In [41]:
```
import pandas as pd

raw_dec_benign = pd.read_csv('../data/raw/dec/decimal_benign.csv')
raw_dec_dos = pd.read_csv('../data/raw/dec/decimal_DoS.csv')
raw_dec_gas = pd.read_csv('../data/raw/dec/decimal_spoofing-GAS.csv')
raw_dec_rpm = pd.read_csv('../data/raw/dec/decimal_spoofing-RPM.csv')
raw_dec_speed = pd.read_csv('../data/raw/dec/decimal_spoofing-SPEED.csv')
raw_dec_wheel = pd.read_csv('../data/raw/dec/decimal_spoofing-STEERING_WH
```

In [55]:
```
import matplotlib.pyplot as plt
import seaborn as sns

def eda(df, nombre, id='ID'):
    print(f'EDA de {nombre}\n')
    # 1. Display first few rows
    print('Primeras filas:')
    print(df.head(), "\n")
    # 2. General info about dataset
    print(df.info(), "\n")
    # 3. Check duplicates
    num_duplicates = df.duplicated().sum()
    print(f'Filas enteras repetidas: {num_duplicates}\n')
    # 4. Check missing values
    print('Valores ausentes:')
    print(df.isna().sum(), "\n")
    # 5. Check unique CAN IDs
    print(f'Cantidad de CAN IDs únicos: {df[id].nunique()}')
    print('Top 10 CAN IDs más frecuentes:')
    print(df[id].value_counts().head(10), "\n")
    # 6. Descriptive statistics for numeric columns
    print('Estadísticas descriptivas:')
    print(df.describe(), "\n")
    # 7. Check for outliers in the DATA fields
    print('Posibles valores atípicos en los bytes de datos:')
    data_cols = [col for col in df.columns if col.startswith("DATA_")]
    print(df[data_cols].describe(percentiles=[0.01, 0.25, 0.75, 0.99]), "
    # 8. Class distribution
    if 'label' in df.columns:
        print('Distribución de etiquetas:')
        print(df['label'].value_counts(), "\n")

    if 'category' in df.columns:
        print('Distribución de categorías:')
        print(df['category'].value_counts(), "\n")

    if 'specific_class' in df.columns:
        print('Distribución de clases específicas:')
        print(df['specific_class'].value_counts(), "\n")
    # 9. Byte Correlation Analysis
    if data_cols:
        correlation_matrix = df[data_cols].corr()
        plt.figure(figsize=(8, 6))
        sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=
        plt.title("Correlación entre bytes de datos")
        plt.show()
    # 10. Visual Insights
    df[id].value_counts().head(10).plot(kind='bar', title='Top 10 CAN IDs
    plt.xlabel("CAN ID")
```

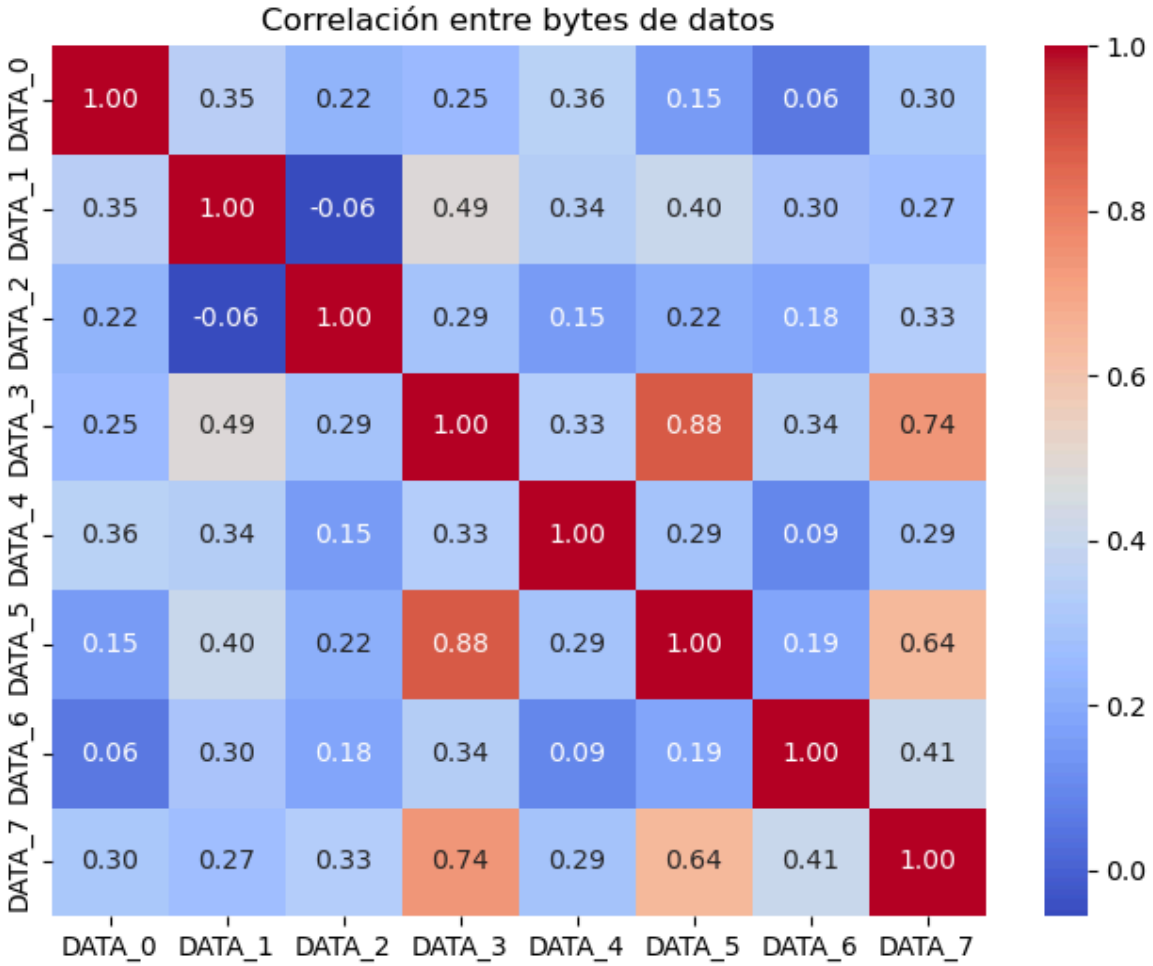
```
    plt.ylabel("Frecuencia")
    plt.show()

    for col in data_cols:
        plt.figure()
        df[col].hist(bins=50)
        plt.title(f'Distribución de valores en {col}')
        plt.xlabel('Valor del byte')
        plt.ylabel('Frecuencia')
        plt.show()
    # 11. Anomaly Detection
    out_of_range = df[data_cols].apply(lambda x: ((x < 0) | (x > 255)).su
    print("Valores fuera de rango (esperado 0-255):")
    print(out_of_range, "\n")
```

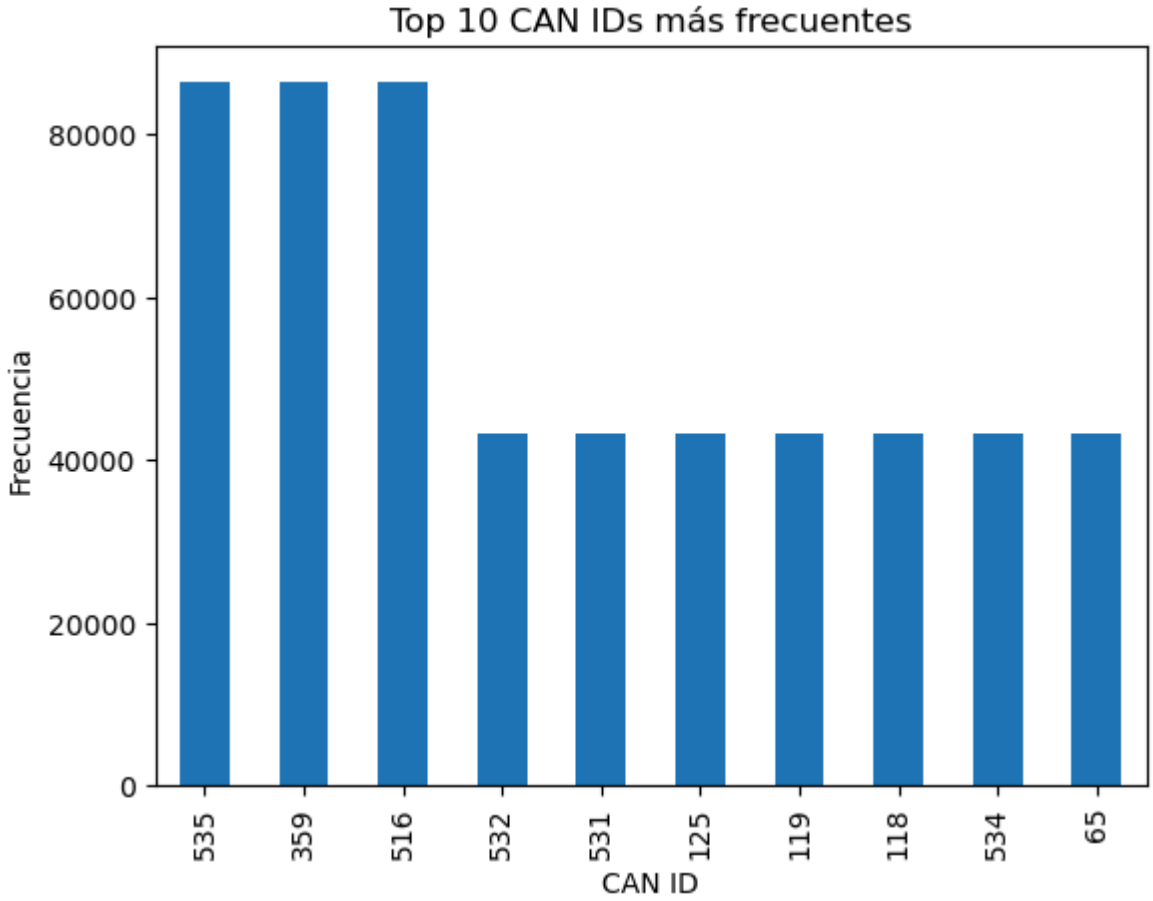
In [53]:
```
eda(raw_dec_benign, 'dec_benign')
```

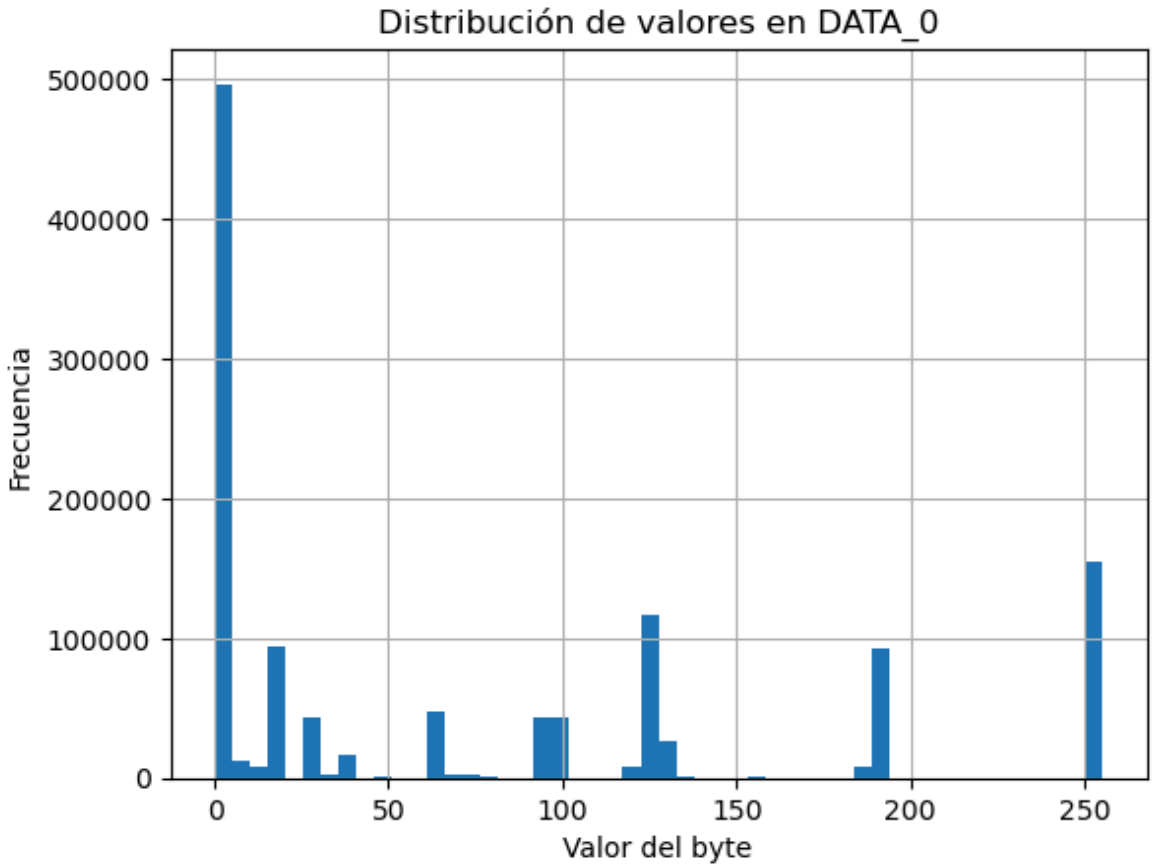
```
EDA de dec_benign

Primeras filas:
     ID  DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5  DATA_6  DATA_7  \
0    65      96       0       0       0       0       0       0       0
1  1068     132      13     160       0       0       0       0       0
2   535     127     255     127     255     127     255     127     255
3   131      15     224       0       0       0       0       0       0
4   936       1       0      39      16       0       0       0       0

    label category specific_class
0  BENIGN   BENIGN         BENIGN
1  BENIGN   BENIGN         BENIGN
2  BENIGN   BENIGN         BENIGN
3  BENIGN   BENIGN         BENIGN
4  BENIGN   BENIGN         BENIGN

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1223737 entries, 0 to 1223736
Data columns (total 12 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   ID              1223737 non-null  int64
 1   DATA_0          1223737 non-null  int64
 2   DATA_1          1223737 non-null  int64
 3   DATA_2          1223737 non-null  int64
 4   DATA_3          1223737 non-null  int64
 5   DATA_4          1223737 non-null  int64
 6   DATA_5          1223737 non-null  int64
 7   DATA_6          1223737 non-null  int64
 8   DATA_7          1223737 non-null  int64
 9   label           1223737 non-null  object
 10  category        1223737 non-null  object
 11  specific_class  1223737 non-null  object
dtypes: int64(9), object(3)
memory usage: 112.0+ MB
None

Filas enteras repetidas: 1220190

Valores ausentes:
ID                0
DATA_0            0
DATA_1            0
DATA_2            0
DATA_3            0
DATA_4            0
DATA_5            0
DATA_6            0
DATA_7            0
label             0
category          0
specific_class    0
dtype: int64

Cantidad de CAN IDs únicos: 72
Top 10 CAN IDs más frecuentes:
ID
535    86385
359    86383
```

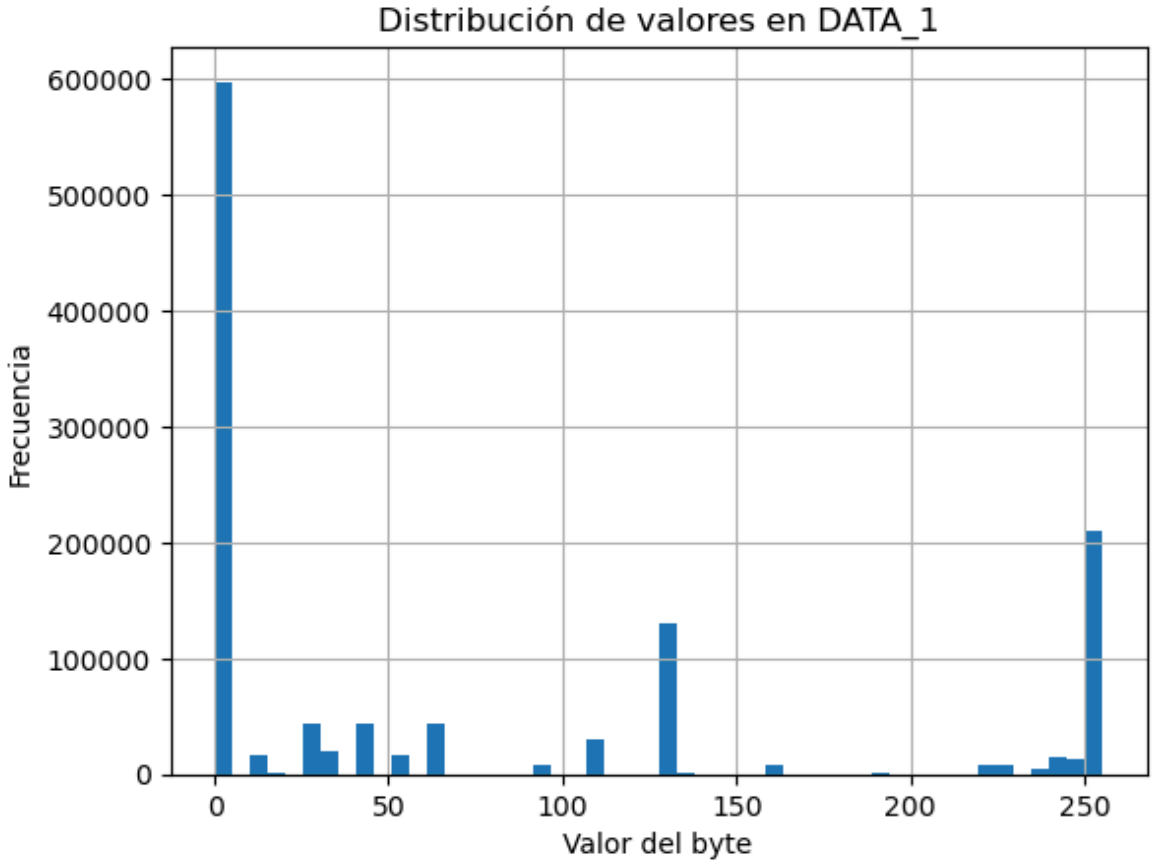
```
516    86383
532    43194
531    43194
125    43194
119    43194
118    43193
534    43193
65     43193
Name: count, dtype: int64

Estadísticas descriptivas:
                ID        DATA_0        DATA_1        DATA_2        DATA_
3  \
count  1.223737e+06  1.223737e+06  1.223737e+06  1.223737e+06  1.223737e+0
6
mean   5.645002e+02  7.691670e+01  7.835954e+01  5.930063e+01  6.111469e+0
1
std    3.343070e+02  9.126328e+01  9.970725e+01  7.371815e+01  9.427600e+0
1
min    6.500000e+01  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+0
0
25%    3.590000e+02  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+0
0
50%    5.310000e+02  2.800000e+01  1.300000e+01  1.600000e+01  0.000000e+0
0
75%    9.360000e+02  1.270000e+02  1.280000e+02  1.270000e+02  1.120000e+0
2
max    1.438000e+03  2.550000e+02  2.550000e+02  2.550000e+02  2.550000e+0
2

            DATA_4        DATA_5        DATA_6        DATA_7
count  1.223737e+06  1.223737e+06  1.223737e+06  1.223737e+06
mean   4.921648e+01  6.038979e+01  7.776605e+01  6.611918e+01
std    6.656173e+01  9.941050e+01  1.057676e+02  1.052630e+02
min    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%    1.500000e+01  0.000000e+00  0.000000e+00  0.000000e+00
75%    1.130000e+02  7.900000e+01  2.270000e+02  1.280000e+02
max    2.550000e+02  2.550000e+02  2.550000e+02  2.550000e+02

Posibles valores atípicos en los bytes de datos:
            DATA_0        DATA_1        DATA_2        DATA_3        DATA_
4  \
count  1.223737e+06  1.223737e+06  1.223737e+06  1.223737e+06  1.223737e+0
6
mean   7.691670e+01  7.835954e+01  5.930063e+01  6.111469e+01  4.921648e+0
1
std    9.126328e+01  9.970725e+01  7.371815e+01  9.427600e+01  6.656173e+0
1
min    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+0
0
1%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+0
0
25%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+0
0
50%    2.800000e+01  1.300000e+01  1.600000e+01  0.000000e+00  1.500000e+0
1
75%    1.270000e+02  1.280000e+02  1.270000e+02  1.120000e+02  1.130000e+0
2
99%    2.550000e+02  2.550000e+02  2.550000e+02  2.550000e+02  2.400000e+0
```

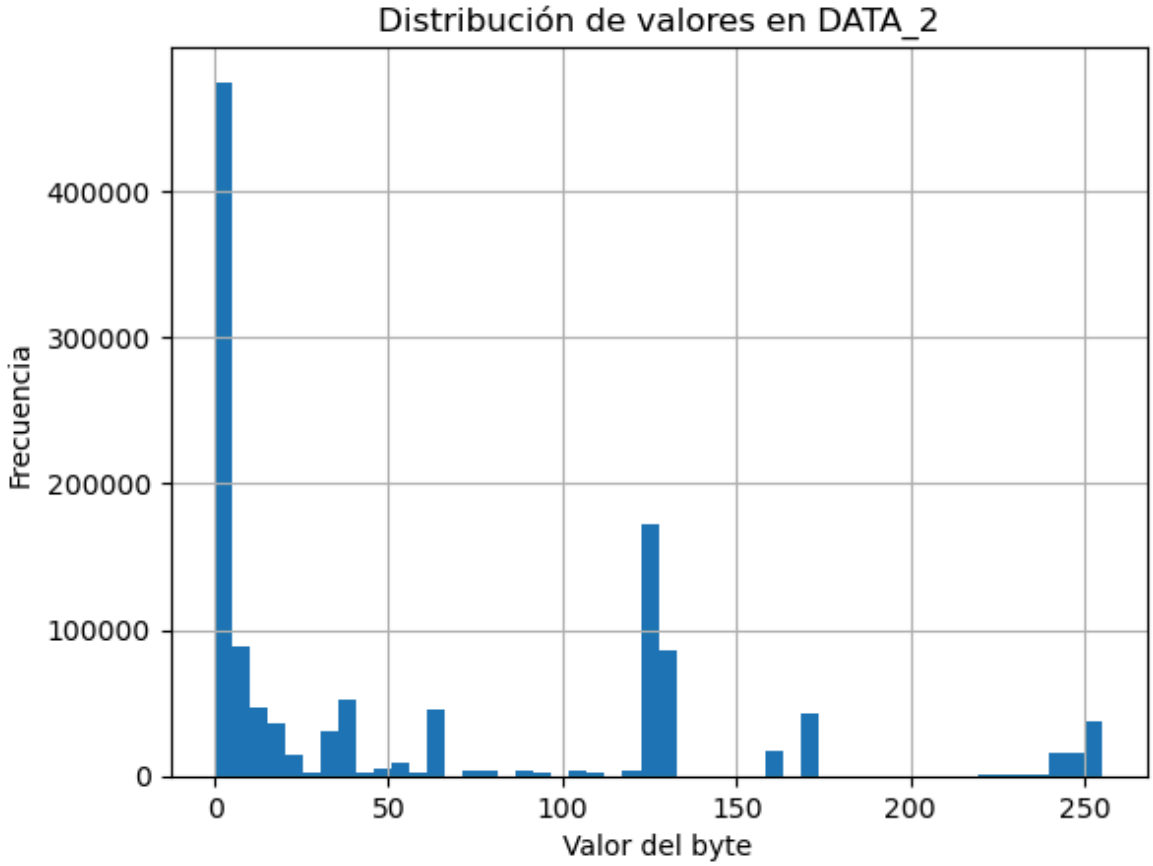
```
2
max    2.550000e+02  2.550000e+02  2.550000e+02  2.550000e+02  2.550000e+0
2

              DATA_5        DATA_6        DATA_7
count   1.223737e+06  1.223737e+06  1.223737e+06
mean    6.038979e+01  7.776605e+01  6.611918e+01
std     9.941050e+01  1.057676e+02  1.052630e+02
min     0.000000e+00  0.000000e+00  0.000000e+00
1%      0.000000e+00  0.000000e+00  0.000000e+00
25%     0.000000e+00  0.000000e+00  0.000000e+00
50%     0.000000e+00  0.000000e+00  0.000000e+00
75%     7.900000e+01  2.270000e+02  1.280000e+02
99%     2.550000e+02  2.550000e+02  2.550000e+02
max     2.550000e+02  2.550000e+02  2.550000e+02

Distribución de etiquetas:
label
BENIGN    1223737
Name: count, dtype: int64

Distribución de categorías:
category
BENIGN    1223737
Name: count, dtype: int64

Distribución de clases específicas:
specific_class
BENIGN    1223737
Name: count, dtype: int64
```

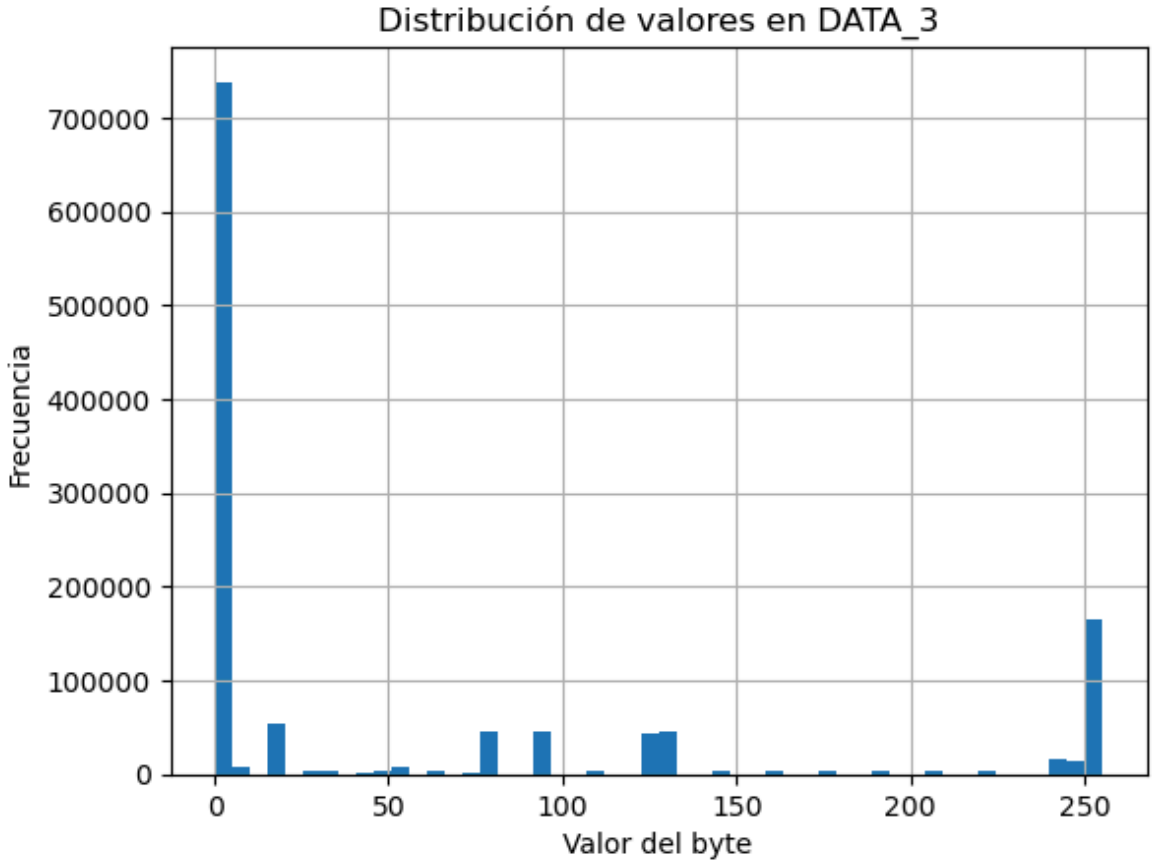


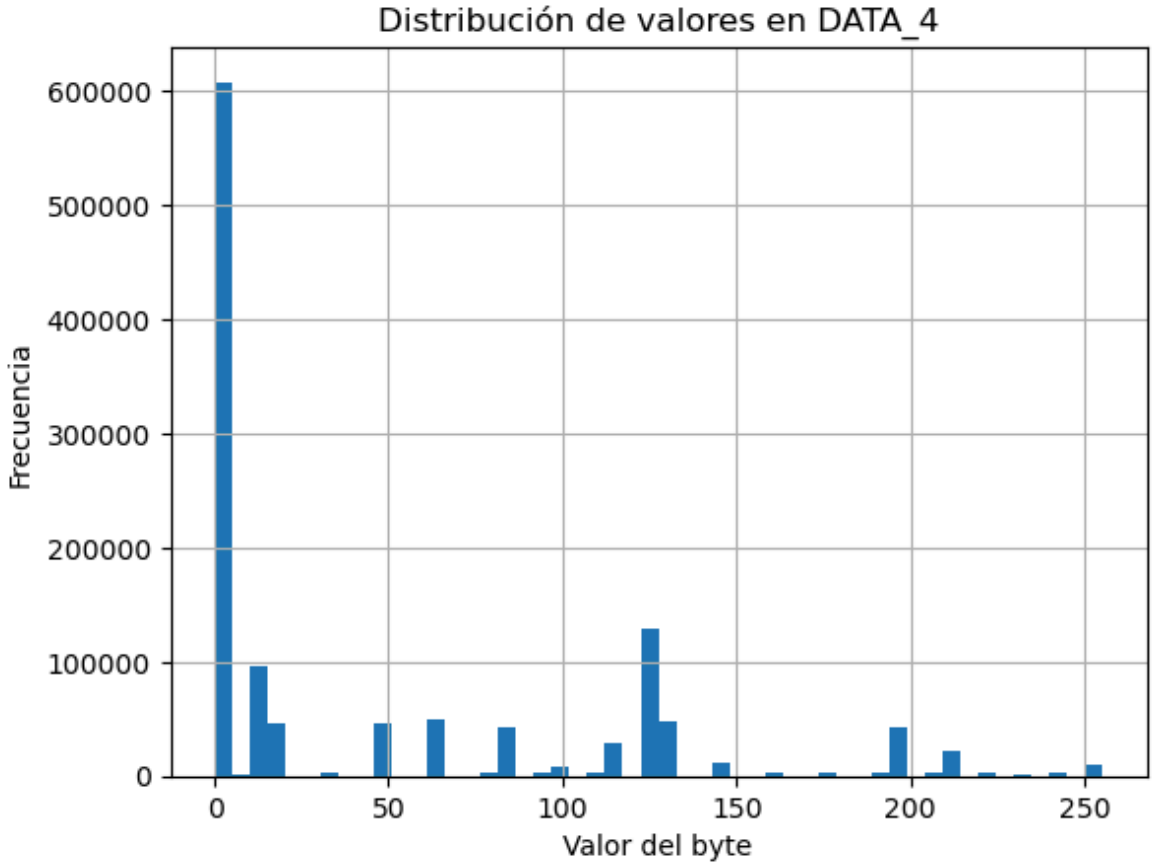

Top 10 CAN IDs más frecuentes
Frecuencia
0
20000
40000
60000
80000
535
359
516
532
531
125
119
118
534
65
CAN ID

