Discoveries on the forty datasets

1- 186 - Wine Quality - Business

Dataset analysed: winequality-red.csv from https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv (dataset_file_url column from Fortydatasets.xlsx) has header

fixed acidity (acidity):
 Numerical format: All 1599 values are numerical in the range (4.6:15.9).

volatile acidity (acidity):
 Numerical format: All 1599 values are numerical in the range (0.12:1.58).

citric acid (acid):
 Numerical format: All 1599 values are numerical in the range (0.0:1.0).

residual sugar (sugar):
 Numerical format: All 1599 values are numerical in the range (0.9:15.5).

chlorides (chloride):
 Numerical format: All 1599 values are numerical in the range (0.012:0.611).

free sulfur dioxide (sulfur):
 Numerical format: All 1599 values are numerical in the range (1.0:72.0).

total sulfur dioxide (total):
 Numerical format: All 1599 values are numerical in the range (6.0:289.0).

density:
 Numerical >=0 format: All 1599 values are numerical and greater or equal to 0 in the range (0.99007:1.00369).

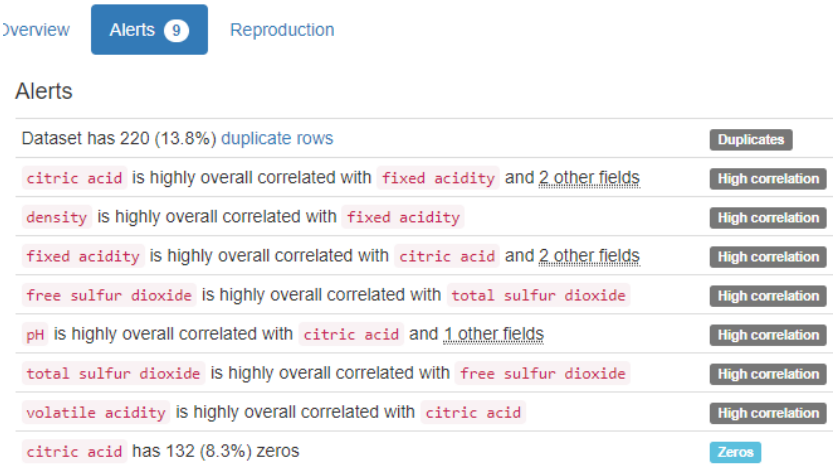
pH:
 pH format: All 1599 values are numerical and valid in the range [0, 14].
Actual range of values: (2.74 : 4.01)

sulphates (sulphate):
 Numerical format: All 1599 values are numerical in the range (0.33:2.0).

alcohol:
 Numerical format: All 1599 values are numerical in the range (8.4:14.9).

quality (score):
 Numerical >=0 format: All 1599 values are numerical and greater or equal to 0 in the range (3:8).

Last run on: 2024-01-26 12:45:06



YData did not find anything that we search for.

2- 222 - Bank Marketing - Business

Dataset analysed: bank-additional/bank-additional-full.csv from bank-additional.zip obtained at https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank-additional.zip (dataset_file_url column from Fortydatasets.xlsx) Has header

```
age:
  Age format: All 41188 values are numerical and valid in the range [0, 130].
Actual range of values: (17 : 98)
```

```
job:
 All 41188 values are correctly categorical.

  Categorical format with 12 unique values:
     Category  Frequency
        admin.      10422
  blue-collar       9254
   technician       6743
     services       3969
   management       2924
      retired       1720
 entrepreneur       1456
self-employed       1421
    housemaid       1060
   unemployed       1014
      student        875
      unknown        330

marital:
All 41188 values are correctly categorical.

  Categorical format with 4 unique values:
Category  Frequency
 married      24928
  single      11568
divorced       4612
 unknown         80

education:
All 41188 values are correctly categorical.

  Categorical format with 8 unique values:
           Category  Frequency
  university.degree      12168
        high.school       9515
           basic.9y       6045
professional.course       5243
           basic.4y       4176
           basic.6y       2292
            unknown       1731
         illiterate         18

default (categorical):
All 41188 values are correctly categorical.

  Categorical format with 3 unique values:
Category  Frequency
      no      32588
 unknown       8597
     yes          3

housing (categorical):
All 41188 values are correctly categorical.

  Categorical format with 3 unique values:
Category  Frequency
     yes      21576
      no      18622
 unknown        990

loan (categorical):
All 41188 values are correctly categorical.

  Categorical format with 3 unique values:
Category  Frequency
      no      33950
     yes       6248
 unknown        990


contact (categorical):
All 41188 values are correctly categorical.

  Categorical format with 2 unique values:
 Category  Frequency
 cellular      26144
telephone      15044


month:
  Month format: All 41188 month values are valid.

Frequency Distribution:
    Month  Frequency
      May      13769
     July       7174
   August       6178
     June       5318
 November       4101
    April       2632
  October        718
September        570
    March        546
```

```
 December        182

day_of_week (day of week):
Weekday format: All 41188 weekday values are valid.

  Weekday format:
Frequency Distribution:
  Weekday  Frequency
 Thursday       8623
   Monday       8514
Wednesday       8134
  Tuesday       8090
   Friday       7827

duration:
  Numerical >=0 format: All 41188 values are correct (numeric and >=0) in the range (0:4918).

campaign (number):
  Numerical format: All 41188 values are numerical in the range (1:56).

pdays (number):
  Numerical format: All 41188 values are numerical in the range (0:999).

previous (number):
  Numerical format: All 41188 values are numerical in the range (0:7).

poutcome (categorical):
All 41188 values are correctly categorical.

  Categorical format with 3 unique values:
   Category  Frequency
nonexistent      35563
    failure       4252
    success       1373


emp.var.rate (rate):
  Numerical format: All 41188 values are numerical in the range (-3.4:1.4).

cons.price.idx (price):
  Numerical format: All 41188 values are numerical in the range (92.201:94.767).

cons.conf.idx (indicator):
  Numerical format: All 41188 values are numerical in the range (-50.8:-26.9).

euribor3m (indicator):
  Numerical format: All 41188 values are numerical in the range (0.634:5.045)

nr.employed (indicator):
  Numerical format: All 41188 values are numerical in the range (4963.6:5228.1).

y - has the client subscribed a term deposit? (binary):
  Binary format: All 41188 binary values are valid.

Frequency Distribution:
Value  Frequency
   no      36548
  yes       4640
```

Last run on: 2024-01-25 00:02:33

Overview | Alerts 5 | Reproduction

Alerts

Dataset has 12 (< 0.1%) duplicate rows	Duplicates
`default` is highly imbalanced (53.3%)	Imbalance
`loan` is highly imbalanced (51.3%)	Imbalance
`poutcome` is highly imbalanced (56.8%)	Imbalance
`previous` has 35563 (86.3%) zeros	Zeros

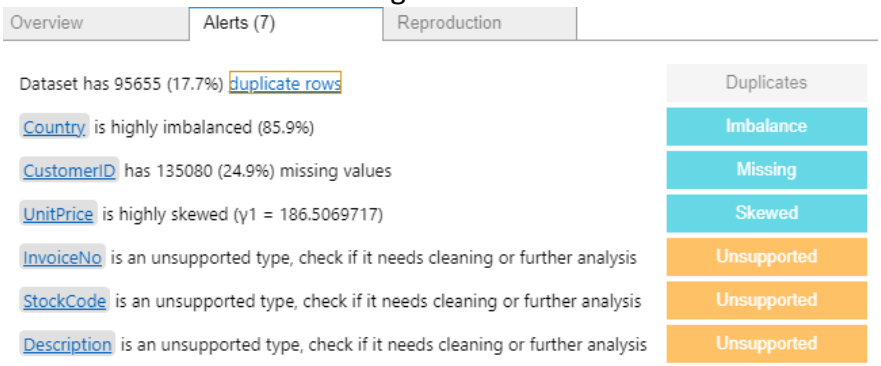
YData did not find anything that we care

3- 352 - Online Retail - Business

Dataset analysed: Online Retail.xlsx from https://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx (dataset_file_url column from Fortydatasets.xlsx) has_header

```
  InvoiceNo StockCode                          Description  Quantity  I am running a few minutes late; my previous meeting is running
over.
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

          InvoiceDate  UnitPrice  CustomerID         Country
0 2010-12-01 08:26:00       2.55     17850.0  United Kingdom
1 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
2 2010-12-01 08:26:00       2.75     17850.0  United Kingdom
3 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
4 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
```

Alerts in Ydata-Profiling



It could not analyse the three last columns above, which are analysed below, with the Data Quality Issues found:

```
InvoiceNo (number):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 9291 Non-numeric value(s) at index(es): [(141, 'C536379'), (154, 'C536383'), (235, 'C536391'), (236, 'C536391'), (237, 'C536391'),
(238, 'C536391'), (239, 'C536391'), (240, 'C536391'), (241, 'C536391'), (939, 'C536506'), ('...', '...'), (540141, 'C581468'), (540142,
'C581468'), (540176, 'C581470'), (540422, 'C581484'), (540448, 'C581490'), (540449, 'C581490'), (541541, 'C581499'), (541715,
'C581568'), (541716, 'C581569'), (541717, 'C581569')] (displaying only the first and last 10 items)

Range of values: (536365.0:581587.0).

StockCode (code):
  Error(s) found:
DQI #10 (Structural Conflicts - Consistency, Uniqueness):
 Data seems not categorical or has too many categories (> 100). Sample values: ['85123A', 71053, '84406B', '84029G', '84029E', 22752,
21730, 22633, 22632, 84879]
Categorical format with 4070 unique values:
Category Frequency
   85123A      2313
    22423      2203
   85099B      2159
    47566      1727
    20725      1639
    84879      1502
    22720      1477
    22197      1476
    21212      1385
    20727      1350
      ...       ...
 DCGS0055         1
 DCGS0057         1
DCGS0066P         1
 DCGS0067         1
 DCGS0068         1
 DCGS0071         1
 DCGS0072         1
 DCGS0073         1
 DCGS0074         1
        m         1

Description:
 String format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1454 Blank/Empty/Null/NaN value(s) at index(es): [(622, ''), (1970, ''), (1971, ''), (1972, ''), (1987, ''), (1988, ''), (2024, ''),
(2025, ''), (2026, ''), (2406, ''), ('...', '...'), (524473, ''), (524475, ''), (529667, ''), (533711, ''), (533712, ''), (535322, ''),
(535326, ''), (535332, ''), (536981, ''), (538554, '')] (displaying only the first and last 10 items)

DQI #17 (Non-String Data Type - Consistency):
 1 Non-string value(s) at index(es): [(420391, 20713)]
```

```
String range (lexicographical): ( 4 PURPLE FLOCK DINNER CANDLES : wrongly sold sets)


Quantity:
 Numerical format: All 541909 values are numerical in the range (-80995:80995).


InvoiceDate (date and time):
  Datetime format: All 541909 datetime values are valid in the YYYYMMDD format in the range 2010/12/01 08:26 to 2011/12/09 12:50.

UnitPrice (price):
  Numerical format: All 541909 values are numerical in the range (-11062.06:38970.0).

CustomerID (id):
 ID column format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 135080 Blank/Empty/Null/NaN value(s) at index(es): [(622, ''), (1443, ''), (1444, ''), (1445, ''), (1446, ''), (1447, ''), (1448, ''),
(1449, ''), (1450, ''), (1451, ''), ('...', '...'), (541531, ''), (541532, ''), (541533, ''), (541534, ''), (541535, ''), (541536, ''),
(541537, ''), (541538, ''), (541539, ''), (541540, '')] (displaying only the first and last 10 items)

DQI #9 (Duplicates - Uniqueness):
 541830 Duplicate value(s) at index(es): [[0, 17850], [1, 17850], [2, 17850], [3, 17850], [4, 17850], [5, 17850], [6, 17850], [7,
17850], [8, 17850], [9, 13047], ('...', '...'), [541899, 12680], [541900, 12680], [541901, 12680], [541902, 12680], [541903, 12680],
[541904, 12680], [541905, 12680], [541906, 12680], [541907, 12680], [541908, 12680]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):
 537536 Uniqueness violation(s) at index(es): [(1, 17850), (2, 17850), (3, 17850), (4, 17850), (5, 17850), (6, 17850), (7, 17850), (8,
17850), (10, 13047), (11, 13047), ('...', '...'), (541899, 12680), (541900, 12680), (541901, 12680), (541902, 12680), (541903, 12680),
(541904, 12680), (541905, 12680), (541906, 12680), (541907, 12680), (541908, 12680)] (displaying only the first and last 10 items)

Alphanumeric range of values: (12346.0 : 18287.0)
```

CustomerID has 135080 (24.9%) missing values

Missing

This code also managed to find the 135080 Missing values, but it also found Duplicates and Uniqueness Violation errors in almost all values of this variable, while Ydata-Profiling found Duplicate rows in only 17,7% of the data:

Dataset has 95655 (17.7%) duplicate rows

Duplicates

```
Country:
  Country format: All 541909 country values are valid.

Frequency Distribution  (showing top and bottom 10 of 38 categories):
              Country Frequency
       United Kingdom    495478
              Germany      9495
               France      8557
                 EIRE      8196
                Spain      2533
          Netherlands      2371
              Belgium      2069
          Switzerland      2002
             Portugal      1519
            Australia      1259
                  ...       ...
               Malta       127
United Arab Emirates        68
  European Community        61
                  RSA        58
              Lebanon        45
            Lithuania        35
               Brazil        32
       Czech Republic        30
              Bahrain        19
         Saudi Arabia        10
```

4- 602 - Dry Bean Dataset - Computer

Dataset analysed: DryBeanDataset/Dry_Bean_Dataset.xlsx from https://archive.ics.uci.edu/ml/machine-learning-databases/00602/DryBeanDataset.zip (dataset_file_url column from Fortydatasets.xlsx) has_header

```
Area:
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (20420:254616).

Perimeter:
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (524.736:1985.37).

Major axis length (length):
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (183.6011650038393:738.8601534818813).

Minor axis length (length):
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (122.51265345074418:460.1984968278401).

Aspect ratio (ratio):
  Numerical format: All 13611 values are numerical in the range (1.0248675960667681:2.430306446836626).
```

```
Eccentricity:
  Numerical format: All 13611 values are numerical in the range (0.21895126335356507:0.9114229684680053).

Convex area (area):
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (20684:263261).

Equivalent diameter (diameter):
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (161.24376423134018:569.3743583287609).

Extent (area):
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (0.55531471681117:0.8661946405648266).

Solidity (ratio):
  Numerical format: All 13611 values are numerical in the range (0.9192461570857022:0.9946774999456888).

Roundness (calculated):
  Numerical format: All 13611 values are numerical in the range (0.4896182562412148:0.9906853996160323).

Compactness:
  Numerical >=0 format: All 13611 values are numerical and greater or equal to 0 in the range (0.6405767589768725:0.9873029693778109).

ShapeFactor1 (factor):
  Numerical format: All 13611 values are numerical in the range (0.002778012668385549 4:0.010451169324378654).

ShapeFactor2 (factor):
  Numerical format: All 13611 values are numerical in the range (0.0005641690180332927:0.0036649719644516834).

ShapeFactor3 (factor):
  Numerical format: All 13611 values are numerical in the range (0.41033858414131424:0.9747671533422431).

ShapeFactor4 (factor):
  Numerical format: All 13611 values are numerical in the range (0.9476874027098624:0.9997325300471389).

Class:
  All 13611 values are correctly categorical.

Categorical format with 7 unique values:
Category  Frequency
DERMASON       3546
    SIRA       2636
   SEKER       2027
   HOROZ       1928
    CALI       1630
BARBUNYA       1322
  BOMBAY        522
```

Overview | Alerts 1 | Reproduction

Alerts

Dataset has 68 (0.5%) duplicate rows — Duplicates

YData did not find anything that we care.

5- 545 - Rice (Cammeo and Osmancik) - Computer

dataset_file_url was empty at Fortydatasets.xlsx

Dataset analysed: @DATA part extracted from Rice_Cammeo_Osmancik.arff that was obtained at rice+cammeo+and+osmancik.zip from https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik in the new data folders from UCI

```
Area:
  Numerical >=0 format: All 3810 values are correct (numeric and >=0).

Perimeter:
  Numerical >=0 format: All 3810 values are correct (numeric and >=0).

Major Axis Length (length):
  Numerical >=0 format: All 3810 values are correct (numeric and >=0).

Minor Axis Length (length):
  Numerical >=0 format: All 3810 values are correct (numeric and >=0).

Eccentricity:
  Numerical format: All 3810 values are numerical.

Convex Area (area):
  Numerical >=0 format: All 3810 values are correct (numeric and >=0).

Extent (area):
  Numerical format: All 3810 values are numerical.

Class:
  All 3810 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
Osmancik       2180
  Cammeo       1630
```

There were no Alerts provided from YDATA

6- 360 - Air quality - Computer

Dataset analysed: AirQualityUCI.xlsx from https://archive.ics.uci.edu/ml/machine-learning-databases/00360/AirQualityUCI.zip (dataset_file_url column from Fortydatasets.xlsx) has_header

```
Date:
  Date format: All 9357 date values are valid in the YYYYMMDD format in the range 2004/03/10 to 2005/04/04.

Time:
  Time format: All 9357 time values are valid in the range 00:00:00 to 23:00:00
```

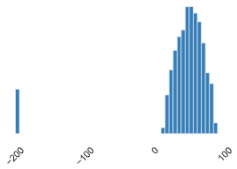
Time is an unsupported type, check if it needs cleaning or further analysis — Unsupported

```
True hourly averaged concentration CO in mg/m^3 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:11.9)

PT08.S1 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:2039.75)

True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200:1189)

True hourly averaged Benzene concentration in microg/m^3 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:63.74147644829163)

PT08.S2 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:2214.0)

True hourly averaged NOx concentration in ppb (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:1479.0)

PT08.S3 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:2682.75)

True hourly averaged NO2 concentration in microg/m^3 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:339.7)

PT08.S4 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:2775.0)

PT08.S5 (averaged):
  Numerical format: All 9357 values are numerical in the range (-200.0:2522.75)

Temperature in °C (temperature):
  Numerical format: All 9357 values are numerical in the range (-200.0:44.60000038147)

Relative Humidity (humidity):
  Numerical >=0 format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 366 Negative value(s) at index(es): [(524, -200.0), (525, -200.0), (526, -200.0), (701, -200.0), (702, -200.0), (703, -200.0), (704, -200.0), (705, -200.0), (706, -200.0), (707, -200.0), ('...', '...'), (8106, -200.0), (8107, -200.0), (8108, -200.0), (8109, -200.0), (8110, -200.0), (8111, -200.0), (8112, -200.0), (8113, -200.0), (8114, -200.0), (8777, -200.0)] (displaying only the first and last 10 items)

Range of values: (-200.0:88.72500038147)
```

▾ RH

RH
Real number (ℝ)

Distinct	4903	Minimum	-200
Distinct (%)	52.4%	Maximum	88.725
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	366
Infinite (%)	0.0%	Negative (%)	3.9%
Mean	39.483611	Memory size	73.2 KiB




```
AH Absolute Humidity (humidity):
  Numerical >=0 format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 366 Negative value(s) at index(es): [(524, -200.0), (525, -200.0), (526, -200.0), (701, -200.0), (702, -200.0), (703, -200.0), (704, -200.0), (705, -200.0), (706, -200.0), (707, -200.0), ('...', '...'), (8106, -200.0), (8107, -200.0), (8108, -200.0), (8109, -200.0), (8110, -200.0), (8111, -200.0), (8112, -200.0), (8113, -200.0), (8114, -200.0), (8777, -200.0)] (displaying only the first and last 10 items)

Range of values: (-200.0:2.2310357155831864)
```

▾ AH

AH
Real number (ℝ)
HIGH CORRELATION

Distinct	8988	Minimum	-200
Distinct (%)	96.1%	Maximum	2.2310357
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	366
Infinite (%)	0.0%	Negative (%)	3.9%
Mean	-6.8376037	Memory size	73.2 KiB

```
Last run on: 2024-01-23 19:24:17
```

Ydata Profiling did find negative values in the two last columns, but it did not give any emphasis on it.

