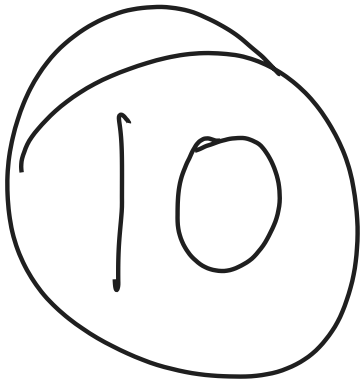
PRÁCTICA DE LABORATORIO:



INTRODUCCIÓN A LA CIENCIA DE DATOS

DÍAZ RODRIGUEZ SANDRO FARID /CANALES CORONA MIA

ANÁLISIS INICIAL: Resumen estadístico de la base de datos antes de la limpieza.

- Edad:
  - Media: 39.72 años, con un rango de 18 a 80 años.
  - La mayoría de los clientes tiene entre 32 y 48 años (percentiles 25% y 75%).
- Duración del préstamo:
  - Media: 54.18 meses, con un rango de 12 a 120 meses.
  - El 50% de los préstamos tiene una duración entre 36 y 72 meses.
- Estado civil:
  - No hay datos.
- Número de dependientes:
  - Media: 1.51 dependientes, con un rango de 0 a 5.
  - El 50% de los clientes tiene entre 0 y 2 dependientes.
- Estado de propiedad de vivienda:
  - No hay datos.
- Número de consultas de crédito:
  - Media: 0.99 consultas, con un rango de 0 a 7 consultas.
  - El 50% de los clientes tiene entre 0 y 2 consultas.
- Historial de quiebra:
  - El 5.18% de los clientes tiene un historial de quiebra.
- Propósito del préstamo:
  - No hay datos.
- Historial de impagos en préstamos previos:
  - El 10% de los clientes tiene impagos previos en préstamos.
- Activos totales:
  - Media: $96,630, con un rango de $2,098 a $2,619,627.

- El 50% de los clientes tiene activos entre $31,169 y $116,764.

 Antigüedad laboral:

- Media: 4.99 años, con un rango de 0 a 16 años.
- El 50% de los clientes tiene entre 3 y 6 años de antigüedad laboral.

 Tasa de interés:

- Media: 23.93%, con un rango de 11.33% a 44.68%.
- El 50% de los préstamos tiene tasas entre 20.93% y 26.57%.

 Puntuación de riesgo:

- Media: 50.76, con un rango de 28.8 a 84.
- El 50% de los clientes tiene un puntaje de riesgo entre 46 y 56.

Tabla que muestre el porcentaje de valores faltantes por columna.

	Porcentaje Valores Faltantes
ApplicationDate	2.484093
Age	4.201168
AnnualIncome	2.453587
CreditScore	2.301055
EmploymentStatus	2.728144
EducationLevel	2.593045
Experience	2.558180
LoanAmount	2.488451
LoanDuration	4.445219
MaritalStatus	100.000000
NumberOfDependents	4.615183
HomeOwnershipStatus	100.000000
MonthlyDebtPayments	2.763009
CreditCardUtilizationRate	2.593045
NumberOfOpenCreditLines	2.505883
NumberOfCreditInquiries	4.266539
DebtToIncomeRatio	2.562538
BankruptcyHistory	4.340626
LoanPurpose	100.000000
PreviousLoanDefaults	4.358058
PaymentHistory	2.688922
LengthOfCreditHistory	2.627909
SavingsAccountBalance	2.654057
CheckingAccountBalance	2.566896
TotalAssets	4.179378
TotalLiabilities	2.562538
MonthlyIncome	2.579970
UtilityBillsPaymentHistory	2.558180
JobTenure	4.549813
NetWorth	2.505883
BaseInterestRate	2.571254
InterestRate	4.405997
MonthlyLoanPayment	2.588686
TotalDebtToIncomeRatio	2.654057
LoanApproved	2.771725
RiskScore	4.392923

Total de filas duplicadas encontradas.

