cleaning

May 13, 2022

[ ]:
```
%load_ext autoreload
%autoreload 2
```

[ ]:
```
import pandas as pd
import os
from cleaning_utils import *
import numpy as np
```

1 Read in data file

[ ]:
```
cwd = os.getcwd()
df = pd.read_csv(os.path.join(cwd, "MetObjects.txt"), low_memory=False)
```

2 Example 1 the "Dynasty" column

[ ]:
```
df["Dynasty"].unique()
```

[ ]:
```
array([nan, 'Dynasty 26 or later', 'Late Dynasty 12-13', 'Dynasty 26',
       'Dynasty 18', 'late Dynasty 12 to Dynasty 13', 'Dynasty 13',
       'Dynasty 21', 'Dynasty 19', 'Dynasty 19-20', 'mid-Dynasty 18',
       'Dynasty 12-Dynasty 17', 'Dynasty 12-13', 'late Dynasty 18',
       'Dynasty 4', 'likely Dynasty 19', 'Early Dynasty 18', 'Dynasty 25',
       'Dynasty 22-23', 'Dynasty 5', 'mid-late Dynasty 18',
       'Dynasty 19-22', 'Early Dynasty 12', 'Dynasty 12',
       'Dynasty 11-early 12', 'late Dynasty 12-Dynasty 13',
       'late Dynasty 18-early 19', 'first half of Dynasty 12',
       'late Dynasty 18-early Dynasty 19', 'late Dynasty 12',
       'Dynasty 30', 'Dynasty 11-12', 'Dynasty 20-21',
       'Probably Dynasty 30', 'Dynasty 22-26', 'Dynasty 18-early 19',
       'mid-or late Dynasty 26', 'Dynasty 19 (?)', 'Dynasty 8-11',
       'early Dynasty 18', 'Dynasty 1', 'Dynasty 18, early', 'Dynasty 6',
       'Dynasty 4-5', 'Dynasty 5-6', 'Dynasty 4-mid 5',
       'Dynasty 4, mid-5th Dynasty', 'Dynasty 3', 'Dynasty 6, first half',
       'Dynasty 6-8', 'Dynasty 6-8', 'Dynasty 4-6', 'early Dynasty 3',
       'Dynasty 11-12', 'Dynasty 27', 'Dynasty 12-17',
```

'Dynasty 17-Early Dynasty 18', 'Dynasty 17-18',
'late Dynasty 12-early 13', 'late Dynasty 12-early Dynasty 13',
'Dynasty 12, late - early 13', 'Dynasty 11',
'Dynasty 12, late-early 13',
'Dynasty 12, late-13 up to 1700 B. C.', 'Dynasty 13-17',
'late Dynasty 12-early Dynasty 13',
'Dynasty 12, late-13 up to 1700 B. C.', 'Dynasty 12, late-13',
'Dynasty 9 or later', 'Dynasty 9-10', 'Dynasty 9-11', 'Dynasty 17',
'Second Intermediate Period', 'Dynasty 12, late - 13 up to 1700',
'Second Intermediate Period- Dynasty 18, early',
'Dynasty 12, early', 'Dynasty 18-19', 'Dynasty 15',
'Dynasty 26-30', 'Dynasty 20', 'Dynasty 26-29', 'Dynasty 12-18',
'Dynasty 25-26', 'Dynasty 18-20', 'Dynasty 15-17',
'Late Dynasty 18', 'Dynasty 12, late-13 to 1700 B.C.',
'Dynasty 12, mid-late', 'Dynasty 11, late-12, early',
'mid-Dynasty 12', 'late Dynasty 12- Dynasty 13', 'Dynasty 12 (?)',
'first part of Dynasty 13', 'Mid-Dynasty 12 or later',
'Late Dynasty 13-early Dynasty 17', 'Dynasty 9', 'Dynasty 10-11',
'Dynasty 11, late-12', 'Dynasty 11-13',
'Late Dynasty 12-early Dynasty 13',
'Dynasty 12, late - Second Intermediate Period',
'Dynasty 12, mid - Dynasty 13', 'mid to end Dynasty 12',
'Dynasty 13, mid', 'Late Dynasty 11 - Dynasty 12',
'early to mid-Dynasty 12', 'Dynasty 11-18', 'Dynasty 9-12',
'Dynasty 12-14', 'Probably Dynasty 12', 'Dynasty 14',
'Late 12 to mid-13 Dynasty', 'Dynasty 13-SIP',
'end of Dynasty 11 - beginning of Dynasty 12', 'Dynasty 12, mid',
'mid to late Dynasty 13', 'Dynasty 13-15', 'mid Dynasty 13',
'Dynasty 6-11', '8-11', 'Dynasty 8', 'Dynasty 12, late-17',
'late Dynasty 12-Dynasty 17', 'Dynasty 16', 'Dynasty 14-15',
'Late Dynasty 17-Early Dynasty 18', 'Dynasty 18, second half',
'mid Dynasty 18', 'Dynasty 18, late', 'original Dynasty 18',
'Original New Kingdom, Dynasty 18', 'Dynasty 15, late',
'Dynasty 15-16', 'Dynasty 22, early', 'Dynasty 19-20',
'Dynasty 19-20 or later', 'Dynasty 17-early Dynasty 18',
'Dynasty 13-18', 'Dynasty 21-25', 'Dynasty 12-early Dynasty 18',
'late Dynasty 13-early dynasty 18',
'late Dynasty 17 - early Dynasty 18', 'Dynasty 22',
'Dynasty 25 (Kushite)', 'Dynasty 21-26', 'Dynasty 25-30',
'Dynasty 12, late', 'Dynasty 13, possibly 15, early',
'Dynasty 12 probably', 'Dynasty 13 to 1700 B.C.',
'Dynasty 13, late-15', 'Dynasty 12, late-13 or later',
'late Dynasty 12 to early Dynasty 13', 'Dynasty 13, late',
'End of Dynasty 12-early Dynasty 13', 'Dynasty 11 or earlier',
'Dynasty 12, early-mid', 'mid-Dynasty 13', 'late Dynasty 12-13',
'Dynasty 19-20 (Ramesside)', 'early Dynasty 13',
'mid to late Dynasty 18', 'early to mid Dynasty 12',

