Projeto3

November 16, 2022

1 Data Science Academy

2 Análise de Dados com Linguagem Python

2.1 Projeto 3

2.2 Análise e Limpeza de Dados de Telecomunicações

Não tenha pressa de chegar ao final. O aprendizado não está no final. O aprendizado está na jornada. Aproveite a jornada!

2.3 Pré-Requisitos

Recomendamos que você tenha concluído pelo menos os 5 primeiros capítulos do curso gratuito de Python Fundamentos Para Análise de Dados.

2.4 Instalando e Carregando os Pacotes

[1]:
```
# Versão da Linguagem Python
from platform import python_version
print('Versão da Linguagem Python Usada Neste Jupyter Notebook:',␣
 ↪python_version())
```

```
Versão da Linguagem Python Usada Neste Jupyter Notebook: 3.8.8
```

[2]:
```
# Para atualizar um pacote, execute o comando abaixo no terminal ou prompt de␣
 ↪comando:
# pip install -U nome_pacote

# Para instalar a versão exata de um pacote, execute o comando abaixo no␣
 ↪terminal ou prompt de comando:
# !pip install nome_pacote==versão_desejada

# Depois de instalar ou atualizar o pacote, reinicie o jupyter notebook.

# Instala o pacote watermark.
# Esse pacote é usado para gravar as versões de outros pacotes usados neste␣
 ↪jupyter notebook.
# !pip install -q -U watermark
```

[3]:
```
# Imports
import math
import sys, os
import numpy as np
import pandas as pd
```

[4]:
```
# Vamos adicionar caminho para os módulos Python
sys.path.append(os.path.abspath(os.path.join('modulos')))
from estrategia1 import *
from estrategia2 import *
from estrategia3 import *
```

[5]:
```
pd.set_option('display.max_columns', 100)
```

[6]:
```
# Versões dos pacotes usados neste jupyter notebook
%reload_ext watermark
%watermark -a "Data Science Academy" --iversions
```

```
Author: Data Science Academy

pandas: 1.3.3
numpy : 1.21.0
sys   : 3.8.8 (default, Apr 13 2021, 12:59:45)
[Clang 10.0.0 ]
```

2.5 Carregando os Dados

https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

[7]:
```
# Criamos uma lista para identificar valores ausentes
lista_labels_valores_ausentes = ["n/a", "na", "undefined"]
```

[8]:
```
# Carrega o dataset
dataset = pd.read_csv("dados/dataset.csv", na_values =␣
 ↪lista_labels_valores_ausentes)
```

[9]:
```
# Shape
dataset.shape
```

[9]:
```
(150001, 55)
```

[10]:
```
# Amostra de dados
dataset.head()
```

[10]:

	Bearer Id	Start	Start ms	End	End ms	\
0	1.311448e+19	4/4/2019 12:01	770.0	4/25/2019 14:35	662.0	
1	1.311448e+19	4/9/2019 13:04	235.0	4/25/2019 8:15	606.0	
2	1.311448e+19	4/9/2019 17:42	1.0	4/25/2019 11:58	652.0	
3	1.311448e+19	4/10/2019 0:31	486.0	4/25/2019 7:36	171.0	
4	1.311448e+19	4/12/2019 20:10	565.0	4/25/2019 10:40	954.0	

	Dur. (ms)	IMSI	MSISDN/Number	IMEI	\
0	1823652.0	2.082014e+14	3.366496e+10	3.552121e+13	
1	1365104.0	2.082019e+14	3.368185e+10	3.579401e+13	
2	1361762.0	2.082003e+14	3.376063e+10	3.528151e+13	
3	1321509.0	2.082014e+14	3.375034e+10	3.535661e+13	
4	1089009.0	2.082014e+14	3.369980e+10	3.540701e+13	

	Last Location Name	Avg RTT DL (ms)	Avg RTT UL (ms)	\
0	9.16456699548519E+015	42.0	5.0	
1	L77566A	65.0	5.0	
2	D42335A	NaN	NaN	
3	T21824A	NaN	NaN	
4	D88865A	NaN	NaN	

	Avg Bearer TP DL (kbps)	Avg Bearer TP UL (kbps)	\
0	23.0	44.0	
1	16.0	26.0	
2	6.0	9.0	
3	44.0	44.0	
4	6.0	9.0	

	TCP DL Retrans. Vol (Bytes)	TCP UL Retrans. Vol (Bytes)	\

```
0                          NaN                           NaN
1                          NaN                           NaN
2                          NaN                           NaN
3                          NaN                           NaN
4                          NaN                           NaN

   DL TP < 50 Kbps (%)  50 Kbps < DL TP < 250 Kbps (%)  \
0                100.0                             0.0
1                100.0                             0.0
2                100.0                             0.0
3                100.0                             0.0
4                100.0                             0.0

   250 Kbps < DL TP < 1 Mbps (%)  DL TP > 1 Mbps (%)  UL TP < 10 Kbps (%)  \
0                            0.0                0.0                100.0
1                            0.0                0.0                100.0
2                            0.0                0.0                100.0
3                            0.0                0.0                100.0
4                            0.0                0.0                100.0

   10 Kbps < UL TP < 50 Kbps (%)  50 Kbps < UL TP < 300 Kbps (%)  \
0                            0.0                             0.0
1                            0.0                             0.0
2                            0.0                             0.0
3                            0.0                             0.0
4                            0.0                             0.0

   UL TP > 300 Kbps (%)  HTTP DL (Bytes)  HTTP UL (Bytes)  \
0                   0.0              NaN              NaN
1                   0.0              NaN              NaN
2                   0.0              NaN              NaN
3                   0.0              NaN              NaN
4                   0.0              NaN              NaN

   Activity Duration DL (ms)  Activity Duration UL (ms)   Dur. (ms).1  \
0                    37624.0                    38787.0  1.823653e+09
1                      168.0                     3560.0  1.365104e+09
2                        0.0                        0.0  1.361763e+09
3                     3330.0                    37882.0  1.321510e+09
4                        0.0                        0.0  1.089009e+09

  Handset Manufacturer                 Handset Type  \
0              Samsung   Samsung Galaxy A5 Sm-A520F
1              Samsung  Samsung Galaxy J5 (Sm-J530)
2              Samsung     Samsung Galaxy A8 (2018)
3                  NaN                          NaN
4              Samsung             Samsung Sm-G390F
```

```
   Nb of sec with 125000B < Vol DL  Nb of sec with 1250B < Vol UL < 6250B  \
0                              NaN                                    NaN
1                              NaN                                    NaN
2                              NaN                                    NaN
3                              NaN                                    NaN
4                              NaN                                    NaN

   Nb of sec with 31250B < Vol DL < 125000B  Nb of sec with 37500B < Vol UL  \
0                                       NaN                            NaN
1                                       NaN                            NaN
2                                       NaN                            NaN
3                                       NaN                            NaN
4                                       NaN                            NaN

   Nb of sec with 6250B < Vol DL < 31250B  \
0                                     NaN
1                                     NaN
2                                     NaN
3                                     NaN
4                                     NaN

   Nb of sec with 6250B < Vol UL < 37500B  Nb of sec with Vol DL < 6250B  \
0                                     NaN                          213.0
1                                     NaN                          971.0
2                                     NaN                          751.0
3                                     NaN                           17.0
4                                     NaN                          607.0

   Nb of sec with Vol UL < 1250B  Social Media DL (Bytes)  \
0                          214.0                1545765.0
1                         1022.0                1926113.0
2                          695.0                1684053.0
3                          207.0                 644121.0
4                          604.0                 862600.0

   Social Media UL (Bytes)  Google DL (Bytes)  Google UL (Bytes)  \
0                  24420.0          1634479.0          1271433.0
1                   7165.0          3493924.0           920172.0
2                  42224.0          8535055.0          1694064.0
3                  13372.0          9023734.0          2788027.0
4                  50188.0          6248284.0          1500559.0

   Email DL (Bytes)  Email UL (Bytes)  Youtube DL (Bytes)  Youtube UL (Bytes)  \
0         3563542.0          137762.0           15854611.0           2501332.0
1          629046.0          308339.0           20247395.0          19111729.0
2         2690151.0          672973.0           19725661.0          14699576.0
```

```
3         1439754.0          631229.0         21388122.0          15146643.0
4         1936496.0          173853.0         15259380.0          18962873.0

   Netflix DL (Bytes)  Netflix UL (Bytes)  Gaming DL (Bytes)  \
0           8198936.0           9656251.0        278082303.0
1          18338413.0          17227132.0        608750074.0
2          17587794.0           6163408.0        229584621.0
3          13994646.0           1097942.0        799538153.0
4          17124581.0            415218.0        527707248.0

   Gaming UL (Bytes)  Other DL (Bytes)  Other UL (Bytes)  Total UL (Bytes)  \
0         14344150.0       171744450.0         8814393.0        36749741.0
1          1170709.0       526904238.0        15055145.0        53800391.0
2           395630.0       410692588.0         4215763.0        27883638.0
3         10849722.0       749039933.0        12797283.0        43324218.0
4          3529801.0       550709500.0        13910322.0        38542814.0

   Total DL (Bytes)
0       308879636.0
1       653384965.0
2       279807335.0
3       846028530.0
4       569138589.0
```

[11]:
```
# Carregando o dicionário de dados
dicionario = pd.read_excel("dados/Dicionario.xlsx")
```

[12]:
```
# Shape
dicionario.shape
```

[12]:
```
(56, 2)
```

[13]:
```
# Amostra de dados
dicionario.head()
```

[13]:
```
      Fields                                        Description
0  bearer id                          xDr session identifier
1  Dur. (ms)                 Total Duration of the xDR (in ms)
2      Start      Start time of the xDR (first frame timestamp)
3   Start ms  Milliseconds offset of start time for the xDR …
4        End         End time of the xDR (last frame timestamp)
```

2.6 Análise Exploratória

[14]:
```
# Info
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150001 entries, 0 to 150000
Data columns (total 55 columns):
 #   Column                                    Non-Null Count   Dtype
---  ------                                    --------------   -----
 0   Bearer Id                                 149010 non-null  float64
 1   Start                                     150000 non-null  object
 2   Start ms                                  150000 non-null  float64
 3   End                                       150000 non-null  object
 4   End ms                                    150000 non-null  float64
 5   Dur. (ms)                                 150000 non-null  float64
 6   IMSI                                      149431 non-null  float64
 7   MSISDN/Number                             148935 non-null  float64
 8   IMEI                                      149429 non-null  float64
 9   Last Location Name                        148848 non-null  object
 10  Avg RTT DL (ms)                           122172 non-null  float64
 11  Avg RTT UL (ms)                           122189 non-null  float64
 12  Avg Bearer TP DL (kbps)                   150000 non-null  float64
 13  Avg Bearer TP UL (kbps)                   150000 non-null  float64
 14  TCP DL Retrans. Vol (Bytes)               61855 non-null   float64
 15  TCP UL Retrans. Vol (Bytes)               53352 non-null   float64
 16  DL TP < 50 Kbps (%)                       149247 non-null  float64
 17  50 Kbps < DL TP < 250 Kbps (%)            149247 non-null  float64
 18  250 Kbps < DL TP < 1 Mbps (%)             149247 non-null  float64
 19  DL TP > 1 Mbps (%)                        149247 non-null  float64
 20  UL TP < 10 Kbps (%)                       149209 non-null  float64
 21  10 Kbps < UL TP < 50 Kbps (%)             149209 non-null  float64
 22  50 Kbps < UL TP < 300 Kbps (%)            149209 non-null  float64
 23  UL TP > 300 Kbps (%)                      149209 non-null  float64
 24  HTTP DL (Bytes)                           68527 non-null   float64
 25  HTTP UL (Bytes)                           68191 non-null   float64
 26  Activity Duration DL (ms)                 150000 non-null  float64
 27  Activity Duration UL (ms)                 150000 non-null  float64
 28  Dur. (ms).1                               150000 non-null  float64
 29  Handset Manufacturer                      140442 non-null  object
 30  Handset Type                              140442 non-null  object
 31  Nb of sec with 125000B < Vol DL           52463 non-null   float64
 32  Nb of sec with 1250B < Vol UL < 6250B     57107 non-null   float64
 33  Nb of sec with 31250B < Vol DL < 125000B  56415 non-null   float64
 34  Nb of sec with 37500B < Vol UL            19747 non-null   float64
 35  Nb of sec with 6250B < Vol DL < 31250B    61684 non-null   float64
 36  Nb of sec with 6250B < Vol UL < 37500B    38158 non-null   float64
 37  Nb of sec with Vol DL < 6250B             149246 non-null  float64
 38  Nb of sec with Vol UL < 1250B             149208 non-null  float64
 39  Social Media DL (Bytes)                   150001 non-null  float64
 40  Social Media UL (Bytes)                   150001 non-null  float64
 41  Google DL (Bytes)                         150001 non-null  float64
 42  Google UL (Bytes)                         150001 non-null  float64
```

```
 43  Email DL (Bytes)                         150001 non-null  float64
 44  Email UL (Bytes)                         150001 non-null  float64
 45  Youtube DL (Bytes)                       150001 non-null  float64
 46  Youtube UL (Bytes)                       150001 non-null  float64
 47  Netflix DL (Bytes)                       150001 non-null  float64
 48  Netflix UL (Bytes)                       150001 non-null  float64
 49  Gaming DL (Bytes)                        150001 non-null  float64
 50  Gaming UL (Bytes)                        150001 non-null  float64
 51  Other DL (Bytes)                         150001 non-null  float64
 52  Other UL (Bytes)                         150001 non-null  float64
 53  Total UL (Bytes)                         150000 non-null  float64
 54  Total DL (Bytes)                         150000 non-null  float64
dtypes: float64(50), object(5)
memory usage: 62.9+ MB
```