Total de filas duplicadas: 156

Descripción de los tipos de datos originales y los problemas encontrados.

```
Resumen estadístico general:
              Edad   IngresoAnual   CreditScore   Experiencia         Monto  \
count  20494.000000   20494.000000  20494.000000  20494.000000   20494.000000
mean      39.772519   56161.411925    541.464575     17.486581   23551.067044
std       11.282680   41443.240046    137.240448     10.955604   14226.083351
min       18.000000       0.000000      0.000000      0.000000       0.000000
25%       32.000000   29175.000000    531.000000     10.000000   14486.000000
50%       40.000000   46663.000000    575.000000     17.000000   21221.500000
75%       47.000000   72412.000000    607.000000     25.000000   30149.000000
max       80.000000  485341.000000    712.000000     61.000000  158686.000000

           Duración  NumberOfDependents  MonthlyDebtPayments  \
count  20494.000000        20494.000000         20494.000000
mean      51.098077            1.432907           432.206060
std       26.890628            1.388819           254.818149
min        0.000000            0.000000             0.000000
25%       36.000000            0.000000           269.000000
50%       48.000000            1.000000           391.000000
75%       60.000000            2.000000           552.000000
max      120.000000            5.000000          2919.000000

       CreditCardUtilizationRate  NumberOfOpenCreditLines  ...  \
count               20494.000000              20494.00000  ...
mean                    0.270727                  2.85425  ...
std                     0.168233                  1.82766  ...
min                     0.000000                  0.00000  ...
25%                     0.142023                  2.00000  ...
50%                     0.253469                  3.00000  ...
75%                     0.381994                  4.00000  ...
max                     0.917380                 13.00000  ...

       CheckingAccountBalance   TotalAssets  TotalLiabilities   IngresoMens  \
count            20494.000000  2.049400e+04      2.049400e+04  20494.000000
mean              1689.083195  9.179241e+04      3.414402e+04   4631.668842
std               2227.688244  1.201573e+05      4.694795e+04   3400.032117
min                  0.000000  0.000000e+00      0.000000e+00      0.000000
25%                484.000000  2.701450e+04      9.795250e+03   2410.375000
```

```
25%                484.000000  2.701450e+04      9.795250e+03   2410.375000
50%               1033.000000  5.609250e+04      2.067800e+04   3864.375000
75%               2046.000000  1.125320e+05      4.122225e+04   5980.062500
max              52572.000000  2.619627e+06      1.417302e+06  25000.000000

       UtilityBillsPaymentHistory     JobTenure       NetWorth  \
count                20494.000000  20494.000000  2.049400e+04
mean                     0.756071      4.719576  6.807870e+04
std                      0.216843      2.443620  1.139525e+05
min                      0.000000      0.000000  0.000000e+00
25%                      0.705927      3.000000  7.397750e+03
50%                      0.810912      5.000000  2.848650e+04
75%                      0.887961      6.000000  8.386575e+04
max                      0.999433     15.000000  2.603208e+06

       MonthlyLoanPayment  LoanApproved     RiskScore
count        20494.000000  20494.000000  20494.000000
mean           861.759970      0.238460     50.777047
std            683.615359      0.426152      7.789086
min              0.000000      0.000000     28.800000
25%            456.018655      0.000000     46.000000
50%            699.545027      0.000000     52.000000
75%           1082.140970      0.000000     56.000000
max          10892.629520      1.000000     84.000000

[8 rows x 27 columns]

Hay 103 filas duplicadas en el DataFrame.

Porcentaje de valores faltantes por columna:
Series([], dtype: float64)

Tipos de datos por columna:
FechaPrestamo                  object
Edad                            int64
IngresoAnual                  float64
CreditScore                     int64
```

```
CreditScore                    int64
Situ_Empleo                   object
Educación                     object
Experiencia                    int64
Monto                          int64
Duración                       int64
MaritalStatus                 object
NumberOfDependents             int64
HomeOwnershipStatus           object
MonthlyDebtPayments            int64
CreditCardUtilizationRate    float64
NumberOfOpenCreditLines        int64
NumberOfCreditInquiries        int64
DebtToIncomeRatio            float64
BankruptcyHistory            float64
LoanPurpose                   object
PreviousLoanDefaults           int64
PaymentHistory                 int64
LengthOfCreditHistory          int64
SavingsAccountBalance          int64
CheckingAccountBalance         int64
TotalAssets                    int64
TotalLiabilities               int64
IngresoMens                  float64
UtilityBillsPaymentHistory   float64
JobTenure                      int64
NetWorth                       int64
MonthlyLoanPayment           float64
LoanApproved                 float64
RiskScore                    float64
dtype: object
```

```
Primeras 5 filas del DataFrame limpio:
  FechaPrestamo  Edad  IngresoAnual  CreditScore Situ_Empleo    Educación  \
0    2018-01-01    45       39948.0          617  No_especif       Master
1    2018-01-02    38       39709.0          628    Empleado    Associate
2    2018-01-03    47       40724.0          570    Empleado     Bachelor
