01_convert_types

March 21, 2022

0.1 Learning Data Wrangling in Python using pandas

Abid Ali
https://github.com/abid-2362

[ ]:
```
import pandas as pd

data = pd.read_csv("./data/artwork_sample.csv")
```

[ ]:
```
data.head()
```

[ ]:
```
     id accession_number          artist artistRole  artistId  \
0  1035           A00001   Blake, Robert     artist        38
1  1036           A00002   Blake, Robert     artist        38
2  1037           A00003   Blake, Robert     artist        38
3  1038           A00004   Blake, Robert     artist        38
4  1039           A00005  Blake, William     artist        39

                                               title               dateText  \
0  A Figure Bowing before a Seated Old Man with h…          date not known
1  Two Drawings of Frightened Figures, Probably f…          date not known
2  The Preaching of Warning. Verso: An Old Man En…                 ?c.1785
3     Six Drawings of Figures with Outstretched Arms          date not known
4  The Circle of the Lustful: Francesca da Rimini…  1826-7, reprinted 1892

                                            medium  \
0  Watercolour, ink, chalk and graphite on paper…
1                                Graphite on paper
2       Graphite on paper. Verso: graphite on paper
3                                Graphite on paper
4                          Line engraving on paper

                                      creditLine    year  acquisitionYear  \
0               Presented by Mrs John Richmond 1922     NaN             1922
1               Presented by Mrs John Richmond 1922     NaN             1922
2               Presented by Mrs John Richmond 1922  1785.0             1922
3               Presented by Mrs John Richmond 1922     NaN             1922
4  Purchased with the assistance of a special gra…  1826.0             1919
```

```
              dimensions  width  height  depth units  inscription  \
0  support: 394 x 419 mm  394.0   419.0    NaN    mm          NaN
1  support: 311 x 213 mm  311.0   213.0    NaN    mm          NaN
2  support: 343 x 467 mm  343.0   467.0    NaN    mm          NaN
3  support: 318 x 394 mm  318.0   394.0    NaN    mm          NaN
4    image: 243 x 335 mm  243.0   335.0    NaN    mm          NaN

   thumbnailCopyright                                    thumbnailUrl  \
0                 NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
1                 NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
2                 NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
3                 NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
4                 NaN  http://www.tate.org.uk/art/images/work/A/A00/A…

                                                 url
0  http://www.tate.org.uk/art/artworks/blake-a-fi…
1  http://www.tate.org.uk/art/artworks/blake-two-…
2  http://www.tate.org.uk/art/artworks/blake-the-…
3  http://www.tate.org.uk/art/artworks/blake-six-…
4  http://www.tate.org.uk/art/artworks/blake-the-…
```

[ ]:
```
data.dtypes
```

[ ]:
```
id                    int64
accession_number     object
artist               object
artistRole           object
artistId              int64
title                object
dateText             object
medium               object
creditLine           object
year                float64
acquisitionYear       int64
dimensions           object
width               float64
height              float64
depth               float64
units                object
inscription         float64
thumbnailCopyright  float64
thumbnailUrl         object
url                  object
dtype: object
```

[ ]:
```
data.acquisitionYear
```

[ ]:
```
0    1922
1    1922
2    1922
3    1922
4    1919
5    1919
6    1919
7    1919
8    1919
9    1919
Name: acquisitionYear, dtype: int64
```

[ ]:
```
# Changing data type using astype
# Note that this does not change the data type of column in place
data.acquisitionYear.astype(float)
```

[ ]:
```
0    1922.0
1    1922.0
2    1922.0
3    1922.0
4    1919.0
5    1919.0
6    1919.0
7    1919.0
8    1919.0
9    1919.0
Name: acquisitionYear, dtype: float64
```

[ ]:
```
# Still int type because astype does not change the column type in place
# rather it returns the brand new series of the new type
data.acquisitionYear.dtype
```

[ ]:
```
dtype('int64')
```

[ ]:
```
# replace column with new data type
data.acquisitionYear = data.acquisitionYear.astype(float)
data.acquisitionYear.dtype
```