[15]:
```
# Estatísticas descritivas
dataset.describe()
```

[15]:

	Bearer Id	Start ms	End ms	Dur. (ms)	IMSI
count	1.490100e+05	150000.000000	150000.000000	1.500000e+05	1.494310e+05
mean	1.013887e+19	499.188200	498.800880	1.046086e+05	2.082016e+14
std	2.893173e+18	288.611834	288.097653	8.103762e+04	2.148809e+10
min	6.917538e+18	0.000000	0.000000	7.142000e+03	2.040471e+14
25%	7.349883e+18	250.000000	251.000000	5.744050e+04	2.082014e+14
50%	7.349883e+18	499.000000	500.000000	8.639900e+04	2.082015e+14
75%	1.304243e+19	749.000000	750.000000	1.324302e+05	2.082018e+14
max	1.318654e+19	999.000000	999.000000	1.859336e+06	2.140743e+14

	MSISDN/Number	IMEI	Avg RTT DL (ms)	Avg RTT UL (ms)
count	1.489350e+05	1.494290e+05	122172.000000	122189.000000
mean	4.188282e+10	4.847455e+13	109.795706	17.662883
std	2.447443e+12	2.241637e+13	619.782739	84.793524
min	3.360100e+10	4.400152e+11	0.000000	0.000000
25%	3.365130e+10	3.546071e+13	32.000000	2.000000
50%	3.366371e+10	3.572201e+13	45.000000	5.000000
75%	3.368349e+10	8.611970e+13	70.000000	15.000000
max	8.823971e+14	9.900120e+13	96923.000000	7120.000000

	Avg Bearer TP DL (kbps)	Avg Bearer TP UL (kbps)
count	150000.000000	150000.000000
mean	13300.045927	1770.428647
std	23971.878541	4625.355500
min	0.000000	0.000000
25%	43.000000	47.000000
50%	63.000000	63.000000
75%	19710.750000	1120.000000
max	378160.000000	58613.000000

```
       TCP DL Retrans. Vol (Bytes)  TCP UL Retrans. Vol (Bytes)  \
count                 6.185500e+04                 5.335200e+04
mean                  2.080991e+07                 7.596587e+05
std                   1.825665e+08                 2.645305e+07
min                   2.000000e+00                 1.000000e+00
25%                   3.565150e+04                 4.694750e+03
50%                   5.687300e+05                 2.094950e+04
75%                   3.768308e+06                 8.402025e+04
max                   4.294426e+09                 2.908226e+09

       DL TP < 50 Kbps (%)  50 Kbps < DL TP < 250 Kbps (%)  \
count        149247.000000                   149247.000000
mean             92.844754                        3.069355
std              13.038031                        6.215233
min               0.000000                        0.000000
25%              91.000000                        0.000000
50%             100.000000                        0.000000
75%             100.000000                        4.000000
max             100.000000                       93.000000

       250 Kbps < DL TP < 1 Mbps (%)  DL TP > 1 Mbps (%)  UL TP < 10 Kbps (%)  \
count                 149247.000000       149247.000000        149209.000000
mean                       1.717341            1.609654            98.530142
std                        4.159538            4.828890             4.634285
min                        0.000000            0.000000             0.000000
25%                        0.000000            0.000000            99.000000
50%                        0.000000            0.000000           100.000000
75%                        1.000000            0.000000           100.000000
max                      100.000000           94.000000           100.000000

       10 Kbps < UL TP < 50 Kbps (%)  50 Kbps < UL TP < 300 Kbps (%)  \
count                 149209.000000                   149209.000000
mean                       0.776749                        0.147987
std                        3.225176                        1.624523
min                        0.000000                        0.000000
25%                        0.000000                        0.000000
50%                        0.000000                        0.000000
75%                        0.000000                        0.000000
max                       98.000000                      100.000000

       UL TP > 300 Kbps (%)  HTTP DL (Bytes)  HTTP UL (Bytes)  \
count         149209.000000     6.852700e+04     6.819100e+04
mean               0.078923     1.144710e+08     3.242301e+06
std                1.295396     9.631946e+08     1.957064e+07
min                0.000000     4.000000e+01     4.000000e+01
25%                0.000000     1.124035e+05     2.432200e+04
```

```
50%                0.000000     1.941949e+06     2.297330e+05
75%                0.000000     2.504290e+07     1.542827e+06
max               96.000000     7.253064e+10     1.491890e+09

       Activity Duration DL (ms)  Activity Duration UL (ms)   Dur. (ms).1  \
count               1.500000e+05               1.500000e+05  1.500000e+05
mean                1.829177e+06               1.408880e+06  1.046091e+08
std                 5.696395e+06               4.643231e+06  8.103761e+07
min                 0.000000e+00               0.000000e+00  7.142988e+06
25%                 1.487775e+04               2.153975e+04  5.744079e+07
50%                 3.930450e+04               4.679350e+04  8.639998e+07
75%                 6.796095e+05               5.990952e+05  1.324308e+08
max                 1.365365e+08               1.449113e+08  1.859336e+09

       Nb of sec with 125000B < Vol DL  Nb of sec with 1250B < Vol UL < 6250B  \
count                     52463.000000                           57107.000000
mean                        989.699998                             340.434395
std                        2546.524440                            1445.365032
min                           1.000000                               1.000000
25%                          20.000000                              10.000000
50%                         128.000000                              52.000000
75%                         693.500000                             203.000000
max                       81476.000000                           85412.000000

       Nb of sec with 31250B < Vol DL < 125000B  \
count                              56415.000000
mean                                 810.837401
std                                 1842.162008
min                                    1.000000
25%                                   26.000000
50%                                  164.000000
75%                                  757.000000
max                                58525.000000

       Nb of sec with 37500B < Vol UL  Nb of sec with 6250B < Vol DL < 31250B  \
count                     19747.000000                           61684.000000
mean                        149.257052                             965.464756
std                        1219.112287                            1946.387608
min                           1.000000                               1.000000
25%                           2.000000                              39.000000
50%                           8.000000                             288.000000
75%                          35.000000                            1092.000000
max                       50553.000000                           66913.000000

       Nb of sec with 6250B < Vol UL < 37500B  Nb of sec with Vol DL < 6250B  \
count                            38158.000000                  149246.000000
mean                               141.304812                    3719.787552
```

std	993.349688	9171.609010
min	1.000000	1.000000
25%	3.000000	87.000000
50%	8.000000	203.000000
75%	31.000000	2650.000000
max	49565.000000	604061.000000

	Nb of sec with Vol UL < 1250B	Social Media DL (Bytes)	\
count	149208.000000	1.500010e+05	
mean	4022.083454	1.795322e+06	
std	10160.324314	1.035482e+06	
min	1.000000	1.200000e+01	
25%	106.000000	8.991480e+05	
50%	217.000000	1.794369e+06	
75%	2451.000000	2.694938e+06	
max	604122.000000	3.586064e+06	

	Social Media UL (Bytes)	Google DL (Bytes)	Google UL (Bytes)	\
count	150001.000000	1.500010e+05	1.500010e+05	
mean	32928.434380	5.750753e+06	2.056542e+06	
std	19006.178256	3.309097e+06	1.189917e+06	
min	0.000000	2.070000e+02	3.000000e+00	
25%	16448.000000	2.882393e+06	1.024279e+06	
50%	32920.000000	5.765829e+06	2.054573e+06	
75%	49334.000000	8.623552e+06	3.088454e+06	
max	65870.000000	1.146283e+07	4.121357e+06	

	Email DL (Bytes)	Email UL (Bytes)	Youtube DL (Bytes)	\
count	1.500010e+05	150001.000000	1.500010e+05	
mean	1.791729e+06	467373.441940	1.163407e+07	
std	1.035840e+06	269969.307031	6.710569e+06	
min	1.400000e+01	2.000000	5.300000e+01	
25%	8.927930e+05	233383.000000	5.833501e+06	
50%	1.793505e+06	466250.000000	1.161602e+07	
75%	2.689327e+06	700440.000000	1.744852e+07	
max	3.586146e+06	936418.000000	2.325910e+07	

	Youtube UL (Bytes)	Netflix DL (Bytes)	Netflix UL (Bytes)	\
count	1.500010e+05	1.500010e+05	1.500010e+05	
mean	1.100941e+07	1.162685e+07	1.100175e+07	
std	6.345423e+06	6.725218e+06	6.359490e+06	
min	1.050000e+02	4.200000e+01	3.500000e+01	
25%	5.517965e+06	5.777156e+06	5.475981e+06	
50%	1.101345e+07	1.164222e+07	1.099638e+07	
75%	1.651556e+07	1.747048e+07	1.650727e+07	
max	2.201196e+07	2.325919e+07	2.201196e+07	

```
       Gaming DL (Bytes)  Gaming UL (Bytes)  Other DL (Bytes)  \
count       1.500010e+05       1.500010e+05      1.500010e+05
mean        4.220447e+08       8.288398e+06      4.211005e+08
std         2.439675e+08       4.782700e+06      2.432050e+08
min         2.516000e+03       5.900000e+01      3.290000e+03
25%         2.104733e+08       4.128476e+06      2.101869e+08
50%         4.234081e+08       8.291208e+06      4.218030e+08
75%         6.331742e+08       1.243162e+07      6.316918e+08
max         8.434419e+08       1.655879e+07      8.434425e+08

       Other UL (Bytes)  Total UL (Bytes)  Total DL (Bytes)
count      1.500010e+05      1.500000e+05      1.500000e+05
mean       8.264799e+06      4.112121e+07      4.546434e+08
std        4.769004e+06      1.127639e+07      2.441429e+08
min        1.480000e+02      2.866892e+06      7.114041e+06
25%        4.145943e+06      3.322201e+07      2.431068e+08
50%        8.267071e+06      4.114331e+07      4.558411e+08
75%        1.238415e+07      4.903424e+07      6.657055e+08
max        1.655882e+07      7.833131e+07      9.029696e+08
```

Não faz sentido calcular estatísticas descritivas para Beared Id, IMSI, MSISDN / Number e IMEI. Mas o método describe() calcula as estatísticas de todas as colunas numéricas. Essas estatísticas estão sendo calculadas antes que os dados sejam limpos. Portanto, pode haver mudanças depois que os valores ausentes e outliers são tratados.

[16]:
```
# Shape
dataset.shape
```

[16]: (150001, 55)

[17]:
```
# Shape
dicionario.shape
```

[17]: (56, 2)

Existem 150.001 linhas e 55 colunas no dataframe. No entanto, temos 56 colunas com seus nomes e descrições no dicionário. Isso significa que há uma coluna descrita, mas não incluída no dataframe. Vamos identificar qual é a coluna faltante.