Below are lines 524 to 526. Observe that many other values have the content -200, besides the two last Humidity values:

Date	Time	CO (GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
1/04/2004	14:00:00	1.7	-200	222	-200.0	-200	99	-200	72	-200	-200	-200	-200	-200
1/04/2004	15:00:00	1.9	-200	197	-200.0	-200	108	-200	81	-200	-200	-200	-200	-200
1/04/2004	16:00:00	2.3	-200	319	-200.0	-200	131	-200	93	-200	-200	-200	-200	-200

7- 242 - Energy efficiency - Computer

Dataset analysed: ENB2012_data.xlsx from https://archive.ics.uci.edu/ml/machine-learning-databases/00242/ENB2012_data.xlsx (dataset_file_url column from Fortydatasets.xlsx) has_header

```
X1 (compactness):
  Numerical >=0 format: All 768 values are numerical and greater or equal to 0 in the range (0.62:0.98)

X2 (area):
  Numerical >=0 format: All 768 values are numerical and greater or equal to 0 in the range (514.5:808.5)

X3 (area):
  Numerical >=0 format: All 768 values are numerical and greater or equal to 0 in the range (245.0:416.5)

X4 (area):
  Numerical >=0 format: All 768 values are numerical and greater or equal to 0 in the range (110.25:220.5)

X5 (height):
  Numerical >=0 format: All 768 values are numerical and greater or equal to 0 in the range (3.5:7.0)

X6 (orientation):
  Numerical format: All 768 values are numerical in the range (2:5)

X7 (area):
  Numerical >=0 format: All 768 values are numerical and greater or equal to 0 in the range (0.0:0.4)

X8 (area):
  Numerical >=0 format: All 768 values are numerical and greater or equal to 0 in the range (0:5)

y1 (load):
  Numerical format: All 768 values are numerical in the range (6.01:43.1)

y2 (load):
  Numerical format: All 768 values are numerical in the range (10.9:48.03)

Last run on: 2024-01-23 19:19:31
```

Overview | Alerts 10 | Reproduction

Alerts

X1 is highly overall correlated with X2 and 4 other fields	High correlation
X2 is highly overall correlated with X1 and 4 other fields	High correlation
X3 is highly overall correlated with X4 and 1 other fields	High correlation
X4 is highly overall correlated with X1 and 5 other fields	High correlation
X5 is highly overall correlated with X1 and 5 other fields	High correlation
y1 is highly overall correlated with X1 and 4 other fields	High correlation
y2 is highly overall correlated with X1 and 4 other fields	High correlation
X5 is uniformly distributed	Uniform
X6 is uniformly distributed	Uniform
X8 has 48 (6.2%) zeros	Zeros

YData did not find anything that we care.

8- 267 - banknote authentication – Computer

Dataset analysed: data_banknote_authentication.txt from https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt (dataset_file_url column from Fortydatasets.xlsx)

```
variance of Wavelet Transformed image (variance):
  Numerical format: All 1372 values are numerical in the range (-7.0421:6.8248).

skewness of Wavelet Transformed image (skewness):
```

```
  Numerical format: All 1372 values are numerical in the range (-13.7731:12.9516).

curtosis of Wavelet Transformed image (curtosis):
  Numerical format: All 1372 values are numerical in the range (-5.2861:17.9274).

entropy of image (entropy):
  Numerical format: All 1372 values are numerical in the range (-8.5482:2.4495).

class (integer):
  Numerical format: All 1372 values are numerical in the range (0:1).

Last run on: 2024-01-28 15:37:37
```

Overview | Alerts 6 | Reproduction

Alerts

Dataset has 11 (0.8%) duplicate rows	Duplicates
`class` is highly overall correlated with `variance of Wavelet Transformed image`	High correlation
`curtosis of Wavelet Transformed image` is highly overall correlated with `skewness of Wavelet Transformed image`	High correlation
`entropy of image` is highly overall correlated with `skewness of Wavelet Transformed image`	High correlation
`skewness of Wavelet Transformed image` is highly overall correlated with `curtosis of Wavelet Transformed image` and 1 other fields	High correlation
`variance of Wavelet Transformed image` is highly overall correlated with `class`	High correlation

YData did not find anything that we care.

9- 29 - Computer Hardware – Computer

Dataset analysed: machine.data from https://archive.ics.uci.edu/ml/machine-learning-databases/cpu-performance/machine.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header

```
vendor name (name):
  Name format: All 209 vendor name values are valid.

Frequency Distribution:
 vendor name Frequency
         ibm        32
         nas        19
   honeywell        13
         nor        13
      sperry        13
     siemens        12
      amdahl         9
         cdc         9
   burroughs         8
          dg         7
         ...       ...
     nixdorf         3
perkin-elmer         3
      apollo         2
        basf         2
         bti         2
        wang         2
     adviser         1
  four-phase         1
   microdata         1
      sratus         1

Range of Values: (adviser to wang)

Model Name:
  Model Name format: All 209 Model Name values are valid.

Frequency Distribution:
Model Name Frequency
       100         1
1100/61-h1         1
   1100/81         1
   1100/82         1
   1100/83         1
   1100/84         1
   1100/93         1
   1100/94         1
```

```
    1636-1         1
   1636-10         1
       ...       ...
     v8635         1
     v8650         1
     v8655         1
     v8665         1
     v8670         1
vax:11/730         1
vax:11/750         1
vax:11/780         1
    vs-100         1
     vs-90         1

Range of Values: (100 to vs-90)

MYCT (integer):
  Numerical format: All 209 values are numerical in the range (17:1500).

MMIN (minimum):
  Numerical format: All 209 values are numerical in the range (64:32000).

MMAX (maximum):
  Numerical format: All 209 values are numerical in the range (64:64000).

CACH (integer):
  Numerical format: All 209 values are numerical in the range (0:256).

CHMIN (minimum):
  Numerical format: All 209 values are numerical in the range (0:52).

CHMAX (maximum):
  Numerical format: All 209 values are numerical in the range (0:176).

PRP (integer):
  Numerical format: All 209 values are numerical in the range (6:1150).

ERP (integer):
  Numerical format: All 209 values are numerical in the range (15:1238).

Last run on: 2024-01-28 19:22:45
```

Overview | Alerts 13 | Reproduction

Alerts

CACH is highly overall correlated with CHMAX and 5 other fields	High correlation
CHMAX is highly overall correlated with CACH and 6 other fields	High correlation
CHMIN is highly overall correlated with CHMAX and 4 other fields	High correlation
ERP is highly overall correlated with CACH and 6 other fields	High correlation
MMAX is highly overall correlated with CACH and 5 other fields	High correlation
MMIN is highly overall correlated with CACH and 5 other fields	High correlation
MYCT is highly overall correlated with CACH and 6 other fields	High correlation
PRP is highly overall correlated with CACH and 6 other fields	High correlation
vendor name is highly overall correlated with CHMAX	High correlation
Model Name has unique values	Unique
CACH has 69 (33.0%) zeros	Zeros
CHMIN has 5 (2.4%) zeros	Zeros
CHMAX has 5 (2.4%) zeros	Zeros

YData found that Model Name has unique values. We have not checked that.

10- 294 - Combined Cycle Power Plant - Computer

Dataset analysed: CCPP/Folds5x2_pp.xlsx from https://archive.ics.uci.edu/ml/machine-learning-databases/00294/CCPP.zip (dataset_file_url column from Fortydatasets.xlsx)

This was the original description of the Attributes Information:
"Features consist of hourly average ambient variables

1. Temperature (T) in the range 1.81°C and 37.11°C,
2. Ambient Pressure (AP) in the range 992.89-1033.30 milibar,
3. Relative Humidity (RH) in the range 25.56% to 100.16%
4. Exhaust Vacuum (V) in teh range 25.36-81.56 cm Hg
5. Net hourly electrical energy output (EP) 420.26-495.76 MW
The averages are taken from various sensors located around the plant that record the ambient variables every second. The variables are given without normalization."

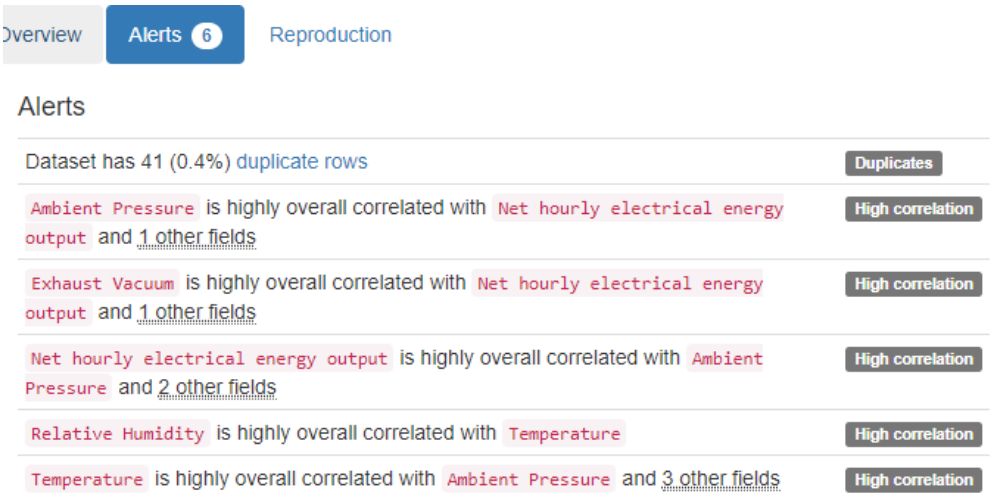
And these were the first 5 items:

	AT	V	AP	RH	PE
0	14.96	41.76	1024.07	73.17	463.26
1	25.18	62.96	1020.04	59.08	444.37
2	5.11	39.40	1012.16	92.14	488.56
3	20.86	57.32	1010.24	76.64	446.48
4	10.82	37.50	1009.23	96.62	473.90

The order of the dataset are not the ones in the description. Only the first and last are correct. The others are changed. The second column in the data is V, which is the fourth column in the description. The third column in the data is AP, which is the second column in the description, and the fourth column in the dataset is RH, which is the third in the description.

After correcting the order of the columns, the result was:

```
Temperature:
  Numerical format: All 9568 values are numerical in the range (1.81:37.11).

Exhaust Vacuum (vacuum):
  Numerical format: All 9568 values are numerical in the range (25.36:81.56).

Ambient Pressure (pressure):
  Numerical format: All 9568 values are numerical in the range (992.89:1033.3).

Relative Humidity (humidity):
  Numerical >=0 format: All 9568 values are numerical and greater or equal to 0 in the range (25.56:100.16).

Net hourly electrical energy output (energy):
  Numerical >=0 format: All 9568 values are numerical and greater or equal to 0 in the range (420.26:495.76).

Last run on: 2024-01-29 12:23:35
```



YData did not find anything that we care.

11- 229 - Skin Segmentation – Computer

Dataset analysed: Skin_NonSkin.txt from https://archive.ics.uci.edu/ml/machine-learning-databases/00229/Skin_NonSkin.txt (dataset_file_url column from Fortydatasets.xlsx)

```
x1 (b):
  Numerical format: All 245057 values are numerical in the range (0:255).

x2 (g):
  Numerical format: All 245057 values are numerical in the range (0:255).

x3 (r):
  Numerical format: All 245057 values are numerical in the range (0:255).

y (class):
  All 245057 values are correctly categorical.
```

```
Categorical format with 2 unique values:
 Category  Frequency
        2     194198
        1      50859

Last run on: 2024-01-29 13:50:39
```

Overview | Alerts 8 | Reproduction

Alerts

Dataset has 20347 (8.3%) duplicate rows	Duplicates
x1 is highly overall correlated with x2	High correlation
x2 is highly overall correlated with x1 and 1 other fields	High correlation
x3 is highly overall correlated with x2 and 1 other fields	High correlation
y is highly overall correlated with x3	High correlation
x1 has 3237 (1.3%) zeros	Zeros
x2 has 2968 (1.2%) zeros	Zeros
x3 has 6539 (2.7%) zeros	Zeros

YData did not find anything that we care.

12- 246 - 3D Road Network (North Jutland, Denmark) – Computer

Dataset analysed: 3D_spatial_network.txt from https://archive.ics.uci.edu/ml/machine-learning-databases/00246/3D_spatial_network.txt (dataset_file_url column from Fortydatasets.xlsx)

```
OSM_ID (id):
  ID column format: Error(s) found:
DQI #9 (Duplicates - Uniqueness):
 421305 Duplicate value(s) at index(es): [[0, 144552912], [1, 144552912], [2, 144552912], [3, 144552912], [4, 144552912], [5,
144552912], [6, 144552912], [7, 144552912], [8, 144552912], [9, 144552912], ('...', '...'), [434864, 93323205], [434865, 93323205],
[434866, 93323205], [434867, 93323205], [434868, 93323205], [434869, 93323205], [434870, 93323205], [434871, 93323205], [434872,
93323209], [434873, 93323209]] (displaying only the first and last 10 items)

DQI #17 (Wrong Data Type - Consistency):
 204080 Inconsistent length in alphanumeric value(s) at index(es): [(19, 42991631), (20, 42991631), (21, 42991631), (22, 42991631),
(23, 42991631), (24, 42991631), (25, 42991631), (26, 42991631), (27, 42991632), (28, 42991632), ('...', '...'), (434864, 93323205),
(434865, 93323205), (434866, 93323205), (434867, 93323205), (434868, 93323205), (434869, 93323205), (434870, 93323205), (434871,
93323205), (434872, 93323209), (434873, 93323209)] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):
 377545 Uniqueness violation(s) at index(es): [(1, 144552912), (2, 144552912), (3, 144552912), (4, 144552912), (5, 144552912), (6,
144552912), (7, 144552912), (8, 144552912), (9, 144552912), (10, 144552912), ('...', '...'), (434863, 93323205), (434864, 93323205),
(434865, 93323205), (434866, 93323205), (434867, 93323205), (434868, 93323205), (434869, 93323205), (434870, 93323205), (434871,
93323205), (434873, 93323209)] (displaying only the first and last 10 items)

Alphanumeric range of values: (100009227 : 99935416)
```

YDATA found nothing!

```
LONGITUDE:
  Longitude format: All 434874 values are numerical and valid in the range [-180, 180].
Actual range of values: (8.1461259 : 11.1993265)

LATITUDE:
  Latitude format: All 434874 values are numerical and valid in the range [-90, 90].
Actual range of values: (56.5824856 : 57.750511)

ALTITUDE:
  Numerical format: All 434874 values are numerical in the range (-8.60818370517905:134.441946906076).

Last run on: 2024-01-29 14:03:03
```

13- 374 - Appliances energy prediction – Computer

Dataset analysed: energydata_complete.csv from https://archive.ics.uci.edu/ml/machine-learning-databases/00374/energydata_complete.csv (dataset_file_url column from Fortydatasets.xlsx) has header

```
date time year-month-day hour (date time):
  Datetime format: All 19735 datetime values are valid in the YYYYMMDD format in the range 2016/01/11 17:00 to 2016/05/27 18:00.

Appliances, energy use in Wh (energy):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (10:1080).

lights, energy use of light fixtures in the house in Wh (energy):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (0:70).

T1, Temperature in kitchen area, in Celsius (temperature):
```

```
  Numerical format: All 19735 values are numerical in the range (16.79:26.26).

RH_1, Humidity in kitchen area, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (27.0233333333333:63.36).

T2, Temperature in living room area, in Celsius (temperature):
  Numerical format: All 19735 values are numerical in the range (16.1:29.8566666666667).

RH_2, Humidity in living room area, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (20.463333333333296:56.0266666666667).

T3, Temperature in laundry room area (temperature):
  Numerical format: All 19735 values are numerical in the range (17.2:29.236).

RH_3, Humidity in laundry room area, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (28.7666666666667:50.1633333333333).

T4, Temperature in office room, in Celsius (temperature):
  Numerical format: All 19735 values are numerical in the range (15.1:26.2).

RH_4, Humidity in office room, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (27.66:51.09).

T5, Temperature in bathroom, in Celsius (temperature):
  Numerical format: All 19735 values are numerical in the range (15.33:25.795).

RH_5, Humidity in bathroom, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (29.815:96.3216666666667).

T6, Temperature outside the building (temperature):
  Numerical format: All 19735 values are numerical in the range (-6.065:28.29).

RH_6, Humidity outside the building (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (1.0:99.9).

T7, Temperature in ironing room , in Celsius (temperature):
  Numerical format: All 19735 values are numerical in the range (15.39:26.0).

RH_7, Humidity in ironing room, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (23.2:51.4).

T8, Temperature in teenager room 2, in Celsius (temperature):
  Numerical format: All 19735 values are numerical in the range (16.3066666666667:27.23).

RH_8, Humidity in teenager room 2, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (29.6:58.78).

T9, Temperature in parents room, in Celsius (temperature):
  Numerical format: All 19735 values are numerical in the range (14.89:24.5).

RH_9, Humidity in parents room, in % (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (29.1666666666667:53.3266666666667).

To, Temperature outside (temperature):
  Numerical format: All 19735 values are numerical in the range (-5.0:26.1).

Pressure:
  Numerical format: All 19735 values are numerical in the range (729.2999999999998:772.2999999999998).

RH_out, Humidity outside (humidity):
  Numerical >=0 format: All 19735 values are numerical and greater or equal to 0 in the range (24.0:100.0).

Wind speed (speed):
  Numerical format: All 19735 values are numerical in the range (0.0:14.0).

Visibility:
  Numerical format: All 19735 values are numerical in the range (1.0:66.0).

Tdewpoint (dewpoint):
  Numerical format: All 19735 values are numerical in the range (-6.6:15.5).

rv1, Random variable 1, nondimensional (random):
  Numerical format: All 19735 values are numerical in the range (0.0053216819651424:49.9965296825394).

rv2, Random variable 2, nondimensional (random):
  Numerical format: All 19735 values are numerical in the range (0.0053216819651424:49.9965296825394).

Last run on: 2024-01-29 14:41:33
```

Overview | Alerts 27 | Reproduction

Alerts

`RH_1, Humidity in kitchen area, in %` is highly overall correlated with `RH_2, Humidity in living room area, in %` and 6 other fields — High correlation

`RH_2, Humidity in living room area, in %` is highly overall correlated with `RH_1, Humidity in kitchen area, in %` and 6 other fields — High correlation

…

`rv2, Random variable 2, nondimensional` is highly overall correlated with `rv1, Random variable 1, nondimensional`	High correlation
`date time year-month-day hour` has unique values	Unique
`rv1, Random variable 1, nondimensional` has unique values	Unique
`rv2, Random variable 2, nondimensional` has unique values	Unique
`lights, energy use of light fixtures in the house in Wh` has 15252 (77.3%) zeros	Zeros

YData found 3 columns that are Unique. We have not checked that.

14- 248 - Buzz in social media - Computer

Dataset analysed: ./regression/TomsHardware/TomsHardware.data from https://archive.ics.uci.edu/ml/machine-learning-databases/00248/regression.tar.gz (dataset_file_url column from Fortydatasets.xlsx) does not have header

```
NCD_0 (number):
  Numerical format: All 28179 values are numerical in the range (0:182).

NCD_1 (number):
  Numerical format: All 28179 values are numerical in the range (0:118).

NCD_2 (number):
  Numerical format: All 28179 values are numerical in the range (0:118).

NCD_3 (number):
  Numerical format: All 28179 values are numerical in the range (0:118).

NCD_4 (number):
  Numerical format: All 28179 values are numerical in the range (0:118).

NCD_5 (number):
  Numerical format: All 28179 values are numerical in the range (0:118).

NCD_6 (number):
  Numerical format: All 28179 values are numerical in the range (0:154).

NCD_7 (number):
  Numerical format: All 28179 values are numerical in the range (0:88).

BL_0 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

BL_1 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

BL_2 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

BL_3 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

BL_4 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

BL_5 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

BL_6 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

BL_7 (burstiness):
  Numerical format: All 28179 values are numerical in the range (0.0:1.0).

NAD_0 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:217).

NAD_1 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:210).

NAD_2 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:210).

NAD_3 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:210).

NAD_4 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:210).

NAD_5 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:194).

NAD_6 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:170).

NAD_7 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0:185).

AI_0 (increase):
```

  Numerical format: All 28179 values are numerical in the range (0:158).

AI_1 (increase):
  Numerical format: All 28179 values are numerical in the range (0:146).

AI_2 (increase):
  Numerical format: All 28179 values are numerical in the range (0:149).

AI_3 (increase):
  Numerical format: All 28179 values are numerical in the range (0:161).

AI_4 (increase):
  Numerical format: All 28179 values are numerical in the range (0:169).

AI_5 (increase):
  Numerical format: All 28179 values are numerical in the range (0:149).

AI_6 (increase):
  Numerical format: All 28179 values are numerical in the range (0:156).

