Class2 Data Quality, 8.16.2019

September 20, 2019

```python
[7]: import pandas as pd
     from pandas import DataFrame
     import numpy as np
     import sqlite3
     import csv
     import datetime
     ## pd.options.mode.chained_assignment = None
```

```python
[8]: pip install numpy
```

```
Requirement already satisfied: numpy in /anaconda3/lib/python3.7/site-packages
(1.16.4)
Note: you may need to restart the kernel to use updated packages.
```

```python
[14]: a=pd.read.csv('/Users/IILSCIN/Desktop')
```

```
        ␣
    ↪---------------------------------------------------------------------------

        AttributeError                          Traceback (most recent call␣
    ↪last)

        <ipython-input-14-1cd79278b304> in <module>
    ----> 1 a=pd.read.csv('/Users/IILSCIN/Desktop')


        AttributeError: module 'pandas' has no attribute 'read'
```

```python
[15]: import pandas as pd
      import sqlite3
      import csv
      a=pd.read_csv('/Users/IILSCIN/Desktop/Arms_Armour_Met.csv')
      a.describe()
```

```
[15]:           Object ID  Period  Dynasty  Reign  Portfolio  Object Begin Date  \
       count  12992.000000     0.0      0.0    0.0        0.0       12966.000000
       mean   28957.044797     NaN      NaN    NaN        NaN        1576.238624
       std    10013.946237     NaN      NaN    NaN        NaN         363.753637
       min    21810.000000     NaN      NaN    NaN        NaN      -12500.000000
       25%    25106.750000     NaN      NaN    NaN        NaN        1530.000000
       50%    28879.500000     NaN      NaN    NaN        NaN        1590.000000
       75%    32254.250000     NaN      NaN    NaN        NaN        1701.000000
       max   473912.000000     NaN      NaN    NaN        NaN        2004.000000

              Object End Date  Geography Type  County  Country  ...  Locale  Locus  \
       count     12966.000000             0.0     0.0      0.0  ...     0.0    0.0
       mean       1729.686488             NaN     NaN      NaN  ...     NaN    NaN
       std         281.985676             NaN     NaN      NaN  ...     NaN    NaN
       min       -5000.000000             NaN     NaN      NaN  ...     NaN    NaN
       25%        1609.000000             NaN     NaN      NaN  ...     NaN    NaN
       50%        1800.000000             NaN     NaN      NaN  ...     NaN    NaN
       75%        1900.000000             NaN     NaN      NaN  ...     NaN    NaN
       max        2006.000000             NaN     NaN      NaN  ...     NaN    NaN

              Excavation  River  Rights and Reproduction  Metadata Date  Column1  \
       count         0.0    0.0                      0.0   1.231900e+04      0.0
       mean          NaN    NaN                      NaN   4.287733e+04      NaN
       std           NaN    NaN                      NaN   1.026679e-08      NaN
       min           NaN    NaN                      NaN   4.287733e+04      NaN
       25%           NaN    NaN                      NaN   4.287733e+04      NaN
       50%           NaN    NaN                      NaN   4.287733e+04      NaN
       75%           NaN    NaN                      NaN   4.287733e+04      NaN
       max           NaN    NaN                      NaN   4.287733e+04      NaN

              Column2  Column3  Column4
       count      0.0      0.0      0.0
       mean       NaN      NaN      NaN
       std        NaN      NaN      NaN
       min        NaN      NaN      NaN
       25%        NaN      NaN      NaN
       50%        NaN      NaN      NaN
       75%        NaN      NaN      NaN
       max        NaN      NaN      NaN

       [8 rows x 22 columns]

[6]: a.describe
```

```
        NameError                              Traceback (most recent call
     ↪last)

        <ipython-input-6-c3ae66b7bc8b> in <module>
     ----> 1 a.describe

        NameError: name 'a' is not defined
```

[17]: `a.dtypes`

```
[17]: Object Number              object
      Is Highlight                 bool
      Is Public Domain             bool
      Object ID                   int64
      Department                 object
      Object Name                object
      Title                      object
      Culture                    object
      Period                    float64
      Dynasty                   float64
      Reign                     float64
      Portfolio                 float64
      Artist Role                object
      Artist Prefix              object
      Artist Display Name        object
      Artist Display Bio         object
      Artist Suffix              object
      Artist Alpha Sort          object
      Artist Nationality         object
      Artist Begin Date          object
      Artist End Date            object
      Object Date                object
      Object Begin Date         float64
      Object End Date           float64
      Medium                     object
      Dimensions                 object
      Credit Line                object
      Geography Type            float64
      City                       object
      State                      object
      County                    float64
      Country                   float64
      Region                    float64
      Subregion                 float64
      Locale                    float64
      Locus                     float64
```

```
Excavation                    float64
River                         float64
Classification                 object
Rights and Reproduction       float64
Link Resource                  object
Metadata Date                 float64
Repository\                    object
Column1                       float64
Column2                       float64
Column3                       float64
Column4                       float64
dtype: object
```

