Famous Paintings Database

An SQL Case Study and Workflow Documentation

Index

I. Introduction

The following paper serves to document the SQL knowledge and skills I acquired through dedicated self-study and practice.

In order to demonstrate my abilities in a practical manner, I have

- acquired a dataset about famous paintings,
- transformed it into a database using Python and PostgreSQL,
- created an Entity Relationship Diagram (ERD) to clearly visualize relations between information tables,
- and performed SQL queries giving insights into the data based on a variety of questions.

The idea for this project was found in this **YouTube tutorial**. The execution and solution of the tasks was all up to me and I furthermore have extended the concepts and approaches introduced in the video to make this SQL Case Study my own.

II. Data Acquisition

The **Famous Paintings Dataset** was provided by the user **mexwell** on **kaggle** and has originally been downloaded from **data.world**. It consists of the following 8 csv files with a total of 41 columns:

artist.csv	9 columns, 920 rows
canvas_size.csv	4 columns, 9648 rows
image_link.csv	4 columns, 210k rows
museum.csv	9 columns, 86 rows
museum_hours.csv	4 columns, 86 rows
product_size.csv	4 columns, 210k rows
subject.csv	2 columns, 210k rows
work.csv	5 columns, 210k rows

No further explanation has been given about the nature of the data.
The respective column contents of each file can be viewed in the ERD on page 6.

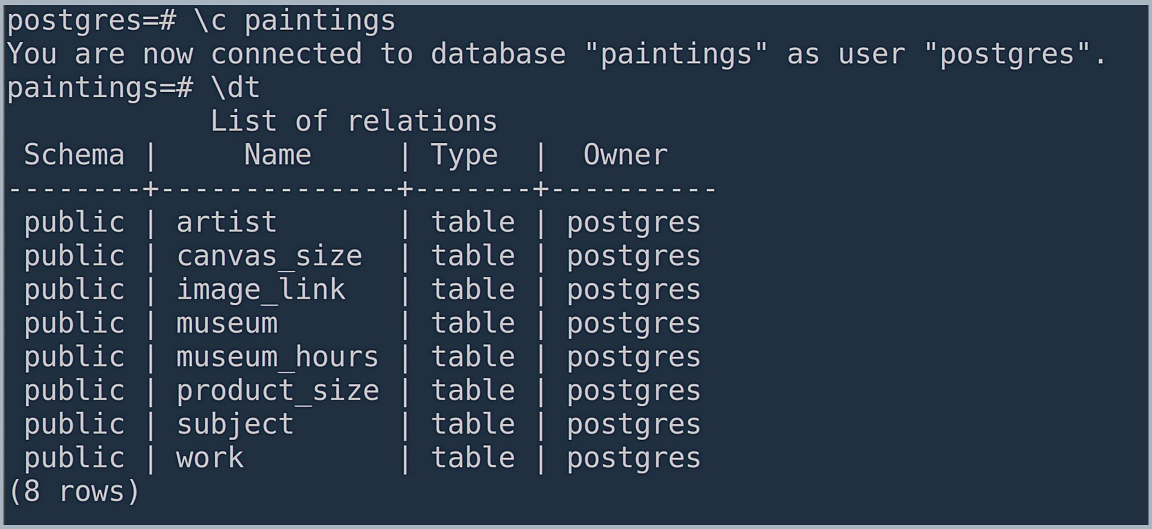
III.   Database Creation

Before being able to gain insights on the data using SQL, the dataset needs to be transformed into a database. This will be achieved using **PostgreSQL** as the object–relational database management system (ORDBMS) and a **Python** script which automatically creates the database and imports each csv file as a table into that database:

```
#!/usr/bin/env python3

import pandas as pd
from sqlalchemy import create_engine
from sqlalchemy.engine.url import URL
import os
import psycopg2

# Get user input
pw = input("Postgres Password: ")
db_name = input("New Database: ")


# Create New Database
conn = psycopg2.connect(host="localhost", dbname=f"postgres",
                        user="postgres", password=f"{pw}", port=5432)
cur = conn.cursor()
conn.autocommit = True

cur.execute(f"CREATE DATABASE {db_name};")

cur.close()
conn.close()


# Use sqlalchemy to connect to new db
url = URL.create(
    drivername='postgresql',
    username='postgres',
    password=pw,
    host='localhost',
    port=5432,
    database=db_name,
).render_as_string(hide_password=False)
db = create_engine(url)
new_db_conn = db.connect()


# Get csv file path and csv file names
path = "/home/lisa/SQL/" + input('CSV File Folder Name: ')
files = [f for f in os.listdir(path) if os.path.isfile(os.path.join(path, f))]
file_names = [name.split(".")[0] for name in files]

# Convert csv files to pd.DataFrame and import into postgres
for file in file_names:
    df = pd.read_csv(path + f"/{file}.csv")
    df.to_sql(file, con=new_db_conn, if_exists='replace', index=False)


# Success Message:
print(f"Created new database: {db_name} \n"
      f"Created {len(file_names)} new tables: \n"
      f"{file_names}")
```

On my **GitHub** you can find the **source code** of my automatic_import_csv_to_postgres_db Python script. The **Readme** file explains necessary user inputs and details usage remarks, as well as improvement ideas on how to make the script more versatile and stable in possible future use cases.

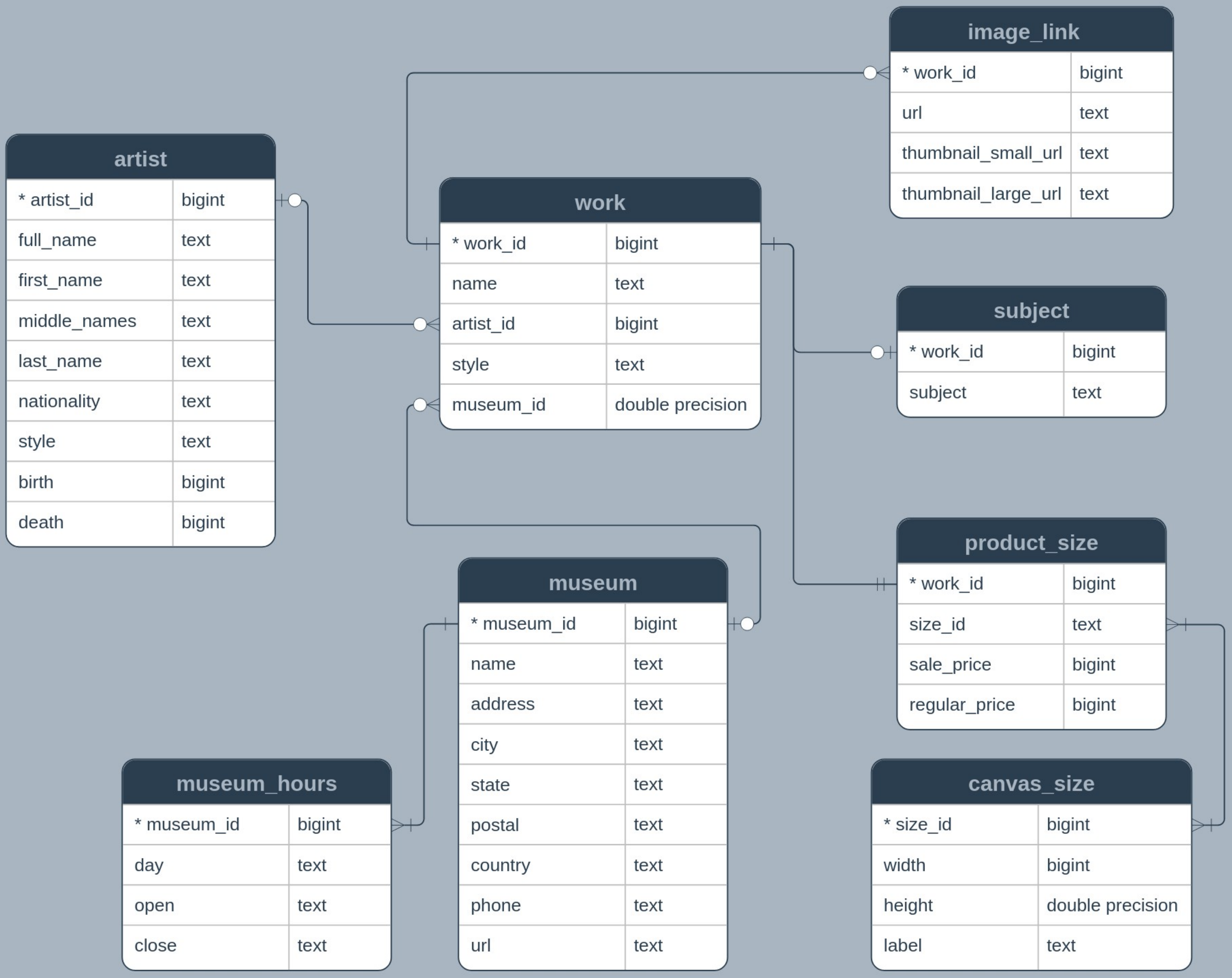
As a result of executing this script, a PostgreSQL database with the name “paintings” with its respective tables has successfully been created and can be connected to:



```
postgres=# \c paintings
You are now connected to database "paintings" as user "postgres".
paintings=# \dt
            List of relations
 Schema |     Name      | Type  |  Owner
--------+--------------+-------+----------
 public | artist        | table | postgres
 public | canvas_size   | table | postgres
 public | image_link    | table | postgres
 public | museum        | table | postgres
 public | museum_hours  | table | postgres
 public | product_size  | table | postgres
 public | subject       | table | postgres
 public | work          | table | postgres
(8 rows)
```

IV. Entity Relationship Diagram (ERD)

The following Famous Paintings Database Schema was created with smartdraw and displays the relations between the various tables.