'early Dynasty 12', 'Dynasty 20-22', 'Dynasty 11, late',
'Late Dynasty 11', 'Late dynasty 11', 'Dynasty 11 or later',
'late Dynasty 13-17', 'Dynasty 15-18', 'Dynasty 14-17',
'Dynasty 27-30', 'Dynasty 19-26', 'Dynasty 18-21',
'Dynasty 19 (Ramesside)', 'Dynasty 12-13, late',
'Dynasty 12 to early Dynasty 13', 'Dynasty 17-early 18',
'Dynasty 5-18 (?)', 'Dynasty 2-3', 'Dynasty 3-4',
'Dynasty 13-18, early', 'Dynasty 2', 'Dynasty 2, second half',
'Dynasty 11, early', 'Dynasty 19, early', 'Dynasty 25?',
'Dynasty 12, late -13', 'Dynasty 25-early 26',
'Dynasty 30 or later', 'Dynasty 22-24', 'Dynasty 26 ?',
'Ptolemaic Dynasty', 'Dynasty 23', 'Dynasty 21-24',
'Dynasty 20-25', 'Dynasty 13?', 'Dynasties 18-20',
'Dynasty 12, late - 13, early', 'Dynasty 12, late -13, early',
'Dynasty 12, early - mid', 'Late Dynasty 12',
'Dynasty 13, late-18, early', 'Dynasty 13 or later',
'Dynasty 13-2nd Intermediate Period', 'Dynasty 12- 13',
'Dynasty 12-13', 'Dynasty 12 mid-13', 'Dynasty 13-18',
'Dynasty 15-18, early', 'Dynasty 12, mid-13', 'Dynasty 12-13 ?',
'Dynasty 20-22', 'Dynasty 6 (?)', 'Dynasty 12, mid- Dynasy 13',
'Dynasty 12,mid-Dynasty 13', 'Dynasty 11-early 18',
'Dynasty 12-Early Dynasty 18', 'late Dynasty 17-early Dynasty 18',
'Dynasty 0-1', 'Probably Dynasty 1', 'Dynasty 0-2',
'Dynasty 2 probably', 'Dynasty 26-4th century',
'probably mid-Dynasty 18', 'Dynasty 18 or 21-22',
'Dynasty 25-early Dynasty 26', 'Dynasty 26, mid to late',
'Dynasty 21-30', 'Dynasty 5-8', 'Dynasty 11-mid 12',
'Dynasty 21, second half', 'late Dynasty 21', 'Dynasty 22-26',
'Dynasty 26-27', 'Dynasty 26-28', 'Dynasty 22-25',
'original Dynasty 19', 'Dynasty 18, early to mid', 'Dynasty 6-9',
'Dynasty 18, mid', 'Dynasty 18, Reign of Amenhotep III',
'Dynasty 18, early, probably', 'Dynasty 30 and later',
'Dynasty 26 and later', 'Dynasty 20 (?)', 'Dynasty 21 (?)',
'Dynasty 18 or later', 'Dynasty 21-26 (?)',
'Late Dynasty 18-early Dynasty 19',
'Late Dynasty 18-early Dynasty 19', 'Dynasty 23 (northern)',
'late Dynasty 18-19', 'Dynasty 27 or later', 'early dynasty 4-5',
'Dynasty 21, late', 'Dynasty 19-30', 'mid to late Dynasty 12',
'Dynasty 1-2', 'Late 18 Dynasty - Early 19', 'Late Dynasty 18-19',
'Dynasty 26 (Saite)?', 'Dynasty 21-22', 'late Dynasty 22',
'Dynasty 27 (Persian)?', 'Dynasty 29', 'Probably Dynasty 26',
'Dynasty 25-29', 'Dynasty 19-25', 'second half of Dynasty 26',
'Dynasty 30-2nd Persian Period', 'Dynasty 5 (?)', '6',
'Dynasty 6 or later', 'Dynasty 8-12', 'Dynasty 3-9',
'Late Dynasty 19', 'Dynasty 12, late-Dynasty 17',
'Dynasty 12, late-17', 'Dynasty 5, second half',
'Dynasty 20-22 (?)', 'Dynasty 12, beginning',

```
'Dynasty 22 / Dynasty 25', 'Dynasty 23-26', 'Dynasty 23-30',
'Dynasty 24-26', 'Ptolemaic', 'Late Dynasty 21-early Dynasty 22',
'early Dynasty 21', 'Dynasty 24', 'Dynasty 20-26',
'Dynasty 26-30 and later',
'Dynasty 12-18; reused 8th century B.C.', 'Dynasty 27-30?',
'Dynasty 26 (Saite)', 'Dynasty 18-19',
'Dynasty 21-early Dynasty 22', 'Dynasty 18 ?', 'Dynasty 21-26 ?',
'Dynasty 21-27', 'Dynasty 21-25 ?', 'Dynasty 19-25 ?',
'Dynasty 25 ?', 'Dynasty 19-21', 'Dynasty 18-20 ?',
'Dynasty 18-25', 'Dynasty 21?', 'Dynasty 21 to early Dynasty 22',
'Dynasty 21 or 22', 'Dynasty 21-23', 'Dynasty 18, possibly later',
'Dynasty 20-24', 'Dynasty 26-Ptolemaic Period',
'Dynasty 25 or later', 'Dynasty 18 (?)', 'Dynasties 19-20',
'Dynasty 18, reign of Amenhotep III', 'Dynasty 17-18',
'Dynasty 18, first half', 'Dynasty 22-25',