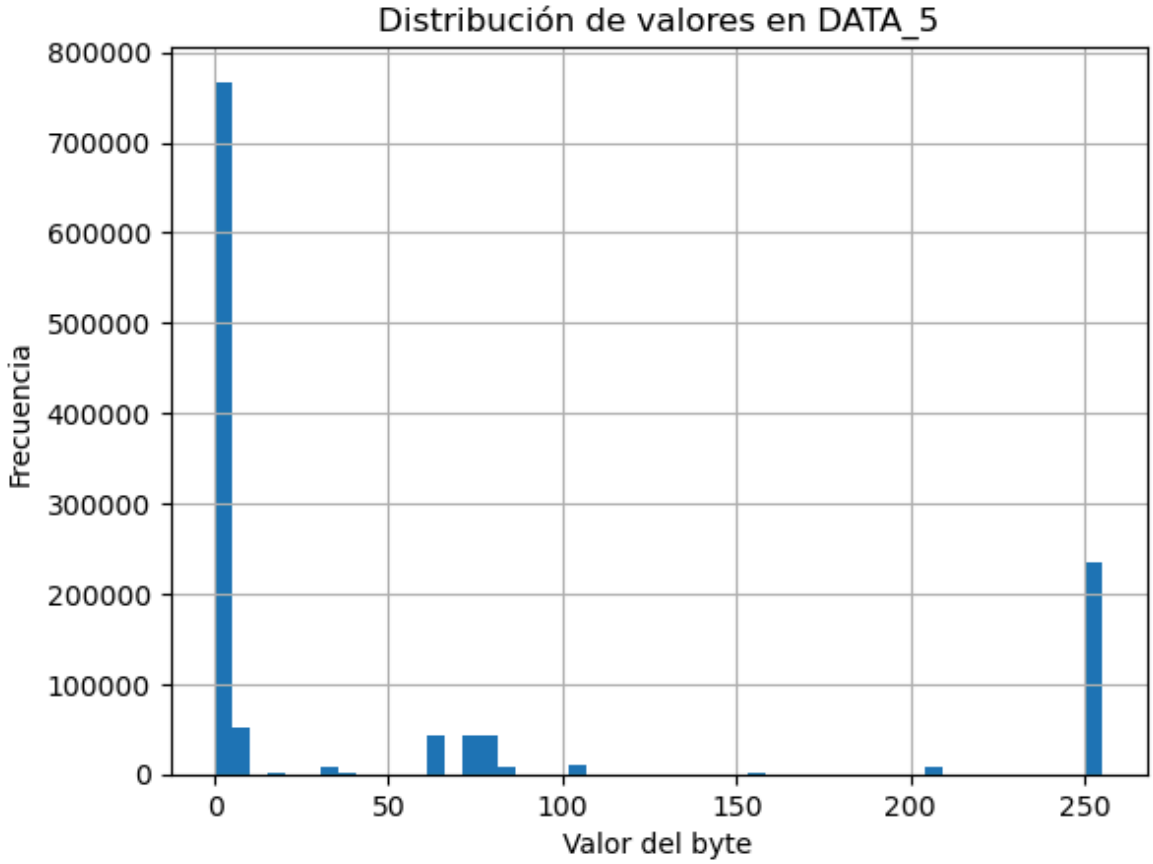

Distribución de valores en DATA_0
Frecuencia
0
100000
200000
300000
400000
500000
0
50
100
150
200
250
Valor del byte

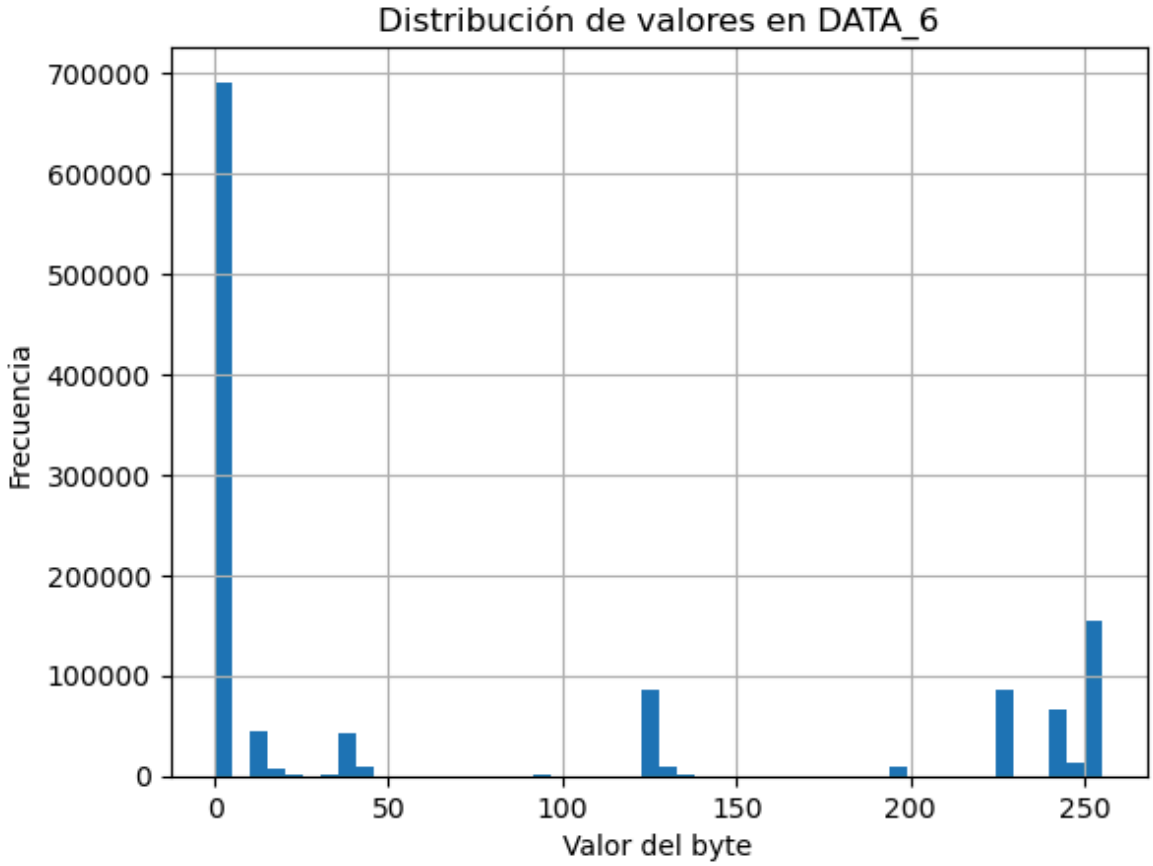

Distribución de valores en DATA_1
Frecuencia
600000
500000
400000
300000
200000
100000
0
0
50
100
150
200
250
Valor del byte

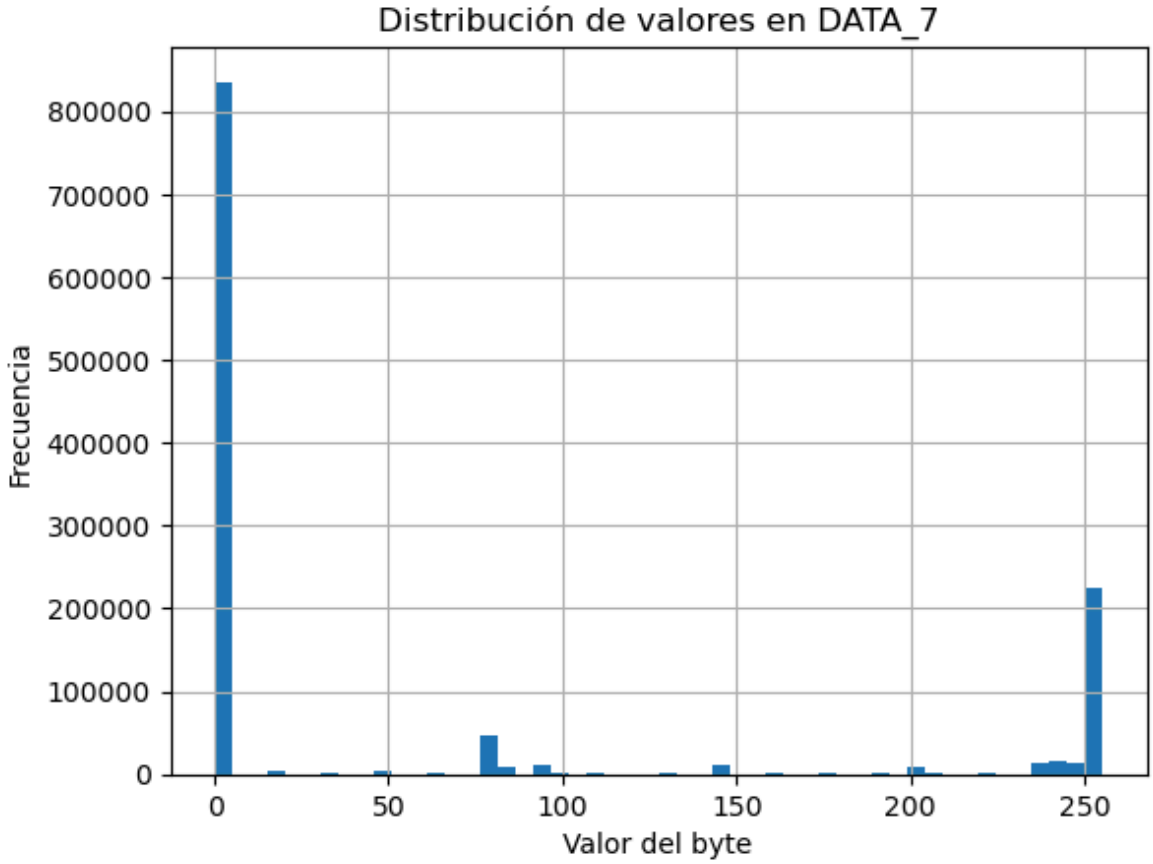

Distribución de valores en DATA_2
Frecuencia
400000
300000
200000
100000
0
0
50
100
150
200
250
Valor del byte


Distribución de valores en DATA_3
Frecuencia
Valor del byte
0
100000
200000
300000
400000
500000
600000
700000
0
50
100
150
200
250


Distribución de valores en DATA_4
Frecuencia
Valor del byte
0
100000
200000
300000
400000
500000
600000
0
50
100
150
200
250


Distribución de valores en DATA_5
Frecuencia
Valor del byte
0
100000
200000
300000
400000
500000
600000
700000
800000
0
50
100
150
200
250


Distribución de valores en DATA_6
Frecuencia
Valor del byte
0
100000
200000
300000
400000
500000
600000
700000
0
50
100
150
200
250



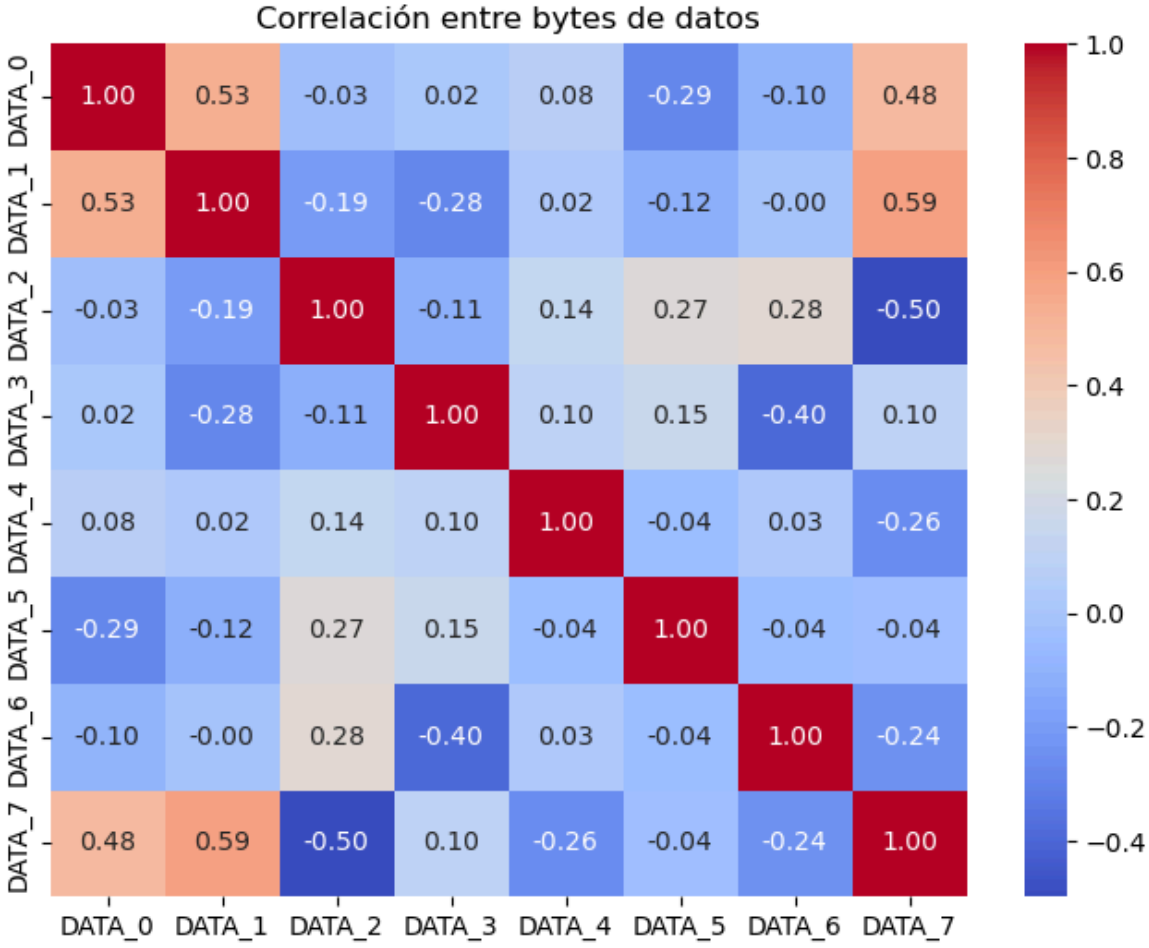

```
Valores fuera de rango (esperado 0–255):
DATA_0    0
DATA_1    0
DATA_2    0
DATA_3    0
DATA_4    0
DATA_5    0
DATA_6    0
DATA_7    0
dtype: int64
```

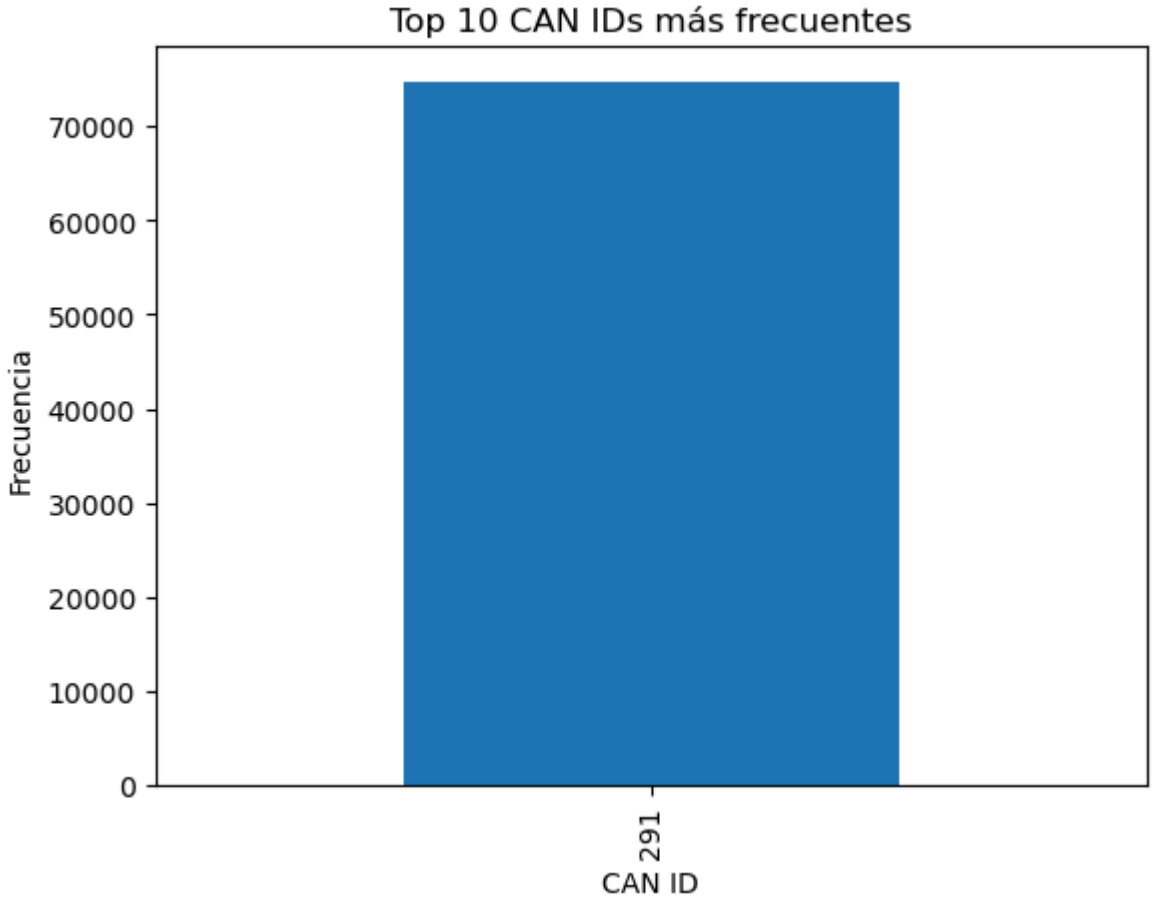
In [57]:
```
eda(raw_dec_dos, 'dec_dos')
```

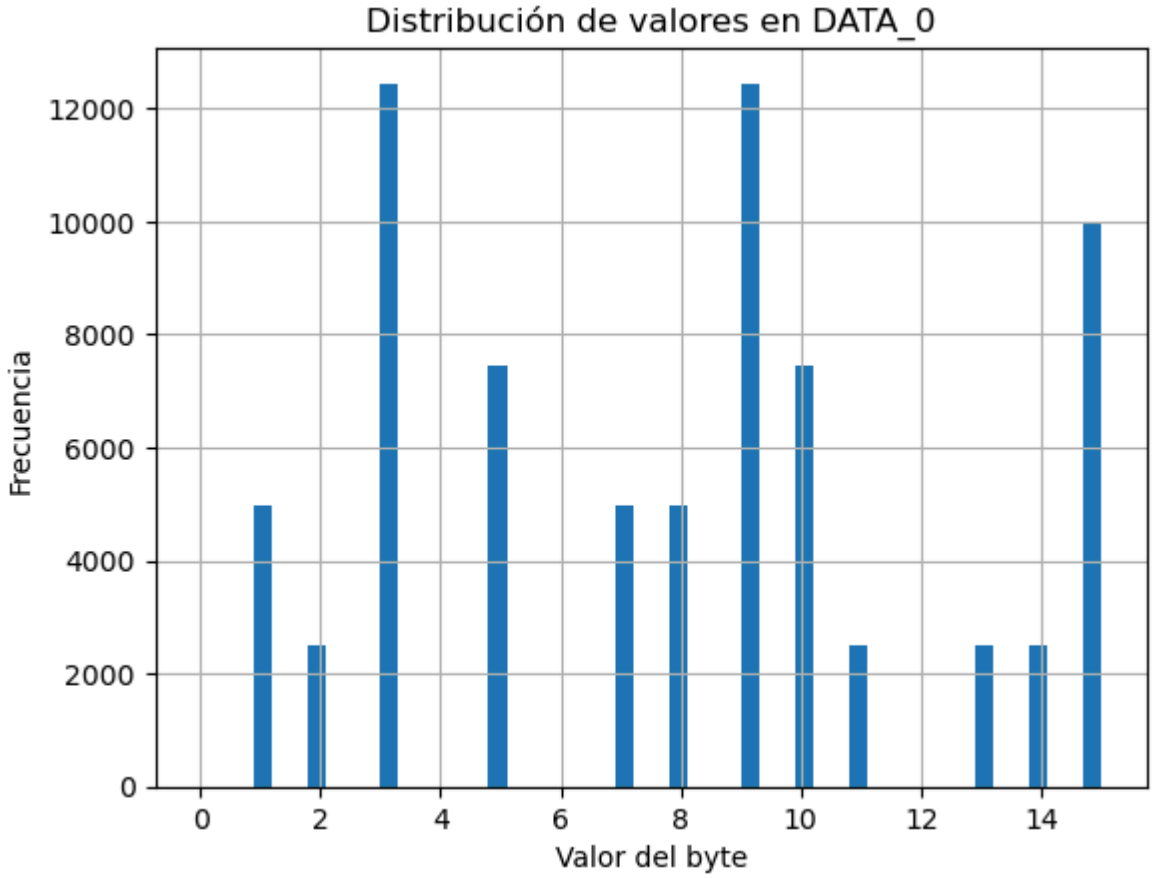
```
EDA de dec_dos

Primeras filas:
    ID  DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5  DATA_6  DATA_7  \
0  291       0       0       0       0       0       0       0       0
1  291      14      11       4       4       3       3       8      12
2  291      14      11       4       4       3       3       8      12
3  291      14      11       4       4       3       3       8      12
4  291      14      11       4       4       3       3       8      12

    label category specific_class
0  ATTACK      DoS            DoS
1  ATTACK      DoS            DoS
2  ATTACK      DoS            DoS
3  ATTACK      DoS            DoS
4  ATTACK      DoS            DoS

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74663 entries, 0 to 74662
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ID              74663 non-null  int64
 1   DATA_0          74663 non-null  int64
 2   DATA_1          74663 non-null  int64
 3   DATA_2          74663 non-null  int64
 4   DATA_3          74663 non-null  int64
 5   DATA_4          74663 non-null  int64
 6   DATA_5          74663 non-null  int64
 7   DATA_6          74663 non-null  int64
 8   DATA_7          74663 non-null  int64
 9   label           74663 non-null  object
 10  category        74663 non-null  object
 11  specific_class  74663 non-null  object
dtypes: int64(9), object(3)
memory usage: 6.8+ MB
None

Filas enteras repetidas: 74642

Valores ausentes:
ID                0
DATA_0            0
DATA_1            0
DATA_2            0
DATA_3            0
DATA_4            0
DATA_5            0
DATA_6            0
DATA_7            0
label             0
category          0
specific_class    0
dtype: int64

Cantidad de CAN IDs únicos: 1
Top 10 CAN IDs más frecuentes:
ID
291    74663
Name: count, dtype: int64
```

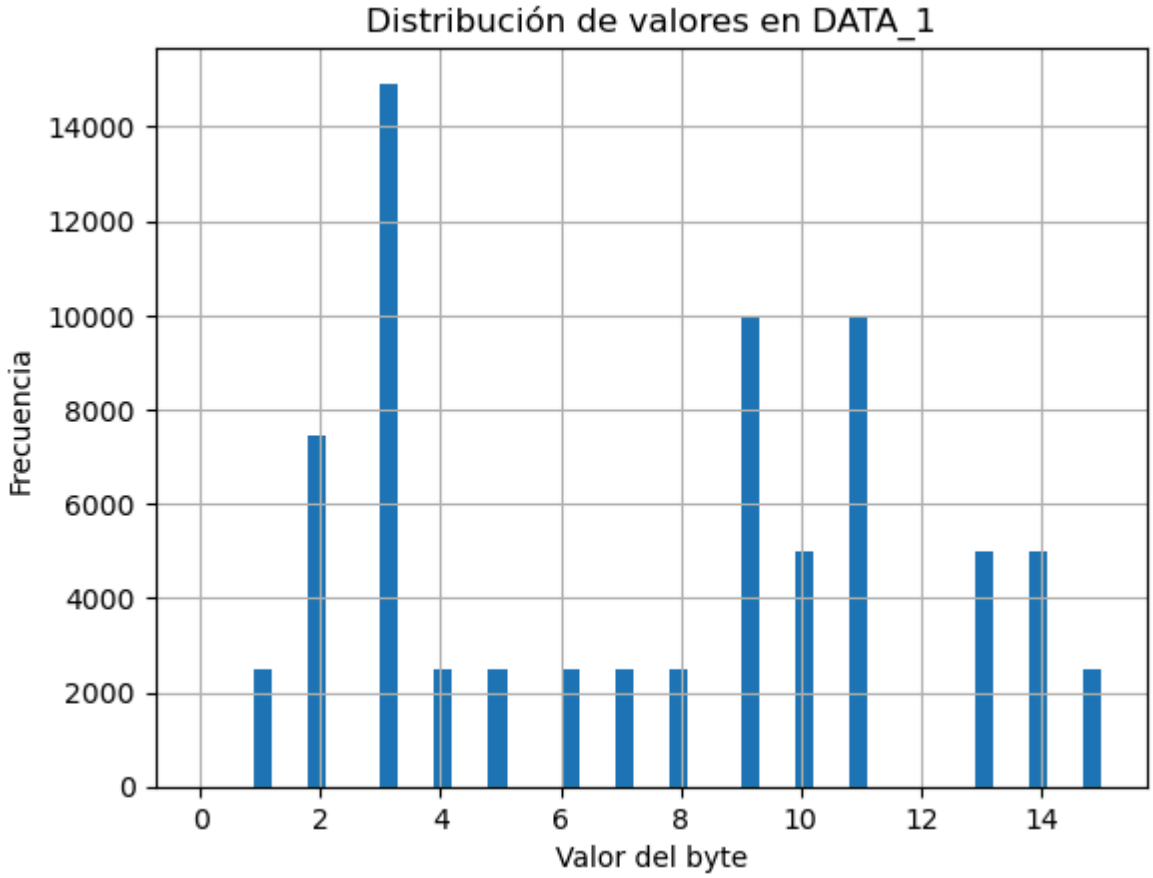
```
Estadísticas descriptivas:
            ID        DATA_0        DATA_1        DATA_2        DATA_3  \
count  74663.0  74663.000000  74663.000000  74663.000000  74663.000000
mean     291.0      7.898973      7.465800      6.864083      7.765627
std        0.0      4.363179      4.271956      4.558480      3.740096
min      291.0      0.000000      0.000000      0.000000      0.000000
25%      291.0      3.000000      3.000000      2.000000      5.000000
50%      291.0      8.000000      8.000000      8.000000      7.000000
75%      291.0     10.000000     11.000000     11.000000     11.000000
max      291.0     15.000000     15.000000     13.000000     14.000000

             DATA_4        DATA_5        DATA_6        DATA_7
count  74663.000000  74663.000000  74663.000000  74663.000000
mean       8.298863      6.233770      7.666729      7.267522
std        3.743825      5.364923      4.707330      4.660512
min        0.000000      0.000000      0.000000      0.000000
25%        6.000000      1.000000      2.000000      4.000000
50%        8.000000      3.000000      8.000000      8.000000
75%       11.000000     11.000000     11.000000     11.000000
max       14.000000     15.000000     15.000000     15.000000

Posibles valores atípicos en los bytes de datos:
             DATA_0        DATA_1        DATA_2        DATA_3        DATA_
4  \
count  74663.000000  74663.000000  74663.000000  74663.000000  74663.00000
0
mean       7.898973      7.465800      6.864083      7.765627      8.29886
3
std        4.363179      4.271956      4.558480      3.740096      3.74382
5
min        0.000000      0.000000      0.000000      0.000000      0.00000
0
1%         1.000000      1.000000      0.000000      0.000000      0.00000
0
25%        3.000000      3.000000      2.000000      5.000000      6.00000
0
50%        8.000000      8.000000      8.000000      7.000000      8.00000
0
75%       10.000000     11.000000     11.000000     11.000000     11.00000
0
99%       15.000000     15.000000     13.000000     14.000000     14.00000
0
max       15.000000     15.000000     13.000000     14.000000     14.00000
0

             DATA_5        DATA_6        DATA_7
count  74663.000000  74663.000000  74663.000000
mean       6.233770      7.666729      7.267522
std        5.364923      4.707330      4.660512
min        0.000000      0.000000      0.000000
1%         0.000000      0.000000      0.000000
25%        1.000000      2.000000      4.000000
50%        3.000000      8.000000      8.000000
75%       11.000000     11.000000     11.000000
99%       15.000000     15.000000     15.000000
max       15.000000     15.000000     15.000000

Distribución de etiquetas:
label
```

