In [ ]:
```
# Essential Liberies
import pandas as pd
from sqlalchemy import create_engine
from urllib.parse import quote
import os
import duckdb
import numpy as np
```

In [ ]:
```
# Connection to connect to MYSQL Database
password = '123456'
encoded_pass = quote(password,safe='')
connection = f"mysql+mysqlconnector://root:{encoded_pass}@localhost:3306/painting"
db = create_engine(connection)
conn = db.connect()
```

In [ ]:
```
files = os.listdir('D:/Project/SQL Project/DataSet')
files
```

Out[ ]:
```
['artist.csv',
 'canvas_size.csv',
 'image_link.csv',
 'museum.csv',
 'museum_hours.csv',
 'product_size.csv',
 'subject.csv',
 'work.csv']
```

In [ ]:
```
artist = pd.read_csv(f'D:/Project/SQL Project/DataSet/artist.csv')
canvas_size = pd.read_csv(f'D:/Project/SQL Project/DataSet/canvas_size.csv')
image_link = pd.read_csv(f'D:/Project/SQL Project/DataSet/image_link.csv')
museum = pd.read_csv(f'D:/Project/SQL Project/DataSet/museum.csv')
museum_hours = pd.read_csv(f'D:/Project/SQL Project/DataSet/museum_hours.csv')
product_size = pd.read_csv(f'D:/Project/SQL Project/DataSet/product_size.csv')
subject = pd.read_csv(f'D:/Project/SQL Project/DataSet/subject.csv')
work = pd.read_csv(f'D:/Project/SQL Project/DataSet/work.csv')
```

In [ ]:
```
# artist
artist.head()
```

Out[ ]:

	artist_id	full_name	first_name	middle_names	last_name	nationality	style	bi
0	500	Pierre-Auguste Renoir	Pierre	Auguste	Renoir	French	Impressionist	18
1	501	Alexandre Cabanel	Alexandre	NaN	Cabanel	French	Classicist	18
2	502	James Ensor	James	NaN	Ensor	Belgian	Expressionist	18
3	503	Maximilien Luce	Maximilien	NaN	Luce	French	Pointillist	18
4	504	August Macke	August	NaN	Macke	German	Expressionist	18

In [ ]:
```
artist.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421 entries, 0 to 420
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   artist_id     421 non-null    int64 
 1   full_name     421 non-null    object
 2   first_name    421 non-null    object
 3   middle_names  148 non-null    object
 4   last_name     421 non-null    object
 5   nationality   421 non-null    object
 6   style         421 non-null    object
 7   birth         421 non-null    int64 
 8   death         421 non-null    int64 
dtypes: int64(3), object(6)
memory usage: 29.7+ KB
```

In [ ]:
```
artist.shape
```

Out[ ]: (421, 9)

In [ ]:
```
artist.duplicated().sum()
```

Out[ ]: 0

In [ ]:
```
artist.isnull().sum()
```

Out[ ]:
```
artist_id         0
full_name         0
first_name        0
middle_names    273
last_name         0
nationality       0
style             0
birth             0
death             0
dtype: int64
```

In [ ]:
```
artist.drop(columns=['middle_names'],axis=1,inplace=True)
artist.head()
```

Out[ ]:

	artist_id	full_name	first_name	last_name	nationality	style	birth	death
0	500	Pierre-Auguste Renoir	Pierre	Renoir	French	Impressionist	1841	1919
1	501	Alexandre Cabanel	Alexandre	Cabanel	French	Classicist	1823	1889
2	502	James Ensor	James	Ensor	Belgian	Expressionist	1860	1949
3	503	Maximilien Luce	Maximilien	Luce	French	Pointillist	1858	1941
4	504	August Macke	August	Macke	German	Expressionist	1887	1914

