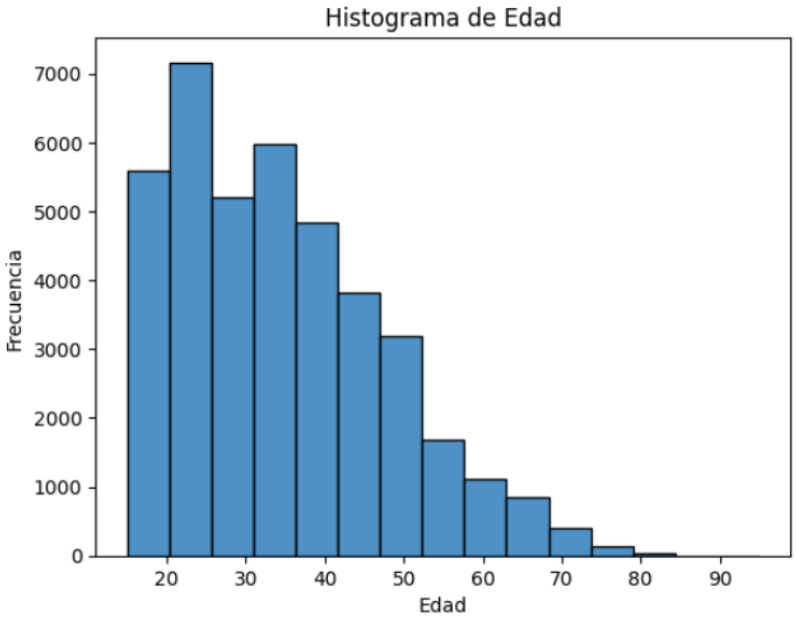

Figura 19: Histograma de edad

Para el caso del día de pago los datos se ven normales, el rango de valores es bastante normal y el histograma muestra información coherente, por lo que no hay razón para hacer un análisis de datos atípicos en esta variable.

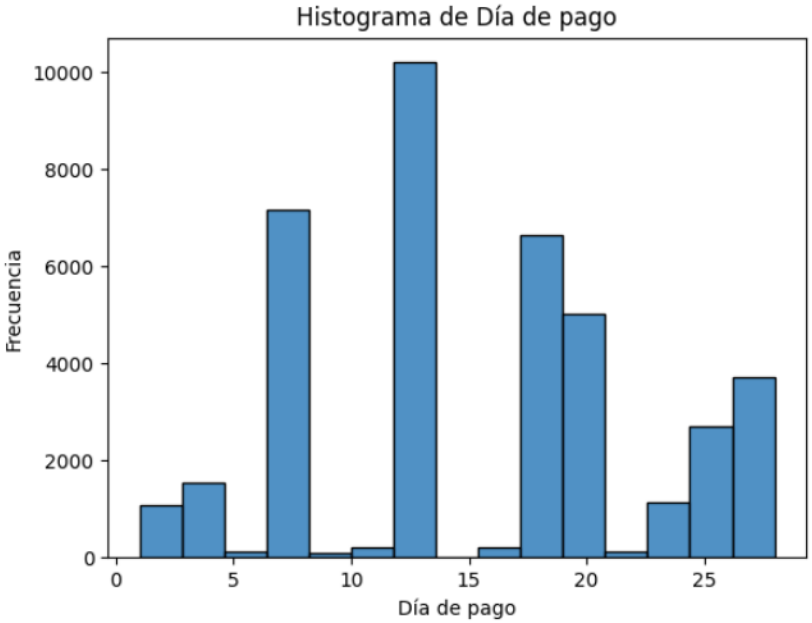

Figura 20: Histograma de día de pago

Para los meses de residencia si hay datos que parecieran ser atípicos, hay muy poco datos arriba de 600, al igual que con la edad, esto puede ser coherente pero dejarlos pudiera sesgar los datos.