[21]: `a('river').drop_duplicates('river')`

```
        ␣
    →---------------------------------------------------------------------------

        TypeError                                Traceback (most recent call␣
    →last)

        <ipython-input-21-9ac5fce5158f> in <module>
    ----> 1 a('river').drop_duplicates('river')


        TypeError: 'DataFrame' object is not callable
```

[16]: `len(a)`

[16]: 12992

[22]: `len(a.columns)`

[22]: 47

[23]: `a.head(10)`

[23]:
	Object Number	Is Highlight	Is Public Domain	Object ID	Department	\
0	X.5	False	True	21810	Arms and Armor	
1	X.11	False	True	21811	Arms and Armor	
2	X.617	False	False	21812	Arms and Armor	
3	14.100.448	False	False	21813	Arms and Armor	
4	91.1.800	False	True	21814	Arms and Armor	
5	91.1.870	False	False	21815	Arms and Armor	
6	91.1.871	False	False	21816	Arms and Armor	
7	91.1.872	False	False	21817	Arms and Armor	
8	91.1.873	False	False	21818	Arms and Armor	
9	91.1.874	False	False	21819	Arms and Armor	

```
                                    Object Name  \
0                                       Gorget
1           Blade and mounting for a sword (Katana)
2                      Thumb of a left gauntlet
3                 Fencing shoulder and arm guard
4                           Sword guard (Tsuba)
5   Blade for a hatchet (Nata) and mounting for a ...
6    Blade and mounting for a dagger (Tant\uc0\u333 )
7     Blade and mounting for double-edge dagger (Ken)
8            Mounting for a dagger (Tant\uc0\u333 )
9      Blade and mounting for Dagger (Tant\uc0\u333 )


                                        Title          Culture  Period  \
0  Gorget of an Officer of the King's American Re...  Anglo-American     NaN
1            Blade and Mounting for a Sword (Katana)        Japanese     NaN
2                       Thumb of a Left Gauntlet          French     NaN
3                  Fencing Shoulder and Arm Guard        Japanese     NaN
4                            Sword Guard (Tsuba)        Japanese     NaN
5  Blade for a Hatchet (Nata) and Mounting for a ...        Japanese     NaN
6         Blade and Mounting for a Dagger (Tant\u333 )        Japanese     NaN
7    Blade and Mounting for Double-Edge Dagger (Ken)        Japanese     NaN
8              Mounting for a Dagger (Tant\u333 )        Japanese     NaN
9         Blade and Mounting for Dagger (Tant\u333 )        Japanese     NaN


   Dynasty  ...  River       Classification Rights and Reproduction  \
0     NaN  ...    NaN    Armor Parts-Gorgets                     NaN
1     NaN  ...    NaN                 Swords                     NaN
2     NaN  ...    NaN    Armor Parts-Gauntlets                   NaN
3     NaN  ...    NaN       Fencing Equipment                    NaN
4     NaN  ...    NaN    Sword Furniture-Tsuba                   NaN
5     NaN  ...    NaN                Daggers                     NaN
6     NaN  ...    NaN                Daggers                     NaN
7     NaN  ...    NaN                Daggers                     NaN
8     NaN  ...    NaN                Daggers                     NaN
9     NaN  ...    NaN                Daggers                     NaN


                                 Link Resource Metadata Date  \
0  http://www.metmuseum.org/art/collection/search...    42877.33341
1  http://www.metmuseum.org/art/collection/search...    42877.33341
2  http://www.metmuseum.org/art/collection/search...    42877.33341
3  http://www.metmuseum.org/art/collection/search...    42877.33341
4  http://www.metmuseum.org/art/collection/search...    42877.33341
5  http://www.metmuseum.org/art/collection/search...    42877.33341
6  http://www.metmuseum.org/art/collection/search...    42877.33341
7  http://www.metmuseum.org/art/collection/search...    42877.33341
8  http://www.metmuseum.org/art/collection/search...    42877.33341
```

```
9  http://www.metmuseum.org/art/collection/search...   42877.33341

                             Repository\ Column1 Column2 Column3 Column4
0  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
1  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
2  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
3  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
4  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
5  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
6  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
7  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
8  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN
9  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN     NaN

[10 rows x 47 columns]
```

