

DSC540-T302 Data Preparation

Weeks 7 & 8: Data Cleaning and Transforming

Saravanan Janarthanan

Assignment

```
In [1]: # import the modules
import pandas as pd
import numpy as np
```

```
In [2]: # Load the meta objects csv file
metObj_df = pd.read_csv("MetObjects_data.csv", encoding='utf-8', encoding_errors='rep']
```

```
In [3]: # print the shape to find number of rows and number of columns
metObj_df.shape
```

```
Out[3]: (12108, 44)
```

There are 12108 rows and 44 columns

```
In [4]: # Find if there rows with no data and if so drop them
filtered_metObj_df = metObj_df[metObj_df.notna()]
```

```
In [5]: filtered_metObj_df.shape
```

```
Out[5]: (12108, 44)
```

```
In [6]: # List number records null in each column
filtered_metObj_df.isnull().sum()
```

```
Out[6]: Object Number      21
         Is Highlight       512
         Is Public Domain    591
         Object ID           705
         Department          1240
         Object Name          1254
         Title                633
         Culture              2975
         Period               12097
         Dynasty              12106
         Reign                12106
         Portfolio             12107
         Artist Role           6573
         Artist Prefix          10130
         Artist Display Name    6556
         Artist Display Bio     6241
         Artist Suffix          11373
         Artist Alpha Sort      5841
         Artist Nationality     9821
         Artist Begin Date      6950
         Artist End Date         6612
         Object Date            1467
         Object Begin Date      1278
         Object End Date         1279
         Medium                1279
         Dimensions             1332
         Credit Line             2025
         Geography Type          4788
         City                  8110
         State                 11454
         County                11453
         Country                4792
         Region                 11752
         Subregion              12108
         Locale                 12107
         Locus                  12108
         Excavation              12107
         River                  12108
         Classification          2028
         Rights and Reproduction 12037
         Link Resource            2029
         Metadata Date            3428
         Repository              2029
         Tags                   8243
dtype: int64
```

```
In [7]: filtered_metObj_df.head()
```

Out[7]:

	Object Number	Is Highlight	Is Public Domain	Object ID	Department	Object Name	Title	Culture	Period	Dynasty	...
0	1979.486.1	False	False	1	The American Wing	Coin	One-dollar Liberty Head Coin	NaN	NaN	NaN	...
1	1980.264.5	False	False	2	The American Wing	Coin	Ten-dollar Liberty Head Coin	NaN	NaN	NaN	...
2	67.265.9	False	False	3	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	NaN	NaN	...
3	67.265.10	False	False	4	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	NaN	NaN	...
4	67.265.11	False	False	5	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	NaN	NaN	...

5 rows × 44 columns

In [8]:

```
# Drop the columns that has nulls more than 3500 rows
filtered_metObj_df = metObj_df.dropna(axis=1, thresh=3500)
```

In [9]:

```
filtered_metObj_df.shape
```

Out[9]:

```
(12108, 28)
```

In [10]:

```
filtered_metObj_df.isnull().sum()
```

```
Out[10]: Object Number      21
          Is Highlight      512
          Is Public Domain    591
          Object ID           705
          Department          1240
          Object Name          1254
          Title                633
          Culture              2975
          Artist Role          6573
          Artist Display Name   6556
          Artist Display Bio    6241
          Artist Alpha Sort     5841
          Artist Begin Date     6950
          Artist End Date       6612
          Object Date           1467
          Object Begin Date     1278
          Object End Date       1279
          Medium                1279
          Dimensions            1332
          Credit Line            2025
          Geography Type        4788
          City                  8110
          Country                4792
          Classification         2028
          Link Resource          2029
          Metadata Date          3428
          Repository             2029
          Tags                  8243
          dtype: int64
```

Chapter 7

Fill in missing data

Find the rows for each column and fill the values with previous row

```
In [11]: # Forward fill the empty rows with a previous row value
fill_missd_data_metObj_df = filtered_metObj_df.fillna(method="ffill")
```

```
In [12]: # Validate the column null value count
fill_missd_data_metObj_df.isnull().sum()
```

```
Out[12]: Object Number      0  
Is Highlight        0  
Is Public Domain    0  
Object ID           0  
Department          0  
Object Name         0  
Title               0  
Culture             14  
Artist Role         0  
Artist Display Name 0  
Artist Display Bio   0  
Artist Alpha Sort    0  
Artist Begin Date    0  
Artist End Date      0  
Object Date          0  
Object Begin Date    0  
Object End Date      0  
Medium              0  
Dimensions          0  
Credit Line          0  
Geography Type      14  
City                32  
Country              14  
Classification        0  
Link Resource         0  
Metadata Date        6  
Repository           0  
Tags                 6  
dtype: int64
```

```
In [13]: # filter the rows with City column having null or NaN value  
null_df = fill_missd_data_metObj_df[fill_missd_data_metObj_df['City'].isnull()]
```

```
In [14]: null_df
```