3    2045-05-10    58       69084.0          545    Empleado  High School
4    2018-01-05    37      103264.0          594    Empleado    Associate

   Experiencia  Monto  Duración MaritalStatus  ...  CheckingAccountBalance  \
0           17  13152        48        Casado  ...                    1202
1           15  26045        48       Soltero  ...                    3460
2           26  17627        36        Casado  ...                     895
3           34  37898        96       Soltero  ...                    1217
4           17   9184        36        Casado  ...                    4981

  TotalAssets  TotalLiabilities  IngresoMens  UtilityBillsPaymentHistory  \
0      146111             19183  3329.000000                    0.724972
1       53204              9595  3309.083333                    0.935132
2       25176            128874  3393.666667                    0.872241
3           0              5370     0.000000                    0.896155
4      244305             17286  8605.333333                    0.941369

   JobTenure  NetWorth  MonthlyLoanPayment LoanApproved  RiskScore
0         11    126928          419.805992          0.0       49.0
1          0     43609          794.054238          0.0       52.0
2          6      5205          666.406688          0.0       52.0
3          5     99452         1047.506980          0.0       54.0
4          5    227019          330.179140          1.0       36.0

[5 rows x 33 columns]
```

```
Data columns (total 36 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   ApplicationDate              20000 non-null  object
 1   Age                          20000 non-null  int64
 2   AnnualIncome                 20000 non-null  int64
 3   CreditScore                  20000 non-null  int64
 4   EmploymentStatus             20000 non-null  object
 5   EducationLevel               20000 non-null  object
 6   Experience                   20000 non-null  int64
 7   LoanAmount                   20000 non-null  int64
 8   LoanDuration                 20000 non-null  int64
 9   MaritalStatus                20000 non-null  object
 10  NumberOfDependents           20000 non-null  int64
 11  HomeOwnershipStatus          20000 non-null  object
 12  MonthlyDebtPayments          20000 non-null  int64
 13  CreditCardUtilizationRate    20000 non-null  float64
 14  NumberOfOpenCreditLines      20000 non-null  int64
 15  NumberOfCreditInquiries      20000 non-null  int64
 16  DebtToIncomeRatio            20000 non-null  float64
 17  BankruptcyHistory            20000 non-null  int64
 18  LoanPurpose                  20000 non-null  object
 19  PreviousLoanDefaults         20000 non-null  int64
 20  PaymentHistory               20000 non-null  int64
 21  LengthOfCreditHistory        20000 non-null  int64
 22  SavingsAccountBalance        20000 non-null  int64
 23  CheckingAccountBalance       20000 non-null  int64
 24  TotalAssets                  20000 non-null  int64
 25  TotalLiabilities             20000 non-null  int64
 26  MonthlyIncome                20000 non-null  float64
 27  UtilityBillsPaymentHistory   20000 non-null  float64
 28  JobTenure                    20000 non-null  int64
 29  NetWorth                     20000 non-null  int64
 30  BaseInterestRate             20000 non-null  float64
 31  InterestRate                 20000 non-null  float64
 32  MonthlyLoanPayment           20000 non-null  float64
```

```
 31  InterestRate               20000 non-null  float64
 32  MonthlyLoanPayment         20000 non-null  float64
 33  TotalDebtToIncomeRatio     20000 non-null  float64
 34  LoanApproved               20000 non-null  int64
 35  RiskScore                  20000 non-null  float64
dtypes: float64(9), int64(21), object(6)
memory usage: 5.5+ MB

Primeras filas de la base de datos:
  ApplicationDate  Age  AnnualIncome  CreditScore EmploymentStatus  \
0      2018-01-01   45         39948          617         Employed
1      2018-01-02   38         39709          628         Employed
2      2018-01-03   47         40724          570         Employed
3      2018-01-04   58         69084          545         Employed
4      2018-01-05   37        103264          594         Employed

  EducationLevel  Experience  LoanAmount  LoanDuration MaritalStatus  ...  \
0         Master          22       13152            48       Married  ...
1      Associate          15       26045            48        Single  ...
2       Bachelor          26       17627            36       Married  ...
3    High School          34       37898            96        Single  ...
4      Associate          17        9184            36       Married  ...

   MonthlyIncome UtilityBillsPaymentHistory  JobTenure  NetWorth  \
0    3329.000000                  0.724972         11    126928
1    3309.083333                  0.935132          3     43609
2    3393.666667                  0.872241          6      5205
3    5757.000000                  0.896155          5     99452
4    8605.333333                  0.941369          5    227019

   BaseInterestRate  InterestRate  MonthlyLoanPayment  TotalDebtToIncomeRatio  \
0          0.199652      0.227590          419.805992                0.181077
1          0.207045      0.201077          794.054238                0.389852
2          0.217627      0.212548          666.406688                0.462157
3          0.300398      0.300911         1047.506980                0.313098
4          0.197184      0.175990          330.179140                0.070210
```