In [ ]:
```
# canvas_size
canvas_size.head()
```

Out[ ]:

	size_id	width	height	label
0	20	20	NaN	20" Long Edge
1	24	24	NaN	24" Long Edge
2	30	30	NaN	30" Long Edge
3	36	36	NaN	36" Long Edge
4	40	40	NaN	40" Long Edge

In [ ]:
```
canvas_size.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   size_id  200 non-null    int64
 1   width    200 non-null    int64
 2   height   193 non-null    float64
 3   label    200 non-null    object
dtypes: float64(1), int64(2), object(1)
memory usage: 6.4+ KB
```

In [ ]:
```
canvas_size.isnull().sum()
```

Out[ ]:
```
size_id    0
width      0
height     7
label      0
dtype: int64
```

In [ ]:
```
canvas_size.shape
```

Out[ ]:
```
(200, 4)
```

In [ ]:
```
canvas_size.fillna(0,inplace=True)
canvas_size.isnull().sum()
```

Out[ ]:
```
size_id    0
width      0
height     0
label      0
dtype: int64
```

In [ ]:
```
# image_link
image_link.head()
```

Out[ ]:

	work_id	url	thumbnai
0	181978	https://v5.airtableusercontent.com/v1/15/15/16...	https://v5.airtableusercontent.com/v1/
1	173188	https://v5.airtableusercontent.com/v1/15/15/16...	https://v5.airtableusercontent.com/v1/
2	194065	https://v5.airtableusercontent.com/v1/15/15/16...	https://v5.airtableusercontent.com/v1/
3	129337	https://v5.airtableusercontent.com/v1/15/15/16...	https://v5.airtableusercontent.com/v1/
4	141073	https://v5.airtableusercontent.com/v1/15/15/16...	https://v5.airtableusercontent.com/v1/

In [ ]:
```
image_link.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14775 entries, 0 to 14774
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   work_id             14775 non-null  int64 
 1   url                 14775 non-null  object
 2   thumbnail_small_url  14773 non-null  object
 3   thumbnail_large_url  14773 non-null  object
dtypes: int64(1), object(3)
memory usage: 461.8+ KB
```

In [ ]:
```
image_link.isnull().sum()
```

Out[ ]:
```
work_id               0
url                   0
thumbnail_small_url    2
thumbnail_large_url    2
dtype: int64
```

In [ ]:
```
image_link.shape
```

Out[ ]:
```
(14775, 4)
```

In [ ]:
```
image_link.dropna(subset=['thumbnail_small_url','thumbnail_large_url'],inplace=True
image_link.isnull().sum()
```

Out[ ]:
```
work_id               0
url                   0
thumbnail_small_url    0
thumbnail_large_url    0
dtype: int64
```

In [ ]:
```
# museum
museum.head()
```

Out[ ]:

	museum_id	name	address	city	state	postal	country	phone	
0	30	The Museum of Modern Art	11 W 53rd St	New York	NY	10019	USA	+1 212 708-9400	
1	31	Pushkin State Museum of Fine Arts	12 Ulitsa Volkhonka	Moscow	NaN	119019	Russia	+7 495 697-95-78	
2	32	National Gallery of Victoria	180 St Kilda Rd	Melbourne	Victoria	3004	Australia	+61 (0)3 8620 2222	
3	33	São Paulo Museum of Art	Av. Paulista, 1578 - Bela Vista	São Paulo	NaN	01310-200	Brazil	+55 11 3149-5959	
4	34	The State Hermitage Museum	Palace Square	2	Sankt-Peterburg	190000	Russia	7 812 710-90-79	https

In [ ]:
```
museum.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 57 entries, 0 to 56
Data columns (total 9 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   museum_id  57 non-null     int64 
 1   name       57 non-null     object
 2   address    57 non-null     object
 3   city       57 non-null     object
 4   state      38 non-null     object
 5   postal     50 non-null     object
 6   country    57 non-null     object
 7   phone      57 non-null     object
 8   url        57 non-null     object
dtypes: int64(1), object(8)
memory usage: 4.1+ KB
```