Out[14]:

	Object Number	Highlight	Is Public Domain	Object ID	Department	Object Name	Title	Culture	Artist Role	A Dis N
0	1979.486.1	False	False	1	The American Wing	Coin	One-dollar Liberty Head Coin	NaN	Maker	Jas Ba Long
1	1980.264.5	False	False	2	The American Wing	Coin	Ten-dollar Liberty Head Coin	NaN	Maker	Chris Gobri
2	67.265.9	False	False	3	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Chris Gobri
3	67.265.10	False	False	4	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Chris Gobri
4	67.265.11	False	False	5	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Chris Gobri
5	67.265.12	False	False	6	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Chris Gobri
6	67.265.13	False	False	7	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Chris Gobri
7	67.265.14	False	False	8	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Chris Gobri
8	67.265.15	False	False	9	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Chris Gobri
9	1979.486.3	False	False	10	The American Wing	Coin	Two-and-a-half-dollar Indian	NaN	Maker	

	Object Number	Is Highlight	Is Public Domain	Object ID	Department	Object Name	Title	Culture	Artist Role	A Dis N
							Head Coin			
10	1979.486.2	FALSE	FALSE	11	The American Wing	Coin	Two-and-a-half-dollar Liberty Head Coin	Nan	Maker	Chri Gobi
11	1979.486.7	FALSE	FALSE	12	The American Wing	Coin	Twenty-dollar Liberty Head Coin	Nan	Maker	Ja Ba Long
12	1979.486.4	FALSE	FALSE	13	The American Wing	Coin	Five-dollar Indian Head Coin	Nan	Maker	
13	1979.486.5	FALSE	FALSE	14	The American Wing	Coin	Five-dollar Liberty Head Coin	Nan	Maker	Chri Gobi
14	16.74.49	FALSE	FALSE	15	The American Wing	Coin	Coin, 1/2 Real	Mexican	Maker	Chri Gobi
15	16.74.27	FALSE	FALSE	16	The American Wing	Peso	Coin, 1/4 Peso	Mexican	Artist	Mex A
16	16.74.28	FALSE	FALSE	17	The American Wing	Peso	Coin, 1/4 Peso	Mexican	Artist	Mex A
17	16.74.29	FALSE	FALSE	18	The American Wing	Peso	Coin, 1/4 Peso	Mexican	Artist	Mex A
18	16.74.30	FALSE	FALSE	19	The American Wing	Peso	Coin, 1/4 Peso	Mexican	Artist	Mex A
19	16.74.31	FALSE	FALSE	20	The American Wing	Peso	Coin, 1/4 Peso	Mexican	Artist	Mex A

	Object Number	Highlight	Is Public Domain	Object ID	Department	Object Name	Title	Culture	Artist Role	A Dis N
20	16.74.32	FALSE	FALSE	21	The American Wing	Peso	Coin, 1/4 Peso	Mexican	Artist	Mex A
21	16.74.43	FALSE	FALSE	22	The American Wing	Coin	Coin, 1/4 Real	Guatemalan	Artist	Mex A
22	16.74.44	FALSE	FALSE	23	The American Wing	Coin	Coin, 1/4 Real	Guatemalan	Artist	Mex A
23	16.74.33	FALSE	FALSE	24	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
24	16.74.34	FALSE	FALSE	25	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
25	16.74.35	FALSE	FALSE	26	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
26	16.74.36	FALSE	FALSE	27	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
27	16.74.38	FALSE	FALSE	28	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
28	16.74.39	FALSE	FALSE	29	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
29	16.74.37	FALSE	FALSE	30	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
30	16.74.40	FALSE	FALSE	31	The American Wing	Centavos	Coin, 10 Centavos	Mexican	Artist	Mex A
31	09.9.15	FALSE	FALSE	32	The American Wing	Pesos	Coin, 20 Pesos	Mexican	Artist	Mex A

32 rows x 28 columns

Chapter 8

Create hierarchical index

Add 'Department' and 'Object id' as the indexes to the dataframe as hierachial index, these indexes are part of the dataframe

```
In [15]: filtered_metObj_df.isnull().sum()
```

```
Out[15]:
```

Object Number	21
Is Highlight	512
Is Public Domain	591
Object ID	705
Department	1240
Object Name	1254
Title	633
Culture	2975
Artist Role	6573
Artist Display Name	6556
Artist Display Bio	6241
Artist Alpha Sort	5841
Artist Begin Date	6950
Artist End Date	6612
Object Date	1467
Object Begin Date	1278
Object End Date	1279
Medium	1279
Dimensions	1332
Credit Line	2025
Geography Type	4788
City	8110
Country	4792
Classification	2028
Link Resource	2029
Metadata Date	3428
Repository	2029
Tags	8243
dtype:	int64

```
In [16]: hier_df = filtered_metObj_df.fillna(method='ffill')
```

```
In [17]: hier_df.isnull().sum()
```

```
Out[17]: Object Number      0  
Is Highlight          0  
Is Public Domain       0  
Object ID              0  
Department             0  
Object Name            0  
Title                  0  
Culture                14  
Artist Role            0  
Artist Display Name    0  
Artist Display Bio     0  
Artist Alpha Sort      0  
Artist Begin Date      0  
Artist End Date        0  
Object Date            0  
Object Begin Date      0  
Object End Date        0  
Medium                 0  
Dimensions             0  
Credit Line            0  
Geography Type         14  
City                   32  
Country                14  
Classification          0  
Link Resource           0  
Metadata Date          6  
Repository             0  
Tags                   6  
dtype: int64
```

```
In [18]: hier_df.head()
```