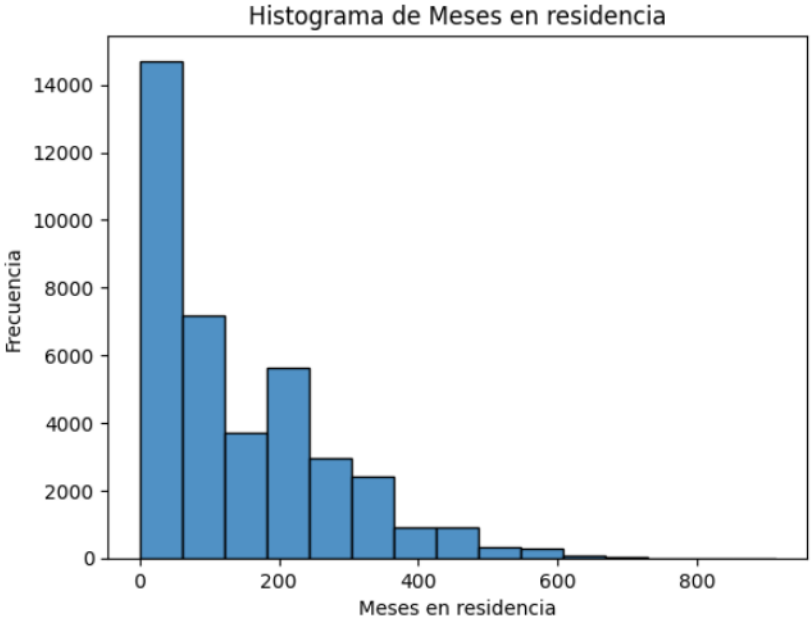

Figura 21: Histograma de meses en residencia

Para los meses en el trabajo observamos lo mismo, hay pocos registros arriba de 300 y tendremos que hacer un análisis para determinar si son atípicos.

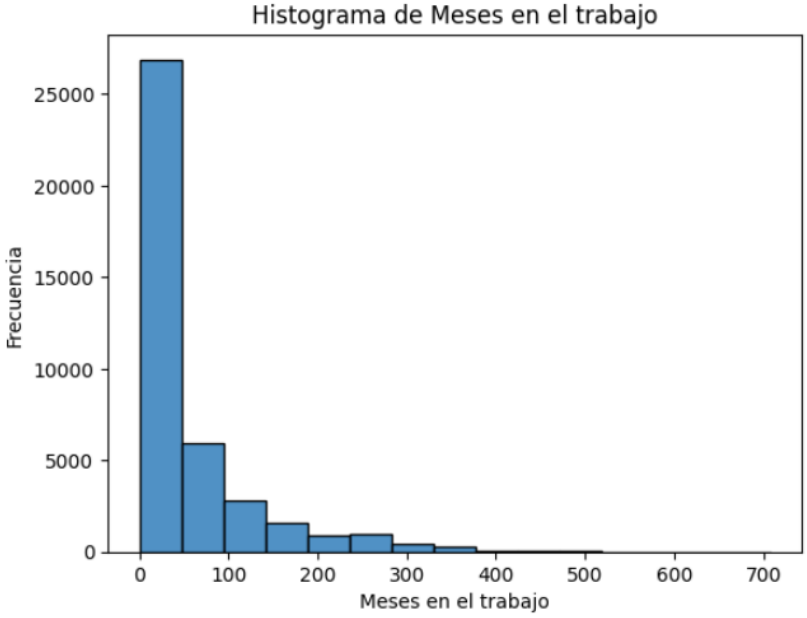

Figura 22: Histograma de meses en el trabajo

Para el ingreso mensual se ve claramente que hay datos atípicos, en gráficas anteriores vimos que los ingresos rondan alrededor de 500, si embargo hay registros que rebasan los 5000000, esto pudiera ser real pero dejarlos sesgaría demasiado nuestros datos.