AI_7 (increase):
  Numerical format: All 28179 values are numerical in the range (0:160).

NAC_0 (number):
  Numerical format: All 28179 values are numerical in the range (0:1734).

NAC_1 (number):
  Numerical format: All 28179 values are numerical in the range (0:1966).

NAC_2 (number):
  Numerical format: All 28179 values are numerical in the range (0:1734).

NAC_3 (number):
  Numerical format: All 28179 values are numerical in the range (0:1734).

NAC_4 (number):
  Numerical format: All 28179 values are numerical in the range (0:1734).

NAC_5 (number):
  Numerical format: All 28179 values are numerical in the range (0:1657).

NAC_6 (number):
  Numerical format: All 28179 values are numerical in the range (0:1403).

NAC_7 (number):
  Numerical format: All 28179 values are numerical in the range (0:1707).

ND_0 (number):
  Numerical format: All 28179 values are numerical in the range (0:235271).

ND_1 (number):
  Numerical format: All 28179 values are numerical in the range (0:197284).

ND_2 (number):
  Numerical format: All 28179 values are numerical in the range (0:225099).

ND_3 (number):
  Numerical format: All 28179 values are numerical in the range (0:207633).

ND_4 (number):
  Numerical format: All 28179 values are numerical in the range (0:272256).

ND_5 (number):
  Numerical format: All 28179 values are numerical in the range (0:272256).

ND_6 (number):
  Numerical format: All 28179 values are numerical in the range (0:330561).

ND_7 (number):
  Numerical format: All 28179 values are numerical in the range (0:272256).

CS_0 (contribution):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

CS_1 (contribution):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

CS_2 (contribution):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

CS_3 (contribution):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

CS_4 (contribution):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

CS_5 (contribution):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

CS_6 (contribution):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

CS_7 (contribution):

```
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:1.0).

AT_0 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:105.0).

AT_1 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:98.0).

AT_2 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:104.0).

AT_3 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:106.0).

AT_4 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:107.0).

AT_5 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:107.0).

AT_6 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:104.0).

AT_7 (interaction):
  Numerical format: All 28179 values are numerical in the range (0.0:106.0).

NA_0 (number):
  Numerical format: All 28179 values are numerical in the range (0:313).

NA_1 (number):
  Numerical format: All 28179 values are numerical in the range (0:313).

NA_2 (number):
  Numerical format: All 28179 values are numerical in the range (0:313).

NA_3 (number):
  Numerical format: All 28179 values are numerical in the range (0:313).

NA_4 (number):
  Numerical format: All 28179 values are numerical in the range (0:313).

NA_5 (number):
  Numerical format: All 28179 values are numerical in the range (0:322).

NA_6 (number):
  Numerical format: All 28179 values are numerical in the range (0:264).

NA_7 (number):
  Numerical format: All 28179 values are numerical in the range (0:309).

ADL_0 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

ADL_1 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

ADL_2 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

ADL_3 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

ADL_4 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

ADL_5 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

ADL_6 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

ADL_7 (length):
  Numerical >=0 format: All 28179 values are numerical and greater or equal to 0 in the range (0.0:150.0).

-- Attention Level (measured with number of authors) (AS(NA))
       (columns [80,87])

          +----------+-----+-------+-------+-------+
          | feature  | min | max   | mean  | std   |
          +----------+-----+-------+-------+-------+
          | AS(NA)_0 | 0   | 0.120 | 0.002 | 0.008 |
          +----------+-----+-------+-------+-------+
          | AS(NA)_1 | 0   | 0.129 | 0.002 | 0.008 |
          +----------+-----+-------+-------+-------+
          | AS(NA)_2 | 0   | 0.129 | 0.002 | 0.008 |
          +----------+-----+-------+-------+-------+
          | AS(NA)_3 | 0   | 0.159 | 0.002 | 0.008 |
          +----------+-----+-------+-------+-------+
          | AS(NA)_4 | 0   | 0.166 | 0.002 | 0.008 |
          +----------+-----+-------+-------+-------+
          | AS(NA)_5 | 0   | 0.166 | 0.002 | 0.008 |
          +----------+-----+-------+-------+-------+
          | AS(NA)_6 | 0   | 0.155 | 0.002 | 0.007 |
```

```
          +----------+-----+-------+-------+-------+
          | AS(NA)_7 | 0   | 0.147 | 0.002 | 0.007 |
          +----------+-----+-------+-------+-------+

AS_NA)_0 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.120455).

AS_NA)_1 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.12885).

AS_NA)_2 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.12885).

AS_NA)_3 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.159019).

AS_NA)_4 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.16569).

AS_NA)_5 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.16569).

AS_NA)_6 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.155071).

AS_NA)_7 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.147081).

-- Attention Level (measured with number of contributions) (AS(NAC))
       (columns [88,95])

          +-----------+-----+-------+-------+-------+
          | feature   | min | max   | mean  | std   |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_0 | 0   | 0.107 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_1 | 0   | 0.130 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_2 | 0   | 0.136 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_3 | 0   | 0.153 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_4 | 0   | 0.153 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_5 | 0   | 0.147 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_6 | 0   | 0.179 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+
          | AS(NAC)_7 | 0   | 0.188 | 0.002 | 0.006 |
          +-----------+-----+-------+-------+-------+

AS_NAC)_0 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.107367).

AS_NAC)_1 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.130337).

AS_NAC)_2 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.135633).

AS_NAC)_3 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.153209).

AS_NAC)_4 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.153209).

AS_NAC)_5 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.147003).

AS_NAC)_6 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.179334).

AS_NAC)_7 (number):
  Numerical format: All 28179 values are numerical in the range (0.0:0.187696).

Feature to predict (feature):
  Error(s) found:
DQI #10 (Structural Conflicts - Consistency, Uniqueness):
 Data seems not categorical or has too many categories (> 100). Sample values: [4.5, 3.5, 2.0, 1.5, 5.0, 2.5, 1.0, 0.5, 3.0, 6.0]
Categorical format with 8895 unique values:
 Category Frequency
      2.0       217
      1.5       182
      0.5       175
      2.5       169
      1.0       157
      4.0       143
      3.0       142
      3.5       134
      6.0       128
      4.5       126
      ...       ...
 206317.5         1
```

```
211342.0         1
211967.0         1
214715.0         1
215597.5         1
219497.5         1
223389.5         1
237281.0         1
242383.0         1
265916.5         1

Last run on: 2024-01-30 14:49:19
```

Feature to predict

Real number (ℝ)

HIGH CORRELATION

Distinct	8895	**Minimum**	0
Distinct (%)	31.6%	**Maximum**	265916.5
Missing	0	**Zeros**	125
Missing (%)	0.0%	**Zeros (%)**	0.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	3486.441	**Memory size**	220.3 KiB

YData confirmed in its analysis that the Feature to predict column had really 8895 distinct values, which seems to be a major problem in such a column. Our code found that to be a problem for a Categorical column should not have over 100 distinct values. Besides that YData found many High Correlation and Zeros.

15- 343 - Occupancy Detection – Computer

Dataset analysed: datatest.txt from https://archive.ics.uci.edu/ml/machine-learning-databases/00357/occupancy_data.zip (dataset_file_url column from Fortydatasets.xlsx)

```
date time year-month-day hour (date time):
  Datetime format: All 2665 datetime values are valid in the YYYYMMDD format in the range 2015/02/02 14:19 to 2015/02/04 10:43.

Temperature, in Celsius (temperature):
  Numerical format: All 2665 values are numerical in the range (20.2:24.4083333333333).

Relative Humidity, % (humidity):
  Numerical >=0 format: All 2665 values are numerical and greater or equal to 0 in the range (22.1:31.4725).

Light, in Lux (light):
  Numerical >=0 format: All 2665 values are numerical and greater or equal to 0 in the range (0.0:1697.25).

CO2, in ppm (co2):
  Numerical >=0 format: All 2665 values are numerical and greater or equal to 0 in the range (427.5:1402.25).

Humidity Ratio, Derived quantity from temperature and relative humidity, in kgwater-vapor/kg-air (temperature):
  Numerical format: All 2665 values are numerical in the range (0.0033033144722347:0.0053777588397133).

Occupancy, 0 or 1, 0 for not occupied, 1 for occupied status (status):
  All 2665 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0       1693
        1        972

Last run on: 2024-01-30 18:56:59
```

Overview | Alerts 8 | Reproduction

Alerts

`CO2, in ppm` is highly overall correlated with `Humidity Ratio, Derived quantity from temperature and relative humidity, in kgwater-vapor` and 4 other fields	High correlation
`Humidity Ratio, Derived quantity from temperature and relative humidity, in kgwater-vapor` is highly overall correlated with `CO2, in ppm` and 4 other fields	High correlation
`Light, in Lux` is highly overall correlated with `CO2, in ppm` and 4 other fields	High correlation
`Occupancy, 0 or 1, 0 for not occupied, 1 for occupied status` is highly overall correlated with `CO2, in ppm` and 4 other fields	High correlation
`Relative Humidity, %` is highly overall correlated with `CO2, in ppm` and 4 other fields	High correlation
`Temperature, in Celsius` is highly overall correlated with `CO2, in ppm` and 4 other fields	High correlation
`date time year-month-day hour` has unique values	Unique
`Light, in Lux` has 1615 (60.6%) zeros	Zeros

YData found that the first column is Unique. We have not checked that.

16- 128 - KDD Cup 1999 Data - Computer

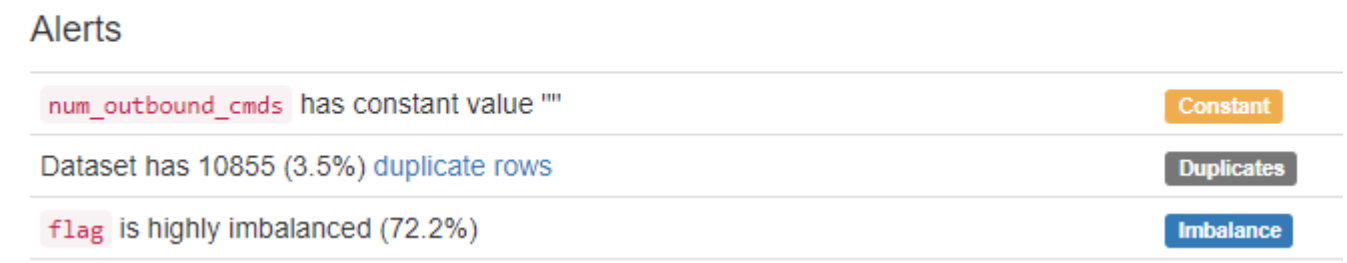
Dataset analysed: corrected (without file ending), renamed to corrected.txt, from http://kdd.ics.uci.edu/databases/kddcup99/corrected.gz obtained at https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, instead of the original dataset_file_url column from Fortydatasets.xlsx which was at https://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/corrected.gz but it is not valid anymore. It does not have header.

```
duration:
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:57715).

protocol_type (type):
  All 311029 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
    icmp     164969
     tcp     119357
     udp      26703

service:
  All 311029 values are correctly categorical.

Categorical format with 65 unique values:
 Category Frequency
    ecr_i    164352
  private     78510
     http     41237
     smtp      8268
    pop_3      3972
 domain_u      3160
 ftp_data      2223
    other      2185
   telnet      2077
      ftp       837
      ...       ...
 csnet_ns        50
   klogin        50
   supdup        50
hostnames        48
      IRC        24
  pm_dump        16
      X11        15
    tim_i         7
     icmp         2
   tftp_u         1

flag:
  All 311029 values are correctly categorical.

Categorical format with 11 unique values:
Category  Frequency
      SF     248379
     REJ      41945
      S0      18012
    RSTO       1393
    RSTR        872
      S3        289
      SH         84
      S1         27
      S2         22
     OTH          4
  RSTOS0          2

src_bytes (bytes):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:62825648).

dst_bytes (bytes):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:5203179).

land (symbolic):
  All 311029 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0     311020
        1          9

wrong_fragment (continuous):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:3).

urgent (continuous):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:3).

hot (continuous):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:101).

num_failed_logins (num):
  Numerical format: All 311029 values are numerical in the range (0:4).
```

```
logged_in (symbolic):
  All 311029 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0     257384
        1      53645

num_compromised (num):
  Numerical format: All 311029 values are numerical in the range (0:796).

root_shell (continuous):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:1).

su_attempted (attempted):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:2).

num_root (num):
  Numerical format: All 311029 values are numerical in the range (0:878).

num_file_creations (num):
  Numerical format: All 311029 values are numerical in the range (0:100).

num_shells (num):
  Numerical format: All 311029 values are numerical in the range (0:5).

num_access_files (num):
  Numerical format: All 311029 values are numerical in the range (0:4).

num_outbound_cmds (num):
  Numerical format: All 311029 values are numerical in the range (0:0).

is_host_login (symbolic):
  All 311029 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0     311017
        1         12

is_guest_login (symbolic):
  All 311029 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0     310275
        1        754

count:
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:511).

srv_count (count):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:511).

serror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

srv_serror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

rerror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

srv_rerror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

same_srv_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

diff_srv_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

srv_diff_host_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

dst_host_count (count):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:255).

dst_host_srv_count (count):
  Numerical >=0 format: All 311029 values are numerical and greater or equal to 0 in the range (0:255).

dst_host_same_srv_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

dst_host_diff_srv_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

dst_host_same_src_port_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

dst_host_srv_diff_host_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).
```

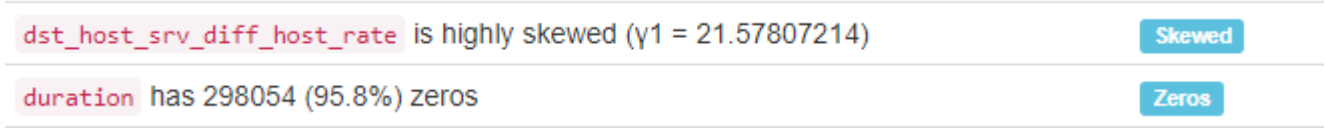
```
dst_host_serror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

dst_host_srv_serror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

dst_host_rerror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

dst_host_srv_rerror_rate (rate):
  Numerical format: All 311029 values are numerical in the range (0.0:1.0).

class:
  All 311029 values are correctly categorical.

Categorical format with 38 unique values:
      Category Frequency
         smurf.    164091
        normal.     60593
       neptune.     58001
snmpgetattack.      7741
     mailbomb.      5000
 guess_passwd.      4367
    snmpguess.      2406
        satan.      1633
  warezmaster.      1602
          back.     1098
           ...       ...
         xlock.         9
        xsnoop.         4
    ftp_write.         3
   loadmodule.         2
          perl.         2
           phf.         2
    sqlattack.         2
     udpstorm.         2
         worm.         2
         imap.         1

Last run on: 2024-02-21 21:03:45
```

Dataset



It shows column num_outbound_cmds with constant value "". We have not checked that. For us it is a numerical with values in the range (0:0).

…



17- 303 - Perfume Data - Computer

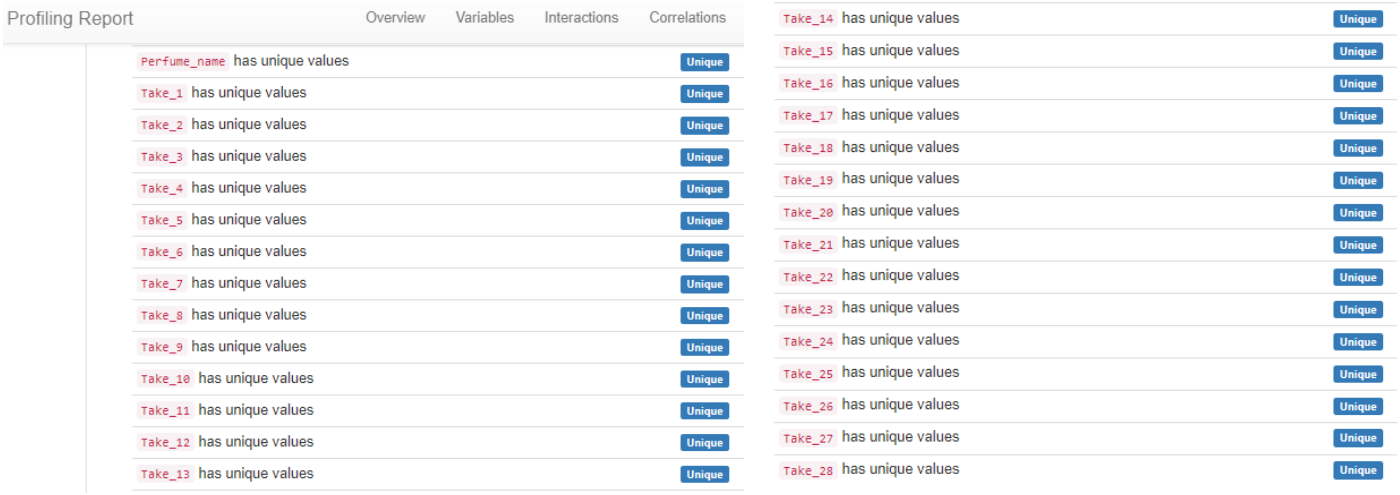
Dataset analysed: perfume_data.xlsx from https://archive.ics.uci.edu/static/public/303/perfume+data.zip instead of the original dataset_file_url column from Fortydatasets.xlsx which was at https://archive.ics.uci.edu/ml/machine-learning-databases/00303/perfume_data.xlsx but it is not valid anymore.

```
Perfume_name (name):
  Name format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 1 Extraneous data value(s) at index(es): [(16, 'constrected2')]

Frequency Distribution:
    Perfume_name  Frequency
        RoseMusk          1
      TeaTreeOil          1
          ajayeb          1
           ajmal          1
          amreaj          1
            aood          1
```

```
       asgar_ali          1
         bukhoor          1
        burberrry         1
carolina_herrera          1
     constrected          1
    constrected2          1
      dehenalaod          1
          junaid          1
          kausar          1
 oudh_ma'alattar          1
       raspberry          1
            rose          1
       solidmusk          1
      strawberry          1

Range of Values: (RoseMusk to strawberry)

Take_1 (take):
  Numerical format: All 20 values are numerical in the range (46014:85056).

Take_2 (take):
  Numerical format: All 20 values are numerical in the range (46014:85056).

Take_3 (take):
  Numerical format: All 20 values are numerical in the range (46014:85056).

Take_4 (take):
  Numerical format: All 20 values are numerical in the range (46014:85056).

Take_5 (take):
  Numerical format: All 20 values are numerical in the range (46014:85056).

Take_6 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_7 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_8 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_9 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_10 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_11 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_12 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_13 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_14 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_15 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_16 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_17 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_18 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_19 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_20 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_21 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_22 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_23 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_24 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_25 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_26 (take):
```

```
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_27 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Take_28 (take):
  Numerical format: All 20 values are numerical in the range (46015:85056).

Last run on: 2024-02-22 13:35:16
```

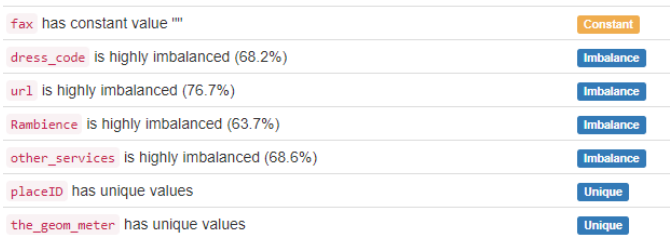
YData found that all columns have Unique values, but did not find the problem in Perfume_name



18- 225 - Restaurant & consumer data – Computer

Dataset analysed: geoplaces2.csv from https://archive.ics.uci.edu/ml/machine-learning-databases/00232/RCdata.zip (dataset_file_url column from Fortydatasets.xlsx) has header but it's format is not UTF-8.

```
placeID (id):
  ID column format: All 130 ID values are unique and valid, and thus suitable for use as a Primary Key.
Alphanumeric range of values: (132560 : 135109)
```