[18]:
```
# Concatena os dataframes
df_compara_colunas = pd.concat([pd.Series(dataset.columns.tolist()),
                                dicionario['Fields']],
                               axis = 1)
```

[19]:
```
# Renomeia as colunas
df_compara_colunas.rename(columns = {0: 'Coluna no Dataset', 'Fields': 'Coluna␣
 ↪no Dicionario'},
                          inplace = True)
```

[20]:
```
# Visualiza
df_compara_colunas
```

[20]:

	Coluna no Dataset
0	Bearer Id
1	Start
2	Start ms
3	End
4	End ms
5	Dur. (ms)
6	IMSI
7	MSISDN/Number
8	IMEI
9	Last Location Name
10	Avg RTT DL (ms)
11	Avg RTT UL (ms)
12	Avg Bearer TP DL (kbps)
13	Avg Bearer TP UL (kbps)
14	TCP DL Retrans. Vol (Bytes)
15	TCP UL Retrans. Vol (Bytes)
16	DL TP < 50 Kbps (%)
17	50 Kbps < DL TP < 250 Kbps (%)
18	250 Kbps < DL TP < 1 Mbps (%)
19	DL TP > 1 Mbps (%)
20	UL TP < 10 Kbps (%)
21	10 Kbps < UL TP < 50 Kbps (%)
22	50 Kbps < UL TP < 300 Kbps (%)
23	UL TP > 300 Kbps (%)
24	HTTP DL (Bytes)
25	HTTP UL (Bytes)
26	Activity Duration DL (ms)
27	Activity Duration UL (ms)
28	Dur. (ms).1
29	Handset Manufacturer
30	Handset Type
31	Nb of sec with 125000B < Vol DL
32	Nb of sec with 1250B < Vol UL < 6250B
33	Nb of sec with 31250B < Vol DL < 125000B
34	Nb of sec with 37500B < Vol UL
35	Nb of sec with 6250B < Vol DL < 31250B
36	Nb of sec with 6250B < Vol UL < 37500B
37	Nb of sec with Vol DL < 6250B
38	Nb of sec with Vol UL < 1250B
39	Social Media DL (Bytes)
40	Social Media UL (Bytes)
41	Google DL (Bytes)
42	Google UL (Bytes)

43	Email DL (Bytes)
44	Email UL (Bytes)
45	Youtube DL (Bytes)
46	Youtube UL (Bytes)
47	Netflix DL (Bytes)
48	Netflix UL (Bytes)
49	Gaming DL (Bytes)
50	Gaming UL (Bytes)
51	Other DL (Bytes)
52	Other UL (Bytes)
53	Total UL (Bytes)
54	Total DL (Bytes)
55	NaN

	Coluna no Dicionario
0	bearer id
1	Dur. (ms)
2	Start
3	Start ms
4	End
5	End ms
6	Dur. (s)
7	IMSI
8	MSISDN/Number
9	IMEI
10	Last Location Name
11	Avg RTT DL (ms)
12	Avg RTT UL (ms)
13	Avg Bearer TP DL (kbps)
14	Avg Bearer TP UL (kbps)
15	TCP DL Retrans. Vol (Bytes)
16	TCP UL Retrans. Vol (Bytes)
17	DL TP < 50 Kbps (%)
18	50 Kbps < DL TP < 250 Kbps (%)
19	250 Kbps < DL TP < 1 Mbps (%)
20	DL TP > 1 Mbps (%)
21	UL TP < 10 Kbps (%)
22	10 Kbps < UL TP < 50 Kbps (%)
23	50 Kbps < UL TP < 300 Kbps (%)
24	UL TP > 300 Kbps (%)
25	HTTP DL (Bytes)
26	HTTP UL (Bytes)
27	Activity Duration DL (ms)
28	Activity Duration UL (ms)
29	Dur. (ms).1
30	Handset Manufacturer
31	Handset Type

```
32           Nb of sec with 125000B < Vol DL
33     Nb of sec with 1250B < Vol UL < 6250B
34  Nb of sec with 31250B < Vol DL < 125000B
35            Nb of sec with 37500B < Vol UL
36    Nb of sec with 6250B < Vol DL < 31250B
37    Nb of sec with 6250B < Vol UL < 37500B
38             Nb of sec with Vol DL < 6250B
39             Nb of sec with Vol UL < 1250B
40                   Social Media DL (Bytes)
41                   Social Media UL (Bytes)
42                        YouTube DL (Bytes)
43                        YouTube UL (Bytes)
44                        Netflix DL (Bytes)
45                        Netflix UL (Bytes)
46                         Google DL (Bytes)
47                         Google UL (Bytes)
48                          Email DL (Bytes)
49                          Email UL (Bytes)
50                         Gaming DL (Bytes)
51                         Gaming UL (Bytes)
52                                  Other DL
53                                  Other UL
54                          Total DL (Bytes)
55                          Total UL (Bytes)
```

"Dur. (Ms)" é ignorado no dataset como visto no índice 1 em **df_compara_colunas**. É aqui que a ordem das colunas começou a mudar.

Mas o mesmo nome de coluna "Dur. (Ms)" aparece no dataset no índice 5, enquanto o arquivo de dicionário nos diz que é "Dur. (S)" no índice 6. Como as medidas de ambas as colunas diferem conforme mostrado em seus nomes, nós precisamos verificar qual está certo. Para investigar isso, usaremos a coluna "Dur. (Ms) .1" que se encontra nos índices 28 e 29 no dataset e no arquivo de dicionário, respectivamente.

[21]:
```
dataset[['Dur. (ms)', 'Dur. (ms).1']]
```

[21]:
```
         Dur. (ms)   Dur. (ms).1
0        1823652.0  1.823653e+09
1        1365104.0  1.365104e+09
2        1361762.0  1.361763e+09
3        1321509.0  1.321510e+09
4        1089009.0  1.089009e+09
…              …            …
149996     81230.0  8.123076e+07
149997     97970.0  9.797070e+07
149998     98249.0  9.824953e+07
149999     97910.0  9.791063e+07
150000         NaN           NaN
```

```
[150001 rows x 2 columns]
```

Parece que a coluna “Dur. (Ms)” é medida em segundos. Portanto, vamos renomeá-la apropriadamente. Vamos também renomear algumas das colunas para que fiquem claras como sua descrição e sigam o estilo de nomenclatura de outras colunas.

[22]:
```
# Renomeia colunas
dataset.rename(columns = {'Dur. (ms)': 'Dur (s)',
                          'Dur. (ms).1': 'Dur (ms)',
                          'Start ms': 'Start Offset (ms)',
                          'End ms': 'End Offset (ms)'},
               inplace = True)
```

[23]:
```
# Lista de colunas do dataset
dataset.columns.tolist()
```

[23]:
```
['Bearer Id',
 'Start',
 'Start Offset (ms)',
 'End',
 'End Offset (ms)',
 'Dur (s)',
 'IMSI',
 'MSISDN/Number',
 'IMEI',
 'Last Location Name',
 'Avg RTT DL (ms)',
 'Avg RTT UL (ms)',
 'Avg Bearer TP DL (kbps)',
 'Avg Bearer TP UL (kbps)',
 'TCP DL Retrans. Vol (Bytes)',
 'TCP UL Retrans. Vol (Bytes)',
 'DL TP < 50 Kbps (%)',
 '50 Kbps < DL TP < 250 Kbps (%)',
 '250 Kbps < DL TP < 1 Mbps (%)',
 'DL TP > 1 Mbps (%)',
 'UL TP < 10 Kbps (%)',
 '10 Kbps < UL TP < 50 Kbps (%)',
 '50 Kbps < UL TP < 300 Kbps (%)',
 'UL TP > 300 Kbps (%)',
 'HTTP DL (Bytes)',
 'HTTP UL (Bytes)',
 'Activity Duration DL (ms)',
 'Activity Duration UL (ms)',
 'Dur (ms)',
 'Handset Manufacturer',
 'Handset Type',
```

```
 'Nb of sec with 125000B < Vol DL',
 'Nb of sec with 1250B < Vol UL < 6250B',
 'Nb of sec with 31250B < Vol DL < 125000B',
 'Nb of sec with 37500B < Vol UL',
 'Nb of sec with 6250B < Vol DL < 31250B',
 'Nb of sec with 6250B < Vol UL < 37500B',
 'Nb of sec with Vol DL < 6250B',
 'Nb of sec with Vol UL < 1250B',
 'Social Media DL (Bytes)',
 'Social Media UL (Bytes)',
 'Google DL (Bytes)',
 'Google UL (Bytes)',
 'Email DL (Bytes)',
 'Email UL (Bytes)',
 'Youtube DL (Bytes)',
 'Youtube UL (Bytes)',
 'Netflix DL (Bytes)',
 'Netflix UL (Bytes)',
 'Gaming DL (Bytes)',
 'Gaming UL (Bytes)',
 'Other DL (Bytes)',
 'Other UL (Bytes)',
 'Total UL (Bytes)',
 'Total DL (Bytes)']
```

2.7 Estratégia de Limpeza 1 - Tratamento de Valores Ausentes

- 1- Identificando Valores Ausentes
- 2- Drop de Colunas
- 3- Imputação com Preenchimento Reverso
- 4- Imputação com Preenchimento Progressivo
- 5- Imputação de Variáveis Categóricas
- 6- Drop de Linhas

[24]:
```
help(func_calc_percentual_valores_ausentes)
```

```
Help on function func_calc_percentual_valores_ausentes in module estrategia1:

func_calc_percentual_valores_ausentes(df)
    # Calcula o percentual de valores ausentes
```

[25]:
```
# Verifica o percentual de valores ausentes
# Função do módulo estratégia 1
func_calc_percentual_valores_ausentes(dataset)
```

```
O dataset tem 12.72 % de valores ausentes.
```

[26]:
```
# Cria tabela com valores ausentes
df_missing = func_calc_percentual_valores_ausentes_coluna(dataset)
```

```
O dataset tem 55 colunas.
Encontrado: 41 colunas que têm valores ausentes.
```

[27]:
```
# Visualiza
df_missing
```

[27]:

	Valores Ausentes	\
Nb of sec with 37500B < Vol UL	130254	
Nb of sec with 6250B < Vol UL < 37500B	111843	
Nb of sec with 125000B < Vol DL	97538	
TCP UL Retrans. Vol (Bytes)	96649	
Nb of sec with 31250B < Vol DL < 125000B	93586	
Nb of sec with 1250B < Vol UL < 6250B	92894	
Nb of sec with 6250B < Vol DL < 31250B	88317	
TCP DL Retrans. Vol (Bytes)	88146	
HTTP UL (Bytes)	81810	
HTTP DL (Bytes)	81474	
Avg RTT DL (ms)	27829	
Avg RTT UL (ms)	27812	
Handset Manufacturer	9559	
Handset Type	9559	
Last Location Name	1153	
MSISDN/Number	1066	
Bearer Id	991	
Nb of sec with Vol UL < 1250B	793	
10 Kbps < UL TP < 50 Kbps (%)	792	
UL TP > 300 Kbps (%)	792	
50 Kbps < UL TP < 300 Kbps (%)	792	
UL TP < 10 Kbps (%)	792	
Nb of sec with Vol DL < 6250B	755	
250 Kbps < DL TP < 1 Mbps (%)	754	
50 Kbps < DL TP < 250 Kbps (%)	754	
DL TP < 50 Kbps (%)	754	
DL TP > 1 Mbps (%)	754	
IMEI	572	
IMSI	570	
Start	1	
End Offset (ms)	1	
Total UL (Bytes)	1	
Start Offset (ms)	1	
End	1	
Avg Bearer TP DL (kbps)	1	
Dur (s)	1	
Avg Bearer TP UL (kbps)	1	
Dur (ms)	1	

Activity Duration UL (ms)	1
Activity Duration DL (ms)	1
Total DL (Bytes)	1

	% de Valores Ausentes	Dtype
Nb of sec with 37500B < Vol UL	86.84	float64
Nb of sec with 6250B < Vol UL < 37500B	74.56	float64
Nb of sec with 125000B < Vol DL	65.02	float64
TCP UL Retrans. Vol (Bytes)	64.43	float64
Nb of sec with 31250B < Vol DL < 125000B	62.39	float64
Nb of sec with 1250B < Vol UL < 6250B	61.93	float64
Nb of sec with 6250B < Vol DL < 31250B	58.88	float64
TCP DL Retrans. Vol (Bytes)	58.76	float64
HTTP UL (Bytes)	54.54	float64
HTTP DL (Bytes)	54.32	float64
Avg RTT DL (ms)	18.55	float64
Avg RTT UL (ms)	18.54	float64
Handset Manufacturer	6.37	object
Handset Type	6.37	object
Last Location Name	0.77	object
MSISDN/Number	0.71	float64
Bearer Id	0.66	float64
Nb of sec with Vol UL < 1250B	0.53	float64
10 Kbps < UL TP < 50 Kbps (%)	0.53	float64
UL TP > 300 Kbps (%)	0.53	float64
50 Kbps < UL TP < 300 Kbps (%)	0.53	float64
UL TP < 10 Kbps (%)	0.53	float64
Nb of sec with Vol DL < 6250B	0.50	float64
250 Kbps < DL TP < 1 Mbps (%)	0.50	float64
50 Kbps < DL TP < 250 Kbps (%)	0.50	float64
DL TP < 50 Kbps (%)	0.50	float64
DL TP > 1 Mbps (%)	0.50	float64
IMEI	0.38	float64
IMSI	0.38	float64
Start	0.00	object
End Offset (ms)	0.00	float64
Total UL (Bytes)	0.00	float64
Start Offset (ms)	0.00	float64
End	0.00	object
Avg Bearer TP DL (kbps)	0.00	float64
Dur (s)	0.00	float64
Avg Bearer TP UL (kbps)	0.00	float64
Dur (ms)	0.00	float64
Activity Duration UL (ms)	0.00	float64
Activity Duration DL (ms)	0.00	float64
Total DL (Bytes)	0.00	float64

Normalmente removemos colunas com mais de 30% de valores ausentes.

