EDA for dataset: Prolexitim_Jan23_En.csv

Dataset version: Jan23_En. Last file modification: 2023-06-01 14:47:29.305594

Report Generated: 05/06/2023 09:39:59

By: raul.arrabales - See: https://github.com/raul-arrabales/DataScience_Enablement

From: data/Prolexitim_Jan23_En.csv

Dataset Dimensions

File size: 2.32 MB.

Number of rows: 1010

Number of columns: 36

Duplicates Detection

There are no duplicate rows in the dataset.

Time Series Data Detection

Column Timestamp was loaded as an object but contains a well-formatted datetime.

A new column of type datetime64[ns, UTC] named Converted_Timestamp has been created successfully.

Possible single time series. Column Timestamp has 1 observation per timestamp

Column Timestamp (1) was loaded as an object but contains a well-formatted datetime.

A new column of type datetime64[ns, UTC] named Converted_Timestamp (1) has been created successfu

Possible single time series. Column Timestamp (1) has 1 observation per timestamp

List of possible time index columns: ['Timestamp', 'Timestamp (1)']

Columns Univariate Analysis

Column: Timestamp

Column Type: object

See Converted_Timestamp column below for a detailed analysis.

Sample value: 2021-03-17T06:55:09.874Z

Number of unique values: 1010

Percentage of unique values: 100.00%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: TAS20

Column Type: int64

Sample value: 71

Number of unique values: 65

Percentage of unique values: 6.44%

Number of missing values: 0

Percentage of missing values: 0.00%

Stats	TAS20
count	1010.00
mean	67.40
std	11.71
min	25.00
25%	61.00
50%	69.00
75%	75.00
max	100.00

Warning: There are outliers in column TAS20

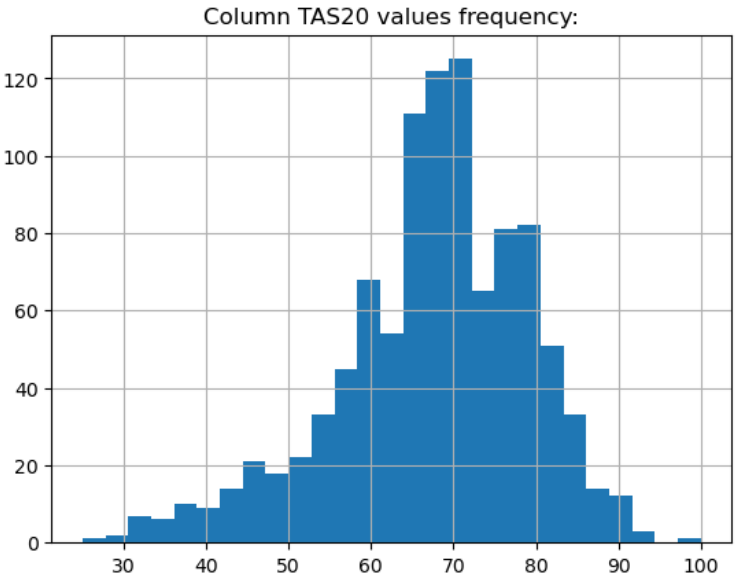
Sample of outliers: [39, 33, 39, 37, 32, 100, 34, 34, 39, 38]

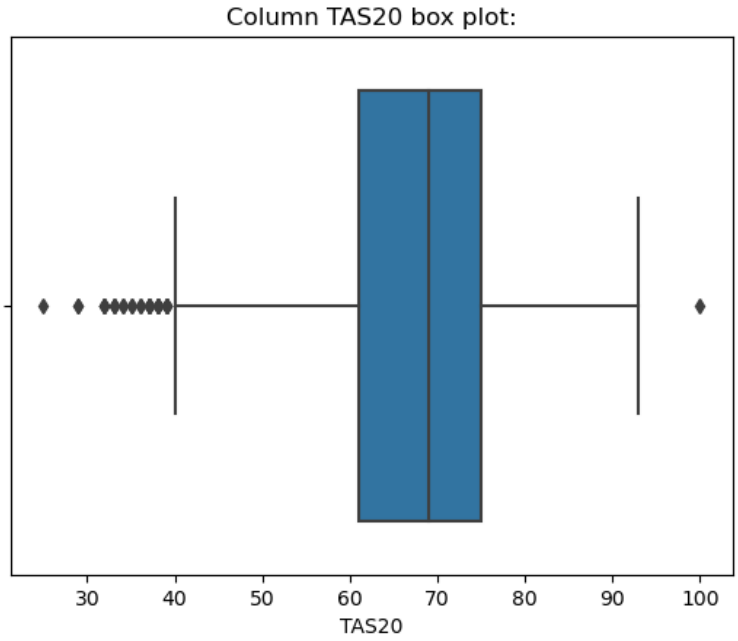
Sample of outliers over Q3: [100]

Sample of outliers below Q1: [39, 39, 39, 39, 39, 38, 38, 38, 38, 38]

No inliers detected in column TAS20

Column TAS20 values frequency:





Column: F1

Column Type: int64

Sample value: 30

Number of unique values: 29

Percentage of unique values: 2.87%

Number of missing values: 0

Percentage of missing values: 0.00%

Stats	F1
count	1010.00
mean	25.52
std	5.55
min	7.00
25%	23.00
50%	26.00
75%	29.00
max	35.00

Warning: There are outliers in column F1

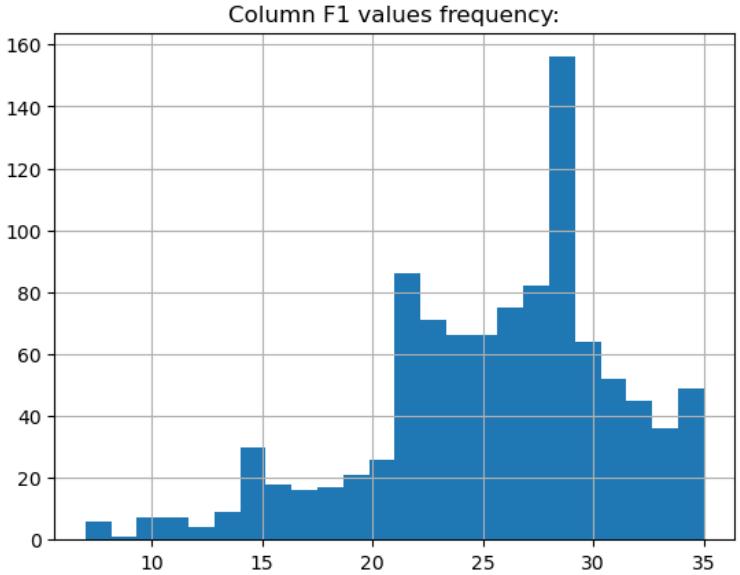
Sample of outliers: [10, 11, 13, 12, 13, 13, 8, 13, 10, 7]

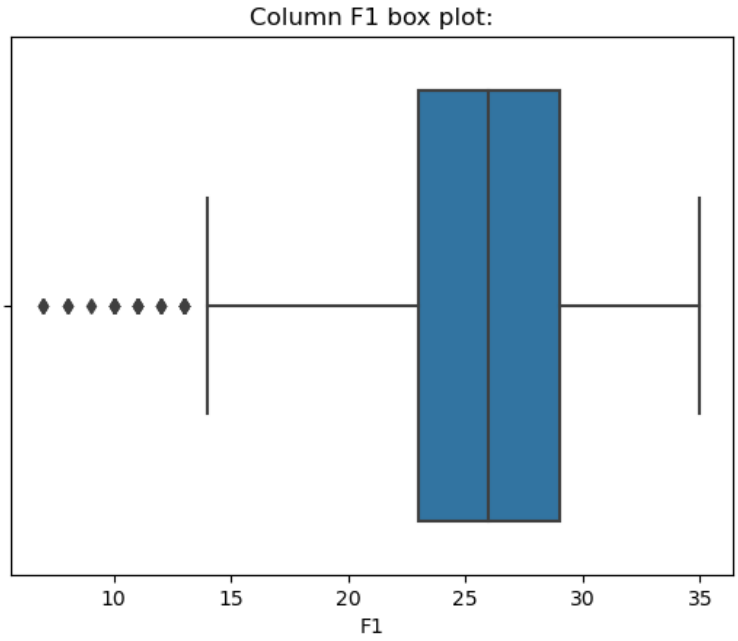
Sample of outliers over Q3: []

Sample of outliers below Q1: [13, 13, 13, 13, 13, 13, 13, 13, 13, 12]

No inliers detected in column F1

Column F1 values frequency:





Column: F2

Column Type: int64

Sample value: 21

Number of unique values: 21

Percentage of unique values: 2.08%

Number of missing values: 0

Percentage of missing values: 0.00%

Stats	F2
count	1010.00
mean	19.11
std	4.17
min	5.00
25%	17.00
50%	20.00
75%	22.00
max	25.00

Warning: There are outliers in column F2

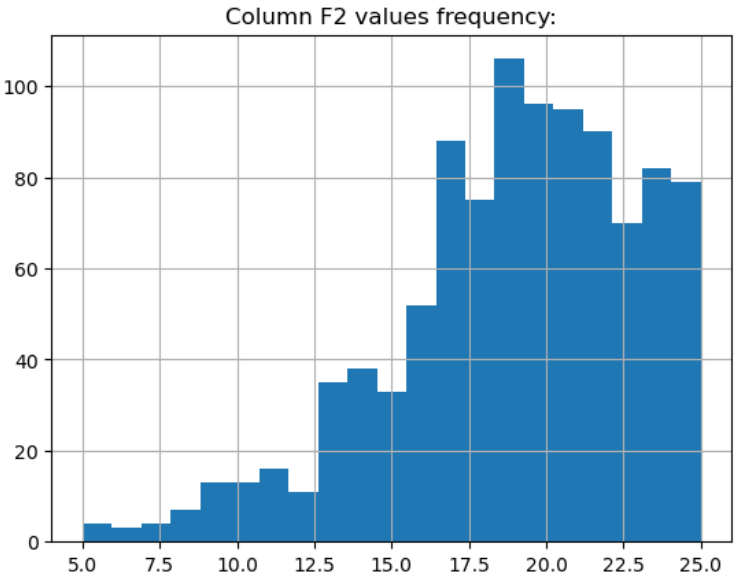
Sample of outliers: [9, 6, 9, 7, 5, 5, 7, 7, 6, 9]

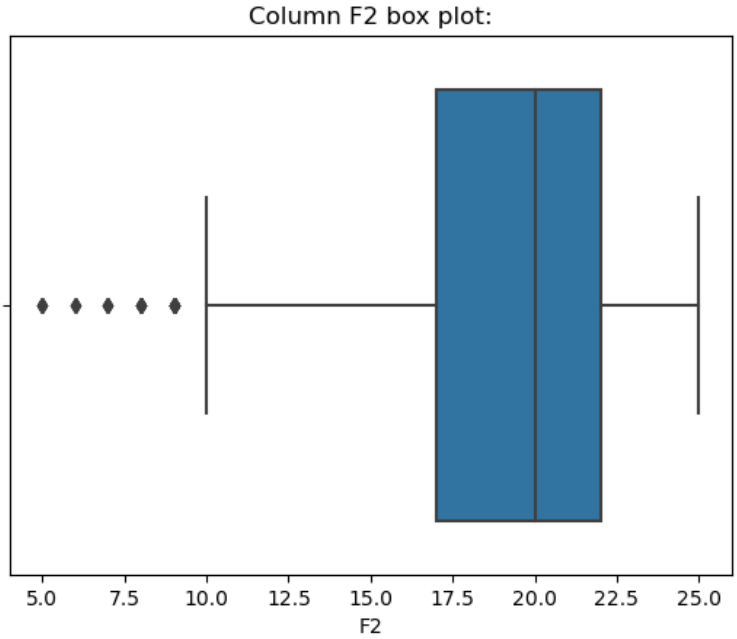
Sample of outliers over Q3: []

Sample of outliers below Q1: [9, 9, 9, 9, 9, 9, 9, 9, 9, 9]

No inliers detected in column F2

Column F2 values frequency:





Column: F3

Column Type: int64

Sample value: 20

Number of unique values: 31

Percentage of unique values: 3.07%

Number of missing values: 0

Percentage of missing values: 0.00%

Stats	F3
count	1010.00
mean	22.77
std	5.36
min	8.00
25%	19.00
50%	23.00
75%	26.75
max	40.00

Warning: There are outliers in column F3

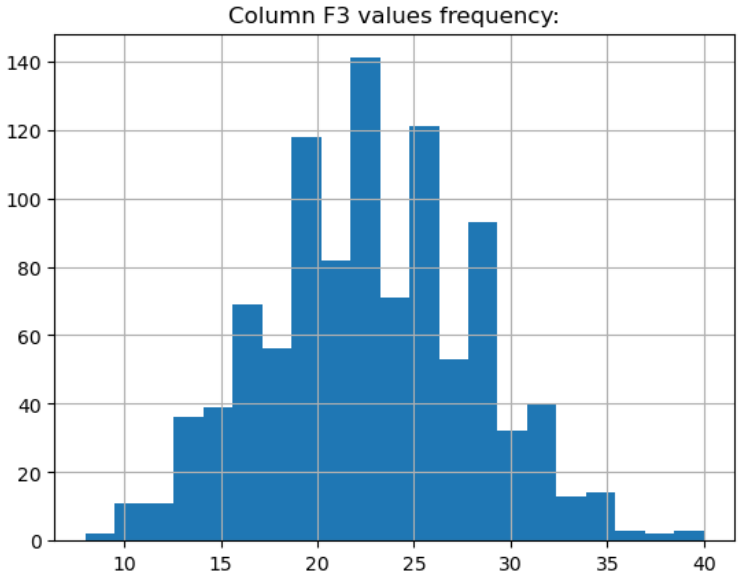
Sample of outliers: [40, 40, 40]

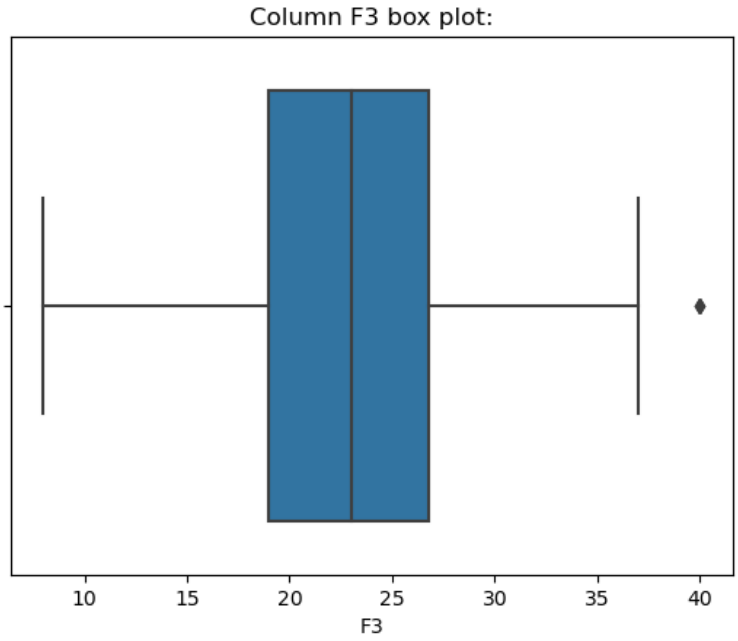
Sample of outliers over Q3: [40, 40, 40]