YData did not find the many '?'s, but found two Unique values that we also found, as ID and in the categorical column below that has all 130 values as unique, but we have not alerted it. And it found that fax column has only one value, considered Constant, but considered for them a different value than the '?' we found.

```
latitude:
  Latitude format: All 130 values are numerical and valid in the range [-90, 90].
Actual range of values: (18.859803 : 23.7602683)

longitude:
  Longitude format: All 130 values are numerical and valid in the range [-180, 180].
Actual range of values: (-101.0286 : -99.1265059)

the_geom_meter (nominal):
  Error(s) found:
DQI #10 (Structural Conflicts - Consistency, Uniqueness):
 Data seems not categorical or has too many categories (> 100). Sample values: ['0101000020957F000088568DE356715AC138C0A525FC464A41',
'0101000020957F00001AD016568C4858C1243261274BA54B41', '0101000020957F0000649D6F21634858C119AE9BF528A34B41',
'0101000020957F00005D67BCDDED8157C1222A2DC8D84D4941', '0101000020957F00008EBA2D06DC8157C194E03B7B504E4941',
'0101000020957F00001B552189B84A58C15A2AAEFD2CA24B41', '0101000020957F00008A20E615808157C16272FECBF84F4941',
'0101000020957F00008A2A0747DE4758C11EB31D2A31A84B41', '0101000020957F0000A478418BBA8057C133851EB22C4E4941',
'0101000020957F0000A29FAF95CD4958C1FEEEBB73A9914B41']
Categorical format with 130 unique values:
                                        Category Frequency
0101000020957F000000DD3546816E5AC119D4BD17FD544A41         1
0101000020957F000003B195E25F8457C1C535BD04614B4941         1
0101000020957F000004457BB7AA8657C15F10835CD9444941         1
0101000020957F000005810F19B84858C136805B2745A74B41         1
0101000020957F00000B6735CA004858C108FD525CB2A44B41         1
0101000020957F00000F14BF6B2C8657C1963CCB8E5C464941         1
```

```
0101000020957F00000F624C60B54958C11757339BCCA24B41         1
0101000020957F000011DFB4E3EE4858C13F78758452AB4B41         1
0101000020957F000011E92CCE714B58C19BF8C0CA75924B41         1
0101000020957F000013696871A24558C11EE432FB04A14B41         1
                                                         ...
0101000020957F0000F9F19DDC3B4858C191B265A83BA44B41         1
0101000020957F0000FA1A0E5A9B4858C17C884C4173AE4B41         1
0101000020957F0000FBE7171F056F5AC1E8A6C0A5AF554A41         1
0101000020957F0000FC60BDA8E88157C1B2C357D6DA4E4941         1
0101000020957F0000FC799354656C5AC1233FCC70E7584A41         1
0101000020957F0000FC8866CF17785AC14CD7055B02514A41         1
0101000020957F0000FD78FFF7AD4458C1896A33F029A04B41         1
0101000020957F0000FDF8D26EE08157C1FEDB6A1FDB4E4941         1
0101000020957F0000FE987FAF936D5AC1E1D486215E524A41         1
0101000020957F0000FEC3FB453E4B58C19B4A463617994B41         1
```

name:
  Name format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 3 Extraneous data value(s) at index(es): [(10, 'Restaurante 75'), (66, 'Cenaduria El RincÃ³n de Tlaquepaque'), (103, 'Carnitas Mata Calle 16 de Septiembre')]

DQI #15 (Domain Violation - Accuracy):
 57 Capitalization/Format issue(s) at index(es): [(1, 'puesto de tacos'), (3, 'little pizza Emilio Portes Gil'), (4, 'carnitas_mata'), (5, 'Restaurant los Compadres'), (6, 'Taqueria EL amigo '), (7, 'shi ro ie'), (9, 'la Estrella de Dimas'), (12, 'El angel Restaurante'), (15, 'Tortas y hamburguesas el gordo'), (17, 'rockabilly '), ('...', '...'), (109, 'carnitas mata calle Emilio Portes Gil'), (110, 'crudalia'), (111, 'tacos de barbacoa enfrente del Tec'), (115, 'Cafeteria cenidet'), (119, 'Cafeteria y Restaurant El Pacifico'), (122, 'tacos abi'), (123, 'la perica hamburguesa'), (124, 'McDonalds Centro'), (128, 'Restaurant Bar Coty y Pablo'), (129, 'sirloin stockade')] (displaying only the first and last 10 items)

Frequency Distribution:

name	Frequency
Gorditas Dona Tota	2
Abondance Restaurante Bar	1
Arrachela Grill	1
Cabana Huasteca	1
Cafe Chaires	1
Cafeteria cenidet	1
Cafeteria y Restaurant El Pacifico	1
Carls Jr	1
Carnitas Mata Calle 16 de Septiembre	1
Carreton de Flautas y Migadas	1
...	...
puesto de tacos	1
rockabilly	1
shi ro ie	1
sirloin stockade	1
tacos abi	1
tacos de barbacoa enfrente del Tec	1
tacos de la estacion	1
tacos los volcanes	1
tortas hawai	1
vips	1

Range of Values: (Abondance Restaurante Bar to vips)

address:
  Street format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 28 Extraneous street data value(s) at index(es): [(7, '?'), (19, '?'), (23, '?'), (27, '?'), (31, '?'), (43, '?'), (45, '?'), (62, '?'), (70, '?'), (72, '?'), ('...', '...'), (95, '?'), (96, '?'), (107, '?'), (110, '?'), (111, '?'), (118, '?'), (122, '?'), (123, '?'), (127, '?'), (129, '?')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):
 26 Capitalization/Format issue(s) at index(es): [(1, 'esquina santos degollado y leon guzman'), (3, 'calle emilio portes gil'), (4, 'lic. Emilio portes gil'), (5, 'Camino a Simon Diaz 155 Centro'), (8, 'tampico'), (16, 'carr. mexico'), (17, 'agustin de iturbide'), (26, 'avenida salvador montiel '), (29, 'r.b. anaya esq. florencia'), (34, '1a. de Lozada 1'), ('...', '...'), (80, 'sevilla y olmedo 715 a'), (81, 'circuito oriente esq. carretera 57'), (85, 'Av. Saan Luis enttre moctezuma y salinas'), (89, 'Zaragoza entre Francisco Zarco y Lopez Velarde'), (93, 'avenida hivno nacional'), (98, 'Paseo de las Fuentes'), (101, 'Av. de los Pintores'), (106, 'frente al tecnologico'), (114, 'Himno nacional esq. Blvd. Juarez'), (124, 'Rayon sn col. Centro')] (displaying only the first and last 10 items)

city:
  City format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 18 Extraneous data value(s) at index(es): [(7, '?'), (27, '?'), (31, '?'), (43, '?'), (45, '?'), (62, '?'), (70, '?'), (72, '?'), (83, '?'), (87, '?'), (95, '?'), (96, '?'), (107, '?'), (110, '?'), (111, '?'), (118, '?'), (127, '?'), (129, '?')]

DQI #15 (Domain Violation - Accuracy):
 23 Capitalization/Format issue(s) at index(es): [(1, 's.l.p.'), (3, 'victoria '), (4, 'victoria'), (8, 'victoria'), (17, 'san luis potosi'), (19, 'victoria'), (26, 'cuernavaca'), (29, 'slp'), (39, 'san luis potosi'), (41, 'victoria'), ('...', '...'), (76, 'victoria'), (80, 'san luis potosi'), (81, 'san luis potosi '), (90, 's.l.p'), (93, 'san luis potosi'), (103, 'victoria '), (106, 'victoria'), (109, 'victoria'), (122, 'victoria'), (123, 'victoria')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 15 categories):

City	Frequency
San Luis Potosi	64
?	18
Cuernavaca	15
victoria	12
san luis potosi	6
Jiutepec	4
Ciudad Victoria	2

```
        Soledad          2
    Cd Victoria          1
   Cd. Victoria          1
     cuernavaca          1
          s.l.p          1
         s.l.p.          1
 san luis potos          1
            slp          1


state:
  State format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 18 Extraneous data value(s) at index(es): [(7, '?'), (27, '?'), (31, '?'), (43, '?'), (45, '?'), (56, '?'), (70, '?'), (72, '?'), (83,
'?'), (87, '?'), (95, '?'), (96, '?'), (107, '?'), (110, '?'), (111, '?'), (118, '?'), (127, '?'), (129, '?')]

DQI #15 (Domain Violation - Accuracy):
 20 Capitalization/Format issue(s) at index(es): [(1, 's.l.p.'), (3, 'tamaulipas'), (17, 'san luis potosi'), (19, 'tamaulipas'), (26,
'morelos'), (29, 'slp'), (39, 'san luis potosi'), (46, 'mexico'), (66, 'san luis potosi'), (73, 'tamaulipas'), (76, 'tamaulipas'), (80,
'san luis potosi'), (81, 'slp'), (90, 'mexico'), (93, 'san luis potos'), (103, 'tamaulipas'), (106, 'tamaulipas'), (109, 'tamaulipas'),
(122, 'tamaulipas'), (123, 'tamaulipas')]


Frequency Distribution (showing top and bottom 10 of 13 categories):
          State  Frequency
            SLP         50
        Morelos         19
              ?         18
San Luis Potosi         14
     tamaulipas          9
     Tamaulipas          7
san luis potosi          4
         S.L.P.          2
         mexico          2
            slp          2
        morelos          1
         s.l.p.          1
 san luis potos          1


country:
  Country format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 28 Extraneous data value(s) at index(es): [(3, '?'), (7, '?'), (27, '?'), (31, '?'), (39, '?'), (43, '?'), (45, '?'), (46, '?'), (56,
'?'), (70, '?'), ('...', '...'), (98, '?'), (103, '?'), (107, '?'), (109, '?'), (110, '?'), (111, '?'), (118, '?'), (123, '?'), (127,
'?'), (129, '?')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):
 13 Capitalization/Format issue(s) at index(es): [(1, 'mexico'), (17, 'mexico'), (19, 'mexico'), (26, 'mexico'), (29, 'mexico'), (66,
'mexico'), (76, 'mexico'), (80, 'mexico'), (81, 'mexico'), (85, 'mexico'), (93, 'mexico'), (106, 'mexico'), (122, 'mexico')]


Frequency Distribution:
Country  Frequency
 Mexico         89
      ?         28
 mexico         13


fax:
  Phone format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 130 Incorrect telephone number format issue(s) at index(es): [(0, '?'), (1, '?'), (2, '?'), (3, '?'), (4, '?'), (5, '?'), (6, '?'),
(7, '?'), (8, '?'), (9, '?'), ('...', '...'), (120, '?'), (121, '?'), (122, '?'), (123, '?'), (124, '?'), (125, '?'), (126, '?'), (127,
'?'), (128, '?'), (129, '?')] (displaying only the first and last 10 items)
```

fax has constant value "" Constant

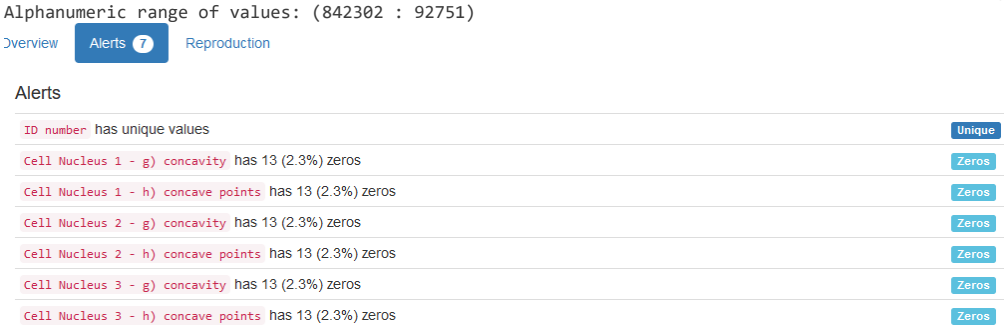
```
zip:
  Postal Code format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 74 Non-alphanumeric value(s) at index(es): [(0, '?'), (3, '?'), (4, '?'), (7, '?'), (8, '?'), (15, '?'), (16, '?'), (17, '?'), (18,
'?'), (19, '?'), ('...', '...'), (115, '?'), (117, '?'), (118, '?'), (120, '?'), (122, '?'), (123, '?'), (125, '?'), (127, '?'), (128,
'?'), (129, '?')] (displaying only the first and last 10 items)


alcohol (nominal):
  All 130 values are correctly categorical.

Categorical format with 3 unique values:
         Category  Frequency
No_Alcohol_Served         87
        Wine-Beer         34
         Full_Bar          9

smoking_area (nominal):
  All 130 values are correctly categorical.

Categorical format with 5 unique values:
     Category  Frequency
         none         70
not permitted         25
```

```
      section         24
    permitted          9
  only at bar          2

dress_code (code):
  All 130 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
informal        118
  casual         10
  formal          2

accessibility (nominal):
  All 130 values are correctly categorical.

Categorical format with 3 unique values:
        Category  Frequency
no_accessibility         76
      completely         45
       partially          9

price (nominal):
  All 130 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
  medium         60
     low         45
    high         25

url:
  URL format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 117 Invalid URL format issue(s) at index(es): [(1, '?'), (2, '?'), (3, '?'), (4, '?'), (5, '?'), (6, '?'), (7, '?'), (8, '?'), (9,
'?'), (10, '?'), ('...', '...'), (119, '?'), (120, '?'), (121, '?'), (122, '?'), (123, '?'), (124, 'no'), (125, '?'), (127, '?'), (128,
'?'), (129, '?')] (displaying only the first and last 10 items)


Rambience (nominal):
  All 130 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
familiar        121
   quiet          9

franchise (ranch):
  All 130 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       f        108
       t         22

area (nominal):
  All 130 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
  closed        115
    open         15

other_services (service):
  All 130 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
    none        119
 variety          7
Internet          4

Last run on: 2024-02-22 15:41:41
```

19- 17 - Breast Cancer Wisconsin (Diagnostic) – Life

Dataset analysed: wdbc.data from https://archive.ics.uci.edu/static/public/17/breast+cancer+wisconsin+diagnostic.zip instead of the https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data that was in the dataset_file_url column from Fortydatasets.xlsx. Curiously this same dataset is being used for a similar dataset, 15 - Breast Cancer Wisconsin (Original) – Life, that is presented later on this document. It does not have header.

```
ID number (id):
  ID column format: All 569 ID values are unique and valid, and thus suitable for use as a Primary Key.
Alphanumeric range of values: (842302 : 92751)
```



YData shows the column 'ID number' as Unique. We consider it not only unique but suitable for use as a Primary Key.

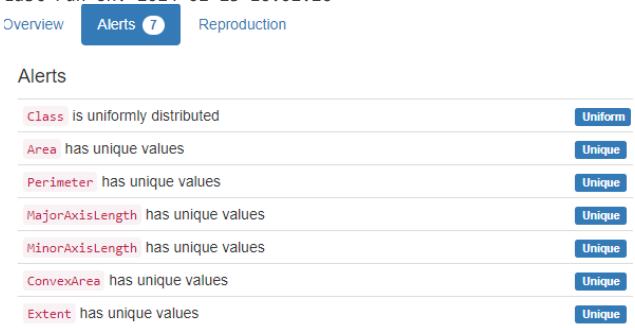
```
Diagnosis:
  All 569 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       B        357
       M        212

Cell Nucleus 1 - a) radius (radius):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (6.981:28.11).

Cell Nucleus 1 - b) texture (texture):
  Numerical format: All 569 values are numerical in the range (9.71:39.28).

Cell Nucleus 1 - c) perimeter (perimeter):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (43.79:188.5).

Cell Nucleus 1 - d) area (area):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (143.5:2501.0).

Cell Nucleus 1 - e) smoothness (smoothness):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.05263:0.1634).

Cell Nucleus 1 - f) compactness (compactness):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.01938:0.3454).

Cell Nucleus 1 - g) concavity (concavity):
  Numerical format: All 569 values are numerical in the range (0.0:0.4268).

Cell Nucleus 1 - h) concave points (points):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.0:0.2012).

Cell Nucleus 1 - i) symmetry (symmetry):
  Numerical format: All 569 values are numerical in the range (0.106:0.304).

Cell Nucleus 1 - j) fractal dimension (dimension):
  Numerical format: All 569 values are numerical in the range (0.04996:0.09744).

Cell Nucleus 2 - a) radius (radius):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.1115:2.873).

Cell Nucleus 2 - b) texture (texture):
  Numerical format: All 569 values are numerical in the range (0.3602:4.885).

Cell Nucleus 2 - c) perimeter (perimeter):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.757:21.98).

Cell Nucleus 2 - d) area (area):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (6.802:542.2).

Cell Nucleus 2 - e) smoothness (smoothness):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.001713:0.03113).

Cell Nucleus 2 - f) compactness (compactness):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.002252:0.1354).

Cell Nucleus 2 - g) concavity (concavity):
  Numerical format: All 569 values are numerical in the range (0.0:0.396).

Cell Nucleus 2 - h) concave points (points):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.0:0.05279).

Cell Nucleus 2 - i) symmetry (symmetry):
  Numerical format: All 569 values are numerical in the range (0.007882:0.07895).
```

```
Cell Nucleus 2 - j) fractal dimension (dimension):
  Numerical format: All 569 values are numerical in the range (0.0008948:0.02984).

Cell Nucleus 3 - a) radius (radius):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (7.93:36.04).

Cell Nucleus 3 - b) texture (texture):
  Numerical format: All 569 values are numerical in the range (12.02:49.54).

Cell Nucleus 3 - c) perimeter (perimeter):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (50.41:251.2).

Cell Nucleus 3 - d) area (area):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (185.2:4254.0).

Cell Nucleus 3 - e) smoothness (smoothness):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.07117:0.2226).

Cell Nucleus 3 - f) compactness (compactness):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.02729:1.058).

Cell Nucleus 3 - g) concavity (concavity):
  Numerical format: All 569 values are numerical in the range (0.0:1.252).

Cell Nucleus 3 - h) concave points (points):
  Numerical >=0 format: All 569 values are numerical and greater or equal to 0 in the range (0.0:0.291).

Cell Nucleus 3 - i) symmetry (symmetry):
  Numerical format: All 569 values are numerical in the range (0.1565:0.6638).

Cell Nucleus 3 - j) fractal dimension (dimension):
  Numerical format: All 569 values are numerical in the range (0.05504:0.2075).

Last run on: 2024-02-23 18:27:01
```

20- 850 - Raisin Dataset – Life

Dataset analysed: Raisin_Dataset\Raisin_Dataset.xlsx from https://archive.ics.uci.edu/ml/machine-learning-databases/00617/Raisin_Dataset.zip (dataset_file_url column from Fortydatasets.xlsx) has header

```
Area:
  Numerical >=0 format: All 900 values are numerical and greater or equal to 0 in the range (25387:235047).

Perimeter:
  Numerical >=0 format: All 900 values are numerical and greater or equal to 0 in the range (225.629541:997.2919406).

MajorAxisLength (length):
  Numerical >=0 format: All 900 values are numerical and greater or equal to 0 in the range (143.7108718:492.2752785).

MinorAxisLength (minimum):
  Numerical format: All 900 values are numerical in the range (0.348729642:0.96212444).

Eccentricity:
  Numerical format: All 900 values are numerical in the range (26139:278217).

ConvexArea (area):
  Numerical >=0 format: All 900 values are numerical and greater or equal to 0 in the range (0.379856115:0.835454545).

Extent (ratio):
  Numerical format: All 900 values are numerical in the range (619.074:2697.753).

Class:
  All 900 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
   Besni        450
 Kecimen        450

Last run on: 2024-02-23 16:02:26
```



YData shows 6 columns as Unique. We have not checked that.

21- 1 - Abalone – Life

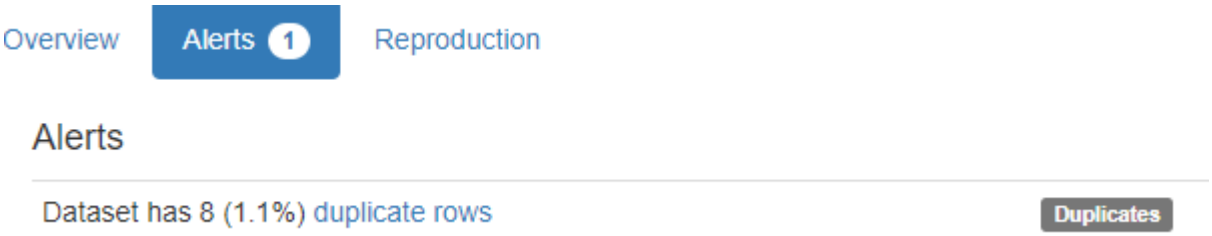
Dataset analysed: abalone.data from https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header

```
Sex:
  All 4177 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
       M       1528
       I       1342
       F       1307

Length:
  Numerical >=0 format: All 4177 values are numerical and greater or equal to 0 in the range (0.075:0.815).

Diameter:
  Numerical >=0 format: All 4177 values are numerical and greater or equal to 0 in the range (0.055:0.65).

Height:
  Numerical format: All 4177 values are numerical in the range (0.0:1.13).

Whole weight (weight):
  Numerical >=0 format: All 4177 values are numerical and greater or equal to 0 in the range (0.002:2.8255).

Shucked weight (weight):
  Numerical >=0 format: All 4177 values are numerical and greater or equal to 0 in the range (0.001:1.488).

Viscera weight (weight):
  Numerical >=0 format: All 4177 values are numerical and greater or equal to 0 in the range (0.0005:0.76).

Shell weight (weight):
  Numerical >=0 format: All 4177 values are numerical and greater or equal to 0 in the range (0.0015:1.005).

Rings (integer):
  Numerical format: All 4177 values are numerical in the range (1:29).

Last run on: 2024-02-23 16:26:24
```

YData did not find any Alerts in this Dataset

22- 15 - Breast Cancer Wisconsin (Original) – Life

Dataset analysed: breast-cancer-wisconsin.data from https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.