image_link

* work_id	bigint
url	text
thumbnail_small_url	text
thumbnail_large_url	text

artist

* artist_id	bigint
full_name	text
first_name	text
middle_names	text
last_name	text
nationality	text
style	text
birth	bigint
death	bigint

work

* work_id	bigint
name	text
artist_id	bigint
style	text
museum_id	double precision

subject

* work_id	bigint
subject	text

product_size

* work_id	bigint
size_id	text
sale_price	bigint
regular_price	bigint

museum

* museum_id	bigint
name	text
address	text
city	text
state	text
postal	text
country	text
phone	text
url	text

museum_hours

* museum_id	bigint
day	text
open	text
close	text

canvas_size

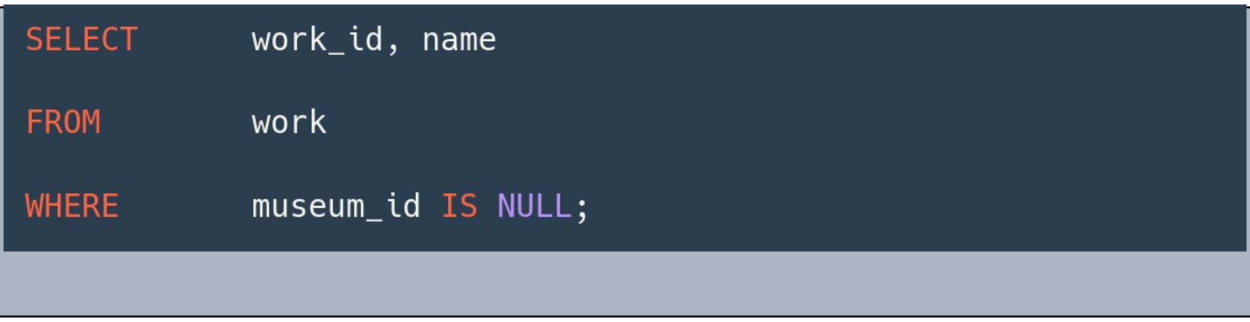
* size_id	bigint
width	bigint
height	double precision
label	text

V. Data Analysis

As a last step, meaningful insights about the data can be gained by performing SQL queries on the Famous Paintings Database. This analysis was conducted focusing on the posterior tasks and questions:

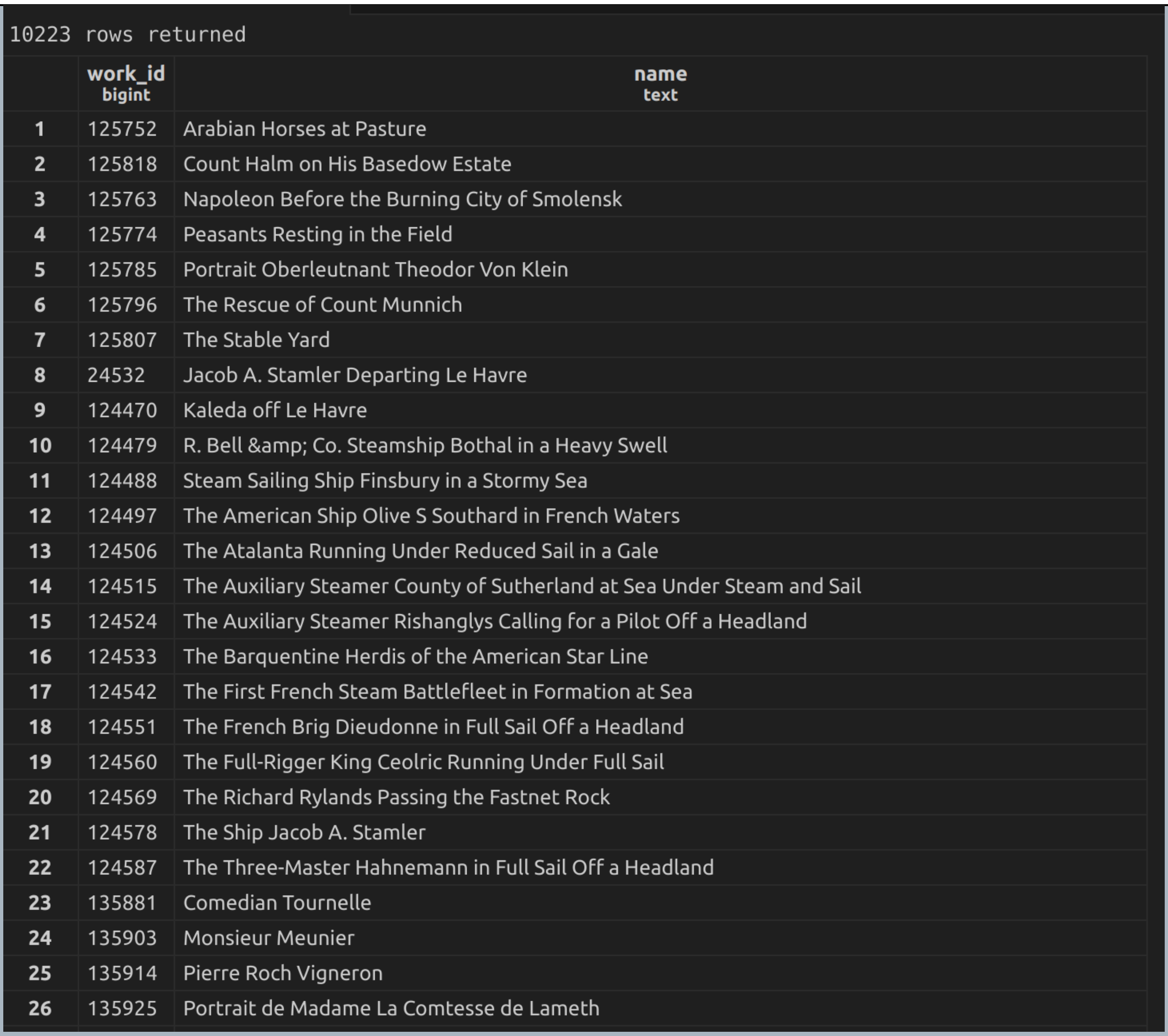
Task #1   Fetch all paintings not on display at any museum.

Input



```
SELECT      work_id, name

FROM        work

WHERE       museum_id IS NULL;
```

Output



```
10223 rows returned
```

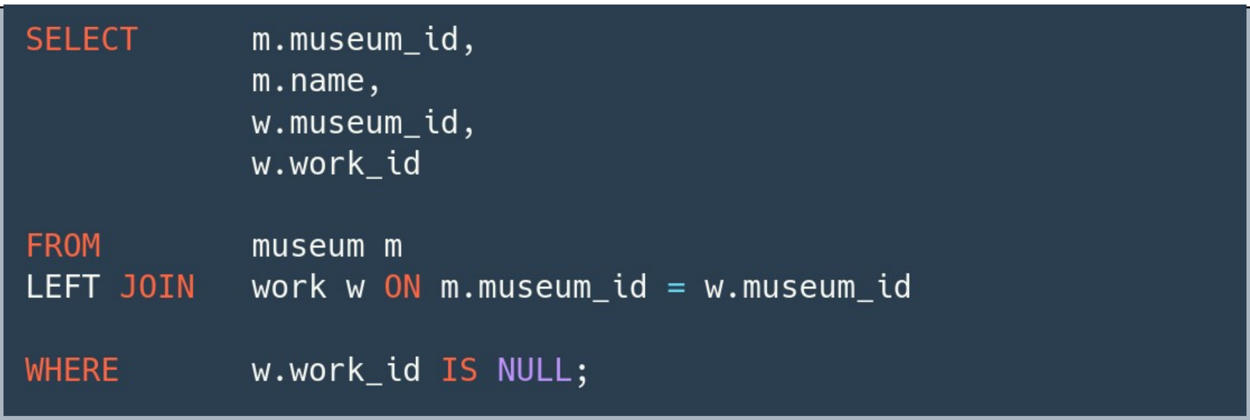
	work_id bigint	name text
1	125752	Arabian Horses at Pasture
2	125818	Count Halm on His Basedow Estate
3	125763	Napoleon Before the Burning City of Smolensk
4	125774	Peasants Resting in the Field
5	125785	Portrait Oberleutnant Theodor Von Klein
6	125796	The Rescue of Count Munnich
7	125807	The Stable Yard
8	24532	Jacob A. Stamler Departing Le Havre
9	124470	Kaleda off Le Havre
10	124479	R. Bell &amp; Co. Steamship Bothal in a Heavy Swell
11	124488	Steam Sailing Ship Finsbury in a Stormy Sea
12	124497	The American Ship Olive S Southard in French Waters
13	124506	The Atalanta Running Under Reduced Sail in a Gale
14	124515	The Auxiliary Steamer County of Sutherland at Sea Under Steam and Sail
15	124524	The Auxiliary Steamer Rishanglys Calling for a Pilot Off a Headland
16	124533	The Barquentine Herdis of the American Star Line
17	124542	The First French Steam Battlefleet in Formation at Sea
18	124551	The French Brig Dieudonne in Full Sail Off a Headland
19	124560	The Full-Rigger King Ceolric Running Under Full Sail
20	124569	The Richard Rylands Passing the Fastnet Rock
21	124578	The Ship Jacob A. Stamler
22	124587	The Three-Master Hahnemann in Full Sail Off a Headland
23	135881	Comedian Tournelle
24	135903	Monsieur Meunier
25	135914	Pierre Roch Vigneron
26	135925	Portrait de Madame La Comtesse de Lameth

There are 10,223 paintings not on display at any museum.

[screenshot limited to 26 results]

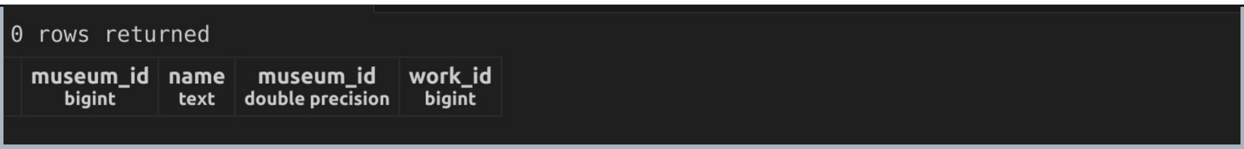
Task #2 Are there any museums without any paintings?

Input



```
SELECT      m.museum_id,
            m.name,
            w.museum_id,
            w.work_id

FROM        museum m
LEFT JOIN   work w ON m.museum_id = w.museum_id

WHERE       w.work_id IS NULL;
```

Output



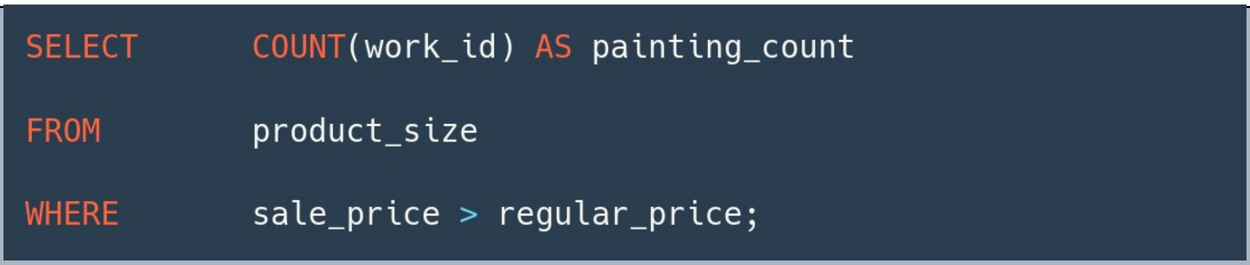
0 rows returned

	museum_id	name	museum_id	work_id
	bigint	text	double precision	bigint

There are no museums without any paintings.