Sample of outliers below Q1: []

No inliers detected in column F3

Column F3 values frequency:





Column: Gender

Column Type: int64

Sample value: 2

Number of unique values: 3

Percentage of unique values: 0.30%

Number of missing values: 0

Percentage of missing values: 0.00%

Stats	Gender
count	1010.00
mean	1.88
std	0.40
min	1.00
25%	2.00
50%	2.00
75%	2.00
max	3.00

Warning: There are outliers in column Gender

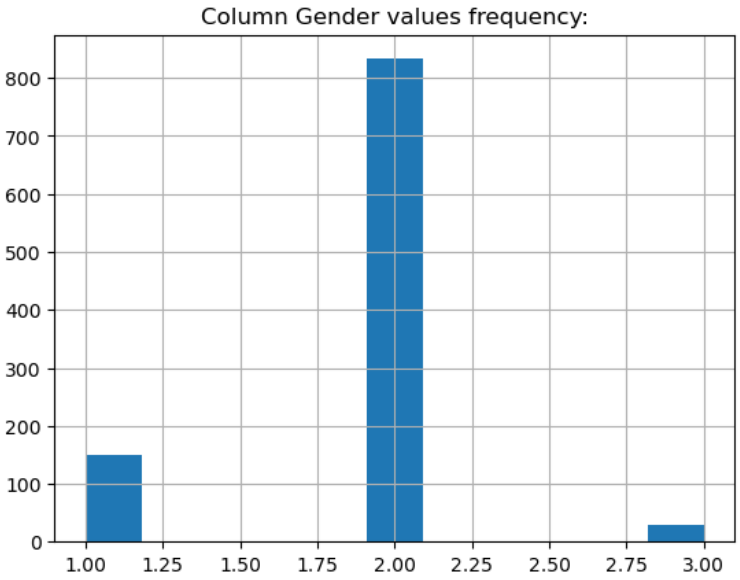
Sample of outliers: [1, 3, 1, 1, 1, 1, 1, 1, 1, 3]

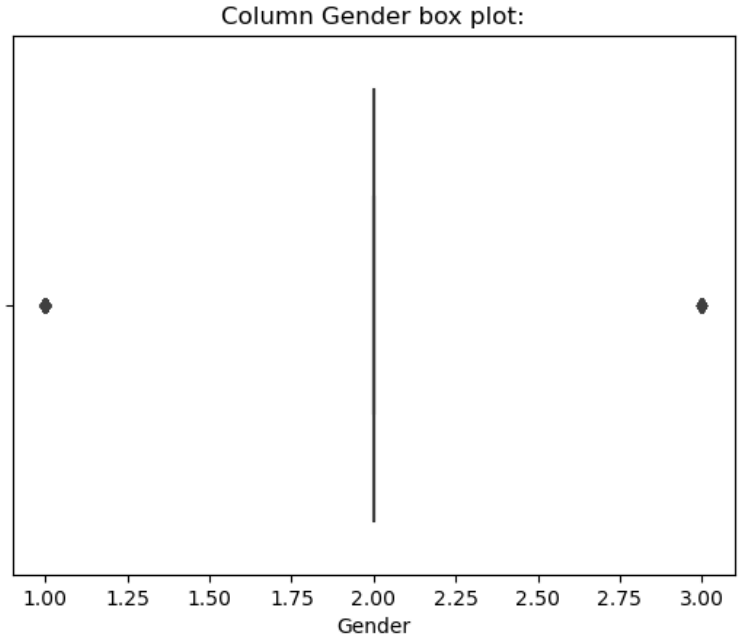
Sample of outliers over Q3: [3, 3, 3, 3, 3, 3, 3, 3, 3, 3]

Sample of outliers below Q1: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

No inliers detected in column Gender

Column Gender values frequency:





Column: Age

Column Type: int64

Sample value: 20

Number of unique values: 45

Percentage of unique values: 4.46%

Number of missing values: 0

Percentage of missing values: 0.00%

Stats	Age
count	1010.00
mean	23.17
std	9.06
min	18.00
25%	18.00
50%	18.00
75%	24.00
max	68.00

Warning: There are outliers in column Age

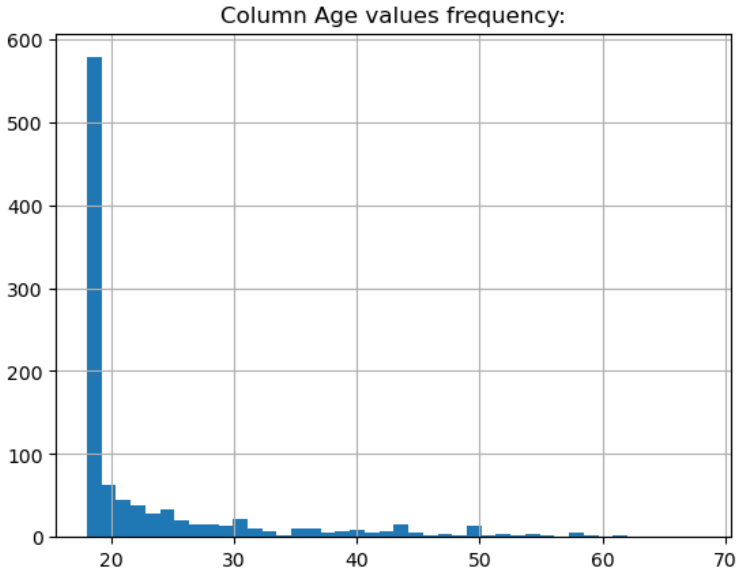
Sample of outliers: [35, 35, 52, 41, 44, 44, 45, 54, 38, 37]

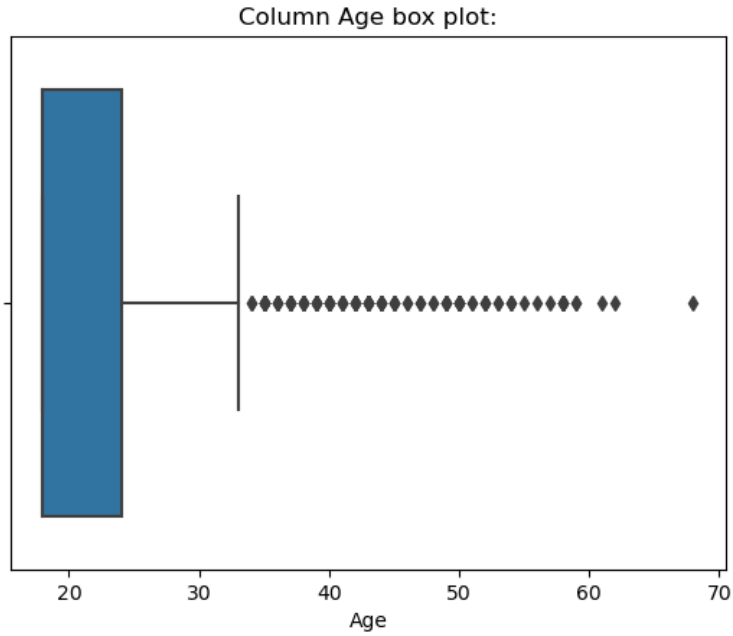
Sample of outliers over Q3: [34, 34, 35, 35, 35, 35, 35, 35, 35, 35]

Sample of outliers below Q1: []

No inliers detected in column Age

Column Age values frequency:





Column: Code

Column Type: object

See Converted_Code column below for a detailed analysis.

Sample value: 8b45a0f1b5fad1e453f0ac5ccb905fa7

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Alexithymia

Column Type: object

See Converted_Alexithymia column below for a detailed analysis.

Sample value: POSITIVE

Number of unique values: 2

Percentage of unique values: 0.20%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Timestamp (1)

Column Type: object

See Converted_Timestamp (1) column below for a detailed analysis.

Sample value: 2021-03-17T07:39:24.457Z

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: DateDiff_TAS20_NLP

Column Type: int64

Sample value: 0

Number of unique values: 43

Percentage of unique values: 4.26%

Number of missing values: 0

Percentage of missing values: 0.00%

Stats	DateDiff_TAS20_NLP
count	1010.00
mean	-1.15
std	31.04
min	-498.00
25%	0.00
50%	0.00
75%	0.00
max	381.00

Warning: There are outliers in column DateDiff_TAS20_NLP

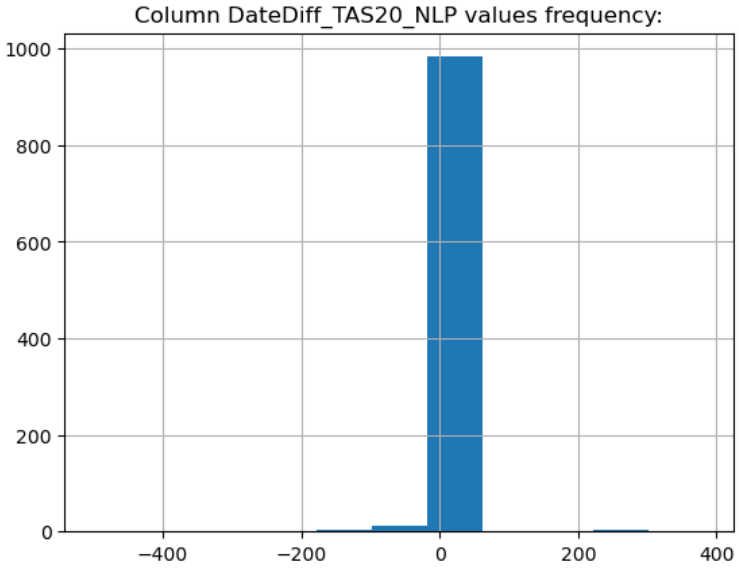
Sample of outliers: [-263, -22, 71, 21, -62, -376, 1, 381, -224, 5]

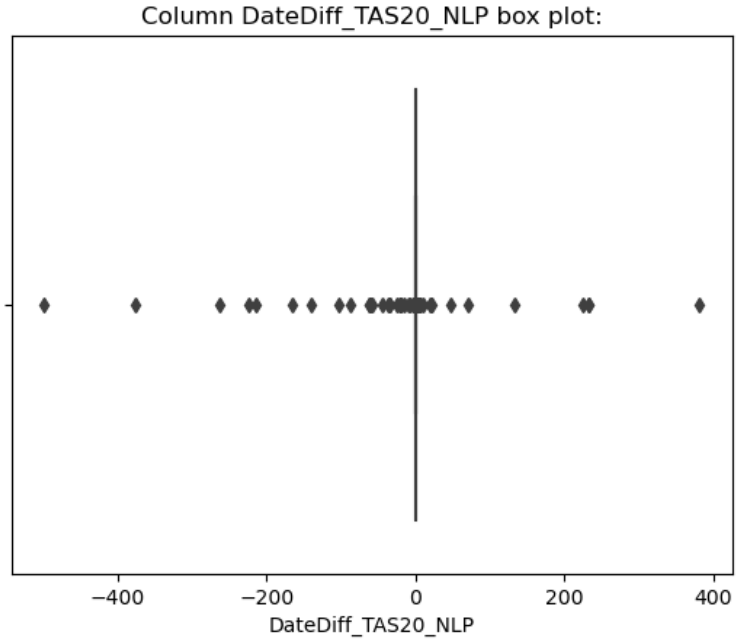
Sample of outliers over Q3: [1, 1, 1, 1, 2, 4, 4, 5, 5, 6]

Sample of outliers below Q1: [-1, -1, -2, -3, -8, -9, -15, -19, -21, -22]

No inliers detected in column DateDiff_TAS20_NLP

Column DateDiff_TAS20_NLP values frequency:





Column: Text_Card1

Column Type: object

See Converted_Text_Card1 column below for a detailed analysis.

Sample value: ... un niño, común, como cualquier otro, salía de casa, iba a la escuela, tenía amigos, su fa

Number of unique values: 945

Percentage of unique values: 93.56%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Text_Card9VH

Column Type: object

See Converted_Text_Card9VH column below for a detailed analysis.

Sample value: ... un hombre que estaba borracho con sus amigos

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Text_Card11

Column Type: object

See Converted_Text_Card11 column below for a detailed analysis.

Sample value: ... un sendero mágico, capaz de llevarte guiarte tu deseo más profundo, no importaba que

Number of unique values: 943

Percentage of unique values: 93.37%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Text_Card13HM

Column Type: object

See Converted_Text_Card13HM column below for a detailed analysis.

Sample value: ... un hombre que se dió cuenta que no era feliz, aún con una esposa que lo amaba, pues

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Text_AllCards

Column Type: object

See Converted_Text_AllCards column below for a detailed analysis.