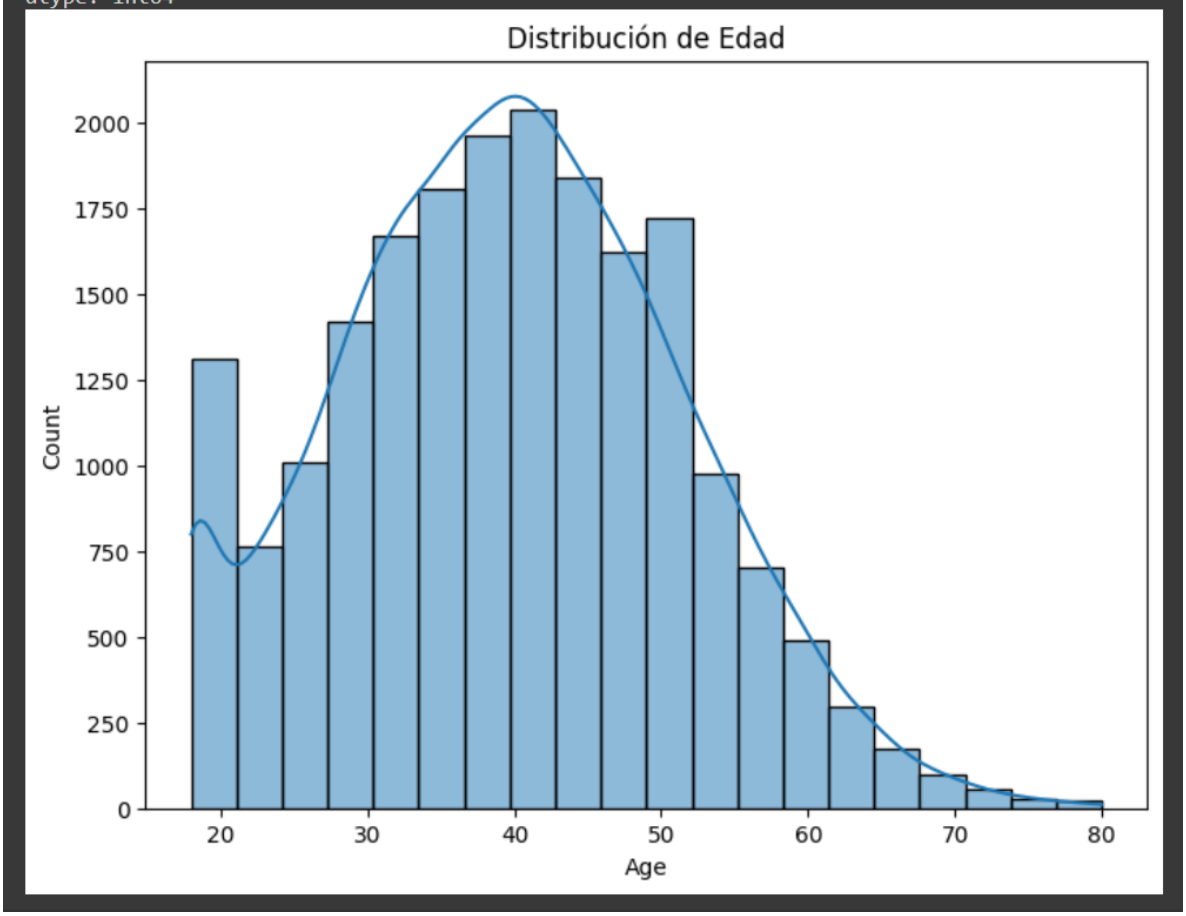
```
  LoanApproved  RiskScore
0            0       49.0
1            0       52.0
2            0       52.0
3            0       54.0
4            1       36.0

[5 rows x 36 columns]

Descripción estadística de las columnas numéricas:
                Age   AnnualIncome   CreditScore    Experience     LoanAmount  \
count  20000.000000   20000.000000  20000.000000  20000.000000   20000.000000
mean      39.752600   59161.473550    571.612400     17.522750   24882.867800
std       11.622713   40350.845168     50.997358     11.316836   13427.421217
min       18.000000   15000.000000    343.000000      0.000000    3674.000000
25%       32.000000   31679.000000    540.000000      9.000000   15575.000000
50%       40.000000   48566.000000    578.000000     17.000000   21914.500000
75%       48.000000   74391.000000    609.000000     25.000000   30835.000000
max       80.000000  485341.000000    712.000000     61.000000  184732.000000

       LoanDuration  NumberOfDependents  MonthlyDebtPayments  \
count  20000.000000        20000.000000         20000.000000
mean      54.057000            1.517300           454.292700
std       24.664857            1.386325           240.507609
min       12.000000            0.000000            50.000000
25%       36.000000            0.000000           286.000000
50%       48.000000            1.000000           402.000000
75%       72.000000            2.000000           564.000000
max      120.000000            5.000000          2919.000000

       CreditCardUtilizationRate  NumberOfOpenCreditLines  ...  MonthlyIncome  \
count               20000.000000             20000.000000  ...   20000.000000
mean                    0.286381                 3.023350  ...    4891.715521
std                     0.159793                 1.736161  ...    3296.771598
min                     0.000974                 0.000000  ...    1250.000000
25%                     0.160794                 2.000000  ...    2629.583333
```

```
25%                    0.160794                2.000000  ...    2629.583333
50%                    0.266673                3.000000  ...    4034.750000
75%                    0.390634                4.000000  ...    6163.000000
max                    0.917380               13.000000  ...   25000.000000

       UtilityBillsPaymentHistory     JobTenure      NetWorth  \
count                20000.000000  20000.000000  2.000000e+04
mean                     0.799918      5.002650  7.229432e+04
std                      0.120665      2.236804  1.179200e+05
min                      0.259203      0.000000  1.000000e+03
25%                      0.727379      3.000000  8.734750e+03
50%                      0.820962      5.000000  3.285550e+04
75%                      0.892333      6.000000  8.882550e+04
max                      0.999433     16.000000  2.603208e+06

       BaseInterestRate  InterestRate  MonthlyLoanPayment  \
count      20000.000000  20000.000000        20000.000000
mean           0.239124      0.239110          911.607052
std            0.035509      0.042205          674.583473
min            0.130101      0.113310           97.030193
25%            0.213889      0.209142          493.763700
50%            0.236157      0.235390          728.511452
75%            0.261533      0.265532         1112.770759
max            0.405029      0.446787        10892.629520

       TotalDebtToIncomeRatio  LoanApproved     RiskScore
count            20000.000000  20000.000000  20000.000000
mean                 0.402182      0.239000     50.766780
std                  0.338924      0.426483      7.778262
min                  0.016043      0.000000     28.800000
25%                  0.179693      0.000000     46.000000
50%                  0.302711      0.000000     52.000000
75%                  0.509214      0.000000     56.000000
max                  4.647657      1.000000     84.000000

[8 rows x 30 columns]
```

```
Descripción de las columnas categóricas:
       ApplicationDate EmploymentStatus EducationLevel MaritalStatus  \
count            20000            20000          20000         20000
unique           20000                3              5             4
top         2018-01-01         Employed       Bachelor       Married
freq                1            17036           6054         10041

       HomeOwnershipStatus LoanPurpose
count                20000       20000
unique                   4           5
top               Mortgage        Home
freq                  7939        5925
```

```
Conteo de valores únicos por columna:
ApplicationDate                20000
Age                               63
AnnualIncome                   17516
CreditScore                      322
EmploymentStatus                   3
EducationLevel                     5
Experience                        62
LoanAmount                     15578
LoanDuration                      10
MaritalStatus                      4
NumberOfDependents                 6
HomeOwnershipStatus                4
MonthlyDebtPayments             1299
CreditCardUtilizationRate      20000
NumberOfOpenCreditLines           14
NumberOfCreditInquiries            8
DebtToIncomeRatio              20000
BankruptcyHistory                  2
LoanPurpose                        5
PreviousLoanDefaults               2
PaymentHistory                    38
LengthOfCreditHistory             29
SavingsAccountBalance           9199
CheckingAccountBalance          5151
TotalAssets                    18814
TotalLiabilities               17163
MonthlyIncome                  17489
UtilityBillsPaymentHistory     20000
JobTenure                         17
NetWorth                       17724
BaseInterestRate               18742
InterestRate                   19999
MonthlyLoanPayment             20000
TotalDebtToIncomeRatio         20000
LoanApproved                       2
```

```
MonthlyLoanPayment          20000
TotalDebtToIncomeRatio      20000
LoanApproved                    2
RiskScore                      73
dtype: int64
```