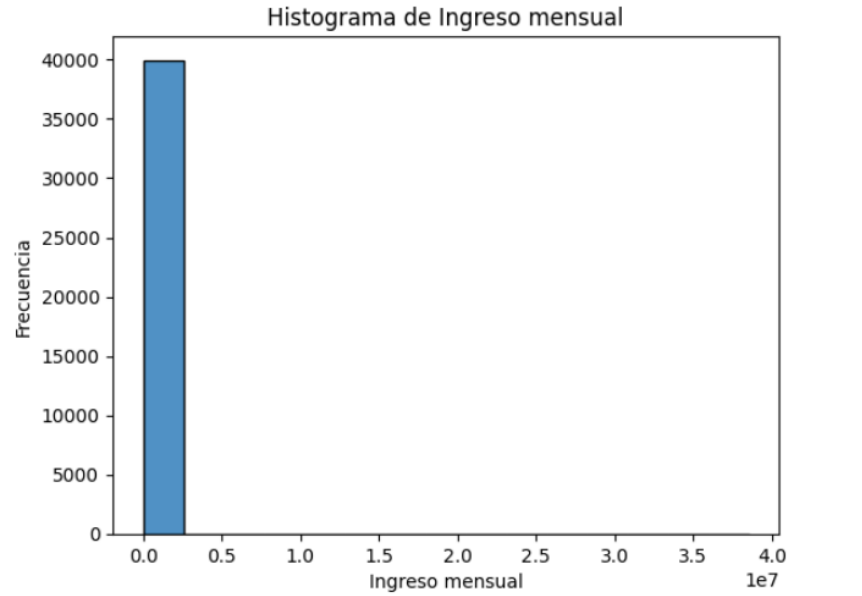

Figura 23: Histograma de ingreso mensual

Para la cantidad de tarjetas adicionales solicitadas en el mismo formulario de solicitud no vemos nada anormal, la variable solo toma 3 valores por lo que no vale la pena aplicar métodos para determinar si hay datos atípicos

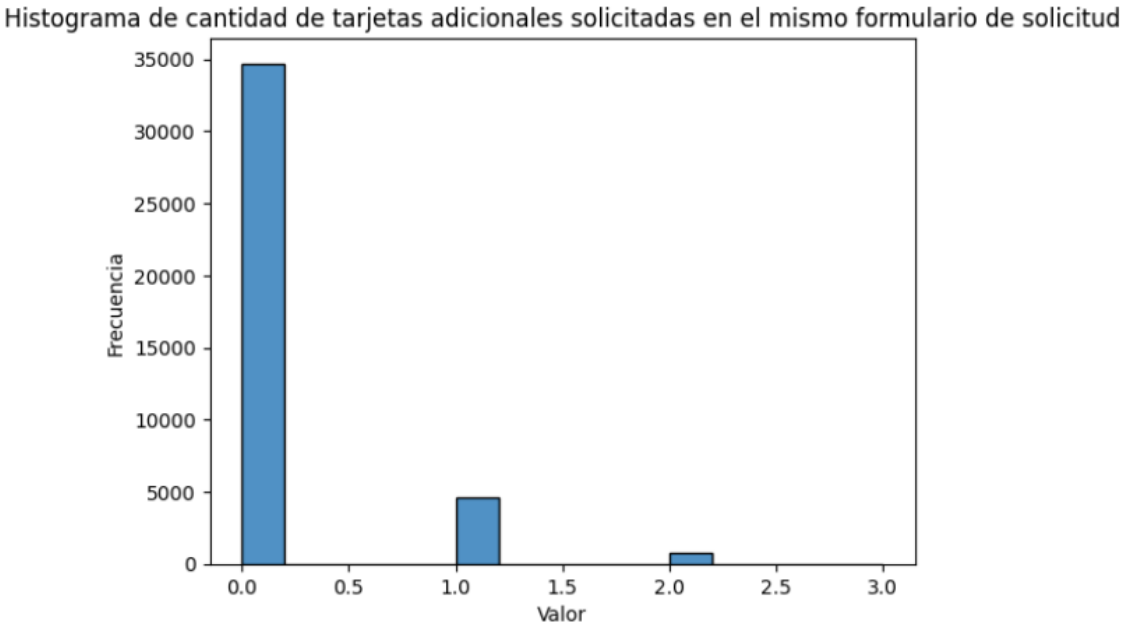

Figura 24: Histograma de tarjetas adicionales solicitadas en el mismo formulario de solicitud