Sample value: ... un niño, común, como cualquier otro, salía de casa, iba a la escuela, tenía amigos, su fa

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: PHQ9

Column Type: float64

Sample value: 19.0

Number of unique values: 29

Percentage of unique values: 2.87%

Number of missing values: 586

Percentage of missing values: 58.02%

Warning: too missing values in column PHQ9

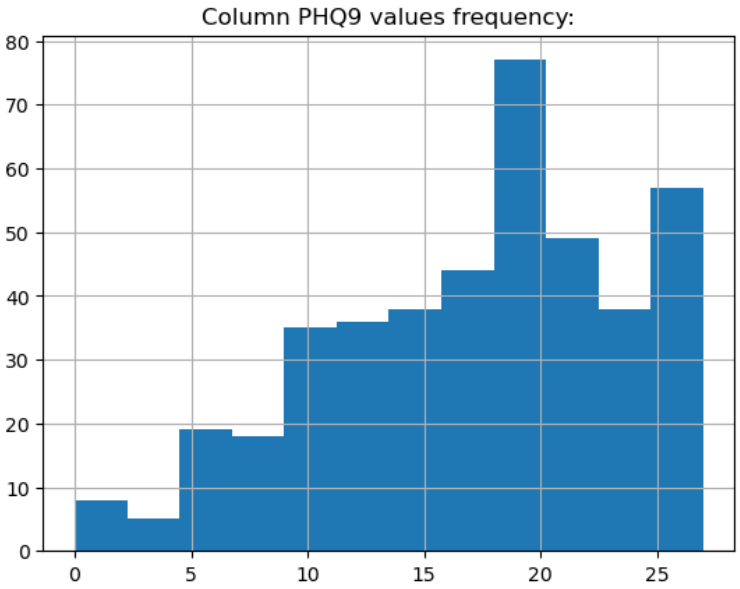
Stats	PHQ9
count	424.00
mean	17.09

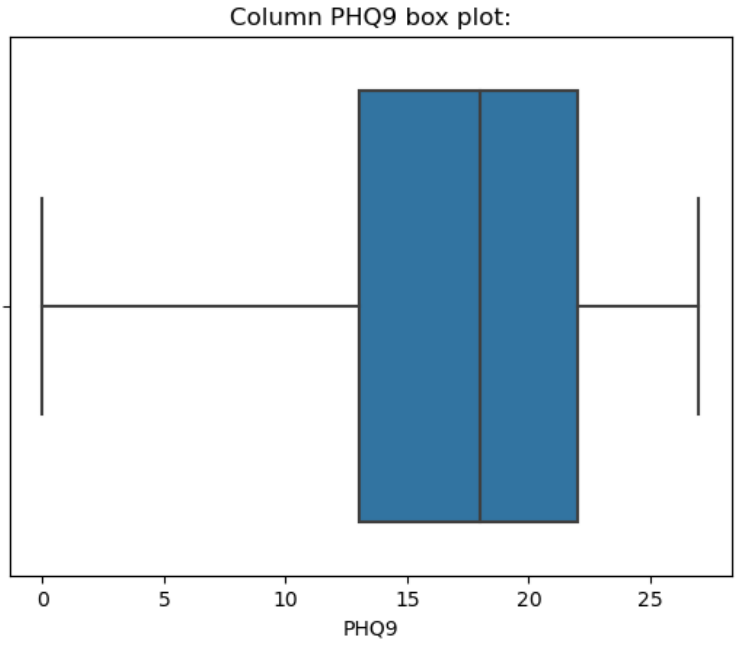
std	6.39
min	0.00
25%	13.00
50%	18.00
75%	22.00
max	27.00

There are no outliers in column PHQ9

No inliers detected in column PHQ9

Column PHQ9 values frequency:





Column: I1

Column Type: float64

Sample value: 3.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

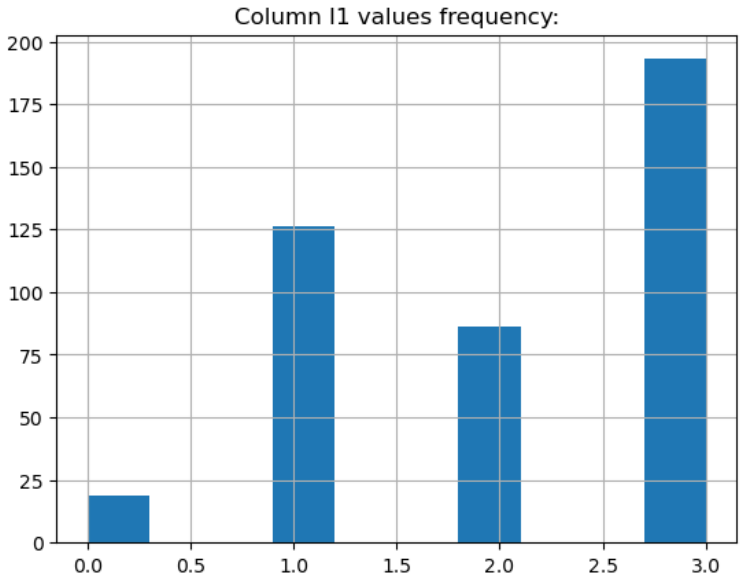
Warning: too missing values in column I1

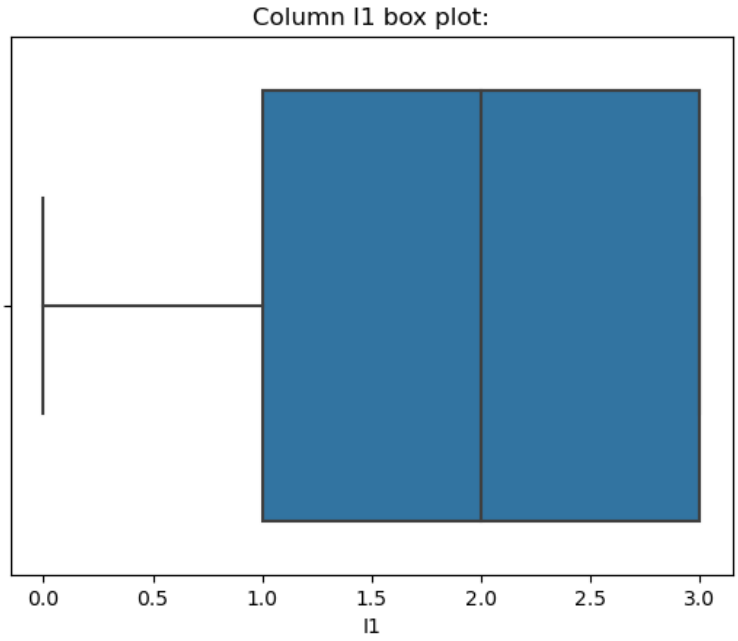
Stats	I1
count	424.00
mean	2.07
std	0.96
min	0.00
25%	1.00
50%	2.00
75%	3.00
max	3.00

There are no outliers in column I1

No inliers detected in column I1

Column I1 values frequency:





Column: I2

Column Type: float64

Sample value: 2.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

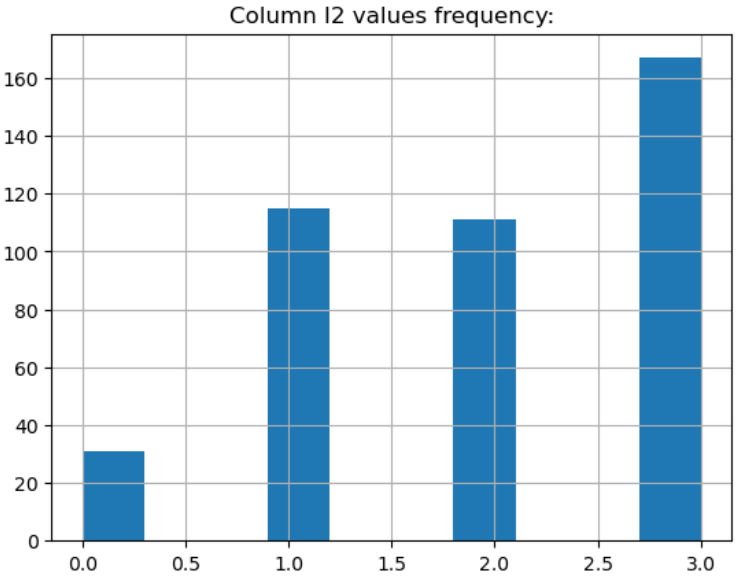
Warning: too missing values in column I2

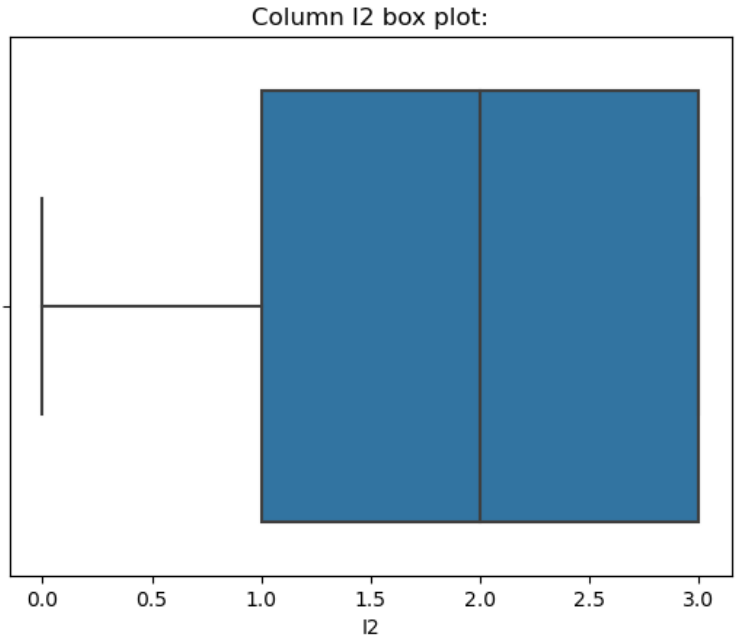
Stats	I2
count	424.00
mean	1.98
std	0.98
min	0.00
25%	1.00
50%	2.00
75%	3.00
max	3.00

There are no outliers in column I2

No inliers detected in column I2

Column I2 values frequency:





Column: I3

Column Type: float64

Sample value: 2.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

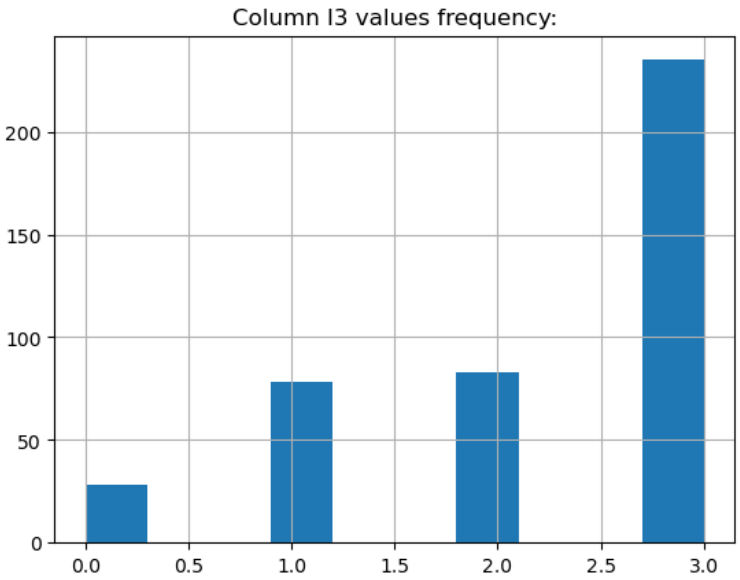
Warning: too missing values in column I3

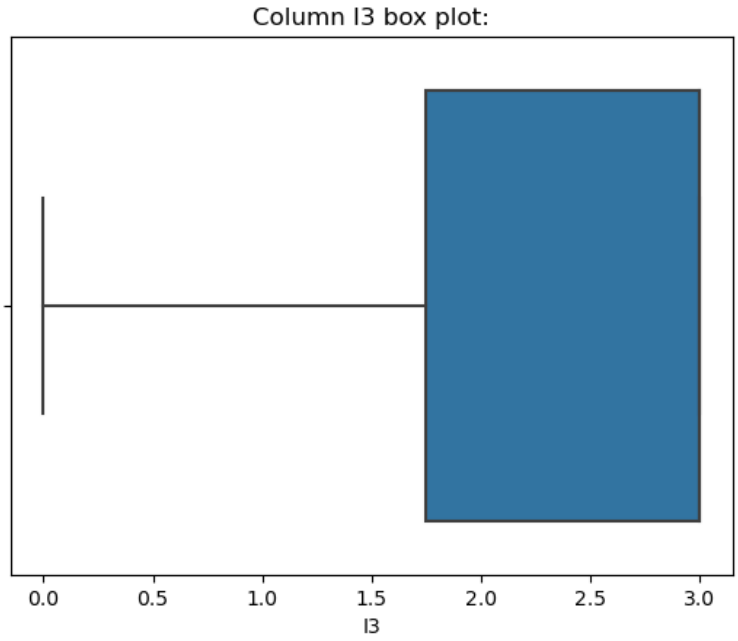
Stats	I3
count	424.00
mean	2.24
std	0.97
min	0.00
25%	1.75
50%	3.00
75%	3.00
max	3.00

There are no outliers in column I3

No inliers detected in column I3

Column I3 values frequency:





Column: I4

Column Type: float64

Sample value: 3.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

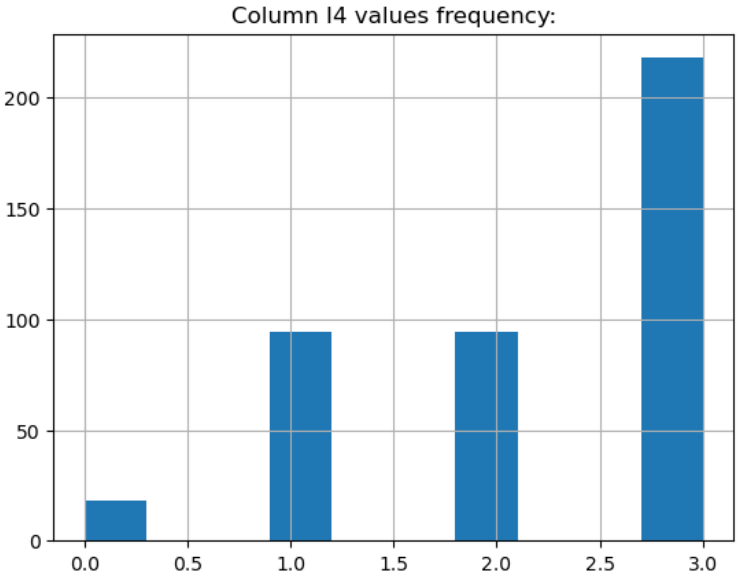
Warning: too missing values in column I4

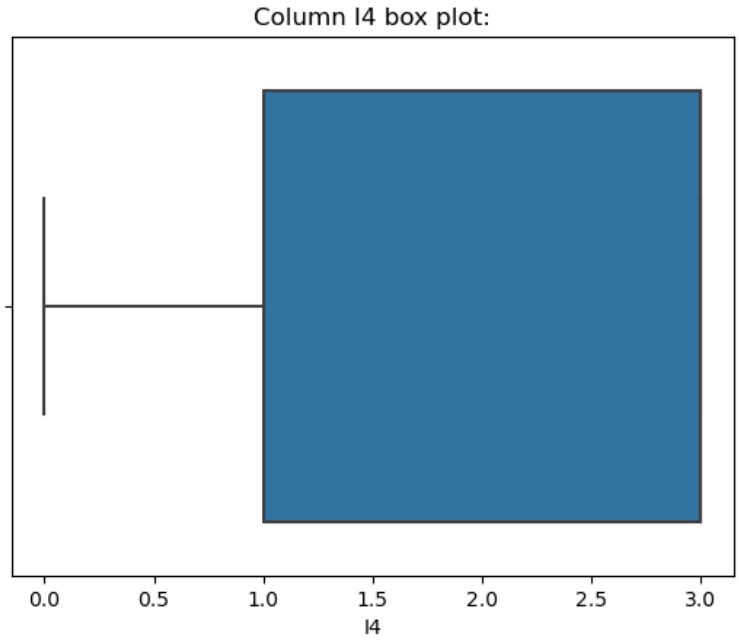
Stats	I4
count	424.00
mean	2.21
std	0.93
min	0.00
25%	1.00
50%	3.00
75%	3.00
max	3.00

There are no outliers in column I4

No inliers detected in column I4

Column I4 values frequency:





Column: I5

Column Type: float64

Sample value: 2.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

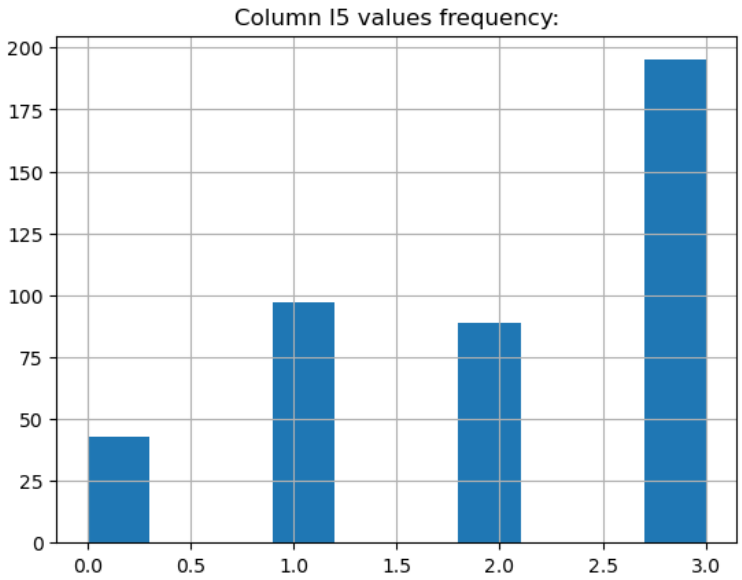
Warning: too missing values in column I5

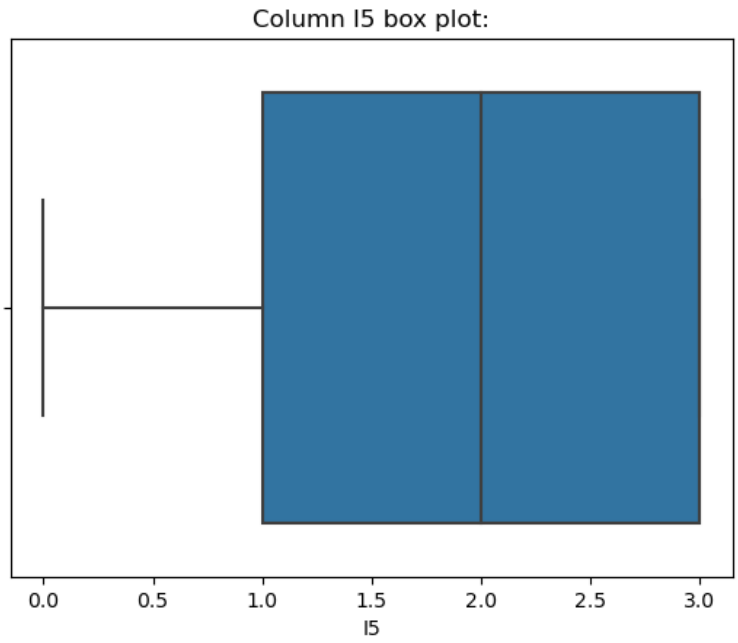
Stats	I5
count	424.00
mean	2.03
std	1.05
min	0.00
25%	1.00
50%	2.00
75%	3.00
max	3.00

There are no outliers in column I5

No inliers detected in column I5

Column I5 values frequency:





Column: I6

Column Type: float64

Sample value: 3.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

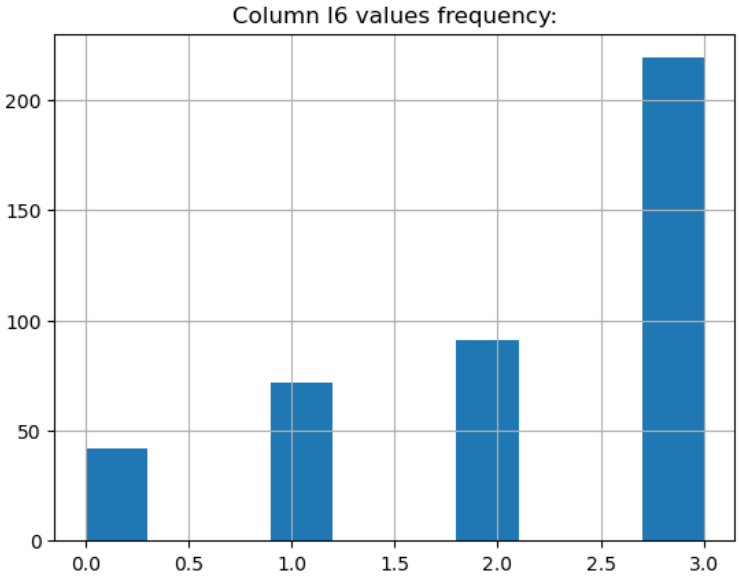
Warning: too missing values in column I6

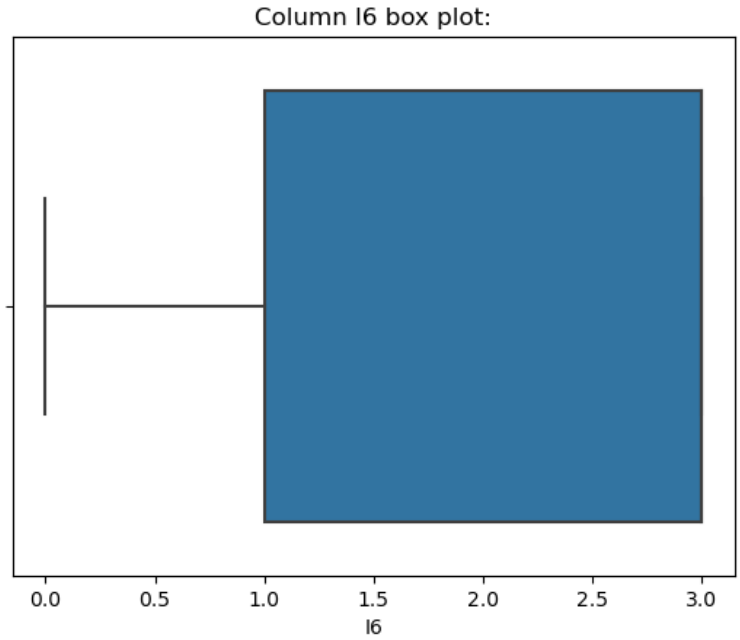
Stats	I6
count	424.00
mean	2.15
std	1.03
min	0.00
25%	1.00
50%	3.00
75%	3.00
max	3.00

There are no outliers in column I6

No inliers detected in column I6

Column I6 values frequency:





Column: I7

Column Type: float64

Sample value: 3.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

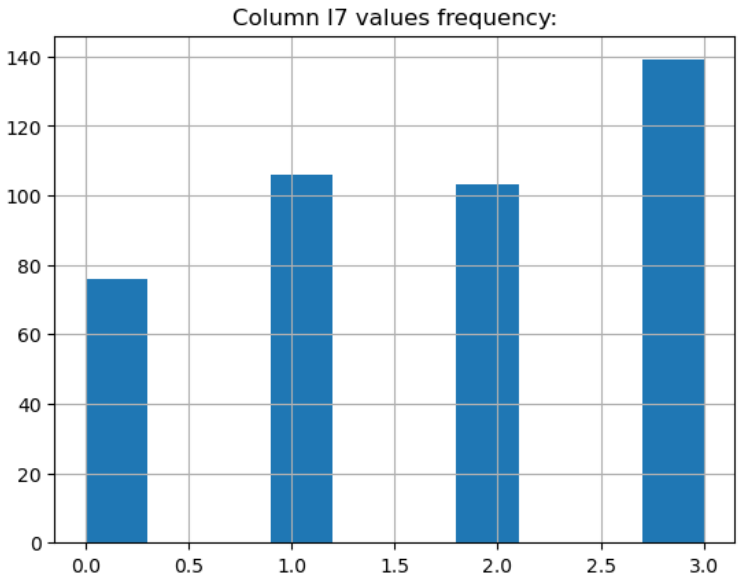
Warning: too missing values in column I7

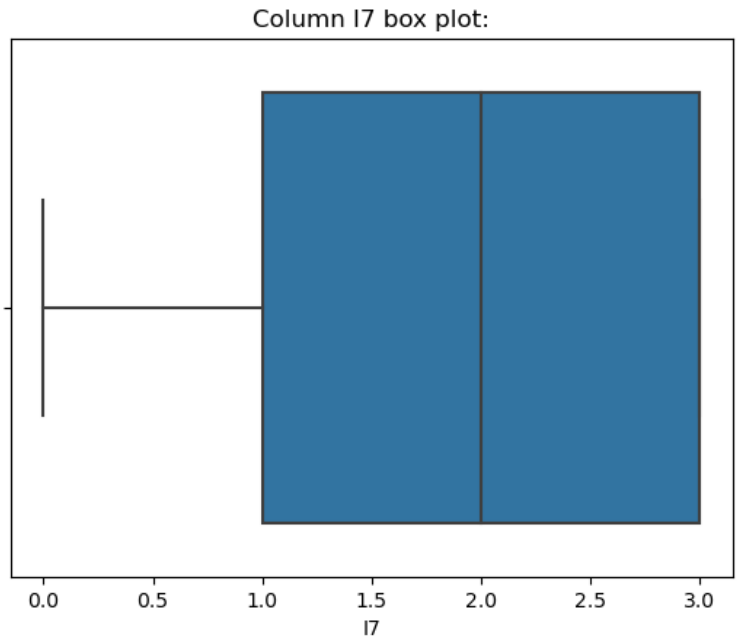
Stats	I7
count	424.00
mean	1.72
std	1.10
min	0.00
25%	1.00
50%	2.00
75%	3.00
max	3.00

There are no outliers in column I7

No inliers detected in column I7

Column I7 values frequency:





Column: I8

Column Type: float64

Sample value: 1.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

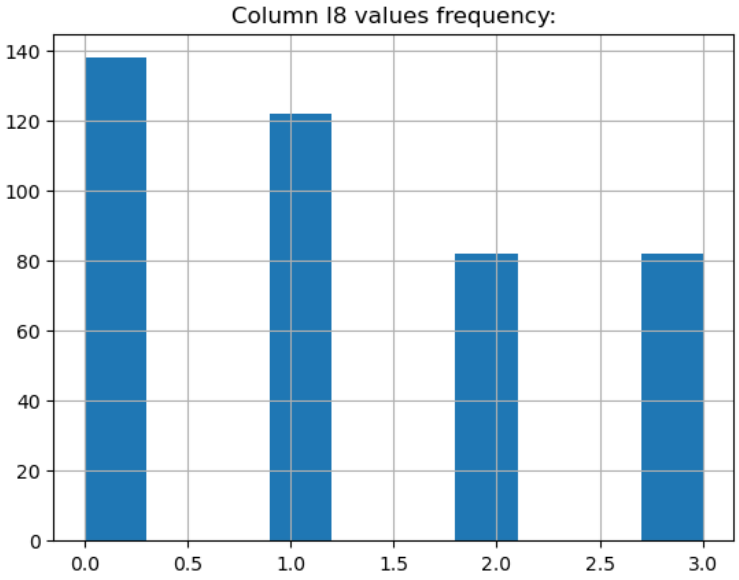
Warning: too missing values in column I8

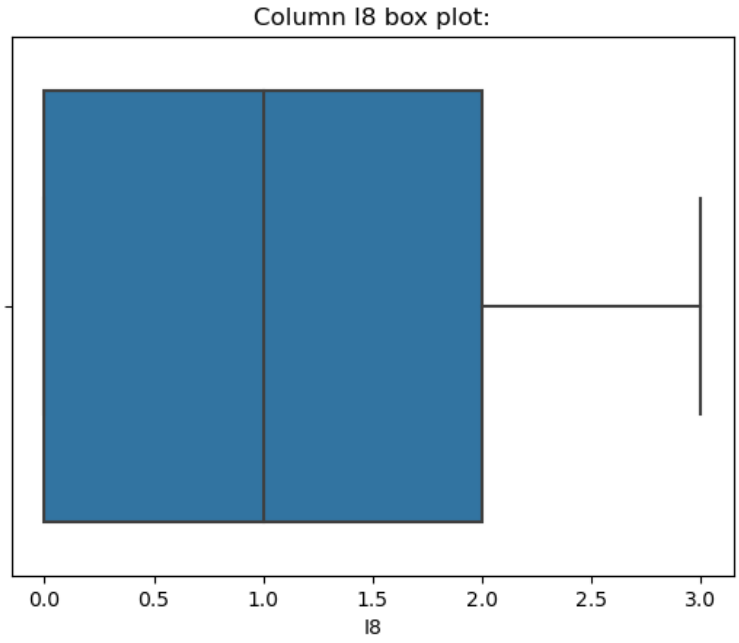
Stats	I8
count	424.00
mean	1.25
std	1.11
min	0.00
25%	0.00
50%	1.00
75%	2.00
max	3.00

There are no outliers in column I8

No inliers detected in column I8

Column I8 values frequency:





Column: I9

Column Type: float64

Sample value: 0.0

Number of unique values: 5

Percentage of unique values: 0.50%

Number of missing values: 586

Percentage of missing values: 58.02%

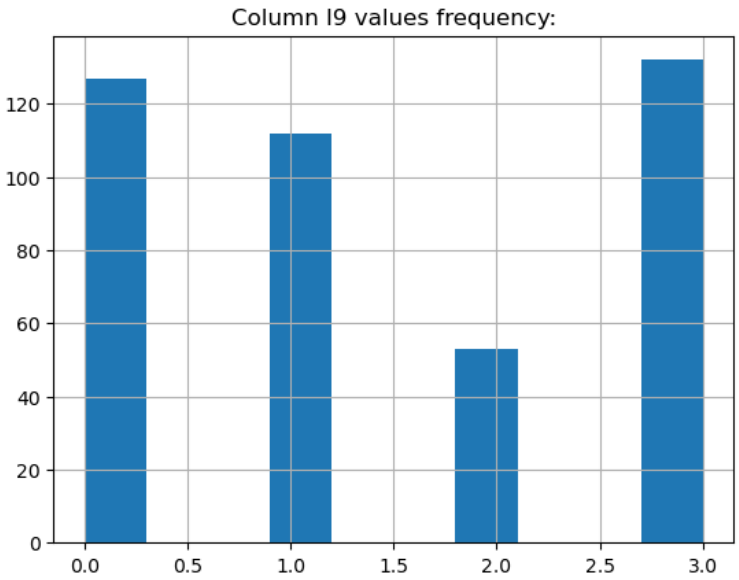
Warning: too missing values in column I9

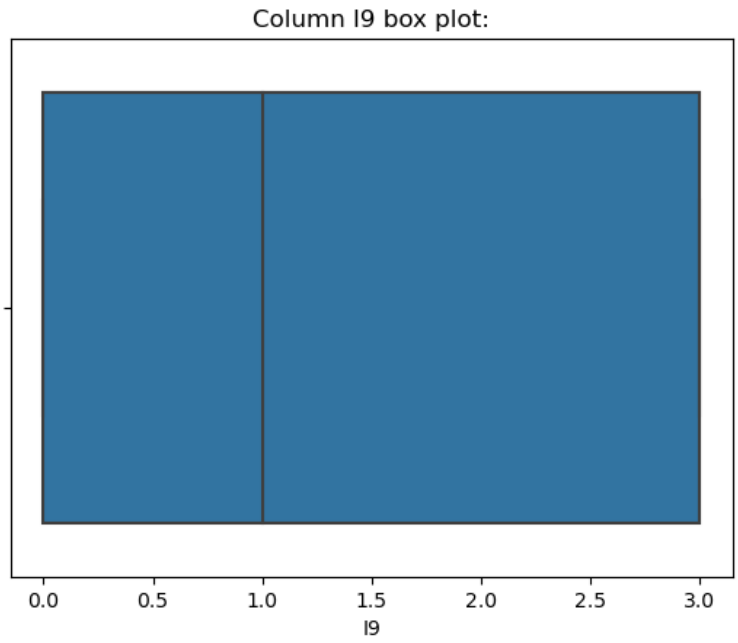
Stats	I9
count	424.00
mean	1.45
std	1.21
min	0.00
25%	0.00
50%	1.00
75%	3.00
max	3.00

There are no outliers in column I9

No inliers detected in column I9

Column I9 values frequency:





Column: Depression

Column Type: object

See Converted_Depression column below for a detailed analysis.