Había muchas filas duplicadas, como datos irrelevantes, “aaaa” en algunos casos en incluso NaN en muchos otros.

PROCESO DE LIMPIEZA:

Describir qué métodos utilizaron para limpiar la base de datos

 Age (Edad):

- Tipo esperado: Numérico (entero).
- Descripción: Representa la edad del cliente en años.

 LoanDuration (Duración del préstamo):

- Tipo esperado: Numérico (entero).
- Descripción: Duración del préstamo en meses.

 MaritalStatus (Estado civil):

- Tipo esperado: Categórico (cadena de texto o valor numérico codificado).
- Descripción: Indica el estado civil del cliente, como soltero, casado, divorciado, etc.

 NumberOfDependents (Número de dependientes):

- Tipo esperado: Numérico (entero).
- Descripción: Número de dependientes que tiene el cliente.

 HomeOwnershipStatus (Estado de propiedad de la vivienda):

- Tipo esperado: Categórico.
- Descripción: Indica si el cliente es propietario o inquilino de su vivienda.

 NumberOfCreditInquiries (Número de consultas de crédito):

- Tipo esperado: Numérico (entero).
- Descripción: Número de veces que se ha consultado el historial crediticio del cliente.

 BankruptcyHistory (Historial de quiebra):

- Tipo esperado: Categórico (binario, 0 o 1).
- Descripción: Indica si el cliente ha tenido un historial de quiebra (1) o no (0).

 LoanPurpose (Propósito del préstamo):

- Tipo esperado: Categórico.
- Descripción: El motivo por el cual el cliente solicita el préstamo, como compra de vehículo, hogar, educación, etc.

 PreviousLoanDefaults (Impagos previos de préstamos):

- Tipo esperado: Categórico (binario, 0 o 1).
- Descripción: Indica si el cliente ha incumplido pagos en préstamos anteriores.

 TotalAssets (Activos totales):

- Tipo esperado: Numérico (flotante).
- Descripción: El valor total de los activos del cliente en dólares.

- JobTenure (Antigüedad laboral):

  - Tipo esperado: Numérico (entero).
  - Descripción: Número de años que el cliente lleva en su trabajo actual.

- InterestRate (Tasa de interés):

  - Tipo esperado: Numérico (flotante).
  - Descripción: Tasa de interés del préstamo.

- RiskScore (Puntuación de riesgo):

  - Tipo esperado: Numérico (flotante o entero).
  - Descripción: Puntuación que refleja el riesgo crediticio del cliente.

Problemas encontrados en la base de datos

1. Valores faltantes (Missing Data):
   - Varias columnas tienen valores faltantes, como:
     - MaritalStatus, HomeOwnershipStatus, y LoanPurpose muestran NaN, indicando que no se cuenta con datos en esas columnas.
     - Algunas filas tienen valores faltantes en LoanDuration, NumberOfDependents, JobTenure, etc.

2. Desbalance en los datos:
   - Algunas columnas, como BankruptcyHistory y PreviousLoanDefaults, tienen muchos valores "0" (sin historial de quiebra o sin impagos previos), lo que puede significar un desbalance en la distribución de estos datos.

3. Errores de tipo de datos:
   - Es posible que algunos datos categóricos, como MaritalStatus o HomeOwnershipStatus, estén codificados de forma inadecuada o faltante, como NaN, lo que podría ser un problema de conversión o errores en la captura de datos.