Para determinar cuales son datos atípicos se aplicaron 3 métodos:

- IQR
- Método de percentiles
- Mean change

Cada método dio resultados diferentes sobre cuales registros son atípicos, por lo que se considero como atípicos aquellos registros que más de un método clasificara como atípico, en la siguiente tabla resumimos la información y observamos cuantos atípicos detecto cada método para cada variable, así como el porcentaje que esto representa del total de datos, se muestra también los registros que si se consideran como atípicos y el porcentaje de atípicos finales para cada variable.

Cuadro 22: Número de valores atípicos en características

Variables	**n_out_IQR**	**n_out_Percentil**	**n_out_Mean_Change**	**total_out**
c_personal_net_income	3099	3298	1	1999
c_months_in_the_job	4326	1836	1	1836
c_months_in_residence	418	2682	1	418
c_age	188	3409	4	191

Cuadro 23: Porcentaje de valores atípicos en características

Variables	**n_out_IQR %**	**n_out_Percentil %**	**n_out_Mean_Change %**	**%out**
c_personal_net_income	7.75 %	8.25 %	0.00 %	5.00 %
c_months_in_the_job	10.82 %	4.59 %	0.00 %	4.59 %
c_months_in_residence	1.05 %	6.71 %	0.00 %	1.05 %
c_age	0.47 %	8.53 %	0.01 %	0.48 %

En resumen, para la variable c_personal_net_income se detectaron 1999 atípicos que corresponde al 5 % de los registros, para la variable c_months_in_the_job se detectaron 1836 atípicos que corresponde al 4.59 % de los registros, para la variable c_months_in_residence se detectaron 418 atípicos que corresponde al 1.05 % de los registros y para la variable c_age se detectaron 191 atípicos que corresponde al 0.48 % de los registros, al final se eliminaran 4004 (no es igual a la suma de cada variable ya que hay registros que más de una variable detecto como atípico), esto corresponde al 10.01 % de los datos.

Una vez eliminados los registros anómalos observamos a continuación los histogramas actualizados, para la edad ya no se ven esos registros arriba de 80 que anteriormente veíamos.

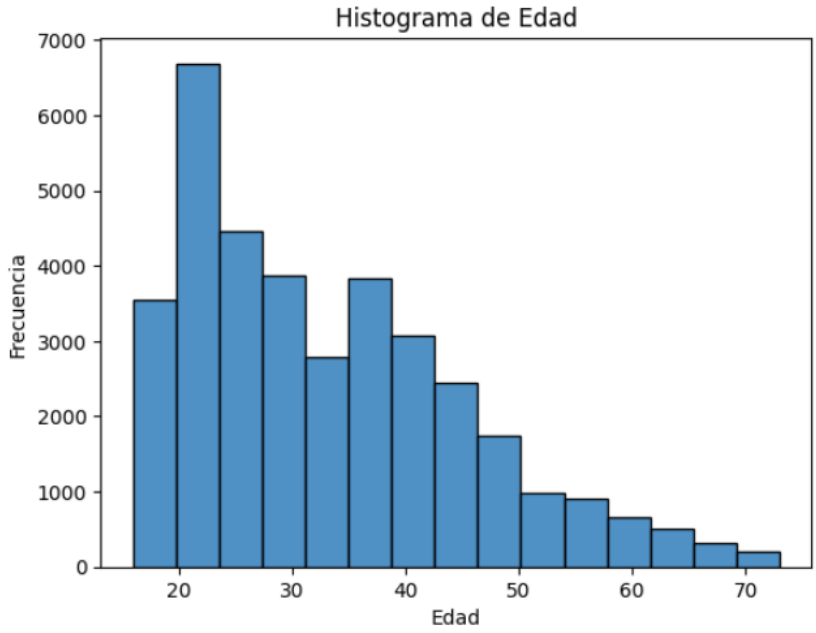

Figura 25: Histograma de edad

Para la variable de meses en residencia observamos una mejor gráfica sin registros arriba de 600 como lo veíamos.

