
# UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

**Facultad de Estudios Superiores Acatlán**

Diplomado de Ciencias de Datos

## Examen II

Diplomado de Ciencia de Datos, Módulo I
Profesora: Carla Paola Malerva Reséndiz
Alumno: Ricardo Paramont Hernández González
Fecha: Jueves 4 de febrero de 2021

## 1. Conjunto de Datos

El examen se hace con el dataset "data_examen_2_.csv" y un complementario de "ocupaciones_examen2.csv".

El primer data set se compone de las variables:

**VARIABLES**

ID_CLIENT
ID_SHOP
SEX
MARITAL_STATUS
AGE
QUANT_DEPENDANTS
EDUCATION
FLAG_RESIDENCIAL_PHONE
AREA_CODE_RESIDENCIAL_PHONE
PAYMENT_DAY
SHOP_RANK
RESIDENCE_TYPE
MONTHS_IN_RESIDENCE
FLAG_MOTHERS_NAME
FLAG_FATHERS_NAME
FLAG_RESIDENCE_TOWN=WORKING_TOWN
FLAG_RESIDENCE_STATE=WORKING_STATE
MONTHS_IN_THE_JOB
PROFESSION_CODE
MATE_INCOME
FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS
FLAG_OTHER_CARD
QUANT_BANKING_ACCOUNTS
PERSONAL_REFERENCE_#1
PERSONAL_REFERENCE_#2
FLAG_MOBILE_PHONE
FLAG_CONTACT_PHONE
PERSONAL_NET_INCOME
COD_APPLICATION_BOOTH
QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION
FLAG_CARD_INSURANCE_OPTION
tgt

# 2. Calidad de Datos

## 2.1 Etiquetado de variables

- Primero se definió un generador que devuelve el próximo nombre de columna del dataframe para poder obtener facilmente datos de la columna que me ayudarían a clasificarla como **continua, discreta, fecha** o **texto.**
- Columnas de del dataframe:
  - 'v_id_shop'
  - 'v_sex'
  - 'v_marital_status',
  - 'c_age'
  - 'v_quant_dependants'
  - 'v_education'
  - 'v_flag_residencial_phone'
  - 'v_area_code_residencial_phone'
  - 'v_payment_day'
  - 'v_shop_rank'
  - 'v_residence_type'
  - 'c_months_in_residence'
  - 'v_flag_mothers_name'
  - 'v_flag_fathers_name'
  - 'v_flag_residence_town=working_town'
  - 'v_flag_residence_state=working_state'
  - 'c_months_in_the_job'
  - 'v_profession_code'
  - 'c_mate_income'
  - 'v_flag_residencial_address=postal_address',
  - 'v_flag_other_card'
  - 'v_quant_banking_accounts'
  - 't_personal_reference_#1'
  - 't_personal_reference_#2'
  - 'v_flag_mobile_phone'
  - 'v_flag_contact_phone'
  - 'c_personal_net_income'
  - 'v_cod_application_booth'
  - 'v_quant_additional_cards_in_the_application'
  - 'v_flag_card_insurance_option'

  *Ni el id del registro ni el tgt necesitan etiqueta.

## 2.2 Duplicados

- Se revisó el número de renglones duplicados. Se encontraron 5 y se eliminaron

## 2.3 Completitud

- Se encuentra que sólo dos columnas contienen valores faltantes:

| | columna | total | completitud |
|---|---|---|---|
| 0 | v_education | 50995 | 0.000000 |
| 1 | t_personal_reference_#1 | 20625 | 59.554858 |
| 2 | t_personal_reference_#2 | 13886 | 72.769879 |
| 3 | c_age | 2472 | 95.152466 |
| 4 | v_flag_contact_phone | 1279 | 97.491911 |
| 5 | c_mate_income | 306 | 99.399941 |
| 6 | v_sex | 3 | 99.994117 |
| 7 | id_client | 0 | 100.000000 |

- Se procede a eliminar la columna de v_education, t_personal_reference#1, t_personal_reference#2

## 2.4 Consistencia

Se revisan las distintas variables, y se nota:

- variables categoricas: varias de ellas tenían errores de registro siendo que las strings contenían espacios o tenían diferencias entre minúsculas y mayúsculas. Se procede a corregir dichos errores.
- Variables continuas de cantidad de tiempo: algunas tenían valores demasiado grandes, por lo que se consideró como fuera de formato aquellas que sobrepasaban cierto límite.
- Variables unarias: se encontraron columnas categóricas que contenían realmente un único valor en todos sus registros. Dichas columnas de borran en reducción de dimensiones y son:
  - 'v_quant_dependants'
  - 'c_mate_income'
  - 'v_flag_other_card'
  - 'v_quant_banking_accounts'
  - 'v_flag_mobile_phone'
  - 'v_flag_contact_phone'
  - 'v_cod_application_booth'

- 'v_flag_card_insurance_option'

- Variables continuas: se encontráron 5 variables continuas:
  - c_age
  - c_months_in_residence
  - c_months_in_the_job
  - c_mate_income
  - c_prsonal_net_income

  Como notación de interés, se observa que más de tres cuartas partes de los registros de ingresos de la pareja tienen valor de cero.