4. Outliers o valores extremos:

- En la columna **TotalAssets**, los valores van desde $2,098 hasta $2,619,627, lo que podría indicar la presencia de valores atípicos extremos que pueden distorsionar el análisis.
- Igualmente, en **InterestRate**, los valores mínimos y máximos pueden ser extremos y requieren análisis adicional para verificar si son válidos.

5. **Distribución de los datos**:
   - En algunos atributos, como **NumberOfCreditInquiries**, hay una gran concentración en valores bajos, mientras que algunos clientes tienen valores muy altos (hasta 7 consultas), lo que también puede indicar outliers o casos especiales.

6. **Posible multicolinealidad**:
   - Atributos como **LoanDuration** y **InterestRate** pueden estar correlacionados con **RiskScore**, lo que podría causar redundancia en los modelos predictivos.

Antes y después de cada paso clave

Traducir columnas

	ApplicationDate	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	MaritalStatus	...	MonthlyIncome
0	2018-01-01	45.0	39948.0	617.0	Employed	Master	22.0	13152.0	NaN	NaN	...	3329.0

	FechaPrestamo	Edad	IngresoAnual	CreditScore	Situ_Empleo	Educación	Experiencia	Monto	Duración	MaritalStatus	.|.	CheckingAccountBalance	TotalAssets	T
0	2018-01-01	45	39948.0	617	No_especif	Master	17	13152	48	Casado	...	1202	146111	
1	2018-01-02	38	39709.0	628	Empleado	Associate	15	26045	48	Soltero		3460	53204	

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22926 entries, 0 to 22925
Data columns (total 36 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   FechaPrestamo              22133 non-null  object
 1   Edad                       22099 non-null  object
 2   IngresoAnual               22170 non-null  object
 3   CreditScore                22114 non-null  object
 4   Situ_Empleo                22118 non-null  object
 5   Educación                  22113 non-null  object
 6   Experiencia                22100 non-null  object
 7   Monto                      22158 non-null  object
 8   Duración                   22139 non-null  object
 9   MaritalStatus              22133 non-null  object
 10  NumberOfDependents         22103 non-null  object
 11  HomeOwnershipStatus        22087 non-null  object
 12  MonthlyDebtPayments        22172 non-null  object
 13  CreditCardUtilizationRate  22126 non-null  object
 14  NumberOfOpenCreditLines    22138 non-null  object
 15  NumberOfCreditInquiries    22138 non-null  object
 16  DebtToIncomeRatio          22144 non-null  object
 17  BankruptcyHistory          22148 non-null  object
```

```
df.isnull().sum()
```

```
ApplicationDate              793
Age                          827
AnnualIncome                 756
CreditScore                  812
EmploymentStatus             808
EducationLevel               813
Experience                   826
LoanAmount                   768
LoanDuration                 787
MaritalStatus                793
NumberOfDependents           823
HomeOwnershipStatus          839
MonthlyDebtPayments          754
CreditCardUtilizationRate    800
NumberOfOpenCreditLines      788
NumberOfCreditInquiries      788
DebtToIncomeRatio            782
BankruptcyHistory            778
LoanPurpose                  811
```

Sustituir los “aaaa” por NaN o por valores 0 para la manipulación de éstos

BaseInterestRate	InterestRate	MonthlyLoanPayment	TotalDebtToIncomeRatio	LoanApproved	RiskScore
0.199652	0.227590	419.8059915607372	0.1810771978253941	NaN	49.0
aaaaa	0.201077	794.0542382198969	0.3898524480253535	0.0	52.0
0.217627	0.212548	666.4066876774697	0.4621569652325321	0.0	52.0
0.300398	0.300911	1047.5069802292967	0.3130983116604649	NaN	54.0
0.197184	0.175990	330.17914048482373	0.0702098474378087	aaaaa	36.0
...	...	...	...	...	...
0.243905	0.253025	648.378237344448	0.2052368314833539	0.0	59.0
0.233298	0.227744	365.8552923172616	0.0933475346223263	1.0	46.4
0.213757	0.222896	1006.7046228952942	0.0810706033924103	1.0	44.0
0.273271	0.310777	902.2443778268477	0.7822543563220032	aaaaa	49.0
0.248847	0.199991	547.2385007481284	0.1960667917356041	1.0	47.2

IngresoMens	UtilityBillsPaymentHistory	JobTenure	NetWorth	MonthlyLoanPayment	LoanApproved	RiskScore
3329.000000	0.724972	11	126928	419.805992	0.0	49.0
3309.083333	0.935132	0	43609	794.054238	0.0	52.0
3393.666667	0.872241	6	5205	666.406688	0.0	52.0
0.000000	0.896155	5	99452	1047.506980	0.0	54.0
8605.333333	0.941369	5	227019	330.179140	1.0	36.0
...	...	...	...	...	...	...
2808.916667	0.787107	0	2245	4713.127814	0.0	58.0
5486.000000	0.696379	4	67943	369.650806	0.0	60.0
5130.083333	0.753478	3	2384	0.000000	0.0	52.0
4054.583333	0.898416	9	5876	649.605095	0.0	53.0
4332.416667	0.837484	6	44452	0.000000	0.0	58.0

Borrar columnas que no son de ayuda

	ApplicationDate	Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	MaritalStatus	...	MonthlyIncome
0	2018-01-01	45.0	39948.0	617.0	Employed	Master	22.0	13152.0	NaN	NaN	...	3329.0
1	2018-01-02	38.0	39709.0	628.0	Employed	Associate	15.0	26045.0	NaN	NaN	...	3309.083333333333
2	2018-01-03	47.0	40724.0	570.0	Employed	Bachelor	26.0	17627.0	36.0	NaN	...	3393.6666666666665
3	2018-01-04	58.0	69084.0	545.0	Employed	High School	34.0	37898.0	96.0	NaN	...	NaN
4	NaN	37.0	103264.0	594.0	Employed	Associate	17.0	9184.0	36.0	NaN	...	8605.333333333334
...	...	...	...	...	...	...	...	...	...	...	...	...
22941	2070-11-22	41.0	46505.0	590.0	Employed	Master	16.0	23905.0	72.0	NaN	...	3875.4166666666665
22942	2054-06-16	43.0	75827.0	592.0	Employed	High School	aaaaa	14298.0	72.0	NaN	...	6318.916666666667
22943	2034-06-22	66.0	206739.0	595.0	Employed	High School	aaaaa	26257.0	36.0	NaN	...	17228.25
22944	2057-02-09	35.0	16019.0	565.0	Employed	High School	13.0	30771.0	84.0	NaN	...	1334.9166666666667
22945	2057-10-29	46.0	84292.0	578.0	Employed	Bachelor	24.0	22847.0	72.0	NaN	...	7024.333333333333