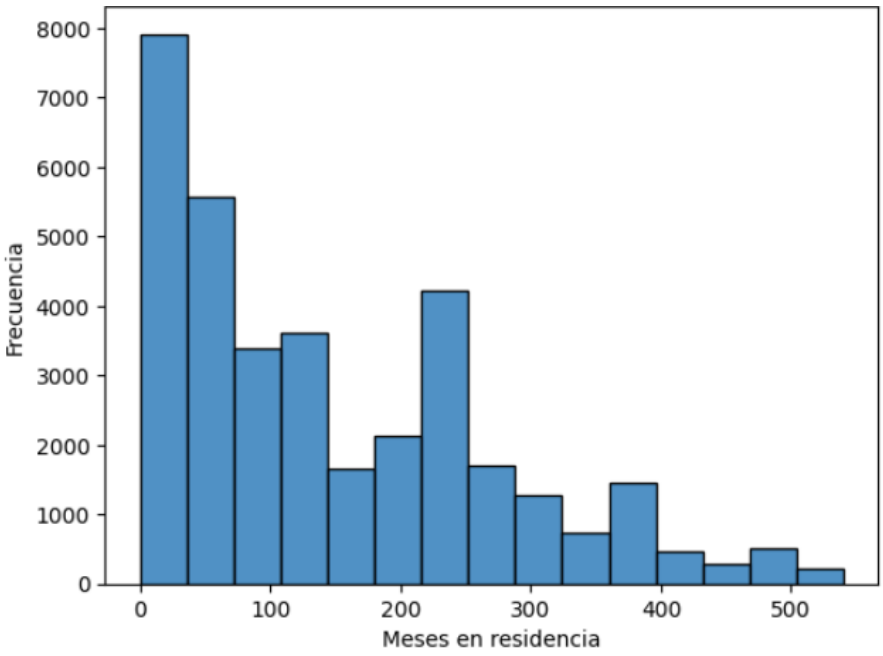

Figura 26: Histograma de meses en residencia

El mismo caso aplica para los meses en el trabajo, ya no tenemos esos registros mayores a 400.

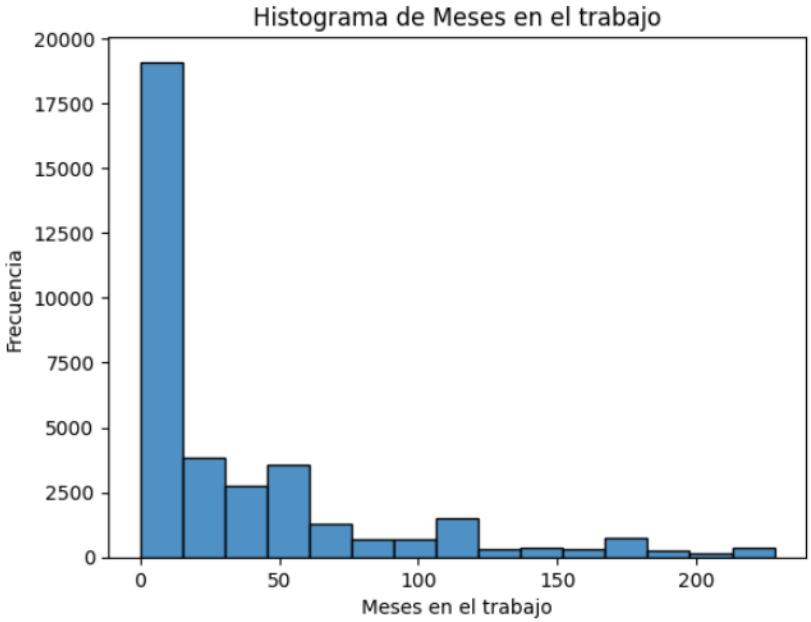

Figura 27: Histograma de meses en el trabajo

Para el ingreso mensual se puede incluso ver de mejor forma el histograma de datos, anteriormente teníamos datos tan grandes que no podíamos ver bien la distribución de los datos.