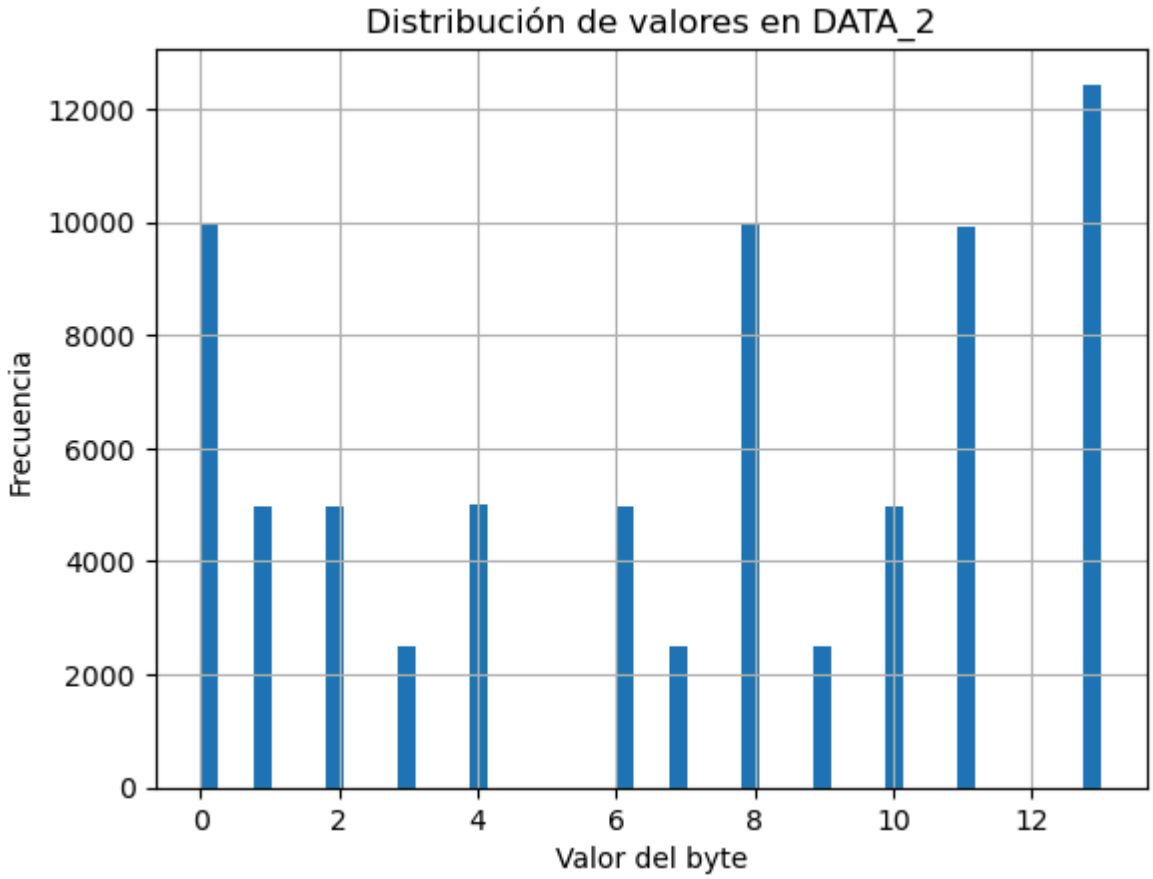
```
ATTACK    74663
Name: count, dtype: int64

Distribución de categorías:
category
DoS    74663
Name: count, dtype: int64

Distribución de clases específicas:
specific_class
DoS    74663
Name: count, dtype: int64
```

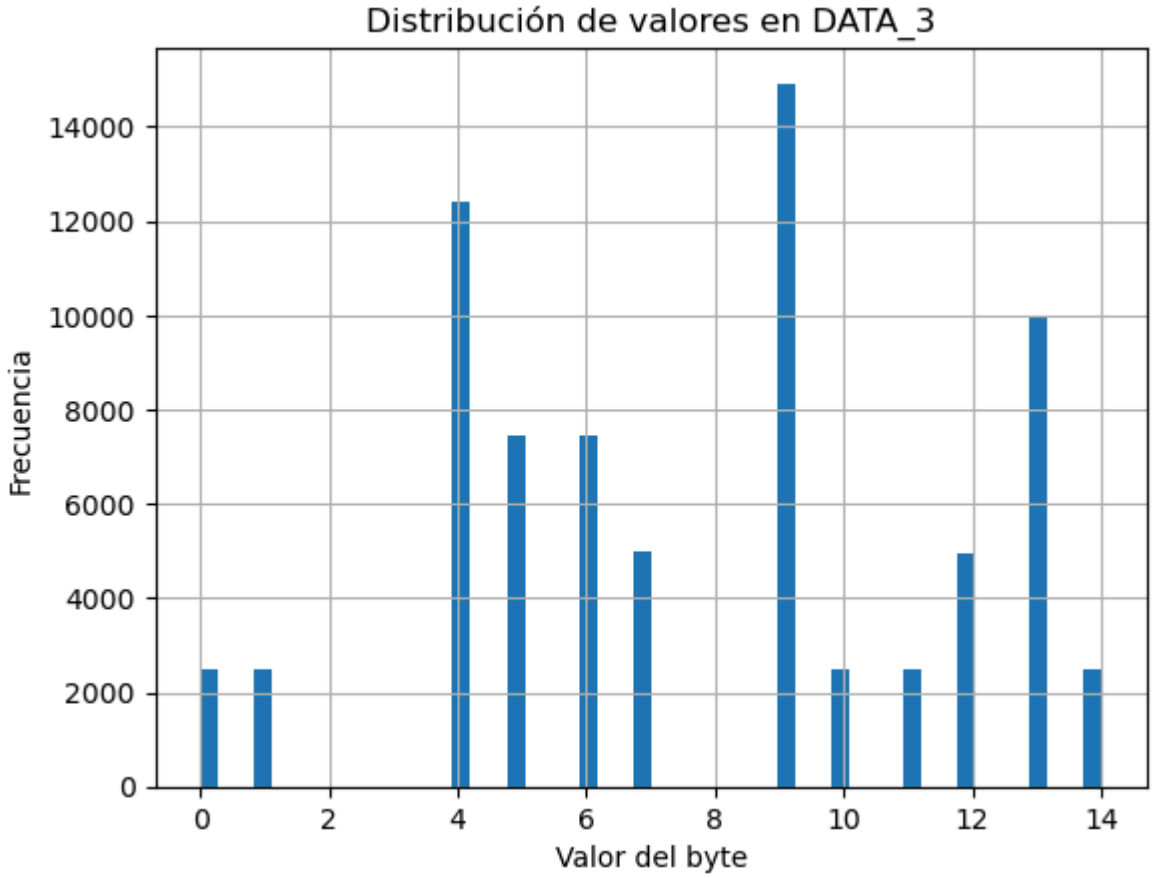


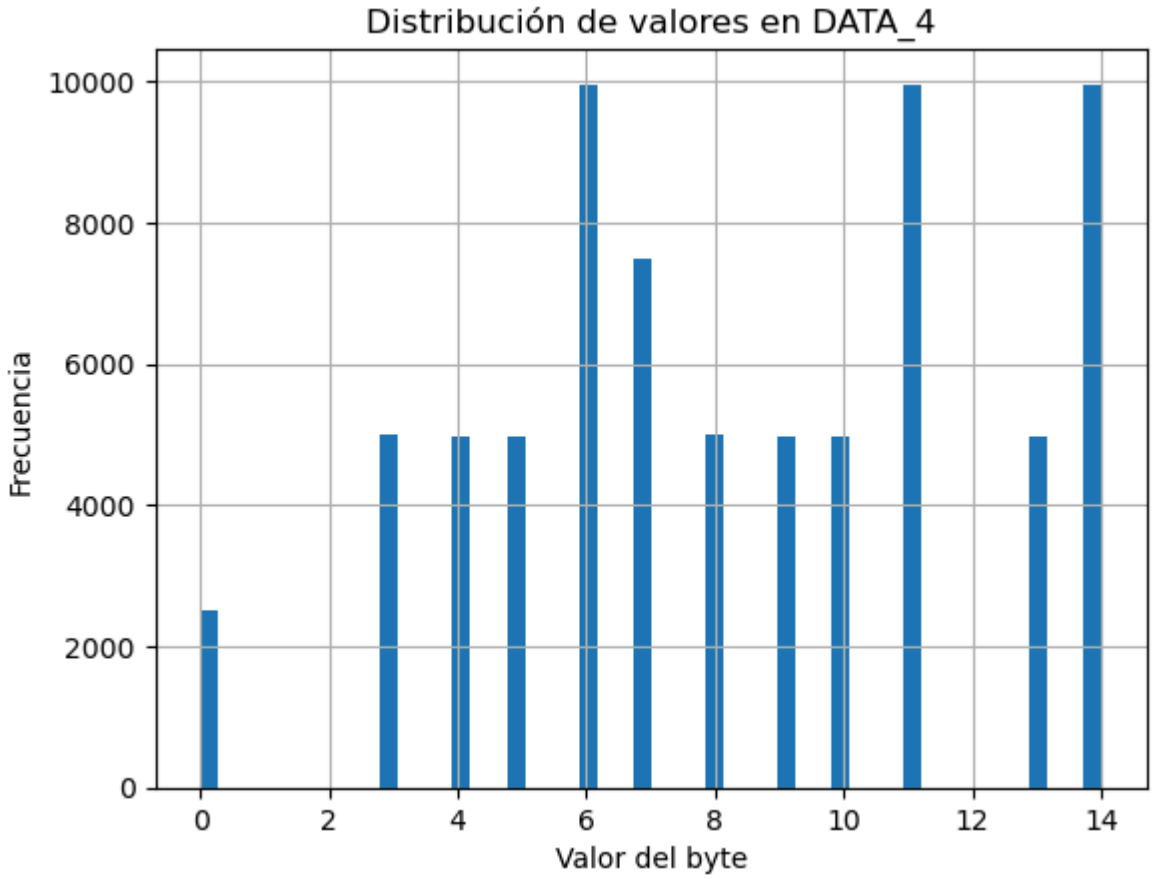

Top 10 CAN IDs más frecuentes
Frecuencia
70000
60000
50000
40000
30000
20000
10000
0
291
CAN ID

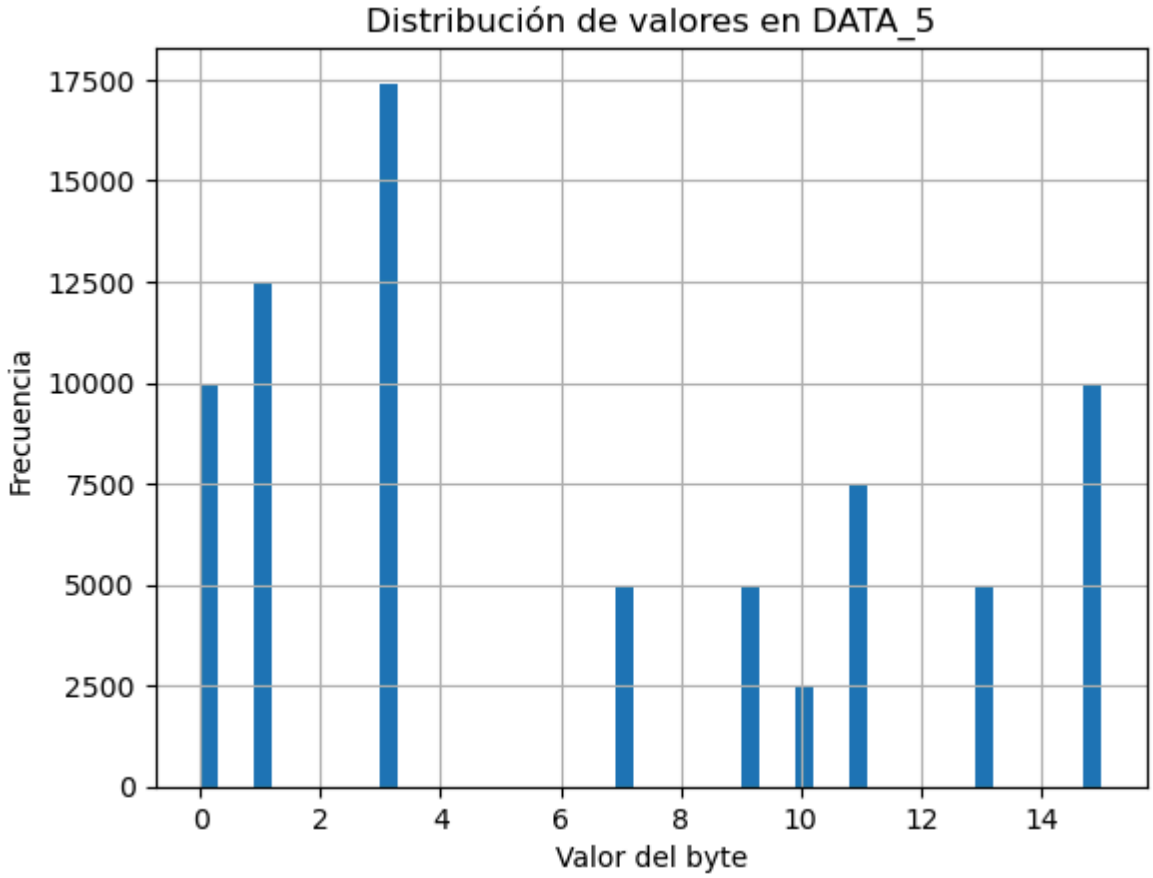

Distribución de valores en DATA_0
Frecuencia
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12
14
Valor del byte

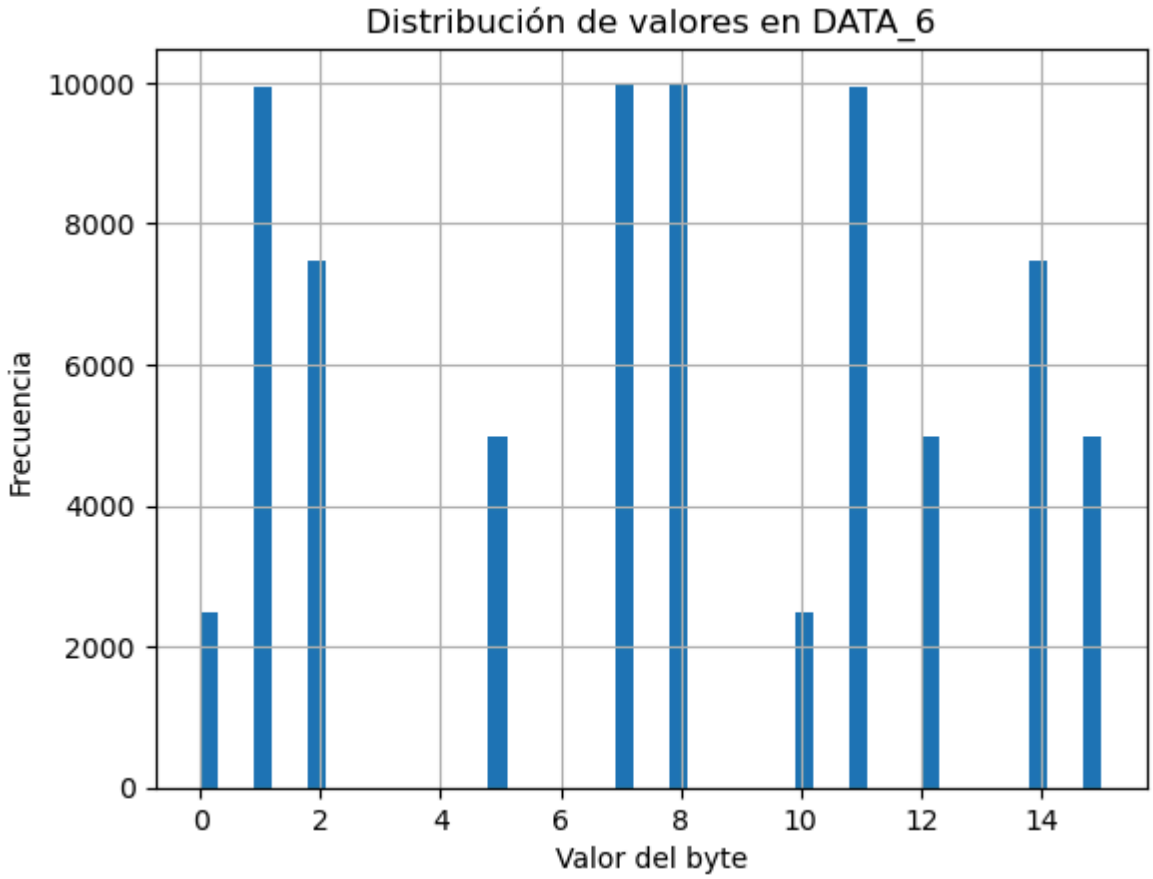

Distribución de valores en DATA_1
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
12000
14000
0
2
4
6
8
10
12
14


Distribución de valores en DATA_2
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
12000
0
2
4
6
8
10
12


Distribución de valores en DATA_3
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
12000
14000
0
2
4
6
8
10
12
14


Distribución de valores en DATA_4
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
0
2
4
6
8
10
12
14


Distribución de valores en DATA_5
Frecuencia
Valor del byte
0
2500
5000
7500
10000
12500
15000
17500
0
2
4
6
8
10
12
14


Distribución de valores en DATA_6
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
0
2
4
6
8
10
12
14

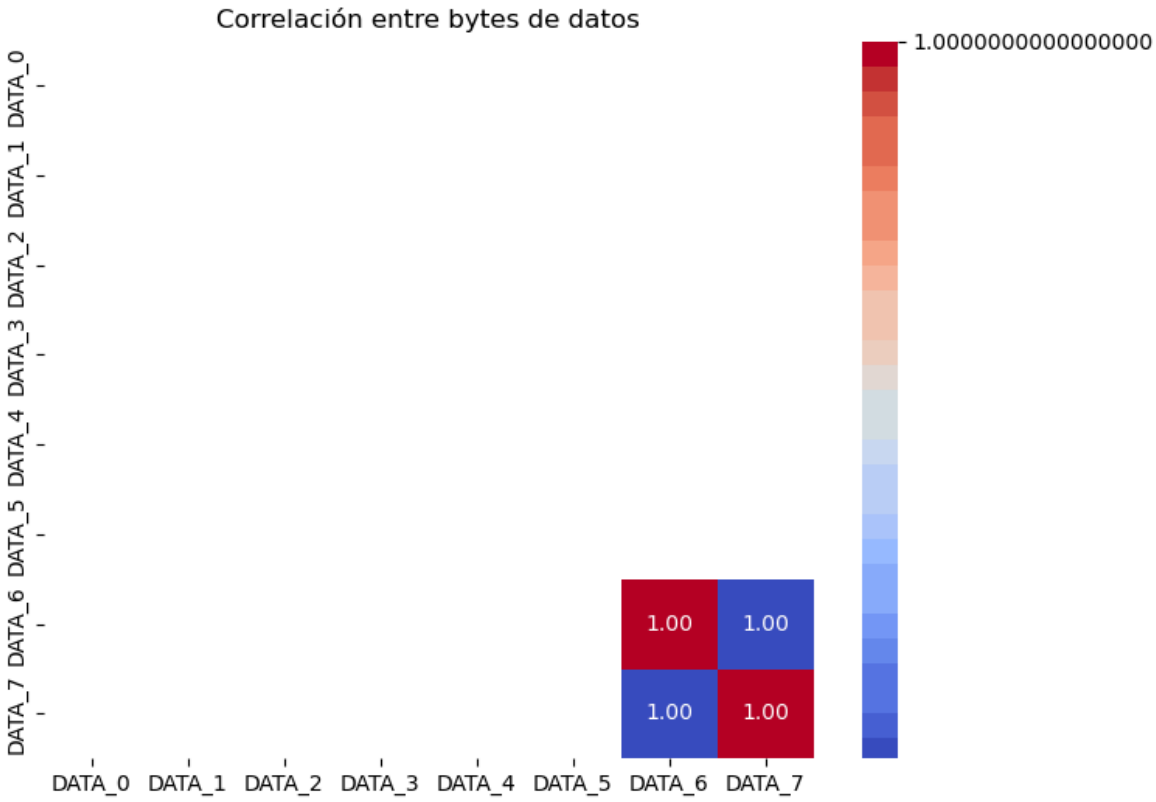

```
Valores fuera de rango (esperado 0–255):
DATA_0    0
DATA_1    0
DATA_2    0
DATA_3    0
DATA_4    0
DATA_5    0
DATA_6    0
DATA_7    0
dtype: int64
```

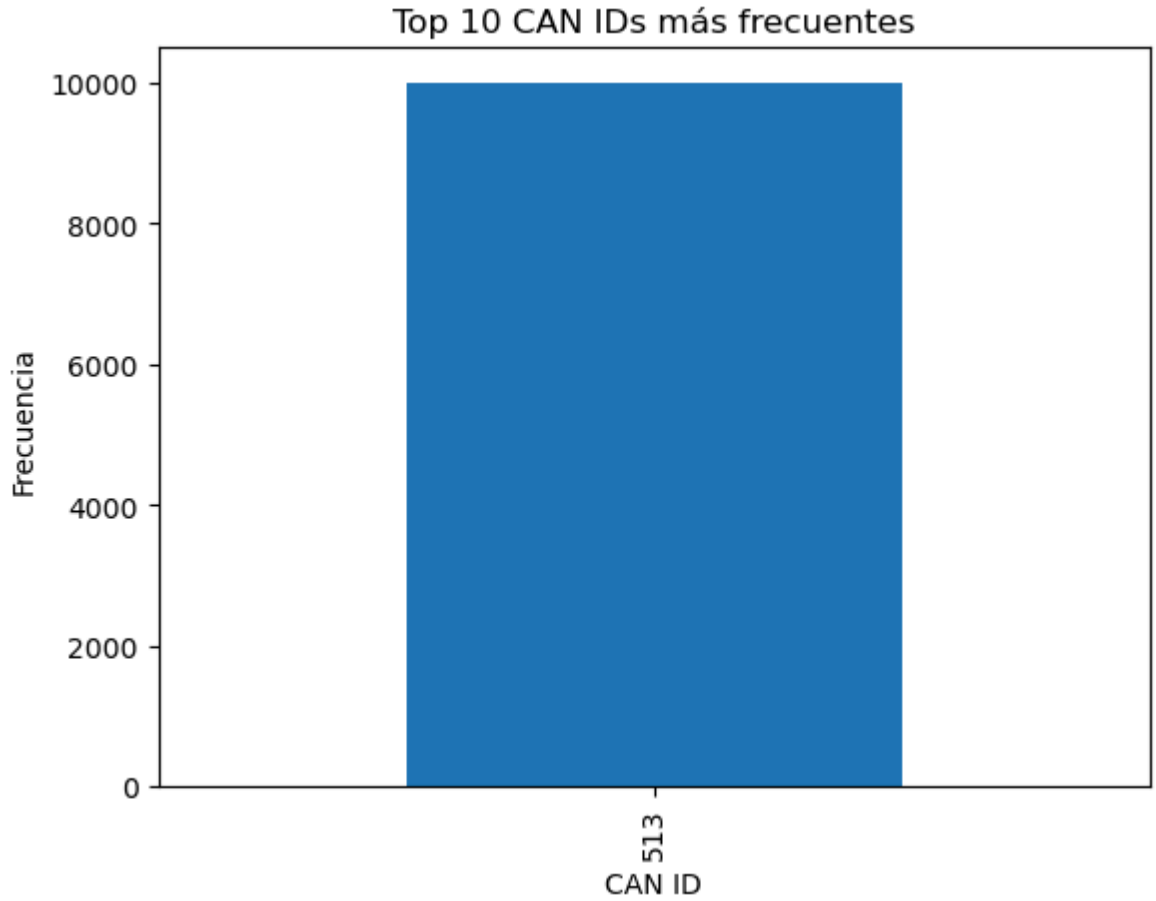
In [59]:
```
eda(raw_dec_gas, 'dec_gas')
```

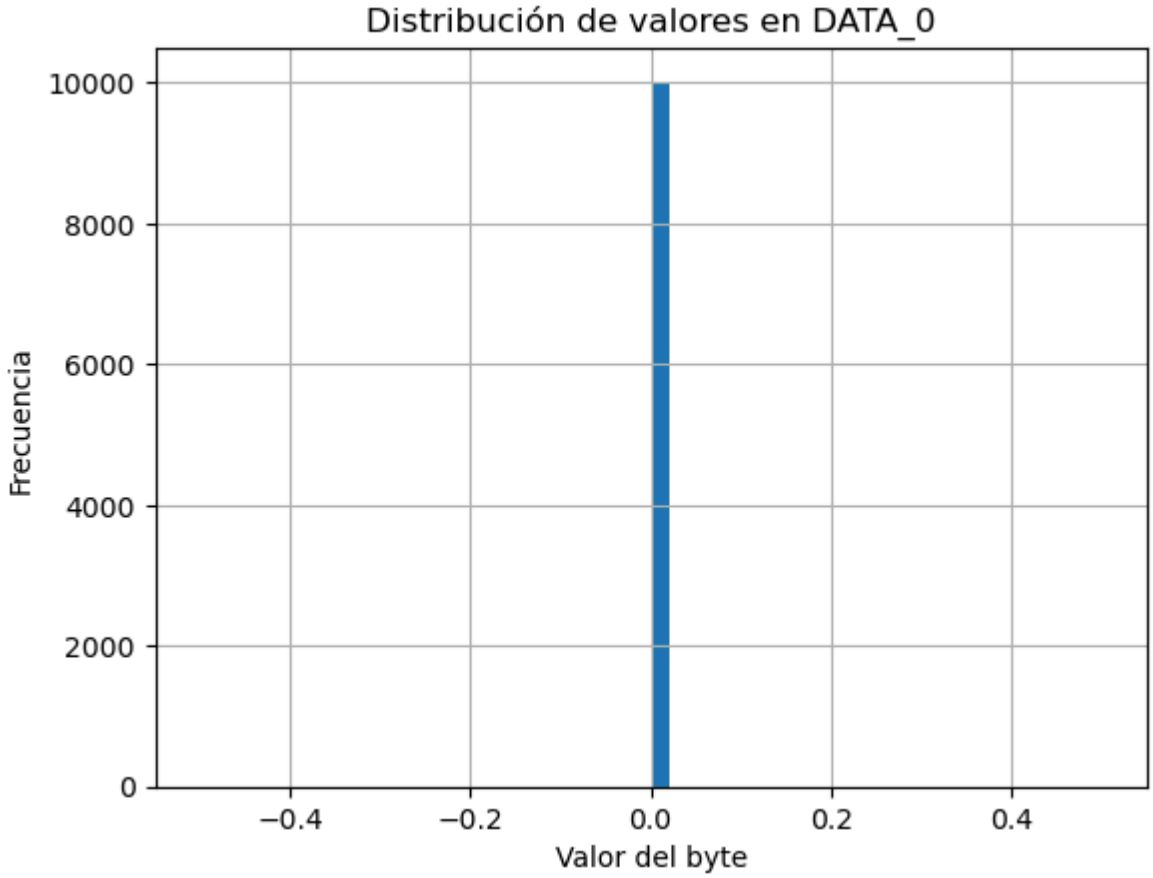
```
EDA de dec_gas

Primeras filas:
    ID  DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5  DATA_6  DATA_7  \
0  513       0       0       0       0       0       0       0     125
1  513       0       0       0       0       0       0       0     125
2  513       0       0       0       0       0       0       0     125
3  513       0       0       0       0       0       0       0     125
4  513       0       0       0       0       0       0       0     125

    label  category specific_class
0  ATTACK  SPOOFING            GAS
1  ATTACK  SPOOFING            GAS
2  ATTACK  SPOOFING            GAS
3  ATTACK  SPOOFING            GAS
4  ATTACK  SPOOFING            GAS

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9991 entries, 0 to 9990
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ID              9991 non-null   int64
 1   DATA_0          9991 non-null   int64
 2   DATA_1          9991 non-null   int64
 3   DATA_2          9991 non-null   int64
 4   DATA_3          9991 non-null   int64
 5   DATA_4          9991 non-null   int64
 6   DATA_5          9991 non-null   int64
 7   DATA_6          9991 non-null   int64
 8   DATA_7          9991 non-null   int64
 9   label           9991 non-null   object
 10  category        9991 non-null   object
 11  specific_class  9991 non-null   object
dtypes: int64(9), object(3)
memory usage: 936.8+ KB
None

Filas enteras repetidas: 9989

Valores ausentes:
ID                0
DATA_0            0
DATA_1            0
DATA_2            0
DATA_3            0
DATA_4            0
DATA_5            0
DATA_6            0
DATA_7            0
label             0
category          0
specific_class    0
dtype: int64

Cantidad de CAN IDs únicos: 1
Top 10 CAN IDs más frecuentes:
ID
513    9991
Name: count, dtype: int64
```