In [ ]:
```
museum.shape
```

Out[ ]: (57, 9)

In [ ]:
```
museum.isnull().sum()
```

Out[ ]:
```
museum_id     0
name          0
address       0
city          0
state        19
postal        7
country       0
phone         0
url           0
dtype: int64
```

In [ ]:
```
# museum_hours
museum_hours.head()
```

Out[ ]:

	museum_id	day	open	close
0	30	Sunday	10:30:AM	05:30:PM
1	30	Monday	10:30:AM	05:30:PM
2	30	Tuesday	10:30:AM	05:30:PM
3	30	Wednesday	10:30:AM	05:30:PM
4	30	Thusday	10:30:AM	05:30:PM

In [ ]:
```
museum_hours.shape
```

Out[ ]:
```
(351, 4)
```

In [ ]:
```
museum_hours.isnull().sum()
```

Out[ ]:
```
museum_id    0
day          0
open         0
close        0
dtype: int64
```

In [ ]:
```
# product_size
product_size.head()
```

Out[ ]:

	work_id	size_id	sale_price	regular_price
0	160228	24	85	85
1	160228	30	95	95
2	160236	24	85	85
3	160236	30	95	95
4	160244	24	85	85

In [ ]:
```
product_size.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110347 entries, 0 to 110346
Data columns (total 4 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   work_id        110347 non-null  int64
 1   size_id        110347 non-null  object
 2   sale_price     110347 non-null  int64
 3   regular_price  110347 non-null  int64
dtypes: int64(3), object(1)
memory usage: 3.4+ MB
```

In [ ]:
```
product_size['size_id'].unique()
```

Out[ ]:
```
array(['24', '30', '3024', '3226', '3629', '4030', '4836', '6048', '6854',
       '3624', '4630', '5436', '6040', '7248', '20', '2430', '2632',
       '2936', '3040', '3648', '4860', '5468', '36', '2436', '3046',
       '3654', '4060', '4872', '1620', '2024', '1624', '2030', '40', '48',
       '4020', '4824', '6030', '7236', '8040', '9648', '2424', '3030',
       '3636', '4040', '4848', '6060', '2016', '2420', '2020', '2416',
       '3020', '#VALUE!', '1632', '2040', '2448', '3060', '3672', '4080',
       '16.223', '28.836', '39.532', '3216', '56', '3628.7', '32.226',
       '3225.5', '4896', '3218', '32.426', '3239.8', '2936.5', '3225.6',
       '3223.7', '36.329', '36.432', '36.532', '27.326', '31.944',
       '28.525', '3232', '57.745', '4434', '3225.9', '2417.8', '7171',
       '5136.5', '5439.4', '22.43', '26.232', '2429', '50.635', '39.632',
       '3225.8', '3730', '3627', '35.652', '4235.5', '5062.6', '2823.2',
       '2419.7', '25.632', '37.428', '2921.7', '2921.5', '2639.8',
       '32.251', '31.525', '5936.2', '2923.6', '2923.7', '25.732',
       '2632.1', '57.542', '3536.5', '38.531', '5691.2', '3225.7',
       '32.326', '39.526', '23.529', '4023.7', '36.429', '36.326',
       '36.829', '32.524', '36.529', '39.435', '79.163', '36.626',
       '3239.4', '39.429', '5844.9', '3239.2', '57.545', '25.832',
       '32.332', '28.824', '3628.6', '4635.4', '4635', '2632.3', '50.465',
       '4835.8', '19.524', '28.936', '3246', '3628.5', '44.534', '3628.8',
       '45.835', '28.537', '32.126', '28.724', '3729.3', '37.53',
       '29.137', '3527.6', '2227.2', '4532', '3123.6', '3729.9', '26.332',
       '2936.6', '35.628'], dtype=object)
```

In [ ]:
```
product_size['size_id'] = pd.to_numeric(product_size['size_id'],errors='coerce')
product_size['size_id'].unique()
```