YData only found Duplicate entire rows. We found a duplicate ID column, and other problem in the first ID column. We also found '?' in 'Bare Nuclei'

```
Sample code number (id):
  ID column format: Error(s) found:
DQI #9 (Duplicates - Uniqueness):
 100 Duplicate value(s) at index(es): [[4, 1017023], [8, 1033078], [9, 1033078], [29, 1070935], [30, 1070935], [42, 1100524], [47,
1105524], [61, 1115293], [62, 1116116], [64, 1116192], ('...', '...'), [661, 1339781], [672, 1354840], [673, 1354840], [683, 466906],
[684, 466906], [689, 654546], [690, 654546], [691, 695091], [697, 897471], [698, 897471]] (displaying only the first and last 10 items)

DQI #17 (Wrong Data Type - Consistency):
 188 Inconsistent length in alphanumeric value(s) at index(es): [(243, 128059), (246, 144888), (247, 145447), (248, 167528), (249,
169356), (250, 183913), (251, 191250), (259, 242970), (260, 255644), (261, 263538), ('...', '...'), (689, 654546), (690, 654546), (691,
695091), (692, 714039), (693, 763235), (694, 776715), (695, 841769), (696, 888820), (697, 897471), (698, 897471)] (displaying only the
first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):
 54 Uniqueness violation(s) at index(es): [(9, 1033078), (30, 1070935), (82, 1143978), (109, 1171710), (116, 1173347), (121, 1174057),
(195, 1212422), (208, 1218860), (252, 1017023), (253, 1100524), ('...', '...'), (618, 1061990), (632, 1238777), (639, 1277792), (644,
1299596), (661, 1339781), (673, 1354840), (684, 466906), (690, 654546), (691, 695091), (698, 897471)] (displaying only the first and
last 10 items)

Alphanumeric range of values: (1000025 : 95719)

Clump Thickness (thickness):
  Numerical format: All 699 values are numerical in the range (1:10).
```

```
Uniformity of Cell Size (size):
  Numerical >=0 format: All 699 values are numerical and greater or equal to 0 in the range (1:10).

Uniformity of Cell Shape (uniformity):
  Numerical format: All 699 values are numerical in the range (1:10).

Marginal Adhesion (adhesion):
  Numerical format: All 699 values are numerical in the range (1:10).

Single Epithelial Cell Size (size):
  Numerical >=0 format: All 699 values are numerical and greater or equal to 0 in the range (1:10).

Bare Nuclei (nuclei):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 16 Non-numeric value(s) at index(es): [(23, '?'), (40, '?'), (139, '?'), (145, '?'), (158, '?'), (164, '?'), (235, '?'), (249, '?'),
(275, '?'), (292, '?'), (294, '?'), (297, '?'), (315, '?'), (321, '?'), (411, '?'), (617, '?')]

Range of values: (1.0:10.0).

Bland Chromatin (chromatin):
  Numerical format: All 699 values are numerical in the range (1:10).

Normal Nucleoli (nucleoli):
  Numerical format: All 699 values are numerical in the range (1:10).

Mitoses:
  Numerical format: All 699 values are numerical in the range (1:10).

Class:
  All 699 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        2        458
        4        241

Last run on: 2024-02-23 18:29:53
```

23- 73 - Mushroom - Life

Dataset analysed: agaricus-lepiota.data from https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

All fields are categorical and have been altered manually in the AllColumnsfromFortyDatasets file.

```
poisonous (categorical):
  All 8124 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       e       4208
       p       3916

cap-shape (categorical):
  All 8124 values are correctly categorical.

Categorical format with 6 unique values:
Category  Frequency
       x       3656
       f       3152
       k        828
       b        452
       s         32
       c          4

cap-surface (categorical):
  All 8124 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
       y       3244
       s       2556
       f       2320
       g          4

cap-color (categorical):
  All 8124 values are correctly categorical.

Categorical format with 10 unique values:
Category  Frequency
       n       2284
       g       1840
       e       1500
       y       1072
       w       1040
       b        168
       p        144
       c         44
       r         16
```

```
       u         16

bruises? (categorical):
  All 8124 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       f       4748
       t       3376

odor (categorical):
  All 8124 values are correctly categorical.

Categorical format with 9 unique values:
Category  Frequency
       n       3528
       f       2160
       s        576
       y        576
       a        400
       l        400
       p        256
       c        192
       m         36

gill-attachment (categorical):
  All 8124 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       f       7914
       a        210

gill-spacing (categorical):
  All 8124 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       c       6812
       w       1312

gill-size (categorical):
  All 8124 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       b       5612
       n       2512

gill-color (color):
  All 8124 values are correctly categorical.

Categorical format with 12 unique values:
Category  Frequency
       b       1728
       p       1492
       w       1202
       n       1048
       g        752
       h        732
       u        492
       k        408
       e         96
       y         86
       o         64
       r         24

stalk-shape (categorical):
  All 8124 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       t       4608
       e       3516

stalk-root (categorical):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 2480 Unacceptable value(s) at index(es): [(3984, '?'), (4023, '?'), (4076, '?'), (4100, '?'), (4104, '?'), (4196, '?'), (4200, '?'),
(4283, '?'), (4291, '?'), (4326, '?'), ('...', '...'), (8113, '?'), (8115, '?'), (8116, '?'), (8117, '?'), (8118, '?'), (8119, '?'),
(8120, '?'), (8121, '?'), (8122, '?'), (8123, '?')] (displaying only the first and last 10 items)
Categorical format with 5 unique values:
Category  Frequency
       b       3776
       ?       2480
       e       1120
       c        556
       r        192

stalk-surface-above-ring (categorical):
  All 8124 values are correctly categorical.

Categorical format with 4 unique values:
```

```
Category  Frequency
       s       5176
       k       2372
       f        552
       y         24

stalk-surface-below-ring (categorical):
  All 8124 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
       s       4936
       k       2304
       f        600
       y        284

stalk-color-above-ring (color):
  All 8124 values are correctly categorical.

Categorical format with 9 unique values:
Category  Frequency
       w       4464
       p       1872
       g        576
       n        448
       b        432
       o        192
       e         96
       c         36
       y          8

stalk-color-below-ring (color):
  All 8124 values are correctly categorical.

Categorical format with 9 unique values:
Category  Frequency
       w       4384
       p       1872
       g        576
       n        512
       b        432
       o        192
       e         96
       c         36
       y         24

veil-type (type):
  All 8124 values are correctly categorical.

Categorical format with 1 unique values:
Category  Frequency
       p       8124
```

Alerts

veil-type has constant value "" Constant

YData just found that this column has only one value. We showed it but have not considered an Alert. Besides that it did not find that ‘stalk-root’ had ‘?’ values

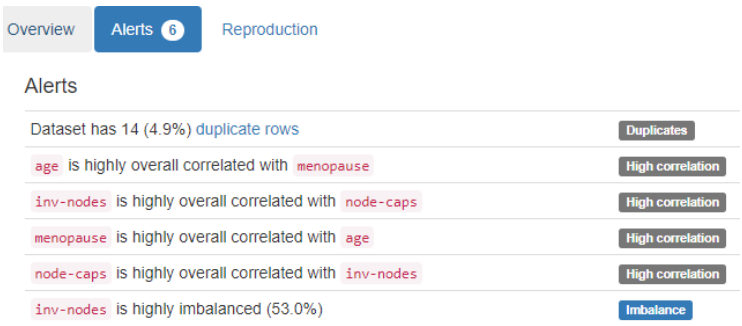
Other Alerts they show are High Correlation and Imbalance

```
veil-color (color):
  All 8124 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
       w       7924
       n         96
       o         96
       y          8

ring-number (categorical):
  All 8124 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
       o       7488
       t        600
       n         36

ring-type (type):
  All 8124 values are correctly categorical.

Categorical format with 5 unique values:
Category  Frequency
       p       3968
       e       2776
       l       1296
       f         48
       n         36
```

```
spore-print-color (color):
  All 8124 values are correctly categorical.

Categorical format with 9 unique values:
Category  Frequency
       w       2388
       n       1968
       k       1872
       h       1632
       r         72
       b         48
       o         48
       u         48
       y         48

population (categorical):
  All 8124 values are correctly categorical.

Categorical format with 6 unique values:
Category  Frequency
       v       4040
       y       1712
       s       1248
       n        400
       a        384
       c        340

habitat (categorical):
  All 8124 values are correctly categorical.

Categorical format with 7 unique values:
Category  Frequency
       d       3148
       g       2148
       p       1144
       l        832
       u        368
       m        292
       w        192

Last run on: 2024-02-23 20:33:28
```

24- 14 - Breast Cancer – Life

Dataset analysed: breast-cancer.data from https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/breast-cancer.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.



YData just found whole Duplicate lines, High correlation and Imbalance. Not the 2 cases of '?' values.

```
Class:
  All 286 values are correctly categorical.

Categorical format with 2 unique values:
            Category  Frequency
no-recurrence-events        201
   recurrence-events         85

age (10-):
  All 286 values are correctly categorical.

Categorical format with 6 unique values:
Category  Frequency
   50-59         96
   40-49         90
   60-69         57
   30-39         36
   70-79          6
   20-29          1

menopause:
  All 286 values are correctly categorical.
```

```
Categorical format with 3 unique values:
Category  Frequency
 premeno        150
    ge40        129
    lt40          7

tumor-size (0-):
  All 286 values are correctly categorical.

Categorical format with 11 unique values:
Category  Frequency
   30-34         60
   25-29         54
   20-24         50
   15-19         30
   10-14         28
   40-44         22
   35-39         19
     0-4          8
   50-54          8
     5-9          4
   45-49          3

inv-nodes (node):
  All 286 values are correctly categorical.

Categorical format with 7 unique values:
Category  Frequency
     0-2        213
     3-5         36
     6-8         17
    9-11         10
   15-17          6
   12-14          3
   24-26          1

node-caps (node):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 8 Unacceptable value(s) at index(es): [(145, '?'), (163, '?'), (164, '?'), (183, '?'), (184, '?'), (233, '?'), (263, '?'), (264, '?')]
Categorical format with 3 unique values:
Category  Frequency
      no        222
     yes         56
       ?          8

deg-malig (degree):
  All 286 values are correctly categorical.

Categorical format with 3 unique values:
 Category  Frequency
        2        130
        3         85
        1         71

breast:
  All 286 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
    left        152
   right        134

breast-quad (breast):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 1 Unacceptable value(s) at index(es): [(206, '?')]
Categorical format with 6 unique values:
 Category  Frequency
 left_low        110
  left_up         97
 right_up         33
right_low         24
  central         21
        ?          1

irradiat (yes):
  All 286 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
      no        218
     yes         68

Last run on: 2024-02-25 11:33:55
```

25- 236 - seeds – Life

Dataset analysed: seeds_dataset.txt from archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

```
15.26    14.84    0.871    5.763    3.312    2.221    5.22     1
14.88    14.57    0.8811   5.554    3.333    1.018    4.956    1
14.29    14.09    0.905    5.291    3.337    2.699    4.825    1
13.84    13.94    0.8955   5.324    3.379    2.259    4.805    1
16.14    14.99    0.9034   5.658    3.562    1.355    5.175    1
14.38    14.21    0.8951   5.386    3.312    2.462    4.956    1
14.69    14.49    0.8799   5.563    3.259    3.586    5.219    1
14.11    14.1     0.8911   5.42     3.302    2.7      5        1
16.63    15.46    0.8747   6.053    3.465    2.04     5.877    1
16.44    15.25    0.888    5.884    3.505    1.969    5.533    1
15.26    14.85    0.8696   5.714    3.242    4.543    5.314    1
```

The code was altered so that two tabs now are just one.

```
area A (area):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (10.59:21.18).

perimeter P (perimeter):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (12.41:17.25).

compactness C = 4*pi*A (compactness):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (0.8081:0.9183).

length of kernel (length):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (4.899:6.675).

width of kernel (width):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (2.63:4.033).

asymmetry coefficient (coefficient):
  Numerical format: All 210 values are numerical in the range (0.7651:8.456).

length of kernel groove (length):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (4.519:6.55).

class:
  All 210 values are correctly categorical.

Categorical format with 3 unique values:
 Category  Frequency
        1         70
        2         70
        3         70

Last run on: 2024-02-25 12:14:03
```

Overview | Alerts 8 | Reproduction

Alerts

area A is highly overall correlated with class and 5 other fields	High correlation
class is highly overall correlated with area A and 1 other fields	High correlation
compactness C = 4*pi*A is highly overall correlated with area A and 3 other fields	High correlation
length of kernel is highly overall correlated with area A and 3 other fields	High correlation
length of kernel groove is highly overall correlated with area A and 3 other fields	High correlation
perimeter P is highly overall correlated with area A and 4 other fields	High correlation
width of kernel is highly overall correlated with area A and 4 other fields	High correlation
class is uniformly distributed	Uniform

YData did not find anything that we care.

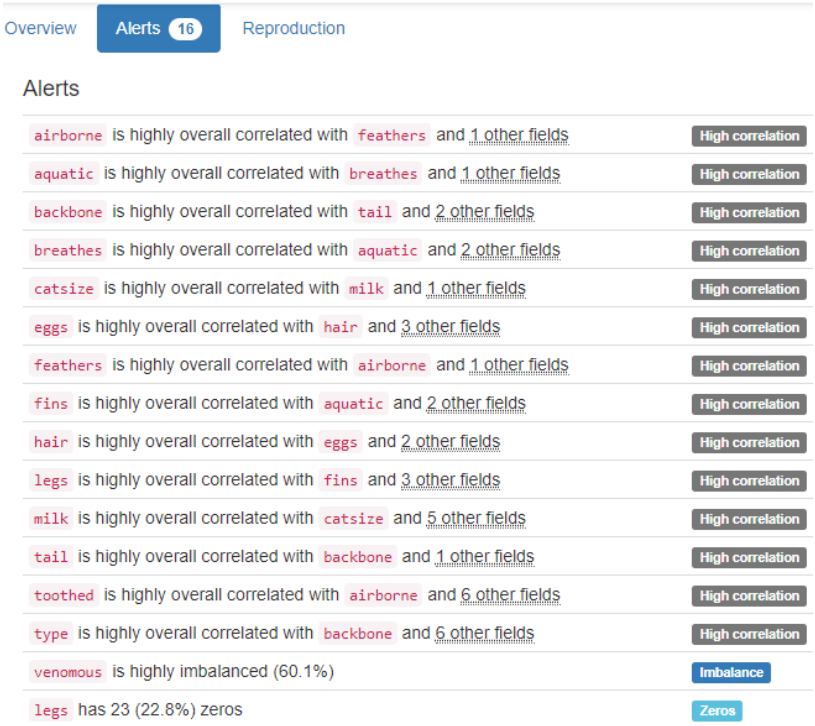
26- 111 - Zoo - Life

Dataset analysed: zoo.data from archive.ics.uci.edu/ml/machine-learning-databases/zoo/zoo.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

```
animal name (name):
  Name format: All 101 animal name values are valid.

Frequency Distribution:
animal name Frequency
       frog         2
   aardvark         1
   antelope         1
       bass         1
       bear         1
```

```
       boar         1
    buffalo         1
       calf         1
       carp         1
    catfish         1
        ...       ...
    tuatara         1
       tuna         1
    vampire         1
       vole         1
    vulture         1
    wallaby         1
       wasp         1
       wolf         1
       worm         1
       wren         1

Range of Values: (aardvark to wren)

hair (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         58
     1         43


feathers (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         81
     1         20


eggs (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     1         59
     0         42


milk (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         60
     1         41


airborne (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         77
     1         24


aquatic (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         65
     1         36


predator (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     1         56
     0         45


toothed (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     1         61
     0         40


backbone (boolean):
  Binary format: All 101 binary values are valid.
```

```
Frequency Distribution:
 Value  Frequency
     1         83
     0         18


breathes (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     1         80
     0         21


venomous (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         93
     1          8


fins (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         84
     1         17


legs (numeric):
  Numerical format: All 101 values are numerical in the range (0:8).

tail (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     1         75
     0         26


domestic (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         88
     1         13


catsize (boolean):
  Binary format: All 101 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0         57
     1         44


type (numeric):
  Numerical format: All 101 values are numerical in the range (1:7).

Last run on: 2024-02-25 16:52:25
```



YData did not find anything that we care.

27- 31 - Covertype – Life

Dataset analysed: covtype.data from https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.data.gz (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

```
Elevation:
  Numerical format: All 581012 values are numerical in the range (1859:3858).

Aspect (speed):
  Numerical format: All 581012 values are numerical in the range (0:360).

Slope:
  Numerical format: All 581012 values are numerical in the range (0:66).

Horizontal_Distance_To_Hydrology (distance):
  Numerical format: All 581012 values are numerical in the range (0:1397).

Vertical_Distance_To_Hydrology (distance):
  Numerical format: All 581012 values are numerical in the range (-173:601).

Horizontal_Distance_To_Roadways (distance):
  Numerical format: All 581012 values are numerical in the range (0:7117).

Hillshade_9am (index):
  Numerical format: All 581012 values are numerical in the range (0:254).

Hillshade_Noon (index):
  Numerical format: All 581012 values are numerical in the range (0:254).

Hillshade_3pm (index):
  Numerical format: All 581012 values are numerical in the range (0:254).

Horizontal_Distance_To_Fire_Points (distance):
  Numerical format: All 581012 values are numerical in the range (0:7173).

Wilderness_Area_1 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     320216
     1     260796


Wilderness_Area_2 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     551128
     1      29884


Wilderness_Area_3 (binary):
  Binary format: All 581012 binary values are valid.
```

```
Frequency Distribution:
 Value  Frequency
     0     327648
     1     253364


Wilderness_Area_4 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     544044
     1      36968


Soil_Type_1 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     577981
     1       3031


Soil_Type_2 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     573487
     1       7525


Soil_Type_3 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     576189
     1       4823


Soil_Type_4 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     568616
     1      12396


Soil_Type_5 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     579415
     1       1597


Soil_Type_6 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     574437
     1       6575


Soil_Type_7 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     580907
     1        105


Soil_Type_8 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     580833
     1        179


Soil_Type_9 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
```

```
     0     579865
     1       1147


Soil_Type_10 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     548378
     1      32634


Soil_Type_11 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     568602
     1      12410


Soil_Type_12 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     551041
     1      29971


Soil_Type_13 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     563581
     1      17431


Soil_Type_14 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     580413
     1        599


Soil_Type_15 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     581009
     1          3


Soil_Type_16 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     578167
     1       2845


Soil_Type_17 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     577590
     1       3422


Soil_Type_18 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     579113
     1       1899


Soil_Type_19 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     576991
     1       4021
```

Soil_Type_20 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   571753
     1   9259

Soil_Type_21 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   580174
     1   838

Soil_Type_22 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   547639
     1   33373

Soil_Type_23 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   523260
     1   57752

Soil_Type_24 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   559734
     1   21278

Soil_Type_25 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   580538
     1   474

Soil_Type_26 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   578423
     1   2589

Soil_Type_27 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   579926
     1   1086

Soil_Type_28 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   580066
     1   946