### 2.5 Completitud trans revisar consistencia

Algunas variables incrementan su cantidad de valores ausentes, pero no sobrepasan el 20% de los mimos.

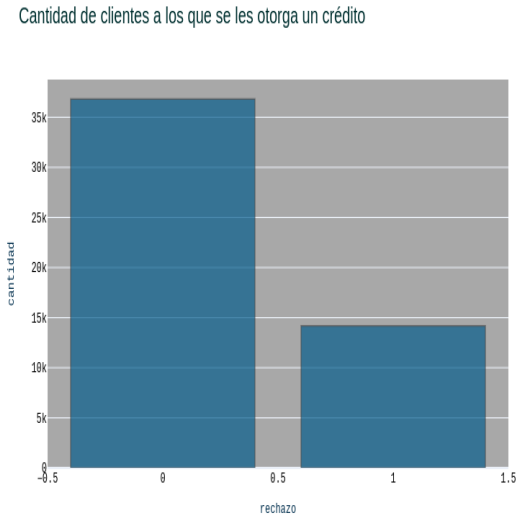
### 2.6 Limpieza, normalización y transformación a tipo numérico

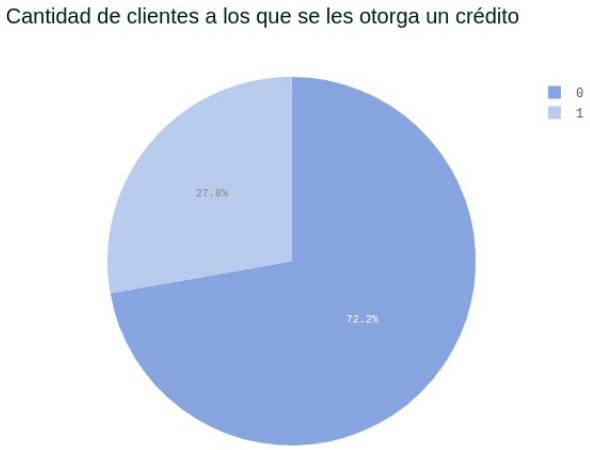
- Las variables fueron previamente normalizadas en la sección de consistencia.
- Se transforma el tipo de dato de cada variable a la más conveniente con el uso de df.convert_dtypes().dtypes

# 3. Análisis Exploratrio de Datos (EDA)

## 3.1 Cantidad de clientes a los que se les otorga un crédito

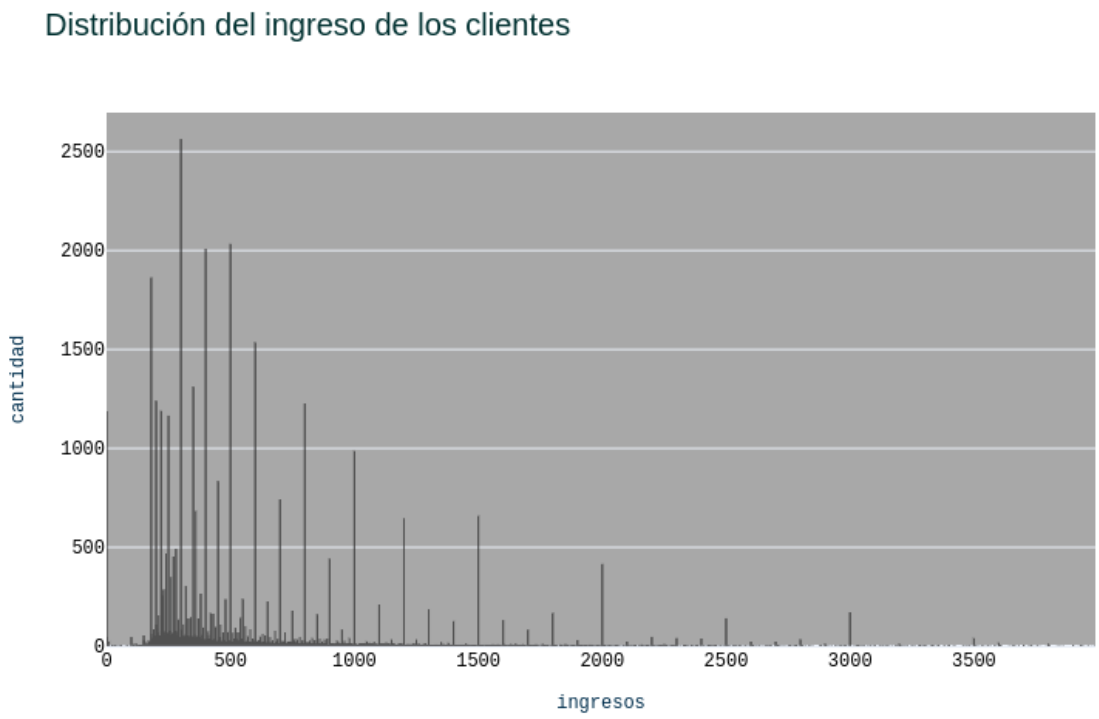
**0**:crédito aprobado
**1**:credito rechazado



