Datasets

index	name
52	Iris
45	Heart Disease
2	Adult
107	Wine
42	Glass Identification
58	Letter Recognition
142	Statlog (German Credit Data)
92	Spambase
103	Congressional Voting Records
27	Credit Approval

52 Iris

```
INFO:root:Successfully assigned column names to the dataset 'Iris' for index 52
```

	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

```
sepal length in cm (length):
  Numerical >=0 format: All 150 values are numerical and greater or equal to 0 in the range (4.3:7.9).

sepal width in cm (width):
  Numerical >=0 format: All 150 values are numerical and greater or equal to 0 in the range (2.0:4.4).

petal length in cm (length):
  Numerical >=0 format: All 150 values are numerical and greater or equal to 0 in the range (1.0:6.9).

petal width in cm (width):
  Numerical >=0 format: All 150 values are numerical and greater or equal to 0 in the range (0.1:2.5).

class:
  All 150 values are correctly categorical.

Categorical format with 3 unique values:
       Category  Frequency
    Iris-setosa         50
Iris-versicolor         50
 Iris-virginica         50
```

Overview Alerts 2 Reproduction

Alerts

Dataset has 1 (0.7%) duplicate rows Duplicates

`class` is uniformly distributed Uniform

```
Ydata did not find anything that we measure.
```

```
CHANGED dataset:
-5.1,3.5,1.4,0.2,Iris-setosa
abc,3.0,1.4,0.2,Iris-setosa
4.7,,1.3,-0.2,Iris-setosa
4.6,3.1,-1.5,0.2,Iris-setosa
5.0,3.6,null,'',Iris-setosa
```

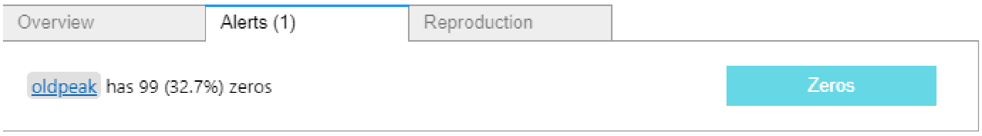
	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	class
0	-5.1	3.5	1.4	0.2	Iris-setosa
1	abc	3.0	1.4	0.2	Iris-setosa
2	4.7		1.3	-0.2	Iris-setosa
3	4.6	3.1	-1.5	0.2	Iris-setosa
4	5.0	3.6	null	''	Iris-setosa

```
sepal length in cm (length):
  Numerical >=0 format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 1 Negative value(s) at index(es): [(0, '-5.1')]

DQI #17 (Wrong Data Type - Consistency):
 1 Non-numeric value(s) at index(es): [(1, 'abc')]

Range of values: (-5.1:7.9).

sepal width in cm (width):
  Numerical >=0 format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(2, '')]

Range of values: (2.0:4.4).

petal length in cm (length):
  Numerical >=0 format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(4, 'null')]

DQI #15 (Domain Violation - Accuracy):
 1 Negative value(s) at index(es): [(3, '-1.5')]

Range of values: (-1.5:6.9).

petal width in cm (width):
  Numerical >=0 format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(4, "''")]

DQI #15 (Domain Violation - Accuracy):
 1 Negative value(s) at index(es): [(2, '-0.2')]

Range of values: (-0.2:2.5).


class unchanged.
```

45 Heart Disease

```
age:
  Age format: All 303 values are numerical and valid in the range [0, 130].
Actual range of values: (29.0 : 77.0)

sex:
  All 303 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
      1.0        206
      0.0         97

cp (type):
  All 303 values are correctly categorical.

Categorical format with 4 unique values:
 Category  Frequency
      4.0        144
      3.0         86
      2.0         50
      1.0         23

trestbps (integer):
  Numerical format: All 303 values are numerical in the range (94.0:200.0).

chol (integer):
  Numerical format: All 303 values are numerical in the range (126.0:564.0).

fbs (categorical):
  All 303 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
      0.0        258
      1.0         45

restecg (categorical):
  All 303 values are correctly categorical.

Categorical format with 3 unique values:
 Category  Frequency
      0.0        151
      2.0        148
      1.0          4

thalach (rate):
  Numerical format: All 303 values are numerical in the range (71.0:202.0).

exang (yes):
  All 303 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
      0.0        204
      1.0         99

oldpeak (integer):
  Numerical format: All 303 values are numerical in the range (0.0:6.2).

slope (categorical):
  All 303 values are correctly categorical.

Categorical format with 3 unique values:
 Category  Frequency
      1.0        142
      2.0        140
      3.0         21

ca (categorical):
   Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 Unacceptable value(s) at index(es):  166, '?'), (192, '?'), (287, '?'), (302, '?')])
Categorical format with 5 unique values:
Category  Frequency
     0.0        176
     1.0         65
     2.0         38
     3.0         20
       ?          4

thal (categorical):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 Unacceptable value(s) at index(es): [(87, '?'), (266, '?')])
Categorical format with 4 unique values:
Category  Frequency
     3.0        166
     7.0        117
     6.0         18
       ?          2
```

```
num (predicted):
  All 303 values are correctly categorical.

  Categorical format with 5 unique values:
 Category  Frequency
        0        164
        1         55
        2         36
        3         35
        4         13
```



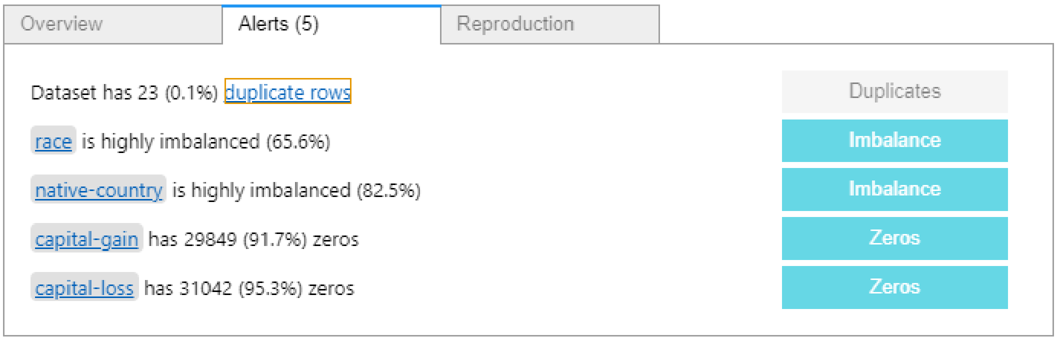
oldpeak has 99 (32.7%) zeros

Ydata Profiling did not find the '?'s in the two columns above where DQI# 4 was found.