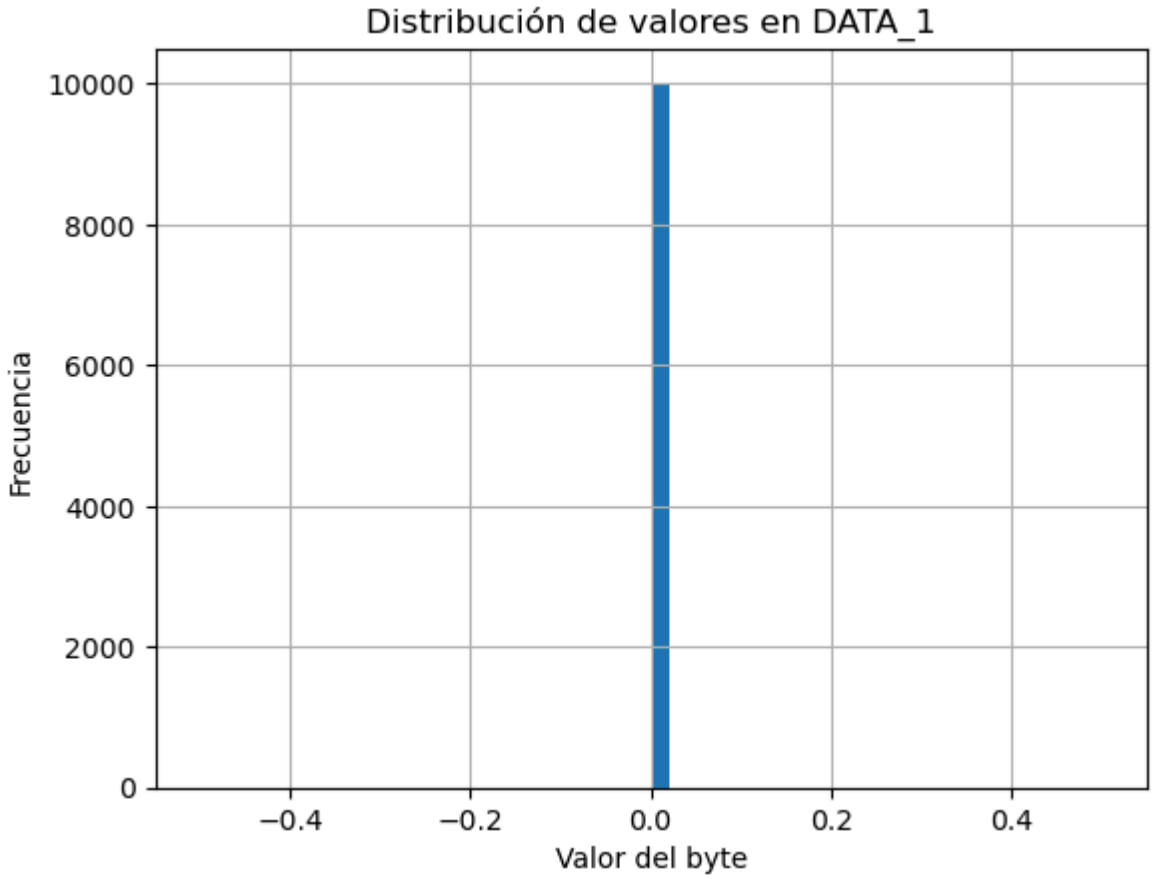
```
Estadísticas descriptivas:
           ID  DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5       DATA_6
\
count  9991.0  9991.0  9991.0  9991.0  9991.0  9991.0  9991.0  9991.000000
mean    513.0     0.0     0.0     0.0     0.0     0.0     0.0    32.022420
std       0.0     0.0     0.0     0.0     0.0     0.0     0.0    32.001594
min     513.0     0.0     0.0     0.0     0.0     0.0     0.0     0.000000
25%     513.0     0.0     0.0     0.0     0.0     0.0     0.0     0.000000
50%     513.0     0.0     0.0     0.0     0.0     0.0     0.0    64.000000
75%     513.0     0.0     0.0     0.0     0.0     0.0     0.0    64.000000
max     513.0     0.0     0.0     0.0     0.0     0.0     0.0    64.000000

            DATA_7
count  9991.000000
mean    140.510860
std      15.500772
min     125.000000
25%     125.000000
50%     156.000000
75%     156.000000
max     156.000000

Posibles valores atípicos en los bytes de datos:
       DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5       DATA_6  \
count  9991.0  9991.0  9991.0  9991.0  9991.0  9991.0  9991.000000
mean      0.0     0.0     0.0     0.0     0.0     0.0    32.022420
std       0.0     0.0     0.0     0.0     0.0     0.0    32.001594
min       0.0     0.0     0.0     0.0     0.0     0.0     0.000000
1%        0.0     0.0     0.0     0.0     0.0     0.0     0.000000
25%       0.0     0.0     0.0     0.0     0.0     0.0     0.000000
50%       0.0     0.0     0.0     0.0     0.0     0.0    64.000000
75%       0.0     0.0     0.0     0.0     0.0     0.0    64.000000
99%       0.0     0.0     0.0     0.0     0.0     0.0    64.000000
max       0.0     0.0     0.0     0.0     0.0     0.0    64.000000

            DATA_7
count  9991.000000
mean    140.510860
std      15.500772
min     125.000000
1%      125.000000
25%     125.000000
50%     156.000000
75%     156.000000
99%     156.000000
max     156.000000

Distribución de etiquetas:
label
ATTACK    9991
Name: count, dtype: int64

Distribución de categorías:
category
SPOOFING    9991
Name: count, dtype: int64

Distribución de clases específicas:
specific_class
```

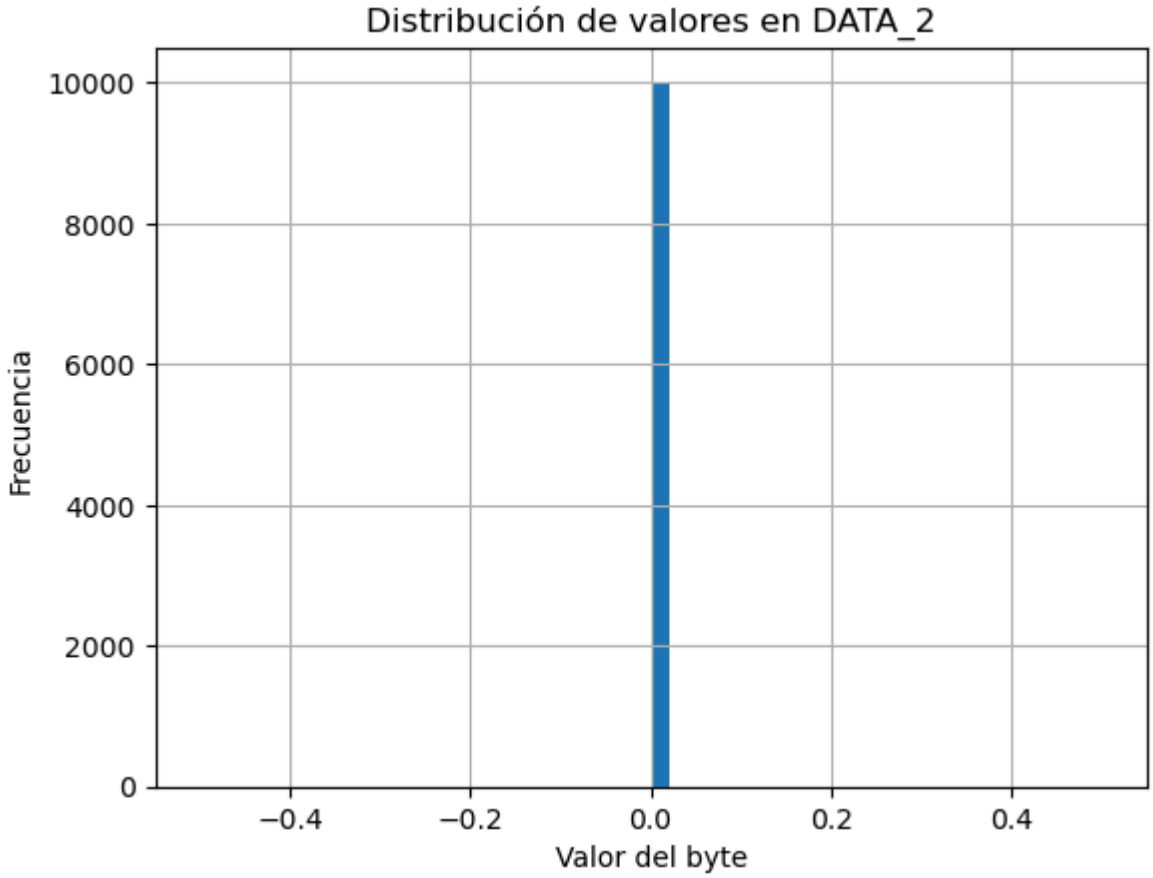
```
GAS    9991
Name: count, dtype: int64
```

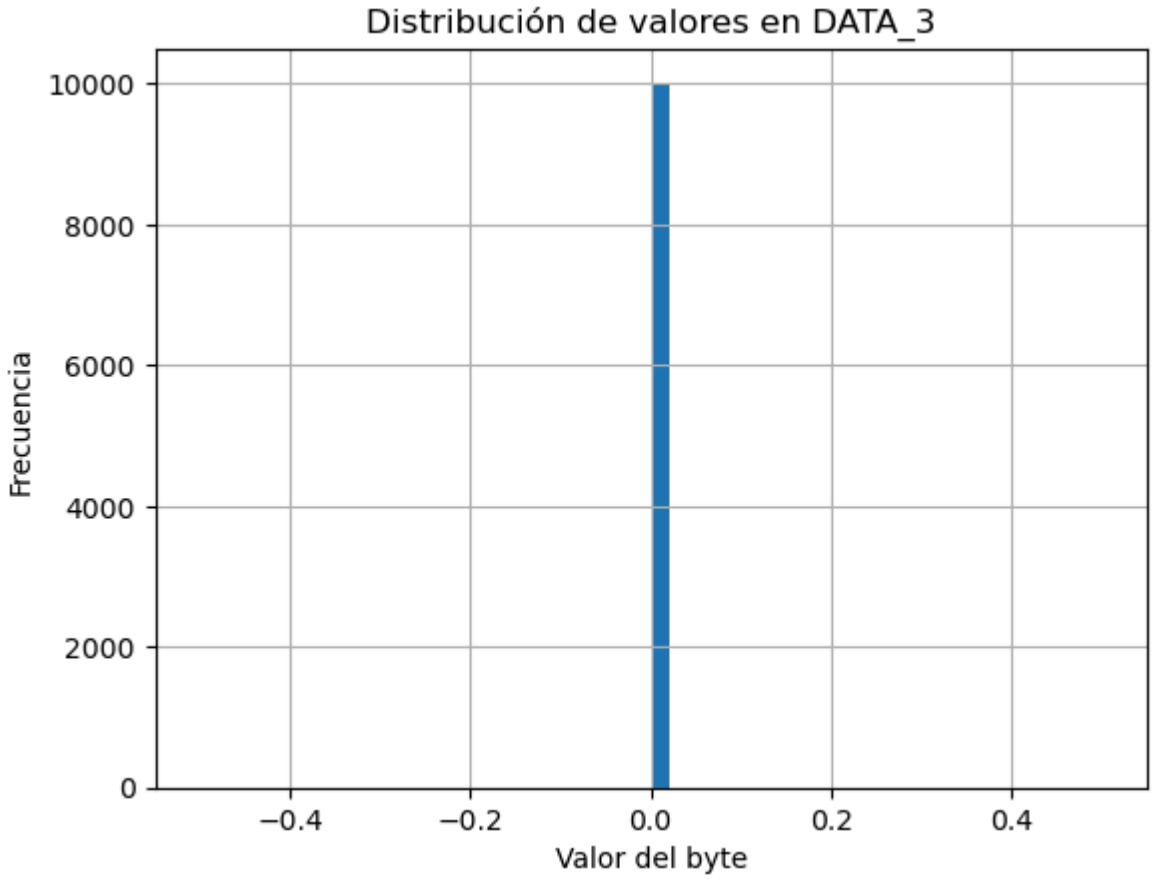


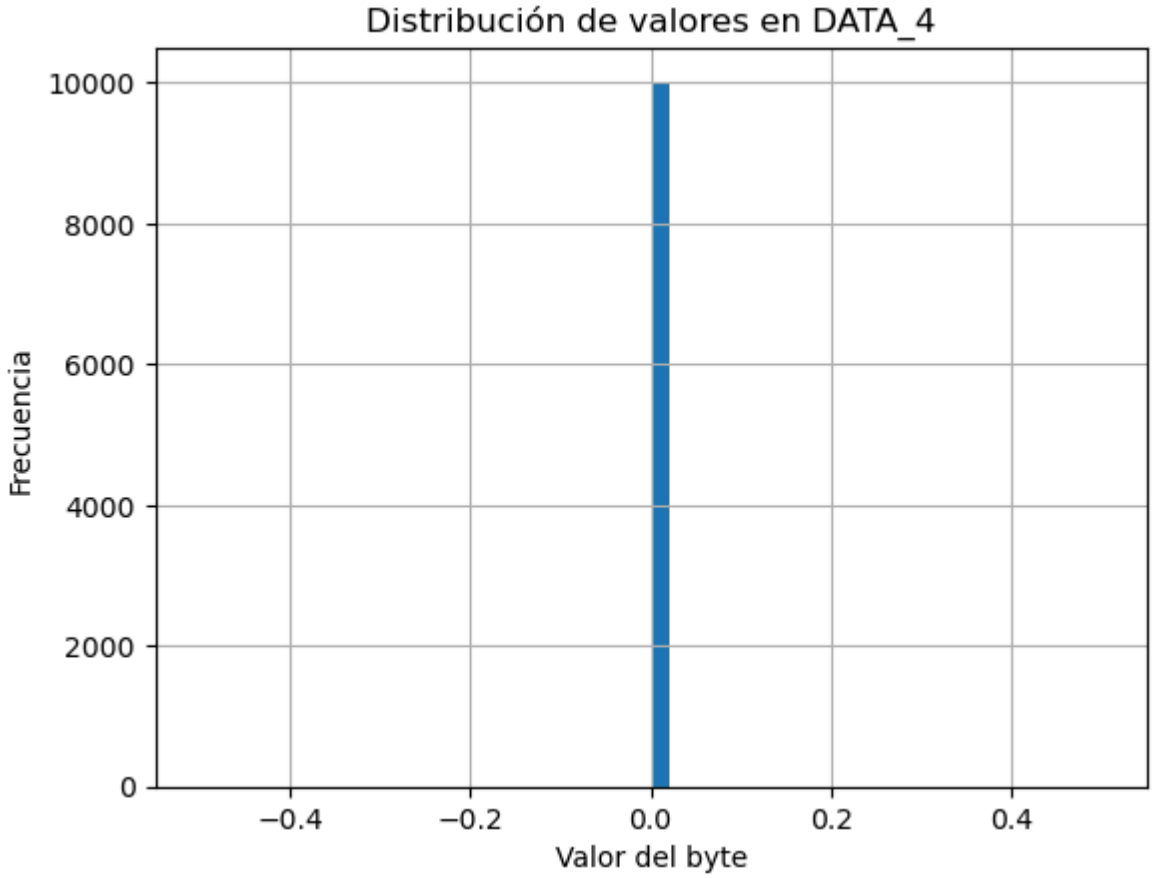


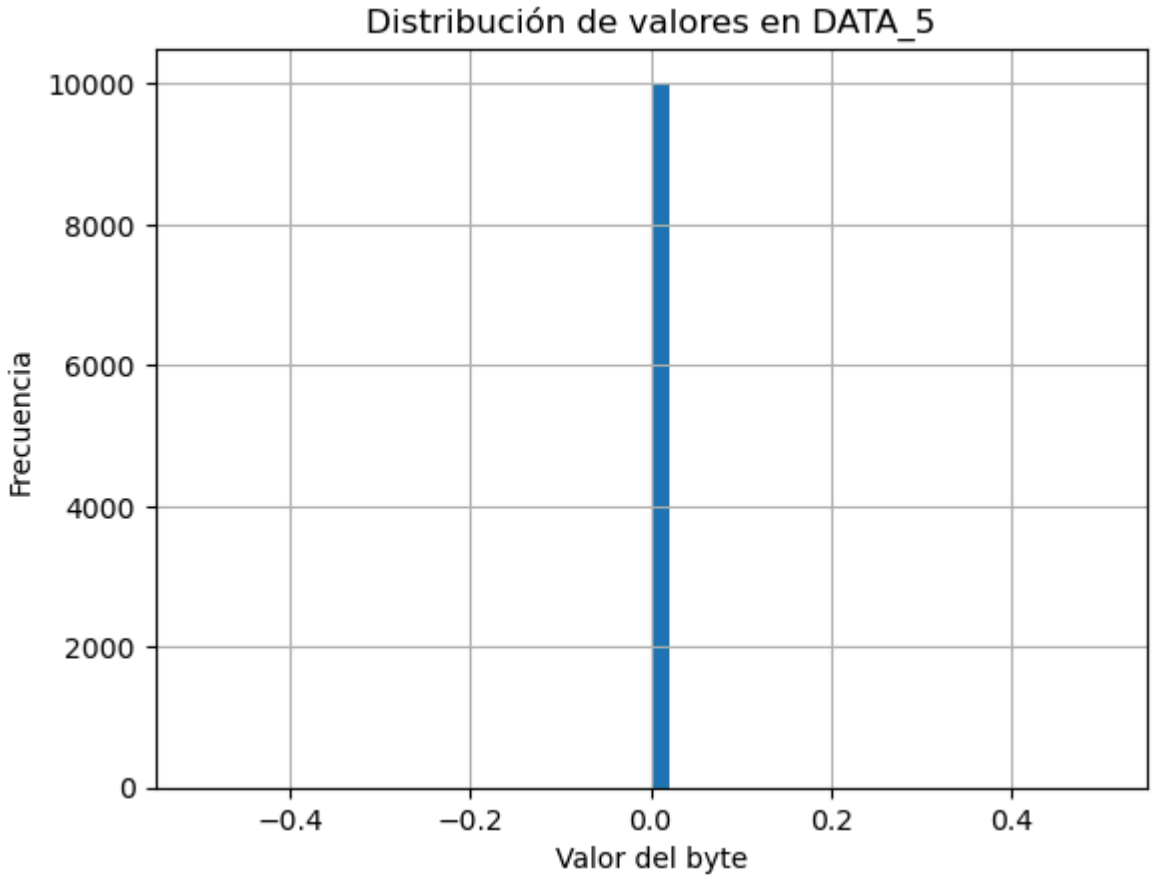

Distribución de valores en DATA_0
Frecuencia
10000
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4
Valor del byte

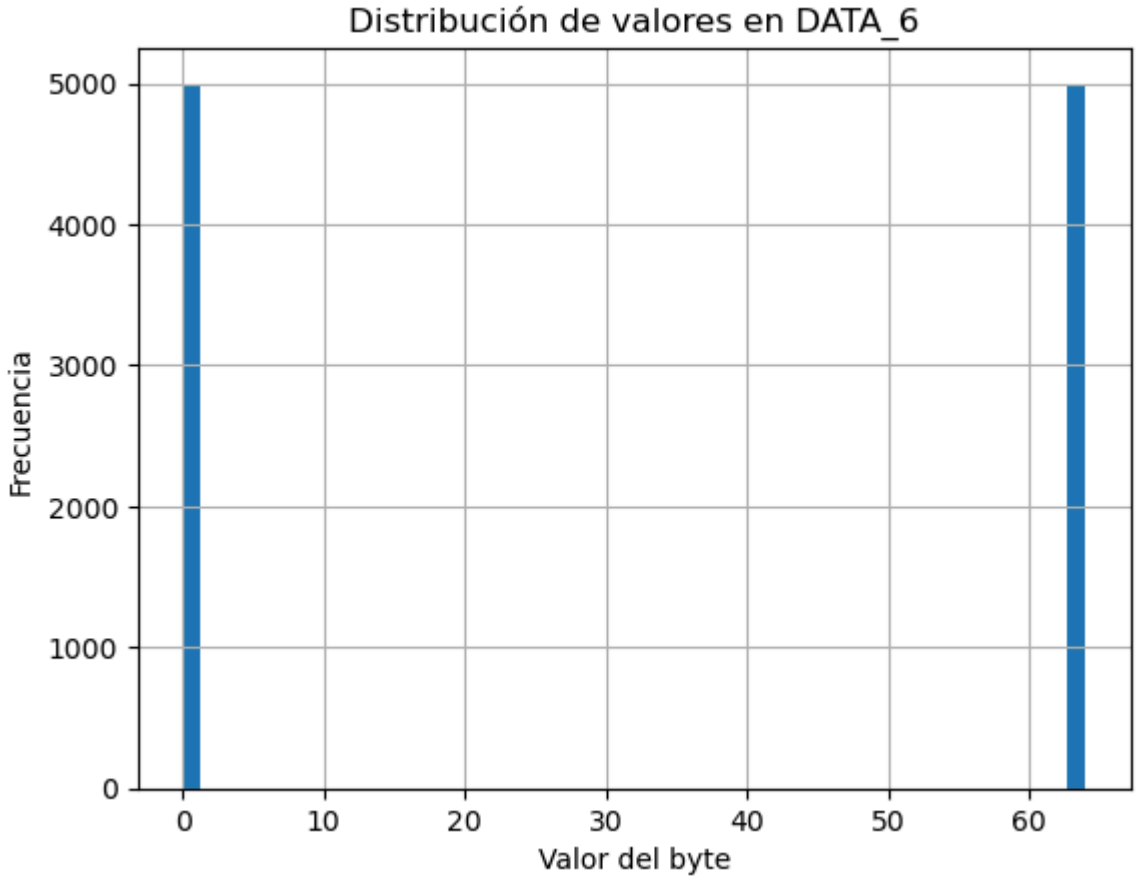

Distribución de valores en DATA_1
Frecuencia
10000
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4
Valor del byte

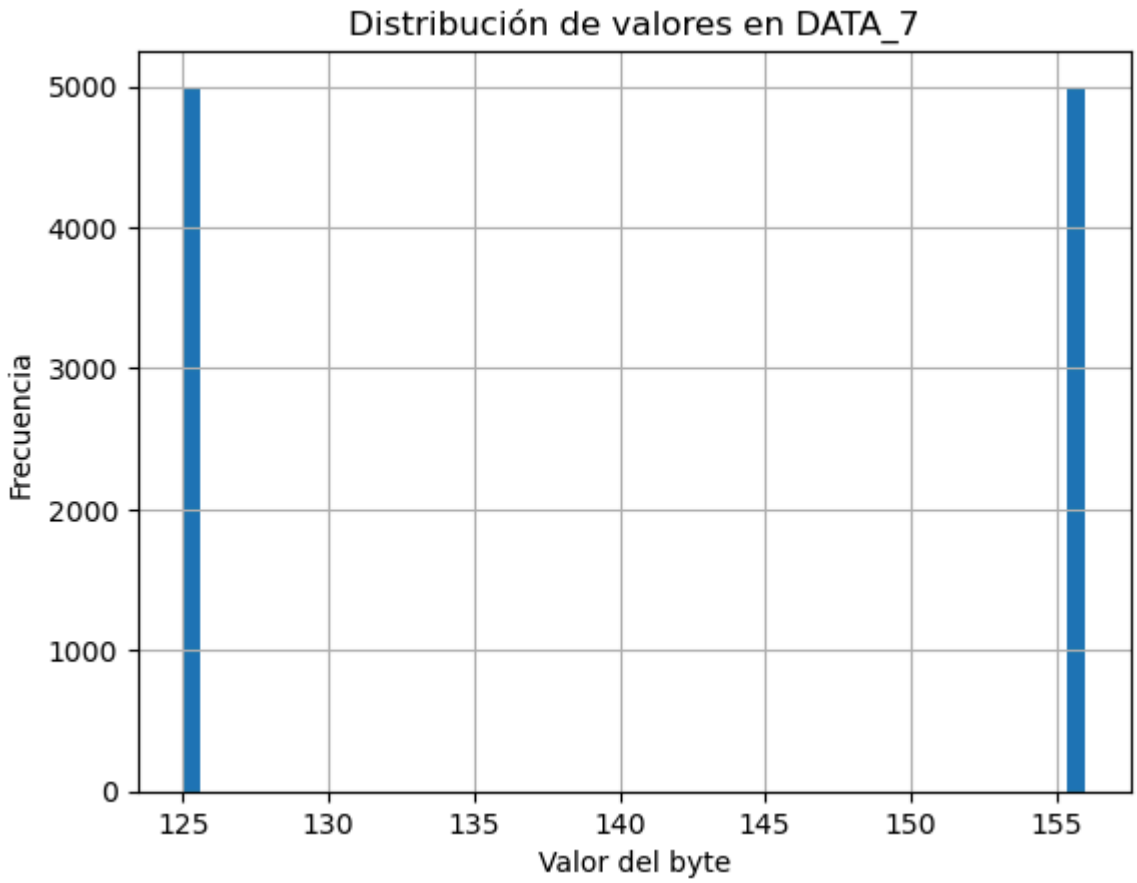

Distribución de valores en DATA_2
Frecuencia
10000
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4
Valor del byte


Distribución de valores en DATA_3
Frecuencia
10000
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4
Valor del byte


Distribución de valores en DATA_4
Frecuencia
10000
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4
Valor del byte


Distribución de valores en DATA_5
Frecuencia
10000
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4
Valor del byte


Distribución de valores en DATA_6
Frecuencia
Valor del byte
0
1000
2000
3000
4000
5000
0
10
20
30
40
50
60


Distribución de valores en DATA_7
Frecuencia
Valor del byte
0
1000
2000
3000
4000
5000
125
130
135
140
145
150
155