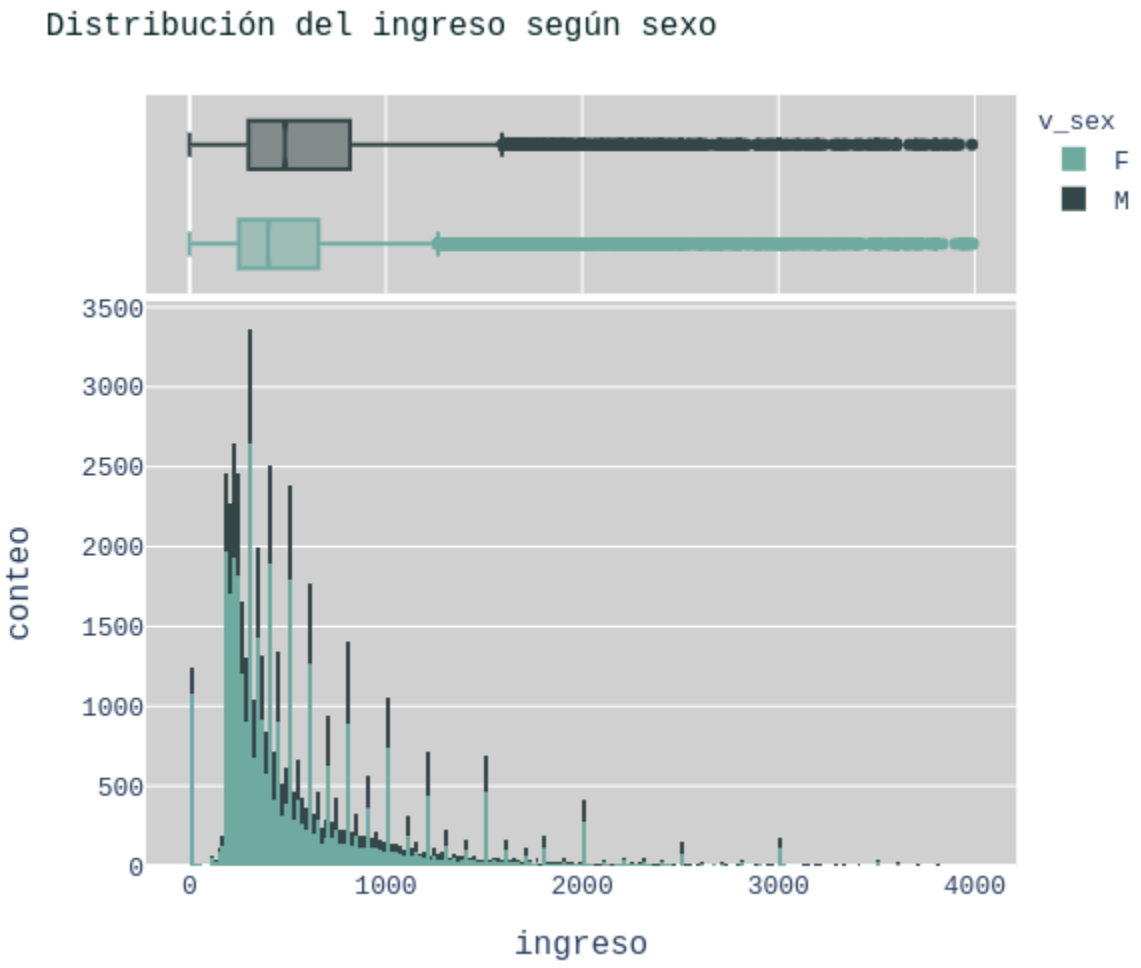

Se observa que a poco más de los clientes no se les aprueba un crédito.