2 Adult

Initial unchanged dataset:

```
age:
  Age format: All 32561 values are numerical and valid in the range [0, 130].
Actual range of values: (17 : 90)

workclass (class):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 1836 Unacceptable value(s) at index(es): Unacceptable value(s) at index(es): [(26, ' ?'), (60, ' ?'), (68, ' ?'), (76, ' ?'), (105, ' ?'), (127, ' ?'), (148, ' ?'), (153, ' ?'), (159, ' ?'), (186, ' ?'), ('...', '...'), (32425, ' ?'), (32476, ' ?'), (32489, ' ?'), (32493, ' ?'), (32524, ' ?'), (32529, ' ?'), (32530, ' ?'), (32538, ' ?'), (32540, ' ?'), (32541, ' ?')] (displaying only the first and last 10 items)
Categorical format with 9 unique values:

         Category  Frequency
          Private      22696
 Self-emp-not-inc       2541
        Local-gov       2093
                ?       1836
        State-gov       1298
     Self-emp-inc       1116
      Federal-gov        960
      Without-pay         14
     Never-worked          7

fnlwgt (weight):
  Numerical >=0 format: All 32561 values are numerical and greater or equal to 0 in the range (12285:1484705).

education:
  All 32561 values are correctly categorical.

  Categorical format with 16 unique values:
     Category  Frequency
      HS-grad      10501
 Some-college       7291
    Bachelors       5355
      Masters       1723
    Assoc-voc       1382
         11th       1175
   Assoc-acdm       1067
         10th        933
      7th-8th        646
  Prof-school        576
          9th        514
         12th        433
    Doctorate        413
      5th-6th        333
      1st-4th        168
    Preschool         51

education-num (num)
  Numerical format: All 32561 values are numerical in the range (1:16).

marital-status (status):
  All 32561 values are correctly categorical.

  Categorical format with 7 unique values:
              Category  Frequency
    Married-civ-spouse      14976
         Never-married      10683
              Divorced       4443
             Separated       1025
               Widowed        993
 Married-spouse-absent        418
     Married-AF-spouse         23

occupation:
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 1843 Unacceptable value(s) at index(es): [(26, ' ?'), (60, ' ?'), (68, ' ?'), (76, ' ?'), (105, ' ?'), (127, ' ?'), (148, ' ?'), (153,
' ?'), (159, ' ?'), (186, ' ?'), ('...', '...'), (32425, ' ?'), (32476, ' ?'), (32489, ' ?'), (32493, ' ?'), (32524, ' ?'), (32529, '
?'), (32530, ' ?'), (32538, ' ?'), (32540, ' ?'), (32541, ' ?')] (displaying only the first and last 10 items)

Categorical format with 15 unique values:
          Category  Frequency
    Prof-specialty       4140
      Craft-repair       4099
   Exec-managerial       4066
      Adm-clerical       3770
             Sales       3650
     Other-service       3295
 Machine-op-inspct       2002
                 ?       1843
  Transport-moving       1597
 Handlers-cleaners       1370
   Farming-fishing        994
      Tech-support        928
   Protective-serv        649
   Priv-house-serv        149
      Armed-Forces          9


relationship:
  All 32561 values are correctly categorical.

Categorical format with 6 unique values:
       Category  Frequency
        Husband      13193
  Not-in-family       8305
      Own-child       5068
      Unmarried       3446
           Wife       1568
 Other-relative        981

race:
  All 32561 values are correctly categorical.

Categorical format with 5 unique values:
           Category  Frequency
              White      27816
              Black       3124
 Asian-Pac-Islander       1039
 Amer-Indian-Eskimo        311
              Other        271
sex:
  All 32561 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
    Male      21790
  Female      10771

capital-gain (gain):
  Numerical format: All 32561 values are numerical in the range (0:99999).

capital-loss (loss):
  Numerical format: All 32561 values are numerical in the range (0:4356).

hours-per-week (hours):
  Numerical >=0 format: All 32561 values are numerical and greater or equal to 0 in the range (1:99).

native-country (country):
  Country format: Error(s) found:
DQI #5 (Extraneous Data - Consistency, Uniqueness):
 583 Extraneous data value(s) at index(es): [(14, ' ?'), (38, ' ?'), (51, ' ?'), (61, ' ?'), (93, ' ?'), (245, ' ?'), (249, ' ?'),
(297, ' ?'), (393, ' ?'), (453, ' ?'), ('...', '...'), (32213, ' ?'), (32232, ' ?'), (32254, ' ?'), (32307, ' ?'), (32413, ' ?'),
(32449, ' ?'), (32469, ' ?'), (32492, ' ?'), (32510, ' ?'), (32525, ' ?')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 42 categories):
                   Country Frequency
             United-States     29170
                    Mexico       643
                         ?       583
               Philippines       198
                   Germany       137
                    Canada       121
               Puerto-Rico       114
               El-Salvador       106
                     India       100
                      Cuba        95
                    ...        ...
               Cambodia         19
         Trinadad&Tobago        19
                    Laos        18
                Thailand        18
              Yugoslavia        16
Outlying-US(Guam-USVI-etc)      14
                Honduras        13
                 Hungary        13
                Scotland        12
        Holand-Netherlands       1

predicted-attribute (predicted):
  Categorical format with 2 unique values:
Category  Frequency
   <=50K      24720
    >50K       7841
```



Dataset has 23 (0.1%) duplicate rows — Duplicates

race is highly imbalanced (65.6%) — Imbalance

native-country is highly imbalanced (82.5%) — Imbalance

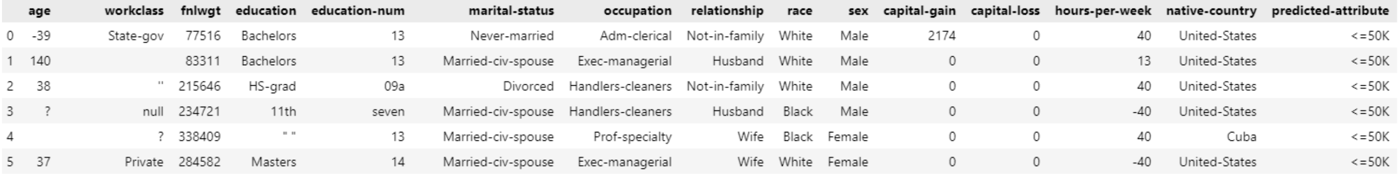
capital-gain has 29849 (91.7%) zeros — Zeros

capital-loss has 31042 (95.3%) zeros — Zeros

Ydata Profiling did not find the errors related to '?' in the four columns presented before.

Changed dataset:

```
-39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
140,, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, '', 215646, HS-grad, 09a, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
?, null, 234721, 11th, seven, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, -40, United-States, <=50K
, ?, 338409, " ", 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, -40, United-States, <=50K
```



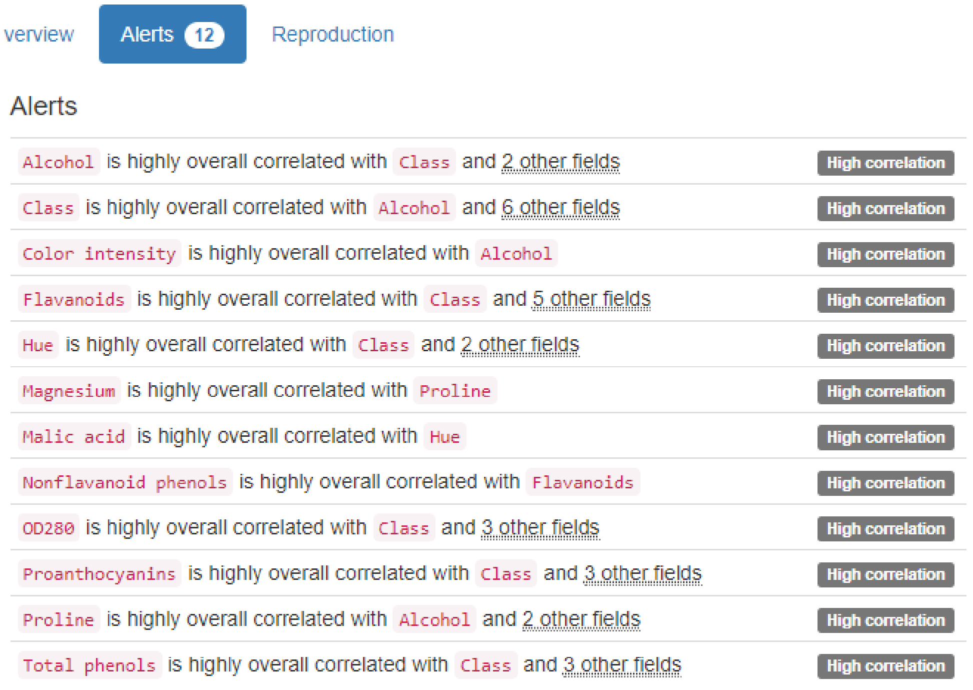
	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	predicted-attribute
0	-39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	140		83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	''	215646	HS-grad	09a	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	?	null	234721	11th	seven	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	-40	United-States	<=50K
4		?	338409	" "	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	-40	United-States	<=50K

```
Analysis results only for columns that had changed, as shown above:

age:
   Age format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(4, '')]

DQI #15 (Domain Violation - Accuracy):
 2 Value(s) outside range [0, 130] at index(es): [(0, '-39'), (1, '140')]

DQI #17 (Wrong Data Type - Consistency):
 1 Non-numeric value(s) at index(es): [(3, '?')]

workclass (class):
   Error(s) found:
DQI #1 (Missing Data - Completeness):
 3 Blank/Empty/Null/NaN value(s) at index(es): [(1, ''), (2, " ''"), (3, ' null')]

DQI #4 (Ambiguous Data - Accuracy, Consistency):
 1837 Unacceptable value(s) at index(es): [(4, ' ?'), (27, ' ?') etc as previously)]
Categorical format with 9 unique values:
         Category  Frequency
          Private      22693
 Self-emp-not-inc       2540
        Local-gov       2093
                ?       1837
        State-gov       1298
     Self-emp-inc       1116
      Federal-gov        960
      Without-pay         14
     Never-worked          7

education:
   Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(4, ' " "')])
Categorical format with 16 unique values:
     Category  Frequency
      HS-grad      10501
 Some-college       7291
    Bachelors       5354
      Masters       1723
    Assoc-voc       1382
         11th       1175
   Assoc-acdm       1067
         10th        933
      7th-8th        646
  Prof-school        576
          9th        514
         12th        433
    Doctorate        413
      5th-6th        333
      1st-4th        168
    Preschool         51

education-num (num):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 2 Non-numeric value(s) at index(es): [(2, ' 09a'), (3, ' seven')]

hours-per-week (hours):
 Numerical >=0 format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 2 Negative value(s) at index(es): [(3, -40), (5, -40)]
```