Out[18]:

	Object Number	Is Highlight	Is Public Domain	Object ID	Department	Object Name	Title	Culture	Artist Role	Artist Display Name	...
0	1979.486.1	False	False	1	The American Wing	Coin	One-dollar Liberty Head Coin	NaN	Maker	James Barton Longacre	...
1	1980.264.5	False	False	2	The American Wing	Coin	Ten-dollar Liberty Head Coin	NaN	Maker	Christian Gobrecht	...
2	67.265.9	False	False	3	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Christian Gobrecht	...
3	67.265.10	False	False	4	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Christian Gobrecht	...
4	67.265.11	False	False	5	The American Wing	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Christian Gobrecht	...

5 rows × 28 columns

In [19]: `# Add the two columns as the index
hier_df.set_index(['Department', 'Object ID'], inplace=True)`

In [20]: `# While printing the few top records those two indexes are visible
hier_df.head(100)`

Out[20]:

		Object Number	Is Highlight	Is Public Domain	Object Name	Title	Culture	Artist Role	Artist Display Name
Department	Object ID								
The American Wing	1	1979.486.1	FALSE	FALSE	Coin	One-dollar Liberty Head Coin	NaN	Maker	James Barton Longacre
	2	1980.264.5	FALSE	FALSE	Coin	Ten-dollar Liberty Head Coin	NaN	Maker	Christian Gobrecht
	3	67.265.9	FALSE	FALSE	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Christian Gobrecht
	4	67.265.10	FALSE	FALSE	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Christian Gobrecht
	5	67.265.11	FALSE	FALSE	Coin	Two-and-a-Half Dollar Coin	NaN	Maker	Christian Gobrecht

102	1971.180.62a	FALSE	TRUE	Andiron	Andiron	American or British	Maker	Union Glass Company	
103	1971.180.62b	FALSE	TRUE	Andiron	Andiron	American or British	Maker	Union Glass Company	
104	1980.499.4	FALSE	TRUE	Andiron	Andiron	American or British	Maker	Union Glass Company	
105	1980.499.5	FALSE	TRUE	Andiron	Andiron	American or British	Maker	Union Glass Company	
106	1984.134.1	FALSE	TRUE	Andiron	Andiron	American	Maker	Union Glass Company	

100 rows × 26 columns

```
In [21]: # Find the unique index values
for level in hier_df.index.levels:
    print(level.name, ":", len(level))
```

```
Department : 28
Object ID : 10910
```

Reshape the dataframe

Reshape the dataframe by using stack function that transposes the columns to the indexes

```
In [22]: #
reshaped_df = hier_df.stack()
```

```
In [23]: hier_df
```

Out[23]:

		Object Number	Is Highlight	Is Public Domain	Object Name	Title	Culture	Artist Role	Art Displ Nar
Department	Object ID								
The American Wing	1	1979.486.1	FALSE	FALSE	Coin	One-dollar Liberty Head Coin	Nan	Maker	Jam Bart Longac
	2	1980.264.5	FALSE	FALSE	Coin	Ten-dollar Liberty Head Coin	Nan	Maker	Christi Gobre
	3	67.265.9	FALSE	FALSE	Coin	Two-and-a-Half Dollar Coin	Nan	Maker	Christi Gobre
	4	67.265.10	FALSE	FALSE	Coin	Two-and-a-Half Dollar Coin	Nan	Maker	Christi Gobre
	5	67.265.11	FALSE	FALSE	Coin	Two-and-a-Half Dollar Coin	Nan	Maker	Christi Gobre

11814	1974.356.1 recto		FALSE	FALSE	Watercolor	Rialto Bridge (Covered Bridge, Venice)	American	Artist	Mauri Bra Prenderg
11815	54.143.8		FALSE	FALSE	Watercolor	The Rider	American	Artist	Mauri Bra Prenderg
11816	1976.201.4		FALSE	FALSE	Watercolor	Umbrellas in the Rain, Venice	American	Artist	Mauri Bra Prenderg
11817	64.118		FALSE	FALSE	Watercolor	Worship of Moloch (The Golden Idol)	American	Artist	Micha Pressm
11817	4		FALSE	FALSE	Watercolor	Worship of	American	Artist	Micha Pressm