### 3.2 Distribución de los ingresos de los clientes



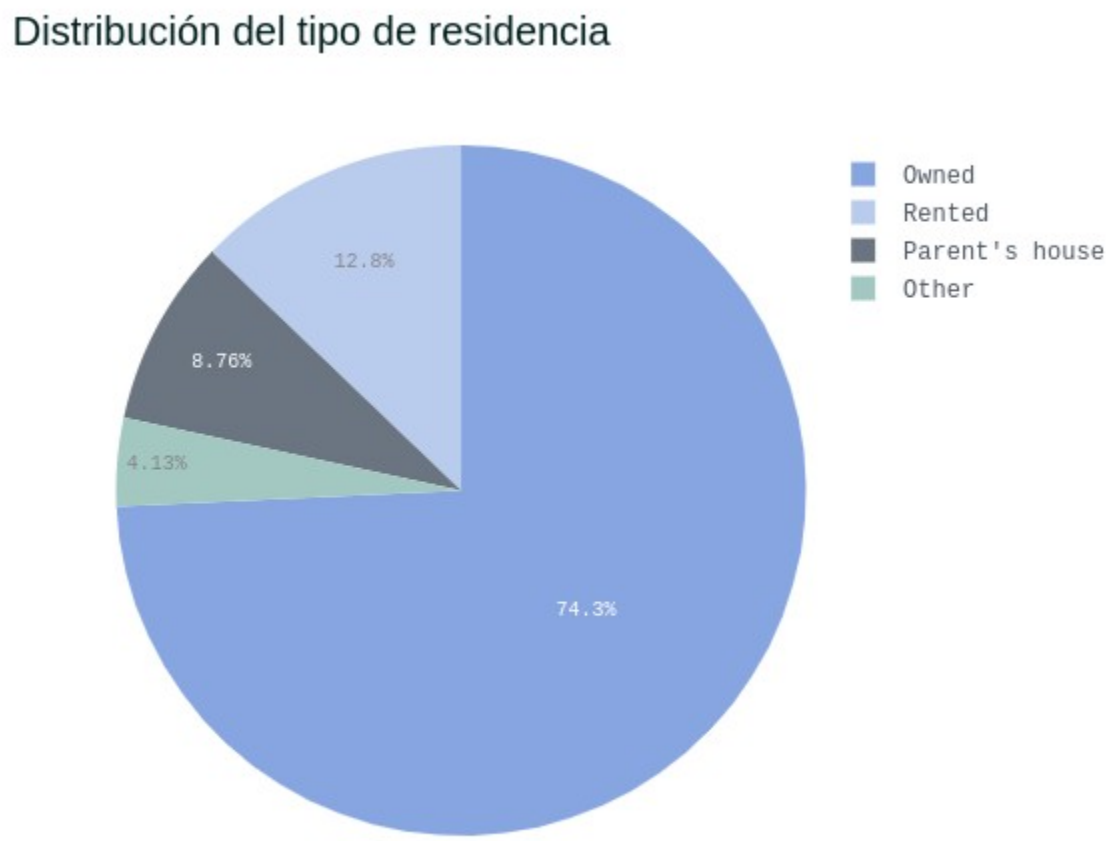

Se puede observar que la gran mayoría de los clientes tiene un ingreso menor a los mil reales mensuales.

### 3.3 Distribución de los ingresos de los clientes según sexo



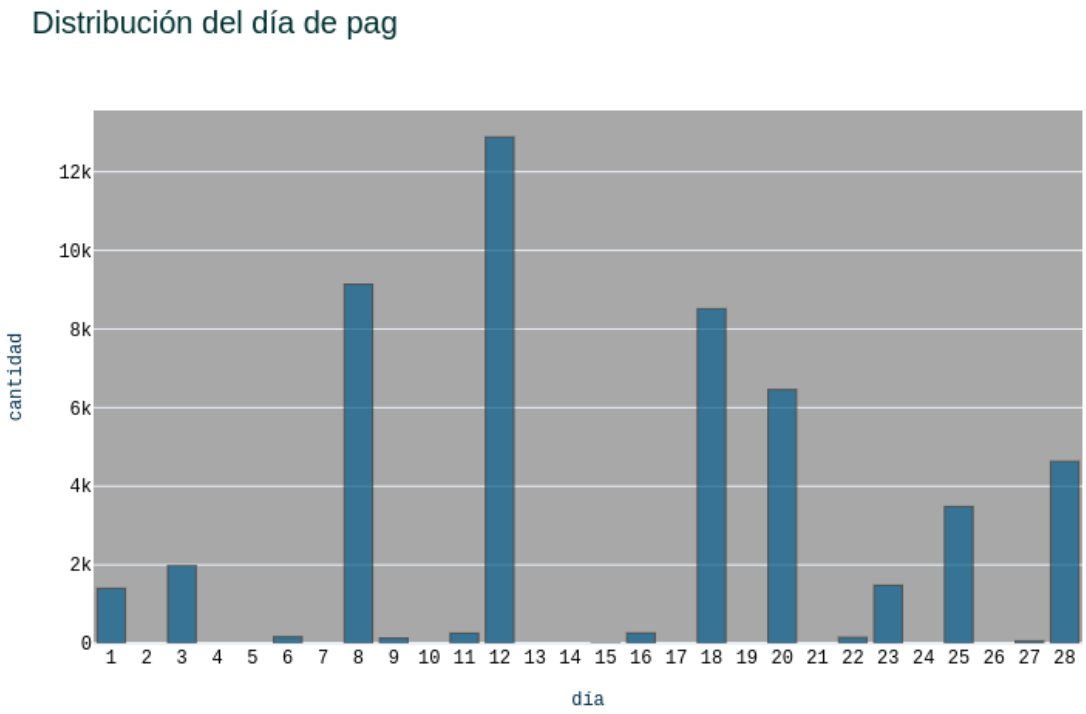

Las mujeres dentro de nuestro dataset tienden a ganar menos que los hombres, sin embargo, representan un mayor volumen de nuestros clientes.