107 Wine

	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	OD280/OD315 of diluted wines	Proline
0	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
2	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
3	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
4	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735

```
Class:
  All 178 values are correctly categorical.

Categorical format with 3 unique values:
 Category  Frequency
        2         71
        1         59
        3         48

Alcohol:
  Numerical format: All 178 values are numerical in the range (11.03:14.83).

Malic acid (acid):
  Numerical format: All 178 values are numerical in the range (0.74:5.8).

Ash:
  Numerical >=0 format: All 178 values are numerical and greater or equal to 0 in the range (1.36:3.23).

Alcalinity of ash (ash):
  Numerical >=0 format: All 178 values are numerical and greater or equal to 0 in the range (10.6:30.0).

Magnesium:
  Numerical format: All 178 values are numerical in the range (70:162).

Total phenols (total):
  Numerical format: All 178 values are numerical in the range (0.98:3.88).

Flavanoids:
  Target word not found, and Format not determined

Nonflavanoid phenols:
  Target word not found, and Format not determined

Proanthocyanins:
  Target word not found, and Format not determined

Color intensity (intensity):
  Numerical format: All 178 values are numerical in the range (1.28:13.0).

Hue:
  Target word not found, and Format not determined

OD280/OD315 of diluted wines:
  Target word not found, and Format not determined

Proline:
  Target word not found, and Format not determined
```



Alerts

`Alcohol` is highly overall correlated with `Class` and 2 other fields — High correlation

`Class` is highly overall correlated with `Alcohol` and 6 other fields — High correlation

`Color intensity` is highly overall correlated with `Alcohol` — High correlation

`Flavanoids` is highly overall correlated with `Class` and 5 other fields — High correlation

`Hue` is highly overall correlated with `Class` and 2 other fields — High correlation

`Magnesium` is highly overall correlated with `Proline` — High correlation

`Malic acid` is highly overall correlated with `Hue` — High correlation

`Nonflavanoid phenols` is highly overall correlated with `Flavanoids` — High correlation

`OD280` is highly overall correlated with `Class` and 3 other fields — High correlation

`Proanthocyanins` is highly overall correlated with `Class` and 3 other fields — High correlation

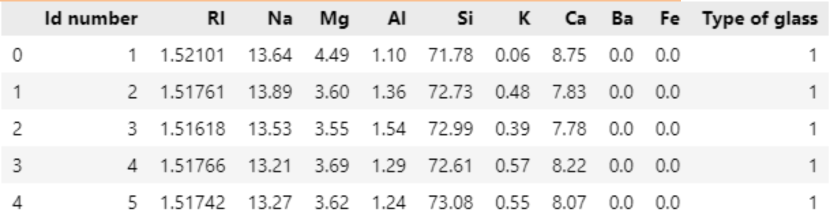
`Proline` is highly overall correlated with `Alcohol` and 2 other fields — High correlation

`Total phenols` is highly overall correlated with `Class` and 3 other fields — High correlation

YData only found High correlation Alerts.

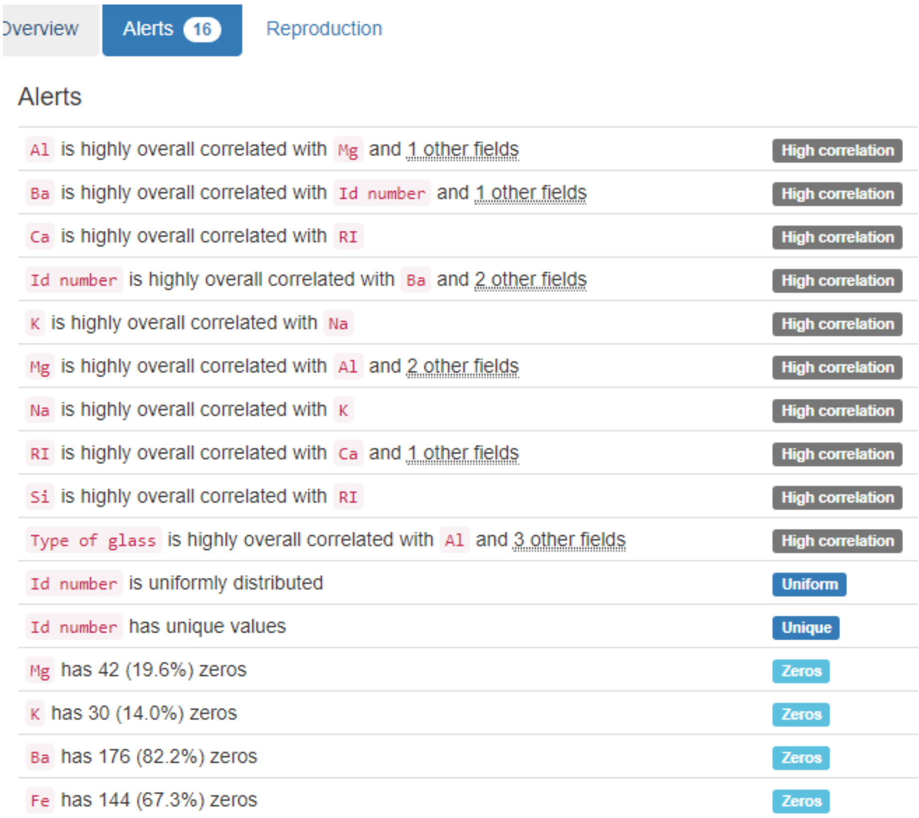
42 Glass Identification

1. Id number	1	Id number	ID column
2. RI: refractive index	2	RI	numerical
3. Na: Sodium	3	Na	numerical
4. Mg: Magnesium	4	Mg	numerical
5. Al: Aluminum	5	Al	numerical
6. Si: Silicon	6	Si	numerical
7. K: Potassium	7	K	numerical
8. Ca: Calcium	8	Ca	numerical
9. Ba: Barium	9	Ba	numerical
10. Fe: Iron	10	Fe	numerical
11. Type of glass (class attribute)	11	Type of glass	categorical



	Id number	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type of glass
0	1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.0	1
1	2	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.0	1
2	3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.0	1
3	4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.0	1
4	5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.0	1

```
Id number (id):
  ID column format: All 214 ID values are unique and valid, and thus suitable for use as a Primary Key.
Alphanumeric range of values: (1 : 99)

RI (index):
  Numerical format: All 214 values are numerical in the range (1.51115:1.53393).

Na (sodium):
  Numerical format: All 214 values are numerical in the range (10.73:17.38).

Mg (magnesium):
  Numerical format: All 214 values are numerical in the range (0.0:4.49).

Al (aluminum):
  Numerical format: All 214 values are numerical in the range (0.29:3.5).

Si (silicon):
  Numerical format: All 214 values are numerical in the range (69.81:75.41).

K (potassium):
  Numerical format: All 214 values are numerical in the range (0.0:6.21).

Ca (calcium):
  Numerical format: All 214 values are numerical in the range (5.43:16.19).

Ba (barium):
  Numerical format: All 214 values are numerical in the range (0.0:3.15).

Fe (iron):
  Numerical format: All 214 values are numerical in the range (0.0:0.51).

Type of glass (type):
  All 214 values are correctly categorical.

Categorical format with 6 unique values:
 Category  Frequency
        2         76
        1         70
        7         29
        3         17
        5         13
        6          9
```