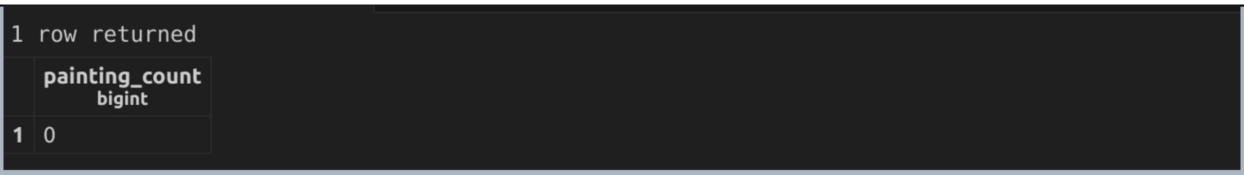
Task #3 How many paintings have a higher sale than regular price?

Input



```
SELECT      COUNT(work_id) AS painting_count

FROM        product_size

WHERE       sale_price > regular_price;
```

Output



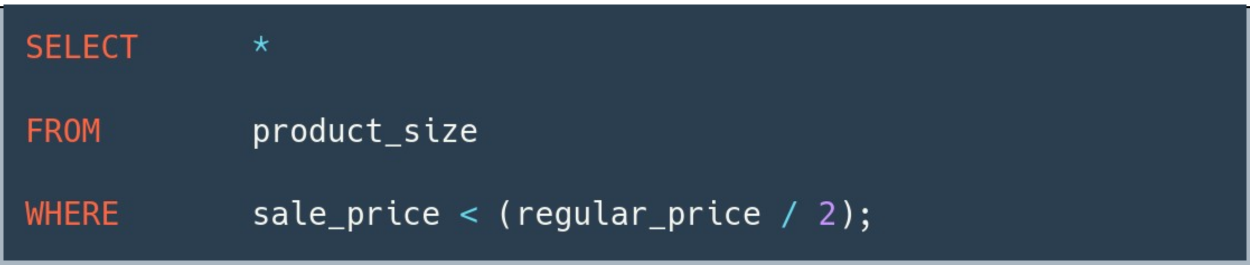
1 row returned

	painting_count
	bigint
1	0

No paintings have a higher sale than regular price.

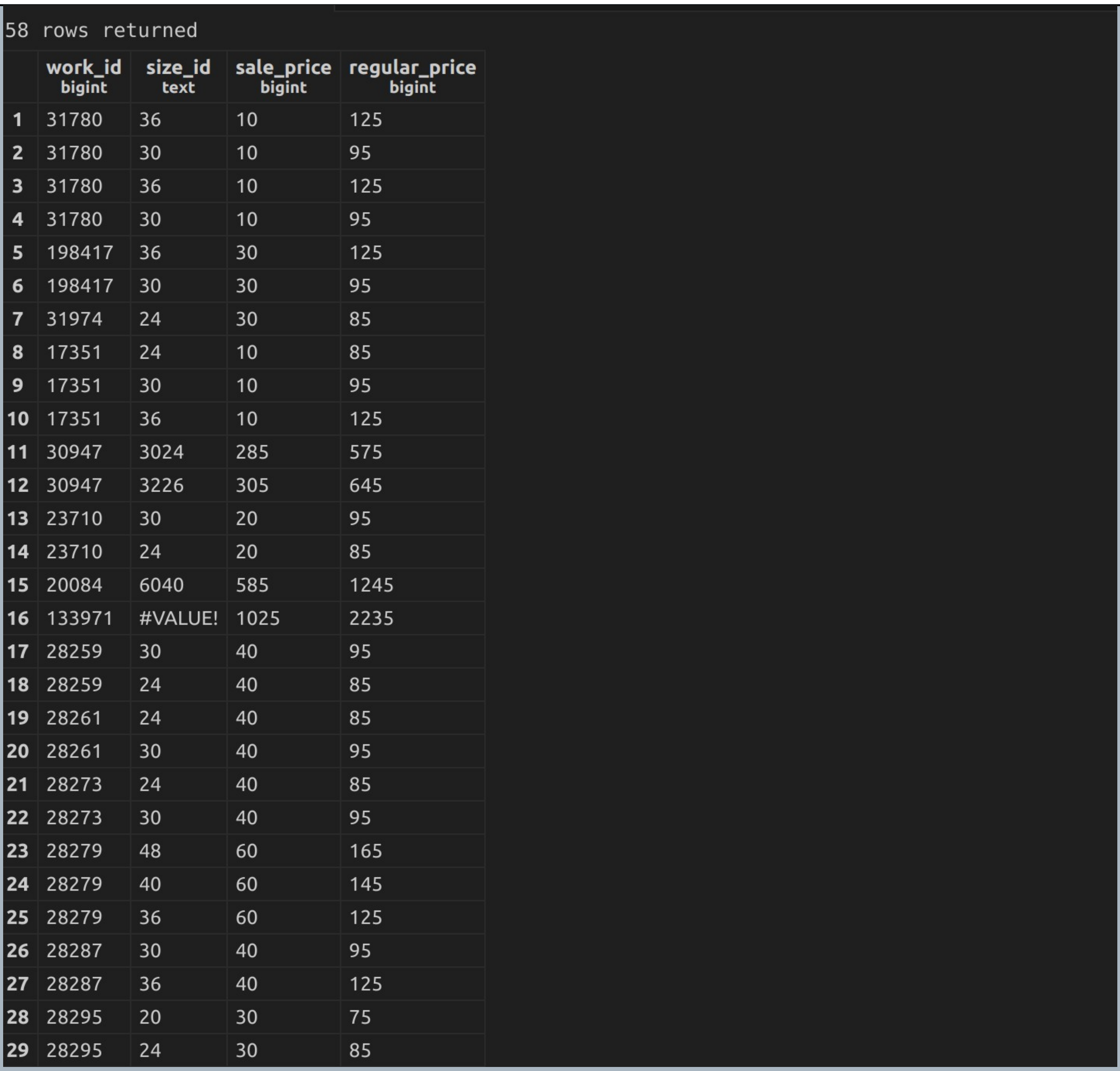
Task #4 Identify all paintings with sale prices less than 50% of their regular prices.

Input



```
SELECT      *

FROM        product_size

WHERE       sale_price < (regular_price / 2);
```

Output



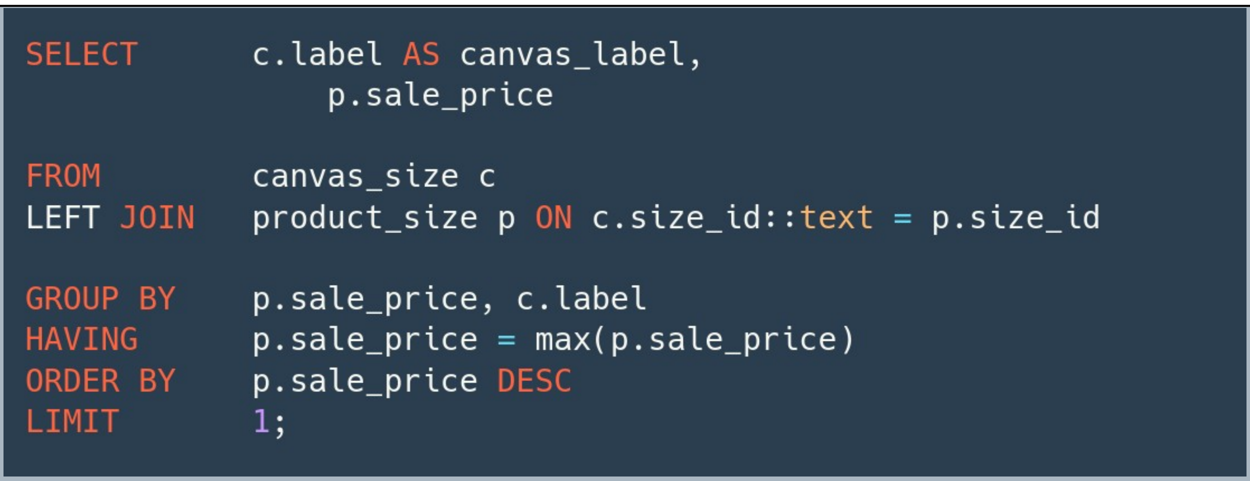
58 rows returned

	work_id bigint	size_id text	sale_price bigint	regular_price bigint
1	31780	36	10	125
2	31780	30	10	95
3	31780	36	10	125
4	31780	30	10	95
5	198417	36	30	125
6	198417	30	30	95
7	31974	24	30	85
8	17351	24	10	85
9	17351	30	10	95
10	17351	36	10	125
11	30947	3024	285	575
12	30947	3226	305	645
13	23710	30	20	95
14	23710	24	20	85
15	20084	6040	585	1245
16	133971	#VALUE!	1025	2235
17	28259	30	40	95
18	28259	24	40	85
19	28261	24	40	85
20	28261	30	40	95
21	28273	24	40	85
22	28273	30	40	95
23	28279	48	60	165
24	28279	40	60	145
25	28279	36	60	125
26	28287	30	40	95
27	28287	36	40	125
28	28295	20	30	75
29	28295	24	30	85

There are 58 paintings with a sale price being less than 50% of their regular price. *[screenshot limited to 29 results]*

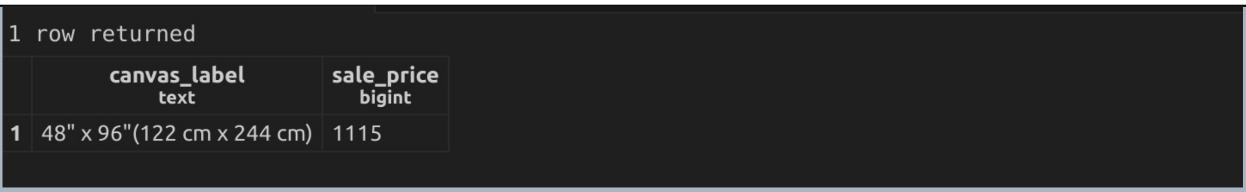
Task #5 Which canvas size costs the most?