Department	Object ID	Object Number	Is Highlight	Is Public Domain	Object Name	Title	Culture	Artist Role	Art Displ Nar
European Paintings	PA.1983.12.1	1983.12.1	True	True	Moloch (The Golden Bull)	The Golden Bull	Egyptian	Painter	Painting, Egyptian, Oil on panel, 1983

13108 rows x 36 columns

```
In [24]: reshaped_df
```

```
Out[24]: Department          Object ID
The American Wing 1          Object Number
1979.486.1
                                         Is Highlight
FALSE
                                         Is Public Domain
FALSE
                                         Object Name
Coin
                                         Title
One-dollar Lib
erty Head Coin
                                         ...
                                         Classification
11817
Drawings
                                         Link Resource
                                         http://www.metmuseum.org/art/collect
tion/search...
                                         Metadata Date
                                         2019-02-01
T10:50:49.477Z
                                         Repository
                                         Metropolitan Museum of Ar
t, New York, NY
                                         Tags
Trees
Length: 314722, dtype: object
```

Chapter 10

Group By

Group the Dataframe by using

- Group by Function
 - Group by Column

Group of function - use the size of the column to group

```
In [25]: grp_df = filtered_metObj_df.fillna(method='ffill')
```

```
In [26]: grp = df.groupby('Object Name').size()
```

```
In [27]: grp_d_data1
```

```
Out[27]: Object Name
1757      |1771          1
32 3/4 x 79 3/8 x 30 1/2 in. (83.2 x 201.6 x 77.5 cm) 3
Advertisement          1
Ale glass              2
Ale pitcher            1
                           ..
Worktable              3
Writing armchair       1
Writing table           2
tag#re                 4
tag#re shelves          1
Length: 1009, dtype: int64
```

```
In [28]: grp_d_data2 = grp_df.groupby('Department').size()
```

```
In [29]: grp_d_data2
```

Out[29]:

Department	
1954"	
3	
glass with	
1	
white pine"	
2	
36 x 49 1/2 x 22 in. (91.4 x 125.7 x 55.9 cm)	
1	
American	
1	
Baltimore	
1	
Berks County	
1	
Boston	
1	
Bucks County	
3	
European Sculpture and Decorative Arts	
3	
Guilford	
1	
Guilford Saybrook	
1	
Made in	
7	
Merseyside	
1	
Modern and Contemporary Art	
1	
Montour County	
1	
New York	
1	
Newport	
1	
Philadelphia	
2	
Plymouth County	
1	
Probably made in	
1	
Probably made in Possibly made in	
1	
Purchase, Friends of the American Wing Fund, Anonymous Gift, George M. Kaufman Gift, Sansbury- Mills Fund; Gifts of the Members of the Committee of the Bertha King Benkar d Memorial Fund, Mrs. Russell Sage, Mrs. Frederick Wildman, F. Ethel Wickham, Edgar W illiam and Bernice Chrysler Garbisch, and Mrs. F. M. Townsend, by exchange; and John Stewart Kennedy Fund and Bequests of Martha S. Tiedeman and W. Gedney Beatty, by exch ange, 1976	1
Queens	
1	
Reading	
1	
South Yorkshire	
2	
Staffordshire	
1	
The American Wing	

```
12066  
dtype: int64
```

Group by column

Group by colum 'Culture' that uses the row values to group the records

```
In [30]: # use the group by column name  
grp_data_3 = grp_df.groupby('Culture')
```

```
In [31]: grp_data_3
```

```
Out[31]: <pandas.core.groupby.generic.DataFrameGroupBy object at 0x0000025D48AE1090>
```

```
In [32]: for name, group in grp_data_3:  
    print(name)  
    print(len(group))
```

American
9620
American or British
162
American or Canadian
3
American or Danish
1
American or European
3
American or French
20
American or German
13
American or Irish
2
American, Japanese
15
American, Shaker
110
American, possibly
7
American, probably
8
Bermudian, possibly
2
Bohemian
2
British
418
British (American Market)
7
British (American market)
309
British or American
2
British or Czech
1
British or Flemish
6
British or Irish
13
British, possibly
12
British, probably
71
Canadian
3
Canadian (British)
4
China
3
Chinese
235
Chinese for export
6
Chinese, for American market
494
Chinese, for Swedish market
2

Czech
2
Dutch
29
Dutch, probably
1
European
12
European, probably
4
European, probably British
1
Flemish
1
French
203
French, possibly
20
French, probably
2
German
27
Guatemalan
3
Irish
7
Italian
3
Italian or German
2
Italian, probably
5
Japan
9
Japanese
5
Mexican
171
Mid-Atlantic|New England
2
New England
3
New England |Mid-Atlantic
1
New York
1
Scottish
6
Spanish
2
United States
11
United States|United States
2
Venezuelan
1
Worcestershire
1
ca. 1800♦1850
1

```
probably American  
2
```

```
In [33]: # print the rows that are grouped by 'American or French' value  
print(grp_data_3.get_group('American or French'))
```