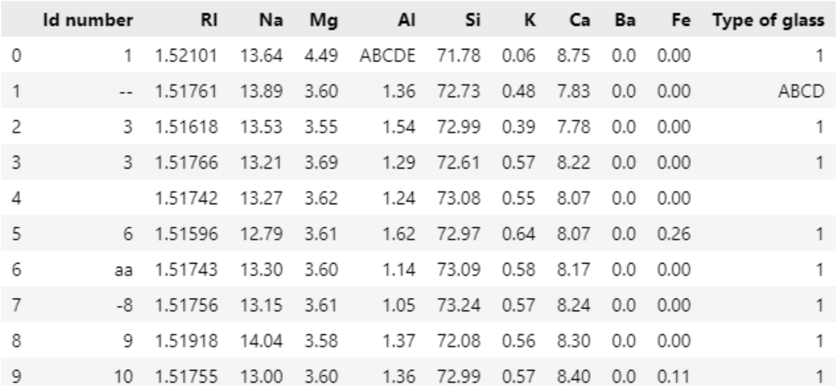
Alerts

`Al` is highly overall correlated with `Mg` and 1 other fields	High correlation
`Ba` is highly overall correlated with `Id number` and 1 other fields	High correlation
`Ca` is highly overall correlated with `RI`	High correlation
`Id number` is highly overall correlated with `Ba` and 2 other fields	High correlation
`K` is highly overall correlated with `Na`	High correlation
`Mg` is highly overall correlated with `Al` and 2 other fields	High correlation
`Na` is highly overall correlated with `K`	High correlation
`RI` is highly overall correlated with `Ca` and 1 other fields	High correlation
`Si` is highly overall correlated with `RI`	High correlation
`Type of glass` is highly overall correlated with `Al` and 3 other fields	High correlation
`Id number` is uniformly distributed	Uniform
`Id number` has unique values	Unique
`Mg` has 42 (19.6%) zeros	Zeros
`K` has 30 (14.0%) zeros	Zeros
`Ba` has 176 (82.2%) zeros	Zeros
`Fe` has 144 (67.3%) zeros	Zeros

YData found that Id number is Unique as we did. No Data Quality Issue as we check.

Glass_CHANGED

```
1,1.52101,13.64,4.49,ABCDE,71.78,0.06,8.75,0.00,0.00,1
--,1.51761,13.89,3.60,1.36,72.73,0.48,7.83,0.00,0.00,ABCD
3,1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0.00,0.00,1
3,1.51766,13.21,3.69,1.29,72.61,0.57,8.22,0.00,0.00,1
 ,1.51742,13.27,3.62,1.24,73.08,0.55,8.07,0.00,0.00,
6,1.51596,12.79,3.61,1.62,72.97,0.64,8.07,0.00,0.26,1
aa,1.51743,13.30,3.60,1.14,73.09,0.58,8.17,0.00,0.00,1
-8,1.51756,13.15,3.61,1.05,73.24,0.57,8.24,0.00,0.00,1
9,1.51918,14.04,3.58,1.37,72.08,0.56,8.30,0.00,0.00,1
10,1.51755,13.00,3.60,1.36,72.99,0.57,8.40,0.00,0.11,1
```



	Id number	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type of glass
0	1	1.52101	13.64	4.49	ABCDE	71.78	0.06	8.75	0.0	0.00	1
1	--	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.00	ABCD
2	3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.00	1
3	3	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.00	1
4		1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.00	
5	6	1.51596	12.79	3.61	1.62	72.97	0.64	8.07	0.0	0.26	1
6	aa	1.51743	13.30	3.60	1.14	73.09	0.58	8.17	0.0	0.00	1
7	-8	1.51756	13.15	3.61	1.05	73.24	0.57	8.24	0.0	0.00	1
8	9	1.51918	14.04	3.58	1.37	72.08	0.56	8.30	0.0	0.00	1
9	10	1.51755	13.00	3.60	1.36	72.99	0.57	8.40	0.0	0.11	1

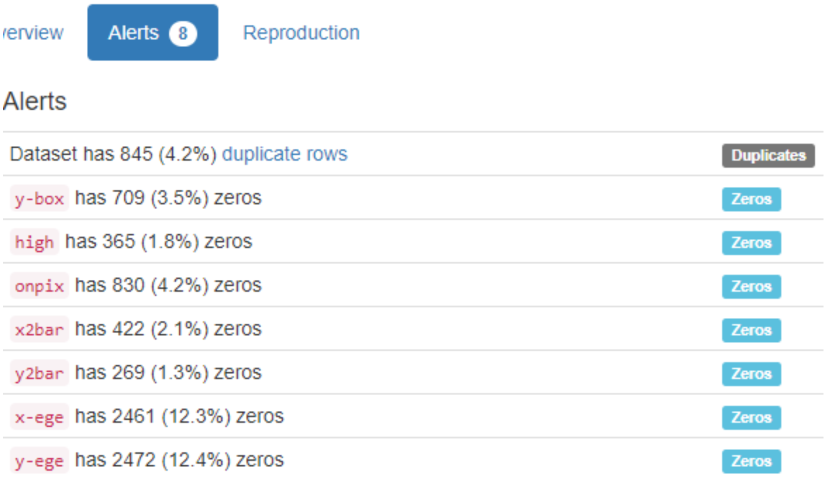
```
Id number (id):
  ID column format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(4, ' ')]

DQI #9 (Duplicates - Uniqueness):
 2 Duplicate value(s) at index(es): [[2, '3'], [3, '3']]

DQI #15 (Domain Violation - Accuracy):
 1 Negative value(s) at index(es): [(7, '-8')]

DQI #19 (Uniqueness Violation - Uniqueness):
 1 Uniqueness violation at index(es): [(3, '3')]

Alphanumeric range of values: (-- : aa)

Al (aluminum):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 1 Non-numeric value(s) at index(es):  0, 'ABCDE')])

Type of glass (type):
   Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(4, '')])
Categorical format with 7 unique values:
Category  Frequency
       2         76
       1         68
       7         29
       3         17
       5         13
       6          9
    ABCD          1
```

58 Letter Recognition

```
INFO:root:Successfully assigned column names to the dataset 'Letter Recognition' for index 58
```

	lettr	x-box	y-box	width	high	onpix	x-bar	y-bar	x2bar	y2bar	xybar	x2ybr	xy2br	x-ege	xegvy	y-ege	yegvx
0	T	2	8	3	5	1	8	13	0	6	6	10	8	0	8	0	8
1	I	5	12	3	7	2	10	5	5	4	13	3	9	2	8	4	10
2	D	4	11	6	8	6	10	6	2	6	10	3	7	3	7	3	9
3	N	7	11	6	6	3	5	9	4	6	4	4	10	6	10	2	8
4	G	2	1	3	1	1	8	6	6	6	6	5	9	1	7	5	10

```
lettr (letter):  String format: All 20000 string values are valid.
String range (lexicographical): (A : Z)

x-box (x):
  Numerical format: All 20000 values are numerical in the range (0:15).

y-box (y):
  Numerical format: All 20000 values are numerical in the range (0:15).

width:
  Numerical >=0 format: All 20000 values are numerical and greater or equal to 0 in the range (0:15).

high (height):
  Numerical >=0 format: All 20000 values are numerical and greater or equal to 0 in the range (0:15).

onpix (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).

x-bar (x):
  Numerical format: All 20000 values are numerical in the range (0:15).

y-bar (y):
  Numerical format: All 20000 values are numerical in the range (0:15).

x2bar (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).

y2bar (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).

xybar (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).

x2ybr (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).

xy2br (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).

x-ege (x):
  Numerical format: All 20000 values are numerical in the range (0:15).

xegvy (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).

y-ege (y):
  Numerical format: All 20000 values are numerical in the range (0:15).

yegvx (integer):
  Numerical format: All 20000 values are numerical in the range (0:15).
```



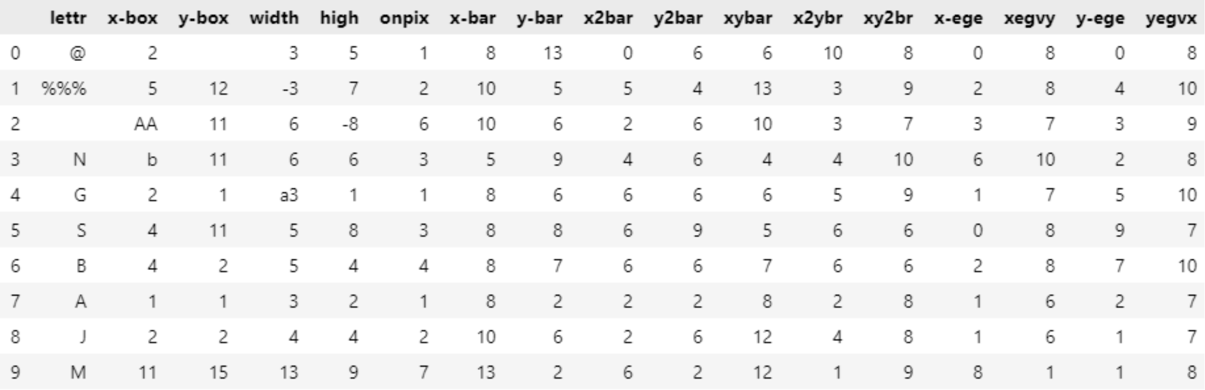
Alerts

Dataset has 845 (4.2%) duplicate rows	Duplicates
y-box has 709 (3.5%) zeros	Zeros
high has 365 (1.8%) zeros	Zeros
onpix has 830 (4.2%) zeros	Zeros
x2bar has 422 (2.1%) zeros	Zeros
y2bar has 269 (1.3%) zeros	Zeros
x-ege has 2461 (12.3%) zeros	Zeros
y-ege has 2472 (12.4%) zeros	Zeros

YData did not find any of the Data Quality Issues we search for.