Input



```
SELECT      c.label AS canvas_label,
                p.sale_price

FROM        canvas_size c
LEFT JOIN   product_size p ON c.size_id::text = p.size_id

GROUP BY    p.sale_price, c.label
HAVING      p.sale_price = max(p.sale_price)
ORDER BY    p.sale_price DESC
LIMIT       1;
```

Output



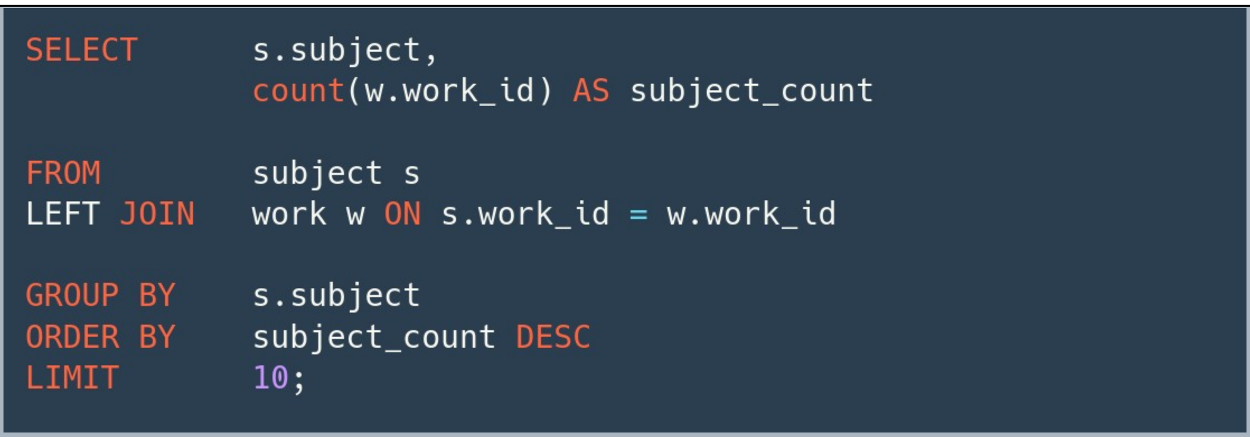
1 row returned

	canvas_label text	sale_price bigint
1	48" x 96"(122 cm x 244 cm)	1115

The most expensive canvas size at a sale price of 1115 is: 48” x 96” (122cm x 244cm)

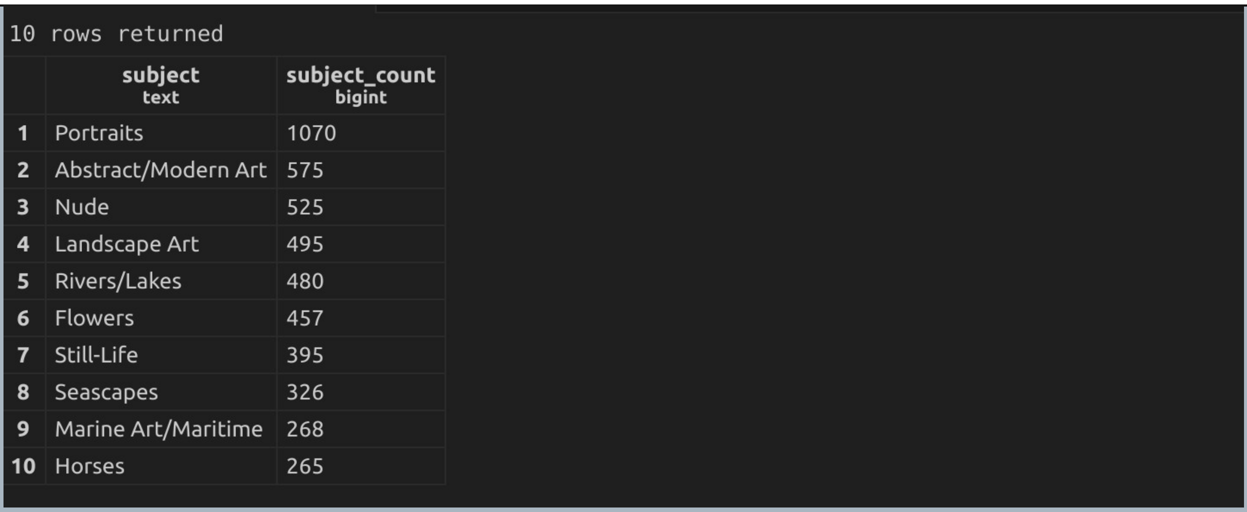
Task #6   Fetch the top 10 most common painting subjects.

Input



```
SELECT      s.subject,
            count(w.work_id) AS subject_count

FROM        subject s
LEFT JOIN   work w ON s.work_id = w.work_id

GROUP BY    s.subject
ORDER BY    subject_count DESC
LIMIT       10;
```

Output



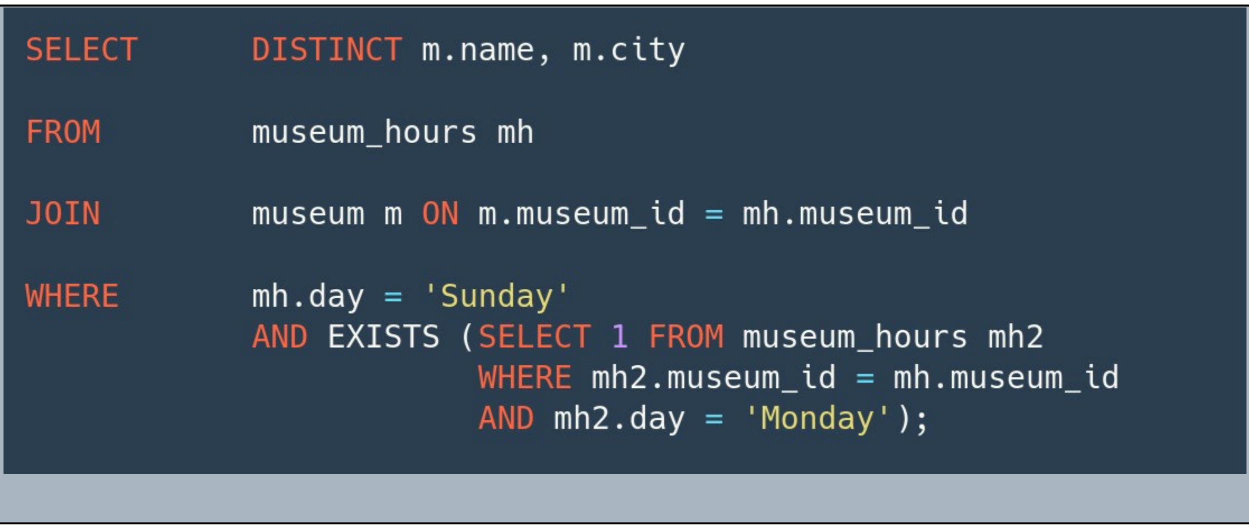
10 rows returned

	subject text	subject_count bigint
1	Portraits	1070
2	Abstract/Modern Art	575
3	Nude	525
4	Landscape Art	495
5	Rivers/Lakes	480
6	Flowers	457
7	Still-Life	395
8	Seascapes	326
9	Marine Art/Maritime	268
10	Horses	265

The top 10 most common painting subjects are the ones provided in the ranking above.

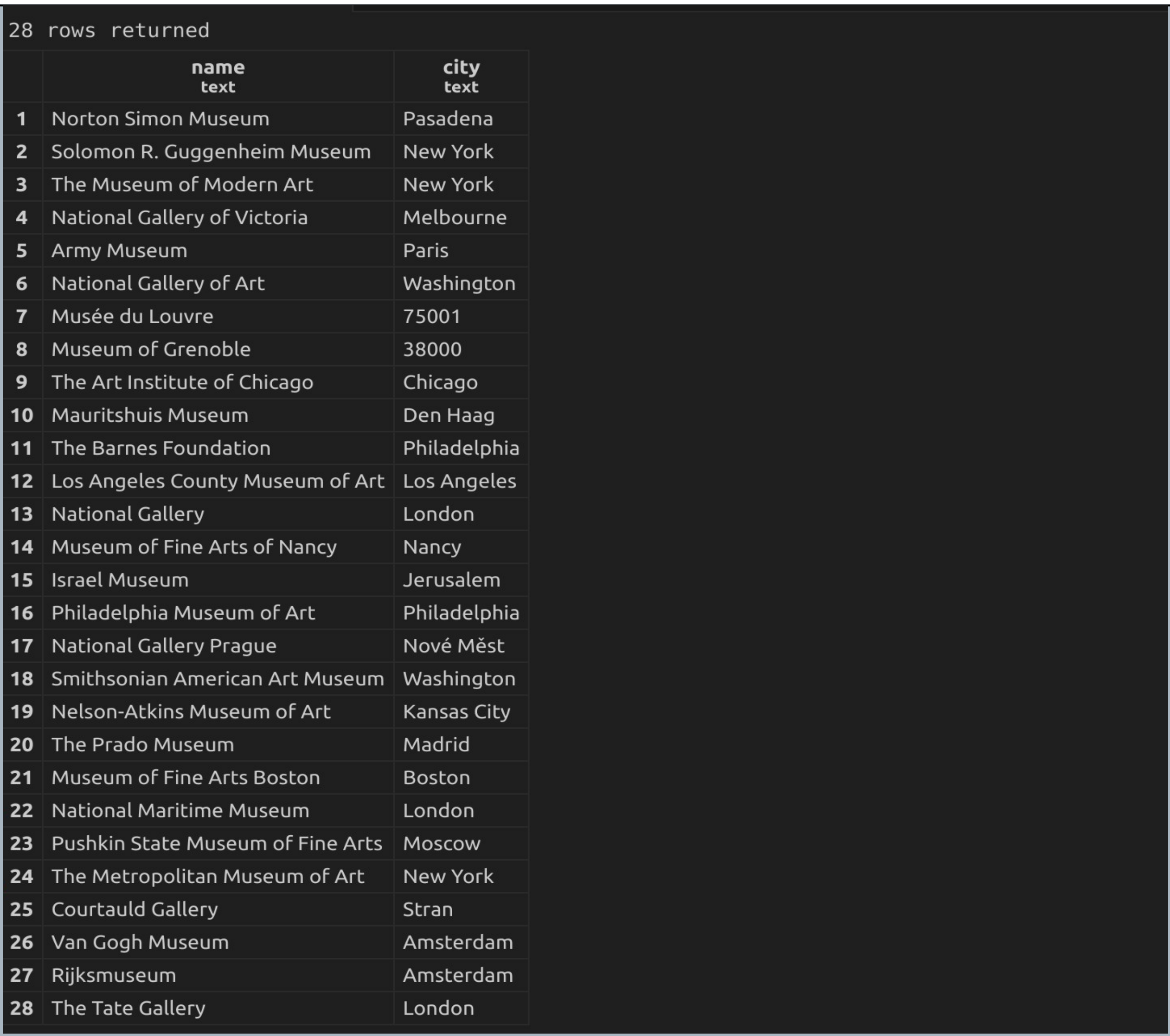
Task #7   Display name and city of the museums that are open on both Sunday and Monday.

