Cleaning AmericanGallery Using Pandas. (SALMAN ARIF)

In [40]:
```
import pandas as pd
import re
```

In [41]:
```
Gallery = pd.read_csv("AmericanGallery.csv")
```

In [42]:
```
Gallery
```

Out[42]:

	Title	Artist	Nationality	BeginDate	EndDate	Gender	Date	Department
0	Dress MacLeod from Tartan Sets	Sarah Charlesworth	(American)	-1947.0	-2013.0	(Female)	1986	Prints & Illustrated Books
1	Duplicate of plate from folio 11 verso (supple...	Pablo Palazuelo	(Spanish)	-1916.0	-2007.0	(Male)	1978	Prints & Illustrated Books
2	Tailpiece (page 55) from SAGESSE	Maurice Denis	(French)	-1870.0	-1943.0	(Male)	1889-1911	Prints & Illustrated Books
3	Headpiece (page 129) from LIVRET DE FOLASTRIES...	Aristide Maillol	(French)	-1861.0	-1944.0	(Male)	1927-1940	Prints & Illustrated Books
4	97 rue du Bac	Eugène Atget	(French)	-1857.0	-1927.0	(Male)	1903	Photography
...	...	...	...	...	...	...	...	...
16724	Oval with Points	Henry Moore	(British)	-1898.0	-1986.0	(Male)	1968-1969	Painting & Sculpture
16725	Cementerio de la Ciudad Abierta, Ritoque, Chile	Juan Baixas	(Chilean)	-1942.0	NaN	(Male)	1975	Architecture & Design
16726	The Catboat	Edward Hopper	(American)	-1882.0	-1967.0	(Male)	1922	Prints & Illustrated Books
16727	Dognat' i peregnat' v tekhniko-ekonomicheskom ...	Unknown	()	NaN	NaN	()	1931	Prints & Illustrated Books
16728	Plate (page 11) from The Dive	Alex Katz	(American)	-1927.0	NaN	(Male)	2011	Prints & Illustrated Books

16729 rows × 8 columns

In [101]:
```
Gallery.rename({"BeginDate":"Begin_Date"}, axis = 1, inplace = True)
Gallery.rename({"EndDate":"End_Date"}, axis = 1, inplace = True)
Gallery.rename({"ActualAge":"Actual_Age"}, axis = 1, inplace = True)

```

In [102]:
```
Gallery["Nationality"] = Gallery["Nationality"].str.strip("(").str.strip(")")
```

In [106]:
```
Gallery["Nationality"]
```

Out[106]:
```
0        American
1         Spanish
2          French
3          French
4          French
           ...
16724     British
16725     Chilean
16726    American
16727          NP
16728    American
Name: Nationality, Length: 16729, dtype: object
```

In [107]:
```
Gallery.dtypes
```

Out[107]:
```
Title             object
Artist            object
Nationality       object
Begin_Date         Int64
End_Date           Int64
Gender            object
Date               int32
Department        object
Actual_Age        object
Submission_Age    object
dtype: object
```

In [46]:
```
Gallery["BeginDate"] = Gallery["BeginDate"].astype(str).str.replace("-", "").astype(float)
```

In [47]:
```
Gallery["EndDate"] = Gallery["EndDate"].astype(str).str.replace("-", "").astype(float)
```

In [48]:
```
Gallery.dtypes
```

Out[48]:
```
Title           object
Artist          object
Nationality     object
BeginDate      float64
EndDate        float64
Gender          object
Date            object
Department      object
dtype: object
```

In [49]:
```
Gallery["Gender"] = Gallery["Gender"].str.strip("(").str.strip(")")
```

In [108]:
```
Gallery["Gender"]
```

Out[108]:
```
0        Female
1          Male
2          Male
3          Male
4          Male
           ...
16724      Male
16725      Male
16726      Male
16727        NP
16728      Male
Name: Gender, Length: 16729, dtype: object
```

In [51]:
```
bad_char = ["(",")",".","c","C","'"," ","S","s"]
```

In [52]:
```
Gallery['Date'] = Gallery['Date'].str.replace('|'.join(map(re.escape, bad_char)), '')
```

```
C:\Users\Salman\AppData\Local\Temp\ipykernel_3716\2401855263.py:1: FutureWarning: The default value of regex will change from Tru
e to False in a future version.
  Gallery['Date'] = Gallery['Date'].str.replace('|'.join(map(re.escape, bad_char)), '')
```

In [53]:
```
Gallery['Date']
```

Out[53]:
```
0             1986
1             1978
2        1889-1911
3        1927-1940
4             1903
            ...
16724    1968-1969
16725         1975
16726         1922
16727         1931
16728         2011
Name: Date, Length: 16729, dtype: object
```