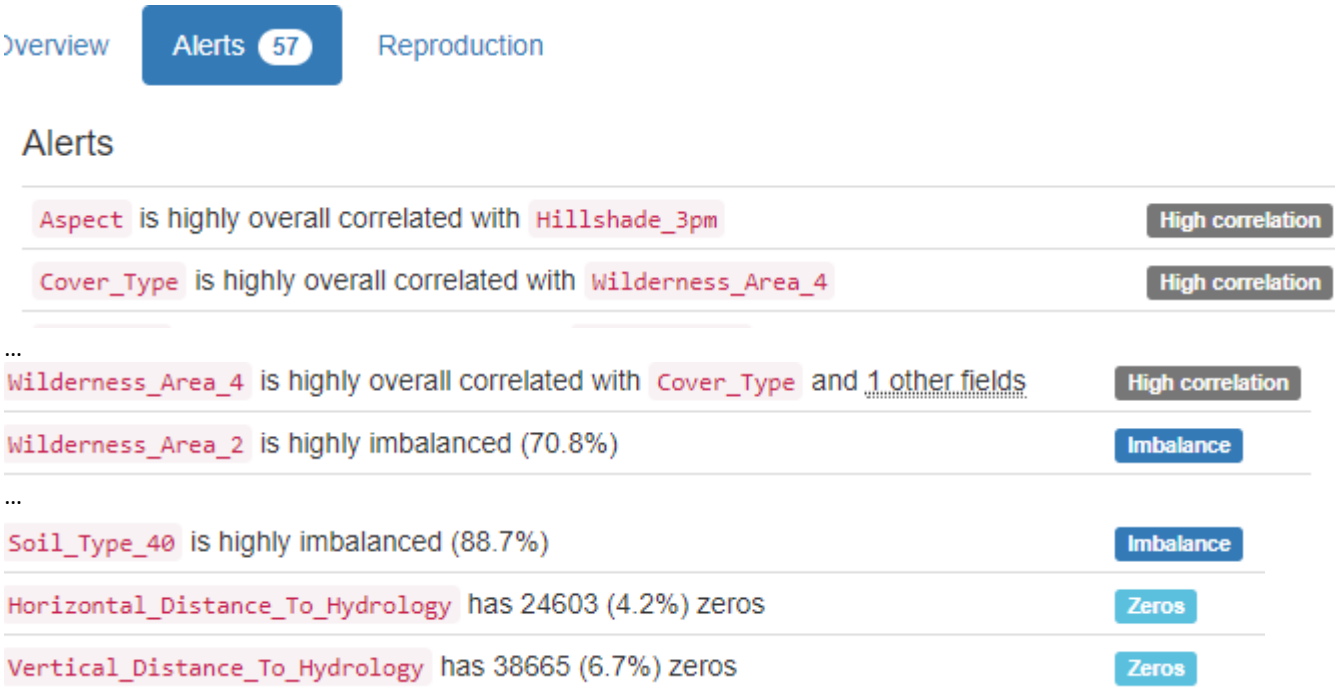
Soil_Type_29 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value Frequency
     0   465765
     1   115247

Soil_Type_30 (binary):
  Binary format: All 581012 binary values are valid.

```
Frequency Distribution:
 Value  Frequency
     0     550842
     1      30170


Soil_Type_31 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     555346
     1      25666


Soil_Type_32 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     528493
     1      52519


Soil_Type_33 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     535858
     1      45154


Soil_Type_34 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     579401
     1       1611


Soil_Type_35 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     579121
     1       1891


Soil_Type_36 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     580893
     1        119


Soil_Type_37 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     580714
     1        298


Soil_Type_38 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     565439
     1      15573


Soil_Type_39 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0     567206
     1      13806


Soil_Type_40 (binary):
  Binary format: All 581012 binary values are valid.

Frequency Distribution:
 Value  Frequency
```

```
     0     572262
     1       8750


Cover_Type (integer):
  Numerical format: All 581012 values are numerical in the range (1:7).

Last run on: 2024-02-25 17:56:27
```



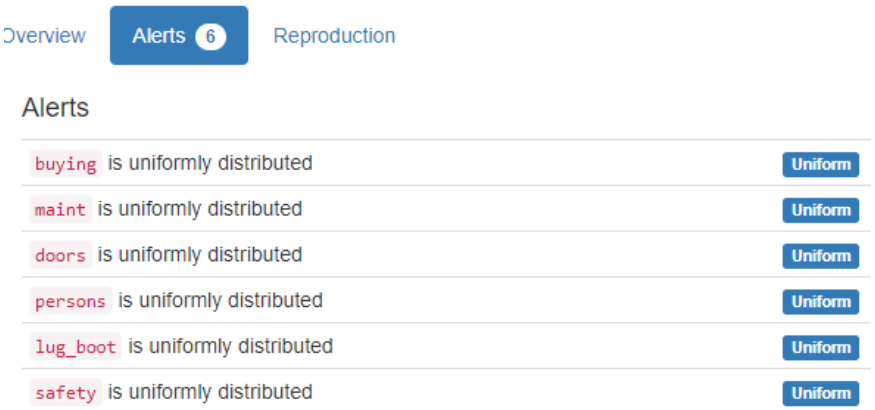
YData did not find anything that we care.

28- 19 - Car Evaluation - N/A

Dataset analysed: car.data from https://archive.ics.uci.edu/ml/machine-learning-databases/car/car.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

```
buying (high):
  All 1728 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
    high        432
     low        432
     med        432
   vhigh        432

maint (high):
  All 1728 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
    high        432
     low        432
     med        432
   vhigh        432

doors (more):
  All 1728 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
       2        432
       3        432
       4        432
   5more        432

persons (more):
  All 1728 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
       2        576
       4        576
    more        576

lug_boot (big):
  All 1728 values are correctly categorical.

Categorical format with 3 unique values:
```

```
Category  Frequency
     big        576
     med        576
   small        576

safety (high):
  All 1728 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
    high        576
     low        576
     med        576

Class Values (class):
  All 1728 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
   unacc       1210
     acc        384
    good         69
   vgood         65
```



YData did not find anything that we care.

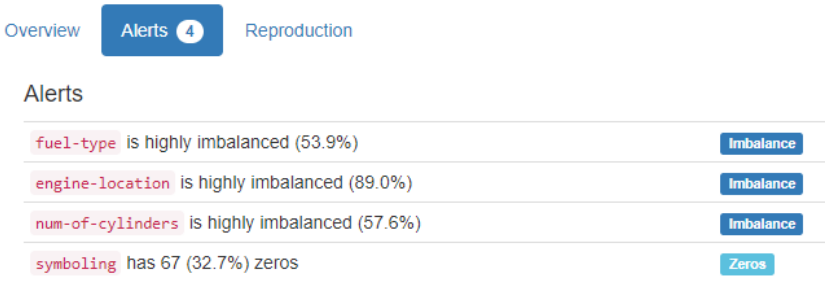
29- 10 - Automobile - N/A

Dataset analysed: imports-85.data from https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

```
symboling (symbol):
  String format: Error(s) found:
DQI #17 (Non-String Data Type - Consistency):
 205 Non-string value(s) at index(es): [(0, 3), (1, 3), (2, 1), (3, 2), (4, 2), (5, 2), (6, 1), (7, 1), (8, 1), (9, 0), ('...', '...'),
(195, -1), (196, -2), (197, -1), (198, -2), (199, -1), (200, -1), (201, -1), (202, -1), (203, -1), (204, -1)] (displaying only the
first and last 10 items)

String range (lexicographical): (-1 : 3)

normalized-losses (loss):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 41 Non-numeric value(s) at index(es): [(0, '?'), (1, '?'), (2, '?'), (5, '?'), (7, '?'), (9, '?'), (14, '?'), (15, '?'), (16, '?'),
(17, '?'), ('...', '...'), (127, '?'), (128, '?'), (129, '?'), (130, '?'), (131, '?'), (181, '?'), (189, '?'), (191, '?'), (192, '?'),
(193, '?')] (displaying only the first and last 10 items)

Range of values: (65.0:256.0).

make:
  All 205 values are correctly categorical.

Categorical format with 22 unique values:
   Category Frequency
     toyota        32
     nissan        18
      mazda        17
      honda        13
 mitsubishi        13
     subaru        12
 volkswagen        12
     peugot        11
      volvo        11
      dodge         9
        ...       ...
       audi         7
   plymouth         7
       saab         6
    porsche         5
      isuzu         4
alfa-romero         3
```

```
  chevrolet         3
     jaguar         3
    renault         2
    mercury         1

fuel-type (type):
  All 205 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
     gas        185
  diesel         20

aspiration:
  All 205 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
     std        168
   turbo         37

num-of-doors (four):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 2 Unacceptable value(s) at index(es): [(27, '?'), (63, '?')]
Categorical format with 3 unique values:
Category  Frequency
    four        114
     two         89
       ?          2

body-style (style):
  All 205 values are correctly categorical.

Categorical format with 5 unique values:
   Category  Frequency
      sedan         96
  hatchback         70
      wagon         25
    hardtop          8
convertible          6

drive-wheels (drive):
  All 205 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
     fwd        120
     rwd         76
     4wd          9

engine-location (location):
  All 205 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
   front        202
    rear          3

wheel-base (continuous):
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (86.6:120.9).

length:
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (141.1:208.1).

width:
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (60.3:72.3).

height:
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (47.8:59.8).

curb-weight (weight):
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (1488:4066).

engine-type (type):
  All 205 values are correctly categorical.

Categorical format with 7 unique values:
Category  Frequency
     ohc        148
    ohcf         15
    ohcv         13
    dohc         12
       l         12
   rotor          4
   dohcv          1

num-of-cylinders (four):
  All 205 values are correctly categorical.

Categorical format with 7 unique values:
Category  Frequency
    four        159
```

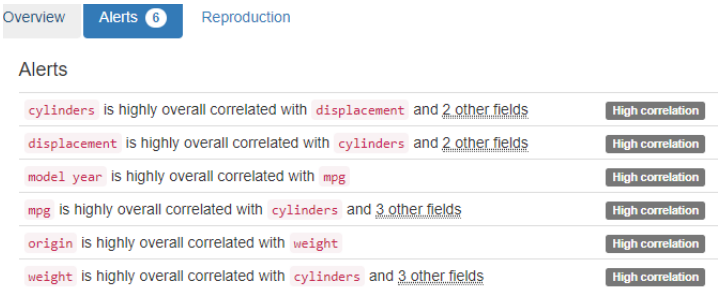
```
     six         24
    five         11
   eight          5
     two          4
   three          1
  twelve          1

engine-size (size):
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (61:326).

fuel-system (system):
  All 205 values are correctly categorical.

Categorical format with 8 unique values:
Category  Frequency
    mpfi         94
    2bbl         66
     idi         20
    1bbl         11
    spdi          9
    4bbl          3
     mfi          1
    spfi          1

bore:
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 4 Non-numeric value(s) at index(es): [(55, '?'), (56, '?'), (57, '?'), (58, '?')]

Range of values: (2.54:3.94).

stroke:
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 4 Non-numeric value(s) at index(es): [(55, '?'), (56, '?'), (57, '?'), (58, '?')]

Range of values: (2.07:4.17).

compression-ratio (ratio):
  Numerical format: All 205 values are numerical in the range (7.0:23.0).

horsepower:
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 2 Non-numeric value(s) at index(es): [(130, '?'), (131, '?')]

Range of values: (48.0:288.0).

peak-rpm (rpm):
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 2 Non-numeric value(s) at index(es): [(130, '?'), (131, '?')]

Range of values: (4150.0:6600.0).

city-mpg (mpg):
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (13:49).

highway-mpg (mpg):
  Numerical >=0 format: All 205 values are numerical and greater or equal to 0 in the range (16:54).

price:
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 4 Non-numeric value(s) at index(es): [(9, '?'), (44, '?'), (45, '?'), (129, '?')]

Range of values: (5118.0:45400.0).

Last run on: 2024-02-25 22:05:41
```



YData did not find the 7 ‘?’s

30- 9 - Auto MPG - N/A

Dataset analysed: auto-mpg.data from https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.



YData just found High Correlation. It did not find '?', Year and Name Data Quality Issues.

```
mpg:
  Numerical >=0 format: All 398 values are numerical and greater or equal to 0 in the range (9.0:46.6).

cylinders (cylinder):
  Numerical >=0 format: All 398 values are numerical and greater or equal to 0 in the range (3:8).

displacement:
  Numerical >=0 format: All 398 values are numerical and greater or equal to 0 in the range (68.0:455.0).

horsepower:
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 6 Non-numeric value(s) at index(es): [(32, '?'), (126, '?'), (330, '?'), (336, '?'), (354, '?'), (374, '?')]

Range of values: (46.0:230.0).

weight:
  Numerical >=0 format: All 398 values are numerical and greater or equal to 0 in the range (1613.0:5140.0).

acceleration:
  Numerical format: All 398 values are numerical in the range (8.0:24.8).

model year (year):
  Year format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 398 Value(s) outside range [1800, 2100] at index(es): [(0, 70), (1, 70), (2, 70), (3, 70), (4, 70), (5, 70), (6, 70), (7, 70), (8,
70), (9, 70), ('...', '...'), (388, 82), (389, 82), (390, 82), (391, 82), (392, 82), (393, 82), (394, 82), (395, 82), (396, 82), (397,
82)] (displaying only the first and last 10 items)

Actual range of values: (70 : 82)

origin:
  All 398 values are correctly categorical.

Categorical format with 3 unique values:
 Category  Frequency
        1        249
        3         79
        2         70

car name (name):
  Name format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 116 Extraneous data value(s) at index(es): [(1, 'buick skylark 320'), (5, 'ford galaxie 500'), (11, "plymouth 'cuda 340"), (18,
'datsun pl510'), (19, 'volkswagen 1131 deluxe sedan'), (20, 'peugeot 504'), (21, 'audi 100 ls'), (22, 'saab 99e'), (23, 'bmw 2002'),
(25, 'ford f250'), ('...', '...'), (354, 'renault 18i'), (357, 'datsun 200sx'), (358, 'mazda 626'), (359, 'peugeot 505s turbo diesel'),
(362, 'datsun 810 maxima'), (369, 'chevrolet cavalier 2-door'), (370, 'pontiac j2000 se hatchback'), (385, 'datsun 310 gx'), (391,
'dodge charger 2.2'), (397, 'chevy s-10')] (displaying only the first and last 10 items)

Frequency Distribution:
            car name Frequency
          ford pinto         6
         amc matador         5
       ford maverick         5
      toyota corolla         5
         amc gremlin         4
          amc hornet         4
  chevrolet chevette         4
    chevrolet impala         4
         peugeot 504         4
       toyota corona         4
                 ...       ...
         volvo 144ea         1
     volvo 145e (sw)         1
         volvo 244dl         1
           volvo 245         1
         volvo 264gl         1
        volvo diesel         1
  vw dasher (diesel)         1
```

```
          vw pickup         1
vw rabbit c (diesel)        1
    vw rabbit custom        1

Range of Values: (amc ambassador brougham to vw rabbit custom)
```

31- 40 - Flags - N/A

Dataset analysed: flag.data from https://archive.ics.uci.edu/ml/machine-learning-databases/flags/flag.data (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

```
name:
  Name format: All 194 name values are valid.

Frequency Distribution:
           name Frequency
    Afghanistan         1
        Albania         1
        Algeria         1
 American-Samoa         1
        Andorra         1
         Angola         1
       Anguilla         1
Antigua-Barbuda         1
      Argentina         1
      Argentine         1
            ...       ...
        Uruguay         1
        Vanuatu         1
   Vatican-City         1
      Venezuela         1
        Vietnam         1
  Western-Samoa         1
     Yugoslavia         1
          Zaire         1
         Zambia         1
       Zimbabwe         1

Range of Values: (Afghanistan to Zimbabwe)

landmass (1=):
  All 194 values are correctly categorical.

Categorical format with 6 unique values:
 Category  Frequency
        4         52
        5         39
        3         35
        1         31
        6         20
        2         17

zone (1=):
  All 194 values are correctly categorical.

Categorical format with 4 unique values:
 Category  Frequency
        1         91
        4         58
        2         29
        3         16

area:
  Numerical >=0 format: All 194 values are numerical and greater or equal to 0 in the range (0:22402).

population:
  Numerical >=0 format: All 194 values are numerical and greater or equal to 0 in the range (0:1008).

language:
  All 194 values are correctly categorical.

Categorical format with 10 unique values:
 Category  Frequency
       10         46
        1         43
        6         30
        2         21
        8         19
        3         17
        4          6
        5          4
        7          4
        9          4

religion:
  All 194 values are correctly categorical.

Categorical format with 8 unique values:
 Category  Frequency
```

```
        1         60
        0         40
        2         36
        5         27
        6         15
        3          8
        4          4
        7          4

bars (number):
  Numerical format: All 194 values are numerical in the range (0:5).

stripes (number):
  Numerical format: All 194 values are numerical in the range (0:14).

colours (number):
  Numerical format: All 194 values are numerical in the range (1:8).

red (0 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        1        153
        0         41

green (0 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0        103
        1         91

blue (0 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        1         99
        0         95

gold (0 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0        103
        1         91

white (0 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        1        146
        0         48

black (0 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0        142
        1         52

orange (0 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0        168
        1         26

mainhue (colour):
  All 194 values are correctly categorical.

Categorical format with 8 unique values:
Category  Frequency
     red         71
    blue         40
   green         31
   white         22
    gold         19
   black          5
  orange          4
   brown          2

circles (circle):
  Numerical >=0 format: All 194 values are numerical and greater or equal to 0 in the range (0:4).

crosses (number):
  Numerical format: All 194 values are numerical in the range (0:2).
```

saltires (number):
  Numerical format: All 194 values are numerical in the range (0:1).

quarters (number):
  Numerical format: All 194 values are numerical in the range (0:4).

sunstars (number):
  Numerical format: All 194 values are numerical in the range (0:50).

crescent (1 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category Frequency
        0   183
        1   11

triangle (1 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category Frequency
        0   167
        1   27

icon (1 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category Frequency
        0   145
        1   49

animate (animation):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category Frequency
        0   155
        1   39

text (1 if):
  All 194 values are correctly categorical.

Categorical format with 2 unique values:
 Category Frequency
        0   178
        1   16

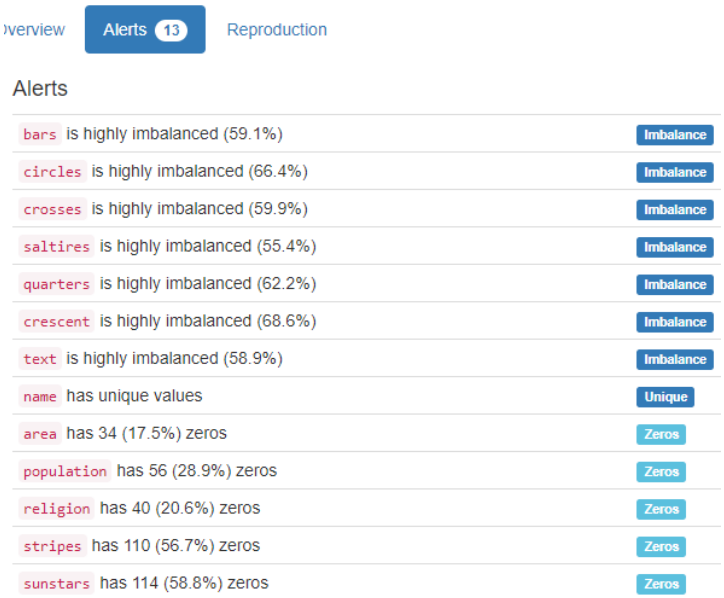
topleft (colour):
  All 194 values are correctly categorical.

Categorical format with 7 unique values:
Category Frequency
     red   56
    blue   43
   white   41
   green   32
   black   12
    gold   6
  orange   4

botright (colour):
  All 194 values are correctly categorical.

Categorical format with 8 unique values:
Category Frequency
     red   69
    blue   47
   green   40
   white   17
   black   9
    gold   9
   brown   2
  orange   1

Last run on: 2024-02-26 16:59:40



YData did not find anything that we care.

32- 336 - Chronic_Kidney_Disease - N/A

Dataset analysed: @DATA part extracted from chronic_kidney_disease_full.arff that was obtained at Chronic_Kidney_Disease.rar from https://archive.ics.uci.edu/ml/machine-learning-databases/00336/Chronic_Kidney_Disease.rar (dataset_file_url column from Fortydatasets.xlsx) It does not have header.