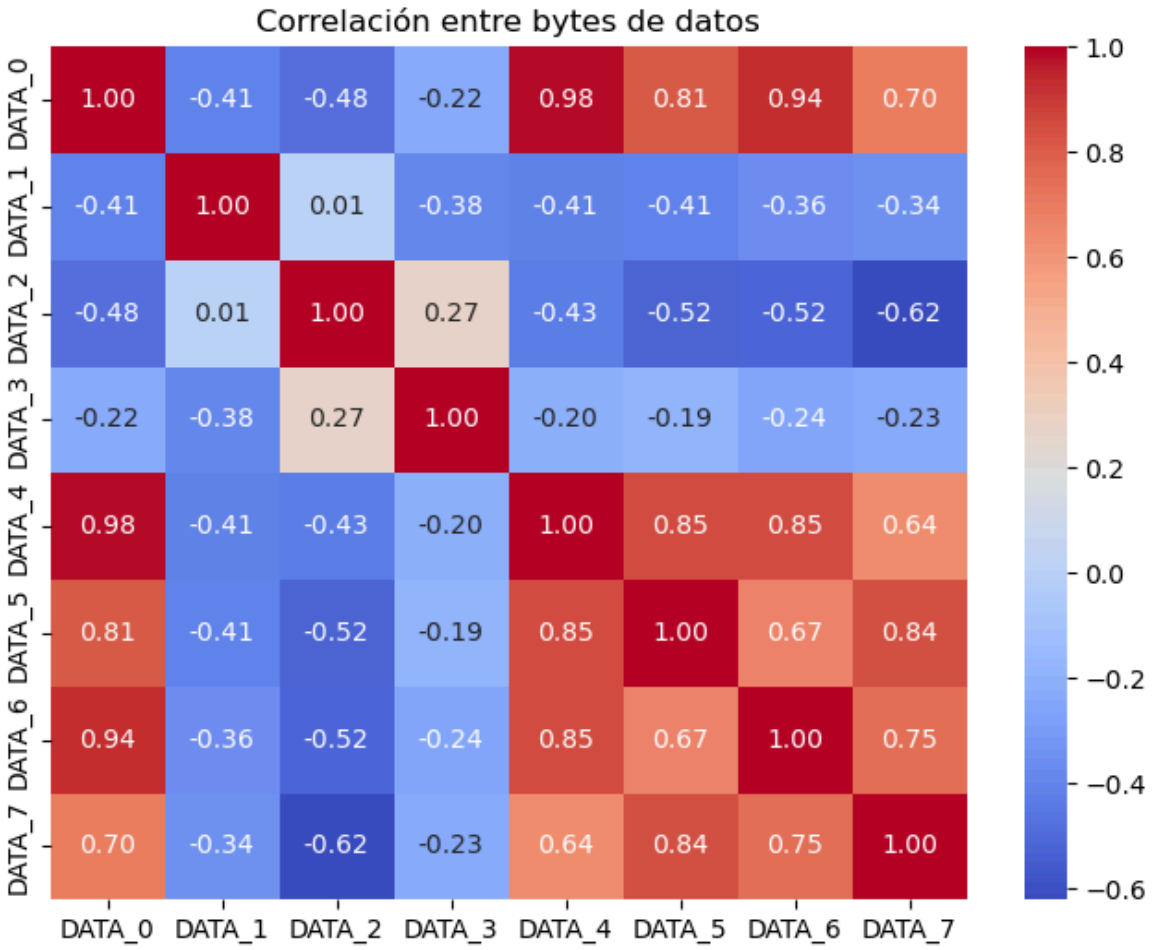
```
Valores fuera de rango (esperado 0-255):
DATA_0    0
DATA_1    0
DATA_2    0
DATA_3    0
DATA_4    0
DATA_5    0
DATA_6    0
DATA_7    0
dtype: int64
```

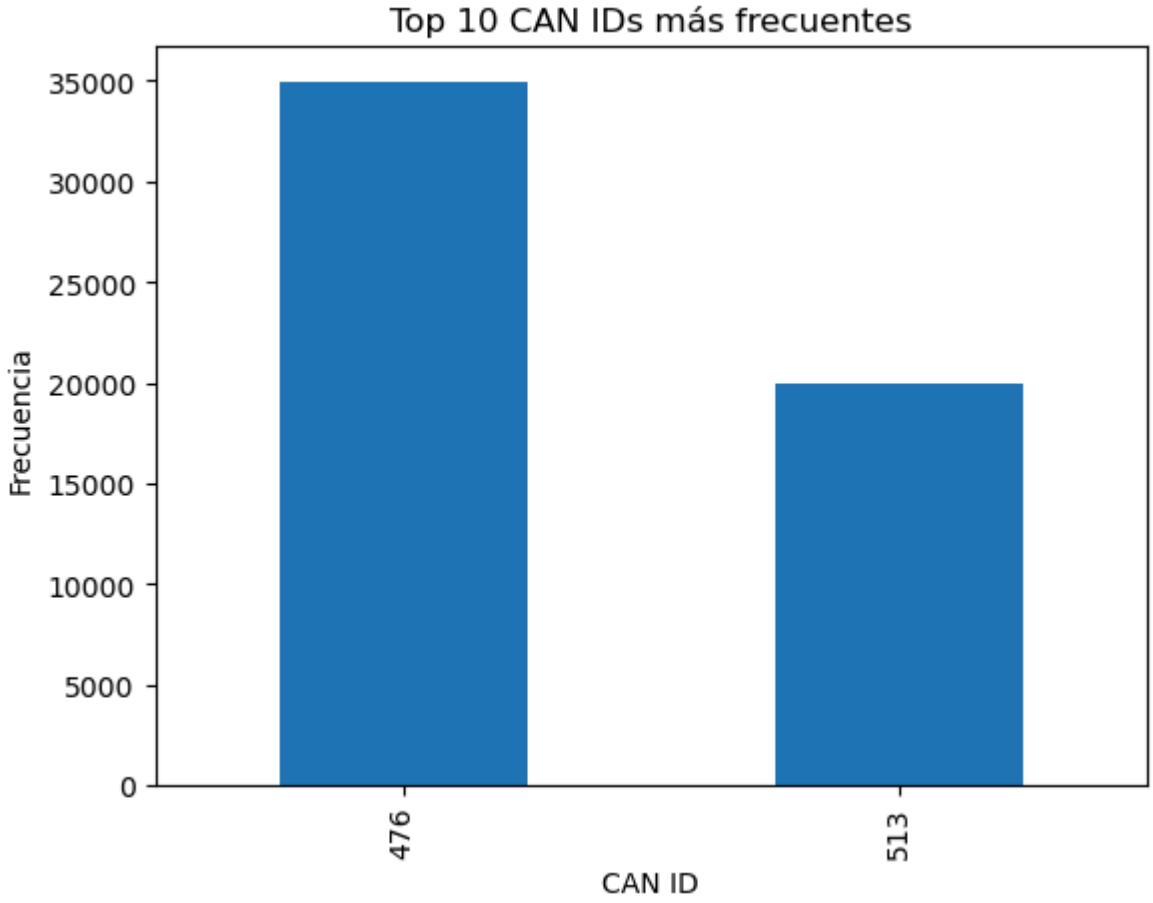
In [61]:
```
eda(raw_dec_rpm, 'dec_rpm')
```

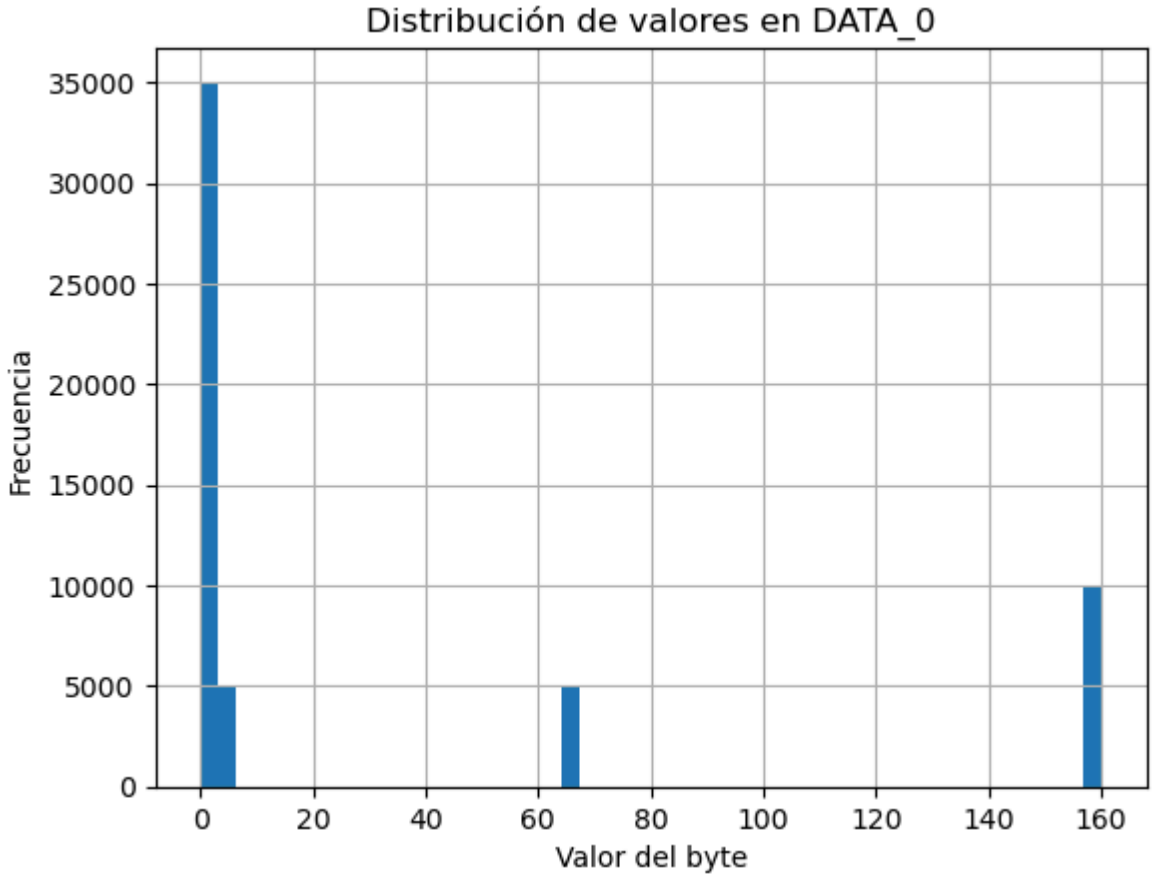
```
EDA de dec_rpm

Primeras filas:
    ID  DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5  DATA_6  DATA_7  \
0  476       2      61      23      19       0       0       0       0
1  476       2      61      23      19       0       0       0       0
2  476       2      61      23      19       0       0       0       0
3  476       2      61      23      19       0       0       0       0
4  476       2      61      23      19       0       0       0       0

    label  category specific_class
0  ATTACK  SPOOFING            RPM
1  ATTACK  SPOOFING            RPM
2  ATTACK  SPOOFING            RPM
3  ATTACK  SPOOFING            RPM
4  ATTACK  SPOOFING            RPM

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54900 entries, 0 to 54899
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ID              54900 non-null  int64
 1   DATA_0          54900 non-null  int64
 2   DATA_1          54900 non-null  int64
 3   DATA_2          54900 non-null  int64
 4   DATA_3          54900 non-null  int64
 5   DATA_4          54900 non-null  int64
 6   DATA_5          54900 non-null  int64
 7   DATA_6          54900 non-null  int64
 8   DATA_7          54900 non-null  int64
 9   label           54900 non-null  object
 10  category        54900 non-null  object
 11  specific_class  54900 non-null  object
dtypes: int64(9), object(3)
memory usage: 5.0+ MB
None

Filas enteras repetidas: 54890

Valores ausentes:
ID                0
DATA_0            0
DATA_1            0
DATA_2            0
DATA_3            0
DATA_4            0
DATA_5            0
DATA_6            0
DATA_7            0
label             0
category          0
specific_class    0
dtype: int64

Cantidad de CAN IDs únicos: 2
Top 10 CAN IDs más frecuentes:
ID
476    34937
513    19963
```

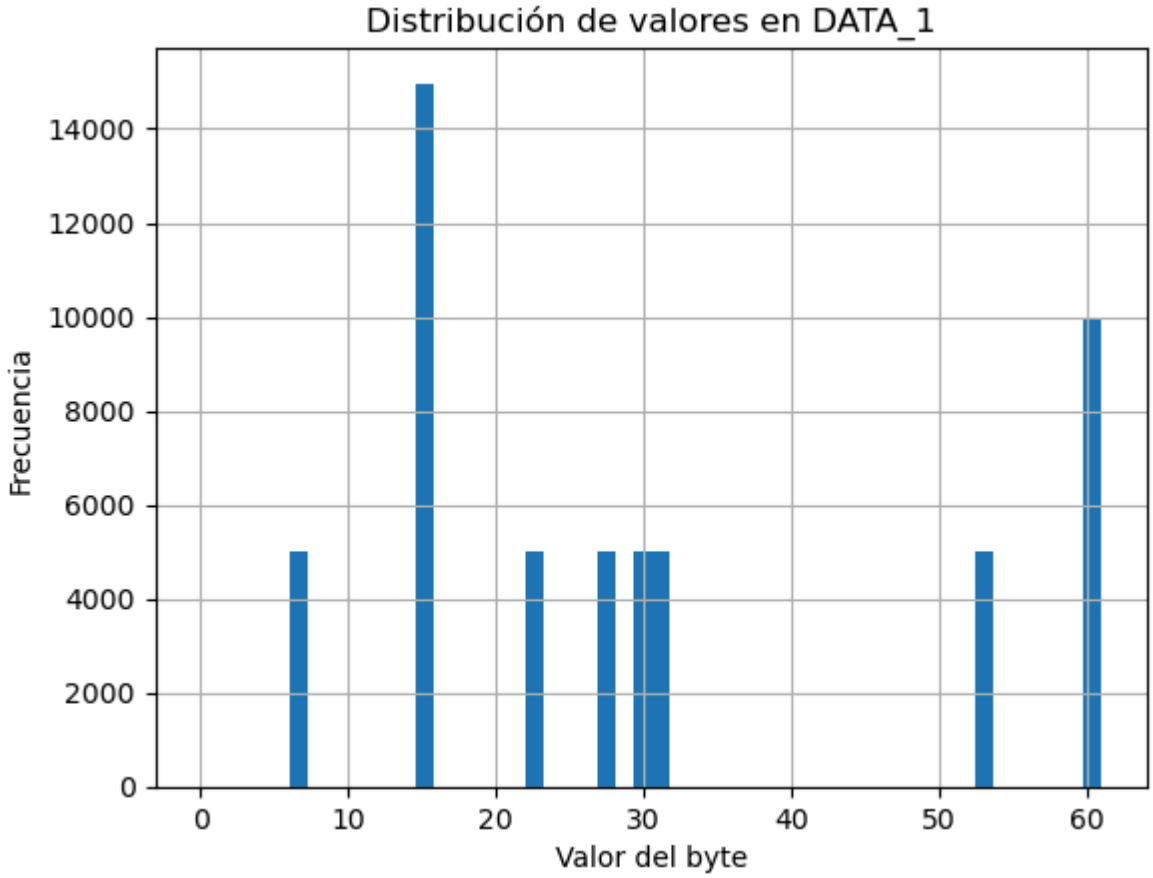
```
Name: count, dtype: int64

Estadísticas descriptivas:
                ID        DATA_0        DATA_1        DATA_2        DATA_
3  \
count  54900.000000  54900.000000  54900.000000  54900.000000  54900.00000
0
mean     489.454117     36.713151     30.821239     77.000656     67.71256
8
std       17.798731     60.685235     18.347934     88.807931     79.15539
7
min      476.000000      0.000000      0.000000      0.000000      0.00000
0
25%      476.000000      2.000000     15.000000      3.000000     19.00000
0
50%      476.000000      2.000000     28.000000     23.000000     37.00000
0
75%      513.000000     64.000000     53.000000    177.000000     69.00000
0
max      513.000000    160.000000     61.000000    245.000000    239.00000
0

            DATA_4        DATA_5        DATA_6        DATA_7
count  54900.000000  54900.000000  54900.000000  54900.000000
mean      33.257814     13.902987     39.623534     12.995647
std       63.865386     21.903687     64.361028     17.293106
min        0.000000      0.000000      0.000000      0.000000
25%        0.000000      0.000000      0.000000      0.000000
50%        0.000000      0.000000      0.000000      0.000000
75%       24.000000     42.000000    138.000000     34.000000
max      168.000000     53.000000    148.000000     41.000000

Posibles valores atípicos en los bytes de datos:
            DATA_0        DATA_1        DATA_2        DATA_3        DATA_
4  \
count  54900.000000  54900.000000  54900.000000  54900.000000  54900.00000
0
mean      36.713151     30.821239     77.000656     67.712568     33.25781
4
std       60.685235     18.347934     88.807931     79.155397     63.86538
6
min        0.000000      0.000000      0.000000      0.000000      0.00000
0
1%         2.000000      7.000000      2.000000      3.000000      0.00000
0
25%        2.000000     15.000000      3.000000     19.000000      0.00000
0
50%        2.000000     28.000000     23.000000     37.000000      0.00000
0
75%       64.000000     53.000000    177.000000     69.000000     24.00000
0
99%      160.000000     61.000000    245.000000    239.000000    168.00000
0
max      160.000000     61.000000    245.000000    239.000000    168.00000
0

            DATA_5        DATA_6        DATA_7
count  54900.000000  54900.000000  54900.000000
mean      13.902987     39.623534     12.995647
std       21.903687     64.361028     17.293106
```

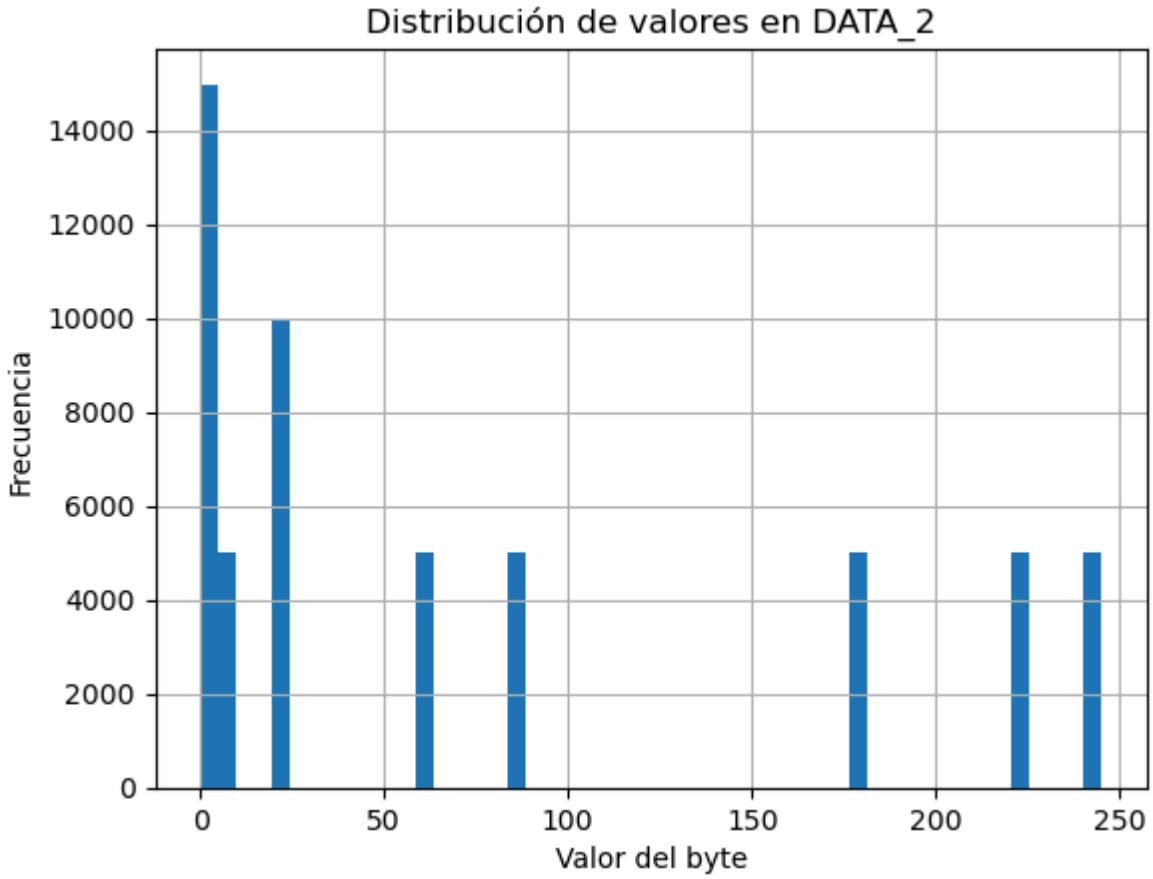
```
min         0.000000      0.000000      0.000000
1%          0.000000      0.000000      0.000000
25%         0.000000      0.000000      0.000000
50%         0.000000      0.000000      0.000000
75%        42.000000    138.000000     34.000000
99%        53.000000    148.000000     41.000000
max        53.000000    148.000000     41.000000

Distribución de etiquetas:
label
ATTACK    54900
Name: count, dtype: int64

Distribución de categorías:
category
SPOOFING    54900
Name: count, dtype: int64

Distribución de clases específicas:
specific_class
RPM    54900
Name: count, dtype: int64
```

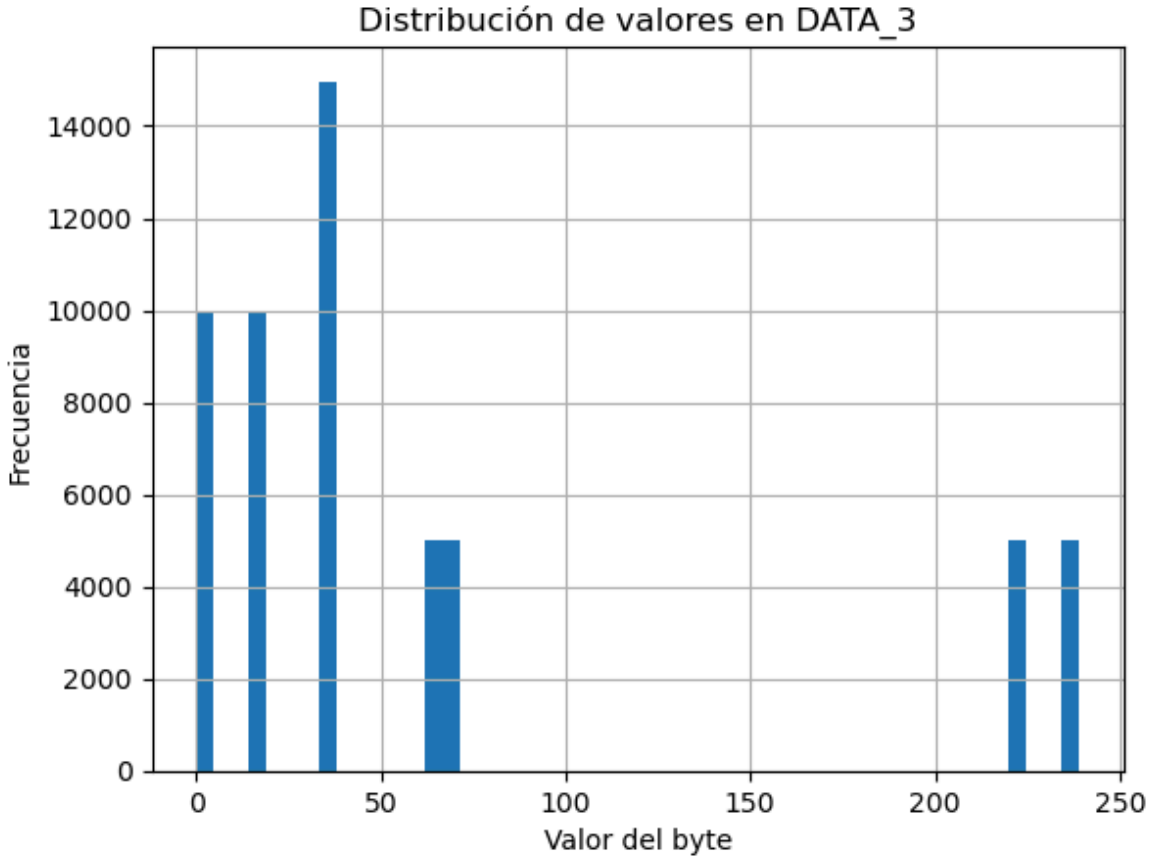


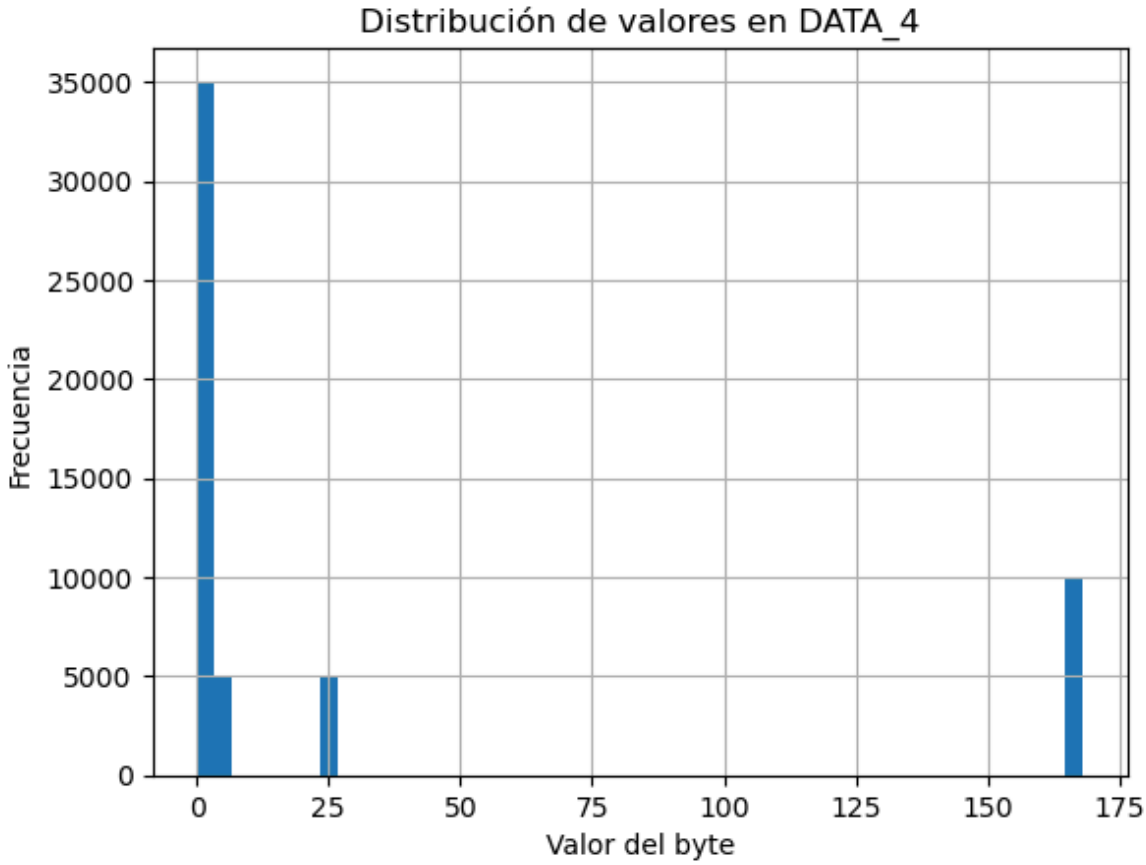

Top 10 CAN IDs más frecuentes
Frecuencia
35000
30000
25000
20000
15000
10000
5000
0
476
513
CAN ID

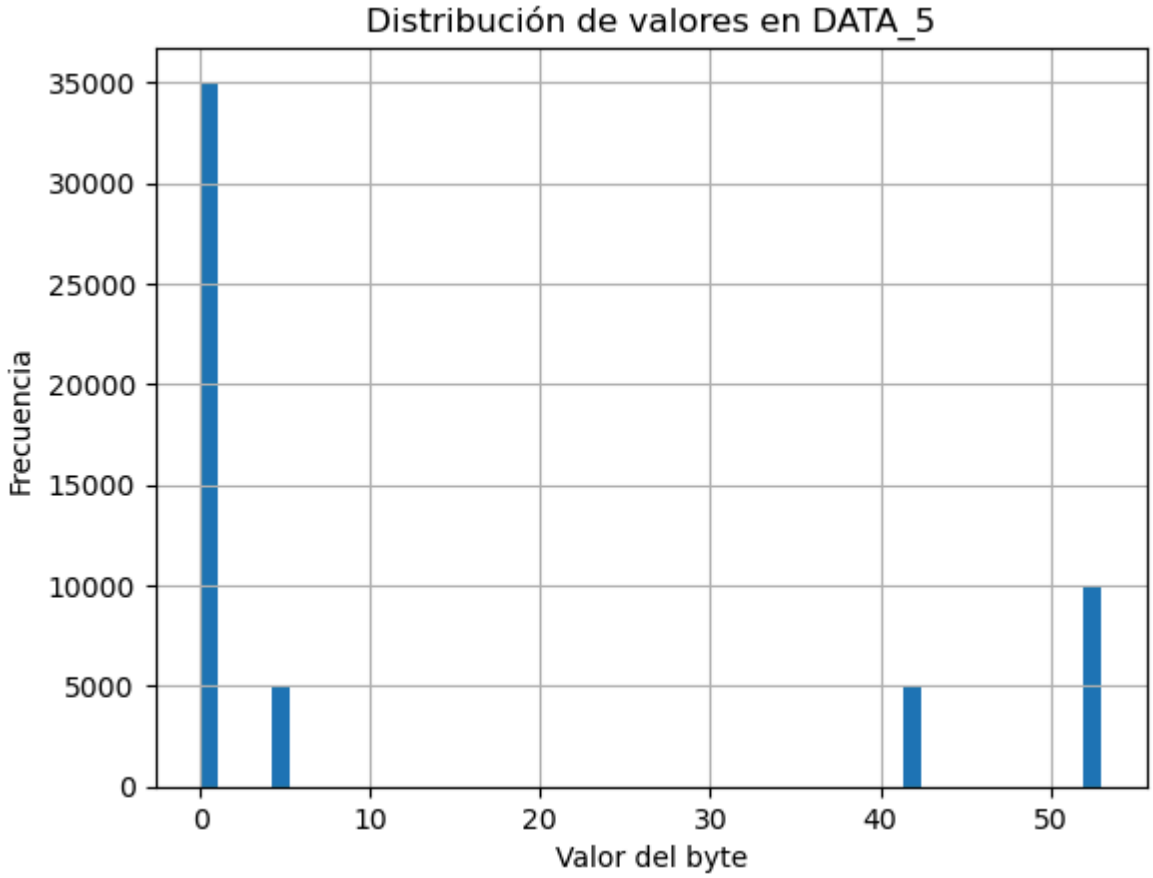

Distribución de valores en DATA_0
Frecuencia
35000
30000
25000
20000
15000
10000
5000
0
0
20
40
60
80
100
120
140
160
Valor del byte