'Dynasty 18, late-Dynasty 19', 'Dynasty 19-20 or later (?)',
'Dynasty 13-17', 'Dynasty 11-17', 'Dynasty 11 or 12',
'Dynasty 13 probably', 'Dynasty 8-11',
'late Dynasty 12 - Dynasty 13', 'Dynasty 13-14', 'Dynasty 13(?)',
'Dynasty 12, late-13, early', 'Dynasty 12, mid to second half',
'Dynasty 12-17 or later', 'late Dynasty 13-17', 'mid Dynasty 12',
'Dynasty 12, late-early 18 or later', 'Dynasty 11?',
'Dynasty 18 or later (?)', 'Dynasty 12-15', 'mid Dynasty 12-13',
'Dynasty 13, reign of Khendjer', 'Mid Dynasty 13',
'late Dynasty 12-13', 'late Dynasty 11-early Dynasty 12',
'Dynasty 19-20 or 21-25', 'Dynasty 12-2nd Intermediate Period',
'Dynasty 12, late - early 13', 'mid to late dynasty 13',
'Dynasty 12, early-to Senwosret II', 'early Dynasty 19',
'Dynasty 12-13, early', 'Dynasty 12-20', 'late 18th through 20th',
'Dynasty 1, 3-4', '1', 'Dynasty 18 (Amarna Period)', 'Dynasty  13',
'Dynasty 12-Dynasty 13', 'late Dynasty 13-early Dynasty 18',
'late Dynasty 20 - early Dynasty 21', 'Dynasty 3-5', 'Dynasty 2-5',
'Dynasty 7-8', 'Dynasty 20 (Ramesside)',
'Dynasty 9-early Dnyasty 11', 'Dynasty 9-early Dynasty 11',
'mid Dynasty 18-20', 'Naqada III-Dynasty 1', 'Naqada II-Dynasty 1',
'Dynasty 12, late - Dynasty 13, early', 'Dynasty 18-20',
'Dynasty 19-20 (?)', 'Dynasty 5 or 6', 'Dynasty 0-2',
'Probably late Dynasty 5', 'Dynasty 6-12', 'Dynasty 8-9',
'Dynasty 7-10 (?)', 'Dynasty 0', 'Dynasty 8-18', '9 or later',
'Dynasty 11 (?)', 'Dynasty 7-10', 'Dyna 19-20 or later',
'Dyn. 19 - 20 or later', 'Dynasty 18, late, or early 19',
'Dynasty 5, end', 'Dynasty 9?', 'Dynasty 26 (?)',
'Dynasty 22-23 or earlier', '26', 'Dynasty 26-30', 'Dynasty 14-22',
'mid-Dynasty 20', 'Reign of Amenhotep I', 'Dynasty 5-6', '11',
'19-20', 'mid Dynasty 21', '18', 'Dynasty 18, late or early 19',
'12', 'Dynasty 3-4'], dtype=object)
```

```
[ ]: df[~df["Dynasty"].str.contains("\d\d", regex=True).map(lambda x : False if␣
      ↪(type(x)==float) else x)]["Dynasty"].unique()
```

```
[ ]: array([nan, 'Dynasty 4', 'Dynasty 5', 'Dynasty 1', 'Dynasty 6',
            'Dynasty 4-5', 'Dynasty 5-6', 'Dynasty 4-mid 5',
            'Dynasty 4, mid-5th Dynasty', 'Dynasty 3', 'Dynasty 6, first half',
            'Dynasty 6-8', 'Dynasty 6-8', 'Dynasty 4-6', 'early Dynasty 3',
            'Dynasty 9 or later', 'Second Intermediate Period', 'Dynasty 9',
            'Dynasty 8', 'Dynasty 2-3', 'Dynasty 3-4', 'Dynasty 2',
            'Dynasty 2, second half', 'Ptolemaic Dynasty', 'Dynasty 6 (?)',
            'Dynasty 0-1', 'Probably Dynasty 1', 'Dynasty 0-2',
            'Dynasty 2 probably', 'Dynasty 5-8', 'Dynasty 6-9',
            'early dynasty 4-5', 'Dynasty 1-2', 'Dynasty 5 (?)', '6',
            'Dynasty 6 or later', 'Dynasty 3-9', 'Dynasty 5, second half',
            'Ptolemaic', 'Dynasty 1, 3-4', '1', 'Dynasty 3-5', 'Dynasty 2-5',
            'Dynasty 7-8', 'Naqada III-Dynasty 1', 'Naqada II-Dynasty 1',
            'Dynasty 5 or 6', 'Dynasty 0-2', 'Probably late Dynasty 5',
            'Dynasty 8-9', 'Dynasty 0', '9 or later', 'Dynasty 5, end',
            'Dynasty 9?', 'Reign of Amenhotep I', 'Dynasty 5-6', 'Dynasty 3-4'],
           dtype=object)
```

```
[ ]: df["Dynasty_clean"] = df["Dynasty"].str.extract("(\d\d)")
     df[~df["Dynasty_clean"].str.contains("\d\d", regex=True).map(lambda x : False␣
      ↪if (type(x)==float) else x)]["Dynasty_clean"].unique()
```

```
[ ]: array([nan], dtype=object)
```