Input



```
SELECT      DISTINCT m.name, m.city

FROM        museum_hours mh

JOIN        museum m ON m.museum_id = mh.museum_id

WHERE       mh.day = 'Sunday'
            AND EXISTS (SELECT 1 FROM museum_hours mh2
                        WHERE mh2.museum_id = mh.museum_id
                        AND mh2.day = 'Monday');
```

Output



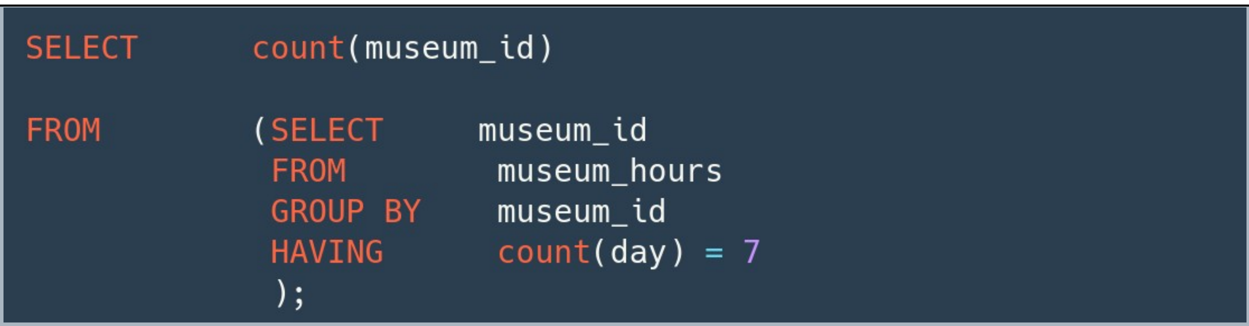
28 rows returned

	name text	city text
1	Norton Simon Museum	Pasadena
2	Solomon R. Guggenheim Museum	New York
3	The Museum of Modern Art	New York
4	National Gallery of Victoria	Melbourne
5	Army Museum	Paris
6	National Gallery of Art	Washington
7	Musée du Louvre	75001
8	Museum of Grenoble	38000
9	The Art Institute of Chicago	Chicago
10	Mauritshuis Museum	Den Haag
11	The Barnes Foundation	Philadelphia
12	Los Angeles County Museum of Art	Los Angeles
13	National Gallery	London
14	Museum of Fine Arts of Nancy	Nancy
15	Israel Museum	Jerusalem
16	Philadelphia Museum of Art	Philadelphia
17	National Gallery Prague	Nové Měst
18	Smithsonian American Art Museum	Washington
19	Nelson-Atkins Museum of Art	Kansas City
20	The Prado Museum	Madrid
21	Museum of Fine Arts Boston	Boston
22	National Maritime Museum	London
23	Pushkin State Museum of Fine Arts	Moscow
24	The Metropolitan Museum of Art	New York
25	Courtauld Gallery	Stran
26	Van Gogh Museum	Amsterdam
27	Rijksmuseum	Amsterdam
28	The Tate Gallery	London

All 28 museums that are open Sunday & Monday are displayed above.

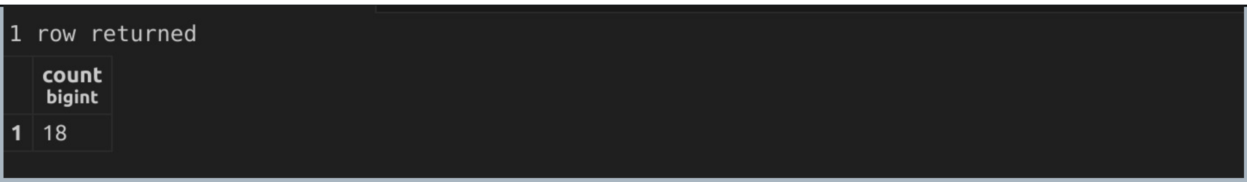
Task #8   How many museums are open every single day?

Input



```
SELECT      count(museum_id)

FROM        (SELECT     museum_id
             FROM        museum_hours
             GROUP BY    museum_id
             HAVING      count(day) = 7
             );
```

Output

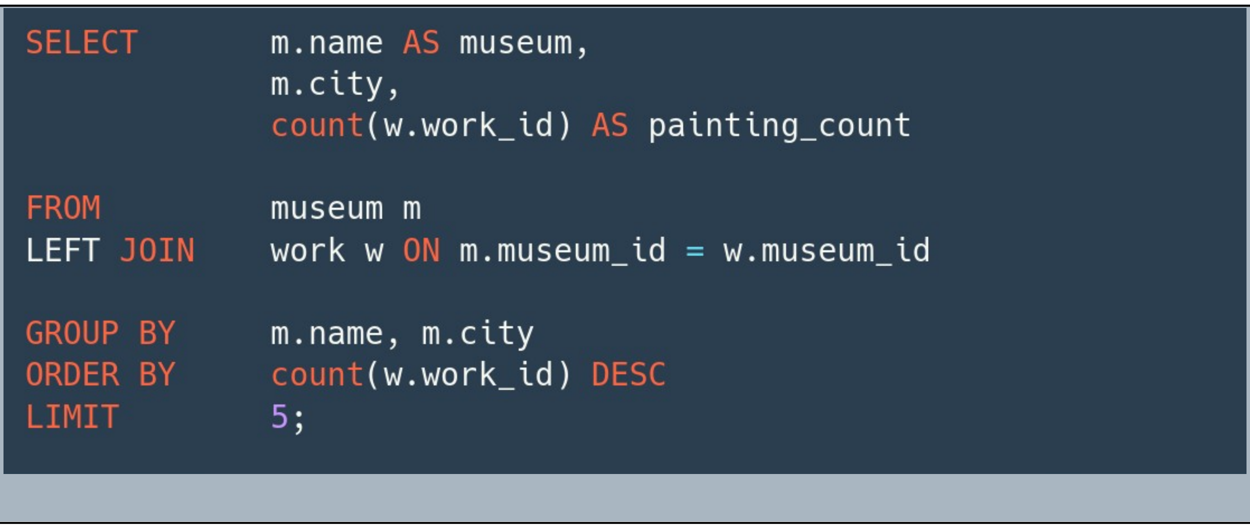


```
1 row returned
   count
   bigint
1  18
```

18 museums are open every single day.

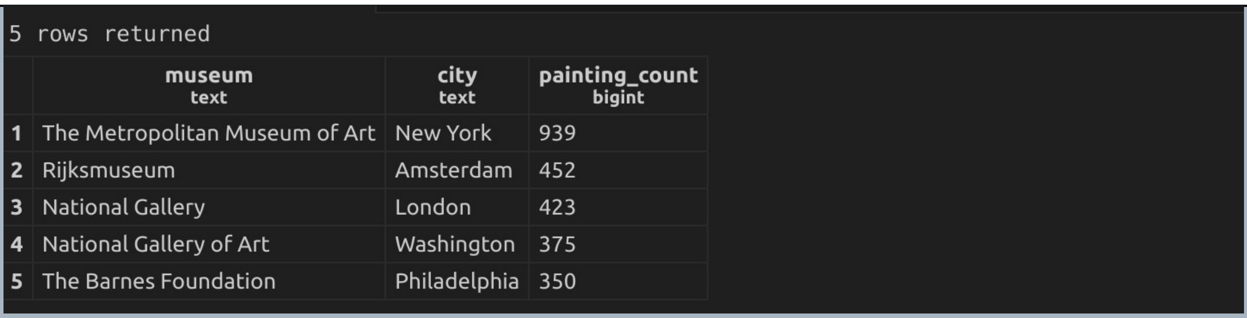
Task #9 Which top 5 museums have the most number of paintings?

Input



```
SELECT      m.name AS museum,
            m.city,
            count(w.work_id) AS painting_count

FROM        museum m
LEFT JOIN   work w ON m.museum_id = w.museum_id

GROUP BY    m.name, m.city
ORDER BY    count(w.work_id) DESC
LIMIT       5;
```

Output



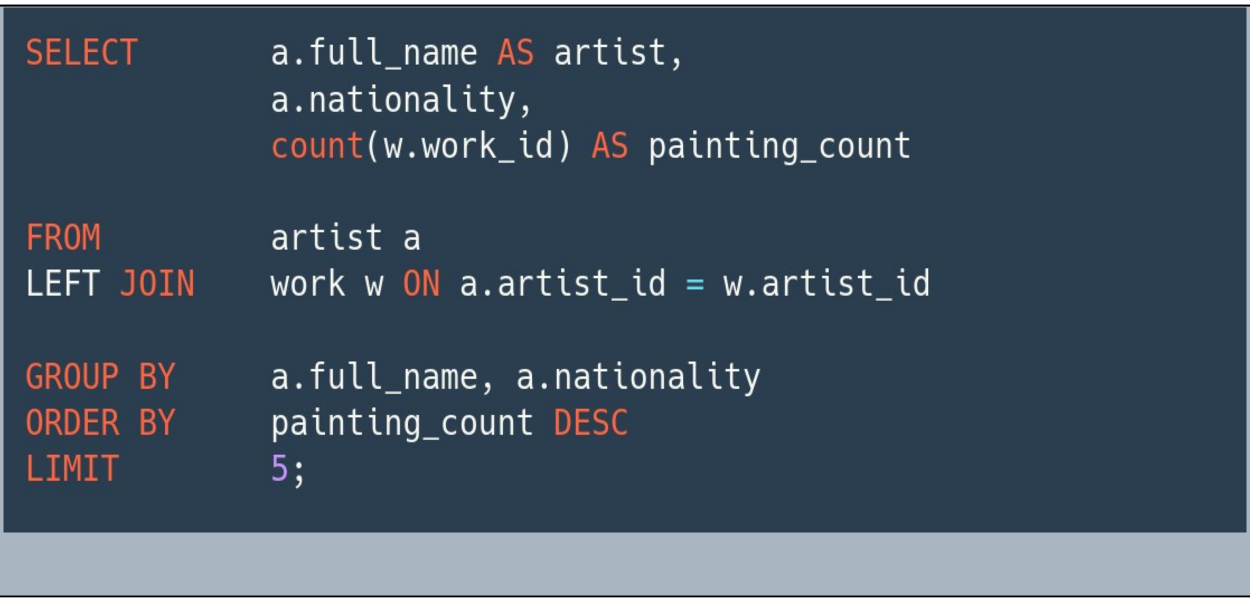
5 rows returned

	museum text	city text	painting_count bigint
1	The Metropolitan Museum of Art	New York	939
2	Rijksmuseum	Amsterdam	452
3	National Gallery	London	423
4	National Gallery of Art	Washington	375
5	The Barnes Foundation	Philadelphia	350

The museums with the top 5 most paintings are ranked above.