YData Profiling just brought High Correlation and Imbalance alerts, no '?' information. We found dozens of cases with '?'.

```
Age:
  Age format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 9 Non-numeric value(s) at index(es): [(30, '?'), (73, '?'), (112, '?'), (116, '?'), (117, '?'), (169, '?'), (191, '?'), (203, '?'),
(268, '?')]

Actual range of values: (2.0 : 90.0)

Blood Pressure (blood):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 12 Non-numeric value(s) at index(es): [(7, '?'), (75, '?'), (132, '?'), (138, '?'), (161, '?'), (164, '?'), (185, '?'), (187, '?'),
(188, '?'), (215, '?'), (293, '?'), (316, '?')]

Range of values: (50.0:180.0).

Specific Gravity (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 47 Unacceptable value(s) at index(es): [(13, '?'), (17, '?'), (21, '?'), (28, '?'), (30, '?'), (37, '?'), (50, '?'), (57, '?'), (59,
'?'), (78, '?'), ('...', '...'), (228, '?'), (231, '?'), (236, '?'), (238, '?'), (245, '?'), (268, '?'), (280, '?'), (295, '?'), (322,
'?'), (346, '?')] (displaying only the first and last 10 items)
Categorical format with 6 unique values:
Category  Frequency
   1.020        106
   1.010         84
   1.025         81
   1.015         75
       ?         47
   1.005          7

Albumin (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 46 Unacceptable value(s) at index(es): [(13, '?'), (17, '?'), (21, '?'), (30, '?'), (37, '?'), (50, '?'), (57, '?'), (59, '?'), (78,
'?'), (81, '?'), ('...', '...'), (228, '?'), (231, '?'), (236, '?'), (238, '?'), (245, '?'), (268, '?'), (280, '?'), (295, '?'), (322,
'?'), (346, '?')] (displaying only the first and last 10 items)
Categorical format with 7 unique values:
Category  Frequency
       0        199
       ?         46
       1         44
       2         43
       3         43
       4         24
       5          1

Sugar (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
```

```
 49 Unacceptable value(s) at index(es): [(13, '?'), (17, '?'), (21, '?'), (30, '?'), (37, '?'), (50, '?'), (57, '?'), (59, '?'), (78,
'?'), (81, '?'), ('...', '...'), (228, '?'), (231, '?'), (236, '?'), (238, '?'), (245, '?'), (268, '?'), (280, '?'), (295, '?'), (322,
'?'), (346, '?')] (displaying only the first and last 10 items)
Categorical format with 7 unique values:
Category  Frequency
       0        290
       ?         49
       2         18
       3         14
       1         13
       4         13
       5          3

Red Blood Cells (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 152 Unacceptable value(s) at index(es): [(0, '?'), (1, '?'), (5, '?'), (6, '?'), (10, '?'), (12, '?'), (13, '?'), (15, '?'), (16,
'?'), (17, '?'), ('...', '...'), (245, '?'), (268, '?'), (280, '?'), (290, '?'), (295, '?'), (309, '?'), (322, '?'), (349, '?'), (350,
'?'), (381, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
  normal        201
       ?        152
abnormal         47

Pus Cell (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 65 Unacceptable value(s) at index(es): [(5, '?'), (13, '?'), (17, '?'), (21, '?'), (28, '?'), (30, '?'), (34, '?'), (37, '?'), (39,
'?'), (43, '?'), ('...', '...'), (245, '?'), (268, '?'), (280, '?'), (290, '?'), (295, '?'), (309, '?'), (322, '?'), (349, '?'), (350,
'?'), (381, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
  normal        259
abnormal         76
       ?         65

Pus Cell clumps (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 4 Unacceptable value(s) at index(es): [(290, '?'), (300, '?'), (316, '?'), (328, '?')]
Categorical format with 3 unique values:
  Category  Frequency
notpresent        354
   present         42
         ?          4

Bacteria (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 4 Unacceptable value(s) at index(es): [(290, '?'), (300, '?'), (316, '?'), (328, '?')]
Categorical format with 3 unique values:
  Category  Frequency
notpresent        374
   present         22
         ?          4

Blood Glucose Random (blood):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 44 Non-numeric value(s) at index(es): [(1, '?'), (21, '?'), (23, '?'), (24, '?'), (29, '?'), (38, '?'), (41, '?'), (47, '?'), (52,
'?'), (54, '?'), ('...', '...'), (209, '?'), (215, '?'), (232, '?'), (234, '?'), (276, '?'), (283, '?'), (312, '?'), (315, '?'), (332,
'?'), (378, '?')] (displaying only the first and last 10 items)

Range of values: (22.0:490.0).

Blood Urea (blood):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 19 Non-numeric value(s) at index(es): [(23, '?'), (54, '?'), (55, '?'), (64, '?'), (67, '?'), (113, '?'), (134, '?'), (161, '?'),
(165, '?'), (209, '?'), (215, '?'), (220, '?'), (232, '?'), (276, '?'), (283, '?'), (312, '?'), (315, '?'), (334, '?'), (378, '?')]

Range of values: (1.5:391.0).

Serum Creatinine (creatinine):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 17 Non-numeric value(s) at index(es): [(23, '?'), (55, '?'), (64, '?'), (67, '?'), (113, '?'), (161, '?'), (165, '?'), (209, '?'),
(215, '?'), (216, '?'), (220, '?'), (232, '?'), (276, '?'), (283, '?'), (312, '?'), (315, '?'), (334, '?')]

Range of values: (0.4:76.0).

Sodium:
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 87 Non-numeric value(s) at index(es): [(0, '?'), (1, '?'), (2, '?'), (4, '?'), (7, '?'), (8, '?'), (10, '?'), (19, '?'), (23, '?'),
(28, '?'), ('...', '...'), (232, '?'), (234, '?'), (235, '?'), (237, '?'), (240, '?'), (283, '?'), (303, '?'), (315, '?'), (336, '?'),
(363, '?')] (displaying only the first and last 10 items)

Range of values: (4.5:163.0).

Potassium:
  Numerical format: Error(s) found:
```

```
DQI #17 (Wrong Data Type - Consistency):
 88 Non-numeric value(s) at index(es): [(0, '?'), (1, '?'), (2, '?'), (4, '?'), (7, '?'), (8, '?'), (10, '?'), (19, '?'), (21, '?'),
(23, '?'), ('...', '...'), (232, '?'), (234, '?'), (235, '?'), (237, '?'), (240, '?'), (283, '?'), (303, '?'), (315, '?'), (336, '?'),
(363, '?')] (displaying only the first and last 10 items)

Range of values: (2.5:47.0).

Hemoglobin:
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 52 Non-numeric value(s) at index(es): [(23, '?'), (28, '?'), (30, '?'), (34, '?'), (41, '?'), (57, '?'), (60, '?'), (61, '?'), (66,
'?'), (67, '?'), ('...', '...'), (230, '?'), (232, '?'), (233, '?'), (247, '?'), (273, '?'), (319, '?'), (324, '?'), (328, '?'), (330,
'?'), (365, '?')] (displaying only the first and last 10 items)

Range of values: (3.1:17.8).

Packed  Cell Volume (volume):
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 71 Non-numeric value(s) at index(es): [(13, '?'), (16, '?'), (17, '?'), (23, '?'), (28, '?'), (30, '?'), (34, '?'), (38, '?'), (41,
'?'), (45, '?'), ('...', '...'), (224, '?'), (228, '?'), (230, '?'), (232, '?'), (233, '?'), (247, '?'), (273, '?'), (319, '?'), (324,
'?'), (365, '?')] (displaying only the first and last 10 items)

Range of values: (9.0:54.0).

White Blood Cell Count (count):
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 106 Non-numeric value(s) at index(es): [(6, '?'), (10, '?'), (13, '?'), (16, '?'), (17, '?'), (23, '?'), (28, '?'), (29, '?'), (30,
'?'), (33, '?'), ('...', '...'), (238, '?'), (239, '?'), (247, '?'), (273, '?'), (274, '?'), (287, '?'), (302, '?'), (319, '?'), (324,
'?'), (330, '?')] (displaying only the first and last 10 items)

Range of values: (2200.0:26400.0).

Red Blood Cell Count (count):
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 131 Non-numeric value(s) at index(es): [(1, '?'), (2, '?'), (6, '?'), (10, '?'), (13, '?'), (16, '?'), (17, '?'), (23, '?'), (28,
'?'), (29, '?'), ('...', '...'), (238, '?'), (239, '?'), (247, '?'), (273, '?'), (274, '?'), (287, '?'), (302, '?'), (319, '?'), (324,
'?'), (330, '?')] (displaying only the first and last 10 items)

Range of values: (2.1:8.0).

Hypertension (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 2 Unacceptable value(s) at index(es): [(288, '?'), (297, '?')]
Categorical format with 3 unique values:
Category  Frequency
      no        251
     yes        147
       ?          2

Diabetes Mellitus (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 2 Unacceptable value(s) at index(es): [(288, '?'), (297, '?')]
Categorical format with 5 unique values:
Category  Frequency
      no        258
     yes        134
      no          3
     yes          3
       ?          2

Coronary Artery Disease (disease):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 2 Unacceptable value(s) at index(es): [(288, '?'), (297, '?')]
Categorical format with 4 unique values:
Category  Frequency
      no        362
     yes         34
      no          2
       ?          2

Appetite:
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 1 Unacceptable value(s) at index(es): [(294, '?')]
Categorical format with 3 unique values:
Category  Frequency
    good        317
    poor         82
       ?          1

Pedal Edema (nominal):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 1 Unacceptable value(s) at index(es): [(294, '?')]
Categorical format with 3 unique values:
Category  Frequency
      no        323
```

```
     yes         76
       ?          1

Anemia:
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 1 Unacceptable value(s) at index(es): [(294, '?')]
Categorical format with 3 unique values:
Category  Frequency
      no        339
     yes         60
       ?          1

Class:
  All 400 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
     ckd        248
  notckd        150
    ckd           2

Last run on: 2024-02-26 20:59:59
```

33- 50 - Image Segmentation - N/A

Dataset analysed: segmentation.data from https://archive.ics.uci.edu/ml/machine-learning-databases/image/segmentation.data (dataset_file_url column from Fortydatasets.xlsx) It has not only a header but also some comments before the header. Therefore it was necessary to copy the file to local directory to extract extraneous lines.

```
;;; -*- Mode:Common-Lisp; Base:10 -*-
;;; *-* Last-edit: 11/21/90 15:28:30 by Brodley; *-*

REGION-CENTROID-COL,REGION-CENTROID-ROW,REGION-PIXEL-COUNT,SHORT-LINE-DENSITY-5,SHORT-LINE-DENSITY-2,VEDGE-MEAN,VEDGE-SD,HEDGE-MEAN,HEDGE-SD,INTENSITY-MEAN,RAWRED-MEAN,RAWBLUE-
MEAN,RAWGREEN-MEAN,EXRED-MEAN,EXBLUE-MEAN,EXGREEN-MEAN,VALUE-MEAN,SATURATION-MEAN,HUE-MEAN

BRICKFACE,140.0,125.0,9,0.0,0.0,0.2777779,0.06296301,0.66666675,0.31111118,6.185185,7.3333335,7.6666665,3.5555556,3.4444444,4.4444447,-7.888889,7.7777777,0.5456349,-1.1218182
BRICKFACE,188.0,133.0,9,0.0,0.0,0.33333334,0.26666674,0.5,0.077777736,6.6666665,8.333334,7.7777777,3.8888888,5.0,3.3333333,-8.333333,8.444445,0.53858024,-0.92481726
BRICKFACE,105.0,139.0,9,0.0,0.0,0.27777782,0.107407436,0.83333325,0.52222216,6.111111,7.5555553,7.2222223,3.5555556,4.3333335,3.3333333,-7.6666665,7.5555553,0.5326279,-0.96594584

class:
  All 210 values are correctly categorical.

Categorical format with 7 unique values:
 Category  Frequency
BRICKFACE         30
   CEMENT         30
  FOLIAGE         30
    GRASS         30
     PATH         30
      SKY         30
   WINDOW         30

region-centroid-col (column):
  Numerical format: All 210 values are numerical in the range (1.0:252.0).

region-centroid-row (row):
  Numerical format: All 210 values are numerical in the range (11.0:250.0).

region-pixel-count (count):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (9:9).

short-line-density-5 (density):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (0.0:0.11111111).

short-line-density-2 (density):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (0.0:0.22222222).

vedge-mean (mean):
  Numerical format: All 210 values are numerical in the range (0.0:25.5).

vegde-sd (standard deviation):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (0.0:572.9964).

hedge-mean (mean):
  Numerical format: All 210 values are numerical in the range (0.0:44.722225).

hedge-sd (standard deviation):
  Numerical >=0 format: All 210 values are numerical and greater or equal to 0 in the range (0.0:1386.3292).

intensity-mean (intensity):
  Numerical format: All 210 values are numerical in the range (0.0:143.44444).

rawred-mean (mean):
  Numerical format: All 210 values are numerical in the range (0.0:136.88889).

rawblue-mean (mean):
  Numerical format: All 210 values are numerical in the range (0.0:150.88889).

rawgreen-mean (mean):
```

```
  Numerical format: All 210 values are numerical in the range (0.0:142.55556).

exred-mean (mean):
  Numerical format: All 210 values are numerical in the range (-48.22222:5.7777777).

exblue-mean (mean):
  Numerical format: All 210 values are numerical in the range (-9.666667:78.77778).

exgreen-mean (mean):
  Numerical format: All 210 values are numerical in the range (-30.555555:21.88889).

value-mean (mean):
  Numerical format: All 210 values are numerical in the range (0.0:150.88889).

saturatoin-mean (mean):
  Numerical format: All 210 values are numerical in the range (0.0:1.0).

hue-mean (mean):
  Numerical format: All 210 values are numerical in the range (-2.5309503:2.8649306).

Last run on: 2024-02-26 22:03:00
```

verview | Alerts 18 | Reproduction

Alerts

region-pixel-count has constant value ""	Constant
short-line-density-5 is highly imbalanced (61.1%)	Imbalance
short-line-density-2 is highly imbalanced (80.4%)	Imbalance
class is uniformly distributed	Uniform
vedge-mean has 5 (2.4%) zeros	Zeros
vegde-sd has 5 (2.4%) zeros	Zeros
hedge-mean has 5 (2.4%) zeros	Zeros
hedge-sd has 5 (2.4%) zeros	Zeros
intensity-mean has 5 (2.4%) zeros	Zeros
rawred-mean has 10 (4.8%) zeros	Zeros
rawblue-mean has 5 (2.4%) zeros	Zeros
rawgreen-mean has 9 (4.3%) zeros	Zeros
exred-mean has 5 (2.4%) zeros	Zeros
exblue-mean has 5 (2.4%) zeros	Zeros
exgreen-mean has 5 (2.4%) zeros	Zeros
value-mean has 5 (2.4%) zeros	Zeros
saturatoin-mean has 5 (2.4%) zeros	Zeros
hue-mean has 5 (2.4%) zeros	Zeros

YData shows column region-pixel-count with constant value "". We have not checked that. For us it is a numerical with values in the range (9:9).

region-pixel-count
Categorical

CONSTANT

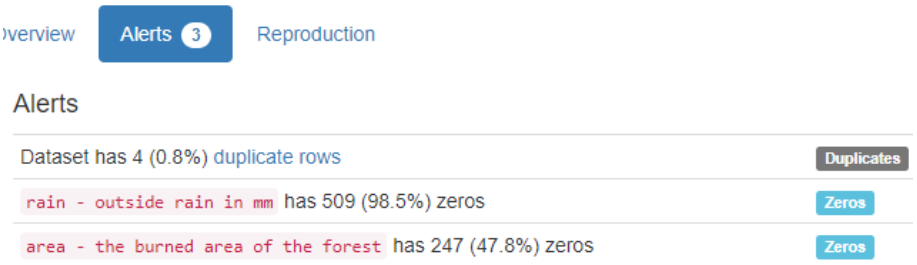
Distinct	1	9	210

34- 162 - Forest Fires – Physical

Dataset analysed: forestfiles.csv from https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv (dataset_file_url column from Fortydatasets.xlsx) has header

```
X - x-axis spatial coordinate within the Montesinho park map (x):
  Numerical format: All 517 values are numerical in the range (1:9).

Y - y-axis spatial coordinate within the Montesinho park map (y):
  Numerical format: All 517 values are numerical in the range (2:9).

month - month of the year (month):
  Month format: All 517 month values are valid.

Frequency Distribution:
    Month  Frequency
```

```
   August        184
September        172
    March         54
     July         32
 February         20
     June         17
  October         15
    April          9
 December          9
  January          2
      May          2
 November          1

day - day of the week (day of the week):
  Weekday format: All 517 weekday values are valid.

Frequency Distribution:
  Weekday  Frequency
   Sunday         95
   Friday         85
 Saturday         84
   Monday         74
  Tuesday         64
 Thursday         61
Wednesday         54

FFMC - FFMC index from the FWI system (index):
  Numerical format: All 517 values are numerical in the range (18.7:96.2).

DMC - DMC index from the FWI system (index):
  Numerical format: All 517 values are numerical in the range (1.1:291.3).

DC - DC index from the FWI system (index):
  Numerical format: All 517 values are numerical in the range (7.9:860.6).

ISI - ISI index from the FWI system (index):
  Numerical format: All 517 values are numerical in the range (0.0:56.1).

temp - temperature in Celsius degrees (temperature):
  Numerical format: All 517 values are numerical in the range (2.2:33.3).

RH - relative humidity in % (humidity):
  Numerical >=0 format: All 517 values are numerical and greater or equal to 0 in the range (15:100).

wind - wind speed in km (speed):
  Numerical format: All 517 values are numerical in the range (0.4:9.4).

rain - outside rain in mm (rain):
  Numerical >=0 format: All 517 values are numerical and greater or equal to 0 in the range (0.0:6.4).

area - the burned area of the forest (area):
  Numerical >=0 format: All 517 values are numerical and greater or equal to 0 in the range (0.0:1090.84).
```



YData did not find anything that we care.