Sample value: POSITIVE

Number of unique values: 3

Percentage of unique values: 0.30%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: BAI

Column Type: float64

Sample value: 19.0

Number of unique values: 61

Percentage of unique values: 6.04%

Number of missing values: 576

Percentage of missing values: 57.03%

Warning: too missing values in column BAI

Stats	BAI
count	434.00
mean	29.86
std	13.42
min	1.00
25%	20.00
50%	29.00
75%	39.00
max	72.00

Warning: There are outliers in column BAI

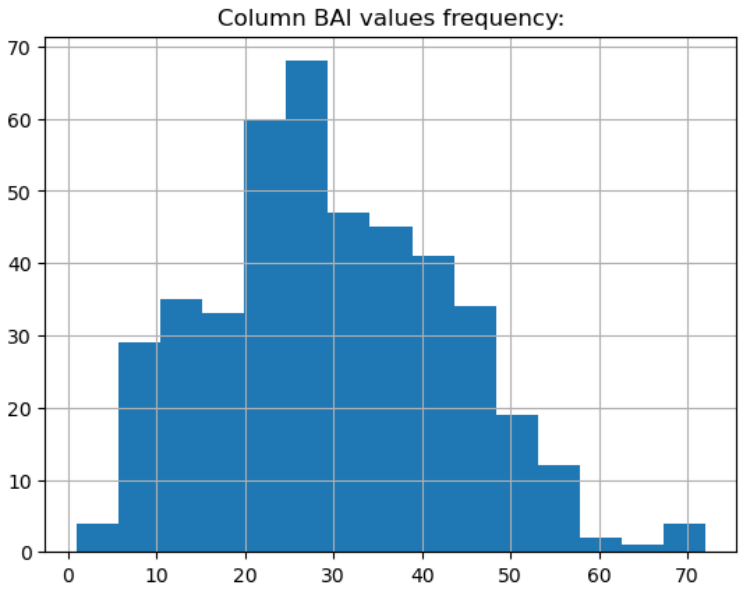
Sample of outliers: [69.0, 72.0, 72.0, 72.0]

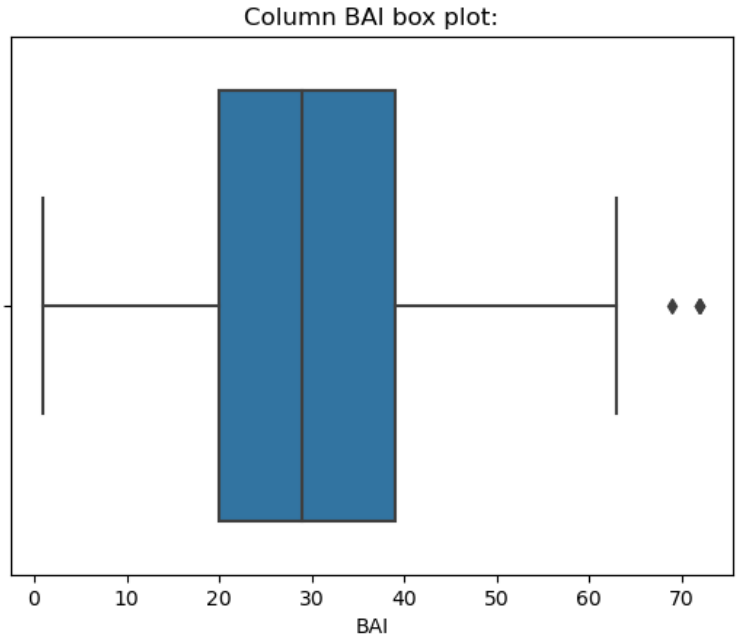
Sample of outliers over Q3: [69.0, 72.0, 72.0, 72.0]

Sample of outliers below Q1: []

No inliers detected in column BAI

Column BAI values frequency:





Column: Anxiety

Column Type: object

See Converted_Anxiety column below for a detailed analysis.

Sample value: POSITIVE

Number of unique values: 3

Percentage of unique values: 0.30%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: PSS

Column Type: float64

Sample value: 41.0

Number of unique values: 44

Percentage of unique values: 4.36%

Number of missing values: 665

Percentage of missing values: 65.84%

Warning: too missing values in column PSS

Stats	PSS
count	345.00
mean	37.61
std	8.30
min	12.00
25%	33.00
50%	38.00
75%	43.00
max	69.00

Warning: There are outliers in column PSS

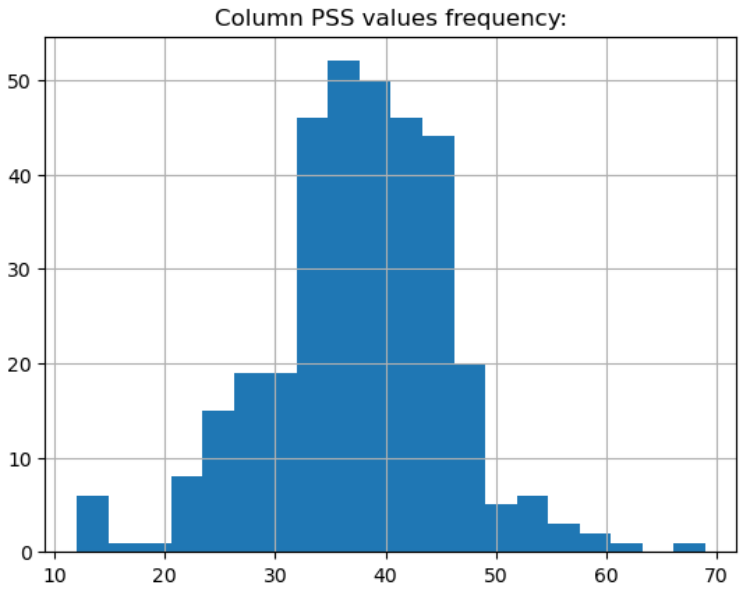
Sample of outliers: [13.0, 62.0, 12.0, 69.0, 60.0, 16.0, 60.0, 13.0, 13.0, 13.0]

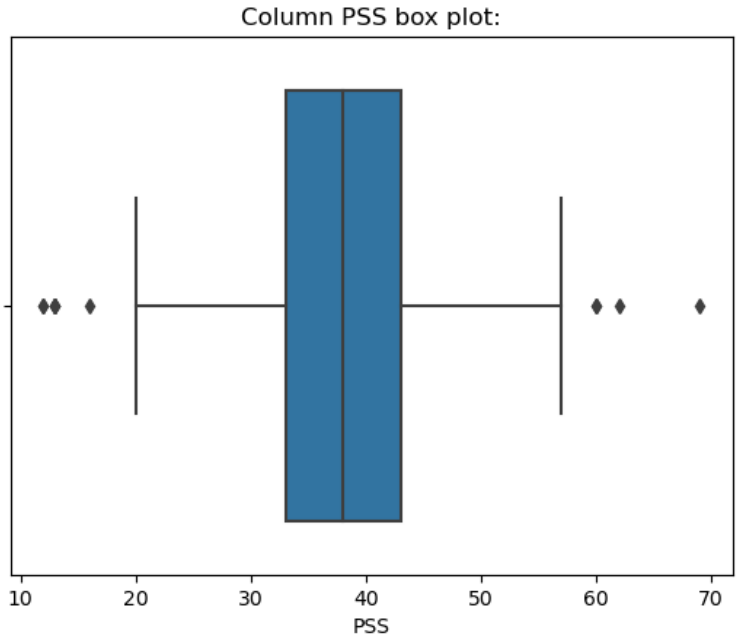
Sample of outliers over Q3: [60.0, 60.0, 62.0, 69.0]

Sample of outliers below Q1: [16.0, 13.0, 13.0, 13.0, 13.0, 12.0, 12.0]

No inliers detected in column PSS

Column PSS values frequency:





Column: AN

Column Type: float64

Sample value: 37.0

Number of unique values: 42

Percentage of unique values: 4.16%

Number of missing values: 570

Percentage of missing values: 56.44%

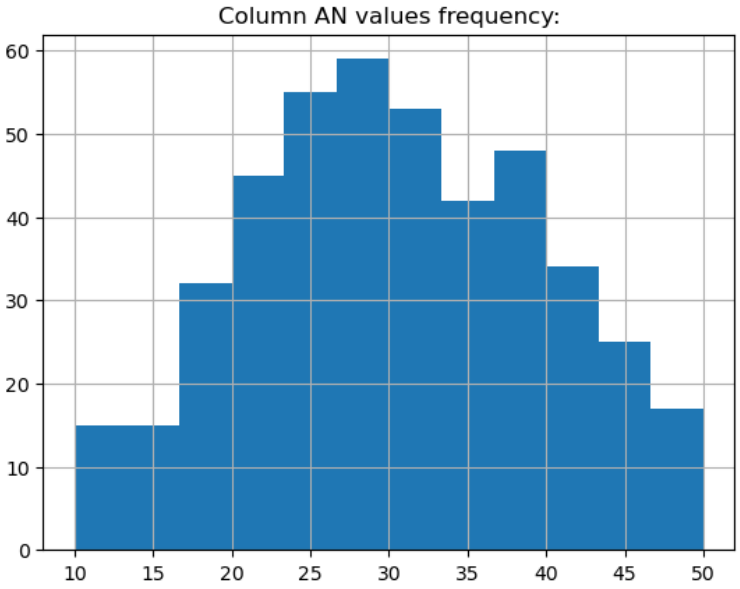
Warning: too missing values in column AN

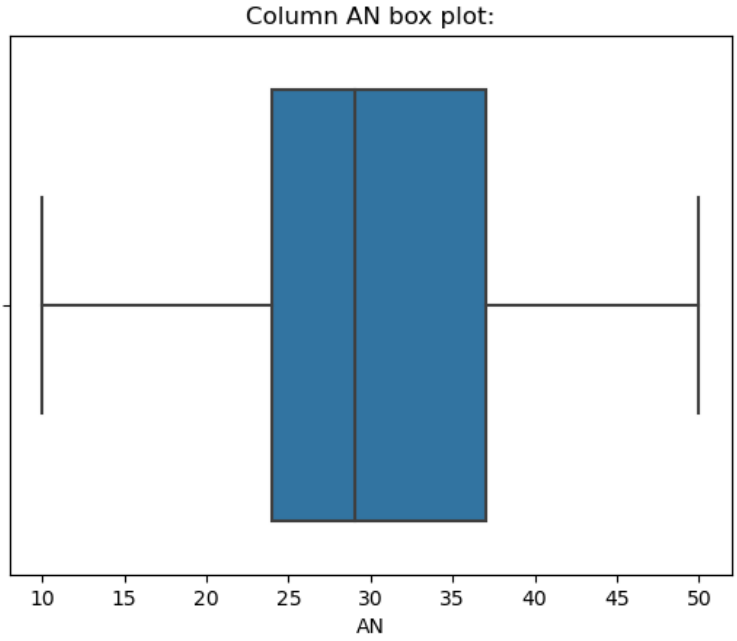
Stats	AN
count	440.00
mean	30.22
std	9.23
min	10.00
25%	24.00
50%	29.00
75%	37.00
max	50.00

There are no outliers in column AN

No inliers detected in column AN

Column AN values frequency:





Column: AP

Column Type: float64

Sample value: 14.0

Number of unique values: 35

Percentage of unique values: 3.47%

Number of missing values: 570

Percentage of missing values: 56.44%

Warning: too missing values in column AP

Stats	AP
count	440.00
mean	23.07
std	7.11
min	10.00
25%	17.00
50%	23.00
75%	28.00
max	47.00

Warning: There are outliers in column AP

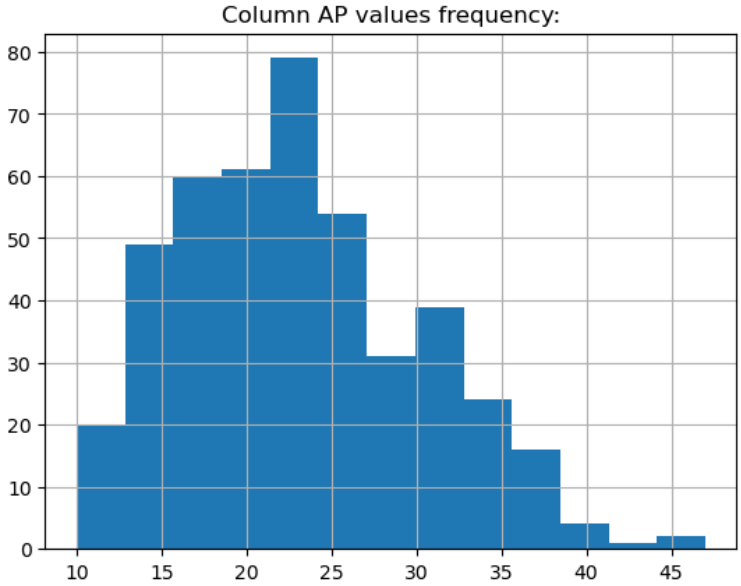
Sample of outliers: [47.0, 46.0]

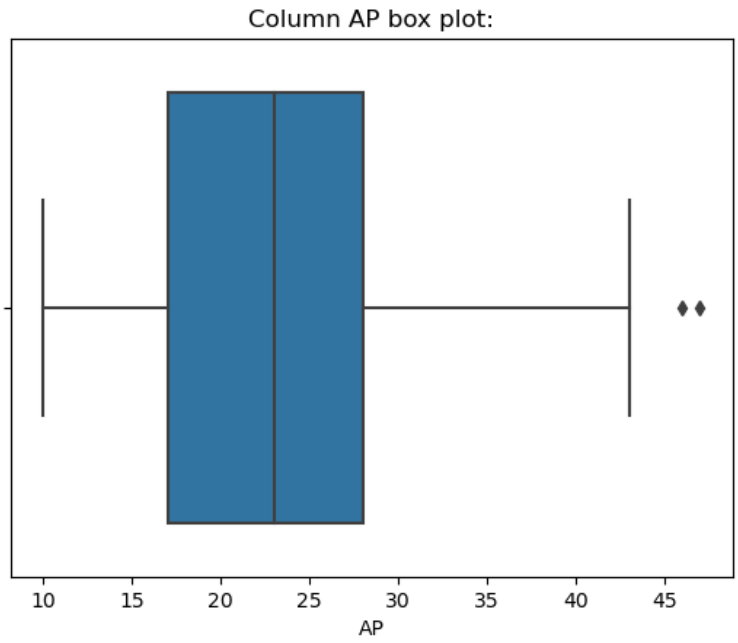
Sample of outliers over Q3: [46.0, 47.0]

Sample of outliers below Q1: []

No inliers detected in column AP

Column AP values frequency:





Column: Text_Card1_en

Column Type: object

See Converted_Text_Card1_en column below for a detailed analysis.