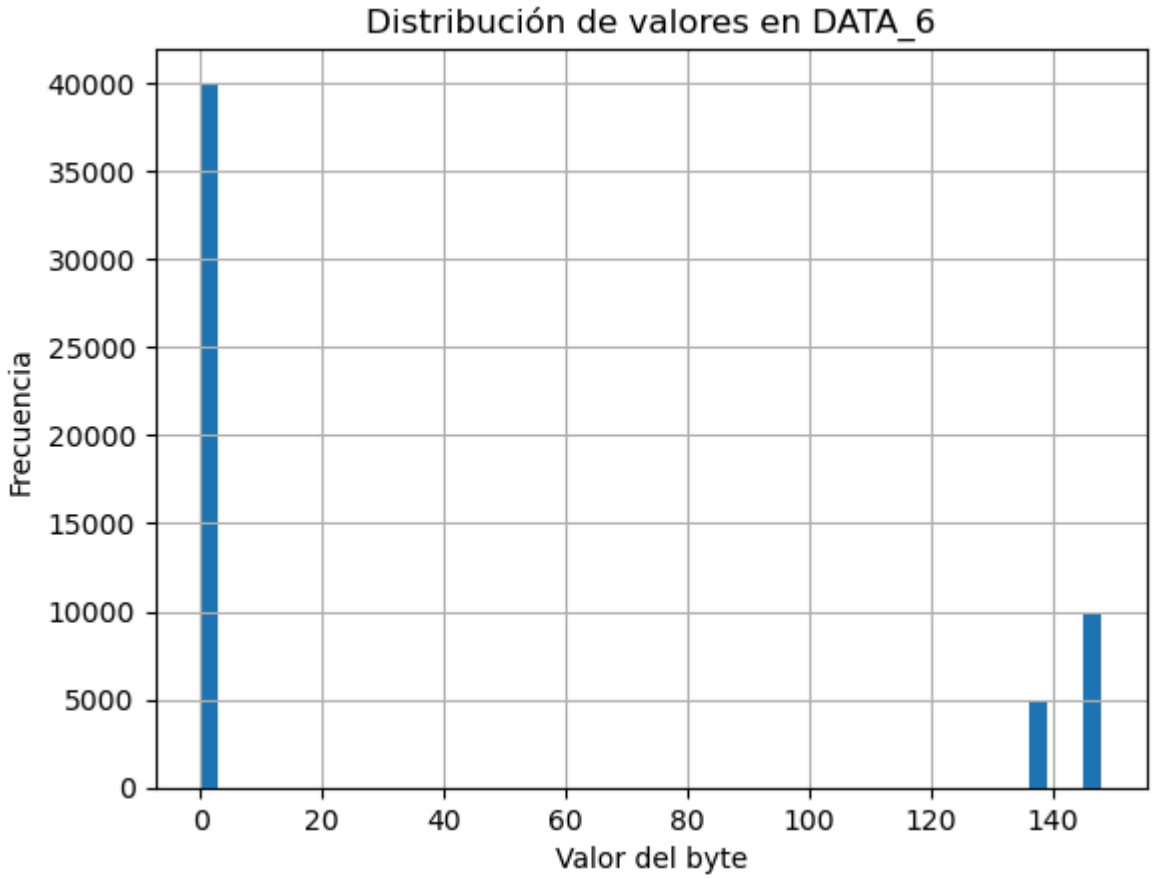

Distribución de valores en DATA_1
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
12000
14000
0
10
20
30
40
50
60

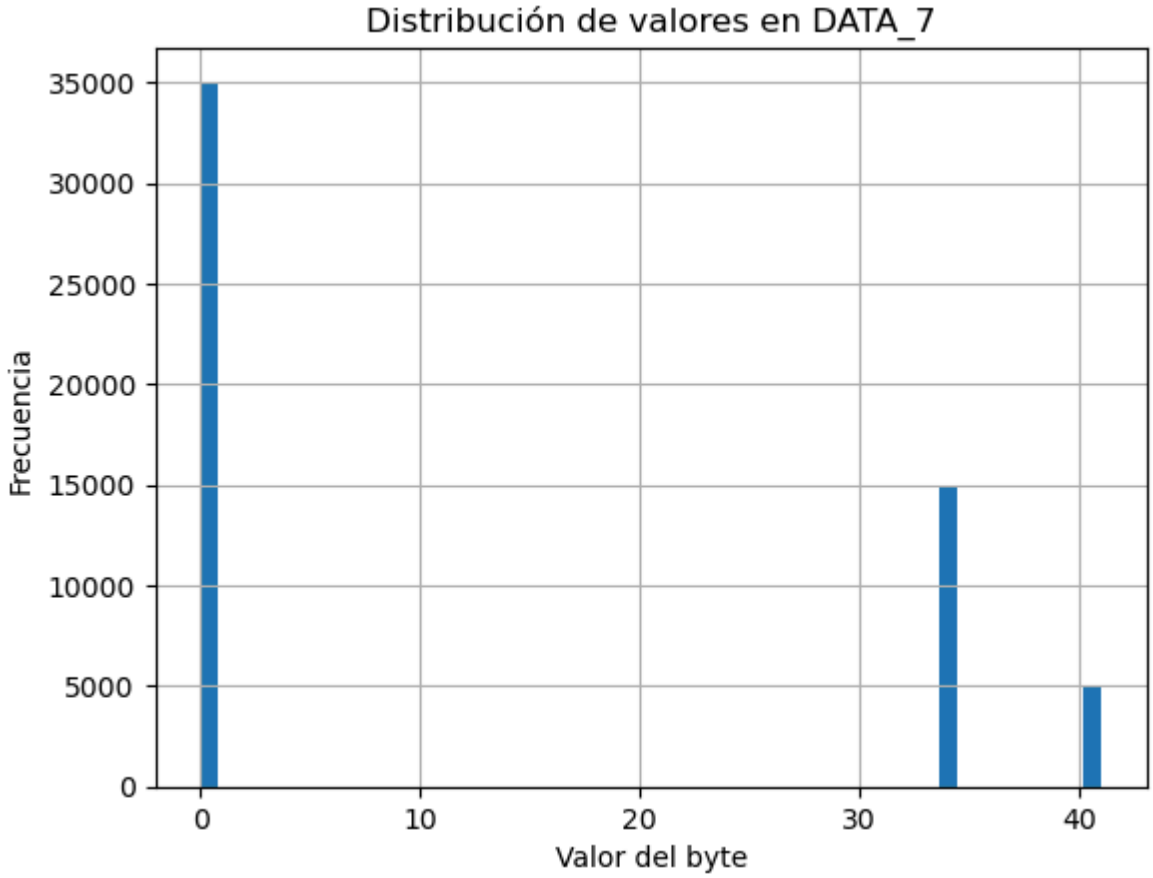

Distribución de valores en DATA_2
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
12000
14000
0
50
100
150
200
250


Distribución de valores en DATA_3
Frecuencia
Valor del byte
0
2000
4000
6000
8000
10000
12000
14000
0
50
100
150
200
250


Distribución de valores en DATA_4
Frecuencia
Valor del byte
0
5000
10000
15000
20000
25000
30000
35000
0
25
50
75
100
125
150
175


Distribución de valores en DATA_5
Frecuencia
Valor del byte
0
5000
10000
15000
20000
25000
30000
35000
0
10
20
30
40
50


Distribución de valores en DATA_6
Frecuencia
Valor del byte
0
5000
10000
15000
20000
25000
30000
35000
40000
0
20
40
60
80
100
120
140



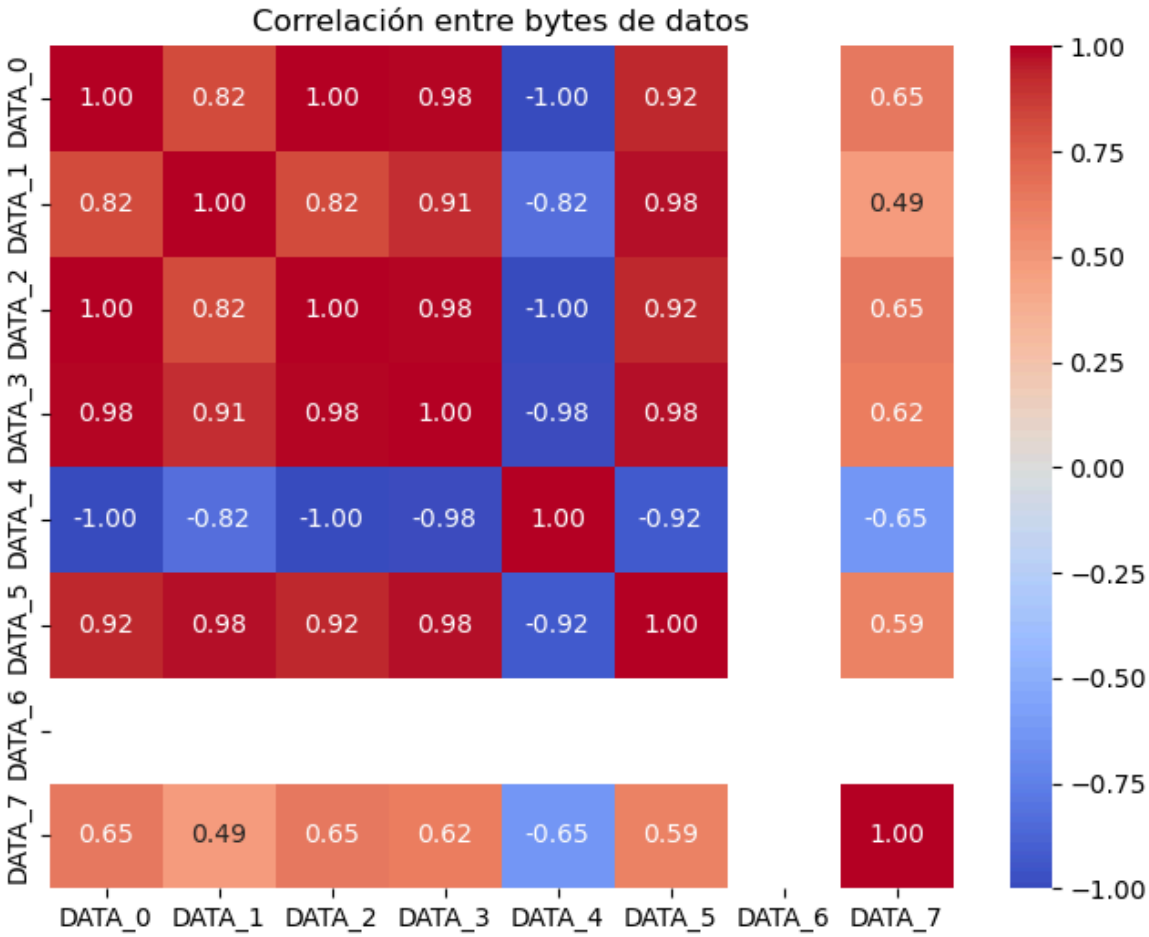

```
Valores fuera de rango (esperado 0–255):
DATA_0    0
DATA_1    0
DATA_2    0
DATA_3    0
DATA_4    0
DATA_5    0
DATA_6    0
DATA_7    0
dtype: int64
```

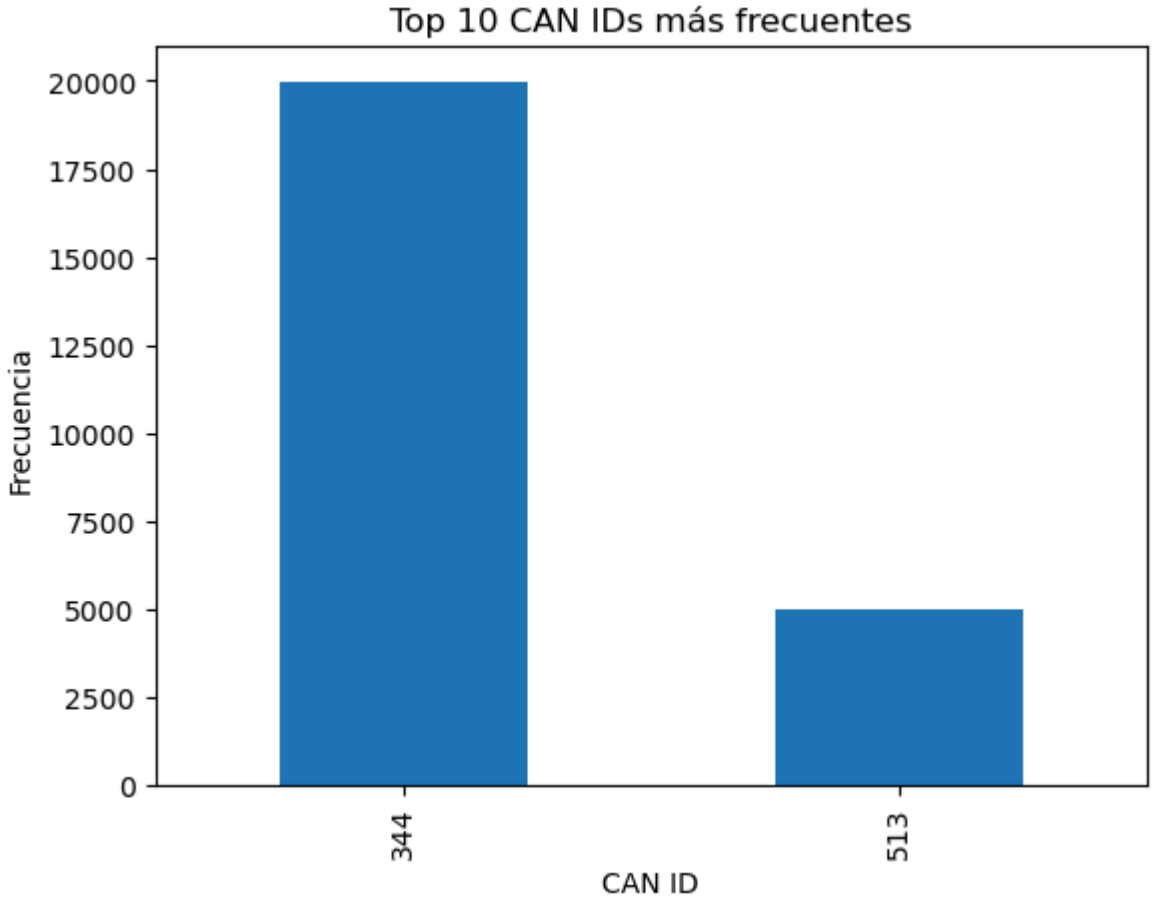
In [63]:
```
eda(raw_dec_speed, 'dec_speed')
```

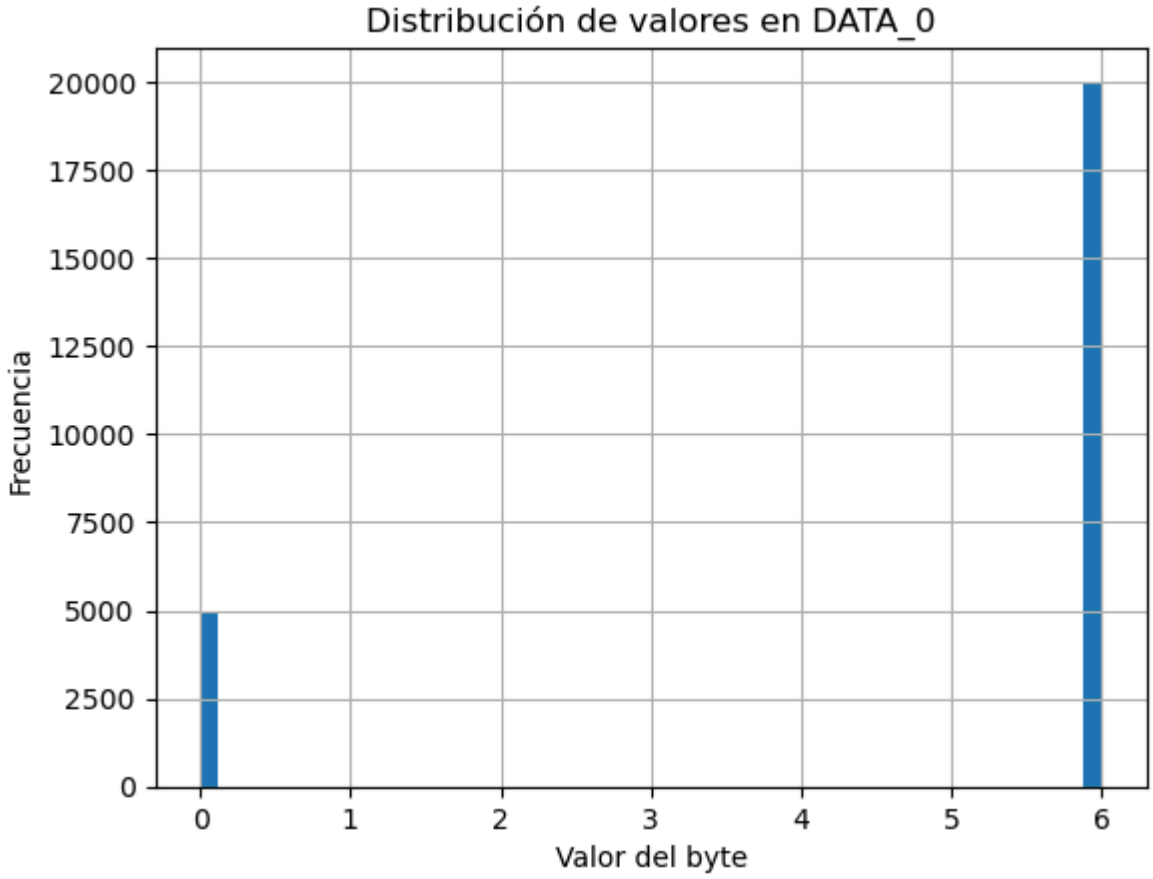
```
EDA de dec_speed

Primeras filas:
    ID  DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5  DATA_6  DATA_7  \
0  344       6      12       6      51       6      26       2      54
1  344       6      12       6      51       6      26       2      54
2  344       6      12       6      51       6      26       2      54
3  344       6      12       6      51       6      26       2      54
4  344       6      12       6      51       6      26       2      54

    label  category specific_class
0  ATTACK  SPOOFING          SPEED
1  ATTACK  SPOOFING          SPEED
2  ATTACK  SPOOFING          SPEED
3  ATTACK  SPOOFING          SPEED
4  ATTACK  SPOOFING          SPEED

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24951 entries, 0 to 24950
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ID              24951 non-null  int64
 1   DATA_0          24951 non-null  int64
 2   DATA_1          24951 non-null  int64
 3   DATA_2          24951 non-null  int64
 4   DATA_3          24951 non-null  int64
 5   DATA_4          24951 non-null  int64
 6   DATA_5          24951 non-null  int64
 7   DATA_6          24951 non-null  int64
 8   DATA_7          24951 non-null  int64
 9   label           24951 non-null  object
 10  category        24951 non-null  object
 11  specific_class  24951 non-null  object
dtypes: int64(9), object(3)
memory usage: 2.3+ MB
None

Filas enteras repetidas: 24946

Valores ausentes:
ID                0
DATA_0            0
DATA_1            0
DATA_2            0
DATA_3            0
DATA_4            0
DATA_5            0
DATA_6            0
DATA_7            0
label             0
category          0
specific_class    0
dtype: int64

Cantidad de CAN IDs únicos: 2
Top 10 CAN IDs más frecuentes:
ID
344    19962
513     4989
```

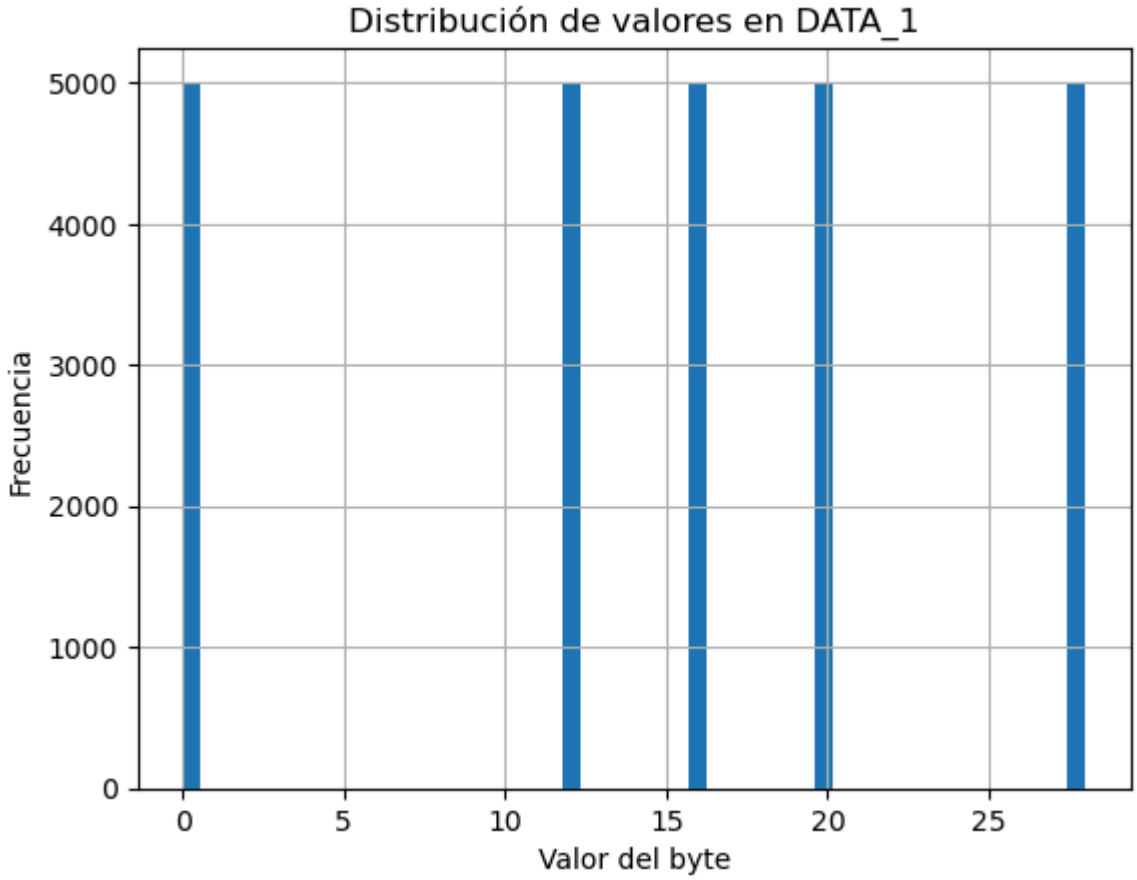
```
Name: count, dtype: int64

Estadísticas descriptivas:
                ID        DATA_0        DATA_1        DATA_2        DATA_
3  \
count  24951.000000  24951.000000  24951.000000  24951.000000  24951.00000
0
mean     377.791872      4.800289     15.200513      4.800289     45.00240
5
std       67.595258      2.399832      9.260055      2.399832     22.84561
4
min      344.000000      0.000000      0.000000      0.000000      0.00000
0
25%      344.000000      6.000000     12.000000      6.000000     51.00000
0
50%      344.000000      6.000000     16.000000      6.000000     54.00000
0
75%      344.000000      6.000000     20.000000      6.000000     57.00000
0
max      513.000000      6.000000     28.000000      6.000000     63.00000
0

            DATA_4        DATA_5   DATA_6        DATA_7
count  24951.000000  24951.000000  24951.0  24951.000000
mean      24.795479     26.001082      2.0     25.198269
std       37.597362     14.070209      0.0     19.527443
min        6.000000      0.000000      2.0      0.000000
25%        6.000000     26.000000      2.0     14.000000
50%        6.000000     29.000000      2.0     17.000000
75%        6.000000     33.000000      2.0     41.000000
max      100.000000     42.000000      2.0     54.000000

Posibles valores atípicos en los bytes de datos:
            DATA_0        DATA_1        DATA_2        DATA_3        DATA_
4  \
count  24951.000000  24951.000000  24951.000000  24951.000000  24951.00000
0
mean       4.800289     15.200513      4.800289     45.002405     24.79547
9
std        2.399832      9.260055      2.399832     22.845614     37.59736
2
min        0.000000      0.000000      0.000000      0.000000      6.00000
0
1%         0.000000      0.000000      0.000000      0.000000      6.00000
0
25%        6.000000     12.000000      6.000000     51.000000      6.00000
0
50%        6.000000     16.000000      6.000000     54.000000      6.00000
0
75%        6.000000     20.000000      6.000000     57.000000      6.00000
0
99%        6.000000     28.000000      6.000000     63.000000    100.00000
0
max        6.000000     28.000000      6.000000     63.000000    100.00000
0

            DATA_5   DATA_6        DATA_7
count  24951.000000  24951.0  24951.000000
mean      26.001082      2.0     25.198269
std       14.070209      0.0     19.527443
```

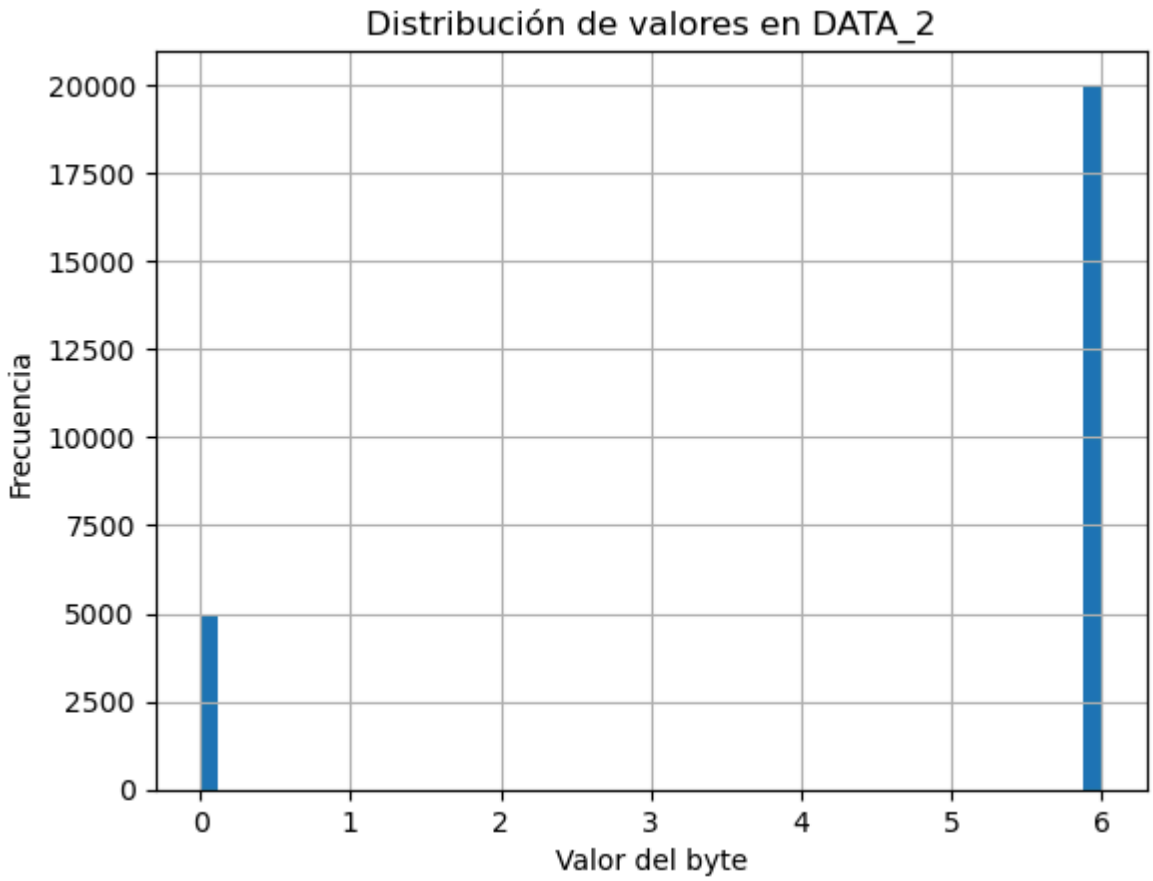
```
min         0.000000      2.0      0.000000
1%          0.000000      2.0      0.000000
25%        26.000000      2.0     14.000000
50%        29.000000      2.0     17.000000
75%        33.000000      2.0     41.000000
99%        42.000000      2.0     54.000000
max        42.000000      2.0     54.000000

Distribución de etiquetas:
label
ATTACK    24951
Name: count, dtype: int64

Distribución de categorías:
category
SPOOFING    24951
Name: count, dtype: int64

Distribución de clases específicas:
specific_class
SPEED    24951
Name: count, dtype: int64
```