[28]:
```
# Colunas que serão removidas
colunas_para_remover = df_missing[df_missing['% de Valores Ausentes'] >= 30.00].
 ↪index.tolist()
```

[29]:
```
# Colunas que serão removidas
colunas_para_remover
```

[29]:
```
['Nb of sec with 37500B < Vol UL',
 'Nb of sec with 6250B < Vol UL < 37500B',
 'Nb of sec with 125000B < Vol DL',
 'TCP UL Retrans. Vol (Bytes)',
 'Nb of sec with 31250B < Vol DL < 125000B',
 'Nb of sec with 1250B < Vol UL < 6250B',
 'Nb of sec with 6250B < Vol DL < 31250B',
 'TCP DL Retrans. Vol (Bytes)',
 'HTTP UL (Bytes)',
 'HTTP DL (Bytes)']
```

Mesmo que as variáveis TCP tenham muitos valores ausentes, em vez de removê-las, iremos imputá-las, uma vez que elas podem ser necessárias para nossa análise posterior.

[30]:
```
# Colunas que serão removidas
colunas_para_remover = [col for col in colunas_para_remover if col not in ['TCP␣
 ↪UL Retrans. Vol (Bytes)',
    'TCP DL Retrans. Vol (Bytes)']]
```

[31]:
```
# Colunas que serão removidas
colunas_para_remover
```

[31]:
```
['Nb of sec with 37500B < Vol UL',
 'Nb of sec with 6250B < Vol UL < 37500B',
 'Nb of sec with 125000B < Vol DL',
 'Nb of sec with 31250B < Vol DL < 125000B',
 'Nb of sec with 1250B < Vol UL < 6250B',
 'Nb of sec with 6250B < Vol DL < 31250B',
 'HTTP UL (Bytes)',
 'HTTP DL (Bytes)']
```

[32]:
```
# Drop das colunas e cria outro dataframe
dataset_clean = dataset.drop(colunas_para_remover, axis = 1)
```

[33]:
```
# Shape
dataset_clean.shape
```

[33]:
```
(150001, 47)
```

Agora vamos verificar o status dos valores ausentes no dataframe modificado.

[34]:
```
func_calc_percentual_valores_ausentes(dataset_clean)
```

O dataset tem 3.85 % de valores ausentes.

[35]:
```
func_calc_percentual_valores_ausentes_coluna(dataset_clean)
```

O dataset tem 47 colunas.
Encontrado: 33 colunas que têm valores ausentes.

[35]:

	Valores Ausentes	% de Valores Ausentes
TCP UL Retrans. Vol (Bytes)	96649	64.43
TCP DL Retrans. Vol (Bytes)	88146	58.76
Avg RTT DL (ms)	27829	18.55
Avg RTT UL (ms)	27812	18.54
Handset Type	9559	6.37
Handset Manufacturer	9559	6.37
Last Location Name	1153	0.77
MSISDN/Number	1066	0.71
Bearer Id	991	0.66
Nb of sec with Vol UL < 1250B	793	0.53
10 Kbps < UL TP < 50 Kbps (%)	792	0.53
UL TP < 10 Kbps (%)	792	0.53
50 Kbps < UL TP < 300 Kbps (%)	792	0.53
UL TP > 300 Kbps (%)	792	0.53
Nb of sec with Vol DL < 6250B	755	0.50
250 Kbps < DL TP < 1 Mbps (%)	754	0.50
DL TP > 1 Mbps (%)	754	0.50
DL TP < 50 Kbps (%)	754	0.50
50 Kbps < DL TP < 250 Kbps (%)	754	0.50
IMEI	572	0.38
IMSI	570	0.38
Start	1	0.00
Avg Bearer TP UL (kbps)	1	0.00
Avg Bearer TP DL (kbps)	1	0.00
Activity Duration DL (ms)	1	0.00
Activity Duration UL (ms)	1	0.00
Dur (ms)	1	0.00
Dur (s)	1	0.00
End Offset (ms)	1	0.00
End	1	0.00
Start Offset (ms)	1	0.00
Total UL (Bytes)	1	0.00
Total DL (Bytes)	1	0.00

	Dtype
TCP UL Retrans. Vol (Bytes)	float64
TCP DL Retrans. Vol (Bytes)	float64
Avg RTT DL (ms)	float64
Avg RTT UL (ms)	float64
Handset Type	object

```
Handset Manufacturer              object
Last Location Name                object
MSISDN/Number                    float64
Bearer Id                        float64
Nb of sec with Vol UL < 1250B    float64
10 Kbps < UL TP < 50 Kbps (%)    float64
UL TP < 10 Kbps (%)              float64
50 Kbps < UL TP < 300 Kbps (%)   float64
UL TP > 300 Kbps (%)             float64
Nb of sec with Vol DL < 6250B    float64
250 Kbps < DL TP < 1 Mbps (%)    float64
DL TP > 1 Mbps (%)               float64
DL TP < 50 Kbps (%)              float64
50 Kbps < DL TP < 250 Kbps (%)   float64
IMEI                             float64
IMSI                             float64
Start                             object
Avg Bearer TP UL (kbps)          float64
Avg Bearer TP DL (kbps)          float64
Activity Duration DL (ms)        float64
Activity Duration UL (ms)        float64
Dur (ms)                         float64
Dur (s)                          float64
End Offset (ms)                  float64
End                               object
Start Offset (ms)                float64
Total UL (Bytes)                 float64
Total DL (Bytes)                 float64
```

Uma vez que as porcentagens ausentes de 'TCP UL Retrans. Vol (Bytes) 'e' TCP DL Retrans. Vol (Bytes) 'são muito altos, iremos imputá-los com o método de preenchimento reverso. Nesse caso, usar um único valor como média ou mediana não é aconselhável, pois pode alterar nossos dados de uma forma indesejada, tornando a maioria dos valores igual a um único valor.

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html

[36]:
```
# Imputação com Preenchimento Reverso
fix_missing_bfill(dataset_clean, 'TCP UL Retrans. Vol (Bytes)')
```

```
96649 valores ausentes na coluna TCP UL Retrans. Vol (Bytes) foram substituídos
usando o método de preenchimento reverso.
```

[36]:
```
0         7230.0
1         7230.0
2         7230.0
3         7230.0
4         7230.0
           ...
149996       NaN
```

```
149997       NaN
149998       NaN
149999       NaN
150000       NaN
Name: TCP UL Retrans. Vol (Bytes), Length: 150001, dtype: float64
```

[37]:
```
# Imputação com Preenchimento Reverso
fix_missing_bfill(dataset_clean, 'TCP DL Retrans. Vol (Bytes)')
```

```
88146 valores ausentes na coluna TCP DL Retrans. Vol (Bytes) foram substituídos
usando o método de preenchimento reverso.
```

[37]:
```
0         19520.0
1         19520.0
2         19520.0
3         19520.0
4         19520.0
            ...
149996       NaN
149997       NaN
149998       NaN
149999       NaN
150000       NaN
Name: TCP DL Retrans. Vol (Bytes), Length: 150001, dtype: float64
```

Avg RTT DL (ms) e Avg RTT UL (ms) têm as próximas porcentagens mais altas de valores ausentes com cerca de 18,5% cada. Vamos verificar se as variáveis estão enviesadas (não seguem uma distribuição normal).

[38]:
```
dataset_clean['Avg RTT DL (ms)'].skew(skipna = True)
```

[38]:
```
62.90782807995961
```

[39]:
```
dataset_clean['Avg RTT UL (ms)'].skew(skipna = True)
```

[39]:
```
28.45741458546382
```

- Se a assimetria estiver entre -0,5 e 0,5, os dados são bastante simétricos
- Se a assimetria estiver entre -1 e - 0,5 ou entre 0,5 e 1, os dados estão moderadamente inclinados
- Se a assimetria for menor que -1 ou maior que 1, os dados estão altamente enviesados

Visto que ambas as colunas Avg RTT DL (ms) e Avg RTT UL (ms) são fortemente enviesadas positivamente, é aconselhável não imputá-las com sua média. Portanto, usaremos o preenchimento progressivo.

[40]:
```
# Imputação com Preenchimento Progressivo
fix_missing_ffill(dataset_clean, 'Avg RTT DL (ms)')
```

```
27829 valores ausentes na coluna Avg RTT DL (ms) foram substituídos usando o
método de preenchimento progressivo.
```

[40]:
```
0         42.0
1         65.0
2         65.0
3         65.0
4         65.0
          …
149996    32.0
149997    27.0
149998    43.0
149999    37.0
150000    37.0
Name: Avg RTT DL (ms), Length: 150001, dtype: float64
```

[41]:
```
# Imputação com Preenchimento Progressivo
fix_missing_ffill(dataset_clean, 'Avg RTT UL (ms)')
```

```
27812 valores ausentes na coluna Avg RTT UL (ms) foram substituídos usando o
método de preenchimento progressivo.
```

[41]:
```
0         5.0
1         5.0
2         5.0
3         5.0
4         5.0
         …
149996    0.0
149997    2.0
149998    6.0
149999    5.0
150000    5.0
Name: Avg RTT UL (ms), Length: 150001, dtype: float64
```

Checamos novamente os valores ausentes.

[42]:
```
func_calc_percentual_valores_ausentes(dataset_clean)
```

```
O dataset tem 0.44 % de valores ausentes.
```

[43]:
```
func_calc_percentual_valores_ausentes_linha(dataset_clean)
```

```
7.96 % das linhas no conjunto de dados contêm pelo menos um valor ausente.
```

[44]:
```
func_calc_percentual_valores_ausentes_coluna(dataset_clean)
```

```
O dataset tem 47 colunas.
Encontrado: 31 colunas que têm valores ausentes.
```

[44]:

	Valores Ausentes	% de Valores Ausentes
Handset Type	9559	6.37
Handset Manufacturer	9559	6.37
Last Location Name	1153	0.77
MSISDN/Number	1066	0.71
Bearer Id	991	0.66
Nb of sec with Vol UL < 1250B	793	0.53
UL TP > 300 Kbps (%)	792	0.53
50 Kbps < UL TP < 300 Kbps (%)	792	0.53
10 Kbps < UL TP < 50 Kbps (%)	792	0.53
UL TP < 10 Kbps (%)	792	0.53
Nb of sec with Vol DL < 6250B	755	0.50
DL TP < 50 Kbps (%)	754	0.50
DL TP > 1 Mbps (%)	754	0.50
250 Kbps < DL TP < 1 Mbps (%)	754	0.50
50 Kbps < DL TP < 250 Kbps (%)	754	0.50
IMEI	572	0.38
IMSI	570	0.38
TCP UL Retrans. Vol (Bytes)	5	0.00
TCP DL Retrans. Vol (Bytes)	5	0.00
Start	1	0.00
Avg Bearer TP UL (kbps)	1	0.00
Avg Bearer TP DL (kbps)	1	0.00
Activity Duration DL (ms)	1	0.00
Activity Duration UL (ms)	1	0.00
Dur (ms)	1	0.00
Dur (s)	1	0.00
End Offset (ms)	1	0.00
End	1	0.00
Start Offset (ms)	1	0.00
Total UL (Bytes)	1	0.00
Total DL (Bytes)	1	0.00

	Dtype
Handset Type	object
Handset Manufacturer	object
Last Location Name	object
MSISDN/Number	float64
Bearer Id	float64
Nb of sec with Vol UL < 1250B	float64
UL TP > 300 Kbps (%)	float64
50 Kbps < UL TP < 300 Kbps (%)	float64
10 Kbps < UL TP < 50 Kbps (%)	float64
UL TP < 10 Kbps (%)	float64
Nb of sec with Vol DL < 6250B	float64
DL TP < 50 Kbps (%)	float64
DL TP > 1 Mbps (%)	float64

```
250 Kbps < DL TP < 1 Mbps (%)   float64
50 Kbps < DL TP < 250 Kbps (%)  float64
IMEI                            float64
IMSI                            float64
TCP UL Retrans. Vol (Bytes)     float64
TCP DL Retrans. Vol (Bytes)     float64
Start                            object
Avg Bearer TP UL (kbps)         float64
Avg Bearer TP DL (kbps)         float64
Activity Duration DL (ms)       float64
Activity Duration UL (ms)       float64
Dur (ms)                        float64
Dur (s)                         float64
End Offset (ms)                 float64
End                              object
Start Offset (ms)               float64
Total UL (Bytes)                float64
Total DL (Bytes)                float64
```