[ ]:
```
dtype('float64')
```

[ ]:
```
fulldf = pd.read_csv("./data/artwork_data.csv", low_memory=False)
```

[ ]:
```
fulldf.head()
```

[ ]:
```
     id accession_number          artist artistRole  artistId  \
0  1035           A00001   Blake, Robert     artist        38
1  1036           A00002   Blake, Robert     artist        38
2  1037           A00003   Blake, Robert     artist        38
```

```
3  1038           A00004   Blake, Robert     artist        38
4  1039           A00005  Blake, William     artist        39

                                                 title               dateText  \
0  A Figure Bowing before a Seated Old Man with h…          date not known
1  Two Drawings of Frightened Figures, Probably f…          date not known
2  The Preaching of Warning. Verso: An Old Man En…                 ?c.1785
3     Six Drawings of Figures with Outstretched Arms           date not known
4  The Circle of the Lustful: Francesca da Rimini…  1826-7, reprinted 1892

                                              medium  \
0  Watercolour, ink, chalk and graphite on paper…
1                                Graphite on paper
2       Graphite on paper. Verso: graphite on paper
3                                Graphite on paper
4                           Line engraving on paper

                                      creditLine  year  acquisitionYear  \
0               Presented by Mrs John Richmond 1922   NaN           1922.0
1               Presented by Mrs John Richmond 1922   NaN           1922.0
2               Presented by Mrs John Richmond 1922  1785           1922.0
3               Presented by Mrs John Richmond 1922   NaN           1922.0
4  Purchased with the assistance of a special gra…  1826          1919.0

             dimensions width height  depth units inscription  \
0  support: 394 x 419 mm   394    419    NaN    mm         NaN
1  support: 311 x 213 mm   311    213    NaN    mm         NaN
2  support: 343 x 467 mm   343    467    NaN    mm         NaN
3  support: 318 x 394 mm   318    394    NaN    mm         NaN
4    image: 243 x 335 mm   243    335    NaN    mm         NaN

  thumbnailCopyright                                      thumbnailUrl  \
0               NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
1               NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
2               NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
3               NaN  http://www.tate.org.uk/art/images/work/A/A00/A…
4               NaN  http://www.tate.org.uk/art/images/work/A/A00/A…

                                              url
0  http://www.tate.org.uk/art/artworks/blake-a-fi…
1  http://www.tate.org.uk/art/artworks/blake-two-…
2  http://www.tate.org.uk/art/artworks/blake-the-…
3  http://www.tate.org.uk/art/artworks/blake-six-…
4  http://www.tate.org.uk/art/artworks/blake-the-…
```

[ ]:
```
fulldf.dtypes
```

[ ]:
```
id                      int64
accession_number       object
artist                 object
artistRole             object
artistId                int64
title                  object
dateText               object
medium                 object
creditLine             object
year                   object
acquisitionYear       float64
dimensions             object
width                  object
height                 object
depth                 float64
units                  object
inscription            object
thumbnailCopyright     object
thumbnailUrl           object
url                    object
dtype: object
```