```
noa = noaa.drop('InterestRate', axis=1 )

noa
```

Python

	FechaPrestamo	Edad	IngresoAnual	CreditScore	Situ_Empleo	Educación	Experiencia	Monto	Duración	MaritalStatus	...	TotalLiabilities	IngresoMens	Utili
0	2018-01-01	45.0	39948.0	617.0	NaN	Master	NaN	13152.0	48.0	Casado	...	19183.0	3329.0	
1	2018-01-02	38.0	39709.0	628.0	Empleado	Associate	15.0	26045.0	48.0	Soltero	...	9595.0	3309.083333333333	
2	2018-01-03	47.0	40724.0	570.0	Empleado	Bachelor	26.0	17627.0	36.0	Casado	...	128874.0	3393.6666666666665	
3	NaN	58.0	69084.0	545.0	Empleado	High School	34.0	37898.0	96.0	Soltero	...	5370.0	NaN	
4	2018-01-05	37.0	103264.0	594.0	Empleado	Associate	17.0	9184.0	36.0	Casado	...	17286.0	8605.333333333334	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
22921	2024-05-03	47.0	33707.0	584.0	Empleado	Bachelor	25.0	86770.0	24.0	Casado	...	12989.0	2808.9166666666665	
22922	2068-04-07	60.0	65832.0	627.0	Empleado	Associate	39.0	10008.0	36.0	Casado	...	29122.0	5486.0	

Contabilizo datos NaN de cada columna y valoro cuantos si son de utilidad y cuales no

```
b.isnull().sum()
```

```
FechaPrestamo                1224
Edad                         1251
IngresoAnual                 1172
CreditScore                  1244
Situ_Empleo                  1254
Educación                    1252
Experiencia                  1306
Monto                        1261
Duración                     1210
MaritalStatus                1259
NumberOfDependents           1250
HomeOwnershipStatus          1279
MonthlyDebtPayments          1161
CreditCardUtilizationRate    1262
NumberOfOpenCreditLines      1238
NumberOfCreditInquiries      1230
DebtToIncomeRatio            1241
BankruptcyHistory            1214
LoanPurpose                  1227
PreviousLoanDefaults         1268
PaymentHistory               1233
LengthOfCreditHistory        1317
SavingsAccountBalance        1236
CheckingAccountBalance       1207
TotalAssets                  1229
TotalLiabilities             1184
IngresoMens                  1247
```

RESULTADOS:

Resumen final de la base de datos después de la limpieza.

```
Resumen estadístico general:
                Edad   IngresoAnual   CreditScore   Experiencia          Monto  \
count  20494.000000   20494.000000  20494.000000  20494.000000   20494.000000
mean      39.772519   56161.411925    541.464575     17.486581   23551.067044
std       11.282680   41443.240046    137.240448     10.955604   14226.083351
min       18.000000       0.000000      0.000000      0.000000       0.000000
25%       32.000000   29175.000000    531.000000     10.000000   14486.000000
50%       40.000000   46663.000000    575.000000     17.000000   21221.500000
75%       47.000000   72412.000000    607.000000     25.000000   30149.000000
max       80.000000  485341.000000    712.000000     61.000000  158686.000000

            Duración  NumberOfDependents  MonthlyDebtPayments  \
count  20494.000000        20494.000000         20494.000000
mean      51.098077            1.432907           432.206060
std       26.890628            1.388819           254.818149
min        0.000000            0.000000             0.000000
25%       36.000000            0.000000           269.000000
50%       48.000000            1.000000           391.000000
75%       60.000000            2.000000           552.000000
max      120.000000            5.000000          2919.000000

       CreditCardUtilizationRate  NumberOfOpenCreditLines  ...  \
count               20494.000000              20494.00000  ...
mean                    0.270727                  2.85425  ...
std                     0.168233                  1.82766  ...
min                     0.000000                  0.00000  ...
25%                     0.142023                  2.00000  ...
50%                     0.253469                  3.00000  ...
75%                     0.381994                  4.00000  ...
max                     0.917380                 13.00000  ...