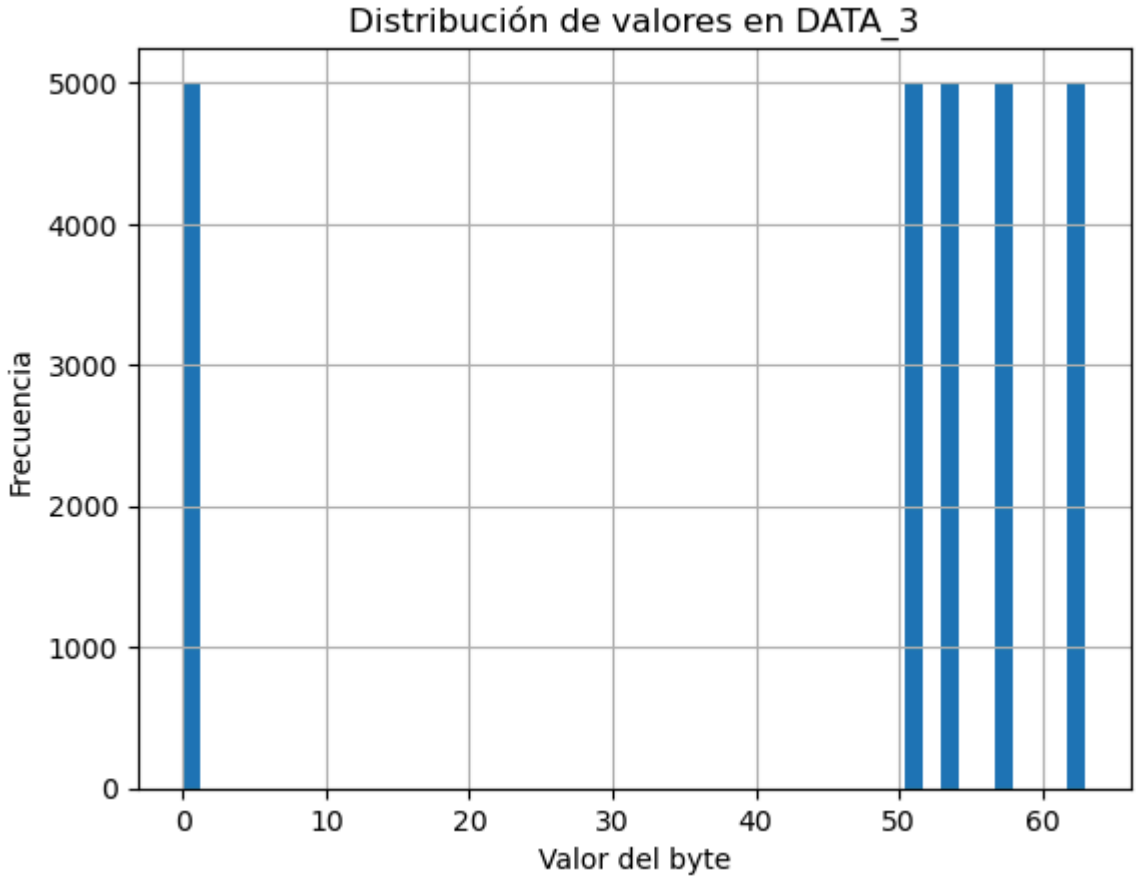


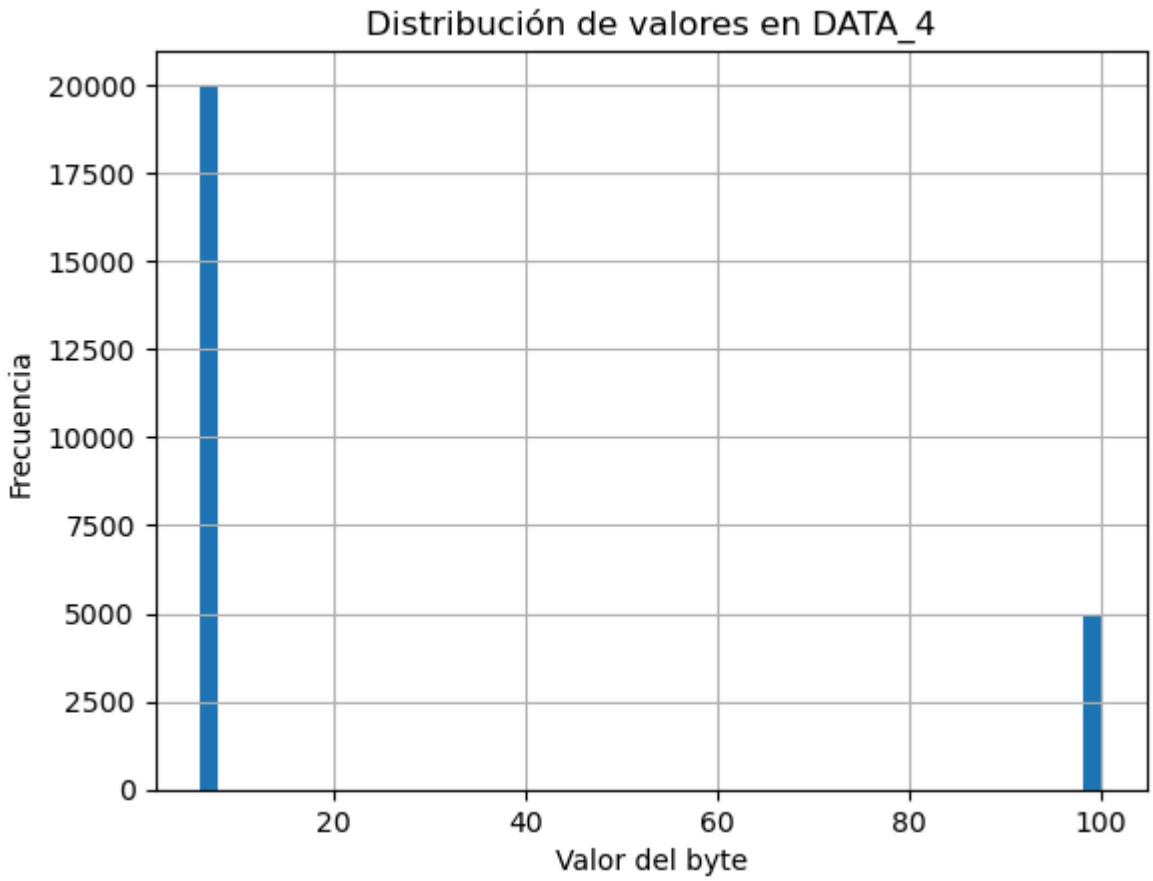

Top 10 CAN IDs más frecuentes
Frecuencia
20000
17500
15000
12500
10000
7500
5000
2500
0
344
513
CAN ID

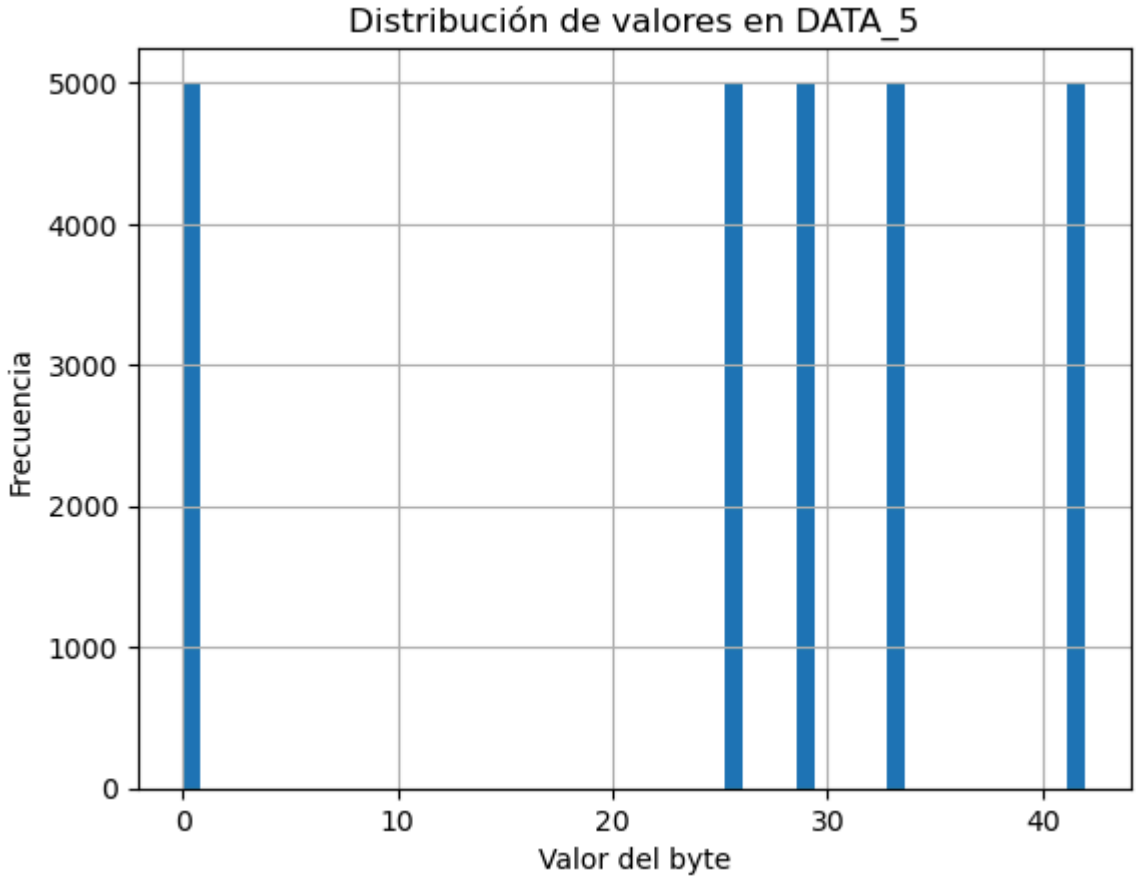

Distribución de valores en DATA_0
Frecuencia
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
6
Valor del byte

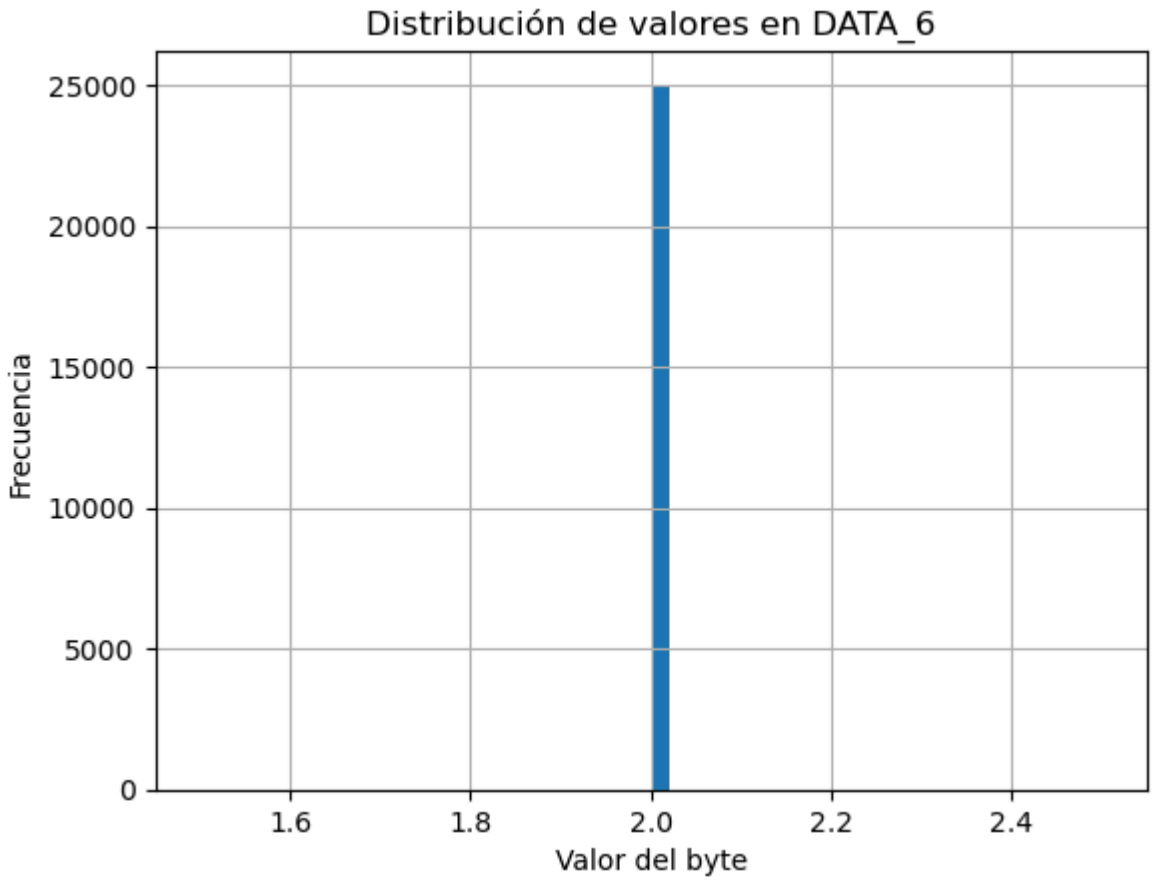

Distribución de valores en DATA_1
Frecuencia
Valor del byte
0
1000
2000
3000
4000
5000
0
5
10
15
20
25

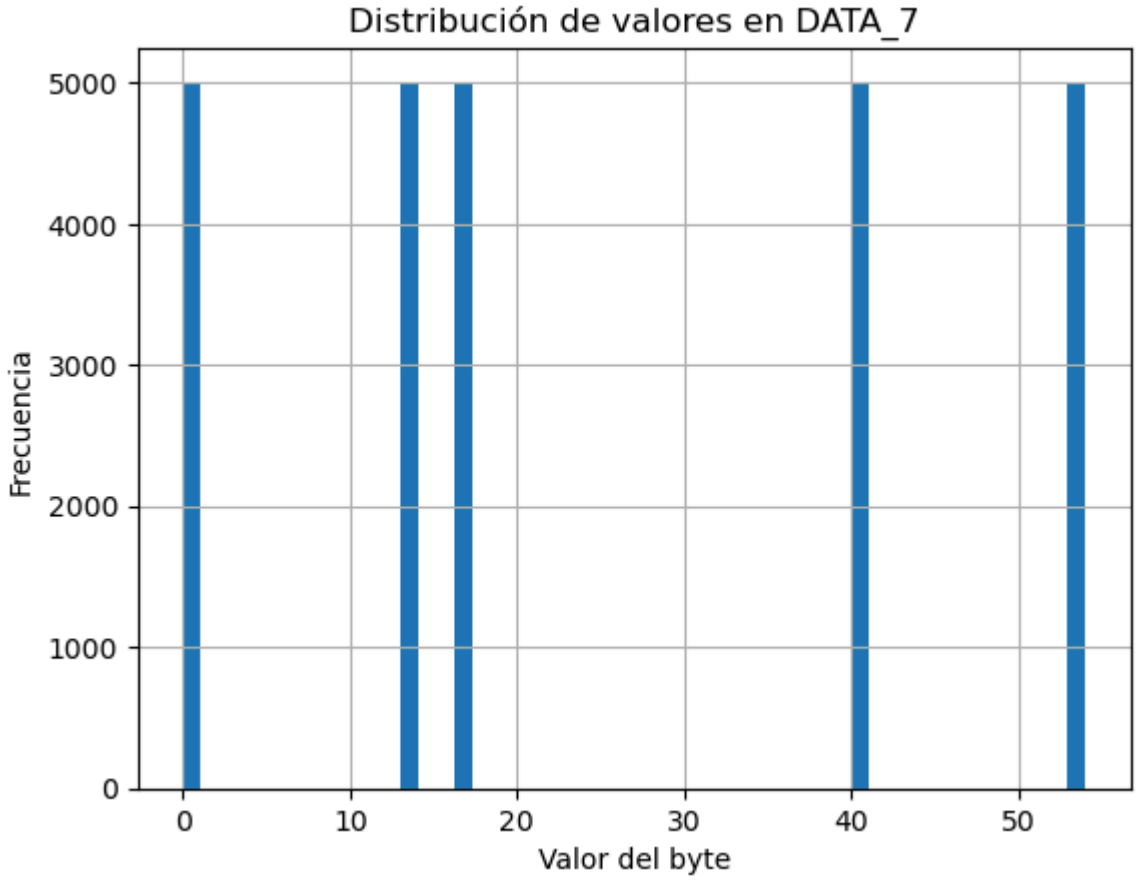

Distribución de valores en DATA_2
Frecuencia
Valor del byte
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1
2
3
4
5
6


Distribución de valores en DATA_3
Frecuencia
Valor del byte
0
1000
2000
3000
4000
5000
0
10
20
30
40
50
60


Distribución de valores en DATA_4
Frecuencia
Valor del byte
0
2500
5000
7500
10000
12500
15000
17500
20000
20
40
60
80
100


Distribución de valores en DATA_5
Frecuencia
Valor del byte
0
1000
2000
3000
4000
5000
0
10
20
30
40


Distribución de valores en DATA_6
Frecuencia
Valor del byte
0
5000
10000
15000
20000
25000
1.6
1.8
2.0
2.2
2.4



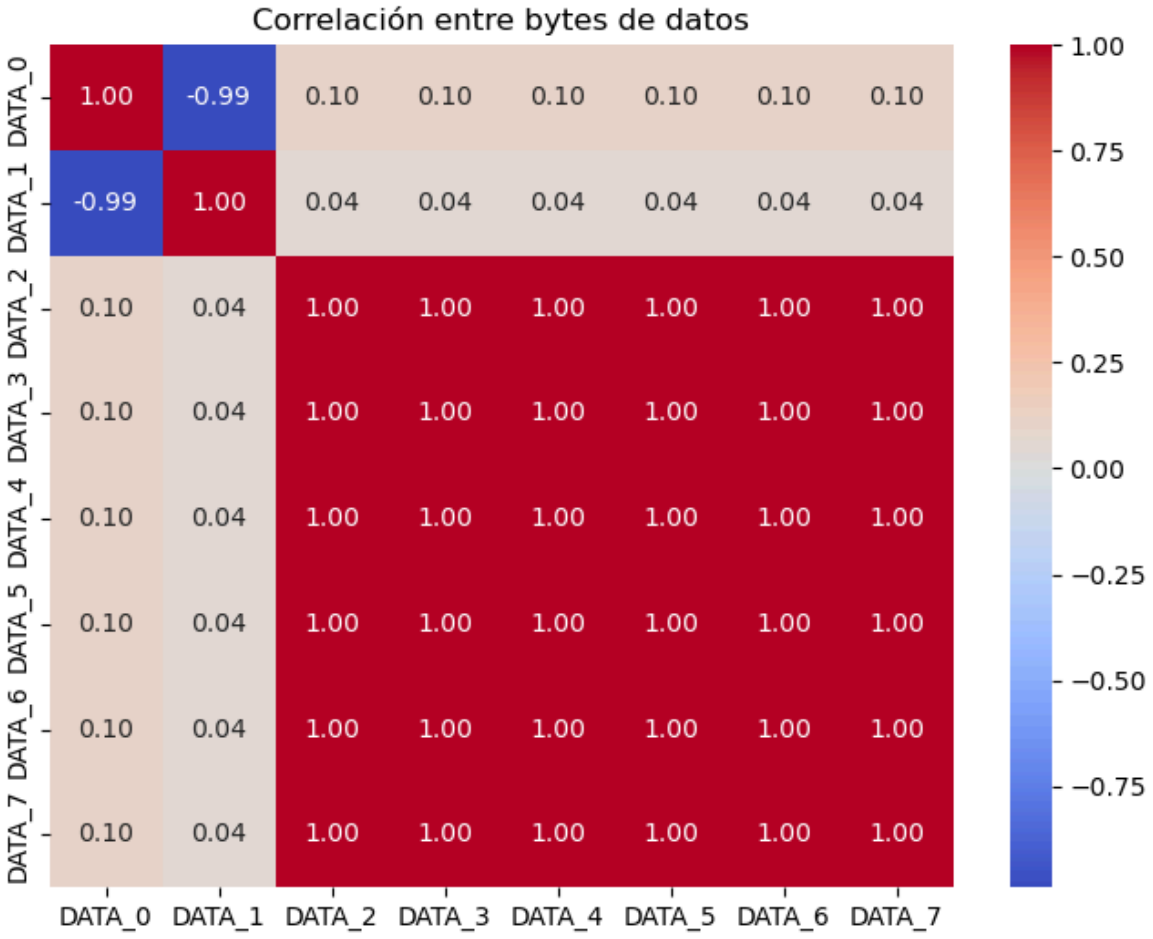

```
Valores fuera de rango (esperado 0–255):
DATA_0    0
DATA_1    0
DATA_2    0
DATA_3    0
DATA_4    0
DATA_5    0
DATA_6    0
DATA_7    0
dtype: int64
```

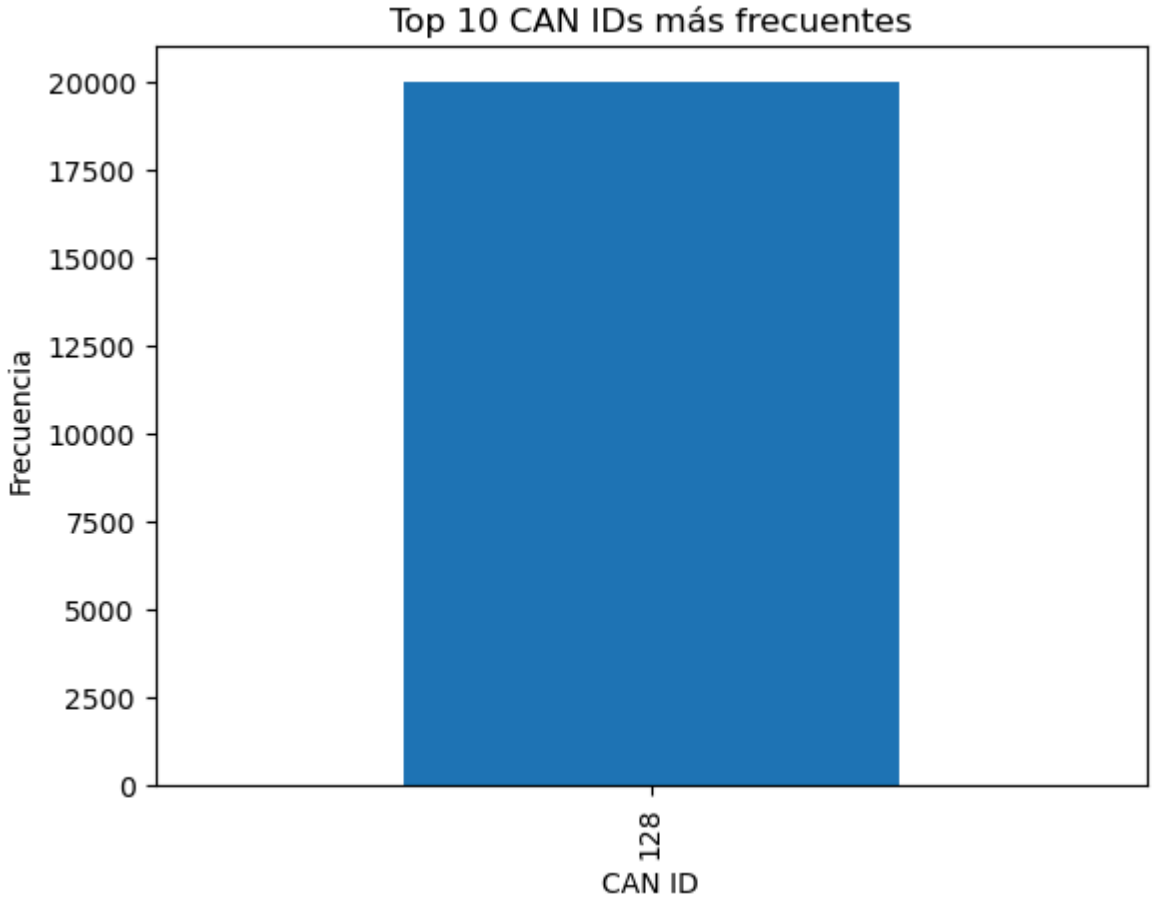
In [65]:
```
eda(raw_dec_wheel, 'dec_wheel')
```

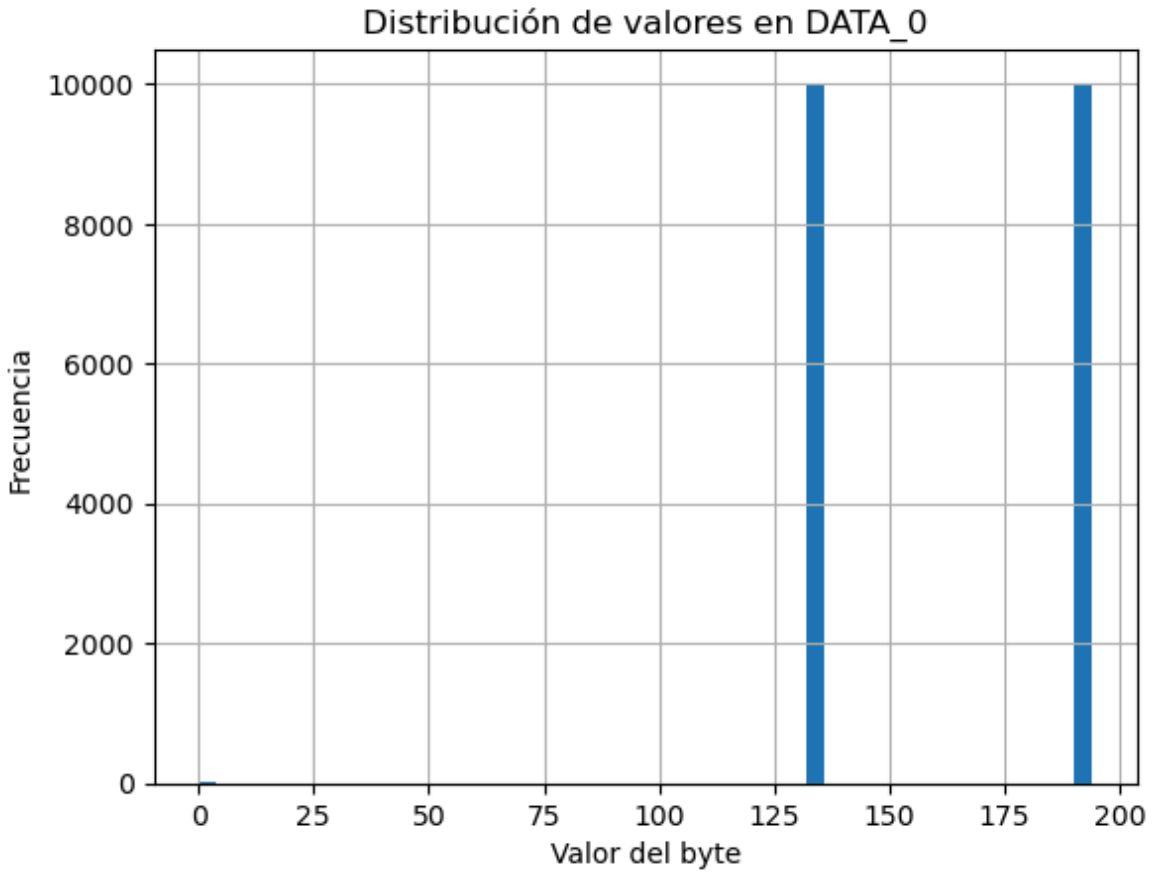
```
EDA de dec_wheel

Primeras filas:
    ID  DATA_0  DATA_1  DATA_2  DATA_3  DATA_4  DATA_5  DATA_6  DATA_7  \
0  128       0       0       0       0       0       0       0       0
1  128     132       3       2      35      24       5     138      34
2  128     132       3       2      35      24       5     138      34
3  128     132       3       2      35      24       5     138      34
4  128     132       3       2      35      24       5     138      34

    label  category  specific_class
0  ATTACK  SPOOFING  STEERING_WHEEL
1  ATTACK  SPOOFING  STEERING_WHEEL
2  ATTACK  SPOOFING  STEERING_WHEEL
3  ATTACK  SPOOFING  STEERING_WHEEL
4  ATTACK  SPOOFING  STEERING_WHEEL

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19977 entries, 0 to 19976
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ID              19977 non-null  int64
 1   DATA_0          19977 non-null  int64
 2   DATA_1          19977 non-null  int64
 3   DATA_2          19977 non-null  int64
 4   DATA_3          19977 non-null  int64
 5   DATA_4          19977 non-null  int64
 6   DATA_5          19977 non-null  int64
 7   DATA_6          19977 non-null  int64
 8   DATA_7          19977 non-null  int64
 9   label           19977 non-null  object
 10  category        19977 non-null  object
 11  specific_class  19977 non-null  object
dtypes: int64(9), object(3)
memory usage: 1.8+ MB
None

Filas enteras repetidas: 19974

Valores ausentes:
ID                0
DATA_0            0
DATA_1            0
DATA_2            0
DATA_3            0
DATA_4            0
DATA_5            0
DATA_6            0
DATA_7            0
label             0
category          0
specific_class    0
dtype: int64

Cantidad de CAN IDs únicos: 1
Top 10 CAN IDs más frecuentes:
ID
128    19977
Name: count, dtype: int64
```