Sample value: ... a common child, like any other, left home, went to school, had friends, his family loved hi

Number of unique values: 943

Percentage of unique values: 93.37%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Text_Card9VH_en

Column Type: object

See Converted_Text_Card9VH_en column below for a detailed analysis.

Sample value: ... a man who was drunk with his friends

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 0

Percentage of missing values: 0.00%

Column: Text_Card11_en

Column Type: object

See Converted_Text_Card11_en column below for a detailed analysis.

Sample value: ... a magical path, capable of leading you to your deepest desire, no matter what it was, eve

Number of unique values: 941

Percentage of unique values: 93.17%

Number of missing values: 1

Percentage of missing values: 0.10%

Column: Text_Card13HM_en

Column Type: object

See Converted_Text_Card13HM_en column below for a detailed analysis.

Sample value: ... a man who realized that he was not happy, even with a wife who loved him, because his

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 1

Percentage of missing values: 0.10%

Column: Converted_Timestamp

Column Type: datetime64[ns, UTC]

Sample value: 2021-03-17 06:55:09.874000+00:00

Number of unique values: 1010

Percentage of unique values: 100.00%

Number of missing values: 0

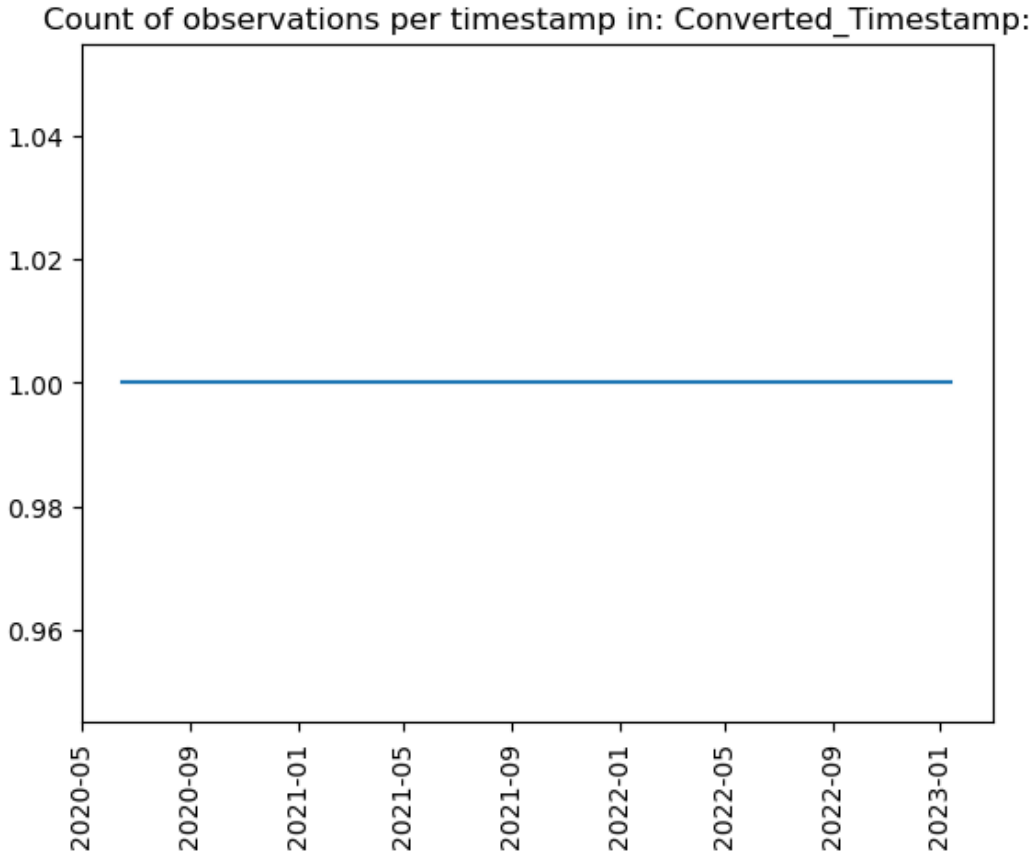
Percentage of missing values: 0.00%

Global Series Start Datetime: 2020-06-15 14:35:58.952000+00:00

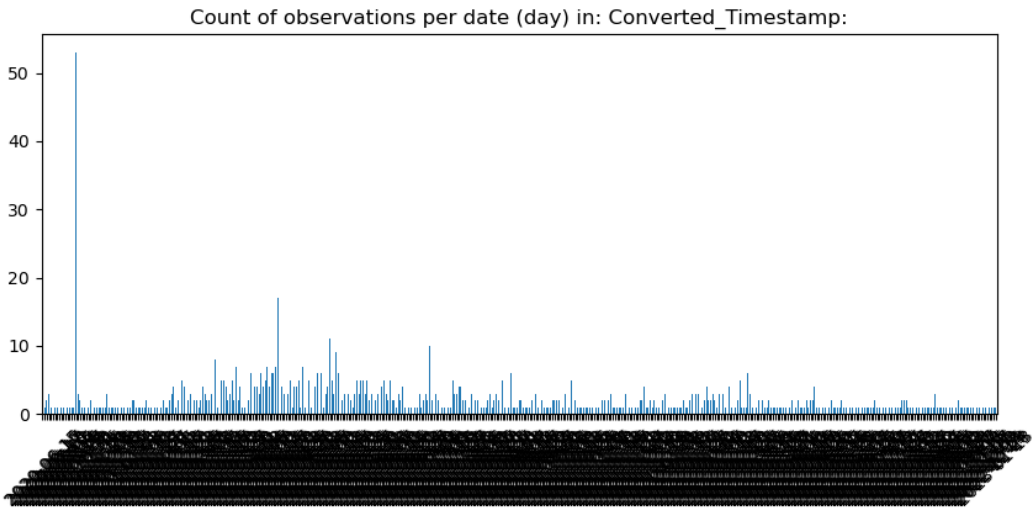
Global Series End Datetime: 2023-01-13 04:44:10.895000+00:00

Global Series Period: 941 days 14:08:11.943000

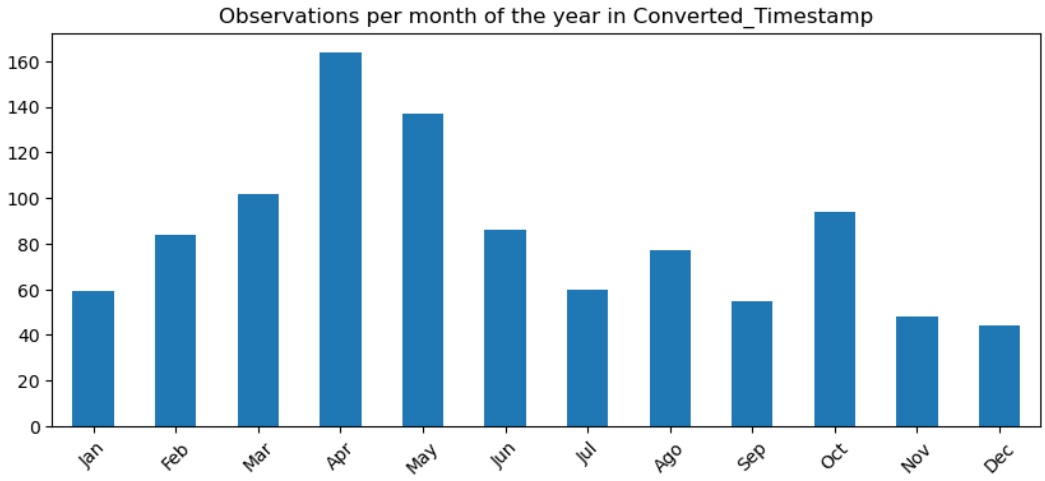
Count of observations per timestamp in: Converted_Timestamp:



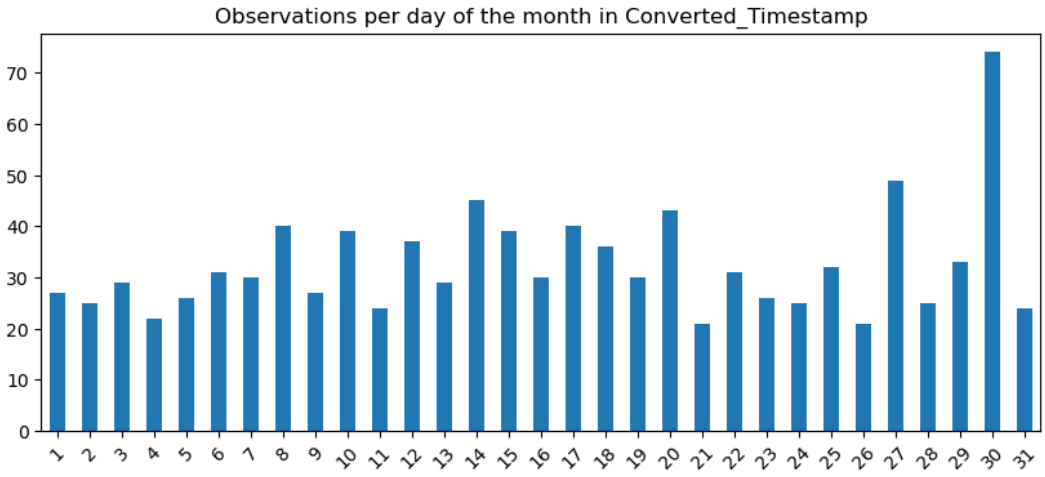
Count of observations per date (day) in: Converted_Timestamp:



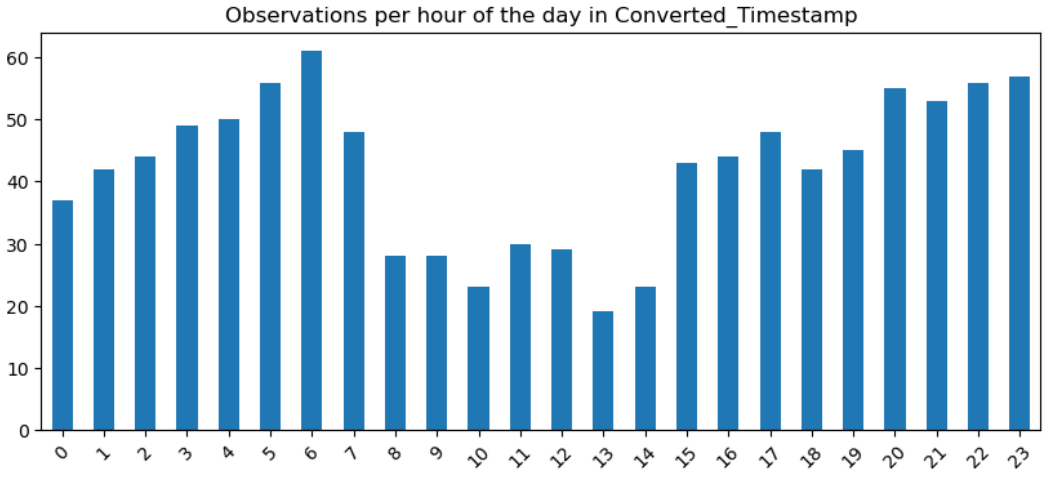
Observations per month of the year in Converted_Timestamp



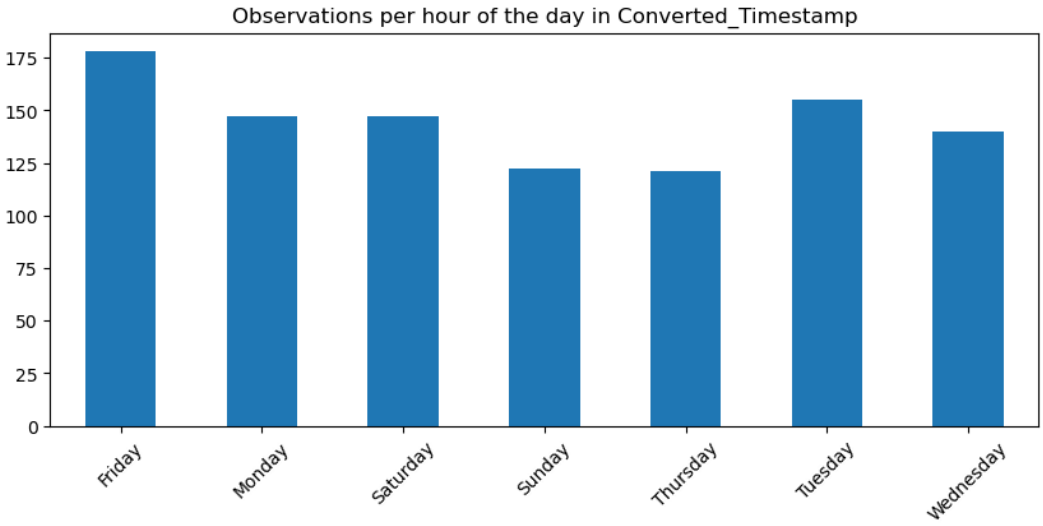
Observations per day of the month in Converted_Timestamp