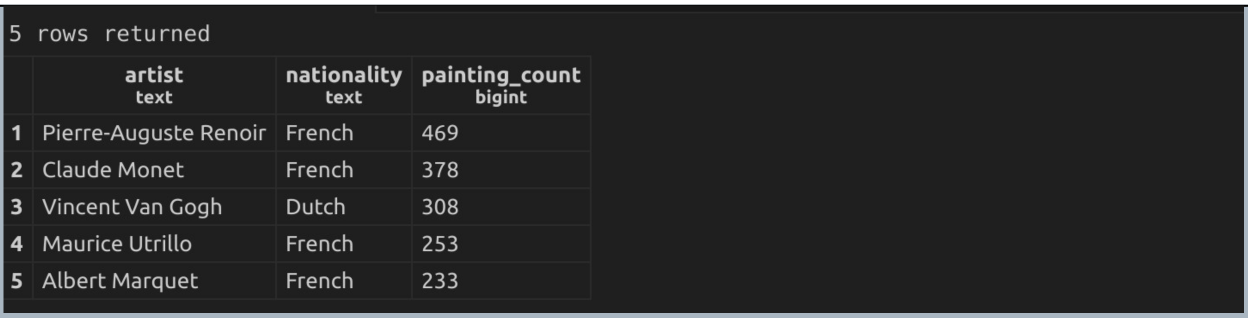
Task #10 Who are the top 5 artists with the most paintings?

Input



```
SELECT      a.full_name AS artist,
            a.nationality,
            count(w.work_id) AS painting_count

FROM        artist a
LEFT JOIN   work w ON a.artist_id = w.artist_id

GROUP BY    a.full_name, a.nationality
ORDER BY    painting_count DESC
LIMIT       5;
```

Output



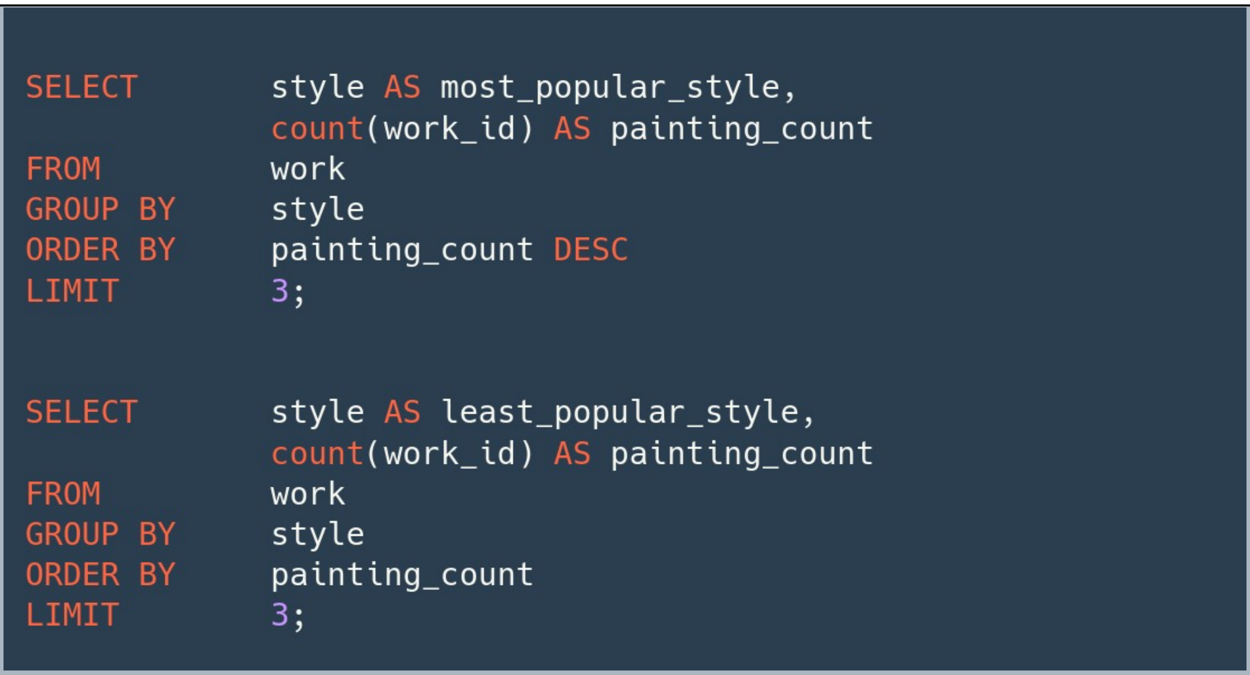
5 rows returned

	artist	nationality	painting_count
	text	text	bigint
1	Pierre-Auguste Renoir	French	469
2	Claude Monet	French	378
3	Vincent Van Gogh	Dutch	308
4	Maurice Utrillo	French	253
5	Albert Marquet	French	233

The artists with the top 5 most paintings are ranked above.

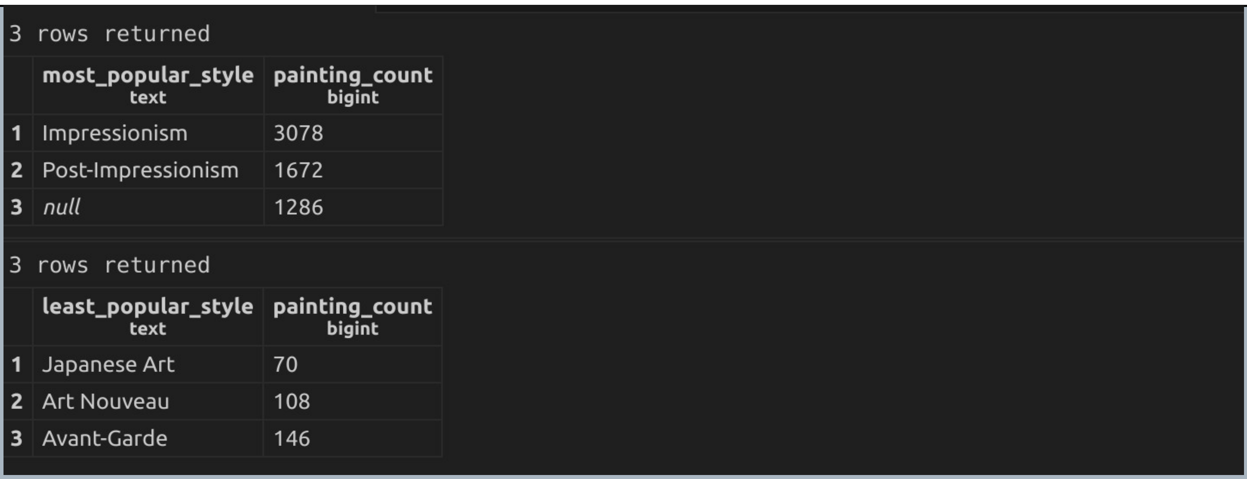
Task #11 Which are the 3 most popular and which are the 3 least popular painting styles?

Input



```
SELECT      style AS most_popular_style,
            count(work_id) AS painting_count
FROM        work
GROUP BY    style
ORDER BY    painting_count DESC
LIMIT       3;


SELECT      style AS least_popular_style,
            count(work_id) AS painting_count
FROM        work
GROUP BY    style
ORDER BY    painting_count
LIMIT       3;
```

Output



3 rows returned

	most_popular_style text	painting_count bigint
1	Impressionism	3078
2	Post-Impressionism	1672
3	*null*	1286

3 rows returned

	least_popular_style text	painting_count bigint
1	Japanese Art	70
2	Art Nouveau	108
3	Avant-Garde	146

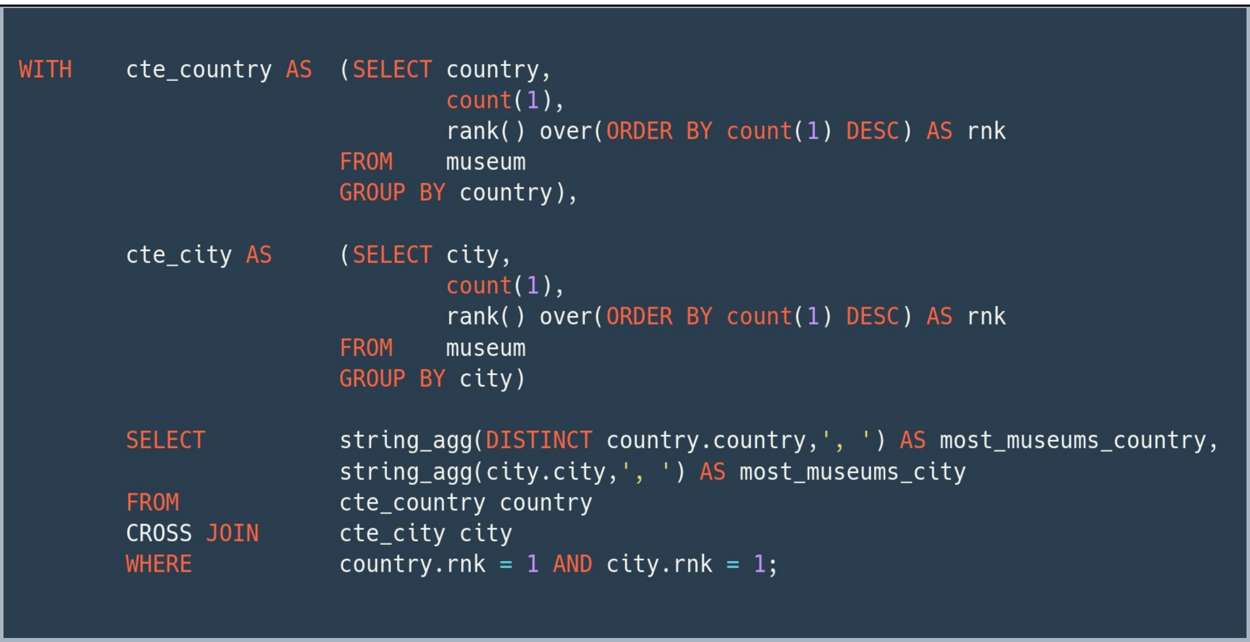
**The 3 most popular painting styles are:
Impressionism, Post-Impressionism and no specific style (null).**

**The 3 least popular painting styles are:
Japanese Art, Art-Nouveau, and Avant-Garde.**

Task #12 Display the country and the city with the most museums.