Changed

```
@,2,,3,5,1,8,13,0,6,6,10,8,0,8,0,8
%%%,5,12,-3,7,2,10,5,5,4,13,3,9,2,8,4,10
,AA,11,6,-8,6,10,6,2,6,10,3,7,3,7,3,9
N,b,11,6,6,3,5,9,4,6,4,4,10,6,10,2,8
G,2,1,a3,1,1,8,6,6,6,6,5,9,1,7,5,10
```



	lettr	x-box	y-box	width	high	onpix	x-bar	y-bar	x2bar	y2bar	xybar	x2ybr	xy2br	x-ege	xegvy	y-ege	yegvx
0	@	2		3	5	1	8	13	0	6	6	10	8	0	8	0	8
1	%%%	5	12	-3	7	2	10	5	5	4	13	3	9	2	8	4	10
2		AA	11	6	-8	6	10	6	2	6	10	3	7	3	7	3	9
3	N	b	11	6	6	3	5	9	4	6	4	4	10	6	10	2	8
4	G	2	1	a3	1	1	8	6	6	6	6	5	9	1	7	5	10
5	S	4	11	5	8	3	8	8	6	9	5	6	6	0	8	9	7
6	B	4	2	5	4	4	8	7	6	6	7	6	6	2	8	7	10
7	A	1	1	3	2	1	8	2	2	2	8	2	8	1	6	2	7
8	J	2	2	4	4	2	10	6	2	6	12	4	8	1	6	1	7
9	M	11	15	13	9	7	13	2	6	2	12	1	9	8	1	1	8

```
lettr (letter):
  String format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(2, '')]

String range (lexicographical): (%%% : Z)

x-box (x):
  Numerical format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 2 Non-numeric value(s) at index(es): [(2, 'AA'), (3, 'b')]

y-box (y):
  Numerical format: Error(s) found:
DQI #1 (Missing Data - Completeness):
 1 Blank/Empty/Null/NaN value(s) at index(es): [(0, '')]


width:
  Numerical >=0 format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
  1 Negative value(s) at index(es): [(1, '-3')]

DQI #17 (Wrong Data Type - Consistency):
 1 Non-numeric value(s) at index(es): [(4, 'a3')]

Range of values: (-3.0:15.0).

high (height):
  Numerical >=0 format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 1 Negative value(s) at index(es): [(2, -8)]

Range of values: (-8:15).
```

142 Statlog (German Credit Data)

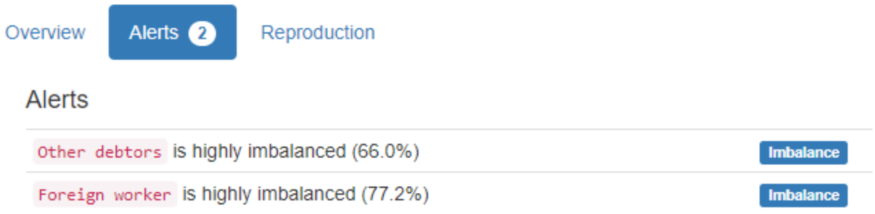
	Status of existing checking account	Duration in months	Credit history	Purpose	Credit amount	Savings account/bonds	Present employment since	Installment rate in percentage of disposable income	Personal status and sex	Other debtors/guarantors	...	Property	Age in years	Other instalment plans	Housing	Number of existing credits at this bank	Job	Number of people being liable to provide maintenance for	Telephone	Foreign worker	predicted-attribute
0	A11	6	A34	A43	1169	A65	A75	4	A93	A101	...	A121	67	A143	A152	2	A173	1	A192	A201	1
1	A12	48	A32	A43	5951	A61	A73	2	A92	A101	...	A121	22	A143	A152	1	A173	1	A191	A201	2
2	A14	12	A34	A46	2096	A61	A74	2	A93	A101	...	A121	49	A143	A152	1	A172	2	A191	A201	1
3	A11	42	A32	A42	7882	A61	A74	2	A93	A103	...	A122	45	A143	A153	1	A173	2	A191	A201	1
4	A11	24	A33	A40	4870	A61	A73	3	A93	A101	...	A124	53	A143	A153	2	A173	2	A191	A201	2

```
Status of existing checking account (status):
  All 1000 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
     A14        394
     A11        274
     A12        269
     A13         63

Duration in months (months):
  Numerical >=0 format: All 1000 values are numerical and greater or equal to 0 in the range (4:72).

Credit history (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 5 unique values:
Category  Frequency
     A32        530
     A34        293
     A33         88
     A31         49
     A30         40

Purpose (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 10 unique values:
Category  Frequency
     A43        280
     A40        234
     A42        181
     A41        103
     A49         97
     A46         50
     A45         22
     A44         12
    A410         12
     A48          9

Credit amount (amount):
  Numerical format: All 1000 values are numerical in the range (250:18424).

Savings account (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 5 unique values:
Category  Frequency
     A61        603
     A65        183
     A62        103
     A63         63
     A64         48

Present employment since (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 5 unique values:
Category  Frequency
     A73        339
     A75        253
     A74        174
     A72        172
     A71         62

Installment rate in percentage of disposable income (percentage):
   Percentage format: All numerical values are valid in the range [0, 100].
Actual range of values: (1 : 4)

Personal status and sex (status):
  All 1000 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
     A93        548
     A92        310
     A94         92
     A91         50

Other debtors (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
    A101        907
    A103         52
    A102         41

Present residence since (since):
  Numerical >=0 format: All 1000 values are numerical and greater or equal to 0 in the range (1:4).

Property (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
    A123        332
    A121        282
    A122        232
    A124        154

Age in years (age):
  Age format: All 1000 values are numerical and valid in the range [0, 130].
Actual range of values: (19 : 75)

Other instalment plans (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
    A143        814
    A141        139
    A142         47

Housing (qualitative):
  All 1000 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
    A152        713
    A151        179
    A153        108

Number of existing credits at this bank (number):
  Numerical format: All 1000 values are numerical in the range (1:4).

Job:
  All 1000 values are correctly categorical.

Categorical format with 4 unique values:
Category  Frequency
    A173        630
    A172        200
    A174        148
    A171         22

Number of people being liable to provide maintenance for (number):
  Numerical format: All 1000 values are numerical in the range (1:2).

Telephone:
  Phone format: Error(s) found:
DQI #15 (Domain Violation - Accuracy):
 999 Incorrect telephone number format at index(es): [(0, 'A192'), (1, 'A191'), (2, 'A191'), (3, 'A191'), (4, 'A191'), (5, 'A192'),
…
(994, 'A192'), (995, 'A191'), (996, 'A192'), (997, 'A191'), (998, 'A192'), (999, 'A191')])
```

This column had to be altered manually in the AnalysedColumns sheet, from ‘phone’ to ‘categorical’ so that it could be analysed as such.