In [54]:
```
Gallery['Date'] = Gallery['Date'].apply(lambda x: x.lstrip('-') if x.startswith('-') else x)

```

In [55]:
```
# fil = Gallery["Date"][Gallery["Date"].str.contains('-') & (Gallery["Date"].str.len() < 9)]
fil = Gallery["Date"][Gallery['Date'].str.contains('-') & (Gallery['Date'].str.len() < 9)]


```

In [56]:
```
fil
```

Out[56]:
```
6021     1910-30
12495    1910-30
14187    1880-90
16003    1910-30
Name: Date, dtype: object
```

In [57]:
```
def combine_date_range(date_range):
    start_year, end_year = date_range.split('-')

    if len(start_year) == 2:
        start_year = start_year.slice[0:2]
    if len(end_year) == 2:
        end_year = start_year[0:2] + end_year

    return f"{start_year}-{end_year}"

fil = fil.apply(combine_date_range)
```

In [58]:
```
fil
```

Out[58]:
```
6021     1910-1930
12495    1910-1930
14187    1880-1890
16003    1910-1930
Name: Date, dtype: object
```

In [109]:
```
Gallery["Date"]
```

Out[109]:
```
0        1986
1        1978
2        1900
3        1933
4        1903
         ...
16724    1968
16725    1975
16726    1922
16727    1931
16728    2011
Name: Date, Length: 16729, dtype: int32
```

In [21]:
```
Gallery.iloc[[6021, 12495, 14187, 16003], 0:8]
```

Out[21]:

	Title	Artist	Nationality	BeginDate	EndDate	Gender	Date	Department
6021	15 postcards of Tiflis (c.1900-30) and one of ...	Unknown		NaN	NaN		1910-30	Prints & Illustrated Books
12495	15 postcards of Tiflis (c.1900-30) and one of ...	Unknown		NaN	NaN		1910-30	Prints & Illustrated Books
14187	STONEHENGE, WILTSHIRE	George Washington Wilson	British	1823.0	1893.0	Male	1880-90	Photography
16003	15 postcards of Tiflis (c.1900-30) and one of ...	Unknown		NaN	NaN		1910-30	Prints & Illustrated Books

In [60]:
```
print("\nIndices of fil:")
print(fil.index)
```

```
Indices of fil:
Int64Index([6021, 12495, 14187, 16003], dtype='int64')
```

In [61]:
```
indices_to_update = [6021, 12495, 14187, 16003]  # Replace with your actual indices

# Update specific rows in Gallery["Date"] with the corrected values from fil
Gallery.loc[indices_to_update, "Date"] = fil.loc[indices_to_update]
```

In [62]:
```
Gallery.iloc[[6021, 12495, 14187, 16003], 0:8]
```

Out[62]:

	Title	Artist	Nationality	BeginDate	EndDate	Gender	Date	Department
6021	15 postcards of Tiflis (c.1900-30) and one of ...	Unknown		NaN	NaN		1910-1930	Prints & Illustrated Books
12495	15 postcards of Tiflis (c.1900-30) and one of ...	Unknown		NaN	NaN		1910-1930	Prints & Illustrated Books
14187	STONEHENGE, WILTSHIRE	George Washington Wilson	British	1823.0	1893.0	Male	1880-1890	Photography
16003	15 postcards of Tiflis (c.1900-30) and one of ...	Unknown		NaN	NaN		1910-1930	Prints & Illustrated Books

In [63]:
```
filt = Gallery["Date"][Gallery['Date'].str.contains('-') & (Gallery['Date'].str.len() == 9)]
```

In [64]:
```
filt
```

Out[64]:
```
2        1889-1911
3        1927-1940
7        1978-1983
10       1949-1950
12       1908-1911
           ...
16705    1889-1911
16706    1880-1910
16707    1945-1951
16709    1964-1965
16724    1968-1969
Name: Date, Length: 3066, dtype: object
```

In [65]:
```
df_split = filt.str.split('-', expand=True)

df_split1 = df_split.astype(int)

mean_values = df_split1.mean(axis=1).astype(int)

result = pd.concat([filt, mean_values.rename('Mean')], axis=1)
```

In [66]:
```
result.drop("Date", axis =1, inplace = True)
```

In [67]:
```
result
```

Out[67]:

	Mean
2	1900
3	1933
7	1980
10	1949
12	1909
...	...
16705	1900
16706	1895
16707	1948
16709	1964
16724	1968

3066 rows × 1 columns

In [68]:
```
result
print("\nIndices of result:")
ind = result.index
```

```
Indices of result:
```

In [69]:
```
ind
```

Out[69]:
```
Int64Index([    2,     3,     7,    10,    12,    15,    18,    20,    25,
                29,
            ...
            16679, 16689, 16692, 16699, 16702, 16705, 16706, 16707, 16709,
            16724],
           dtype='int64', length=3066)
```