Observations per hour of the day in Converted_Timestamp

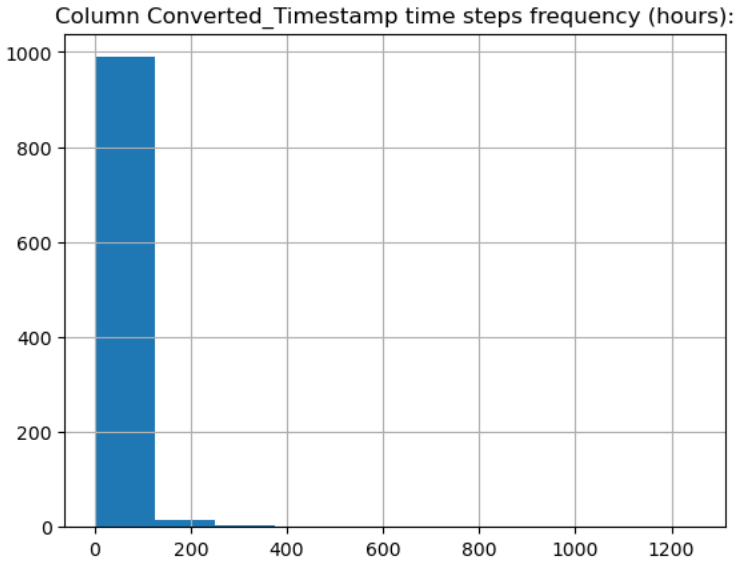


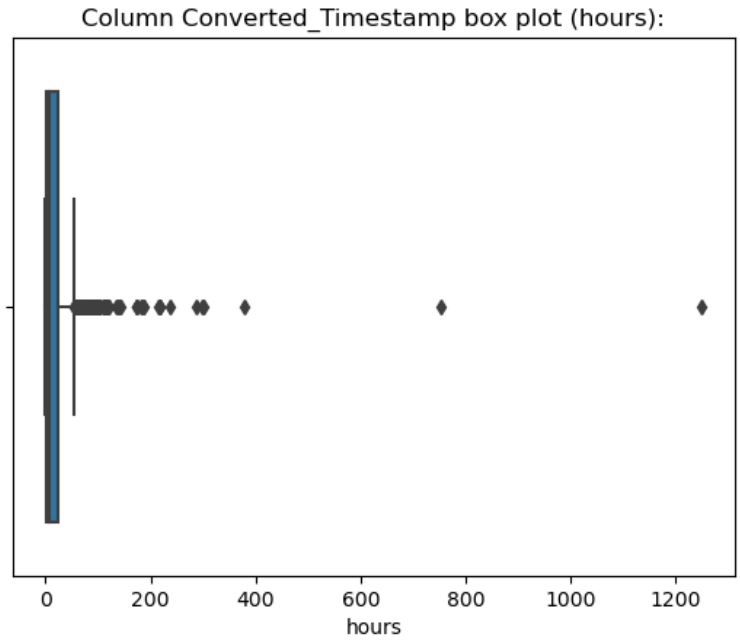
Observations per day of the week in Converted_Timestamp



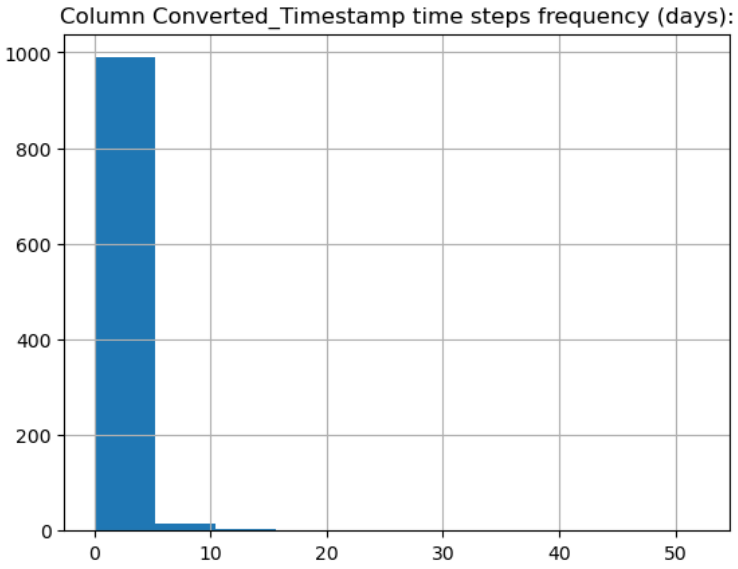
Time step is irregular

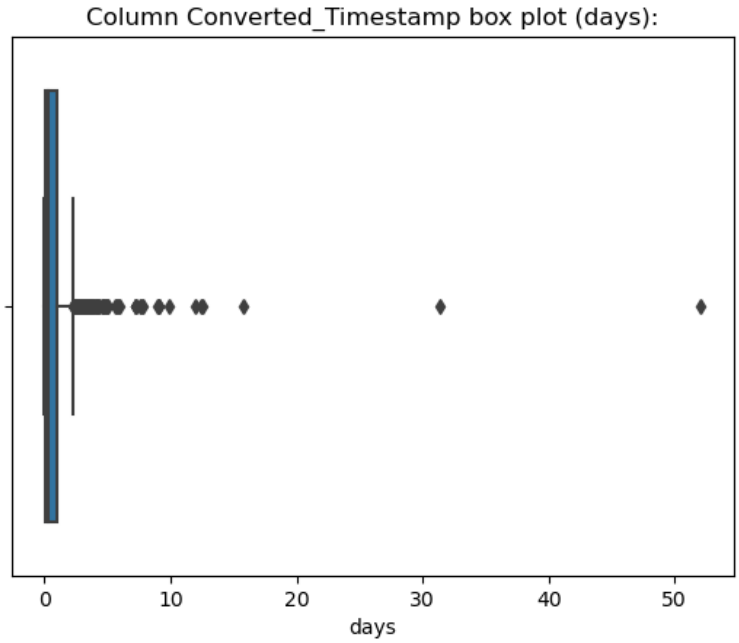
Column Converted_Timestamp time steps frequency (hours):





Column Converted_Timestamp time steps frequency (days):





Column: Converted_Timestamp (1)

Column Type: datetime64[ns, UTC]

Sample value: 2021-03-17 07:39:24.457000+00:00

Number of unique values: 947

Percentage of unique values: 93.76%

Number of missing values: 0

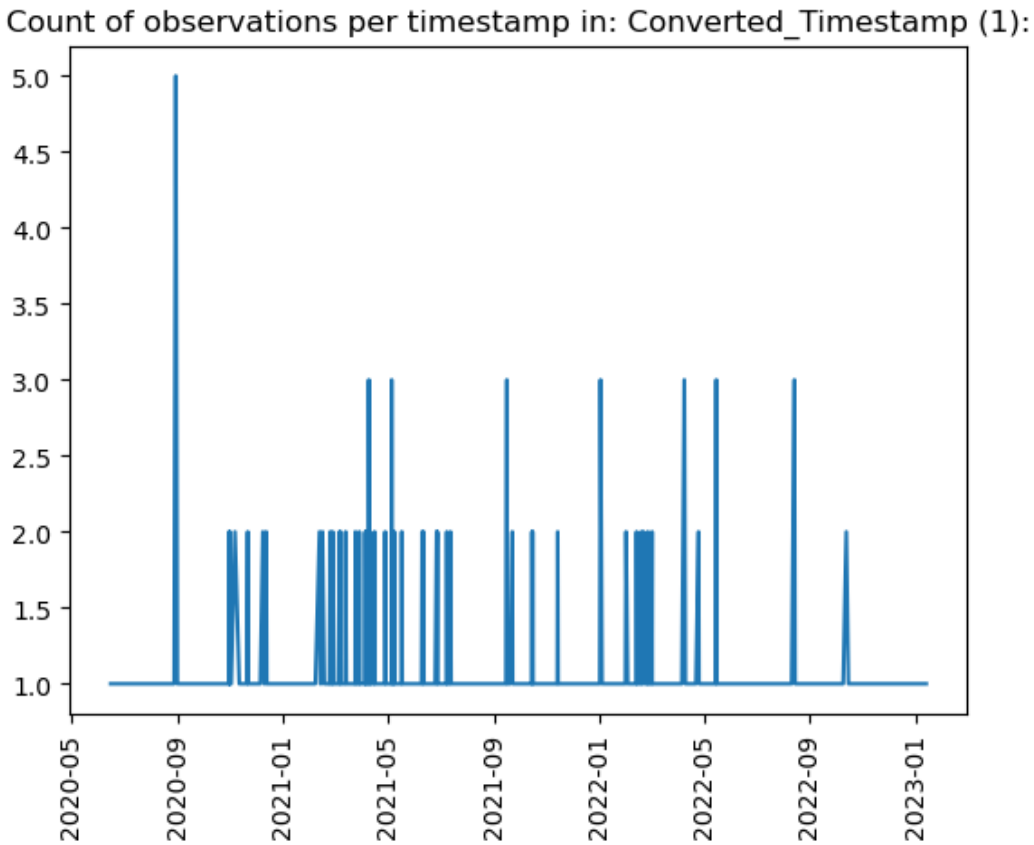
Percentage of missing values: 0.00%

Global Series Start Datetime: 2020-06-15 14:28:25.641000+00:00

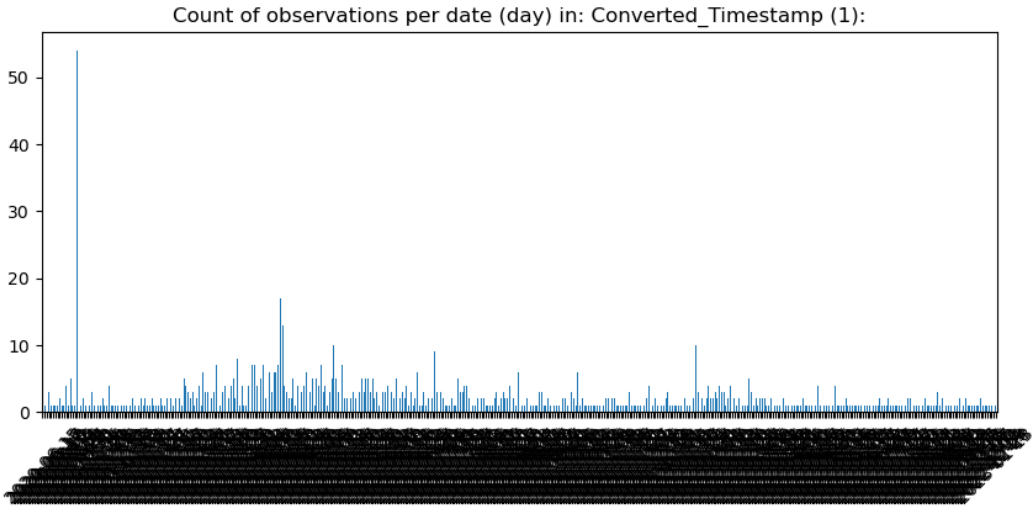
Global Series End Datetime: 2023-01-13 04:51:51.436000+00:00

Global Series Period: 941 days 14:23:25.795000

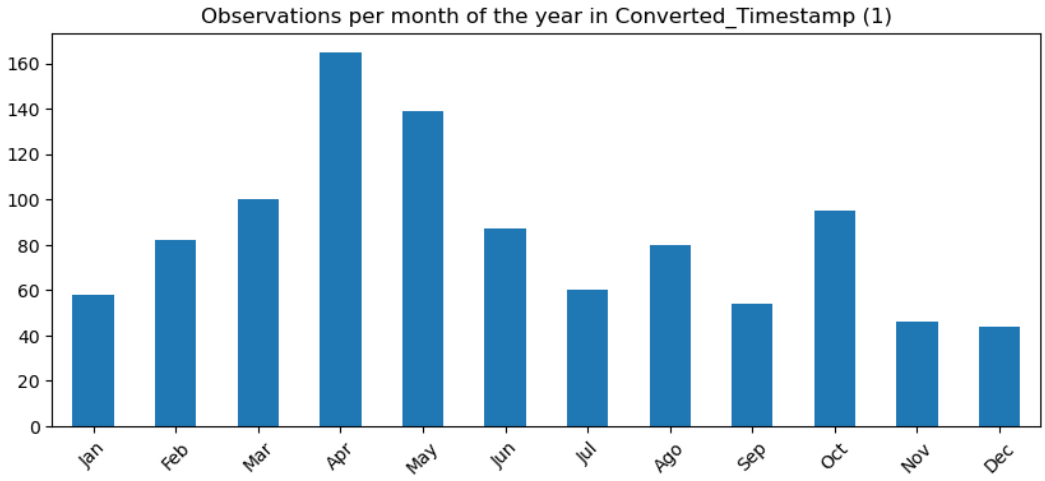
Count of observations per timestamp in: Converted_Timestamp (1):



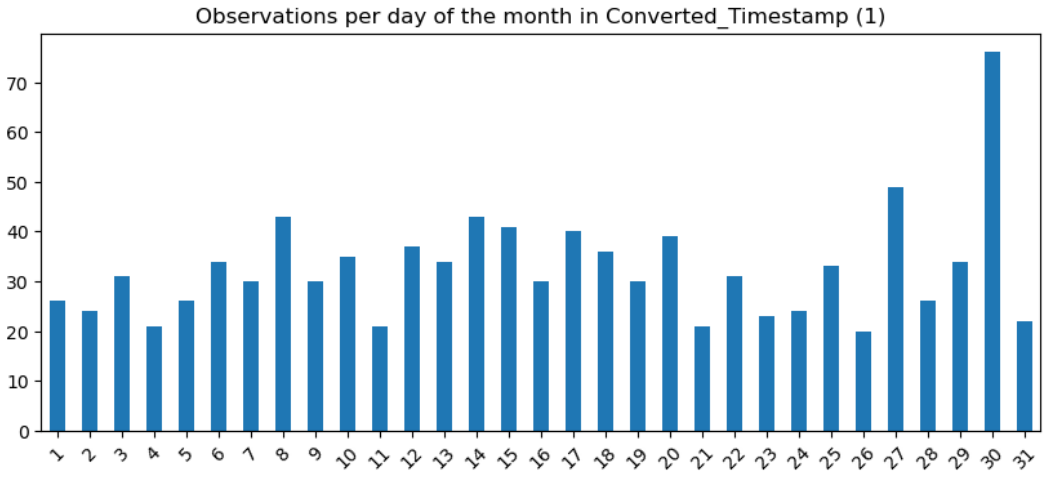
Count of observations per date (day) in: Converted_Timestamp (1):



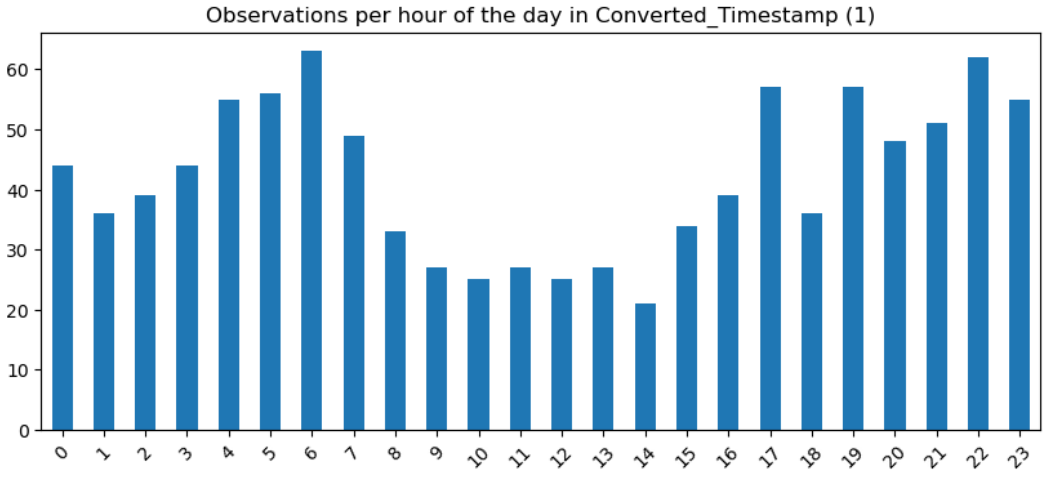
Observations per month of the year in Converted_Timestamp (1)