This is the new output:

```
Telephone:
  Categorical format with 2 unique values:
Category  Frequency
    A191        596
    A192        404

Foreign worker (worker):
  All 1000 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
    A201        963
    A202         37

predicted-attribute (predicted):
  All 1000 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        1        700
        2        300
```

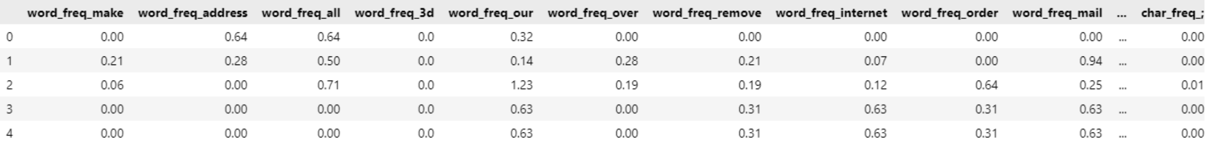


Alerts

Other debtors is highly imbalanced (66.0%)	Imbalance
Foreign worker is highly imbalanced (77.2%)	Imbalance

YData only found Imbalance alerts.

92 Spambase



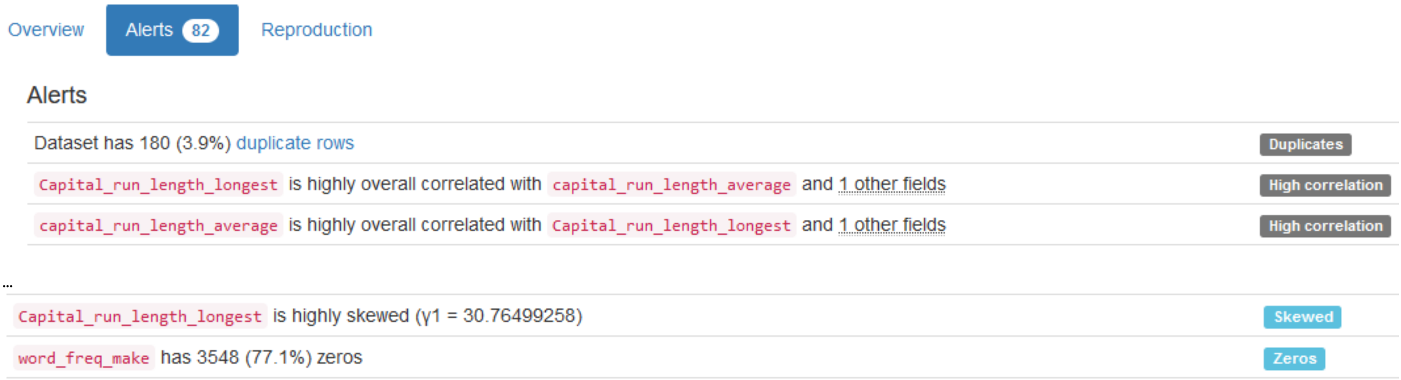
	word_freq_make	word_freq_address	word_freq_all	word_freq_3d	word_freq_our	word_freq_over	word_freq_remove	word_freq_internet	word_freq_order	word_freq_mail	...	char_freq_;
0	0.00	0.64	0.64	0.0	0.32	0.00	0.00	0.00	0.00	0.00	...	0.00
1	0.21	0.28	0.50	0.0	0.14	0.28	0.21	0.07	0.00	0.94	...	0.00
2	0.06	0.00	0.71	0.0	1.23	0.19	0.19	0.12	0.64	0.25	...	0.01
3	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31	0.63	...	0.00
4	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31	0.63	...	0.00

5 rows × 58 columns

```
word_freq_make (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:4.54).

word_freq_address (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:14.28).

word_freq_all (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:5.1).

word_freq_3d (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:42.81).

word_freq_our (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:10.0).

word_freq_over (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:5.88).

word_freq_remove (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:7.27).

word_freq_internet (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:11.11).

word_freq_order (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:5.26).

word_freq_mail (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:18.18).

word_freq_receive (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:2.61).

word_freq_will (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:9.67).

word_freq_people (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:5.55).

word_freq_report (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:10.0).

word_freq_addresses (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:4.41).

word_freq_free (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:20.0).

word_freq_business (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:7.14).

word_freq_email (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:9.09).

word_freq_you (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:18.75).

word_freq_credit (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:18.18).

word_freq_your (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:11.11).

word_freq_font (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:17.1).

word_freq_000 (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:5.45).
```

```
word_freq_money (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:12.5).

word_freq_hp (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:20.83).

word_freq_hpl (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:16.66).

word_freq_george (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:33.33).

word_freq_650 (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:9.09).

word_freq_lab (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:14.28).

word_freq_labs (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:5.88).

word_freq_telnet (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:12.5).

word_freq_857 (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:4.76).

word_freq_data (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:18.18).

word_freq_415 (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:4.76).

word_freq_85 (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:20.0).

word_freq_technology (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:7.69).

word_freq_1999 (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:6.89).

word_freq_parts (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:8.33).

word_freq_pm (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:11.11).

word_freq_direct (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:4.76).

word_freq_cs (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:7.14).

word_freq_meeting (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:14.28).

word_freq_original (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:3.57).

word_freq_project (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:20.0).

word_freq_re (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:21.42).

word_freq_edu (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:22.05).

word_freq_table (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:2.17).

word_freq_conference (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:10.0).

char_freq_; (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:4.385).

char_freq_ (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:9.752).

char_freq_[ (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:4.081).

char_freq_! (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:32.478).

char_freq_$ (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:6.003).

char_freq_* (freq):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (0.0:19.829).
```

```
capital_run_length_average (length):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (1.0:1102.5).

Capital_run_length_longest (length):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (1:9989).

capital_run_length_total (length):
  Numerical >=0 format: All 4601 values are numerical and greater or equal to 0 in the range (1:15841).

spam (nominal):
  All 4601 values are correctly categorical.

Categorical format with 2 unique values:
 Category  Frequency
        0       2788
        1       1813
```



Alerts

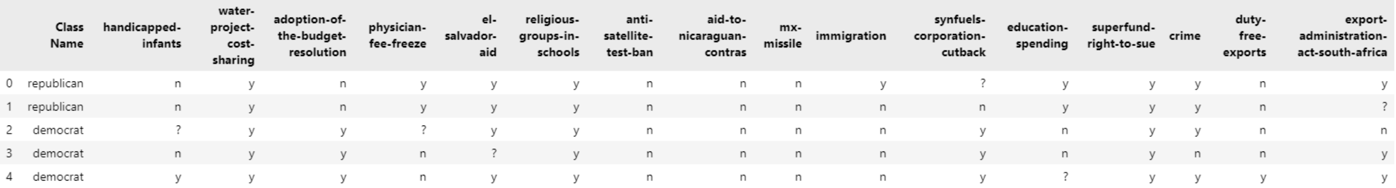
Dataset has 180 (3.9%) duplicate rows	Duplicates
Capital_run_length_longest is highly overall correlated with capital_run_length_average and 1 other fields	High correlation
capital_run_length_average is highly overall correlated with Capital_run_length_longest and 1 other fields	High correlation

…

Capital_run_length_longest is highly skewed (γ1 = 30.76499258)	Skewed
word_freq_make has 3548 (77.1%) zeros	Zeros

Even though there are 82 alerts from YData, there is no one that is interesting to us.