[45]:
```
dataset_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150001 entries, 0 to 150000
Data columns (total 47 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   Bearer Id                       149010 non-null  float64
 1   Start                           150000 non-null  object
 2   Start Offset (ms)               150000 non-null  float64
 3   End                             150000 non-null  object
 4   End Offset (ms)                 150000 non-null  float64
 5   Dur (s)                         150000 non-null  float64
 6   IMSI                            149431 non-null  float64
 7   MSISDN/Number                   148935 non-null  float64
 8   IMEI                            149429 non-null  float64
 9   Last Location Name              148848 non-null  object
 10  Avg RTT DL (ms)                 150001 non-null  float64
 11  Avg RTT UL (ms)                 150001 non-null  float64
 12  Avg Bearer TP DL (kbps)         150000 non-null  float64
 13  Avg Bearer TP UL (kbps)         150000 non-null  float64
 14  TCP DL Retrans. Vol (Bytes)     149996 non-null  float64
 15  TCP UL Retrans. Vol (Bytes)     149996 non-null  float64
 16  DL TP < 50 Kbps (%)             149247 non-null  float64
 17  50 Kbps < DL TP < 250 Kbps (%)  149247 non-null  float64
 18  250 Kbps < DL TP < 1 Mbps (%)   149247 non-null  float64
 19  DL TP > 1 Mbps (%)              149247 non-null  float64
 20  UL TP < 10 Kbps (%)             149209 non-null  float64
 21  10 Kbps < UL TP < 50 Kbps (%)   149209 non-null  float64
```

```
 22  50 Kbps < UL TP < 300 Kbps (%)  149209 non-null  float64
 23  UL TP > 300 Kbps (%)            149209 non-null  float64
 24  Activity Duration DL (ms)       150000 non-null  float64
 25  Activity Duration UL (ms)       150000 non-null  float64
 26  Dur (ms)                        150000 non-null  float64
 27  Handset Manufacturer            140442 non-null  object
 28  Handset Type                    140442 non-null  object
 29  Nb of sec with Vol DL < 6250B   149246 non-null  float64
 30  Nb of sec with Vol UL < 1250B   149208 non-null  float64
 31  Social Media DL (Bytes)         150001 non-null  float64
 32  Social Media UL (Bytes)         150001 non-null  float64
 33  Google DL (Bytes)               150001 non-null  float64
 34  Google UL (Bytes)               150001 non-null  float64
 35  Email DL (Bytes)                150001 non-null  float64
 36  Email UL (Bytes)                150001 non-null  float64
 37  Youtube DL (Bytes)              150001 non-null  float64
 38  Youtube UL (Bytes)              150001 non-null  float64
 39  Netflix DL (Bytes)              150001 non-null  float64
 40  Netflix UL (Bytes)              150001 non-null  float64
 41  Gaming DL (Bytes)               150001 non-null  float64
 42  Gaming UL (Bytes)               150001 non-null  float64
 43  Other DL (Bytes)                150001 non-null  float64
 44  Other UL (Bytes)                150001 non-null  float64
 45  Total UL (Bytes)                150000 non-null  float64
 46  Total DL (Bytes)                150000 non-null  float64
dtypes: float64(42), object(5)
memory usage: 53.8+ MB
```

Visto que "Handset Type" e "Handset Manufacturer" são colunas categóricas, é melhor imputá-los com o valor "unknown" para que não enviesemos os dados.

[46]:
```
# Imputação de variáveis categóricas
fix_missing_value(dataset_clean, 'Handset Type', 'unknown')
fix_missing_value(dataset_clean, 'Handset Manufacturer', 'unknown')
```

```
9559 valores ausentes na coluna Handset Type foram substituídos por unknown.
9559 valores ausentes na coluna Handset Manufacturer foram substituídos por
unknown.
```

[46]:
```
0         Samsung
1         Samsung
2         Samsung
3         unknown
4         Samsung
           ...
149996      Apple
149997      Apple
149998      Apple
```

```
149999     Huawei
150000    unknown
Name: Handset Manufacturer, Length: 150001, dtype: object
```

Checamos novamente os valores ausentes.

[47]:
```
func_calc_percentual_valores_ausentes(dataset_clean)
```

```
O dataset tem 0.17 % de valores ausentes.
```

[48]:
```
func_calc_percentual_valores_ausentes_linha(dataset_clean)
```

```
2.08 % das linhas no conjunto de dados contêm pelo menos um valor ausente.
```

Uma vez que apenas 2,08% das linhas contêm pelo menos um valor ausente e o número total de linhas é de cerca de 150000, descartar essas linhas não terá um impacto negativo perceptível.

[49]:
```
# Drop de linhas com valores ausentes
drop_rows_with_missing_values(dataset_clean)
```

```
3114 linhas contendo valores ausentes foram descartadas.
```

[50]:
```
func_calc_percentual_valores_ausentes(dataset_clean)
```

```
O dataset tem 0.0 % de valores ausentes.
```

[51]:
```
# Shape
dataset_clean.shape
```

[51]:
```
(146887, 47)
```

2.8 Estratégia 2 - Conversão de Tipos de Dados

[52]:
```
dataset_clean.dtypes
```

[52]:
```
Bearer Id                        float64
Start                             object
Start Offset (ms)                float64
End                               object
End Offset (ms)                  float64
Dur (s)                          float64
IMSI                             float64
MSISDN/Number                    float64
IMEI                             float64
Last Location Name                object
Avg RTT DL (ms)                  float64
Avg RTT UL (ms)                  float64
Avg Bearer TP DL (kbps)          float64
Avg Bearer TP UL (kbps)          float64
TCP DL Retrans. Vol (Bytes)      float64
```

```
TCP UL Retrans. Vol (Bytes)       float64
DL TP < 50 Kbps (%)               float64
50 Kbps < DL TP < 250 Kbps (%)    float64
250 Kbps < DL TP < 1 Mbps (%)     float64
DL TP > 1 Mbps (%)                float64
UL TP < 10 Kbps (%)               float64
10 Kbps < UL TP < 50 Kbps (%)     float64
50 Kbps < UL TP < 300 Kbps (%)    float64
UL TP > 300 Kbps (%)              float64
Activity Duration DL (ms)         float64
Activity Duration UL (ms)         float64
Dur (ms)                          float64
Handset Manufacturer               object
Handset Type                       object
Nb of sec with Vol DL < 6250B     float64
Nb of sec with Vol UL < 1250B     float64
Social Media DL (Bytes)           float64
Social Media UL (Bytes)           float64
Google DL (Bytes)                 float64
Google UL (Bytes)                 float64
Email DL (Bytes)                  float64
Email UL (Bytes)                  float64
Youtube DL (Bytes)                float64
Youtube UL (Bytes)                float64
Netflix DL (Bytes)                float64
Netflix UL (Bytes)                float64
Gaming DL (Bytes)                 float64
Gaming UL (Bytes)                 float64
Other DL (Bytes)                  float64
Other UL (Bytes)                  float64
Total UL (Bytes)                  float64
Total DL (Bytes)                  float64
dtype: object
```

[53]:
```
dataset_clean
```

[53]:
```
           Bearer Id            Start  Start Offset (ms)              End  \
0       1.311448e+19   4/4/2019 12:01              770.0  4/25/2019 14:35
1       1.311448e+19   4/9/2019 13:04              235.0   4/25/2019 8:15
2       1.311448e+19   4/9/2019 17:42                1.0  4/25/2019 11:58
3       1.311448e+19   4/10/2019 0:31              486.0   4/25/2019 7:36
4       1.311448e+19  4/12/2019 20:10              565.0  4/25/2019 10:40
…                ...              ...                ...              ...
149991  7.349883e+18   4/29/2019 7:28              794.0   4/30/2019 0:36
149992  1.311448e+19   4/29/2019 7:28              114.0   4/30/2019 7:04
149993  1.311448e+19   4/29/2019 7:28               79.0  4/30/2019 18:22
149994  1.311448e+19   4/29/2019 7:28               83.0  4/30/2019 15:01
```

```
149995  1.304243e+19   4/29/2019 7:28              615.0   4/30/2019 0:01

        End Offset (ms)    Dur (s)          IMSI  MSISDN/Number          IMEI  \
0                 662.0  1823652.0  2.082014e+14   3.366496e+10  3.552121e+13
1                 606.0  1365104.0  2.082019e+14   3.368185e+10  3.579401e+13
2                 652.0  1361762.0  2.082003e+14   3.376063e+10  3.528151e+13
3                 171.0  1321509.0  2.082014e+14   3.375034e+10  3.535661e+13
4                 954.0  1089009.0  2.082014e+14   3.369980e+10  3.540701e+13
…                 …         …           …             …           …
149991            523.0    61661.0  2.082017e+14   3.376215e+10  3.586061e+13
149992            724.0    84940.0  2.082014e+14   3.376127e+10  3.532701e+13
149993            512.0   125622.0  2.082014e+14   3.362611e+10  3.573531e+13
149994            268.0   113545.0  2.082003e+14   3.364566e+10  3.515541e+13
149995            407.0    59587.0  2.082014e+14   3.366865e+10  3.533251e+13

           Last Location Name  Avg RTT DL (ms)  Avg RTT UL (ms)  \
0       9.16456699548519E+015             42.0              5.0
1                   L77566A             65.0              5.0
2                   D42335A             65.0              5.0
3                   T21824A             65.0              5.0
4                   D88865A             65.0              5.0
…                     …               …               …
149991              D10033B             27.0              2.0
149992              D78058B             37.0              4.0
149993              D73542B             46.0              6.0
149994              T88383B             46.0              6.0
149995              T85721A            313.0              9.0

        Avg Bearer TP DL (kbps)  Avg Bearer TP UL (kbps)  \
0                          23.0                     44.0
1                          16.0                     26.0
2                           6.0                      9.0
3                          44.0                     44.0
4                           6.0                      9.0
…                          …                        …
149991                     62.0                     54.0
149992                     23.0                     40.0
149993                     43.0                     41.0
149994                     55.0                     54.0
149995                  63420.0                   1393.0

        TCP DL Retrans. Vol (Bytes)  TCP UL Retrans. Vol (Bytes)  \
0                           19520.0                       7230.0
1                           19520.0                       7230.0
2                           19520.0                       7230.0
3                           19520.0                       7230.0
4                           19520.0                       7230.0
```

```
…                              …                              …
149991                  16552848.0                     162614.0
149992                  16552848.0                     162614.0
149993                  16552848.0                     162614.0
149994                  16552848.0                     162614.0
149995                  16552848.0                     162614.0

        DL TP < 50 Kbps (%)  50 Kbps < DL TP < 250 Kbps (%)  \
0                     100.0                             0.0
1                     100.0                             0.0
2                     100.0                             0.0
3                     100.0                             0.0
4                     100.0                             0.0
…                     …                                 …
149991                100.0                             0.0
149992                100.0                             0.0
149993                100.0                             0.0
149994                100.0                             0.0
149995                 54.0                            23.0

        250 Kbps < DL TP < 1 Mbps (%)  DL TP > 1 Mbps (%)  \
0                                 0.0                 0.0
1                                 0.0                 0.0
2                                 0.0                 0.0
3                                 0.0                 0.0
4                                 0.0                 0.0
…                                 …                   …
149991                            0.0                 0.0
149992                            0.0                 0.0
149993                            0.0                 0.0
149994                            0.0                 0.0
149995                            9.0                12.0

        UL TP < 10 Kbps (%)  10 Kbps < UL TP < 50 Kbps (%)  \
0                     100.0                             0.0
1                     100.0                             0.0
2                     100.0                             0.0
3                     100.0                             0.0
4                     100.0                             0.0
…                     …                                 …
149991                100.0                             0.0
149992                100.0                             0.0
149993                100.0                             0.0
149994                100.0                             0.0
149995                 96.0                             3.0

        50 Kbps < UL TP < 300 Kbps (%)  UL TP > 300 Kbps (%)  \
```

```
0                                  0.0                  0.0
1                                  0.0                  0.0
2                                  0.0                  0.0
3                                  0.0                  0.0
4                                  0.0                  0.0
…                                    …                    …
149991                             0.0                  0.0
149992                             0.0                  0.0
149993                             0.0                  0.0
149994                             0.0                  0.0
149995                             0.0                  0.0

        Activity Duration DL (ms)  Activity Duration UL (ms)      Dur (ms)  \
0                         37624.0                    38787.0  1.823653e+09
1                           168.0                     3560.0  1.365104e+09
2                             0.0                        0.0  1.361763e+09
3                          3330.0                    37882.0  1.321510e+09
4                             0.0                        0.0  1.089009e+09
…                               …                          …             …
149991                    22520.0                    26980.0  6.166173e+07
149992                    26597.0                    30281.0  8.494061e+07
149993                    30493.0                    35262.0  1.256224e+08
149994                    32668.0                    36271.0  1.135452e+08
149995                  3380381.0                  2355730.0  5.958779e+07

       Handset Manufacturer                  Handset Type  \
0                   Samsung    Samsung Galaxy A5 Sm-A520F
1                   Samsung   Samsung Galaxy J5 (Sm-J530)
2                   Samsung      Samsung Galaxy A8 (2018)
3                   unknown                       unknown
4                   Samsung             Samsung Sm-G390F
…                     …                             …
149991                Apple  Apple iPhone 6S Plus (A1687)
149992                Apple       Apple iPhone 6S (A1688)
149993                Apple       Apple iPhone Xr (A2105)
149994              Samsung   Samsung Galaxy J5 (Sm-J530)
149995              unknown                       unknown

        Nb of sec with Vol DL < 6250B  Nb of sec with Vol UL < 1250B  \
0                               213.0                          214.0
1                               971.0                         1022.0
2                               751.0                          695.0
3                                17.0                          207.0
4                               607.0                          604.0
…                                 …                              …
149991                          136.0                          149.0
149992                           94.0                           97.0
```

```
149993                        100.0                         108.0
149994                        150.0                         170.0
149995                       4801.0                        8448.0

        Social Media DL (Bytes)  Social Media UL (Bytes)  Google DL (Bytes)  \
0                     1545765.0                  24420.0          1634479.0
1                     1926113.0                   7165.0          3493924.0
2                     1684053.0                  42224.0          8535055.0
3                      644121.0                  13372.0          9023734.0
4                      862600.0                  50188.0          6248284.0
…                         …                      …                …
149991                 857971.0                  57778.0          8438995.0
149992                1717476.0                  35240.0          3455682.0
149993                2297830.0                  10129.0         10844751.0
149994                 843776.0                  49091.0          8169508.0
149995                 962097.0                  31078.0          1139573.0

        Google UL (Bytes)  Email DL (Bytes)  Email UL (Bytes)  \
0               1271433.0         3563542.0          137762.0
1                920172.0          629046.0          308339.0
2               1694064.0         2690151.0          672973.0
3               2788027.0         1439754.0          631229.0
4               1500559.0         1936496.0          173853.0
…                   …                 …                 …
149991          3597665.0         1079443.0          651174.0
149992          3827662.0           60320.0          894557.0
149993          1615125.0         1803603.0          751975.0
149994           753106.0         1310061.0          255481.0
149995           126061.0         3459965.0          820142.0

        Youtube DL (Bytes)  Youtube UL (Bytes)  Netflix DL (Bytes)  \
0               15854611.0           2501332.0           8198936.0
1               20247395.0          19111729.0          18338413.0
2               19725661.0          14699576.0          17587794.0
3               21388122.0          15146643.0          13994646.0
4               15259380.0          18962873.0          17124581.0
…                   …                   …                   …
149991           7404819.0           9864285.0          18954727.0
149992          22428728.0          14385815.0           6940672.0
149993          18144703.0           8161828.0          20559422.0
149994          23043782.0          19176074.0          18762809.0
149995           6550499.0          18003146.0          22468983.0

        Netflix UL (Bytes)  Gaming DL (Bytes)  Gaming UL (Bytes)  \
0                9656251.0        278082303.0         14344150.0
1               17227132.0        608750074.0          1170709.0
2                6163408.0        229584621.0           395630.0
```

```
3                1097942.0        799538153.0         10849722.0
4                 415218.0        527707248.0          3529801.0
…                       …                  …                  …
149991          19382787.0        258688968.0          9542486.0
149992           6801943.0         35738570.0          7842728.0
149993          20415237.0        569668054.0          1618788.0
149994           9591310.0        810666072.0         14858904.0
149995           7149728.0        784435351.0         12724751.0

        Other DL (Bytes)  Other UL (Bytes)  Total UL (Bytes)  Total DL (Bytes)
0             171744450.0         8814393.0        36749741.0       308879636.0
1             526904238.0        15055145.0        53800391.0       653384965.0
2             410692588.0         4215763.0        27883638.0       279807335.0
3             749039933.0        12797283.0        43324218.0       846028530.0
4             550709500.0        13910322.0        38542814.0       569138589.0
…                     …                 …                 …                 …
149991        249358264.0        14134902.0        57231077.0       295424923.0
149992        808482329.0        12135519.0        45923464.0        70341448.0
149993        168292600.0        11056036.0        43629118.0       623318363.0
149994        777865837.0        14693970.0        59377936.0       862796008.0
149995        321383162.0        14890486.0        53745392.0       819016468.0

[146887 rows x 47 columns]
```

