```
Fernando Anaya Delgado         - A01793832
Christian Emilio Saldaña Lopez - A506509
Equipo 95

Parte 1:

Fundamentos de bases de datos y para cienc
Se llama base de datos, o también banco de
perteneciente a un mismo contexto, ordenad
posterior recuperación, análisis y/o trans

Fundamentos de almacenes de datos (Data Wa
El Data Warehouse concentra y almacena de
información obtenida a partir de las múlti
organización, permitiendo así una rápida i
minería de datos, análisis y reportes (das

En ambos la limpieza de datos también cono
varios procesos destinados a mejorar la ca
utilizan para corregir o eliminar registro
conjunto de datos.
```

Fernando Anaya Delgado - A01793832 Christian Emilio Saldaña Lopez - A506509 Equipo 95

Parte 1:

Fundamentos de bases de datos y para ciencia de datos.
Se llama base de datos, o también banco de datos, a un conjunto de información perteneciente a un mismo contexto, ordenada de modo sistemático para su posterior recuperación, análisis y/o transmisión.

Fundamentos de almacenes de datos (Data Warehouse) para ciencia de datos. El Data Warehouse concentra y almacena de forma estructurada toda la información obtenida a partir de las múltiples fuentes de datos en nuestra organización, permitiendo así una rápida integración con herramientas de minería de datos, análisis y reportes (dashboards).

En ambos la limpieza de datos también conocida como «Data Cleansing» engloba varios procesos destinados a mejorar la calidad de los datos. Estos procesos se utilizan para corregir o eliminar registros inexactos en una base de datos o conjunto de datos.

Parte 2: Selección y limpieza de los Datos en Python

```
import pandas as pd
import numpy as np
import random
```

```
# Cargar el archivo y mostrar las primeras columnas
df = pd.read_csv ('https://raw.githubusercontent.com/PosgradoMNA/Actividades_Aprendiz
df
```

	ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X15	X16	X
0	1	20000	2.0	2.0	1.0	24.0	2.0	2.0	-1.0	-1.0	...	0.0	0.0	
1	2	120000	2.0	2.0	2.0	26.0	-1.0	2.0	0.0	0.0	...	3272.0	3455.0	326
2	3	90000	2.0	2.0	2.0	34.0	0.0	0.0	0.0	0.0	...	14331.0	14948.0	1554
3	4	50000	2.0	2.0	1.0	37.0	0.0	0.0	0.0	0.0	...	28314.0	28959.0	2954
4	5	50000	1.0	2.0	1.0	57.0	-1.0	0.0	-1.0	0.0	...	20940.0	19146.0	1913
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
29995	29996	220000	1.0	3.0	1.0	39.0	0.0	0.0	0.0	0.0	...	88004.0	31237.0	1598
29996	29997	150000	1.0	3.0	2.0	43.0	-1.0	-1.0	-1.0	-1.0	...	8979.0	5190.0	
29997	29998	30000	1.0	2.0	2.0	37.0	4.0	3.0	2.0	-1.0	...	20878.0	20582.0	1935
29998	29999	80000	1.0	3.0	1.0	41.0	1.0	-1.0	0.0	0.0	...	52774.0	11855.0	4894
29999	30000	50000	1.0	2.0	1.0	46.0	0.0	0.0	0.0	0.0	...	36535.0	32428.0	1531