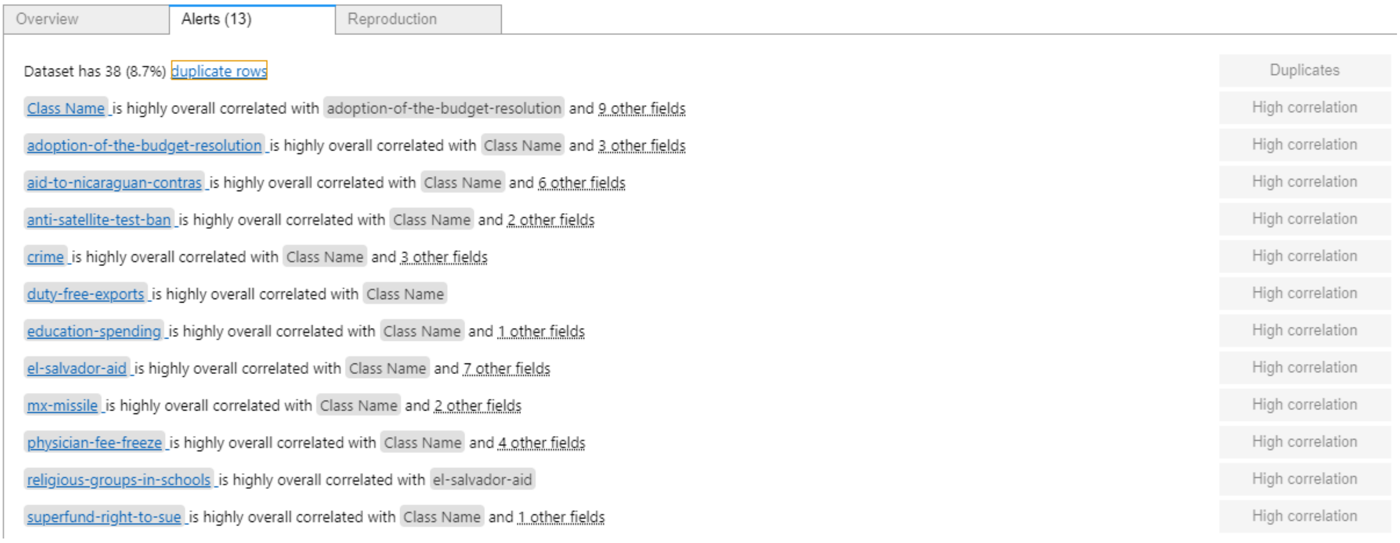
103 Congressional Voting Records



	Class Name	handicapped-infants	water-project-cost-sharing	adoption-of-the-budget-resolution	physician-fee-freeze	el-salvador-aid	religious-groups-in-schools	anti-satellite-test-ban	aid-to-nicaraguan-contras	mx-missile	immigration	synfuels-corporation-cutback	education-spending	superfund-right-to-sue	crime	duty-free-exports	export-administration-act-south-africa
0	republican	n	y	n	y	y	y	n	n	n	y	?	y	y	y	n	y
1	republican	n	y	n	y	y	y	n	n	n	n	n	y	y	y	n	?
2	democrat	?	y	y	?	y	y	n	n	n	n	y	n	y	y	n	n
3	democrat	n	y	y	n	?	y	n	n	n	n	y	n	y	n	n	y
4	democrat	y	y	y	n	y	y	n	n	n	n	y	?	y	y	y	y

```
Class Name (name):
  Name format: All 435 Class Name values are valid.

Frequency Distribution:
Class Name  Frequency
  democrat        267
republican        168

handicapped-infants (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 12 Unacceptable value(s) at index(es): [(2, '?'), (104, '?'), (129, '?'), (143, '?'), (178, '?'), (180, '?'), (183, '?'), (248, '?'),
(390, '?'), (393, '?'), (402, '?'), (428, '?')])
Categorical format with 3 unique values:
Category  Frequency
       n        236
       y        187
       ?         12

water-project-cost-sharing (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 48 Unacceptable value(s) at index(es): [(17, '?'), (22, '?'), (36, '?'), (49, '?'), (68, '?'), (84, '?'), (104, '?'), (107, '?'),
(108, '?'), (109, '?'), ('...', '...'), (248, '?'), (329, '?'), (331, '?'), (341, '?'), (376, '?'), (386, '?'), (390, '?'), (393, '?'),
(428, '?'), (432, '?')] (displaying only the first and last 10 items)
Category  Frequency

       y        195
       n        192
       ?         48

adoption-of-the-budget-resolution (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 11 Unacceptable value(s) at index(es): [(96, '?'), (104, '?'), (107, '?'), (120, '?'), (151, '?'), (183, '?'), (248, '?'), (301, '?'),
(393, '?'), (394, '?'), (428, '?')])
Categorical format with 3 unique values:
Category  Frequency
       y        253
       n        171
       ?         11
```

```
physician-fee-freeze (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 11 Unacceptable value(s) at index(es): [(2, '?'), (104, '?'), (107, '?'), (183, '?'), (248, '?'), (287, '?'), (341, '?'), (373, '?'),
(393, '?'), (394, '?'), (395, '?')])
Categorical format with 3 unique values:
Category  Frequency
       n        247
       y        177
       ?         11

el-salvador-aid (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 15 Unacceptable value(s) at index(es): 3, '?'), (107, '?'), (115, '?'), (130, '?'), (159, '?'), (183, '?'), (248, '?'), (290, '?'),
(322, '?'), (370, '?'), (371, '?'), (390, '?'), (394, '?'), (399, '?'), (429, '?')])
Categorical format with 3 unique values:
Category  Frequency
       y        212
       n        208
       ?         15

religious-groups-in-schools (religion):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 11 Unacceptable value(s) at index(es): [(20, '?'), (60, '?'), (107, '?'), (115, '?'), (130, '?'), (183, '?'), (248, '?'), (261, '?'),
(342, '?'), (397, '?'), (424, '?')])
Categorical format with 3 unique values:
Category  Frequency
       y        272
       n        152
       ?         11

anti-satellite-test-ban (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 14 Unacceptable value(s) at index(es): [(54, '?'), (103, '?'), (107, '?'), (170, '?'), (183, '?'), (247, '?'), (248, '?'), (286, '?'),
(290, '?'), (295, '?'), (377, '?'), (380, '?'), (390, '?'), (433, '?')])
Categorical format with 3 unique values:
Category  Frequency
       y        239
       n        182
       ?         14

aid-to-nicaraguan-contras (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 15 Unacceptable value(s) at index(es): [(51, '?'), (95, '?'), (107, '?'), (155, '?'), (183, '?'), (191, '?'), (194, '?'), (240, '?'),
(248, '?'), (295, '?'), (315, '?'), (322, '?'), (377, '?'), (400, '?'), (433, '?')])
Categorical format with 3 unique values:
Category  Frequency
       y        242
       n        178
       ?         15

mx-missile (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 22 Unacceptable value(s) at index(es): [(13, '?'), (16, '?'), (45, '?'), (47, '?'), (81, '?'), (95, '?'), (102, '?'), (103, '?'),
(107, '?'), (129, '?'), ('...', '...'), (238, '?'), (243, '?'), (248, '?'), (249, '?'), (286, '?'), (323, '?'), (325, '?'), (334, '?'),
(415, '?'), (433, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
       y        207
       n        206
       ?         22

immigration (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 7 Unacceptable value(s) at index(es): [(107, '?'), (170, '?'), (180, '?'), (183, '?'), (248, '?'), (389, '?'), (433, '?')])
Categorical format with 3 unique values:
Category  Frequency
       y        216
       n        212
       ?          7

synfuels-corporation-cutback (corporation):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 21 Unacceptable value(s) at index(es): [(0, '?'), (10, '?'), (40, '?'), (99, '?'), (104, '?'), (107, '?'), (115, '?'), (129, '?'),
(180, '?'), (183, '?'), ('...', '...'), (248, '?'), (261, '?'), (271, '?'), (275, '?'), (359, '?'), (370, '?'), (373, '?'), (377, '?'),
(400, '?'), (413, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
       n        264
       y        150
       ?         21

education-spending (education):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 31 Unacceptable value(s) at index(es): [(4, '?'), (10, '?'), (11, '?'), (13, '?'), (18, '?'), (22, '?'), (64, '?'), (89, '?'), (95,
'?'), (107, '?'), ('...', '...'), (248, '?'), (261, '?'), (282, '?'), (291, '?'), (354, '?'), (370, '?'), (377, '?'), (395, '?'), (400,
'?'), (413, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
       n        233
       y        171
       ?         31

superfund-right-to-sue (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 25 Unacceptable value(s) at index(es): [(6, '?'), (14, '?'), (16, '?'), (21, '?'), (36, '?'), (95, '?'), (107, '?'), (129, '?'), (157,
'?'), (172, '?'), ('...', '...'), (295, '?'), (297, '?'), (315, '?'), (341, '?'), (372, '?'), (400, '?'), (403, '?'), (408, '?'), (413,
'?'), (424, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
       y        209
       n        201
       ?         25

crime (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 17 Unacceptable value(s) at index(es): [(14, '?'), (15, '?'), (21, '?'), (92, '?'), (95, '?'), (129, '?'), (157, '?'), (159, '?'),
(183, '?'), (228, '?'), (248, '?'), (261, '?'), (295, '?'), (315, '?'), (341, '?'), (377, '?'), (429, '?')])
Categorical format with 3 unique values:
Category  Frequency
       y        248
       n        170
       ?         17

duty-free-exports (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 28 Unacceptable value(s) at index(es): [(7, '?'), (9, '?'), (11, '?'), (12, '?'), (52, '?'), (103, '?'), (107, '?'), (120, '?'), (129,
'?'), (141, '?'), ('...', '...'), (257, '?'), (261, '?'), (273, '?'), (284, '?'), (287, '?'), (295, '?'), (336, '?'), (373, '?'), (377,
'?'), (434, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
       n        233
       y        174
       ?         28

export-administration-act-south-africa (yes):
  Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 104 Unacceptable value(s) at index(es): [(1, '?'), (9, '?'), (11, '?'), (12, '?'), (13, '?'), (14, '?'), (15, '?'), (24, '?'), (31,
'?'), (40, '?'), ('...', '...'), (381, '?'), (382, '?'), (386, '?'), (387, '?'), (388, '?'), (389, '?'), (390, '?'), (393, '?'), (400,
'?'), (425, '?')] (displaying only the first and last 10 items)
Categorical format with 3 unique values:
Category  Frequency
       y        269
       ?        104
       n         62
```