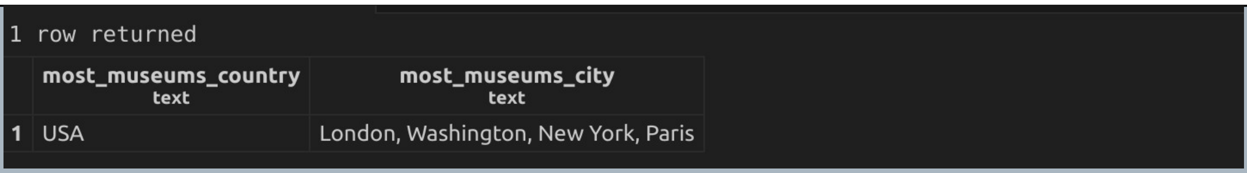
Output 2 separate columns for city and country;
separate multiple values with commas.

Input



```
WITH    cte_country AS  (SELECT country,
                                count(1),
                                rank() over(ORDER BY count(1) DESC) AS rnk
                        FROM    museum
                        GROUP BY country),

        cte_city AS     (SELECT city,
                                count(1),
                                rank() over(ORDER BY count(1) DESC) AS rnk
                        FROM    museum
                        GROUP BY city)

        SELECT          string_agg(DISTINCT country.country,', ') AS most_museums_country,
                        string_agg(city.city,', ') AS most_museums_city
        FROM            cte_country country
        CROSS JOIN      cte_city city
        WHERE           country.rnk = 1 AND city.rnk = 1;
```

Output



1 row returned

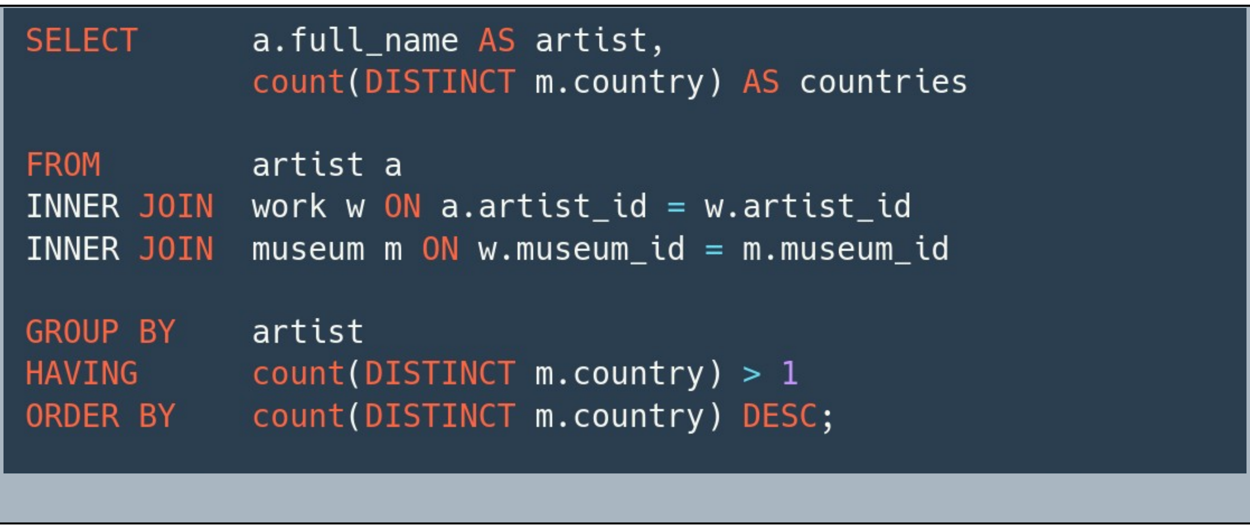
	most_museums_country text	most_museums_city text
1	USA	London, Washington, New York, Paris

Country with the most museums:
USA

Cities with the most museums:
London, Washington, New York, Paris

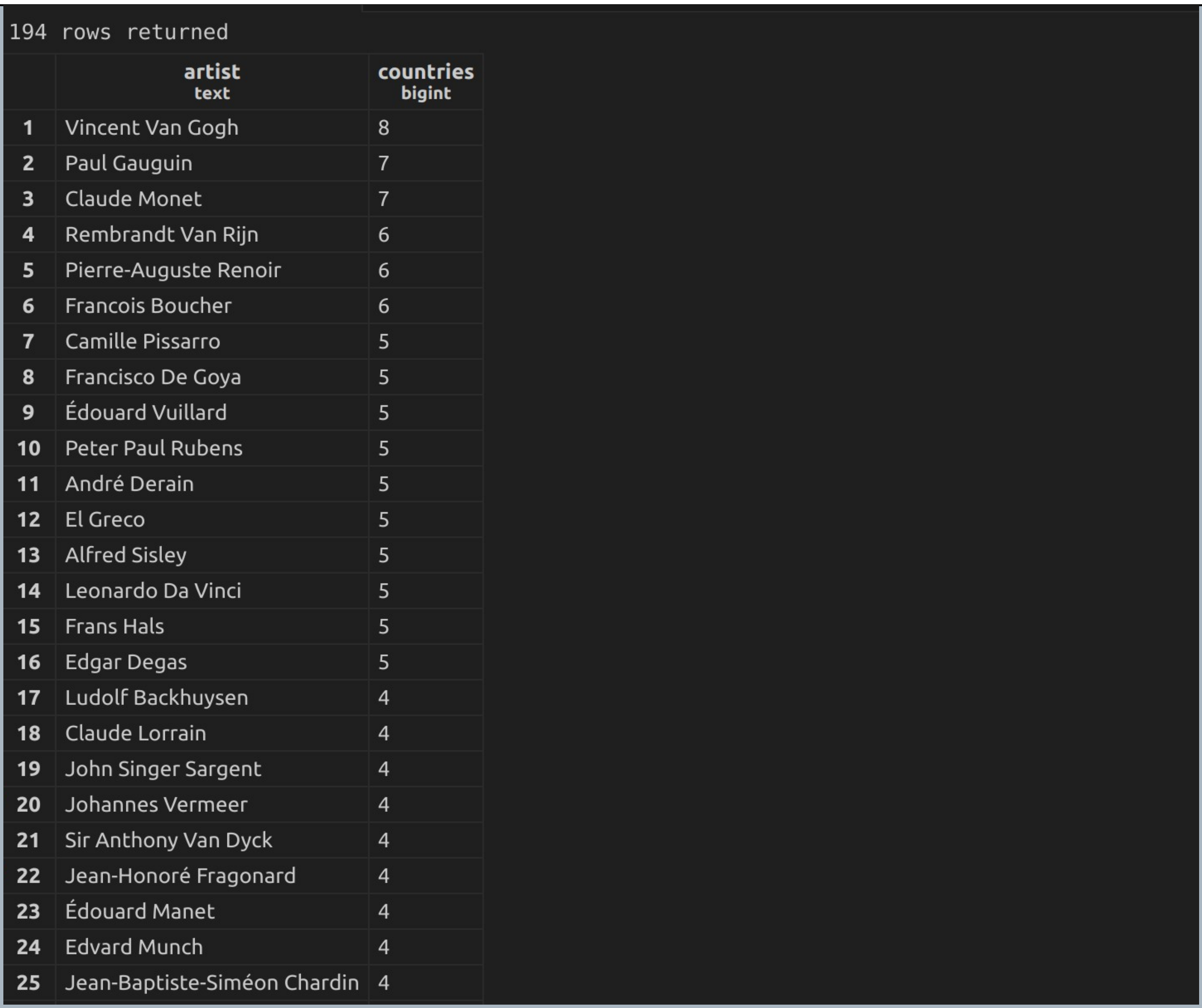
Task #13 Identify artists whose paintings are displayed in multiple countries.

Input



```
SELECT      a.full_name AS artist,
            count(DISTINCT m.country) AS countries

FROM        artist a
INNER JOIN  work w ON a.artist_id = w.artist_id
INNER JOIN  museum m ON w.museum_id = m.museum_id

GROUP BY    artist
HAVING      count(DISTINCT m.country) > 1
ORDER BY    count(DISTINCT m.country) DESC;
```