3 Example 2 the 'End Date' column

```
[ ]: df["Artist End Date"].unique()
```

```
[ ]: array(['1869      ', '1844      ', nan, …, '1808      |1809      ',
            '1952      |1952      ',
            '1942      |1954      |          |          '], dtype=object)
```

```
[ ]: df[~df["Artist End Date"].str.contains("(-?\d\d\d\d\|?){1,10}", regex=True).
      ↪map(lambda x : False if (type(x)==float) else x)]["Artist End Date"].unique()
```

```
C:\Users\mschm\AppData\Local\Temp\ipykernel_15812\1923886761.py:1: UserWarning:
This pattern is interpreted as a regular expression, and has match groups. To
actually get the groups, use str.extract.
  df[~df["Artist End Date"].str.contains("(-?\d\d\d\d\|?){1,10}",
regex=True).map(lambda x : False if (type(x)==float) else x)]["Artist End
Date"].unique()
```

[ ]:
```
array([nan, '          ', '          |          ',
       '          |          |          ',
       '          |          |          |          ',
       '          |          |          |          |          ',
       '          |          |          |          |          |          |
',
       '          |          |          |          |          |          ',
       '0         ', '          |0         ', '0         |          ',
       '0         |0         ',
       '          |          |          |          |          |          |
|          '],
      dtype=object)
```

[ ]:
```
df["Artist End Date_clean"] = df["Artist End Date"].str.extract("(-?\d\d\d\d\|?
↪){1,10}")
df[~df["Artist End Date_clean"].str.contains("(-?\d\d\d\d\|?){1,10}",␣
↪regex=True).map(lambda x : False if (type(x)==float) else x)]["Artist End␣
↪Date_clean"].unique()
```

```
C:\Users\mschm\AppData\Local\Temp\ipykernel_15812\781919347.py:2: UserWarning:
This pattern is interpreted as a regular expression, and has match groups. To
actually get the groups, use str.extract.
  df[~df["Artist End Date_clean"].str.contains("(-?\d\d\d\d\|?){1,10}",
regex=True).map(lambda x : False if (type(x)==float) else x)]["Artist End
Date_clean"].unique()
```

[ ]:
```
array([nan], dtype=object)
```