### 3.4 Distribución del tipo de residencia



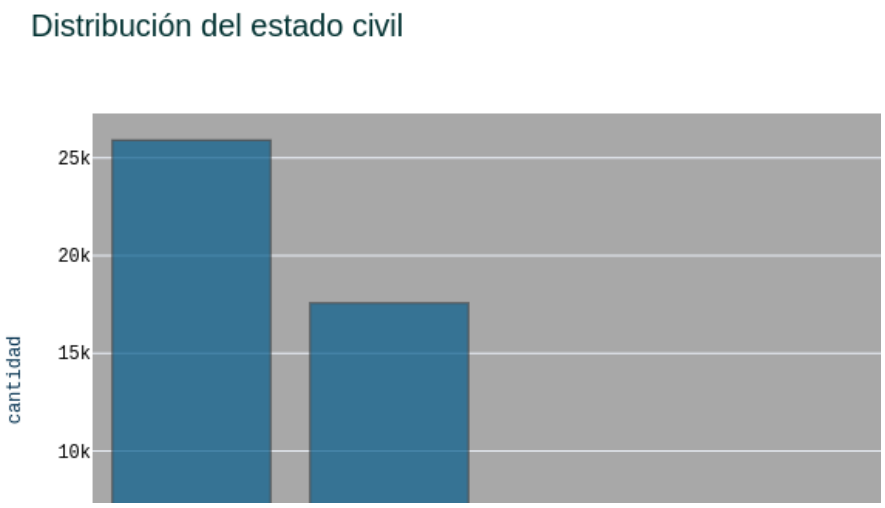

Casi tres de cada uno de los clientes es dueño de su vivienda, lo cual puede ser un poco sorpresivo teniendo en cuenta los bajos ingresos de los mismos.

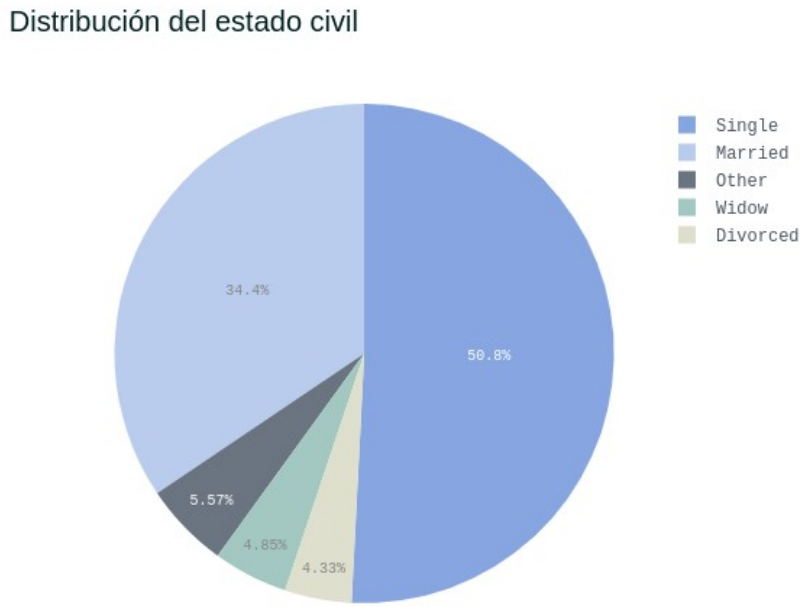
### 3.5 Distribución del día de pago del mes




El día del mes que más pagos registra es el 12. Curiosamente, hay días que no poseen ni un solo registro, como el caso del 2, 7 o 19.

**2.6 En la fiscalía de "juzgados familiares" , ¿Cuáles son los delitos más frecuentes?**




La mitad de nuestros clientes son solteros, mientras que sólo un tercio está casado. Casi la misma cantidad de personas en relaciones de otros tipos, como unión libre, corresponde a la cantidad de viudos.

# 4. Datos anómalos

## 4.1 Análisis de valores anómalos

Para encontrar outliers, se consideraron los registros que fueron identificados tanto por la consideración del rango intercuantílico y los percentiles 0.05 y 0.95. Si un registro fue identificado por un sólo método, dicho registro se conserva.