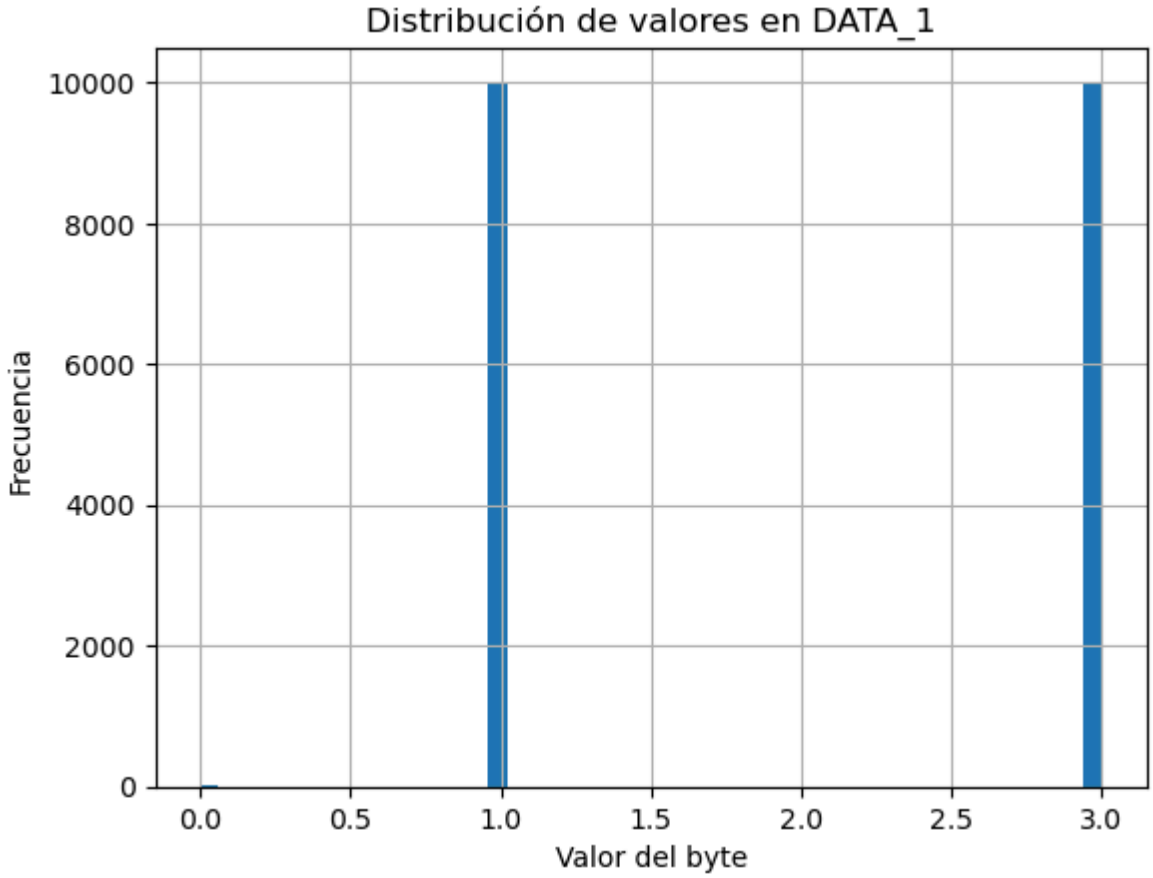
```
Estadísticas descriptivas:
            ID        DATA_0        DATA_1        DATA_2        DATA_3  \
count  19977.0  19977.000000  19977.000000  19977.000000  19977.000000
mean     128.0    162.917655      1.999750      1.999199     34.985984
std        0.0     31.165635      1.000626      0.040016      0.700280
min      128.0      0.000000      0.000000      0.000000      0.000000
25%      128.0    132.000000      1.000000      2.000000     35.000000
50%      128.0    132.000000      3.000000      2.000000     35.000000
75%      128.0    194.000000      3.000000      2.000000     35.000000
max      128.0    194.000000      3.000000      2.000000     35.000000

             DATA_4        DATA_5        DATA_6        DATA_7
count  19977.000000  19977.000000  19977.000000  19977.000000
mean      23.990389      4.997998    137.944736     33.986384
std        0.480192      0.100040      2.761104      0.680272
min        0.000000      0.000000      0.000000      0.000000
25%       24.000000      5.000000    138.000000     34.000000
50%       24.000000      5.000000    138.000000     34.000000
75%       24.000000      5.000000    138.000000     34.000000
max       24.000000      5.000000    138.000000     34.000000

Posibles valores atípicos en los bytes de datos:
             DATA_0        DATA_1        DATA_2        DATA_3        DATA_
4  \
count  19977.000000  19977.000000  19977.000000  19977.000000  19977.00000
0
mean     162.917655      1.999750      1.999199     34.985984     23.99038
9
std       31.165635      1.000626      0.040016      0.700280      0.48019
2
min        0.000000      0.000000      0.000000      0.000000      0.00000
0
1%       132.000000      1.000000      2.000000     35.000000     24.00000
0
25%      132.000000      1.000000      2.000000     35.000000     24.00000
0
50%      132.000000      3.000000      2.000000     35.000000     24.00000
0
75%      194.000000      3.000000      2.000000     35.000000     24.00000
0
99%      194.000000      3.000000      2.000000     35.000000     24.00000
0
max      194.000000      3.000000      2.000000     35.000000     24.00000
0

             DATA_5        DATA_6        DATA_7
count  19977.000000  19977.000000  19977.000000
mean       4.997998    137.944736     33.986384
std        0.100040      2.761104      0.680272
min        0.000000      0.000000      0.000000
1%         5.000000    138.000000     34.000000
25%        5.000000    138.000000     34.000000
50%        5.000000    138.000000     34.000000
75%        5.000000    138.000000     34.000000
99%        5.000000    138.000000     34.000000
max        5.000000    138.000000     34.000000

Distribución de etiquetas:
label
```

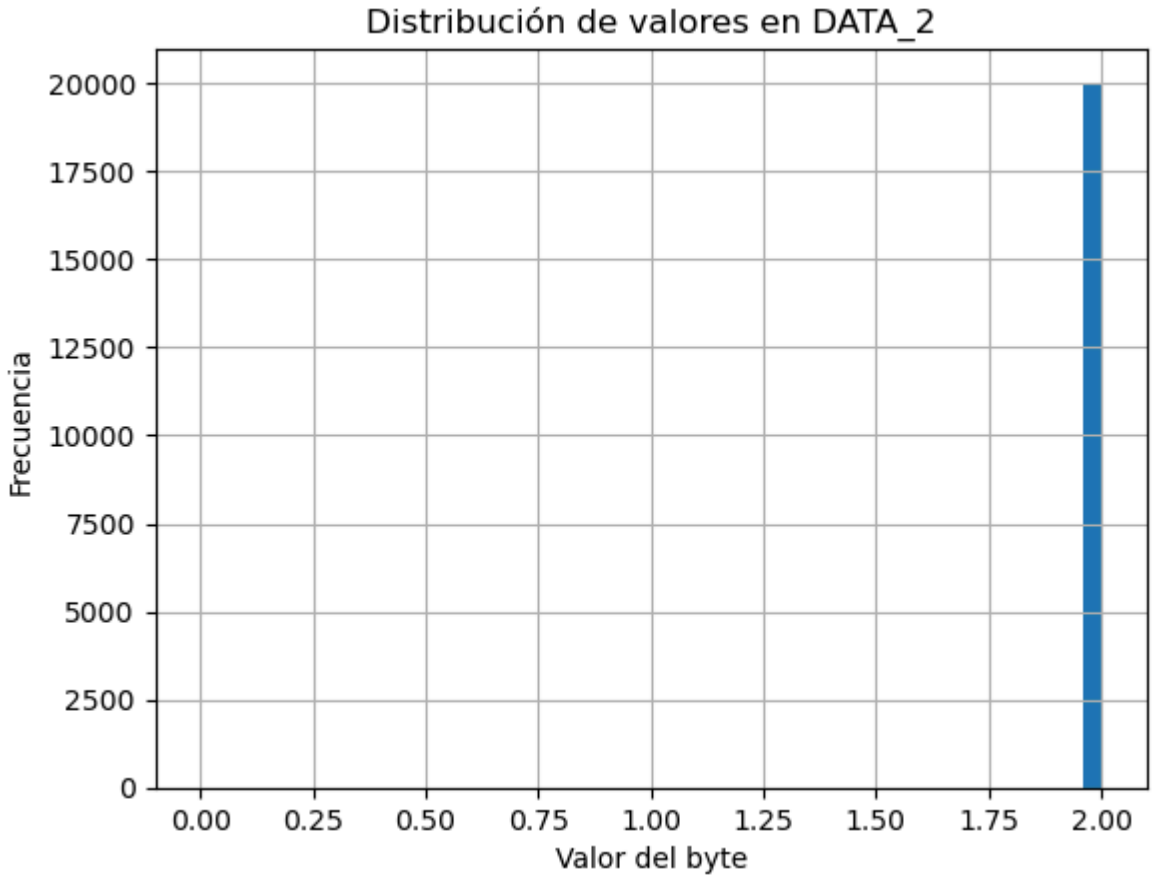
```
ATTACK    19977
Name: count, dtype: int64

Distribución de categorías:
category
SPOOFING    19977
Name: count, dtype: int64

Distribución de clases específicas:
specific_class
STEERING_WHEEL    19977
Name: count, dtype: int64
```

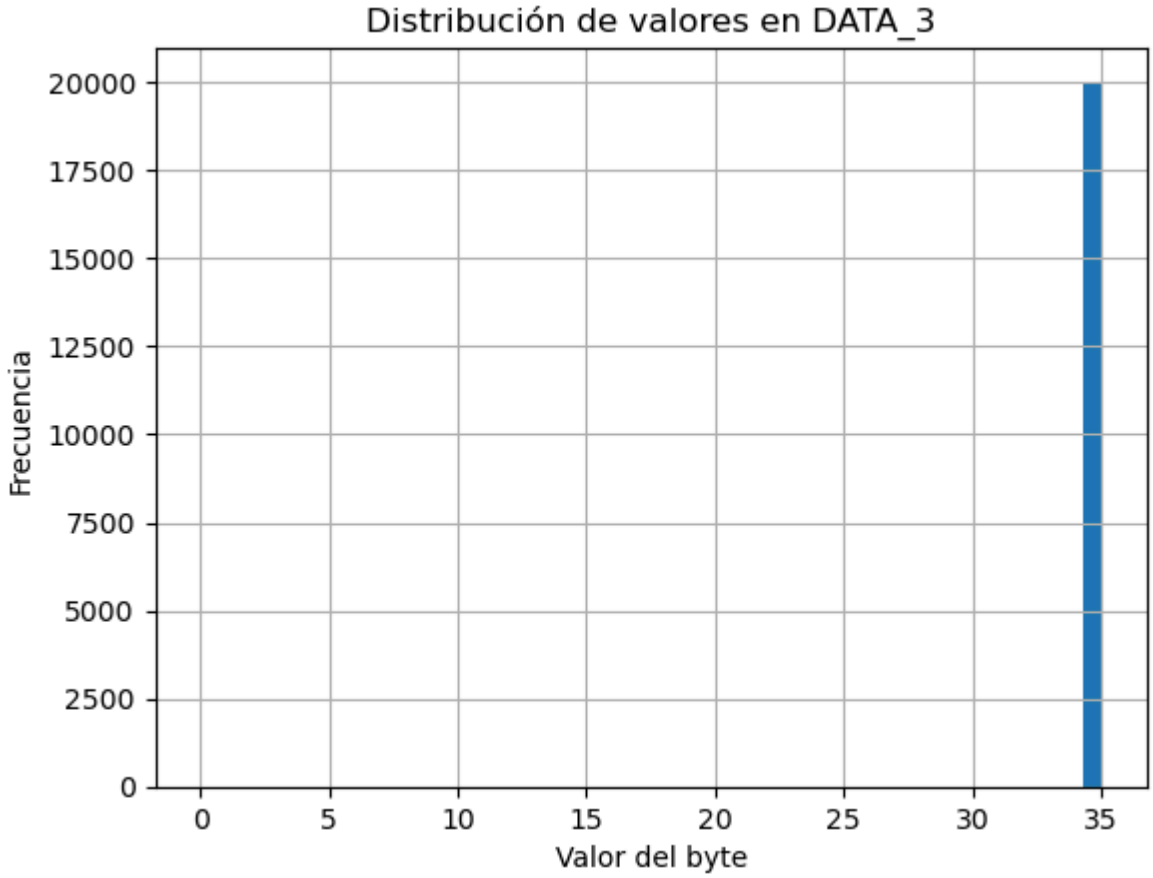


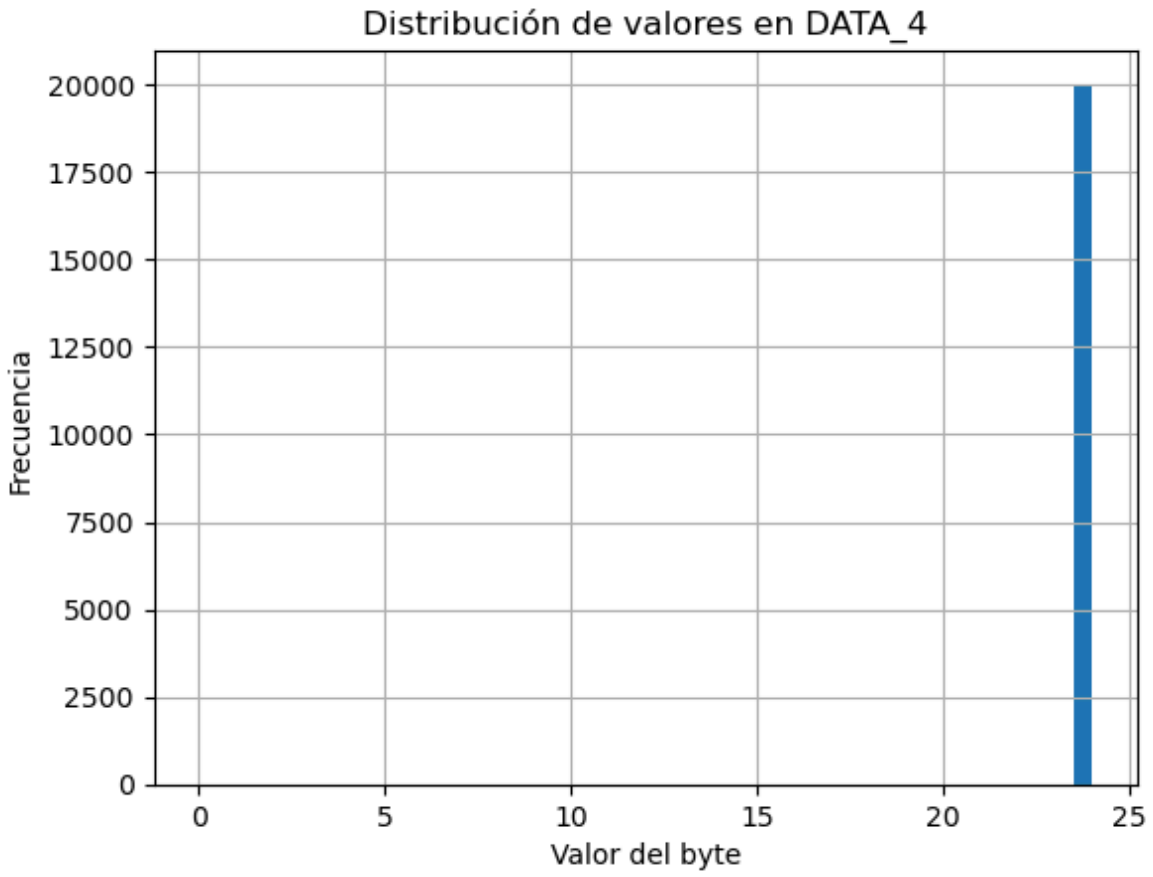

Top 10 CAN IDs más frecuentes
Frecuencia
20000
17500
15000
12500
10000
7500
5000
2500
0
128
CAN ID

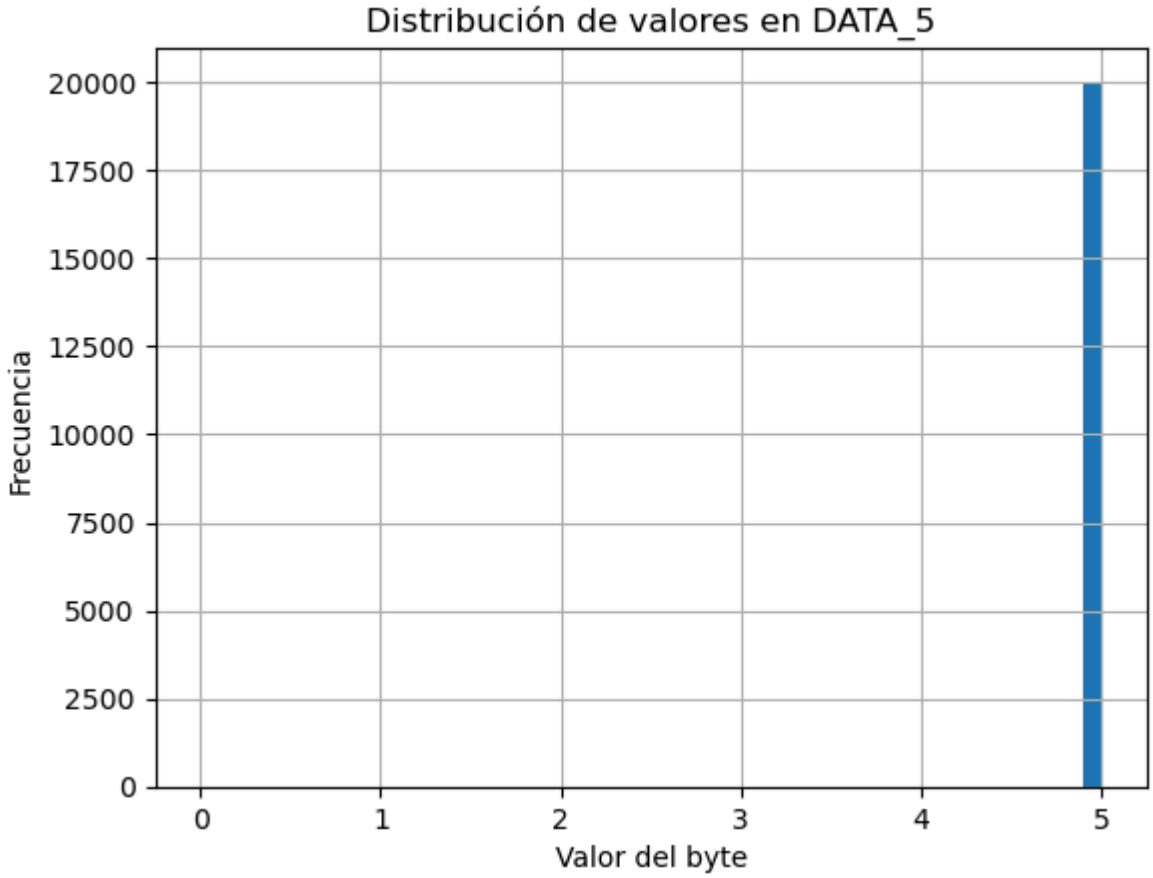

Distribución de valores en DATA_0
Frecuencia
10000
8000
6000
4000
2000
0
0
25
50
75
100
125
150
175
200
Valor del byte

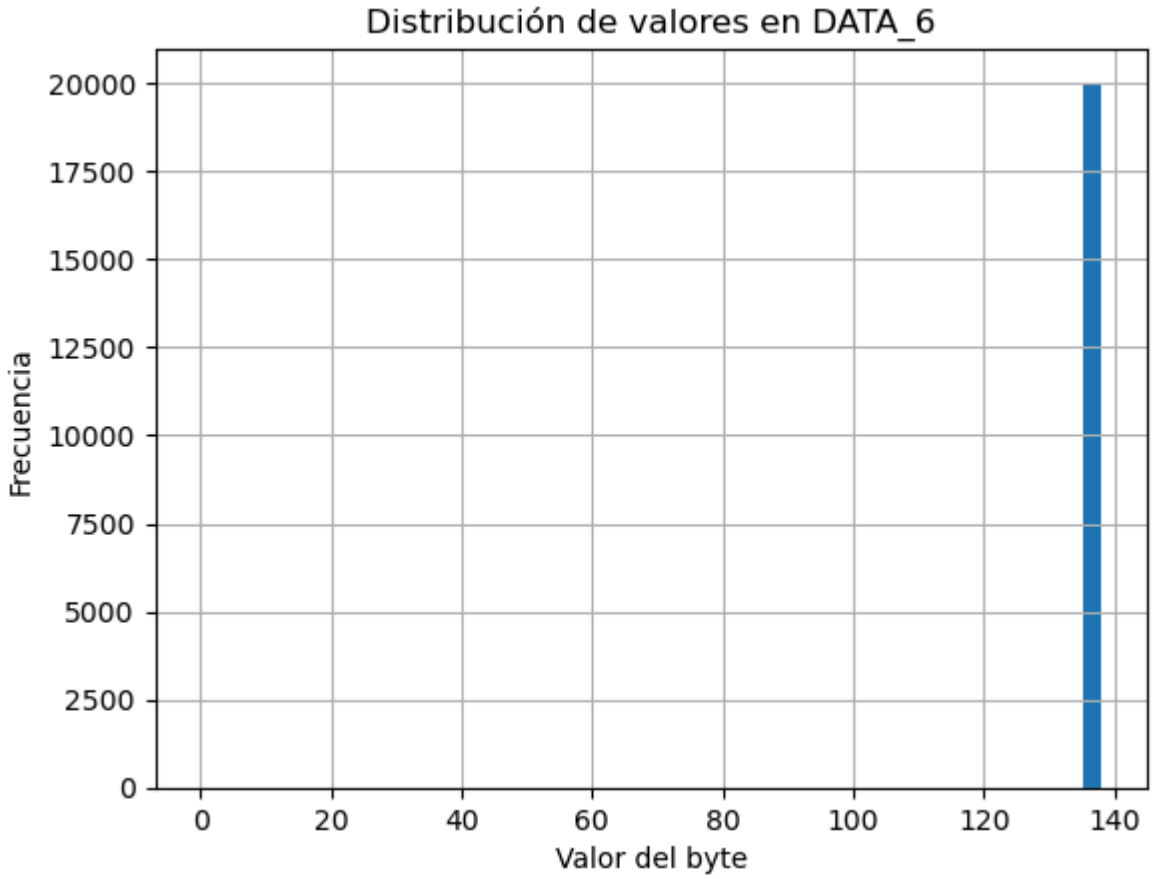

Distribución de valores en DATA_1
Frecuencia
10000
8000
6000
4000
2000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Valor del byte

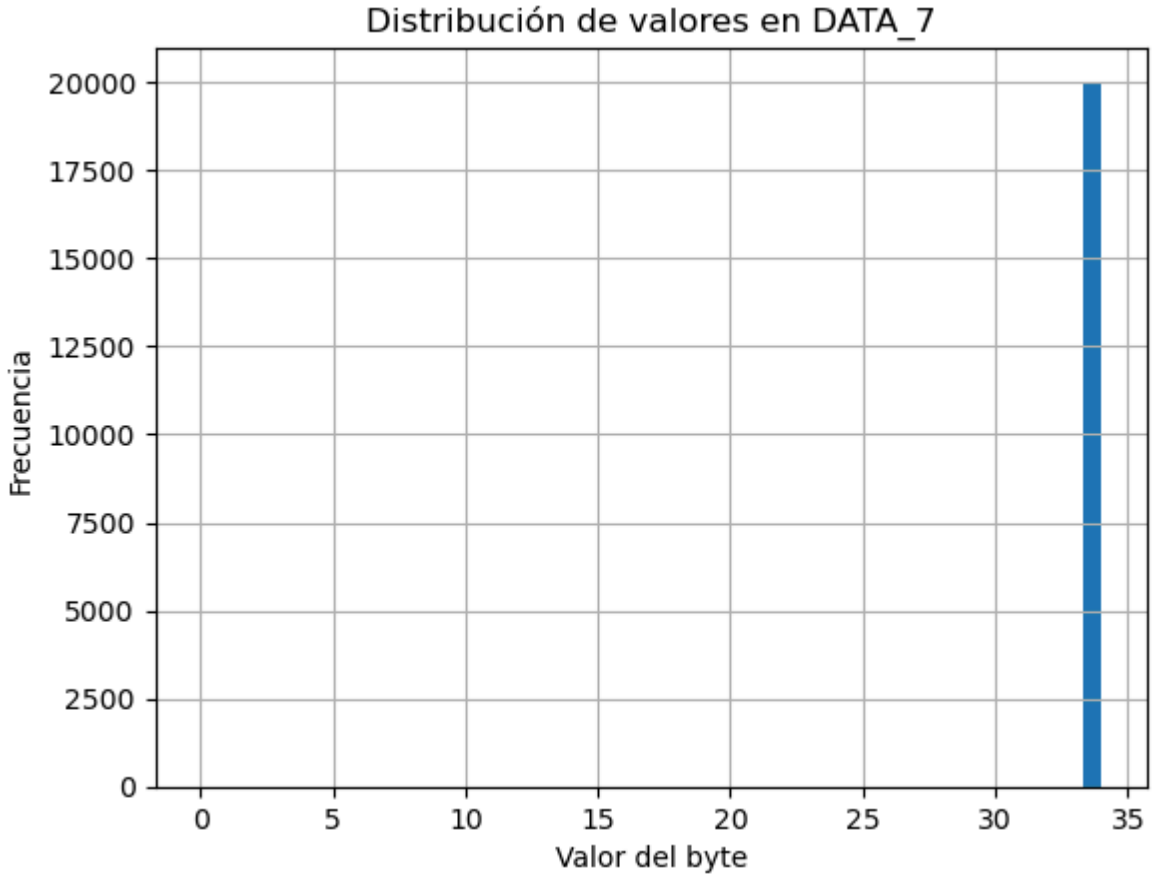

Distribución de valores en DATA_2
Frecuencia
20000
17500
15000
12500
10000
7500
5000
2500
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Valor del byte


Distribución de valores en DATA_3
Frecuencia
Valor del byte
0
2500
5000
7500
10000
12500
15000
17500
20000
0
5
10
15
20
25
30
35


Distribución de valores en DATA_4
Frecuencia
Valor del byte
0
2500
5000
7500
10000
12500
15000
17500
20000
0
5
10
15
20
25


Distribución de valores en DATA_5
Frecuencia
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
Valor del byte


Distribución de valores en DATA_6
Frecuencia
20000
17500
15000
12500
10000
7500
5000
2500
0
0
20
40
60
80
100
120
140
Valor del byte




```
Valores fuera de rango (esperado 0–255):
DATA_0    0
DATA_1    0
DATA_2    0
DATA_3    0
DATA_4    0
DATA_5    0
DATA_6    0
DATA_7    0
dtype: int64
```

In [ ]:
```
#Why are there so many duplicate rows in the dataset?
#The data was collected directly from the vehicle and logged into the CSV
#However, it is important to consider that the CAN BUS packet structure i
```