	Object Number	Is Highlight	Is Public Domain	Object ID	Department \
226	69.262.3	FALSE	TRUE	265	The American Wing
227	69.262.4	FALSE	TRUE	266	The American Wing
1157	53.179.17a o o	FALSE	TRUE	1117	The American Wing
1163	53.179.16a o o	FALSE	TRUE	1123	The American Wing
1164	67.262.8	FALSE	TRUE	1124	The American Wing
1911	69.262.5	FALSE	TRUE	1834	The American Wing
1912	69.262.6	FALSE	TRUE	1835	The American Wing
1913	69.262.7	FALSE	TRUE	1836	The American Wing
1914	69.262.8	FALSE	TRUE	1837	The American Wing
2279	40.150.105	FALSE	TRUE	2171	The American Wing
3516	51.171.52	FALSE	TRUE	3254	The American Wing
3517	51.171.53	FALSE	TRUE	3255	The American Wing
3518	51.171.54	FALSE	TRUE	3256	The American Wing
3937	69.262.9	FALSE	TRUE	3683	The American Wing
7372	51.171.115	FALSE	FALSE	7021	The American Wing
7397	51.171.208	FALSE	FALSE	7046	The American Wing
7966	69.262.1	FALSE	TRUE	7602	The American Wing
7967	69.262.2	FALSE	TRUE	7603	The American Wing
8533	69.262.10	FALSE	TRUE	8082	The American Wing
9532	39.56.4	FALSE	TRUE	9020	The American Wing

	Object Name	Title	Culture \
226	Armchair	Armchair	American or French
227	Armchair	Armchair	American or French
1157	Candelabrum	Candelabrum	American or French
1163	Candelabrum	Candelabrum	American or French
1164	Candelabrum	Candelabrum	American or French
1911	Side Chair	Side Chair	American or French
1912	Side Chair	Side Chair	American or French
1913	Side Chair	Side Chair	American or French
1914	Side Chair	Side Chair	American or French
2279	Perfume bottle	Perfume Bottle	American or French
3516	Dish	Dish	American or French
3517	Dish	Dish	American or French
3518	Dish	Dish	American or French
3937	Firescreen	Firescreen	American or French
7372	Salt	Salt	American or French
7397	Salt	Salt	American or French
7966	Sofa	Sofa	American or French
7967	Sofa	Sofa	American or French
8533	Table	Table	American or French
9532	Tumbler	Tumbler	American or French

	Artist Role \
226	Maker
227	Maker
1157	Maker
1163	Maker
1164	Maker
1911	Maker
1912	Maker
1913	Maker
1914	Maker
2279	Retailer
3516	Maker
3517	Maker
3518	Maker
3937	Maker
7372	Manufacturer Manufacturer

7397	Manufacturer
7966	Maker
7967	Maker
8533	Maker
9532	Maker

	Artist Display Name	...	\
226	Auguste- Emile Rinquet-Leprince	...	
227	Auguste- Emile Rinquet-Leprince	...	
1157	Union Porcelain Works	...	
1163	Union Porcelain Works	...	
1164	Union Porcelain Works	...	
1911	Auguste- Emile Rinquet-Leprince	...	
1912	Auguste- Emile Rinquet-Leprince	...	
1913	Auguste- Emile Rinquet-Leprince	...	
1914	Auguste- Emile Rinquet-Leprince	...	
2279	Henry Kellam Hancock	...	
3516	Union Porcelain Works	...	
3517	Union Porcelain Works	...	
3518	Union Porcelain Works	...	
3937	Auguste- Emile Rinquet-Leprince	...	
7372	Boston & Sandwich Glass Company New England Gl...	...	
7397	Boston & Sandwich Glass Company	...	
7966	Auguste- Emile Rinquet-Leprince	...	
7967	Auguste- Emile Rinquet-Leprince	...	
8533	Auguste- Emile Rinquet-Leprince	...	
9532	Boston & Sandwich Glass Company	...	