Output



194 rows returned

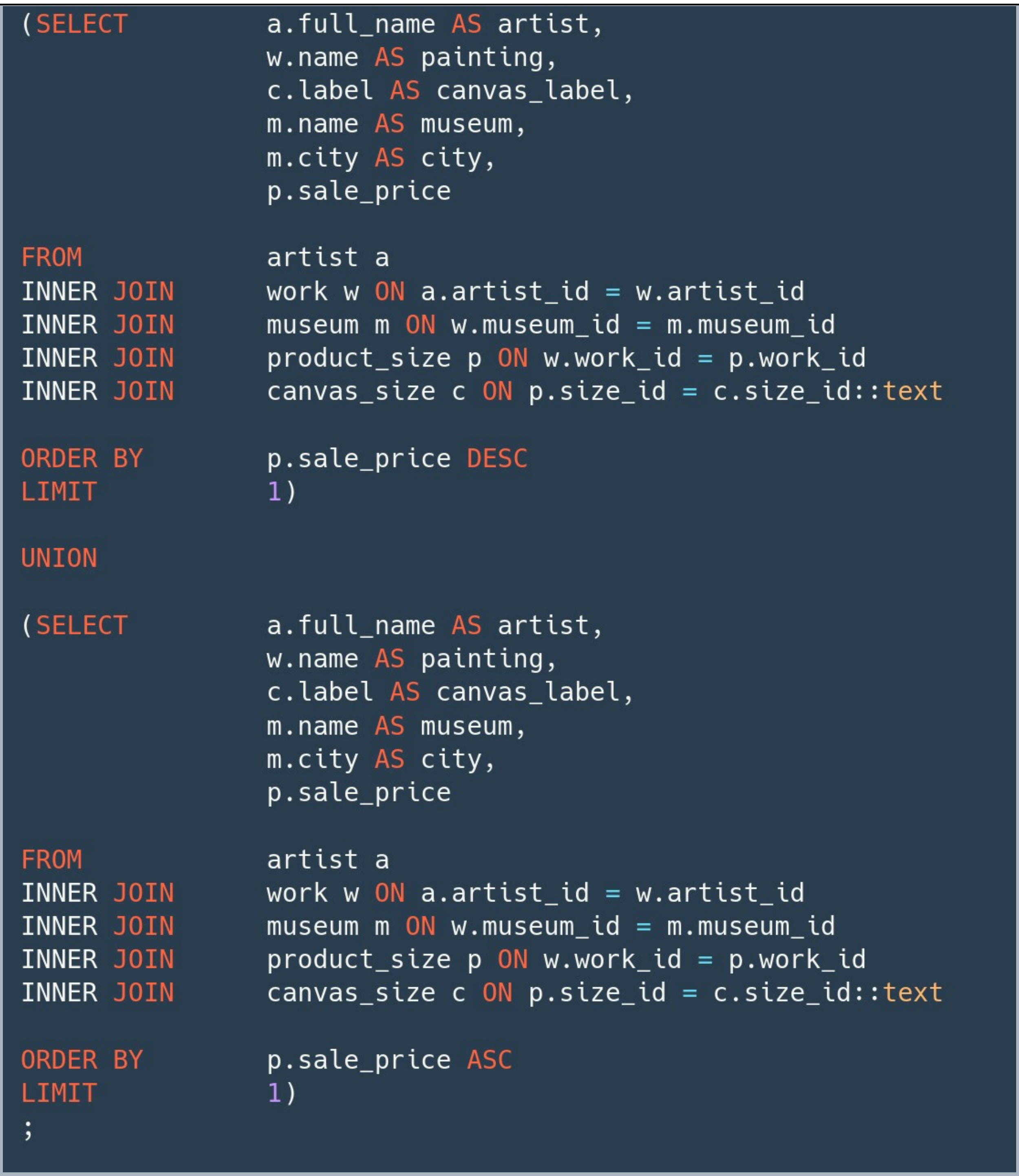
	artist text	countries bigint
1	Vincent Van Gogh	8
2	Paul Gauguin	7
3	Claude Monet	7
4	Rembrandt Van Rijn	6
5	Pierre-Auguste Renoir	6
6	Francois Boucher	6
7	Camille Pissarro	5
8	Francisco De Goya	5
9	Édouard Vuillard	5
10	Peter Paul Rubens	5
11	André Derain	5
12	El Greco	5
13	Alfred Sisley	5
14	Leonardo Da Vinci	5
15	Frans Hals	5
16	Edgar Degas	5
17	Ludolf Backhuysen	4
18	Claude Lorrain	4
19	John Singer Sargent	4
20	Johannes Vermeer	4
21	Sir Anthony Van Dyck	4
22	Jean-Honoré Fragonard	4
23	Édouard Manet	4
24	Edvard Munch	4
25	Jean-Baptiste-Siméon Chardin	4

194 artists have their paintings displayed in museums in more than one country. *[screenshot limited to 25 results]*

Task #14 Identify the museums with the most and least expensive paintings as well as their respective artists.

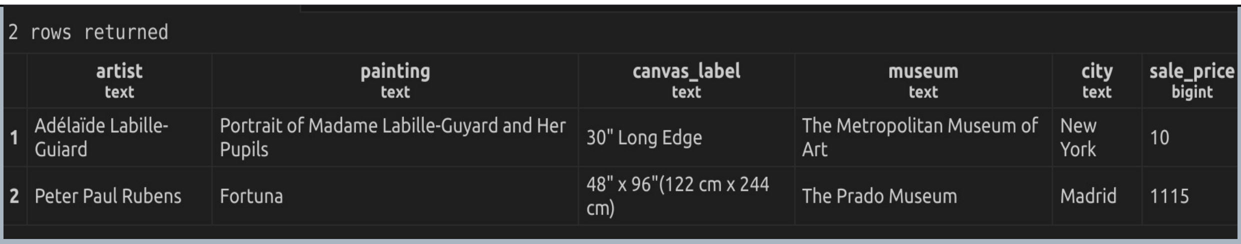
Display artist name, painting name, canvas label
museum name, museum city and sale price.

Input



```
(SELECT         a.full_name AS artist,
                w.name AS painting,
                c.label AS canvas_label,
                m.name AS museum,
                m.city AS city,
                p.sale_price

FROM            artist a
INNER JOIN      work w ON a.artist_id = w.artist_id
INNER JOIN      museum m ON w.museum_id = m.museum_id
INNER JOIN      product_size p ON w.work_id = p.work_id
INNER JOIN      canvas_size c ON p.size_id = c.size_id::text

ORDER BY        p.sale_price DESC
LIMIT           1)

UNION

(SELECT         a.full_name AS artist,
                w.name AS painting,
                c.label AS canvas_label,
                m.name AS museum,
                m.city AS city,
                p.sale_price

FROM            artist a
INNER JOIN      work w ON a.artist_id = w.artist_id
INNER JOIN      museum m ON w.museum_id = m.museum_id
INNER JOIN      product_size p ON w.work_id = p.work_id
INNER JOIN      canvas_size c ON p.size_id = c.size_id::text

ORDER BY        p.sale_price ASC
LIMIT           1)
;
```

Output



2 rows returned

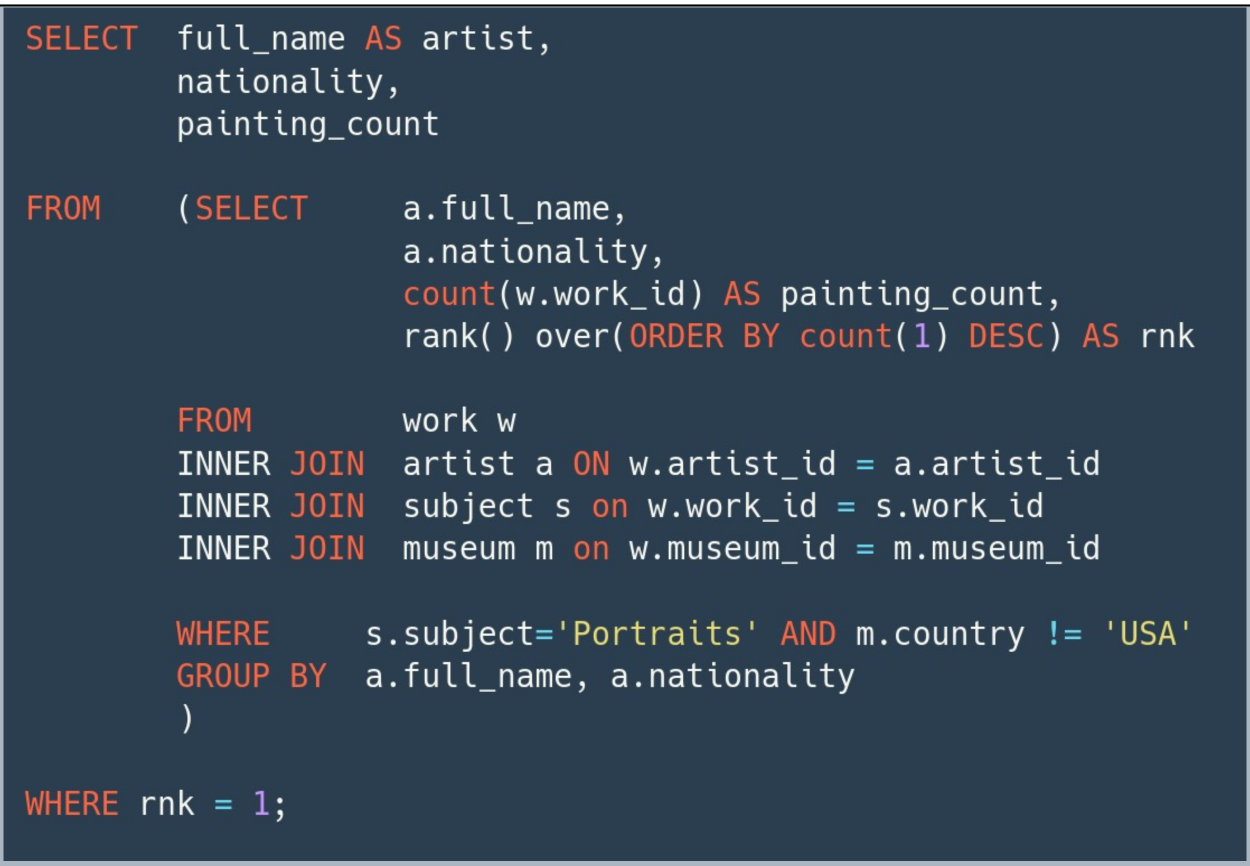
	artist text	painting text	canvas_label text	museum text	city text	sale_price bigint
1	Adélaïde Labille-Guiard	Portrait of Madame Labille-Guyard and Her Pupils	30" Long Edge	The Metropolitan Museum of Art	New York	10
2	Peter Paul Rubens	Fortuna	48" x 96"(122 cm x 244 cm)	The Prado Museum	Madrid	1115

The museums and artists with the most and least expensive paintings are as outputted above.

Task #15 Which artist has the most number of Portrait paintings outside the US?

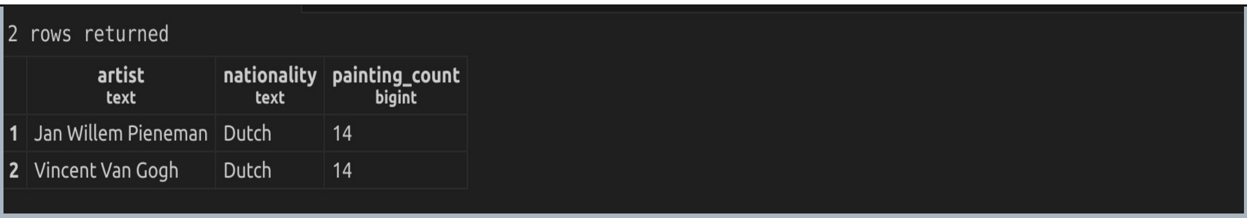
Display artist name, nationality and painting number.

Input



```
SELECT  full_name AS artist,
        nationality,
        painting_count

FROM    (SELECT     a.full_name,
                    a.nationality,
                    count(w.work_id) AS painting_count,
                    rank() over(ORDER BY count(1) DESC) AS rnk

        FROM        work w
        INNER JOIN  artist a ON w.artist_id = a.artist_id
        INNER JOIN  subject s on w.work_id = s.work_id
        INNER JOIN  museum m on w.museum_id = m.museum_id

        WHERE     s.subject='Portraits' AND m.country != 'USA'
        GROUP BY  a.full_name, a.nationality
        )

WHERE rnk = 1;
```

Output



2 rows returned

	artist text	nationality text	painting_count bigint
1	Jan Willem Pieneman	Dutch	14
2	Vincent Van Gogh	Dutch	14

The Dutch artists Jan Willem Pieneman and Vincent Van Gogh have the most Portrait paintings on display in museums that are not located in the US (14 paintings each).