Out[ ]:
```
array([  24.   ,   30.   , 3024.   , 3226.   , 3629.   , 4030.   ,
       4836.   , 6048.   , 6854.   , 3624.   , 4630.   , 5436.   ,
       6040.   , 7248.   ,   20.   , 2430.   , 2632.   , 2936.   ,
       3040.   , 3648.   , 4860.   , 5468.   ,   36.   , 2436.   ,
       3046.   , 3654.   , 4060.   , 4872.   , 1620.   , 2024.   ,
       1624.   , 2030.   ,   40.   ,   48.   , 4020.   , 4824.   ,
       6030.   , 7236.   , 8040.   , 9648.   , 2424.   , 3030.   ,
       3636.   , 4040.   , 4848.   , 6060.   , 2016.   , 2420.   ,
       2020.   , 2416.   , 3020.   ,      nan, 1632.   , 2040.   ,
       2448.   , 3060.   , 3672.   , 4080.   ,   16.223,   28.836,
         39.532, 3216.   ,   56.   , 3628.7  ,   32.226, 3225.5  ,
       4896.   , 3218.   ,   32.426, 3239.8  , 2936.5  , 3225.6  ,
       3223.7  ,   36.329,   36.432,   36.532,   27.326,   31.944,
         28.525, 3232.   ,   57.745, 4434.   , 3225.9  , 2417.8  ,
       7171.   , 5136.5  , 5439.4  ,   22.43 ,   26.232, 2429.   ,
         50.635,   39.632, 3225.8  , 3730.   , 3627.   ,   35.652,
       4235.5  , 5062.6  , 2823.2  , 2419.7  ,   25.632,   37.428,
       2921.7  , 2921.5  , 2639.8  ,   32.251,   31.525, 5936.2  ,
       2923.6  , 2923.7  ,   25.732, 2632.1  ,   57.542, 3536.5  ,
         38.531, 5691.2  , 3225.7  ,   32.326,   39.526,   23.529,
       4023.7  ,   36.429,   36.326,   36.829,   32.524,   36.529,
         39.435,   79.163,   36.626, 3239.4  ,   39.429, 5844.9  ,
       3239.2  ,   57.545,   25.832,   32.332,   28.824, 3628.6  ,
       4635.4  , 4635.   , 2632.3  ,   50.465, 4835.8  ,   19.524,
         28.936, 3246.   , 3628.5  ,   44.534, 3628.8  ,   45.835,
         28.537,   32.126,   28.724, 3729.3  ,   37.53 ,   29.137,
       3527.6  , 2227.2  , 4532.   , 3123.6  , 3729.9  ,   26.332,
       2936.6  ,   35.628])
```

In [ ]:
```
product_size.shape
```

Out[ ]:
```
(110347, 4)
```

In [ ]:
```
product_size.isnull().sum()
```

Out[ ]:
```
work_id           0
size_id         212
sale_price        0
regular_price     0
dtype: int64
```

In [ ]:
```
product_size.dropna(subset=['size_id'],inplace=True)
```

In [ ]:
```
product_size.shape
```

Out[ ]:
```
(110135, 4)
```

In [ ]:
```
# subject
subject.head()
```

Out[ ]:

	work_id	subject
0	160228	Still-Life
1	160236	Still-Life
2	160244	Still-Life
3	160252	Still-Life
4	160260	Still-Life

In [ ]:
```
subject.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6771 entries, 0 to 6770
Data columns (total 2 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   work_id  6771 non-null   int64 
 1   subject  6771 non-null   object
dtypes: int64(1), object(1)
memory usage: 105.9+ KB
```