35- 235 - Individual household electric power consumption – Physical

Dataset analysed: household_power_consumption.txt from https://archive.ics.uci.edu/ml/machine-learning-databases/00235/household_power_consumption.zip (dataset_file_url column from Fortydatasets.xlsx) has header and over 2 million lines !

```
Start of Analysis on: 2024-02-29 14:08:46
date:
  Date format: All 2075259 date values are valid in the DDMMYYYY format in the range 16/12/2006 to 26/11/2010.

time:
  Time format: All 2075259 time values are valid in the range 00:00:00 to 23:59:00
```

Overview | Alerts 7 | Reproduction

Alerts

`global_active_power` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`global_reactive_power` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`voltage` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`global_intensity` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`sub_metering_1` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`sub_metering_2` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`sub_metering_3` is an unsupported type, check if it needs cleaning or further analysis	Unsupported

```
global_active_power (power):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 25979 Non-numeric value(s) at index(es): [(6839, '?'), (6840, '?'), (19724, '?'), (19725, '?'), (41832, '?'), (61909, '?'), (98254,
'?'), (98255, '?'), (142588, '?'), (190497, '?'), ('...', '...'), (1990180, '?'), (1990181, '?'), (1990182, '?'), (1990183, '?'),
(1990184, '?'), (1990185, '?'), (1990186, '?'), (1990187, '?'), (1990188, '?'), (2027411, '?')] (displaying only the first and last 10
items)

Range of values: (0.076:11.122).

global_reactive_power (power):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 25979 Non-numeric value(s) at index(es): [(6839, '?'), (6840, '?'), (19724, '?'), (19725, '?'), (41832, '?'), (61909, '?'), (98254,
'?'), (98255, '?'), (142588, '?'), (190497, '?'), ('...', '...'), (1990180, '?'), (1990181, '?'), (1990182, '?'), (1990183, '?'),
(1990184, '?'), (1990185, '?'), (1990186, '?'), (1990187, '?'), (1990188, '?'), (2027411, '?')] (displaying only the first and last 10
items)

Range of values: (0.0:1.39).

voltage:
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 25979 Non-numeric value(s) at index(es): [(6839, '?'), (6840, '?'), (19724, '?'), (19725, '?'), (41832, '?'), (61909, '?'), (98254,
'?'), (98255, '?'), (142588, '?'), (190497, '?'), ('...', '...'), (1990180, '?'), (1990181, '?'), (1990182, '?'), (1990183, '?'),
(1990184, '?'), (1990185, '?'), (1990186, '?'), (1990187, '?'), (1990188, '?'), (2027411, '?')] (displaying only the first and last 10
items)

Range of values: (223.2:254.15).

global_intensity (intensity):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 25979 Non-numeric value(s) at index(es): [(6839, '?'), (6840, '?'), (19724, '?'), (19725, '?'), (41832, '?'), (61909, '?'), (98254,
'?'), (98255, '?'), (142588, '?'), (190497, '?'), ('...', '...'), (1990180, '?'), (1990181, '?'), (1990182, '?'), (1990183, '?'),
(1990184, '?'), (1990185, '?'), (1990186, '?'), (1990187, '?'), (1990188, '?'), (2027411, '?')] (displaying only the first and last 10
items)

Range of values: (0.2:48.4).

sub_metering_1 (metering):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 25979 Non-numeric value(s) at index(es): [(6839, '?'), (6840, '?'), (19724, '?'), (19725, '?'), (41832, '?'), (61909, '?'), (98254,
'?'), (98255, '?'), (142588, '?'), (190497, '?'), ('...', '...'), (1990180, '?'), (1990181, '?'), (1990182, '?'), (1990183, '?'),
(1990184, '?'), (1990185, '?'), (1990186, '?'), (1990187, '?'), (1990188, '?'), (2027411, '?')] (displaying only the first and last 10
items)

Range of values: (0.0:88.0).

sub_metering_2 (metering):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 25979 Non-numeric value(s) at index(es): [(6839, '?'), (6840, '?'), (19724, '?'), (19725, '?'), (41832, '?'), (61909, '?'), (98254,
'?'), (98255, '?'), (142588, '?'), (190497, '?'), ('...', '...'), (1990180, '?'), (1990181, '?'), (1990182, '?'), (1990183, '?'),
(1990184, '?'), (1990185, '?'), (1990186, '?'), (1990187, '?'), (1990188, '?'), (2027411, '?')] (displaying only the first and last 10
items)

Range of values: (0.0:80.0).

sub_metering_3 (metering):
  Numerical format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 25979 Blank/Empty/Null/NaN value(s) at index(es): [(6839, ''), (6840, ''), (19724, ''), (19725, ''), (41832, ''), (61909, ''), (98254,
''), (98255, ''), (142588, ''), (190497, ''), ('...', '...'), (1990180, ''), (1990181, ''), (1990182, ''), (1990183, ''), (1990184,
''), (1990185, ''), (1990186, ''), (1990187, ''), (1990188, ''), (2027411, '')] (displaying only the first and last 10 items)


Range of values: (0.0:31.0).

Last run on: 2024-02-29 14:16:20 - 7 MINUTES AND 34 SECONDS!
```

36- 165 - Concrete Compressive Strength – Physical

Dataset analysed: Concrete_Data.xls from https://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/Concrete_Data.xls (dataset_file_url column from Fortydatasets.xlsx) has header

```
Cement:
  Numerical >=0 format: All 1030 values are numerical and greater or equal to 0 in the range (102.0:540.0).

Blast Furnace Slag (slag):
  Numerical format: All 1030 values are numerical in the range (0.0:359.4).

Fly Ash (ash):
  Numerical >=0 format: All 1030 values are numerical and greater or equal to 0 in the range (0.0:200.1).

Water:
  Numerical format: All 1030 values are numerical in the range (121.75:247.0).

Superplasticizer (quantitative):
  Numerical format: All 1030 values are numerical in the range (0.0:32.2).

Coarse Aggregate (quantitative):
  Numerical format: All 1030 values are numerical in the range (801.0:1145.0).

Fine Aggregate (quantitative):
  Numerical format: All 1030 values are numerical in the range (594.0:992.6).

Age -- quantitative  -- Day (day):
  Day format: All 1030 values are numerical and valid in the range [1, 366].
Actual range of values: (1 : 365)

Concrete compressive strength -- quantitative -- MPa -- Output Variable (quantitative):
  Numerical format: All 1030 values are numerical in the range (2.331807832:82.5992248).

Last run on: 2024-02-29 14:35:20
```

Overview | Alerts 8 | Reproduction

Alerts

Dataset has 11 (1.1%) duplicate rows	Duplicates
`Age -- quantitative -- Day` is highly overall correlated with `Concrete compressive strength -- quantitative -- MPa -- Output Variable`	High correlation
`Concrete compressive strength -- quantitative -- MPa -- Output Variable` is highly overall correlated with `Age -- quantitative -- Day`	High correlation
`Superplasticizer` is highly overall correlated with `Water`	High correlation
`Water` is highly overall correlated with `Superplasticizer`	High correlation
`Blast Furnace Slag` has 466 (45.2%) zeros	Zeros
`Fly Ash` has 566 (55.0%) zeros	Zeros
`Superplasticizer` has 379 (36.8%) zeros	Zeros

YData did not find anything that we care.

138 - Robot Execution Failures – archive.ics.uci.edu/ml/machine-learning-databases/robotfailure-mld/lp1.data
Physical had data set in different lines for the same instance of data. There is no way to read such dataset. See two instances below:

```
normal
          -1          -1          63          -3          -1           0
           0           0          62          -3          -1           0
          -1          -1          61          -3           0           0
          -1          -1          63          -2          -1           0
          -1          -1          63          -3          -1           0
          -1          -1          63          -3          -1           0
          -1          -1          63          -3           0           0
          -1          -1          63          -3          -1           0
          -1          -1          63          -3          -1           0
          -1          -1          61          -3           0           0
          -1          -1          61          -3           0           0
          -1          -1          64          -3          -1           0
          -1          -1          64          -3          -1           0
          -1          -1          60          -3           0           0
          -1           0          64          -2          -1           0


normal
          -1          -1          63          -2          -1           0
          -1          -1          63          -3          -1           0
          -1          -1          61          -3           0           0
```

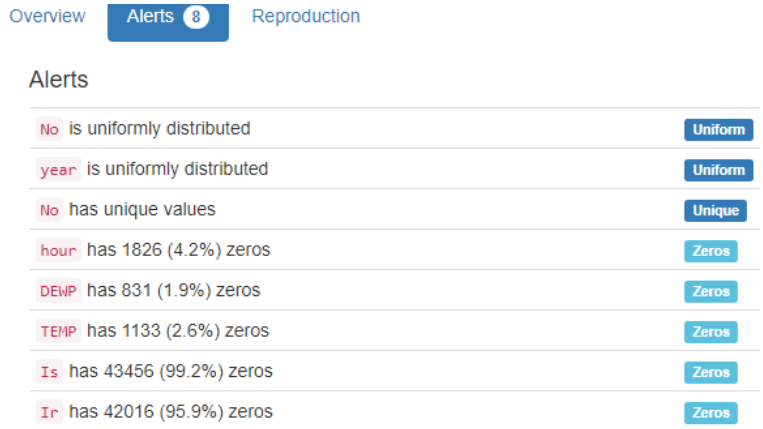
```
0    -4    63    1    0    0
0    -1    59    -2   0    -1
-3   3     57    -8   -3   -1
-1   3     70    -10  -2   -1
0    -3    61    0    0    0
0    -2    53    -1   -2   0
0    -3    66    1    4    0
-3   3     58    -10  -5   0
-1   -1    66    -4   -2   0
-1   -2    67    -3   -1   0
0    1     66    -6   -3   -1
-1   -1    59    -3   -4   0
```

52 – Ionosphere - Physical https://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/ionosphere.data

This dataset does not exhibit names/labels of columns/attributes.

37- 381 - Beijing PM2.5 Data - Physical

Dataset analysed: PRSA_data_2010.1.1-2014.12.31.csv from https://archive.ics.uci.edu/ml/machine-learning-databases/00381/PRSA_data_2010.1.1-2014.12.31.csv (dataset_file_url column from Fortydatasets.xlsx) has header



YData found that the column 'No' has unique values. We just found that it is a numerical format with nothing wrong. Besides that we found that the column pm2.5 should have only numerical values but we found over 2000 values with the content 'NA'. They also found that but did not consider an alert.

```
No (number):
  Numerical format: All 43824 values are numerical in the range (1:43824).

year:
  Year format: All 43824 values are numerical and valid in the range [1800, 2100].
Actual range of values: (2010 : 2014)

month:
  Month format: All 43824 month values are valid.

Frequency Distribution:
     Month  Frequency
   January       3720
     March       3720
       May       3720
      July       3720
    August       3720
   October       3720
  December       3720
     April       3600
      June       3600
 September       3600
  November       3600
  February       3384

day:
  Day format: All 43824 values are numerical and valid in the range [1, 366].
Actual range of values: (1 : 31)

hour:
  Numerical (between 0 and 24) format: All 43824 values are numerical and valid in the range [0, 24].
Actual range of values: (0 : 23)

pm2.5 (concentration):
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 2067 Non-numeric value(s) at index(es): [(0, 'NA'), (1, 'NA'), (2, 'NA'), (3, 'NA'), (4, 'NA'), (5, 'NA'), (6, 'NA'), (7, 'NA'), (8,
'NA'), (9, 'NA'), ('...', '...'), (43283, 'NA'), (43544, 'NA'), (43545, 'NA'), (43546, 'NA'), (43547, 'NA'), (43548, 'NA'), (43549,
'NA'), (43550, 'NA'), (43551, 'NA'), (43552, 'NA')] (displaying only the first and last 10 items)

Range of values: (0.0:994.0).
```

```
DEWP (dew point):
  Numerical format: All 43824 values are numerical in the range (-40:28).

TEMP (temperature):
  Numerical format: All 43824 values are numerical in the range (-19.0:42.0).

PRES (pressure):
  Numerical format: All 43824 values are numerical in the range (991.0:1046.0).

cbwd (direction):
  All 43824 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
      SE      15290
      NW      14150
      cv       9387
      NE       4997

Iws (speed):
  Numerical format: All 43824 values are numerical in the range (0.45:585.6).

Is (hours):
  Numerical >=0 format: All 43824 values are numerical and greater or equal to 0 in the range (0:27).

Ir (hours):
  Numerical >=0 format: All 43824 values are numerical and greater or equal to 0 in the range (0:36).

Last run on: 2024-02-29 15:32:41
```

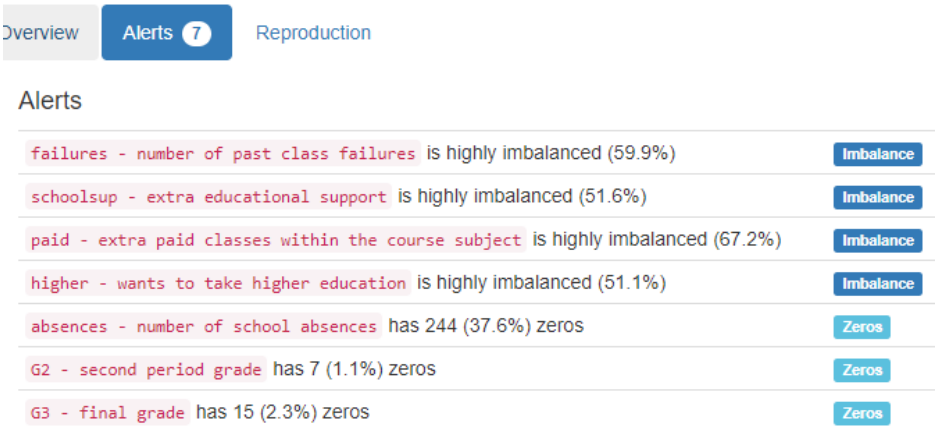
38- 320 - Student Performance - Social

Dataset analysed: student-por.csv from https://archive.ics.uci.edu/ml/machine-learning-databases/00320/.student.zip_old (dataset_file_url column from Fortydatasets.xlsx), but, as the original file ends with zip_old, the file had to be renamed to only .zip to extract the final dataset has header

```
school - student's school (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
   GP        423
   MS        226


sex - student's sex (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
    F        383
    M        266


age - student's age (age):
  Age format: All 649 values are numerical and valid in the range [0, 130].
Actual range of values: (15 : 22)

address - student's home address type (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
    U        452
    R        197


famsize - family size (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
  GT3        457
  LE3        192


Pstatus - parent's cohabitation status (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
    T        569
    A         80
```

```
Medu - mother's education (numeric):
  Numerical format: All 649 values are numerical in the range (0:4).

Fedu - father's education (numeric):
  Numerical format: All 649 values are numerical in the range (0:4).

Mjob - mother's job (job):
  All 649 values are correctly categorical.

Categorical format with 5 unique values:
Category  Frequency
   other        258
services        136
 at_home        135
 teacher         72
  health         48

Fjob - father's job (job):
  All 649 values are correctly categorical.

Categorical format with 5 unique values:
Category  Frequency
   other        367
services        181
 at_home         42
 teacher         36
  health         23

reason - reason to choose this school (reason):
  All 649 values are correctly categorical.

Categorical format with 4 unique values:
  Category  Frequency
    course        285
      home        149
reputation        143
     other         72

guardian - student's guardian (guardian):
  All 649 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
  mother        455
  father        153
   other         41

traveltime - home to school travel time (numeric):
  Numerical format: All 649 values are numerical in the range (1:4).

studytime - weekly study time (numeric):
  Numerical format: All 649 values are numerical in the range (1:4).

failures - number of past class failures (number):
  Numerical format: All 649 values are numerical in the range (0:3).

schoolsup - extra educational support (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
   no        581
  yes         68


famsup - family educational support (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
  yes        398
   no        251


paid - extra paid classes within the course subject (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
   no        610
  yes         39


activities - extra-curricular activities (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
   no        334
  yes        315
```

```
nursery - attended nursery school (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
  yes        521
   no        128


higher - wants to take higher education (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
  yes        580
   no         69


internet - Internet access at home (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
  yes        498
   no        151


romantic - with a romantic relationship (binary):
  Binary format: All 649 binary values are valid.

Frequency Distribution:
Value  Frequency
   no        410
  yes        239


famrel - quality of family relationships (numeric):
  Numerical format: All 649 values are numerical in the range (1:5).

freetime - free time after school (numeric):
  Numerical format: All 649 values are numerical in the range (1:5).

goout - going out with friends (numeric):
  Numerical format: All 649 values are numerical in the range (1:5).

Dalc - workday alcohol consumption (alcohol):
  Numerical format: All 649 values are numerical in the range (1:5).

Walc - weekend alcohol consumption (alcohol):
  Numerical format: All 649 values are numerical in the range (1:5).

health - current health status (numeric):
  Numerical format: All 649 values are numerical in the range (1:5).

absences - number of school absences (number):
  Numerical format: All 649 values are numerical in the range (0:32).

G1 - first period grade (grade):
  Numerical >=0 format: All 649 values are numerical and greater or equal to 0 in the range (0:19).

G2 - second period grade (grade):
  Numerical >=0 format: All 649 values are numerical and greater or equal to 0 in the range (0:19).

G3 - final grade (grade):
  Numerical >=0 format: All 649 values are numerical and greater or equal to 0 in the range (0:19).

Last run on: 2024-02-29 21:26:03
```



YData did not find anything that we care.

39- 275 - Bike Sharing Dataset - Social

Dataset analysed: hour.csv from https://archive.ics.uci.edu/ml/machine-learning-databases/00275/Bike-Sharing-Dataset.zip (dataset_file_url column from Fortydatasets.xlsx) has header

```
instant:
  Numerical >=0 format: All 17379 values are numerical and greater or equal to 0 in the range (1:17379).

dteday (date):
  Date format: All 17379 date values are valid in the YYYYMMDD format in the range 2011-01-01 to 2012-12-31.

season:
  All 17379 values are correctly categorical.

Categorical format with 4 unique values:
 Category  Frequency
        3       4496
        2       4409
        1       4242
        4       4232

The following column had to be altered in the AllColumnsfromFortyDatasets sheet because it does not contain appropriate year values, just the values 0 and 1. So, it was marked as nominal.
yr (nominal):
  All 17379 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        1       8734
        0       8645

mnth (month):
  Month format: All 17379 month values are valid.

Frequency Distribution:
    Month  Frequency
      May       1488
     July       1488
 December       1483
   August       1475
    March       1473
  October       1451
     June       1440
    April       1437
September       1437
 November       1437
  January       1429
 February       1341

hr (hour):
  Numerical (between 0 and 24) format: All 17379 values are numerical and valid in the range [0, 24].
Actual range of values: (0 : 23)

holiday ( or not):
  Binary format: All 17379 binary values are valid.

Frequency Distribution:
 Value  Frequency
     0      16879
     1        500


weekday:
  Weekday format: All 17379 weekday values are valid.

Frequency Distribution:
  Weekday  Frequency
   Friday       2512
 Thursday       2487
   Sunday       2479
  Tuesday       2475
Wednesday       2471
   Monday       2453

workingday (neither):
  Binary format: All 17379 binary values are valid.

Frequency Distribution:
 Value  Frequency
     1      11865
     0       5514


weathersit (weathersit):
  All 17379 values are correctly categorical.

Categorical format with 4 unique values:
 Category  Frequency
        1      11413
        2       4544
        3       1419
        4          3
```

```
temp (normalized):
  Normalized format: All 17379 values are numerical and valid in the range [0, 1].
Actual range of values: (0.02 : 1.0)

atemp (normalized):
  Normalized format: All 17379 values are numerical and valid in the range [0, 1].
Actual range of values: (0.0 : 1.0)

hum (normalized):
  Normalized format: All 17379 values are numerical and valid in the range [0, 1].
Actual range of values: (0.0 : 1.0)

windspeed (normalized):
  Normalized format: All 17379 values are numerical and valid in the range [0, 1].
Actual range of values: (0.0 : 0.8507)

casual (count):
  Numerical >=0 format: All 17379 values are numerical and greater or equal to 0 in the range (0:367).

registered (count):
  Numerical >=0 format: All 17379 values are numerical and greater or equal to 0 in the range (0:886).

cnt (count):
  Numerical >=0 format: All 17379 values are numerical and greater or equal to 0 in the range (1:977)

Last run on: 2024-03-01 12:49:47.
```

Overview | Alerts 19 | Reproduction

Alerts

atemp is highly overall correlated with casual and 2 other fields	High correlation
casual is highly overall correlated with atemp and 3 other fields	High correlation
cnt is highly overall correlated with casual and 2 other fields	High correlation
hr is highly overall correlated with cnt and 1 other fields	High correlation
instant is highly overall correlated with season and 1 other fields	High correlation
mnth is highly overall correlated with season	High correlation
registered is highly overall correlated with casual and 2 other fields	High correlation
season is highly overall correlated with atemp and 2 other fields	High correlation
temp is highly overall correlated with atemp and 1 other fields	High correlation
weekday is highly overall correlated with workingday	High correlation
workingday is highly overall correlated with weekday	High correlation
yr is highly overall correlated with instant	High correlation
holiday is highly imbalanced (81.2%)	Imbalance
instant is uniformly distributed	Uniform
instant has unique values	Unique
hr has 726 (4.2%) zeros	Zeros
weekday has 2502 (14.4%) zeros	Zeros
windspeed has 2180 (12.5%) zeros	Zeros
casual has 1581 (9.1%) zeros	Zeros

YData shows the column 'instant' as Unique. We have not checked that.

40- 13 - Balloons - Social

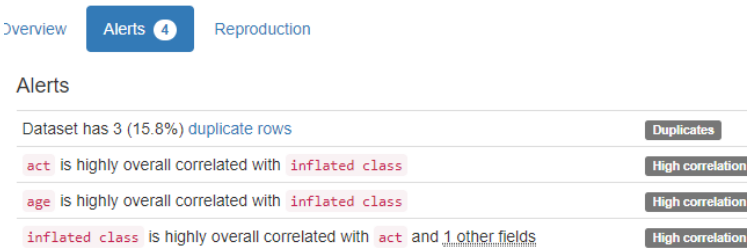
Dataset analysed: adult+stretch.data from https://archive.ics.uci.edu/ml/machine-learning-databases/balloons/adult+stretch.data (dataset_file_url column from Fortydatasets.xlsx) has

```
Color:
  All 19 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
  PURPLE         10
  YELLOW          9

size (large):
  All 19 values are correctly categorical.

Categorical format with 2 unique values:
```

```
Category  Frequency
   LARGE         10
   SMALL          9

act (stretch):
  All 19 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
 STRETCH         11
     DIP          8
```

The following column had to be altered in the AllColumnsfromFortyDatasets sheet because it does not contain appropriate age values, just the values ADULT and CHILD. So, it was marked as nominal.

```
age (nominal):
  All 19 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
   ADULT         11
   CHILD          8

inflated class (class):
  All 19 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       F         12
       T          7

Last run on: 2024-03-01 14:29:19
```



YData did not find anything that we care.