Worldwide Masterpieces :

Data Analytics of Museum Paintings

Analyzed by Sankalp Raj

This report offers a comprehensive analysis of global art collections using data analytics. Leveraging various Data Analytic tools for understanding sophisticated data, this project presents critical insights into the distribution, value, and characteristics of paintings in museums around the world. The goal is to provide a data-driven perspective on art collections, showcasing the power of data analytics in understanding and appreciating the rich tapestry of artistic heritage globally.

Overview

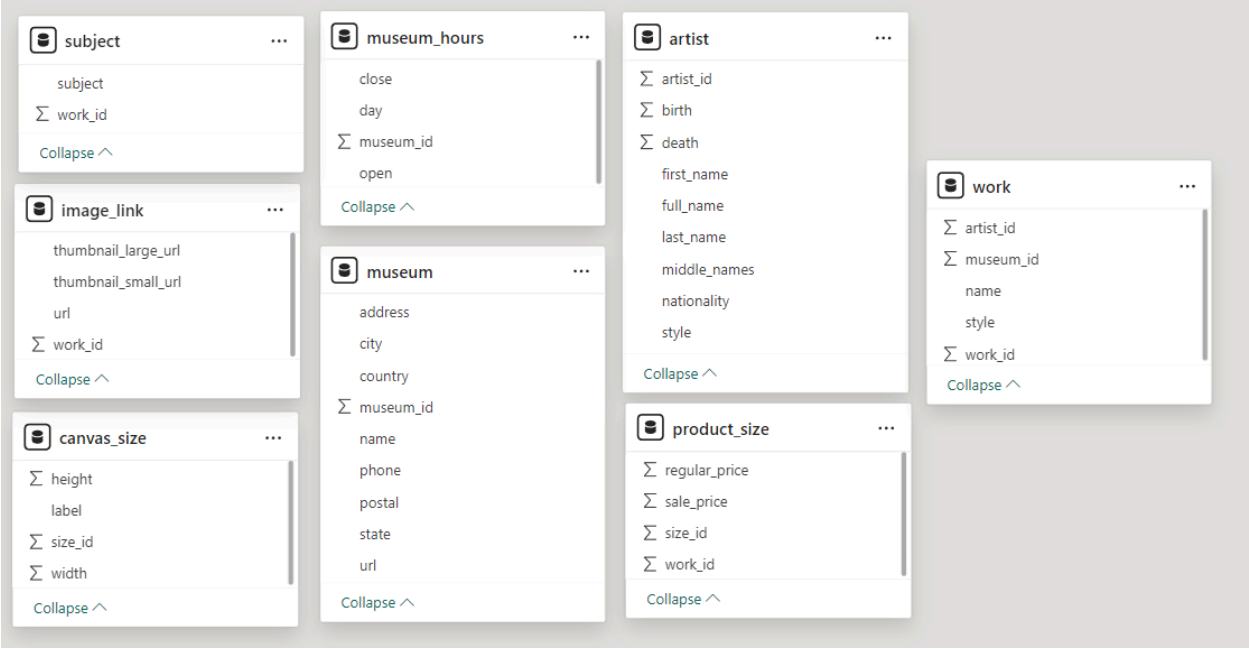
This project, "Worldwide Masterpieces: Data Analytics of Museum Paintings," utilizes **Python for data cleaning**, **SQL for querying**, and **Power BI for visualization**. We start by preprocessing the dataset using Pandas to ensure accuracy. Next, we import the data into a SQL database to answer key questions such as identifying the most diverse museums, total paintings by artists, and top-selling artworks. Finally, we present our findings in an interactive Power BI dashboard, offering insights into the distribution of paintings, popular styles, and artist sales.

This integrated approach provides a comprehensive exploration of the dataset, highlighting significant trends and patterns in the art world.

Content

Understanding the Dataset

The dataset utilized for this project is sourced from [Kaggle], titled "**Famous Paintings**." This dataset encompasses detailed information about various paintings housed in museums worldwide. The tables within the dataset and their corresponding columns are as follows:



1. **Artist** - Provides detailed biographical information about each artist.
2. **Canvas_size** - Describes the dimensions and labels of canvas sizes used for paintings.
3. **Image_link** - Provides URLs for the images and thumbnails of the paintings.
4. **Museum** - Contains detailed information about each museum, including location and contact details.
5. **Museum_hours** - Lists the opening and closing hours of the museums.
6. **Product_size** - Details the sale and regular prices of the paintings along with their canvas sizes.
7. **Subject** - Describes the subject matter of each painting.
8. **Work** - Provides information about the paintings, including their names, styles, and associated artists and museums.