In [ ]:
```
subject.isnull().sum()
```

Out[ ]:
```
work_id    0
subject    0
dtype: int64
```

In [ ]:
```
subject.shape
```

Out[ ]:
```
(6771, 2)
```

In [ ]:
```
# work
work.head()
```

Out[ ]:

	work_id	name	artist_id	style	museum_id
0	160228	Still Life with Flowers and a Watch	615	Baroque	43.0
1	160236	Still Life with Fruit and a Beaker on a Cock's...	615	Baroque	43.0
2	160244	Still Life with Fruit and a Goldfinch	615	Baroque	43.0
3	160252	Still Life with Fruit and Oysters	615	Baroque	43.0
4	160260	Still Life with Fruit, Oysters, and a Porcelai...	615	Baroque	43.0

In [ ]:
```
work.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14776 entries, 0 to 14775
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   work_id    14776 non-null  int64  
 1   name       14776 non-null  object 
 2   artist_id  14776 non-null  int64  
 3   style      13490 non-null  object 
 4   museum_id  4553 non-null   float64
dtypes: float64(1), int64(2), object(2)
memory usage: 577.3+ KB
```

In [ ]:
```
work.isnull().sum()
```

Out[ ]:
```
work_id          0
name             0
artist_id        0
style         1286
museum_id    10223
dtype: int64
```

In [ ]:
```
work.shape
```

Out[ ]:
```
(14776, 5)
```

In [ ]:
```
work[work['style'].isnull()].head()
```

Out[ ]:

	work_id	name	artist_id	style	museum_id
14	24532	Jacob A. Stamler Departing Le Havre	563	NaN	NaN
15	124470	Kaleda off Le Havre	563	NaN	NaN
16	124479	R. Bell & Co. Steamship Bothal in a Heavy ...	563	NaN	NaN
17	124488	Steam Sailing Ship Finsbury in a Stormy Sea	563	NaN	NaN
18	124497	The American Ship Olive S Southard in French W...	563	NaN	NaN

In [ ]:
```
work['style'].unique()
```

Out[ ]:
```
array(['Baroque', nan, 'Neo-Classicism', 'Renaissance', 'Expressionism',
       'American Landscape', 'Post-Impressionism', 'Classicism',
       'Avant-Garde', 'Impressionism', 'Cubism', 'Rococo', 'Realism',
       'Fauvism', 'Nabi', 'Symbolism', 'Naturalism', 'American Art',
       'Romanticism', 'Orientalism', 'Art Nouveau', 'Surrealism',
       'Pointillism', 'Japanese Art'], dtype=object)
```

In [ ]:
```
work['style'].replace(np.nan, work['style'].mode()[0],inplace=True)
work['style'].unique()
```

Out[ ]:
```
array(['Baroque', 'Impressionism', 'Neo-Classicism', 'Renaissance',
       'Expressionism', 'American Landscape', 'Post-Impressionism',
       'Classicism', 'Avant-Garde', 'Cubism', 'Rococo', 'Realism',
       'Fauvism', 'Nabi', 'Symbolism', 'Naturalism', 'American Art',
       'Romanticism', 'Orientalism', 'Art Nouveau', 'Surrealism',
       'Pointillism', 'Japanese Art'], dtype=object)
```

In [ ]:
```
files
```

Out[ ]:
```
['artist.csv',
 'canvas_size.csv',
 'image_link.csv',
 'museum.csv',
 'museum_hours.csv',
 'product_size.csv',
 'subject.csv',
 'work.csv']
```

In [ ]:
```
artist.to_sql('artist',con=conn,if_exists='replace',index=False)
canvas_size.to_sql('canvas_size',con=conn,if_exists='replace',index=False)
image_link.to_sql('image_link',con=conn,if_exists='replace',index=False)
museum.to_sql('museum',con=conn,if_exists='replace',index=False)
museum_hours.to_sql('museum_hours',con=conn,if_exists='replace',index=False)
product_size.to_sql('product_size',con=conn,if_exists='replace',index=False)
subject.to_sql('subject',con=conn,if_exists='replace',index=False)
work.to_sql('work',con=conn,if_exists='replace',index=False)
```

Out[ ]:
```
14776
```