	Dimensions	\
226	38 1/2 x 23 1/4 x 26 1/8 in. (97.8 x 59.1 x 66...)	
227	38 1/2 x 23 1/4 x 26 1/8 in. (97.8 x 59.1 x 66...)	
1157	H. 27 7/8 in. (70.8 cm)	
1163	H. 27 3/4 in. (70.5 cm)	
1164	H. 32 1/2 in. (82.6 cm); Diam. 17 1/2 in. (44....)	
1911	36 x 18 1/2 x 20 in. (91.4 x 47 x 50.8 cm)	
1912	38 1/2 x 23 1/4 x 26 1/8 in. (97.8 x 59.1 x 66...)	
1913	38 1/2 x 23 1/4 x 26 1/8 in. (97.8 x 59.1 x 66...)	
1914	38 1/2 x 23 1/4 x 26 1/8 in. (97.8 x 59.1 x 66...)	
2279	H. 5 1/4 in. (13.3 cm)	
3516	Diam. 5 1/8 in. (13 cm)	
3517	H. 4 5/8 in. (11.7 cm)	
3518	Diam. 4 5/8 in. (11.7 cm)	
3937	40 1/2 x 26 1/4 x 15 in. (102.9 x 66.7 x 38.1 cm)	
7372	Diam. 2 13/16 in. (7.1 cm)	
7397	2 3/16 x 2 x 3 3/8 in. (5.6 x 5.1 x 8.6 cm)	
7966	36 1/4 x 62 1/2 x 29 in. (92.1 x 158.8 x 73.7 cm)	
7967	36 1/4 x 62 1/2 x 29 in. (92.1 x 158.8 x 73.7 cm)	
8533	31 x 59 x 30 in. (78.7 x 149.9 x 76.2 cm)	
9532	H. 3 5/8 in. (9.2 cm); Diam. 3 in. (7.6 cm)	

	Credit Line	\
226	Gift of Mrs. Douglas Williams, 1969	
227	Gift of Mrs. Douglas Williams, 1969	
1157	Gift of Mrs. Screven Lorillard, 1953	
1163	Gift of Mrs. Screven Lorillard, 1953	
1164	Gift of John C. Cattus, 1967	
1911	Gift of Mrs. Douglas Williams, 1969	
1912	Gift of Mrs. Douglas Williams, 1969	
1913	Gift of Mrs. Douglas Williams, 1969	
1914	Gift of Mrs. Douglas Williams, 1969	

2279	Bequest of Constance R. Brown, 1939
3516	Gift of Mrs. Charles W. Green, in memory of Dr...
3517	Gift of Mrs. Charles W. Green, in memory of Dr...
3518	Gift of Mrs. Charles W. Green, in memory of Dr...
3937	Gift of Mrs. Douglas Williams, 1969
7372	Gift of Mrs. Charles W. Green, in memory of Dr...
7397	Gift of Mrs. Charles W. Green, in memory of Dr...
7966	Gift of Mrs. Douglas Williams, 1969
7967	Gift of Mrs. Douglas Williams, 1969
8533	Gift of Mrs. Douglas Williams, 1969
9532	Rogers Fund, 1939

	Geography	Type	\
226	Possibly made in	Possibly made in	
227	Possibly made in	Possibly made in	
1157	Possibly made in	Possibly made in	
1163	Possibly made in	Possibly made in	
1164	Possibly made in	Possibly made in	
1911	Possibly made in	Possibly made in	
1912	Possibly made in	Possibly made in	
1913	Possibly made in	Possibly made in	
1914	Possibly made in	Possibly made in	
2279	Possibly made in	Possibly made in	
3516	Possibly made in	Possibly made in	
3517	Possibly made in	Possibly made in	Possibly mad...
3518	Possibly made in	Possibly made in	Possibly mad...
3937	Possibly made in	Possibly made in	
7372	Possibly made in	Possibly made in	
7397	Possibly made in	Possibly made in	
7966	Possibly made in	Possibly made in	
7967	Possibly made in	Possibly made in	
8533	Possibly made in	Possibly made in	
9532		Made in	

	City	Country	Classification	\
226	New York Paris	United States France	Furniture	
227	New York Paris	United States France	Furniture	
1157	Brooklyn	United States France	Glass	
1163	Brooklyn	United States France	Glass	
1164	Brooklyn	United States France	Metal	
1911	New York Paris	United States France	Furniture	
1912	New York Paris	United States France	Furniture	
1913	New York Paris	United States France	Furniture	
1914	New York Paris	United States France	Furniture	
2279	London	United States France	Glass	
3516	Nuremberg	United States France	Glass	
3517	Nuremberg	Belgium England France	Glass	
3518	Nuremberg	Belgium England France	Glass	
3937	New York Paris	United States France	Furniture	
7372	East Cambridge Sandwich	United States France	Glass	
7397	Sandwich	United States France	Glass	
7966	New York Paris	United States France	Furniture	
7967	New York Paris	United States France	Furniture	
8533	New York Paris	United States France	Furniture	
9532	Pittsburgh	United States	Glass	

	Link	Resource	\
226	http://www.metmuseum.org/art/collection/search...		
227	http://www.metmuseum.org/art/collection/search...		
1157	http://www.metmuseum.org/art/collection/search...		

1163 http://www.metmuseum.org/art/collection/search...
1164 http://www.metmuseum.org/art/collection/search...
1911 http://www.metmuseum.org/art/collection/search...
1912 http://www.metmuseum.org/art/collection/search...
1913 http://www.metmuseum.org/art/collection/search...
1914 http://www.metmuseum.org/art/collection/search...
2279 http://www.metmuseum.org/art/collection/search...
3516 http://www.metmuseum.org/art/collection/search...
3517 http://www.metmuseum.org/art/collection/search...
3518 http://www.metmuseum.org/art/collection/search...
3937 http://www.metmuseum.org/art/collection/search...
7372 http://www.metmuseum.org/art/collection/search...
7397 http://www.metmuseum.org/art/collection/search...
7966 http://www.metmuseum.org/art/collection/search...
7967 http://www.metmuseum.org/art/collection/search...
8533 http://www.metmuseum.org/art/collection/search...
9532 http://www.metmuseum.org/art/collection/search...

