

Business Case

Juan Pablo Hernández Urrego

Contenido

1. Conocimiento de la muestra de datos.
2. Manejo de *Missing Data*.
3. Respuesta preguntas sobre la data.
4. Modelo de nivel de riesgo.

Conocimiento de la muestra

Age
Sex
Job
Housing
Saving accounts
Checking account
Credit amount
Duration
Purpose

Risk



Tenemos 10 variables que ubicamos en 3 categorías:

Carácter del usuario



Información financiera



Información del crédito

Conocimiento de la muestra

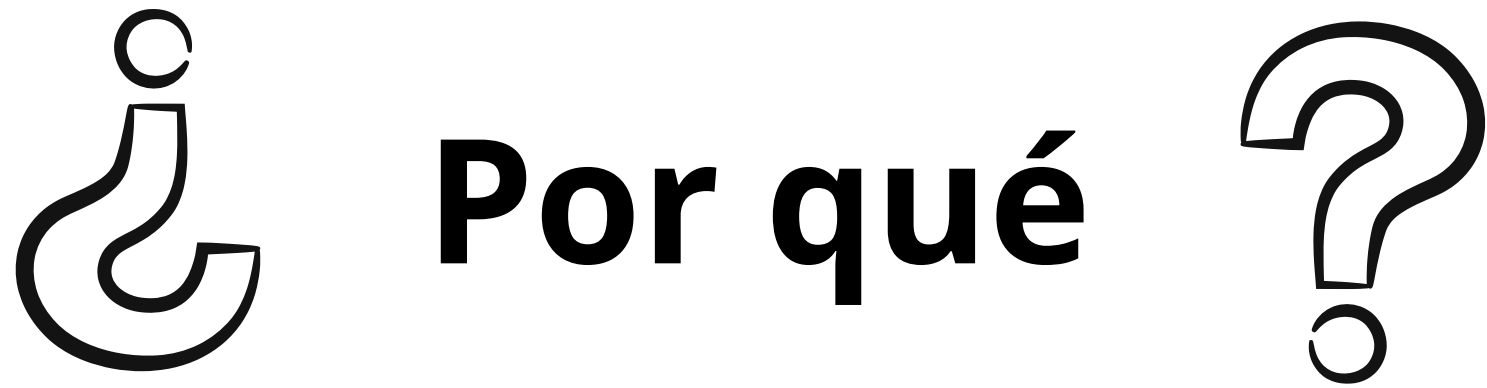
Lo primero que hice fue ver el tipo de cada variable.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1000 entries, 0 to 999
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Age                   1000 non-null   int64
1   Sex                   1000 non-null   object
2   Job                   1000 non-null   int64
3   Housing               1000 non-null   object
4   Saving accounts       817 non-null    object
5   Checking account      606 non-null    object
6   Credit amount         1000 non-null   int64
7   Duration              1000 non-null   int64
8   Purpose               1000 non-null   object
9   Risk                  1000 non-null   object
dtypes: int64(4), object(6)
memory usage: 85.9+ KB
```

Tenemos *Missing Data* en las variables que definimos como "Información financiera"

Manejo de *Missing Data*

Antes de continuar con el conocimiento de la muestra necesito determinar cómo voy a manejar la información faltante.