30000 rows × 25 columns

```
# Validar si hay valores que falten
df.isnull().any()
```

```
ID     False
X1     False
X2      True
X3      True
X4      True
X5      True
X6      True
X7      True
X8      True
X9      True
X10     True
X11     True
X12     True
X13     True
X14     True
X15     True
X16     True
X17     True
X18     True
X19     True
X20     True
X21     True
X22     True
X23     True
Y       True
dtype: bool
```

```
# Eliminar renglones donde falte demasiada informacion.
nonEmptyDataFrame = df.copy()
nonEmptyDataFrame.dropna(thresh=6, inplace = True)



#print(random.choice([1,2]))
# reemplazar el <x2 - gender> con valores random
nonGenderDataFrame = nonEmptyDataFrame.copy()
nonGenderDataFrame = nonGenderDataFrame.astype({'X2':'Int64'})
nonGenderDataFrame['X2'].fillna(value = random.choice([1,2]), inplace = True)
nonGenderDataFrame['X2'].value_counts()
```

```
2    18112
1    11887
Name: X2, dtype: Int64
```

```
# reemplazar el <X3 - Education> con valor 'other' y los valores que tenga valores ba
nonEducationDT = nonGenderDataFrame.copy()
nonEducationDT = nonEducationDT.astype({'X3':'Int64'})
nonEducationDT['X3'].replace({5:4, 0:4, 6:4}, inplace = True)
nonEducationDT['X3'].fillna(4, inplace = True)
nonEducationDT['X3'].value_counts()
```

```
2    14030
1    10585
3     4915
4      469
Name: X3, dtype: Int64
```

```
# reemplazar el <X4 - Marital> con valor 'other' y valores basura tambien con 'other'
nonMaritalDT = nonEducationDT.copy()
nonMaritalDT = nonMaritalDT.astype({'X4':'Int64'})
nonMaritalDT['X4'].replace(0, 3, inplace = True)
nonMaritalDT['X4'].fillna(3, inplace = True)
nonMaritalDT['X4'].value_counts()
```

```
2    15964
1    13657
3      378
Name: X4, dtype: Int64
```

```
# reemplazar el <X5: Age (year)> con valor mediana
nonAgeDT = nonMaritalDT.copy()
nonAgeDT = nonAgeDT.astype({'X5':'Int64'})
edadMediana = nonAgeDT.X5.median()
```

```
nonAgeDT['X5'].fillna(value = edadMediana, inplace = True)
nonAgeDT['X5'].value_counts()
```

```
29    1605
27    1477
28    1409
30    1395
26    1256
31    1217
25    1186
34    1166
32    1158
33    1146
24    1127
35    1113
36    1108
37    1041
39     954
38     943
23     931
40     870
41     823
42     794
44     700
43     669
45     617
46     570
22     560
47     499
48     466
49     452
50     411
51     340
53     325
52     304
54     247
55     209
56     178
58     122
57     122
59      83
60      67
21      67
61      56
62      44
64      31
63      31
66      25
65      24
67      16
69      15
70      10
68       5
73       4
```

```
72       3
75       3
71       3
79       1
74       1
Name: X5, dtype: Int64
```

```
# finalmente quitar todos los renglones que sobren porque no contener datos para proc
finalDT = nonAgeDT.copy()
finalDT.fillna(0, inplace = True)
finalDT.isnull().any()
```

```
ID     False
X1     False
X2     False
X3     False
X4     False
X5     False
X6     False
X7     False
X8     False
X9     False
X10    False
X11    False
X12    False
X13    False
X14    False
X15    False
X16    False
X17    False
X18    False
X19    False
X20    False
X21    False
X22    False
X23    False
Y      False
dtype: bool
```

Parte 3: Con base en los resultados de tu libreta de Google Colab de la Parte 2 responde detalladamente las siguientes preguntas:

¿Qué datos considero mas importantes? ¿Por qué? Consideramos que los datos mas importantes con los pagos porque ahi podemos conseguir informacion para otorgar credito o rechazarlo. Tambien consideramos importante la parte demografica para entender mejor a los clientes.

¿Se eliminaron o reemplazaron datos nulos? ¿Qué se hizo y por qué? En la parte demografica (x1-x6) reemplazamos los valores nulos con los 'other' y corregimos los valores que estaban fuera del rango.

¿Es necesario ordenar los datos para el análisis? Sí / No / ¿Por qué? Ordenar (sort) siento que no influye aun. Ordenar (clean) si es necesario porque la tabla traia muchos valores basura y vacio. Si se ocupaba corregir muchos valores.

¿Existen problemas de formato que deban solucionar antes del proceso de modelado? Sí / No / Por qué. si, muchos valores para matrimonio y genero venian en tipo flotante. Los tuvimos que corregir a int. Un resultado en flotante para este tipo de variables no tendria mucho sentido.

¿Qué ajustes se realizaron en el proceso de limpieza de datos (agregar, integrar, eliminar, modificar registros (filas), cambiar atributos (columnas)? En el caso de matrimonio fue agregar. Para matrimonio y estudios se tuvo que definir como 'other'. Para valores de pagos e historiales de pagos que estaban vacios se reemplazaron con zero.