	Metadata	Date	Repository	\
226	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
227	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
1157	2019-07-31T03:00:40.447Z	Metropolitan Museum of Art, New York, NY		
1163	2019-07-31T03:00:40.447Z	Metropolitan Museum of Art, New York, NY		
1164	2019-07-31T03:00:40.447Z	Metropolitan Museum of Art, New York, NY		
1911	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
1912	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
1913	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
1914	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
2279	2019-07-31T03:00:40.447Z	Metropolitan Museum of Art, New York, NY		
3516	2019-05-22T03:00:45.347Z	Metropolitan Museum of Art, New York, NY		
3517	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
3518	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
3937	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
7372	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
7397	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		
7966	2019-07-31T03:00:40.447Z	Metropolitan Museum of Art, New York, NY		
7967	2019-07-31T03:00:40.447Z	Metropolitan Museum of Art, New York, NY		
8533	2019-07-31T03:00:40.447Z	Metropolitan Museum of Art, New York, NY		
9532	2019-07-26T03:00:41.71Z	Metropolitan Museum of Art, New York, NY		

	Tags
226	Grapes
227	Grapes
1157	Men Profiles
1163	Grapes Children Flowers Leaves
1164	Women
1911	Squirrels Flowers Trees
1912	Squirrels Flowers Trees
1913	Squirrels Flowers Trees
1914	Squirrels Flowers Trees
2279	George Washington Profiles
3516	Bridges Men Trees
3517	Bridges Men Trees
3518	Bridges Men Trees
3937	Flowers
7372	Horses Chariots
7397	Horses Chariots
7966	Hills Landscapes
7967	Hills Landscapes
8533	Flowers

9532

Birds

[20 rows x 28 columns]

Chapter 11

Convert between string and date time / timestamps

Convert the timestamp value in 'Metadata Date' column and convert to millisecond, convert to UAT date time/ timestamp and EST timezone format date time/ timestamp by creating 3 columns to hold those values

```
In [34]: grp_df = grp_df.fillna(method='bfill')
```

```
In [35]: grp_df['Metadata Date']
```

```
Out[35]: 0      2019-02-01T10:50:49.477Z  
1      2019-02-01T10:50:49.477Z  
2      2019-02-01T10:50:49.477Z  
3      2019-02-01T10:50:49.477Z  
4      2019-02-01T10:50:49.477Z  
      ...  
12103    2019-02-01T10:50:49.477Z  
12104    2019-02-01T10:50:49.477Z  
12105    2019-10-10T14:54:23.72Z  
12106    2019-02-01T10:50:49.477Z  
12107    2019-02-01T10:50:49.477Z  
Name: Metadata Date, Length: 12108, dtype: object
```

```
In [36]: # check the date data type  
dt_temp = grp_df['Metadata Date'][0]  
type(dt_temp)
```

```
Out[36]: str
```