Variables con datos anómalos:

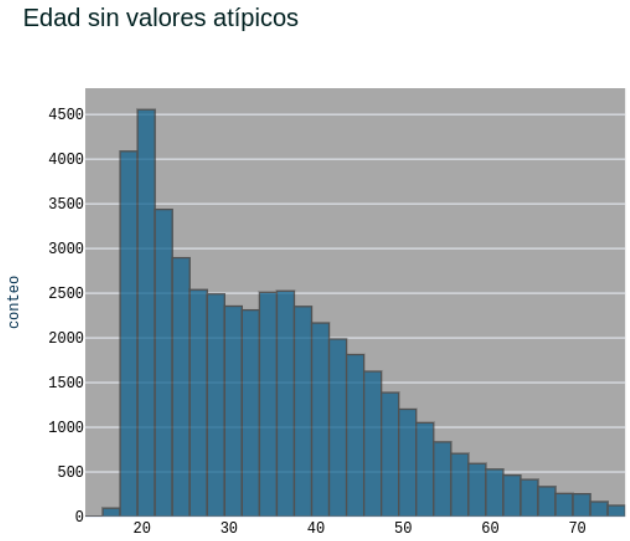
- c_age: 421 outliers $(0.83\%)$
- c_months_in_residence: 599 outliers $(1.17\%)$
- c_months_in_the_job: 2352 outliers $(4.61\%)$
- c_mate_income: 2007 outliers $(3.94\%)$
- c_personal_net_income: 2540 outliers $(4.98\%)$

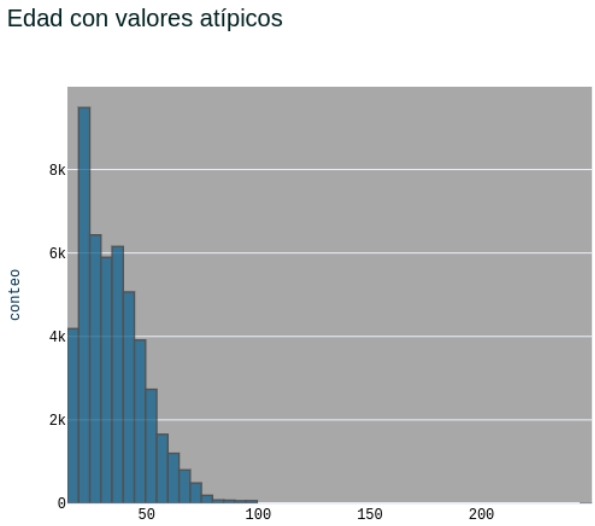
Observaciones:

- Como una observación interesante, más de tres cuartas partes de los registros indican que los ingresos de la pareja son nulos.
- La variable que más outliers encontró tomando en cuenta el rango intercuantílico es el de c_months_in_the_job, lo que demuestra un gran varianza de los datos.
- Existe una cantidad notoria de clientes cuyos ingresos están en el orden de los millones reales mensuales.

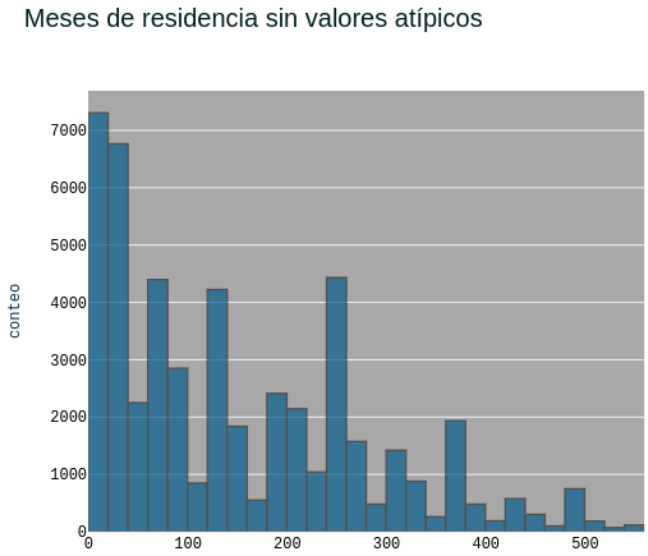
## 4.2 Visualización de datos anómalos con histogramas