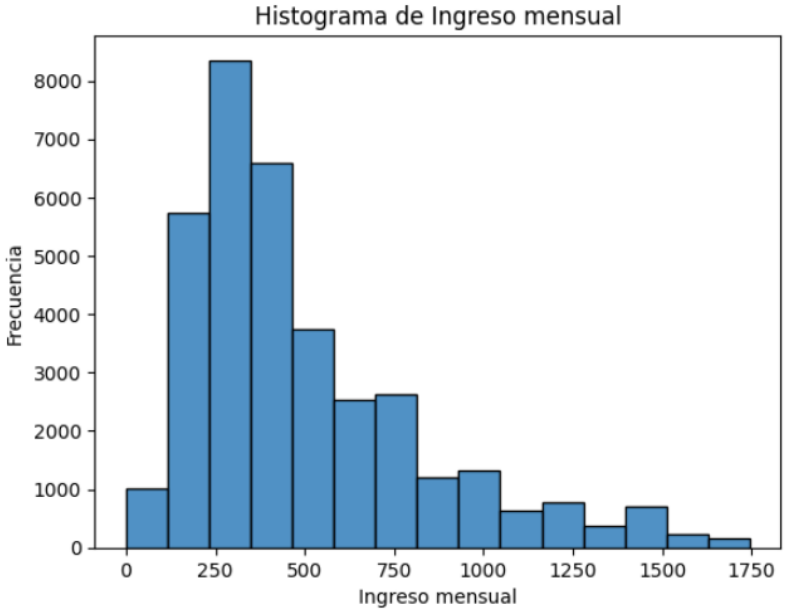

Figura 28: Histograma de ingreso mensual

5. Datos Nulos

A partir de este momento se hizo un dicción entre conjunto de entrenamiento y conjunto de prueba de los datos, los conjuntos se generaron mediante una división estratificada a partir de la target (es decir las proporciones en la target se mantuvieron en cada conjunto) y el conjunto de entrenamiento se quedo con el 70 % mientras que el conjunto de prueba se quedo con el 30 %, hasta este punto teníamos 35983 registros y el conjunto de entrenamiento se quedo con 25188 registros y el conjunto de prueba con 10795 registros. A continuación se muestra el total de registros con datos nulos por variable en el conjunto de entrenamiento y el conjunto de prueba.

Cuadro 24: Resumen de datos faltantes

Columna	**nulos_train**	**nulos_test**	**pct_nulos_train**	**pct_nulos_test**
id_client	0	0	0.000000	0.000000
id_shop	0	0	0.000000	0.000000
v_sex	1	0	0.000040	0.000000
v_marital_status	0	0	0.000000	0.000000
c_age	0	0	0.000000	0.000000
v_flag_residencial_phone	0	0	0.000000	0.000000
v_area_code_residencial _phone	0	0	0.000000	0.000000
c_payment_day	0	0	0.000000	0.000000
v_residence_type	0	0	0.000000	0.000000
c_months_in_residence	573	241	0.022749	0.022325
v_flag_residence_town_eq _working_town	0	0	0.000000	0.000000
c_months_in_the_job	20	2	0.000794	0.000185
v_profession_code	0	0	0.000000	0.000000
c_personal_net_income	3	5	0.000119	0.000463
c_quant_additional_cards _in_the_application	0	0	0.000000	0.000000
v_target_label_bad=1	0	0	0.000000	0.000000

Para la variable v_sex se imputo con la moda, es decir con mujer, mientras que para las variables numéricas se realizaron pruebas imputando la media, la mediana y la moda y se realizo el Test no parametrico de Kolmogórov-Smirnov para determinar cual era el mejor método, se comparo la distribución de las variables antes y después de imputar, el objetivo era encontrar un método que después de imputar mantuviera la distribución original, a continuación se muestra el resumen de los datos

Cuadro 25: Resultados de las pruebas Kolmogorov-Smirnov para imputación de variables