[ ]:
```
fulldf.height.astype(float)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Input In [26], in <cell line: 1>()
----> 1 fulldf.height.astype(float)

File C:\Python310\lib\site-packages\pandas\core\generic.py:5920, in NDFrame.
 ↪astype(self, dtype, copy, errors)
   5913     results = [
   5914         self.iloc[:, i].astype(dtype, copy=copy)
   5915         for i in range(len(self.columns))
   5916     ]
   5918 else:
   5919     # else, only a single dtype is given
-> 5920     new_data = self._mgr.astype(dtype=dtype, copy=copy, errors=errors)
   5921     return self._constructor(new_data).__finalize__(self,␣
 ↪method="astype")
   5923 # GH 33113: handle empty frame or series

File C:\Python310\lib\site-packages\pandas\core\internals\managers.py:419, in␣
 ↪BaseBlockManager.astype(self, dtype, copy, errors)
    418 def astype(self: T, dtype, copy: bool = False, errors: str = "raise") -␣
 ↪T:
--> 419     return self.apply("astype", dtype=dtype, copy=copy, errors=errors)
```

```
File C:\Python310\lib\site-packages\pandas\core\internals\managers.py:304, in␣
 ↪BaseBlockManager.apply(self, f, align_keys, ignore_failures, **kwargs)
    302         applied = b.apply(f, **kwargs)
    303     else:
--> 304         applied = getattr(b, f)(**kwargs)
    305 except (TypeError, NotImplementedError):
    306     if not ignore_failures:

File C:\Python310\lib\site-packages\pandas\core\internals\blocks.py:580, in␣
 ↪Block.astype(self, dtype, copy, errors)
    562 """
    563 Coerce to the new dtype.
    564
   (…)
    576 Block
    577 """
    578 values = self.values
--> 580 new_values = astype_array_safe(values, dtype, copy=copy, errors=errors)
    582 new_values = maybe_coerce_values(new_values)
    583 newb = self.make_block(new_values)

File C:\Python310\lib\site-packages\pandas\core\dtypes\cast.py:1292, in␣
 ↪astype_array_safe(values, dtype, copy, errors)
   1289     dtype = dtype.numpy_dtype
   1291 try:
-> 1292     new_values = astype_array(values, dtype, copy=copy)
   1293 except (ValueError, TypeError):
   1294     # e.g. astype_nansafe can fail on object-dtype of strings
   1295     #  trying to convert to float
   1296     if errors == "ignore":

File C:\Python310\lib\site-packages\pandas\core\dtypes\cast.py:1237, in␣
 ↪astype_array(values, dtype, copy)
   1234     values = values.astype(dtype, copy=copy)
   1236 else:
-> 1237     values = astype_nansafe(values, dtype, copy=copy)
   1239 # in pandas we don't store numpy str dtypes, so convert to object
   1240 if isinstance(dtype, np.dtype) and issubclass(values.dtype.type, str):

File C:\Python310\lib\site-packages\pandas\core\dtypes\cast.py:1181, in␣
 ↪astype_nansafe(arr, dtype, copy, skipna)
   1177     raise ValueError(msg)
   1179 if copy or is_object_dtype(arr.dtype) or is_object_dtype(dtype):
   1180     # Explicit copy, or required since NumPy can't view from / to objec
-> 1181     return arr.astype(dtype, copy=True)
   1183 return arr.astype(dtype, copy=copy)
```

```
ValueError: could not convert string to float: 'mm'
```

[ ]:
```
pd.to_numeric(fulldf.height)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
File C:\Python310\lib\site-packages\pandas\_libs\lib.pyx:2315, in pandas._libs.
 ↪lib.maybe_convert_numeric()

ValueError: Unable to parse string "mm"

During handling of the above exception, another exception occurred:

ValueError                                Traceback (most recent call last)
Input In [27], in <cell line: 1>()
----> 1 pd.to_numeric(fulldf.height)

File C:\Python310\lib\site-packages\pandas\core\tools\numeric.py:184, in␣
 ↪to_numeric(arg, errors, downcast)
    182 coerce_numeric = errors not in ("ignore", "raise")
    183 try:
--> 184     values, _ = lib.maybe_convert_numeric(
    185         values, set(), coerce_numeric=coerce_numeric
    186     )
    187 except (ValueError, TypeError):
    188     if errors == "raise":

File C:\Python310\lib\site-packages\pandas\_libs\lib.pyx:2357, in pandas._libs.
 ↪lib.maybe_convert_numeric()

ValueError: Unable to parse string "mm" at position 41339
```

[ ]:
```
pd.to_numeric(fulldf.height, errors="coerce")
```

[ ]:
```
0         419.0
1         213.0
2         467.0
3         394.0
4         335.0
          ...
69196     305.0
69197     305.0
69198    2410.0
69199       NaN
69200     660.0
```

```
Name: height, Length: 69201, dtype: float64
```

[ ]:
```
# errors="coerce" has converted into numbers, and errors becomes nan
pd.to_numeric(fulldf.height, errors="coerce")[41339]
```

[ ]:
```
nan
```

[ ]:
```
fulldf.height = pd.to_numeric(fulldf.height, errors="coerce")
```

[ ]:
```
fulldf.height.dtypes
```

[ ]:
```
dtype('float64')
```

[ ]:
```

```