**1. INFORMARSE DEL NEGOCIO => Saber la repercusión de cada variable**

**sobre el negocio y la variable objetivo => rangos de cada variable, valores**

**medios, etc. (en general)**

**Datos Nominales:** ‘Id’, ‘user\_id’, ‘deleted\_account\_id‘

dtype: int

Estos datos indican sobre que cuenta son los datos descritos, parece que tenemos una id personal (seguramente la que se muestra al cliente) siendo esta la ‘user\_id’, y una que será para la base de datos privada ‘id’

En ‘deleted\_account\_id ’ vemos una id existente en el monto de datos ‘id’ pero aparentemente no se podría dar una correlación de porque la encontramos en los datos de otra ‘id’, a revisar.

**Datos Cuantitativos:** ‘amount’

dtype: float

Esta variable nos indica la cantidad de dinero que se pide de préstamo

**Datos Qualificativos Nominales:** ‘status’ ‘transfer\_type’ ‘recovery\_status’

dtype: string

*‘status‘*

money\_back 16397

rejected 6568

direct\_debit\_rejected 831

active 59

transaction\_declined 48

direct\_debit\_sent 34

canceled 33

*‘transfer\_type‘*

Instant 13882

regular 10088

*‘recovery\_status‘*

completed 2468

pending 845

pending\_direct\_debit 16

cancelled 1

**Datos De Fechas**

Dentro de esta familia de datos encontramos lo siguiente

*Formato de fecha y hora con zona horaria*

created\_at

updated\_at

moderated\_at

reimbursement\_date

money\_back\_date

send\_at

reco\_creation

reco\_last\_update

*Formato de fecha simple*

cash\_request\_received\_date

**2. ANALASIS EXPLORATORIO**

a. Var objetivo: ***amount***

i. Tipo de dato=> clasificación (nm,n,i): ***dtype('float64')*, numerical**

ii. Estadística descriptiva: ***Numérica, cuantitativa de razón***

count 23970.000000

mean 82.720818

std 26.528065

min 1.000000

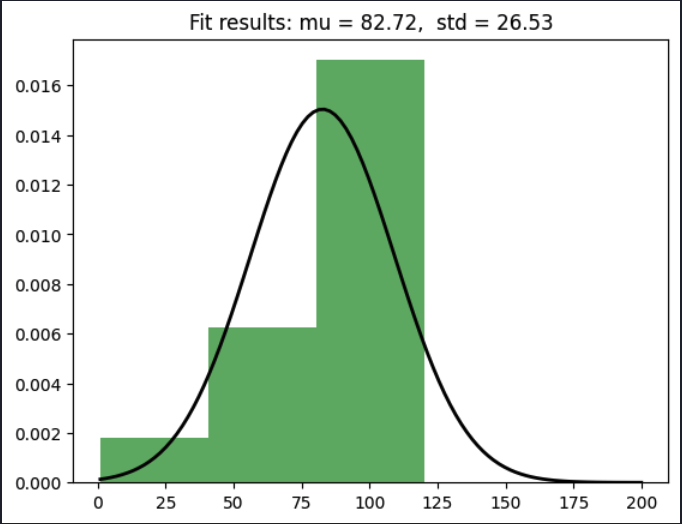
25% 50.000000

50% 100.000000

75% 100.000000

max 200.000000

Name: amount, dtype: float64

iii. Fun. distribución

b. Variables independientes

i. Tipo de dato=> clasificación (nm,n,i)

ii. Estadística descriptiva

iii. Fun. distribución

Variables cuantitativa o numerica:

\* created\_at: cuantitativa, continua de intervalo,

\* updated\_at: cuantitativa, continua de intervalo,

\* moderated\_at: cuantitativa, continua de intervalo,

\* reimbursement\_date: cuantitativa, continua de intervalo,

\* cash\_request\_received\_date: cuantitativa, continua de intervalo,

\* money\_back\_date: cuantitativa, continua de intervalo,

\* send\_at: cuantitativa, continua de intervalo,

Variables categoricas:

\* user\_id: cualitativa, dicotomica, nomina

\* id: cualitativa, dicotomica, nomina

\* deleted\_account\_id: cualitativa, dicotomica, nominal

\* status: cualitativa, dicotomica, nominal

\* transfer\_type: cualitativa, dicotomica, nomina

\* recovery\_status: cualitativa, dicotomica, nominal

\* reco\_creation: cualitativa, dicotomica, nominal

\* reco\_last\_update: cualitativa, dicotomica, nominal

iv. Interpretar la variable en base al a Fun.Dist de la objetivo

3. CALIDAD DEL DATO

a. Presencia de outliers => corregir o imputar a perdido

i. Corregir por:

• Media

• Moda

• Mediana

• Regresión

ii. Imputar a NaN

iii. Eliminar el registro

b. Presencia de misings (perdidos):

i. Corregir por:

• Media

• Moda

• Mediana

• Regresión

ii. Eliminar el registro

4. SELECCIÓN DE VARIABLES ALTAMENTE CORRELADAS

a. CREACIÓN DE VARIABLES => Coger variables correladas, e inventar

una nueva

5. BINARIZACION DE CATEGORICAS

6. DIVISION TRAIN TEST

7. TRANSFORMACION (ESCALADO)

a. MinMax

b. Standarizacion

c. Normalizacion

d. Yeo-Jhonson

8. MODELO

9. ENTRENAMIENTO

10. EVALUACIÓN DEL MODELO

11. PREDICCION CON NUEVOS DATOS