Observando as colunas, podemos perceber que as colunas “Start” e “End” são, na verdade, valores de data e hora, embora sejam rotuladas como objetos pelo pandas. Além dessas duas colunas, todas as outras colunas com tipos de dados de objeto são, na verdade, valores de string. Portanto, vamos converter essas colunas em seus tipos de dados apropriados.

[54]:
```
# Converte para datetime
convert_to_datetime(dataset_clean, ['Start', 'End'])
```

[55]:
```
# Extrai as colunas do tipo object
string_columns = dataset_clean.select_dtypes(include = 'object').columns.
 ↪tolist()
```

[56]:
```
# Visualiza
string_columns
```

[56]:
```
['Last Location Name', 'Handset Manufacturer', 'Handset Type']
```

[57]:
```
# Converte para string
convert_to_string(dataset_clean, string_columns)
```

Também sabemos que Bearer Id, IMSI, MSISDN / Number, IMEI são números únicos usados para identificação. Portanto, para melhor legibilidade (e facilitar os filtros usados mais frente), vamos alterá-los de float64 para int64.

[58]:
```
# Lista de colunas para conversão
int_cols = ['Bearer Id', 'IMSI', 'MSISDN/Number', 'IMEI',]
```

[59]:
```
# Converte para int
convert_to_int(dataset_clean, int_cols)
```

[60]:
```
dataset_clean.dtypes
```

[60]:
```
Bearer Id                                int64
Start                            datetime64[ns]
Start Offset (ms)                       float64
End                              datetime64[ns]
End Offset (ms)                         float64
Dur (s)                                 float64
IMSI                                      int64
MSISDN/Number                             int64
IMEI                                      int64
Last Location Name                       string
Avg RTT DL (ms)                         float64
Avg RTT UL (ms)                         float64
Avg Bearer TP DL (kbps)                 float64
Avg Bearer TP UL (kbps)                 float64
TCP DL Retrans. Vol (Bytes)             float64
TCP UL Retrans. Vol (Bytes)             float64
DL TP < 50 Kbps (%)                     float64
50 Kbps < DL TP < 250 Kbps (%)          float64
250 Kbps < DL TP < 1 Mbps (%)           float64
DL TP > 1 Mbps (%)                      float64
UL TP < 10 Kbps (%)                     float64
10 Kbps < UL TP < 50 Kbps (%)           float64
50 Kbps < UL TP < 300 Kbps (%)          float64
UL TP > 300 Kbps (%)                    float64
Activity Duration DL (ms)               float64
Activity Duration UL (ms)               float64
Dur (ms)                                float64
Handset Manufacturer                     string
Handset Type                             string
Nb of sec with Vol DL < 6250B           float64
Nb of sec with Vol UL < 1250B           float64
Social Media DL (Bytes)                 float64
Social Media UL (Bytes)                 float64
Google DL (Bytes)                       float64
Google UL (Bytes)                       float64
Email DL (Bytes)                        float64
Email UL (Bytes)                        float64
Youtube DL (Bytes)                      float64
Youtube UL (Bytes)                      float64
```

```
Netflix DL (Bytes)                    float64
Netflix UL (Bytes)                    float64
Gaming DL (Bytes)                     float64
Gaming UL (Bytes)                     float64
Other DL (Bytes)                      float64
Other UL (Bytes)                      float64
Total UL (Bytes)                      float64
Total DL (Bytes)                      float64
dtype: object
```

[61]:
```
# Vamos checar se há registros duplicados
drop_duplicates(dataset_clean)
```

```
Nenhuma linha duplicada foi encontrada.
```

Como vimos na seção de limpeza da coluna, temos duas colunas de duração, uma em segundos e a outra em microssegundos. Vamos verificar se os valores são iguais convertendo os microssegundos em segundos.

[62]:
```
# Conversão e comparação
temp_df = dataset_clean[['Dur (s)', 'Dur (ms)']].copy()
multiply_by_factor(temp_df, ['Dur (ms)'], 1/1000)
temp_df['comparison'] = (temp_df['Dur (s)'] == temp_df['Dur (ms)'].apply(math.
 ↪floor))
```

[63]:
```
temp_df
```

[63]:

	Dur (s)	Dur (ms)	comparison
0	1823652.0	1823652.892	True
1	1365104.0	1365104.371	True
2	1361762.0	1361762.651	True
3	1321509.0	1321509.685	True
4	1089009.0	1089009.389	True
…	…	…	…
149991	61661.0	61661.729	True
149992	84940.0	84940.610	True
149993	125622.0	125622.433	True
149994	113545.0	113545.185	True
149995	59587.0	59587.792	True

```
[146887 rows x 3 columns]
```

[64]:
```
# As duas colunas são iguais?
print(all(temp_df['comparison']))
```

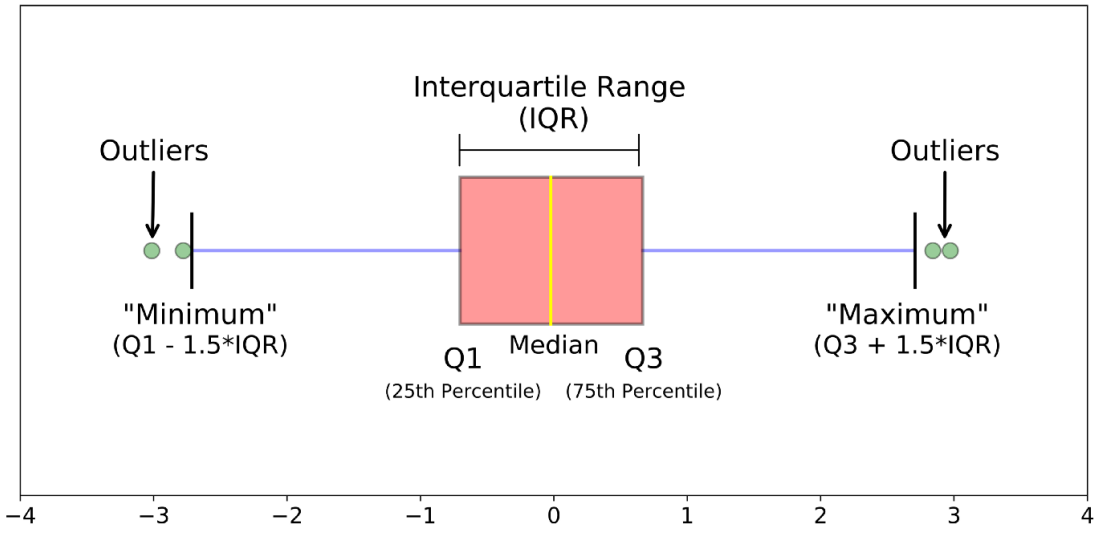
```
True
```

Isso prova que, quando arredondadas, essas duas colunas são iguais. Portanto, manteremos “Dur (ms)”, pois é mais preciso, e removeremos “Dur (s)”.