Variable	**Estadística**	**Valor de prueba**	**Valor p**
c_months_in_residence	Media	0.0133	0.0240
	Mediana	0.0113	0.0805
	Moda	0.0194	0.0002
c_months_in_the_job	Media	0.0006	1.0000
	Mediana	0.0004	1.0000
	Moda	0.0004	1.0000
c_personal_net_income	Media	0.0001	1.0000
	Mediana	0.0001	1.0000
	Moda	0.0001	1.0000

Dado los valores anteriores, para c_personal_net_income se decidió imputar con la media(también pudo ser mediana o moda, el p value es 1), para c_months_in_the_job se imputo con la mediana (también pudo ser media o moda, el p value es 1) y para c_months_in_residence se decidió imputar con la mediana ya que la mediana proporciona el p value más alto, la siguiente tabla muestra con que valor se imputo cada variable.

Cuadro 26: Valores de Variables

Variable	**Valor**
v_sex	F
c_personal_net_income	496.72192615
c_months_in_residence	120
c_months_in_the_job	12

6. Ingeniería de Variables

Las siguientes variables son categóricas:

- v_sex
- v_marital_status
- v_flag_residencial_phone
- v_area_code_residencial_phone
- v_residence_type
- v_flag_residence_town_eq_working_town
- v_profession_code

Como se tienen que convertir a numéricas, se utilizo one-hot encoding para convertirlas a numéricas, las nuevas variables son las siguientes, y las anteriores se eliminaron

- v_sex_M
- v_marital_status_O
- v_marital_status_S
- v_flag_residencial_phone_Y
- v_residence_type_C
- v_residence_type_O
- v_residence_type_P
- v_flag_residence_town_eq_working_town_Y
- v_profession_code_201-400
- v_profession_code_401-600
- v_profession_code_601-800
- v_profession_code_801-999

Adicional a esto, se crearon 4 variables nuevas

- c_dif_age_job: la diferencia entre años de trabajo y edad (c_age - $\frac{c\_months\_in\_the\_job}{12}$)
- c_dif_age_res: la diferencia entre años de residencia y edad (c_age - $\frac{c\_months\_in\_the\_residence}{12}$))
- c_income_tot_job: ingreso total en el ultimo trabajo (c_months_in_the_job*c_personal_net_income)
- c_rat_income_age: razón entre el ingreso y la edad ($\frac{c\_personal\_net\_income}{c\_age}$)

Despues se escalaron todos los datos a fin de tener todas las variables en una escala igual, a continuación se muestran el total de registros y columnas de cada conjunto.

Cuadro 27: Datos finales

Conjunto	**Registros**	**Columnas**
Conjunto de Entrenamiento	25,188	25
Conjunto de Prueba	10,795	25

Finalmente en la siguiente tabla se muestra las nuevas columnas con un ejemplo de los registros y el tipo de dato de cada columna.

Cuadro 28: Variables finales

Variable	**Valor**	**Tipo de Dato**
id_client	0.659040	float64
id_shop	0.568421	float64
c_age	0.403509	float64
c_payment_day	0.629630	float64
c_months_in_residence	0.200000	float64
c_months_in_the_job	0.105263	float64
c_personal_net_income	0.395642	float64
c_quant_additional_cards_in_the_application	0.0	float64
v_target_label_bad=1	0.0	float64
v_sex_M	1.0	float64
v_marital_status_O	0.0	float64
v_marital_status_S	0.0	float64
v_flag_residencial_phone_Y	1.0	float64
v_residence_type_C	0.0	float64
v_residence_type_O	0.0	float64
v_residence_type_P	0.0	float64
v_flag_residence_town_eq_working_town_Y	1.0	float64
v_profession_code_201-400	0.0	float64
v_profession_code_401-600	0.0	float64
v_profession_code_601-800	0.0	float64
v_profession_code_801-999	1.0	float64
c_dif_age_job	0.478261	float64
c_dif_age_res	0.319444	float64
c_income_tot_job	0.042775	float64
c_rat_income_age	0.199973	float64