       CheckingAccountBalance   TotalAssets  TotalLiabilities   IngresoMens  \
count            20494.000000  2.049400e+04      2.049400e+04  20494.000000
mean              1689.083195  9.179241e+04      3.414402e+04   4631.668842
std               2227.688244  1.201573e+05      4.694795e+04   3400.032117
min                  0.000000  0.000000e+00      0.000000e+00      0.000000
25%                484.000000  2.701450e+04      9.795250e+03   2410.375000
```

```
25%                484.000000  2.701450e+04      9.795250e+03   2410.375000
50%               1033.000000  5.609250e+04      2.067800e+04   3864.375000
75%               2046.000000  1.125320e+05      4.122225e+04   5980.062500
max              52572.000000  2.619627e+06      1.417302e+06  25000.000000

       UtilityBillsPaymentHistory     JobTenure       NetWorth  \
count                20494.000000  20494.000000  2.049400e+04
mean                     0.756071      4.719576  6.807870e+04
std                      0.216843      2.443620  1.139525e+05
min                      0.000000      0.000000  0.000000e+00
25%                      0.705927      3.000000  7.397750e+03
50%                      0.810912      5.000000  2.848650e+04
75%                      0.887961      6.000000  8.386575e+04
max                      0.999433     15.000000  2.603208e+06

       MonthlyLoanPayment  LoanApproved     RiskScore
count        20494.000000  20494.000000  20494.000000
mean           861.759970      0.238460     50.777047
std            683.615359      0.426152      7.789086
min              0.000000      0.000000     28.800000
25%            456.018655      0.000000     46.000000
50%            699.545027      0.000000     52.000000
75%           1082.140970      0.000000     56.000000
max          10892.629520      1.000000     84.000000

[8 rows x 27 columns]

Hay 103 filas duplicadas en el DataFrame.

Porcentaje de valores faltantes por columna:
Series([], dtype: float64)

Tipos de datos por columna:
FechaPrestamo                  object
Edad                            int64
IngresoAnual                  float64
CreditScore                     int64
```

```
CreditScore                    int64
Situ_Empleo                   object
Educación                     object
Experiencia                    int64
Monto                          int64
Duración                       int64
MaritalStatus                 object
NumberOfDependents             int64
HomeOwnershipStatus           object
MonthlyDebtPayments            int64
CreditCardUtilizationRate    float64
NumberOfOpenCreditLines        int64
NumberOfCreditInquiries        int64
DebtToIncomeRatio            float64
BankruptcyHistory            float64
LoanPurpose                   object
PreviousLoanDefaults           int64
PaymentHistory                 int64
LengthOfCreditHistory          int64
SavingsAccountBalance          int64
CheckingAccountBalance         int64
TotalAssets                    int64
TotalLiabilities               int64
IngresoMens                  float64
UtilityBillsPaymentHistory   float64
JobTenure                      int64
NetWorth                       int64
MonthlyLoanPayment           float64
LoanApproved                 float64
RiskScore                    float64
dtype: object
```

```
Primeras 5 filas del DataFrame limpio:
  FechaPrestamo  Edad  IngresoAnual  CreditScore Situ_Empleo    Educación  \
0    2018-01-01    45       39948.0          617  No_especif       Master
1    2018-01-02    38       39709.0          628    Empleado    Associate
2    2018-01-03    47       40724.0          570    Empleado     Bachelor
3    2045-05-10    58       69084.0          545    Empleado  High School
4    2018-01-05    37      103264.0          594    Empleado    Associate

   Experiencia  Monto  Duración MaritalStatus  ...  CheckingAccountBalance  \
0           17  13152        48        Casado  ...                    1202
1           15  26045        48       Soltero  ...                    3460
2           26  17627        36        Casado  ...                     895
3           34  37898        96       Soltero  ...                    1217
4           17   9184        36        Casado  ...                    4981

  TotalAssets  TotalLiabilities  IngresoMens  UtilityBillsPaymentHistory  \
0      146111             19183  3329.000000                    0.724972
1       53204              9595  3309.083333                    0.935132
2       25176            128874  3393.666667                    0.872241
3           0              5370     0.000000                    0.896155
4      244305             17286  8605.333333                    0.941369

   JobTenure  NetWorth  MonthlyLoanPayment LoanApproved  RiskScore
0         11    126928          419.805992          0.0       49.0
1          0     43609          794.054238          0.0       52.0
2          6      5205          666.406688          0.0       52.0
3          5     99452         1047.506980          0.0       54.0
4          5    227019          330.179140          1.0       36.0

[5 rows x 33 columns]
```

Tabla que muestre el porcentaje de valores faltantes final por columna.

```
Empty DataFrame
Columns: [Column, MissingPercentage]
Index: []
```

Comprobación de que no hay duplicados ni valores inválidos.

```
Hay 103 filas duplicadas.
No hay valores nulos.
```