[65]:
```
# Drop de coluna
drop_columns(dataset_clean, ['Dur (s)'])
```

```
1 coluna foi descartada.
```

2.9 Estratégia 3 - Tratamento de Outliers



[66]:
```
# Cria o objeto trata outlier
trata_outlier = TrataOutlier(dataset_clean)
```

[67]:
```
# Lista de colunas float64
lista_colunas = dataset_clean.select_dtypes('float64').columns.tolist()
```

[68]:
```
lista_colunas
```

[68]:
```
['Start Offset (ms)',
 'End Offset (ms)',
 'Avg RTT DL (ms)',
 'Avg RTT UL (ms)',
 'Avg Bearer TP DL (kbps)',
 'Avg Bearer TP UL (kbps)',
 'TCP DL Retrans. Vol (Bytes)',
 'TCP UL Retrans. Vol (Bytes)',
 'DL TP < 50 Kbps (%)',
 '50 Kbps < DL TP < 250 Kbps (%)',
 '250 Kbps < DL TP < 1 Mbps (%)',
 'DL TP > 1 Mbps (%)',
 'UL TP < 10 Kbps (%)',
 '10 Kbps < UL TP < 50 Kbps (%)',
 '50 Kbps < UL TP < 300 Kbps (%)',
```

```
 'UL TP > 300 Kbps (%)',
 'Activity Duration DL (ms)',
 'Activity Duration UL (ms)',
 'Dur (ms)',
 'Nb of sec with Vol DL < 6250B',
 'Nb of sec with Vol UL < 1250B',
 'Social Media DL (Bytes)',
 'Social Media UL (Bytes)',
 'Google DL (Bytes)',
 'Google UL (Bytes)',
 'Email DL (Bytes)',
 'Email UL (Bytes)',
 'Youtube DL (Bytes)',
 'Youtube UL (Bytes)',
 'Netflix DL (Bytes)',
 'Netflix UL (Bytes)',
 'Gaming DL (Bytes)',
 'Gaming UL (Bytes)',
 'Other DL (Bytes)',
 'Other UL (Bytes)',
 'Total UL (Bytes)',
 'Total DL (Bytes)']
```

[69]:
```
# Visão geral dos outliers
trata_outlier.getOverview(lista_colunas)
```

[69]:
```
Nome de Coluna   DL TP > 1 Mbps (%) 10 Kbps < UL TP < 50 Kbps (%)  \
Min                              0.0                            0.0
Q1                               0.0                            0.0
Median                           0.0                            0.0
Q3                               0.0                            0.0
Max                             94.0                           98.0
IQR                              0.0                            0.0
Lower fence                      0.0                            0.0
Upper fence                      0.0                            0.0
Skew                        5.345557                      10.941071
Num_Outliers                   36594                          31509
Percent_Outliers               24.4%                         21.01%

Nome de Coluna   250 Kbps < DL TP < 1 Mbps (%) Activity Duration DL (ms)  \
Min                                      0.0                        0.0
Q1                                       0.0                    15418.0
Median                                   0.0                    39726.0
Q3                                       1.0                   697298.5
Max                                     76.0                136536461.0
IQR                                      1.0                   681880.5
Lower fence                             -1.5                -1007402.75
```

Upper fence	2.5	1720119.25
Skew	4.503234	5.821286
Num_Outliers	29343	26126
Percent_Outliers	19.56%	17.42%

Nome de Coluna	Activity Duration UL (ms)	Nb of sec with Vol UL < 1250B
Min	0.0	1.0
Q1	22073.0	107.0
Median	47180.0	217.0
Q3	611803.0	2466.5
Max	144911293.0	604122.0
IQR	589730.0	2359.5
Lower fence	-862522.0	-3432.25
Upper fence	1496398.0	6005.75
Skew	7.321154	7.467915
Num_Outliers	25501	24870
Percent_Outliers	17.0%	16.58%

Nome de Coluna	Nb of sec with Vol DL < 6250B	UL TP < 10 Kbps (%)
Min	1.0	0.0
Q1	87.0	99.0
Median	201.0	100.0
Q3	2612.5	100.0
Max	604061.0	100.0
IQR	2525.5	1.0
Lower fence	-3701.25	97.5
Upper fence	6400.75	101.5
Skew	9.009273	-8.958674
Num_Outliers	24604	21811
Percent_Outliers	16.4%	14.54%

Nome de Coluna	TCP DL Retrans. Vol (Bytes)	Avg Bearer TP UL (kbps)
Min	2.0	0.0
Q1	24895.0	47.0
Median	391635.0	63.0
Q3	3078146.0	1136.0
Max	4294425570.0	58613.0
IQR	3053251.0	1089.0
Lower fence	-4554981.5	-1586.5
Upper fence	7658022.5	2769.5
Skew	16.755094	4.473627
Num_Outliers	21625	21104
Percent_Outliers	14.42%	14.07%

Nome de Coluna	TCP UL Retrans. Vol (Bytes)	DL TP < 50 Kbps (%)
Min	1.0	0.0
Q1	3299.5	91.0

Median	16436.0	100.0
Q3	68622.0	100.0
Max	2908226006.0	100.0
IQR	65322.5	9.0
Lower fence	-94684.25	77.5
Upper fence	166605.75	113.5
Skew	96.168844	-2.293488
Num_Outliers	20367	18114
Percent_Outliers	13.58%	12.08%

Nome de Coluna	Avg RTT UL (ms)	Avg RTT DL (ms)	\
Min	0.0	0.0	
Q1	2.0	31.0	
Median	5.0	44.0	
Q3	13.0	67.0	
Max	7120.0	96923.0	
IQR	11.0	36.0	
Lower fence	-14.5	-23.0	
Upper fence	29.5	121.0	
Skew	26.544463	68.322057	
Num_Outliers	16278	16181	
Percent_Outliers	10.85%	10.79%	

Nome de Coluna	50 Kbps < DL TP < 250 Kbps (%)	Avg Bearer TP DL (kbps)	\
Min	0.0	0.0	
Q1	0.0	43.0	
Median	0.0	64.0	
Q3	4.0	20024.0	
Max	93.0	378160.0	
IQR	4.0	19981.0	
Lower fence	-6.0	-29928.5	
Upper fence	10.0	49995.5	
Skew	3.273684	2.57405	
Num_Outliers	14955	12678	
Percent_Outliers	9.97%	8.45%	

Nome de Coluna	Dur (ms)	50 Kbps < UL TP < 300 Kbps (%)	\
Min	7142988.0	0.0	
Q1	58526863.0	0.0	
Median	86399988.0	0.0	
Q3	133159382.0	0.0	
Max	1859336442.0	85.0	
IQR	74632519.0	0.0	
Lower fence	-53421915.5	0.0	
Upper fence	245108160.5	0.0	
Skew	3.946116	21.051707	
Num_Outliers	7057	4476	

Percent_Outliers	4.7%	2.98%

Nome de Coluna	UL TP > 300 Kbps (%)	Total UL (Bytes)	Youtube UL (Bytes)
Min	0.0	2866892.0	105.0
Q1	0.0	33218346.0	5516549.0
Median	0.0	41143144.0	11013447.0
Q3	0.0	49034880.0	16514278.0
Max	96.0	78331311.0	22011962.0
IQR	0.0	15816534.0	10997729.0
Lower fence	0.0	9493545.0	-10980044.5
Upper fence	0.0	72759681.0	33010871.5
Skew	35.993888	-0.001944	-0.001265
Num_Outliers	2425	241	0
Percent_Outliers	1.62%	0.16%	0.0%

Nome de Coluna	Other UL (Bytes)	Other DL (Bytes)	Gaming UL (Bytes)
Min	148.0	3290.0	59.0
Q1	4144551.0	210155459.0	4132622.0
Median	8265817.0	421633232.0	8294516.0
Q3	12382039.5	631361047.5	12432390.5
Max	16558816.0	843442489.0	16558794.0
IQR	8237488.5	421205588.5	8299768.5
Lower fence	-8211681.75	-421652923.75	-8317030.75
Upper fence	24738272.25	1263169430.25	24882043.25
Skew	0.001481	0.001832	-0.002566
Num_Outliers	0	0	0
Percent_Outliers	0.0%	0.0%	0.0%

Nome de Coluna	Gaming DL (Bytes)	Netflix UL (Bytes)	Netflix DL (Bytes)
Min	2516.0	35.0	42.0
Q1	210381659.0	5480202.0	5776625.5
Median	423492394.0	10999348.0	11642708.0
Q3	633333998.5	16503222.5	17469653.0
Max	843441889.0	22011955.0	23259189.0
IQR	422952339.5	11023020.5	11693027.5
Lower fence	-424046850.25	-11054328.75	-11762915.75
Upper fence	1267762507.75	33037753.25	35009194.25
Skew	-0.003914	-0.000273	-0.002172
Num_Outliers	0	0	0
Percent_Outliers	0.0%	0.0%	0.0%

Nome de Coluna	Start Offset (ms)	Youtube DL (Bytes)	Email UL (Bytes)
Min	0.0	53.0	2.0
Q1	250.0	5833462.0	233439.0
Median	499.0	11616334.0	466357.0
Q3	749.0	17440635.5	700297.5
Max	999.0	23259098.0	936418.0

IQR	499.0	11607173.5	466858.5
Lower fence	-498.5	-11577298.25	-466848.75
Upper fence	1497.5	34851395.75	1400585.25
Skew	0.000823	0.000119	0.007312
Num_Outliers	0	0	0
Percent_Outliers	0.0%	0.0%	0.0%

Nome de Coluna	Email DL (Bytes)	Google UL (Bytes)	Google DL (Bytes)	\
Min	14.0	3.0	207.0	
Q1	892573.0	1024371.5	2882586.5	
Median	1793613.0	2054793.0	5766724.0	
Q3	2689818.0	3088071.5	8625551.0	
Max	3586146.0	4121357.0	11462832.0	
IQR	1797245.0	2063700.0	5742964.5	
Lower fence	-1803294.5	-2071178.5	-5731860.25	
Upper fence	5385685.5	6183621.5	17239997.75	
Skew	-0.002659	0.002242	-0.008414	
Num_Outliers	0	0	0	
Percent_Outliers	0.0%	0.0%	0.0%	

Nome de Coluna	Social Media UL (Bytes)	Social Media DL (Bytes)	\
Min	0.0	12.0	
Q1	16431.0	898089.0	
Median	32908.0	1793409.0	
Q3	49326.5	2694640.0	
Max	65870.0	3586064.0	
IQR	32895.5	1796551.0	
Lower fence	-32912.25	-1796737.5	
Upper fence	98669.75	5389466.5	
Skew	0.000258	-0.001633	
Num_Outliers	0	0	
Percent_Outliers	0.0%	0.0%	

Nome de Coluna	End Offset (ms)	Total DL (Bytes)
Min	0.0	7114041.0
Q1	251.0	243073402.5
Median	500.0	455963875.0
Q3	750.0	665783068.0
Max	999.0	902969616.0
IQR	499.0	422709665.5
Lower fence	-497.5	-390991095.75
Upper fence	1498.5	1299847566.25
Skew	-0.001251	-0.003579
Num_Outliers	0	0
Percent_Outliers	0.0%	0.0%

[70]:
```
# Replace dos outliers
trata_outlier.replace_outliers_with_fences(lista_colunas)
```

[71]:
```
# Visão geral dos outliers
trata_outlier.getOverview(lista_colunas)
```

[71]:
```
Nome de Coluna   Start Offset (ms) Nb of sec with Vol DL < 6250B  \
Min                          0.0                          1.0
Q1                         250.0                         87.0
Median                     499.0                        201.0
Q3                         749.0                       2612.5
Max                        999.0                       6400.75
IQR                        499.0                       2525.5
Lower fence               -498.5                     -3701.25
Upper fence               1497.5                       6400.75
Skew                    0.000823                     1.205814
Num_Outliers                   0                            0
Percent_Outliers            0.0%                         0.0%

Nome de Coluna   Social Media DL (Bytes) Social Media UL (Bytes)  \
Min                              12.0                     0.0
Q1                           898089.0                 16431.0
Median                      1793409.0                 32908.0
Q3                          2694640.0                 49326.5
Max                         3586064.0                 65870.0
IQR                         1796551.0                 32895.5
Lower fence                -1796737.5                -32912.25
Upper fence                 5389466.5                 98669.75
Skew                       -0.001633                 0.000258
Num_Outliers                       0                        0
Percent_Outliers                0.0%                     0.0%

Nome de Coluna   Google DL (Bytes) Google UL (Bytes) Email DL (Bytes)  \
Min                          207.0              3.0             14.0
Q1                         2882586.5        1024371.5         892573.0
Median                     5766724.0        2054793.0        1793613.0
Q3                         8625551.0        3088071.5        2689818.0
Max                       11462832.0        4121357.0        3586146.0
IQR                        5742964.5        2063700.0        1797245.0
Lower fence               -5731860.25       -2071178.5       -1803294.5
Upper fence               17239997.75        6183621.5        5385685.5
Skew                       -0.008414         0.002242        -0.002659
Num_Outliers                       0                0                0
Percent_Outliers                0.0%             0.0%             0.0%

Nome de Coluna   Email UL (Bytes) Youtube DL (Bytes) Youtube UL (Bytes)  \
Min                          2.0              53.0              105.0
```