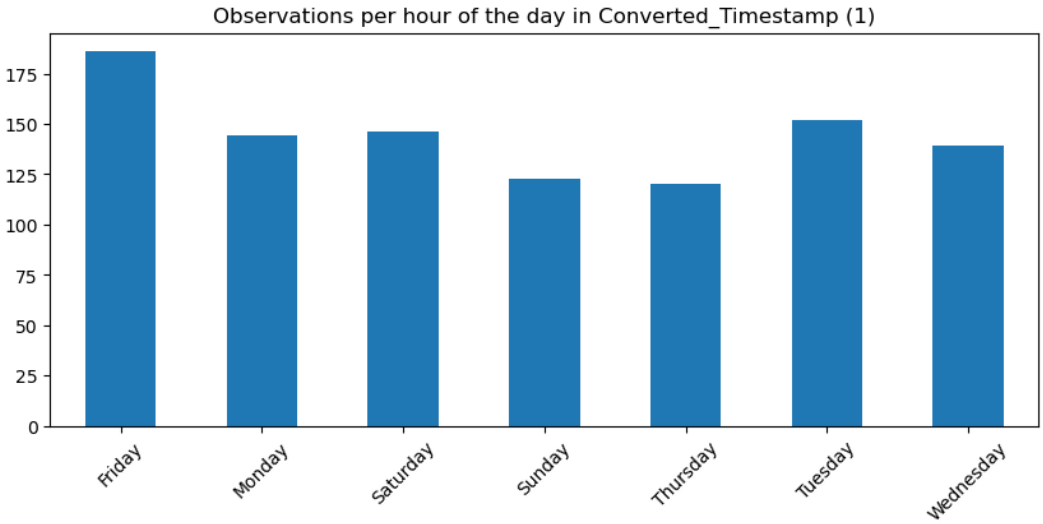
Observations per day of the month in Converted_Timestamp (1)



Observations per hour of the day in Converted_Timestamp (1)

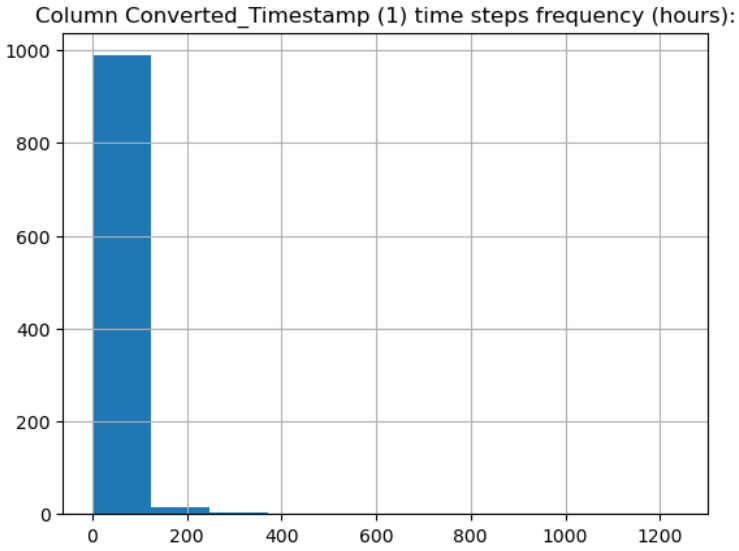


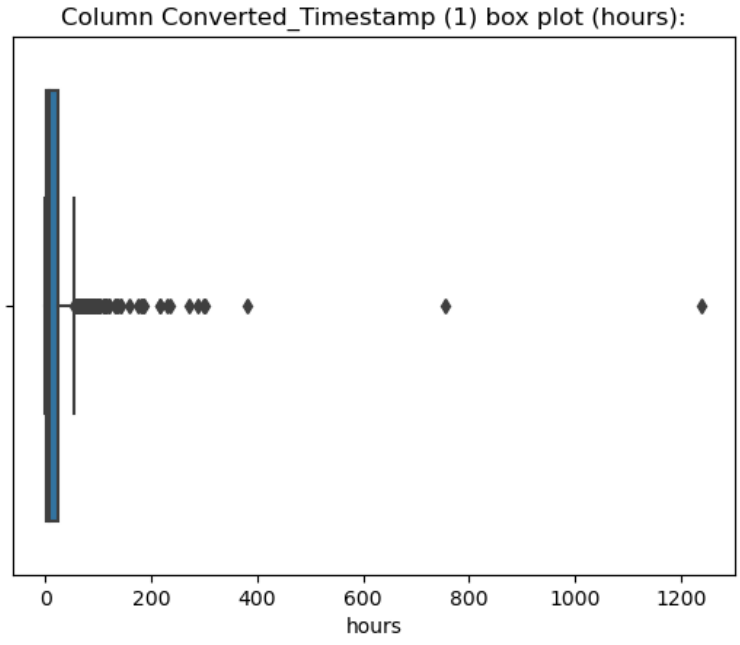
Observations per day of the week in Converted_Timestamp (1)



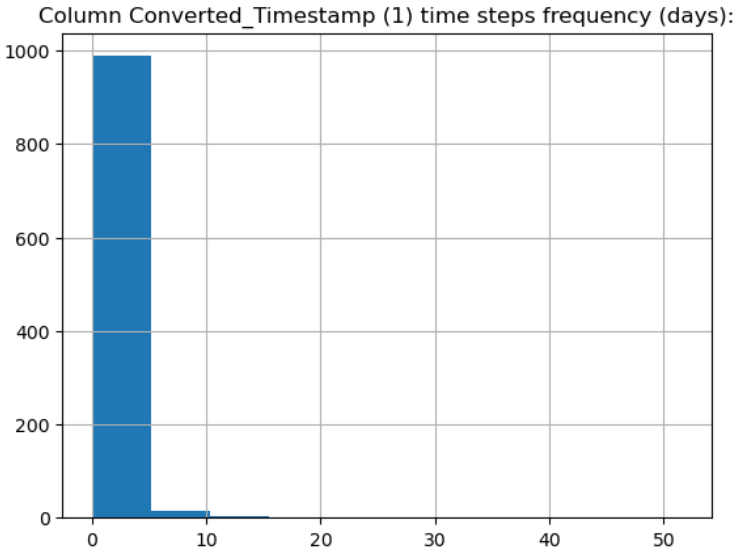
Time step is irregular

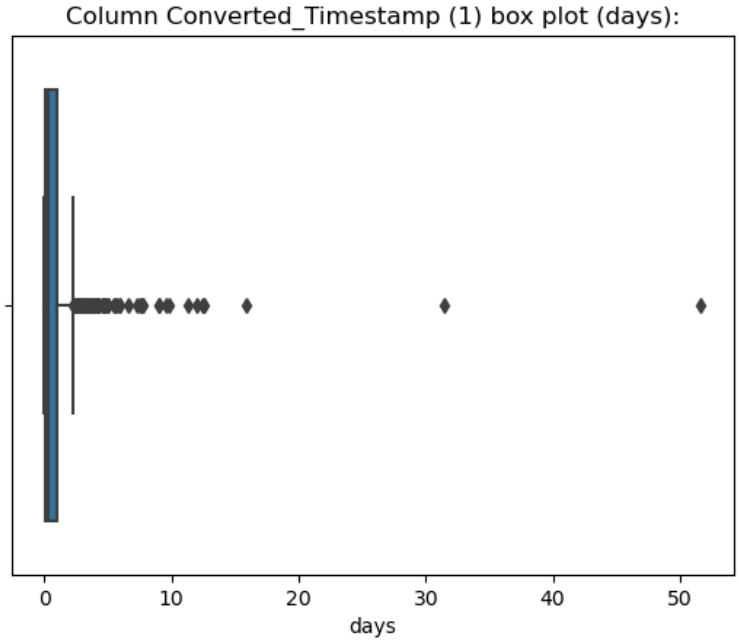
Column Converted_Timestamp (1) time steps frequency (hours):



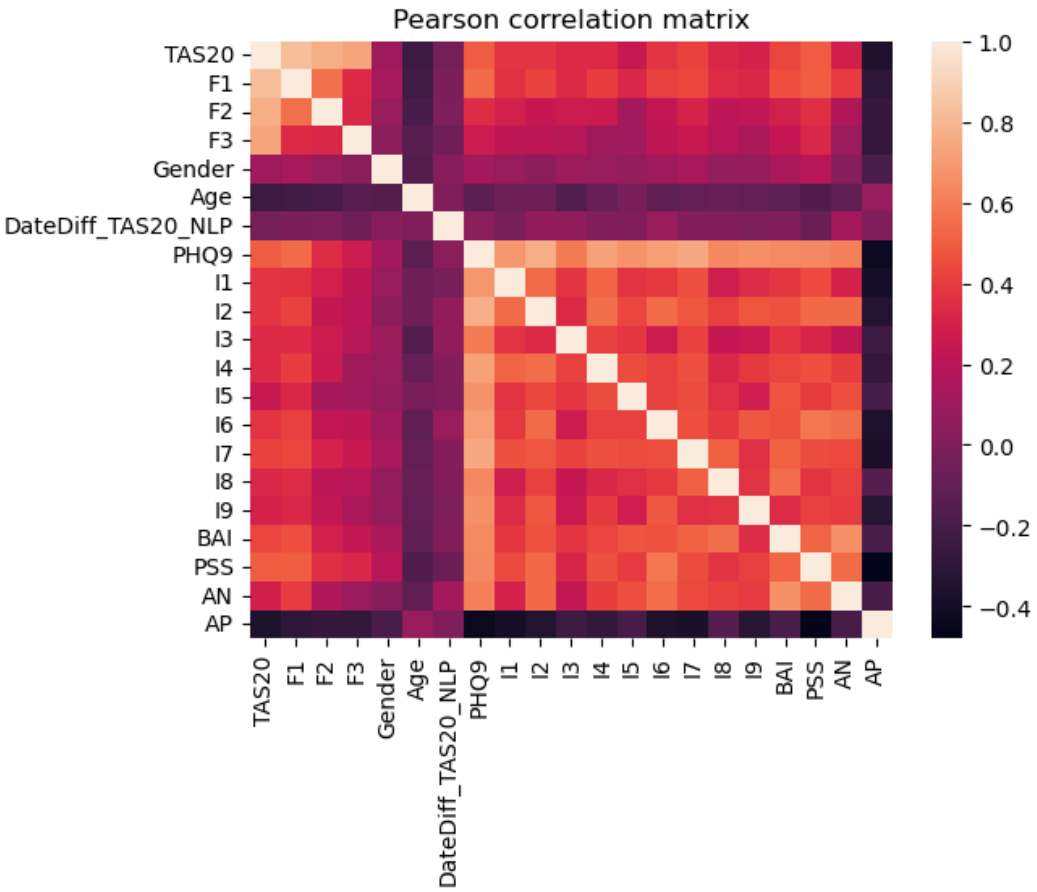


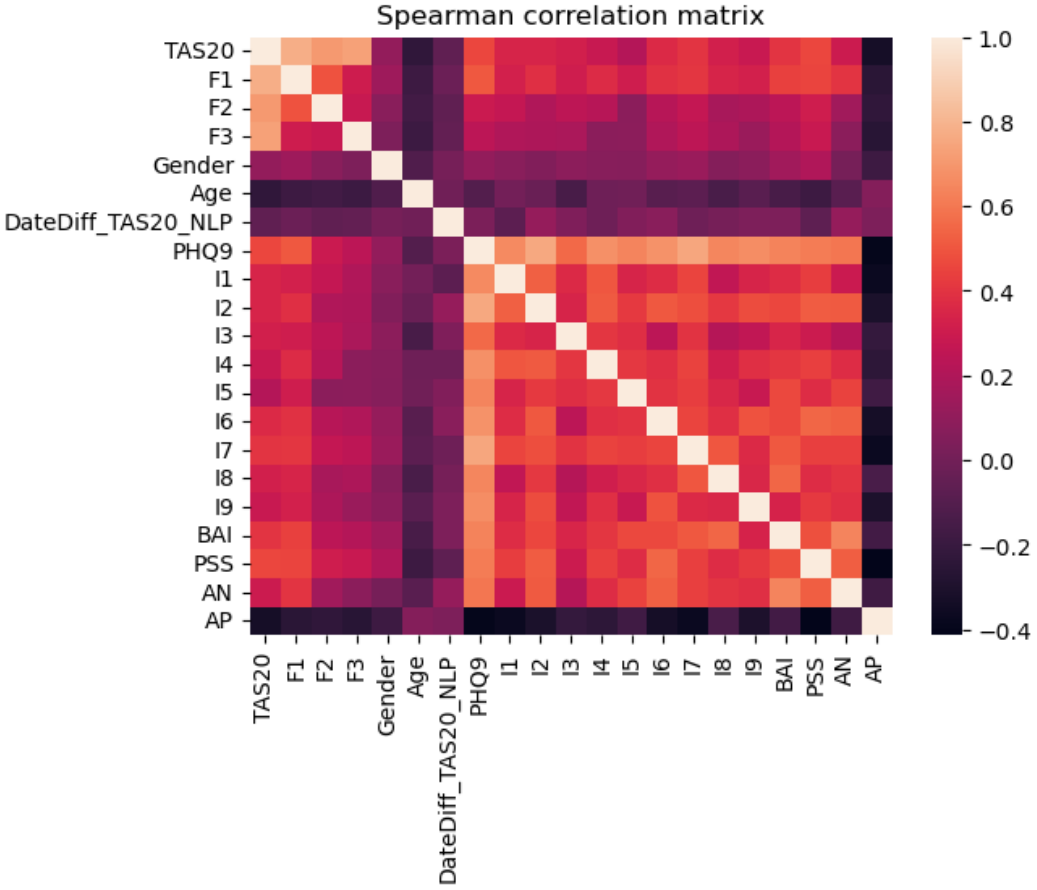
Column Converted_Timestamp (1) time steps frequency (days):





Correlational Analysis


Pearson correlation matrix
TAS20
F1
F2
F3
Gender
Age
DateDiff_TAS20_NLP
PHQ9
I1
I2
I3
I4
I5
I6
I7
I8
I9
BAI
PSS
AN
AP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Spearman correlation matrix
TAS20
F1
F2
F3
Gender
Age
DateDiff_TAS20_NLP
PHQ9
I1
I2
I3
I4
I5
I6
I7
I8
I9
BAI
PSS
AN
AP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4