In [70]:
```
print(result.index)
```

```
Int64Index([    2,     3,     7,    10,    12,    15,    18,    20,    25,
                29,
            ...
            16679, 16689, 16692, 16699, 16702, 16705, 16706, 16707, 16709,
            16724],
           dtype='int64', length=3066)
```

In [71]:
```
indices_to_update = (ind)

Gallery.loc[indices_to_update, "Date"] = result.loc[indices_to_update, "Mean"].values
```

In [72]:
```
48,1252, 255, 257, 258, 263, 267, 272, 276, 277, 282, 286, 289, 295, 296, 16679, 16689, 16692, 16699, 16702, 16705, 16706, 16707, 1
```

Out[72]:

	Title	Artist	Nationality	BeginDate	EndDate	Gender	Date	Department
2	Tailpiece (page 55) from SAGESSE	Maurice Denis	French	1870.0	1943.0	Male	1900	Prints & Illustrated Books
3	Headpiece (page 129) from LIVRET DE FOLASTRIES...	Aristide Maillol	French	1861.0	1944.0	Male	1933	Prints & Illustrated Books
7	Los Angeles Airport	Garry Winogrand	American	1928.0	1984.0	Male	1980	Photography
10	In-text plate (folio 11) from LA MÉTROMANIE	Jean Dubuffet	French	1901.0	1985.0	Male	1949	Prints & Illustrated Books
12	In-text plate (page 108) from LYSISTRATA	František Kupka	Czech	1871.0	1957.0	Male	1909	Prints & Illustrated Books
...	...	...	...	...	...	...	...	...
16702	Plate (folio 4 verso) from Boundary	Sam Francis	American	1923.0	1994.0	Male	1986	Prints & Illustrated Books
16705	Tailpiece (page 8) from SAGESSE	Maurice Denis	French	1870.0	1943.0	Male	1900	Prints & Illustrated Books
16706	Salon de la Plume	Unknown	Nationality Unknown	NaN	NaN		1895	Architecture & Design
16707	Farnsworth House, Plano, Illinois (Plan, secti...	Ludwig Mies van der Rohe	American	1886.0	1969.0	Male	1948	Architecture & Design
16709	Red over Yellow (Rouge sur jaune) from Suite o...	Ellsworth Kelly	American	1923.0	2015.0	Male	1964	Prints & Illustrated Books

66 rows × 8 columns

In [73]:
```
Gallery["EndDate"] = Gallery["EndDate"].astype("Int64")
```

In [81]:
```
Gallery["BeginDate"] = Gallery["BeginDate"].astype("Int64")
```

In [88]:
```
Gallery["Date"] = Gallery["Date"].astype(int)
```

In [89]:
```
Gallery.dtypes
```

Out[89]:
```
Title             object
Artist            object
Nationality       object
BeginDate          Int64
EndDate            Int64
Gender            object
Date               int32
Department        object
ActualAge         object
Submission_Age    object
dtype: object
```

In [90]:
```
Gallery["ActualAge"] = ""

for index, row in Gallery.iterrows():
    begin_date = row["BeginDate"]
    end_date = row["EndDate"]

    if pd.notna(begin_date) and pd.notna(end_date):
        Gallery.at[index, "ActualAge"] = (end_date - begin_date)
    else:
        Gallery.at[index, "ActualAge"] = "NP"
```

In [91]:
```
Gallery
```

Out[91]:

	Title	Artist	Nationality	BeginDate	EndDate	Gender	Date	Department	ActualAge	Submission_Age
0	Dress MacLeod from Tartan Sets	Sarah Charlesworth	American	1947	2013	Female	1986	Prints & Illustrated Books	66	
1	Duplicate of plate from folio 11 verso (supple...	Pablo Palazuelo	Spanish	1916	2007	Male	1978	Prints & Illustrated Books	91	
2	Tailpiece (page 55) from SAGESSE	Maurice Denis	French	1870	1943	Male	1900	Prints & Illustrated Books	73	
3	Headpiece (page 129) from LIVRET DE FOLASTRIES...	Aristide Maillol	French	1861	1944	Male	1933	Prints & Illustrated Books	83	
4	97 rue du Bac	Eugène Atget	French	1857	1927	Male	1903	Photography	70	
...	...	...	...	...	...	...	...	...	...	...
16724	Oval with Points	Henry Moore	British	1898	1986	Male	1968	Painting & Sculpture	88	
16725	Cementerio de la Ciudad Abierta, Ritoque, Chile	Juan Baixas	Chilean	1942	<NA>	Male	1975	Architecture & Design	NP	
16726	The Catboat	Edward Hopper	American	1882	1967	Male	1922	Prints & Illustrated Books	85	
16727	Dognat' i peregnat' v tekhniko-ekonomicheskom ...	Unknown	NP	<NA>	<NA>		1931	Prints & Illustrated Books	NP	
16728	Plate (page 11) from The Dive	Alex Katz	American	1927	<NA>	Male	2011	Prints & Illustrated Books	NP	