Q1	233439.0	5833462.0	5516549.0
Median	466357.0	11616334.0	11013447.0
Q3	700297.5	17440635.5	16514278.0
Max	936418.0	23259098.0	22011962.0
IQR	466858.5	11607173.5	10997729.0
Lower fence	-466848.75	-11577298.25	-10980044.5
Upper fence	1400585.25	34851395.75	33010871.5
Skew	0.007312	0.000119	-0.001265
Num_Outliers	0	0	0
Percent_Outliers	0.0%	0.0%	0.0%

Nome de Coluna	Netflix DL (Bytes)	Netflix UL (Bytes)	Gaming DL (Bytes)
Min	42.0	35.0	2516.0
Q1	5776625.5	5480202.0	210381659.0
Median	11642708.0	10999348.0	423492394.0
Q3	17469653.0	16503222.5	633333998.5
Max	23259189.0	22011955.0	843441889.0
IQR	11693027.5	11023020.5	422952339.5
Lower fence	-11762915.75	-11054328.75	-424046850.25
Upper fence	35009194.25	33037753.25	1267762507.75
Skew	-0.002172	-0.000273	-0.003914
Num_Outliers	0	0	0
Percent_Outliers	0.0%	0.0%	0.0%

Nome de Coluna	Gaming UL (Bytes)	Other DL (Bytes)	Other UL (Bytes)
Min	59.0	3290.0	148.0
Q1	4132622.0	210155459.0	4144551.0
Median	8294516.0	421633232.0	8265817.0
Q3	12432390.5	631361047.5	12382039.5
Max	16558794.0	843442489.0	16558816.0
IQR	8299768.5	421205588.5	8237488.5
Lower fence	-8317030.75	-421652923.75	-8211681.75
Upper fence	24882043.25	1263169430.25	24738272.25
Skew	-0.002566	0.001832	0.001481
Num_Outliers	0	0	0
Percent_Outliers	0.0%	0.0%	0.0%

Nome de Coluna	Total UL (Bytes)	Nb of sec with Vol UL < 1250B	Dur (ms)
Min	9493545.0	1.0	7142988.0
Q1	33218346.0	107.0	58526863.0
Median	41143144.0	217.0	86399988.0
Q3	49034880.0	2466.5	133159382.0
Max	72759681.0	6005.75	245108160.5
IQR	15816534.0	2359.5	74632519.0
Lower fence	9493545.0	-3432.25	-53421915.5
Upper fence	72759681.0	6005.75	245108160.5
Skew	-0.001978	1.209533	0.810422

Num_Outliers	0	0	0
Percent_Outliers	0.0%	0.0%	0.0%

Nome de Coluna	End Offset (ms)	Activity Duration UL (ms)	Avg RTT DL (ms)	\
Min	0.0	0.0	0.0	
Q1	251.0	22073.0	31.0	
Median	500.0	47180.0	44.0	
Q3	750.0	611803.0	67.0	
Max	999.0	1496398.0	121.0	
IQR	499.0	589730.0	36.0	
Lower fence	-497.5	-862522.0	-23.0	
Upper fence	1498.5	1496398.0	121.0	
Skew	-0.001251	1.207483	1.069479	
Num_Outliers	0	0	0	
Percent_Outliers	0.0%	0.0%	0.0%	

Nome de Coluna	Avg RTT UL (ms)	Avg Bearer TP DL (kbps)	\
Min	0.0	0.0	
Q1	2.0	43.0	
Median	5.0	64.0	
Q3	13.0	20024.0	
Max	29.5	49995.5	
IQR	11.0	19981.0	
Lower fence	-14.5	-29928.5	
Upper fence	29.5	49995.5	
Skew	1.165443	1.280613	
Num_Outliers	0	0	
Percent_Outliers	0.0%	0.0%	

Nome de Coluna	Avg Bearer TP UL (kbps)	TCP DL Retrans. Vol (Bytes)	\
Min	0.0	2.0	
Q1	47.0	24895.0	
Median	63.0	391635.0	
Q3	1136.0	3078146.0	
Max	2769.5	7658022.5	
IQR	1089.0	3053251.0	
Lower fence	-1586.5	-4554981.5	
Upper fence	2769.5	7658022.5	
Skew	1.230407	1.216388	
Num_Outliers	0	0	
Percent_Outliers	0.0%	0.0%	

Nome de Coluna	TCP UL Retrans. Vol (Bytes)	DL TP < 50 Kbps (%)	\
Min	1.0	77.5	
Q1	3299.5	91.0	
Median	16436.0	100.0	
Q3	68622.0	100.0	

Max	166605.75	100.0
IQR	65322.5	9.0
Lower fence	-94684.25	77.5
Upper fence	166605.75	113.5
Skew	1.20234	-1.22817
Num_Outliers	0	0
Percent_Outliers	0.0%	0.0%

Nome de Coluna	50 Kbps < DL TP < 250 Kbps (%)	250 Kbps < DL TP < 1 Mbps (%)	\
Min	0.0	0.0	
Q1	0.0	0.0	
Median	0.0	0.0	
Q3	4.0	1.0	
Max	10.0	2.5	
IQR	4.0	1.0	
Lower fence	-6.0	-1.5	
Upper fence	10.0	2.5	
Skew	1.315047	1.098748	
Num_Outliers	0	0	
Percent_Outliers	0.0%	0.0%	

Nome de Coluna	DL TP > 1 Mbps (%)	UL TP < 10 Kbps (%)	\
Min	0.0	97.5	
Q1	0.0	99.0	
Median	0.0	100.0	
Q3	0.0	100.0	
Max	0.0	100.0	
IQR	0.0	1.0	
Lower fence	0.0	97.5	
Upper fence	0.0	101.5	
Skew	0.0	-0.947267	
Num_Outliers	0	0	
Percent_Outliers	0.0%	0.0%	

Nome de Coluna	10 Kbps < UL TP < 50 Kbps (%)	50 Kbps < UL TP < 300 Kbps (%)	\
Min	0.0	0.0	
Q1	0.0	0.0	
Median	0.0	0.0	
Q3	0.0	0.0	
Max	0.0	0.0	
IQR	0.0	0.0	
Lower fence	0.0	0.0	
Upper fence	0.0	0.0	
Skew	0.0	0.0	
Num_Outliers	0	0	
Percent_Outliers	0.0%	0.0%	

```
Nome de Coluna   UL TP > 300 Kbps (%) Activity Duration DL (ms)  \
Min                              0.0                       0.0
Q1                               0.0                   15418.0
Median                           0.0                   39726.0
Q3                               0.0                  697298.5
Max                              0.0                1720119.25
IQR                              0.0                  681880.5
Lower fence                      0.0               -1007402.75
Upper fence                      0.0                1720119.25
Skew                             0.0                  1.205281
Num_Outliers                       0                         0
Percent_Outliers                0.0%                      0.0%

Nome de Coluna   Total DL (Bytes)
Min                       7114041.0
Q1                      243073402.5
Median                  455963875.0
Q3                      665783068.0
Max                     902969616.0
IQR                     422709665.5
Lower fence         -390991095.75
Upper fence         1299847566.25
Skew                      -0.003579
Num_Outliers                    0
Percent_Outliers             0.0%
```

2.10 Bônus: Gerando Novas Colunas

Ter a soma dos volumes de dados de upload e download para cada aplicativo como um total pode ser necessário para análises?

[72]:
```
dataset_clean['Social Media Data Volume (Bytes)'] = dataset_clean['Social Media␣
↪UL (Bytes)'] + dataset_clean['Social Media DL (Bytes)']
```

[73]:
```
dataset_clean['Google Data Volume (Bytes)'] = dataset_clean['Google UL␣
↪(Bytes)'] + dataset_clean['Google DL (Bytes)']
```

[74]:
```
dataset_clean['Email Data Volume (Bytes)'] = dataset_clean['Email UL (Bytes)']␣
↪+ dataset_clean['Email DL (Bytes)']
```

[75]:
```
dataset_clean['Youtube Data Volume (Bytes)'] = dataset_clean['Youtube UL␣
↪(Bytes)'] + dataset_clean['Youtube DL (Bytes)']
```

[76]:
```
dataset_clean['Netflix Data Volume (Bytes)'] = dataset_clean['Netflix UL␣
↪(Bytes)'] + dataset_clean['Netflix DL (Bytes)']
```

[77]:
```
dataset_clean['Gaming Data Volume (Bytes)'] = dataset_clean['Gaming UL␣
↪(Bytes)'] + dataset_clean['Gaming DL (Bytes)']
```

[78]:
```
dataset_clean['Other Data Volume (Bytes)'] = dataset_clean['Other UL (Bytes)']␣
↪+ dataset_clean['Other DL (Bytes)']
```

[79]:
```
dataset_clean['Total Data Volume (Bytes)'] = dataset_clean['Total UL (Bytes)']␣
↪+ dataset_clean['Total DL (Bytes)']
```

[80]:
```
dataset_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 146887 entries, 0 to 149995
Data columns (total 54 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   Bearer Id                           146887 non-null  int64
 1   Start                               146887 non-null  datetime64[ns]
 2   Start Offset (ms)                   146887 non-null  float64
 3   End                                 146887 non-null  datetime64[ns]
 4   End Offset (ms)                     146887 non-null  float64
 5   IMSI                                146887 non-null  int64
 6   MSISDN/Number                       146887 non-null  int64
 7   IMEI                                146887 non-null  int64
 8   Last Location Name                  146887 non-null  string
 9   Avg RTT DL (ms)                     146887 non-null  float64
 10  Avg RTT UL (ms)                     146887 non-null  float64
 11  Avg Bearer TP DL (kbps)             146887 non-null  float64
 12  Avg Bearer TP UL (kbps)             146887 non-null  float64
 13  TCP DL Retrans. Vol (Bytes)         146887 non-null  float64
 14  TCP UL Retrans. Vol (Bytes)         146887 non-null  float64
 15  DL TP < 50 Kbps (%)                 146887 non-null  float64
 16  50 Kbps < DL TP < 250 Kbps (%)      146887 non-null  float64
 17  250 Kbps < DL TP < 1 Mbps (%)       146887 non-null  float64
 18  DL TP > 1 Mbps (%)                  146887 non-null  float64
 19  UL TP < 10 Kbps (%)                 146887 non-null  float64
 20  10 Kbps < UL TP < 50 Kbps (%)       146887 non-null  float64
 21  50 Kbps < UL TP < 300 Kbps (%)      146887 non-null  float64
 22  UL TP > 300 Kbps (%)                146887 non-null  float64
 23  Activity Duration DL (ms)           146887 non-null  float64
 24  Activity Duration UL (ms)           146887 non-null  float64
 25  Dur (ms)                            146887 non-null  float64
 26  Handset Manufacturer                146887 non-null  string
 27  Handset Type                        146887 non-null  string
 28  Nb of sec with Vol DL < 6250B       146887 non-null  float64
 29  Nb of sec with Vol UL < 1250B       146887 non-null  float64
 30  Social Media DL (Bytes)             146887 non-null  float64
 31  Social Media UL (Bytes)             146887 non-null  float64
 32  Google DL (Bytes)                   146887 non-null  float64
 33  Google UL (Bytes)                   146887 non-null  float64
 34  Email DL (Bytes)                    146887 non-null  float64
```

```
 35  Email UL (Bytes)                  146887 non-null  float64
 36  Youtube DL (Bytes)                146887 non-null  float64
 37  Youtube UL (Bytes)                146887 non-null  float64
 38  Netflix DL (Bytes)                146887 non-null  float64
 39  Netflix UL (Bytes)                146887 non-null  float64
 40  Gaming DL (Bytes)                 146887 non-null  float64
 41  Gaming UL (Bytes)                 146887 non-null  float64
 42  Other DL (Bytes)                  146887 non-null  float64
 43  Other UL (Bytes)                  146887 non-null  float64
 44  Total UL (Bytes)                  146887 non-null  float64
 45  Total DL (Bytes)                  146887 non-null  float64
 46  Social Media Data Volume (Bytes)  146887 non-null  float64
 47  Google Data Volume (Bytes)        146887 non-null  float64
 48  Email Data Volume (Bytes)         146887 non-null  float64
 49  Youtube Data Volume (Bytes)       146887 non-null  float64
 50  Netflix Data Volume (Bytes)       146887 non-null  float64
 51  Gaming Data Volume (Bytes)        146887 non-null  float64
 52  Other Data Volume (Bytes)         146887 non-null  float64
 53  Total Data Volume (Bytes)         146887 non-null  float64
dtypes: datetime64[ns](2), float64(45), int64(4), string(3)
memory usage: 61.6 MB
```

[81]:
```
dataset_clean.shape
```

[81]:
```
(146887, 54)
```

2.11 Salvando os Dados Após a Limpeza

[82]:
```
# Salvando os dados
dataset_clean.to_csv('dados/dataset_clean.csv')
```

3 Fim