The Data type is a String, convert the string to datetime Convert it to a UTC , the universal time zone and then to a desired time zone in this it is EST , America/ New_York timezone

```
In [37]: from datetime import datetime  
import pytz  
def convertStrDateToDateTimObject(timestamp_str, TmZnType):  
  
    timestamp_str = timestamp_str[:19]  
  
    datetime_obj = datetime.strptime(timestamp_str, '%Y-%m-%dT%H:%M:%S')  
  
    if TmZnType == 'millis':  
        # Convert datetime object to milliseconds  
        milliseconds = int(datetime_obj.timestamp()) * 1000  
        return milliseconds  
  
    if TmZnType == 'utc':  
        # Set the timezone to UTC  
        utc_datetime = datetime_obj.replace(tzinfo=pytz.utc)  
        return utc_datetime
```

```
if TmZnType == 'est':
    # Set the timezone to EST
    est_datetime = datetime_obj.replace(tzinfo=pytz.utc).astimezone(pytz.timezone('EST'))
    return est_datetime
```

```
In [38]: # create a 3 columns, each to hold milliseconds, UTC timezone and EST timestamps
grp_df['Metadata_Date_Millis'] = [convertStrDateToDateTimObject(x, 'millis') for x in grp_df]
grp_df['Metadata_Date_UTC'] = [convertStrDateToDateTimObject(x, 'utc') for x in grp_df]
grp_df['Metadata_Date_EST'] = [convertStrDateToDateTimObject(x, 'est') for x in grp_df]
```

```
In [39]: grp_df['Metadata_Date_Millis']
```

```
Out[39]: 0      1549039849000
          1      1549039849000
          2      1549039849000
          3      1549039849000
          4      1549039849000
          ...
          12103     1549039849000
          12104     1549039849000
          12105     1570737263000
          12106     1549039849000
          12107     1549039849000
Name: Metadata_Date_Millis, Length: 12108, dtype: int64
```

```
In [40]: grp_df['Metadata_Date_UTC']
```

```
Out[40]: 0      2019-02-01 10:50:49+00:00
          1      2019-02-01 10:50:49+00:00
          2      2019-02-01 10:50:49+00:00
          3      2019-02-01 10:50:49+00:00
          4      2019-02-01 10:50:49+00:00
          ...
          12103     2019-02-01 10:50:49+00:00
          12104     2019-02-01 10:50:49+00:00
          12105     2019-10-10 14:54:23+00:00
          12106     2019-02-01 10:50:49+00:00
          12107     2019-02-01 10:50:49+00:00
Name: Metadata_Date_UTC, Length: 12108, dtype: datetime64[ns, UTC]
```

```
In [41]: grp_df['Metadata_Date_EST']
```

```
Out[41]: 0      2019-02-01 05:50:49-05:00
          1      2019-02-01 05:50:49-05:00
          2      2019-02-01 05:50:49-05:00
          3      2019-02-01 05:50:49-05:00
          4      2019-02-01 05:50:49-05:00
          ...
          12103     2019-02-01 05:50:49-05:00
          12104     2019-02-01 05:50:49-05:00
          12105     2019-10-10 10:54:23-04:00
          12106     2019-02-01 05:50:49-05:00
          12107     2019-02-01 05:50:49-05:00
Name: Metadata_Date_EST, Length: 12108, dtype: datetime64[ns, America/New_York]
```

```
In [42]: temp_dt = grp_df['Metadata_Date_Millis'][0]
print( "Date string value converted to Milliseconds is ", type(temp_dt) )
temp_dt = grp_df['Metadata_Date_UTC'][0]
print( "Date string value converted to UTC timezone timepzone type is ", type(temp_dt)
temp_dt = grp_df['Metadata_Date_EST'][0]
print( "Date string value converted to EST timezone timepzone type is ", type(temp_dt)
```

```
Date string value converted to Milliseconds is <class 'numpy.int64'>
Date string value converted to UTC timezone timepzone type is <class 'pandas._libs.tslibs.timestamps.Timestamp'>
Date string value converted to EST timezone timepzone type is <class 'pandas._libs.tslibs.timestamps.Timestamp'>
```

```
In [ ]:
```

Generate date range

Generate date range by using 2 column values 'Object Begin Date' and 'Object End Date' that had year values. Using the year values from these two columns

create date values - use month january and date 1st for begin date and month december and date 31 for end date.

To avoid same date , a random number is created between 4 to 10 and that number will be used to split the date between start date and end date for each row so they have unique values if the the difference between begin date and end date values were duplicated across rows.

Also the 2nd of the date range created is assigned to new column Object_begin_dt_rng and last but one of the date range created is assigned to Object_end_dt_rng

```
In [43]: import random
```

```
def generteDateRange(DataFrm, isFirstRange):
    try:
        int_st_yr = int(DataFrm['Object Begin Date'])
        int_end_yr = int(DataFrm['Object End Date'])
    except:
        return
    if(int_st_yr < 1700):
        return
    start_year = str(DataFrm['Object Begin Date']) + '-01-01'
    end_yr = str(DataFrm['Object End Date']) + '-12-31'
    # Generate a random number between 4 and 10
    random_number = random.randint(4, 10)

    dt_range = pd.date_range(start_year, end_yr, periods=random_number)

    if isFirstRange:
        return dt_range[1].strftime("%m/%d/%Y")
```

```
    else:  
        return dt_range[random_number-2].strftime("%m/%d/%Y")
```

```
In [44]: grp_df['Object_begin_dt_rng'] = grp_df.apply(lambda row: generateDateRange(row, True),  
grp_df['Object_end_dt_rng'] = grp_df.apply(lambda row: generateDateRange(row, False),
```

```
In [45]: grp_df['Object_begin_dt_rng']
```

```
Out[45]: 0      03/02/1853  
1      02/15/1901  
2      02/10/1911  
3      02/10/1911  
4      10/19/1912  
       ...  
12103   05/27/1911  
12104   10/20/1909  
12105   03/23/1898  
12106   05/02/1950  
12107   02/15/1950  
Name: Object_begin_dt_rng, Length: 12108, dtype: object
```

```
In [46]: grp_df['Object_end_dt_rng']
```

```
Out[46]: 0      11/20/1853  
1      08/31/1901  
2      10/30/1924  
3      11/20/1925  
4      03/13/1924  
       ...  
12103   05/01/1912  
12104   07/21/1912  
12105   05/02/1899  
12106   11/15/1950  
12107   11/20/1950  
Name: Object_end_dt_rng, Length: 12108, dtype: object
```