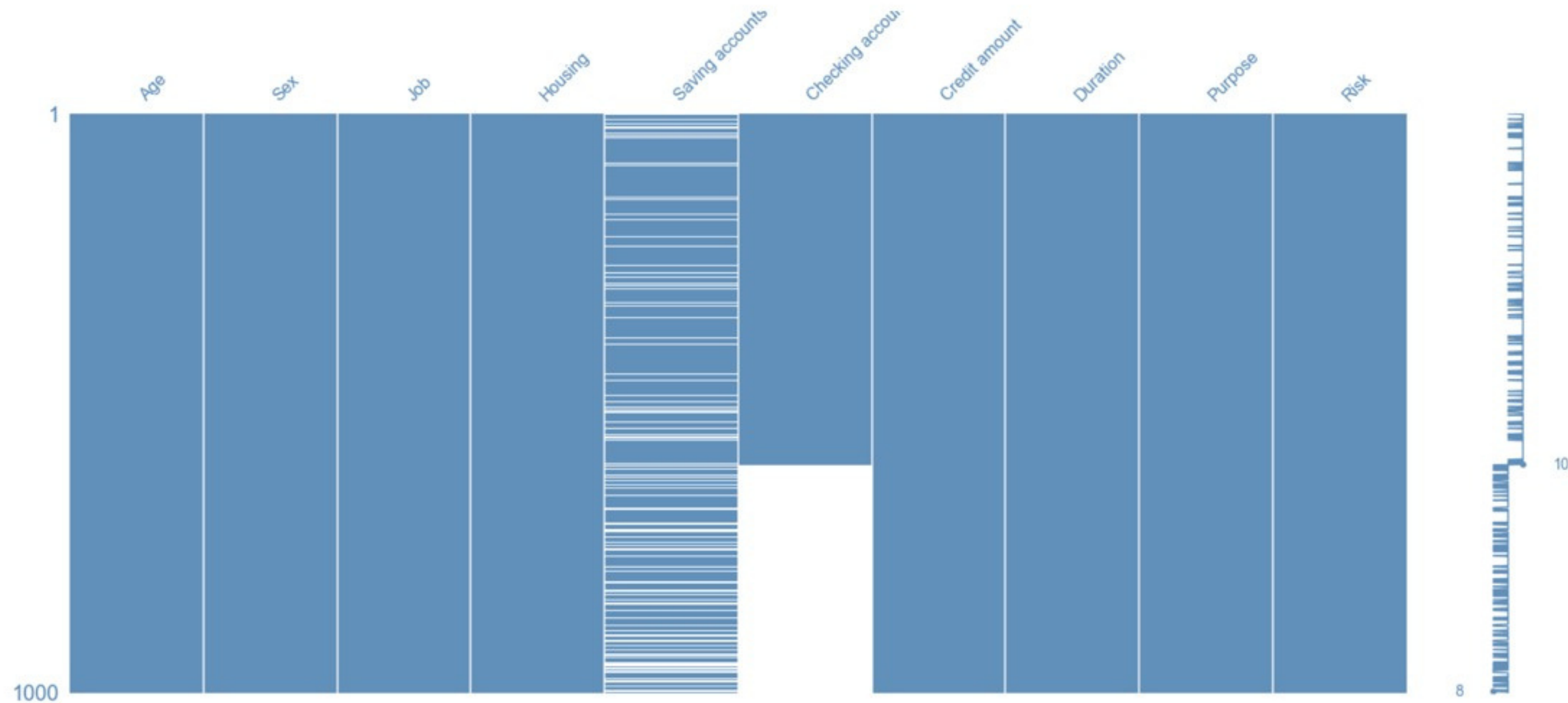


Porque **"Sí pasa con Conekta"**.

Queremos minimizar que clientes potenciales terminen como "falsos negativos" porque no tenemos la información suficiente.

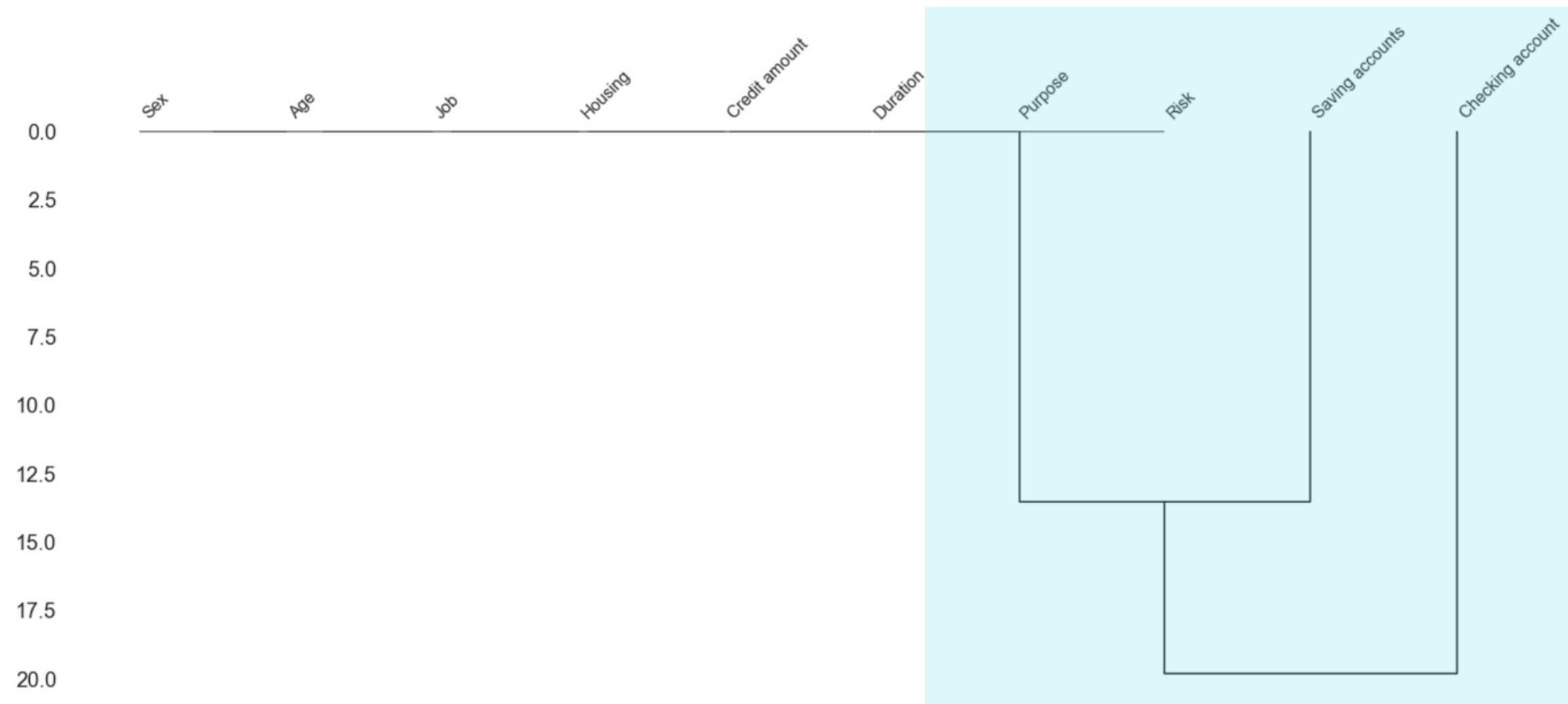
Manejo de *Missing Data*

Como los valores faltantes se encuentran en dos variables semi-continuas es difícil observar correlación.



Manejo de *Missing Data*

Con este dendrograma se puede observar que existe una correlación entre las variables Purpose y Saving accounts.



Manejo de *Missing Data*

Entonces tenemos las siguientes opciones:

~~Eliminar las observaciones
con *Missing Data*~~

K-Nearest Neighbor

**Reemplazar la *Missing Data*
con el promedio**

**Multiple Imputation by
Chained Equations**

Manejo de *Missing Data*

Luego de analizar cuál método se ajusta más a las observaciones de la muestra tenemos que:

~~Eliminar las observaciones
con *Missing Data*~~

K-Nearest Neighbor

Reemplazar la *Missing Data*
con el promedio

Multiple Imputation by
Chained Equations



Manejo de *Missing Data*

El método MICE produjo el R-squared más grande entre todas las opciones.

	No_NaN	Mean	KNN	MICE
R_squared_adj	0.119232	0.088166	0.090625	0.097345

	No_NaN	Mean	KNN	MICE
const	0.632379	0.633370	0.635189	0.616173
Age	0.001542	0.002478	0.002426	0.002369
Sex	0.084035	0.075786	0.075668	0.075168
Job	-0.021631	-0.001005	-0.001104	-0.000743
Housing	-0.015408	-0.012568	-0.011500	-0.011544