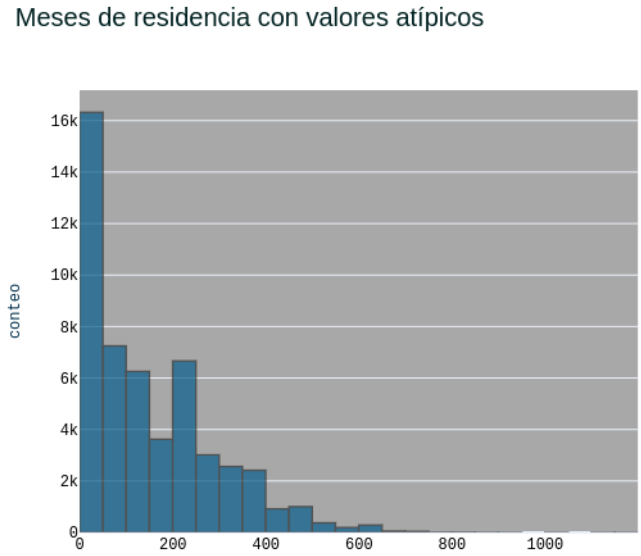
### 4.2.1 c_age



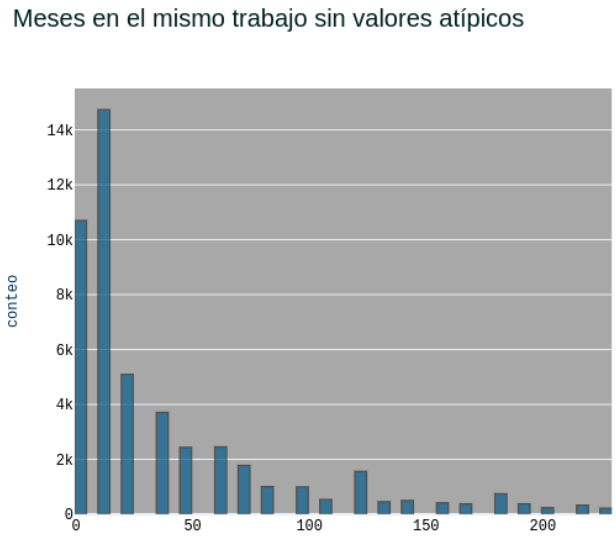



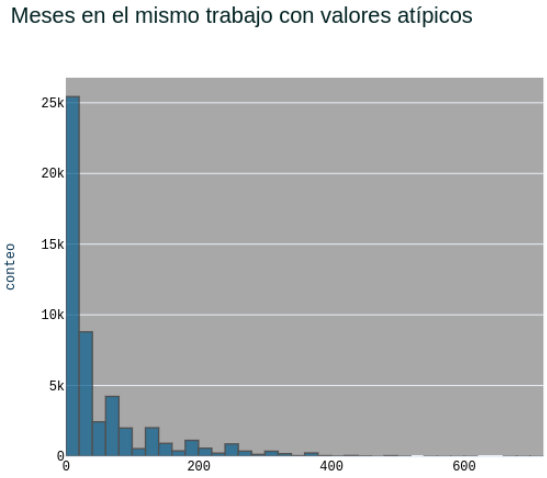
### 4.2.2 c_months_in_residence



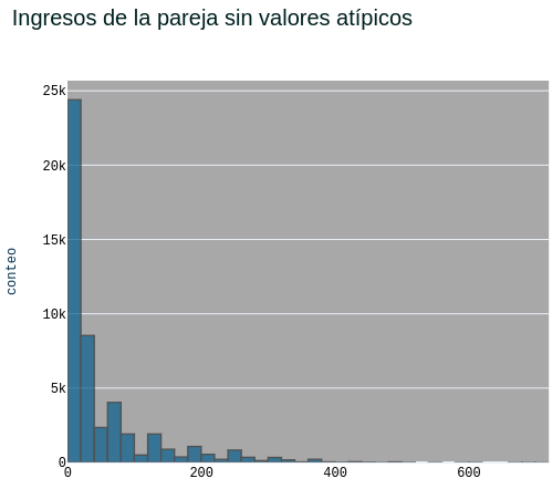



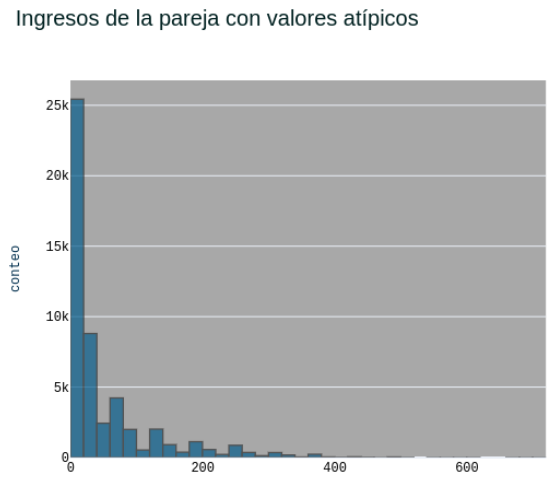
### 4.2.3 c_months_in_the_job



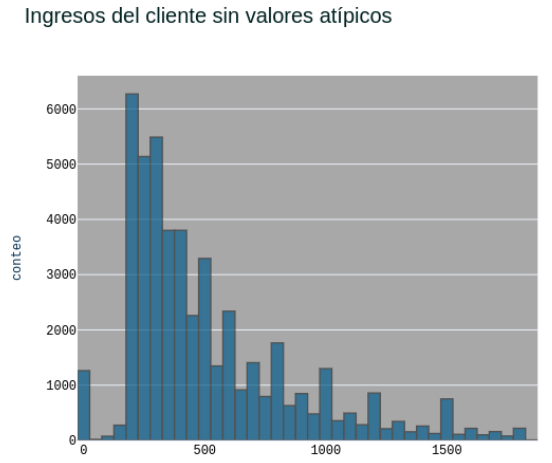



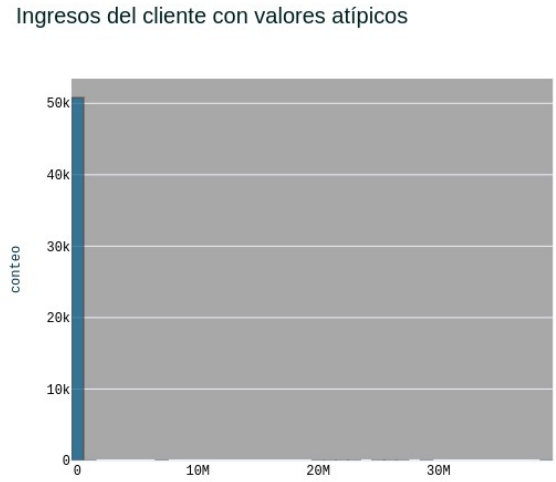
### 4.2.4 c_mate_income



### 4.2.5 c_personal_net_income



# 5. Datos Faltantes

## 5.1 Categoricos

Se imputa la moda en todos los casos.