[24]: `a.tail(10)`

[24]:
```
        Object Number  Is Highlight  Is Public Domain  Object ID  \
12982        2013.618         False              True      35960
12983         36.25.5a        False              True      35961
12984         2013.59         False              True      35962
12985  2007.181, .182         False              True      35963
12986      14.25.1870         False              True     436343
12987      14.25.1878         False              True     436699
12988      14.25.1874         False              True     436729
12989      1984.184.1         False              True     472596
12990      1984.184.2         False              True     472597
12991      25.120.363         False             False     473912


            Department          Object Name  \
12982  Arms and Armor             Backplate
12983  Arms and Armor                Helmet
12984  Arms and Armor  Album of bit designs
12985  Arms and Armor           Two stirrups
12986  Arms and Armor               Painting
12987  Arms and Armor               Painting
12988  Arms and Armor               Painting
12989  Arms and Armor            Shield boss
12990  Arms and Armor     Shield boss (Umbo)
12991  Arms and Armor          Lock with key


                                          Title  \
12982  Backplate of an Armor for Vincenzo Luigi di Ca...
12983                            Ceremonial Helmet
12984                           Album of Bit Designs
12985                                 Two Stirrups
12986  Alfonso II d'Este (1533\'961597), Duke of Ferrara
```

```
12987  William H. Riggs (1837\'961924) in Sixteenth-C...
12988   Alfonso I d'Este (1476\'961534), Duke of Ferrara
12989                                       Shield Boss
12990                                 Shield Boss (Umbo)
12991                                     Lock with Key


                 Culture  Period  Dynasty  ...  River  \
12982       Italian, Milan     NaN      NaN  ...    NaN
12983              Chinese     NaN      NaN  ...    NaN
12984             Bohemian     NaN      NaN  ...    NaN
12985   Mongolian or Tibetan   NaN      NaN  ...    NaN
12986      Italian, Ferrara    NaN      NaN  ...    NaN
12987         French, Paris     NaN      NaN  ...    NaN
12988      Italian, Ferrara    NaN      NaN  ...    NaN
12989           Langobardic     NaN      NaN  ...    NaN
12990           Langobardic     NaN      NaN  ...    NaN
12991                French     NaN      NaN  ...    NaN


                  Classification Rights and Reproduction  \
12982         Armor Parts-Backplates                  NaN
12983                  Armor for Man                  NaN
12984          Works on Paper-Prints                  NaN
12985  Equestrian Equipment-Stirrups                  NaN
12986                       Paintings                  NaN
12987                       Paintings                  NaN
12988                       Paintings                  NaN
12989                 Metalwork-Iron                  NaN
12990                 Metalwork-Iron                  NaN
12991                      Metalwork                  NaN


                                           Link Resource  Metadata Date  \
12982  http://www.metmuseum.org/art/collection/search...    42877.33341
12983  http://www.metmuseum.org/art/collection/search...    42877.33341
12984  http://www.metmuseum.org/art/collection/search...    42877.33341
12985  http://www.metmuseum.org/art/collection/search...    42877.33341
12986  http://www.metmuseum.org/art/collection/search...    42877.33341
12987  http://www.metmuseum.org/art/collection/search...    42877.33341
12988  http://www.metmuseum.org/art/collection/search...    42877.33341
12989  http://www.metmuseum.org/art/collection/search...    42877.33341
12990  http://www.metmuseum.org/art/collection/search...    42877.33341
12991  http://www.metmuseum.org/art/collection/search...    42877.33341


                                Repository\ Column1 Column2 Column3  \
12982  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN
12983  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN
12984  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN
12985  Metropolitan Museum of Art, New York, NY\     NaN     NaN     NaN
```

```
12986  Metropolitan Museum of Art, New York, NY\      NaN      NaN      NaN
12987  Metropolitan Museum of Art, New York, NY\      NaN      NaN      NaN
12988  Metropolitan Museum of Art, New York, NY\      NaN      NaN      NaN
12989  Metropolitan Museum of Art, New York, NY\      NaN      NaN      NaN
12990  Metropolitan Museum of Art, New York, NY\      NaN      NaN      NaN
12991  Metropolitan Museum of Art, New York, NY\      NaN      NaN      NaN

       Column4
12982     NaN
12983     NaN
12984     NaN
12985     NaN
12986     NaN
12987     NaN
12988     NaN
12989     NaN
12990     NaN
12991     NaN

[10 rows x 47 columns]
```