16729 rows × 10 columns

In [92]:
```
Gallery["Submission_Age"] = ""

for index, row in Gallery.iterrows():
    begin_date = row["BeginDate"]
    date = row["Date"]

    if pd.notna(begin_date) and pd.notna(date):
        Gallery.at[index, "Submission_Age"] = (date - begin_date)
    else:
        Gallery.at[index, "Submission_Age"] = "NP"
```

In [93]:
```
Gallery
```

Out[93]:

	Title	Artist	Nationality	BeginDate	EndDate	Gender	Date	Department	ActualAge	Submission_Age
0	Dress MacLeod from Tartan Sets	Sarah Charlesworth	American	1947	2013	Female	1986	Prints & Illustrated Books	66	39
1	Duplicate of plate from folio 11 verso (supple...	Pablo Palazuelo	Spanish	1916	2007	Male	1978	Prints & Illustrated Books	91	62
2	Tailpiece (page 55) from SAGESSE	Maurice Denis	French	1870	1943	Male	1900	Prints & Illustrated Books	73	30
3	Headpiece (page 129) from LIVRET DE FOLASTRIES...	Aristide Maillol	French	1861	1944	Male	1933	Prints & Illustrated Books	83	72
4	97 rue du Bac	Eugène Atget	French	1857	1927	Male	1903	Photography	70	46
...	...	...	...	...	...	...	...	...	...	...
16724	Oval with Points	Henry Moore	British	1898	1986	Male	1968	Painting & Sculpture	88	70
16725	Cementerio de la Ciudad Abierta, Ritoque, Chile	Juan Baixas	Chilean	1942	<NA>	Male	1975	Architecture & Design	NP	33
16726	The Catboat	Edward Hopper	American	1882	1967	Male	1922	Prints & Illustrated Books	85	40
16727	Dognat' i peregnat' v tekhniko-ekonomicheskom ...	Unknown	NP	<NA>	<NA>		1931	Prints & Illustrated Books	NP	NP
16728	Plate (page 11) from The Dive	Alex Katz	American	1927	<NA>	Male	2011	Prints & Illustrated Books	NP	84

16729 rows × 10 columns

In [94]:
```
for index, row in Gallery.iterrows():
    if row["Nationality"] == "":
        Gallery.at[index, "Nationality"] = "NP"
```

In [98]:
```
for index, row in Gallery.iterrows():
    if row["Gender"] == "":
        Gallery.at[index, "Gender"] = "NP"
```

In [104]:

```
1 Gallery
```

Out[104]:

	Title	Artist	Nationality	Begin_Date	End_Date	Gender	Date	Department	Actual_Age	Submission_Age
0	Dress MacLeod from Tartan Sets	Sarah Charlesworth	American	1947	2013	Female	1986	Prints & Illustrated Books	66	39
1	Duplicate of plate from folio 11 verso (supple...	Pablo Palazuelo	Spanish	1916	2007	Male	1978	Prints & Illustrated Books	91	62
2	Tailpiece (page 55) from SAGESSE	Maurice Denis	French	1870	1943	Male	1900	Prints & Illustrated Books	73	30
3	Headpiece (page 129) from LIVRET DE FOLASTRIES...	Aristide Maillol	French	1861	1944	Male	1933	Prints & Illustrated Books	83	72
4	97 rue du Bac	Eugène Atget	French	1857	1927	Male	1903	Photography	70	46
...	...	...	...	...	...	...	...	...	...	...
16724	Oval with Points	Henry Moore	British	1898	1986	Male	1968	Painting & Sculpture	88	70
16725	Cementerio de la Ciudad Abierta, Ritoque, Chile	Juan Baixas	Chilean	1942	<NA>	Male	1975	Architecture & Design	NP	33
16726	The Catboat	Edward Hopper	American	1882	1967	Male	1922	Prints & Illustrated Books	85	40
16727	Dognat' i peregnat' v tekhniko-ekonomicheskom ...	Unknown	NP	<NA>	<NA>	NP	1931	Prints & Illustrated Books	NP	NP
16728	Plate (page 11) from The Dive	Alex Katz	American	1927	<NA>	Male	2011	Prints & Illustrated Books	NP	84

16729 rows × 10 columns

In [105]:

```
1 Gallery.to_csv("American_Gallery_updated.csv", index = False)
```

In [ ]:

```
1
```