5.2 Continuas

Se imputa el valor que represente mejores resultados de la prueba KS.

# 6. Ingeniería de variables/Normalización

## 6.1 Transformando variables flag a dummies

Se ftransforman dichas variables de forma que contengan valores numéricos. La transformación es:

- 'Y' = 1
- 'N'=0

## 6.2 One hot encoding de categoricas restantes con más de una categoría o que no son ordinales

Se crean variables dummy para las categorías de las variables:

- 'v_id_shop'
- 'v_sex'
- 'v_marital_status'
- 'v_residence_type'
- 'v_quant_additional_cards_in_the_application'

## 6.3 Count Vectorizer

Se vectoriza la única variabe de texto $\left(\text{t\_profession}\right)$, la cual fue agregada al hacer un join con la tabla de ocupaciones.

El método utilizado es el de count vectorizing y se incluyen sólo las palabras con un 5% o más de ocurrencia.

# 7. Reducción de dimensiones

## 7.1 Variables unarias

Se eliminan variables con un sólo valor.

## 7.2 Filtro de correlación

Se eliminan 10 variables por tener una alta correlación con otras.

## 7.3 Correlación con el objetivo

Se eliminan 20 variables por tener muy baja correlación con el objetivo.

## 7.4 Multicolinealidad

Se eliminan 3 variables por tener un VIF mayor a 15.

# 8. Cuestionario

### 8.1 Defina multicolinealidad con sus palabras y ¿cómo podemos medirla?

La multicolinealidad se refiere a cuando más de dos variables independientes en nuestro modelo están altamente linealmente relacionadas. Se puede medir con el Variable Inflation Factor $(\text{VIF})$

### 8.2 ¿Cómo reduce dimensiones PCA?

El método de Pricipal Component Analysis $(\text{PCA})$ consiste en una técnica linear insupervisada que nos permite identificar patrones en el conjunto de datos entre las variables. PCA tiene como objetivo encontrar las direcciones de maxima varianza en datos de gran dimensionalidad para proyectarlos en un subespacio con igual o menor número de dimensiones.

### 8.3 Explique como funciona el filtro de alta correlación y filtro de correlación con la variable objetivo.

El filtro de alta correlación se encarga de encontrar variables independientes, o explicativas, con una alta correlación entre sí. Se supone entonces que incluir estas variables altamente correlacionadas significa incluir datos repetidos, pues a través de una variable, se pueden describir las demás que tienen alta correlación con ella. De esta forma, en un conjunto de variables altamente relacionadas, se deja únicamente una variable.

El fltro de correlación con la variable objetivo tiene como objetivo encontrar las variables con una correlación demasiado baja a la variable a predecir con el modelo que se aplicará con la tabla de datos. Ya que dichas variables tienen muy poca capacidad descriptiva de la variable objetivo, se eliminan.

### 8.4 ¿Por qué es necesario el escalado de variables al utilizar PCA?

Por que la dimensionalidad de las variables afecta grandemente al funcionamiento de la técnica. Es necesario tener variables dentro del mismo rango de datos $(\text{comunmente entre } 0 \text{ y } 1)$, para que la relevancia de las variables no se vea distorcionada.

### 8.5 ¿Cómo funciona el método de IQR para outliers?

Es primordialmente una regla heurística, donde se considera que los datos que se encuentran 1.5 veces el IQR sobre el tercer cuartil son datos anómalos altos y los que se encuentran 1.5 veces el IQR por debajo del primer cuartil son datos anómalos bajos.

### 8.6 Para imputar valores , ¿que opciones se tienen para continuas y cuáles para discretas?

Discretas:

- Moda (probando con chi cuadrada)
- Predicción (ajustando un modelo)
- Missing (dejando variables en categoría ausente)

Continua:

- Media
- Mediana
- Moda (estas últimas tres dependiendo de la prueba KS)
- Modelo
- Interpolación(si contamos con una serie de tiempo)

### 8.7 ¿Qué es WOE y IV?, ¿Cómo los podemos obtener?

WOE: Weight of Evidence, nos indica el poder predictivo de una variable independiente, se puede obtener como:

$$WOE = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right)$$

IV: Information Value, parámetro que rankea variables con base en su importancia, se puede obtener como:

$$IV = \sum (\%\ of\ non\text{-}events - \%\ of\ events) * WOE$$