[25]:
```
import statistics
```

[26]:
```
a1 = a
```

[36]:
```
a1_col=a[['Object ID']]
```

[ ]:
```
a2_col=a[['Object ID']]
```

[37]:
```
a1_col.mean()
```

[37]:
```
Object ID    28957.044797
dtype: float64
```

1

[39]:
```
a1_col.median
```

[39]:
```
<bound method DataFrame.median of        Object ID
0            21810
1            21811
2            21812
3            21813
4            21814
5            21815
6            21816
7            21817
8            21818
9            21819
```

```
10          21820
11          21821
12          21822
13          21823
14          21824
15          21825
16          21826
17          21827
18          21828
19          21829
20          21830
21          21831
22          21832
23          21833
24          21834
25          21835
26          21836
27          21837
28          21839
29          21840
...         ...
12962       35934
12963       35935
12964       35936
12965       35937
12966       35938
12967       35939
12968       35942
12969       35944
12970       35945
12971       35947
12972       35948
12973       35949
12974       35952
12975       35953
12976       35954
12977       35955
12978       35956
12979       35957
12980       35958
12981       35959
12982       35960
12983       35961
12984       35962
12985       35963
12986       436343
12987       436699
```

```
12988      436729
12989      472596
12990      472597
12991      473912

[12992 rows x 1 columns]>
```

1.1

1.2

```
[40]: a2 = [['Culture', 'Object ID']]
```

```
[42]: a2
```

```
[42]: [['Culture', 'Object ID']]
```

```
[45]: a2 = a1[['Culture', 'Object ID']]
```

1.3

```
[46]: a2.head(10)
```

```
[46]:          Culture  Object ID
      0  Anglo-American      21810
      1        Japanese      21811
      2          French      21812
      3        Japanese      21813
      4        Japanese      21814
      5        Japanese      21815
      6        Japanese      21816
      7        Japanese      21817
      8        Japanese      21818
      9        Japanese      21819
```

```
[47]: a2.groupby(['Object ID']).size()
```

```
[47]: Object ID
      21810    1
      21811    1
      21812    1
      21813    1
      21814    1
      21815    1
      21816    1
      21817    1
      21818    1
      21819    1
      21820    1
      21821    1
```

```
21822    1
21823    1
21824    1
21825    1
21826    1
21827    1
21828    1
21829    1
21830    1
21831    1
21832    1
21833    1
21834    1
21835    1
21836    1
21837    1
21839    1
21840    1
         ..
35934    1
35935    1
35936    1
35937    1
35938    1
35939    1
35942    1
35944    1
35945    1
35947    1
35948    1
35949    1
35952    1
35953    1
35954    1
35955    1
35956    1
35957    1
35958    1
35959    1
35960    1
35961    1
35962    1
35963    1
436343   1
436699   1
436729   1
472596   1
```

```
472597    1
473912    1
Length: 12992, dtype: int64
```

1.4

```
[48]: a2.isnull().sum()
```

```
[48]: Culture      25
      Object ID     0
      dtype: int64
```

```
[49]: a1.isnull().sum()
```

```
[49]: Object Number            0
      Is Highlight             0
      Is Public Domain         0
      Object ID                0
      Department               0
      Object Name              0
      Title                   13
      Culture                 25
      Period               12992
      Dynasty              12992
      Reign                12992
      Portfolio            12992
      Artist Role          12271
      Artist Prefix        12692
      Artist Display Name  12271
      Artist Display Bio   12289
      Artist Suffix        12976
      Artist Alpha Sort    12271
      Artist Nationality   12351
      Artist Begin Date    12333
      Artist End Date      12323
      Object Date             32
      Object Begin Date       26
      Object End Date         26
      Medium                 473
      Dimensions             952
      Credit Line            450
      Geography Type       12992
      City                 11920
      State                12111
      County               12992
      Country              12992
      Region               12992
      Subregion            12992
```

```
Locale                   12992
Locus                    12992
Excavation               12992
River                    12992
Classification             673
Rights and Reproduction  12992
Link Resource              673
Metadata Date              673
Repository\                 673
Column1                  12992
Column2                  12992
Column3                  12992
Column4                  12992
dtype: int64
```

[ ]: `a2['yr']=`