Saving accounts	0.072457	0.071053	0.076436	0.070798
Checking account	0.082023	0.101757	0.091168	0.120611
Credit amount	0.000297	0.000150	0.000155	0.000156
Duration	-0.024907	-0.015973	-0.015930	-0.015907
Purpose	0.013838	0.014638	0.014523	0.014175

Respuesta preguntas sobre la data

Una vez solucionamos el problema de *Missing Data* en las variables de "Información financiera" podemos continuar con el conocimiento de la muestra.

- **El 75% de los usuarios es menor de 42 años.**

¿Cuáles son las diferencias con un usuario de más de 42 años?

- **2/3 partes de los usuarios son hombres.**

¿Tiene relación con el propósito de crédito más frecuente?

- **El ticket promedio de un crédito es de alrededor 3.200 MXN.**

¿Por qué hay una solicitud de crédito por 18.424 MXN?

- **El plazo promedio de un crédito es de 21 meses.**

¿Cuál es el destino de un crédito a 72 meses?

Modelo de nivel de riesgo

Por último, con el previo conocimiento de la data se construye un modelo Logit para hallar la probabilidad de, dadas las variables anteriormente expuestas, se clasifique como un usuario de riesgo "good".

$$\begin{aligned} \text{Risk} = & \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Sex} + \beta_3 \text{Job} + \beta_4 \text{Housing} + \beta_5 \text{Saving accounts} \\ & + \beta_6 \text{Checking account} + \beta_7 \text{Credit amount} + \beta_8 \text{Duration} \\ & + \beta_9 \text{Purpose} + v \end{aligned}$$

Muchas gracias