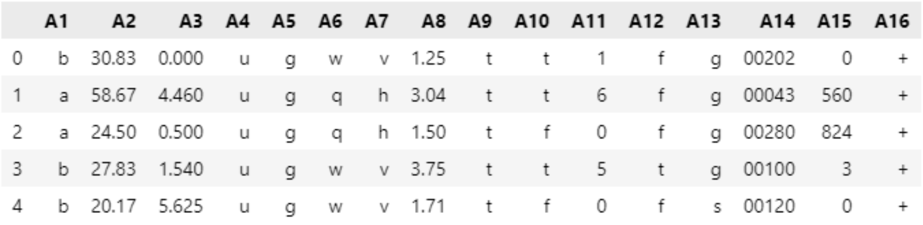


Dataset has 38 (8.7%) duplicate rows	Duplicates
Class Name is highly overall correlated with adoption-of-the-budget-resolution and 9 other fields	High correlation
adoption-of-the-budget-resolution is highly overall correlated with Class Name and 3 other fields	High correlation
aid-to-nicaraguan-contras is highly overall correlated with Class Name and 6 other fields	High correlation
anti-satellite-test-ban is highly overall correlated with Class Name and 2 other fields	High correlation
crime is highly overall correlated with Class Name and 3 other fields	High correlation
duty-free-exports is highly overall correlated with Class Name	High correlation
education-spending is highly overall correlated with Class Name and 1 other fields	High correlation
el-salvador-aid is highly overall correlated with Class Name and 7 other fields	High correlation
mx-missile is highly overall correlated with Class Name and 2 other fields	High correlation
physician-fee-freeze is highly overall correlated with Class Name and 4 other fields	High correlation
religious-groups-in-schools is highly overall correlated with el-salvador-aid	High correlation
superfund-right-to-sue is highly overall correlated with Class Name and 1 other fields	High correlation

Ydata Profiling did not find any of the many columns with '?' on it.

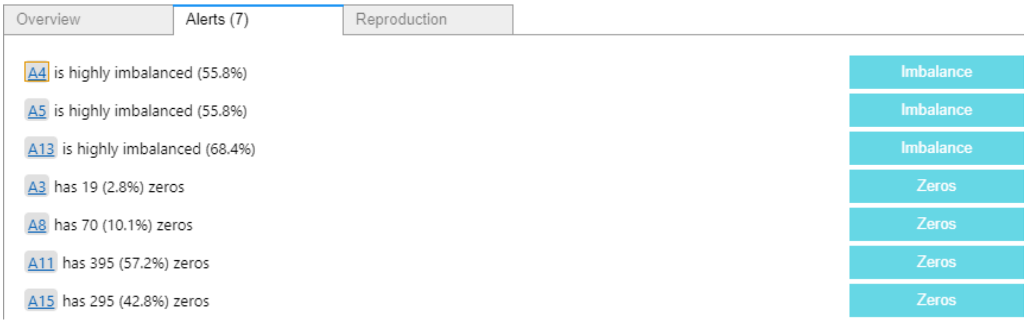
27 Credit Approval

```
INFO:root:Successfully assigned column names to the dataset 'Credit Approval' for index 27
```



	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
0	b	30.83	0.000	u	g	w	v	1.25	t	t	1	f	g	00202	0	+
1	a	58.67	4.460	u	g	q	h	3.04	t	t	6	f	g	00043	560	+
2	a	24.50	0.500	u	g	q	h	1.50	t	f	0	f	g	00280	824	+
3	b	27.83	1.540	u	g	w	v	3.75	t	t	5	t	g	00100	3	+
4	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	00120	0	+

```
A1 (categorical):
    Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 12 Unacceptable value(s) at index(es): [(248, '?'), (327, '?'), (346, '?'), (374, '?'), (453, '?'), (479, '?'), (489, '?'), (520,
'?'), (598, '?'), (601, '?'), (641, '?'), (673, '?')])
Categorical with 3 unique values:
Category  Frequency
       b        468
       a        210
       ?         12


A2 (continuous):
   Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 12 Non-numeric value(s) at index(es): [(83, '?'), (86, '?'), (92, '?'), (97, '?'), (254, '?'), (286, '?'), (329, '?'), (445, '?'),
(450, '?'), (500, '?'), (515, '?'), (608, '?')])

Range of values: (13.75:80.25).

A3 (continuous):
  Numerical >=0 format: All 690 values are numerical and greater or equal to 0 in the range (0.0:28.0).

A4 (categorical):
    Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 6 Unacceptable value(s) at index(es): [(206, '?'), (270, '?'), (330, '?'), (456, '?'), (592, '?'), (622, '?')])
Categorical with 4 unique values:
Category  Frequency
       u        519
       y        163
       ?          6
       l          2

A5 (categorical):
    Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 6 Unacceptable value(s) at index(es): [(206, '?'), (270, '?'), (330, '?'), (456, '?'), (592, '?'), (622, '?')])
Categorical with 4 unique values:
Category  Frequency
       g        519
       p        163
       ?          6
      gg          2

A6 (categorical):
    Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 9 Unacceptable value(s) at index(es): [(206, '?'), (270, '?'), (330, '?'), (456, '?'), (479, '?'), (539, '?'), (592, '?'), (601, '?'),
(622, '?')])
Categorical with 15 unique values:
Category  Frequency
       c        137
       q         78
       w         64
       i         59
      aa         54
      ff         53
       k         51
      cc         41
       m         38
       x         38
       d         30
       e         25
       j         10
       ?          9
       r          3


A7 (categorical):
    Error(s) found:
DQI #4 (Ambiguous Data - Accuracy, Consistency):
 9 Unacceptable value(s) at index(es): [(206, '?'), (270, '?'), (330, '?'), (456, '?'), (479, '?'), (539, '?'), (592, '?'), (601, '?'),
(622, '?')])
Categorical with 10 unique values:
Category  Frequency
       v        399
       h         138
      bb          59
      ff          57
       ?           9
       j           8
       z           8
      dd           6
       n           4
       o           2

A8 (continuous):
  Numerical >=0 format: All 690 values are numerical and greater or equal to 0 in the range (0.0:28.5).

A9 (categorical):
  All 690 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       t         361
       f         329

A10 (categorical):
  All 690 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       f         395
       t         295

A11 (continuous):
  Numerical >=0 format: All 690 values are numerical and greater or equal to 0 in the range (0:67).

A12 (categorical):
  All 690 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       f         374
       t         316

A13 (categorical):
  All 690 values are correctly categorical.

Categorical format with 3 unique values:
Category  Frequency
       g         625
       s          57
       p           8

A14 (continuous):
  Numerical >=0 format: Error(s) found:
DQI #17 (Wrong Data Type - Consistency):
 13 Non-numeric value(s) at index(es): [(71, '?'), (202, '?'), (206, '?'), (243, '?'), (270, '?'), (278, '?'), (330, '?'), (406, '?'),
(445, '?'), (456, '?'), (592, '?'), (622, '?'), (626, '?')])

Range of values: (0.0:2000.0).

A15 (continuous):
  Numerical >=0 format: All 690 values are numerical and greater or equal to 0 in the range (0:100000).

A16 (categorical):
  All 690 values are correctly categorical.

Categorical format with 2 unique values:
Category  Frequency
       -         383
       +         307
```



A4 is highly imbalanced (55.8%)

A5 is highly imbalanced (55.8%)

A13 is highly imbalanced (68.4%)

A3 has 19 (2.8%) zeros

A8 has 70 (10.1%) zeros

A11 has 395 (57.2%) zeros

A15 has 295 (42.8%) zeros

Ydata Profiling did not find any of the many columns with '?' on it.