Data Cleaning using Python

To ensure accuracy and consistency in our analysis, we utilized **Pandas** for data cleaning and preprocessing. This process involved handling missing values and ensuring consistency across all tables. Initial visualizations using **Matplotlib** were performed to gain preliminary insights, laying a solid foundation for subsequent analysis and visualizations.

1)Analyzing Artist Table

The dataset includes columns such as artist_id, full_name, nationality, style, birth, and death.

```
[22] import pandas as pd
     artists = pd.read_csv('/content/artist.csv')
     print(artists.columns)
```

```
Index(['artist_id', 'full_name', 'first_name', 'middle_names', 'last_name',
       'nationality', 'style', 'birth', 'death'],
      dtype='object')
```

Counting Null Values in each column:

```
[8] null_counts = artists.isnull().sum()
    print(null_counts)
```

```
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
```

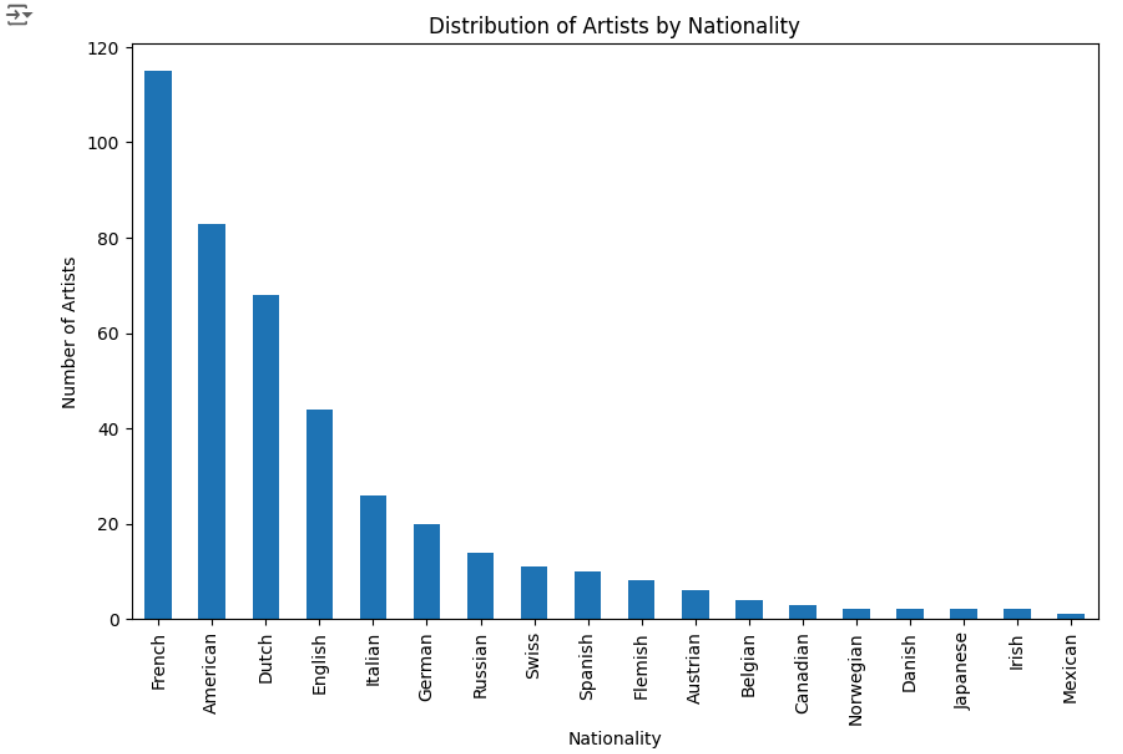
After initial inspection, columns like first_name, middle_names, and last_name were deemed redundant and dropped to simplify the dataset.

```
artists = artists.drop(['first_name', 'middle_names','last_name'], axis=1)
```

Now that the artist table has been cleaned and is ready for analysis, let's dive into some insightful questions:

1. Which nationality has the most and least number of artists?

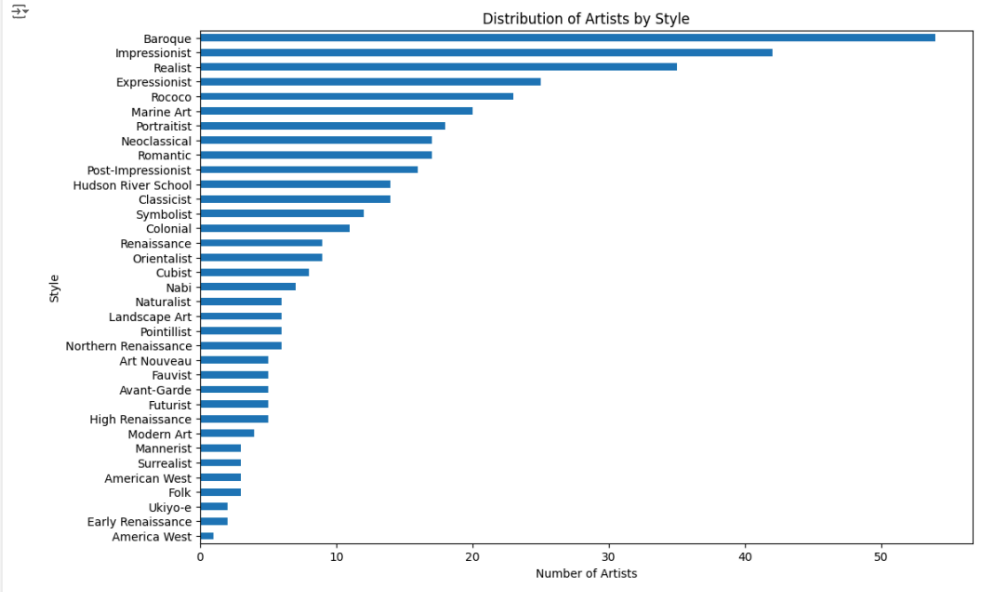
```
nationality_counts = artists['nationality'].value_counts()
plt.figure(figsize=(10, 6))
nationality_counts.plot(kind='bar')
plt.title('Distribution of Artists by Nationality')
plt.xlabel('Nationality')
plt.ylabel('Number of Artists')
plt.show()
```



The chart reveals that **French** artists are the most numerous, while **Mexican** artists are the least represented in the dataset.

2. What are the most preferred styles among artists?

```
plt.figure(figsize=(12, 8))
style_counts.plot(kind='barh')
plt.title('Distribution of Artists by Style')
plt.xlabel('Number of Artists')
plt.ylabel('Style')
plt.gca().invert_yaxis()
plt.show()
```



The analysis shows that the most preferred art style among artists is **Baroque**. This indicates a significant concentration of artists favoring this style, suggesting its prominence and popularity among artists.

2)Analyzing Work Table

```
[49] work = pd.read_csv('/content/work.csv')
     print(work.columns)
```

```
Index(['work_id', 'name', 'artist_id', 'style', 'museum_id'], dtype='object')
```

The work table includes columns such as work_id, name, artist_id, style, and museum_id.

The table seems to contain null values at styles columns

```
work['style'].fillna('Unknown', inplace=True)
```

The work table has been cleaned and is now ready for further analysis.

3)Analyzing Subject Table

```
subject = pd.read_csv('/content/subject.csv')
subject.head(200)
```

	work_id	subject
0	160228	Still-Life
1	160236	Still-Life
2	160244	Still-Life
3	160252	Still-Life
4	160260	Still-Life
...	...	...
195	185411	Landscape Art
196	185659	Landscape Art
197	185419	Landscape Art
198	185427	Landscape Art
199	185435	Landscape Art

200 rows × 2 columns

```
subject.isnull().sum()
```

	0
work_id	0
subject	0

dtype: int64

No null values were detected in any column, indicating complete data.

```
subject['subject'].unique()
```

```
array(['Still-Life', 'Horses', 'Marine Art/Maritime', 'Portraits',
       'Musics', 'Dancers', 'U.S. Presidents', 'Dogs', 'Rivers/Lakes',
       'Flowers', 'Bridges', 'Spring', 'Landscape Art', 'Autumn/Fall',
       'Deers', 'Winter', 'Jesus Christ', 'Christianity', 'Nude',
       'Gardens', 'Seascapes', 'Churches/Temples/Mosques', 'Cafes/Bars',
       'Summer', 'Architectures', 'Lovers', 'Abstract/Modern Art',
       'Water Lilies', 'Tigers'], dtype=object)
```

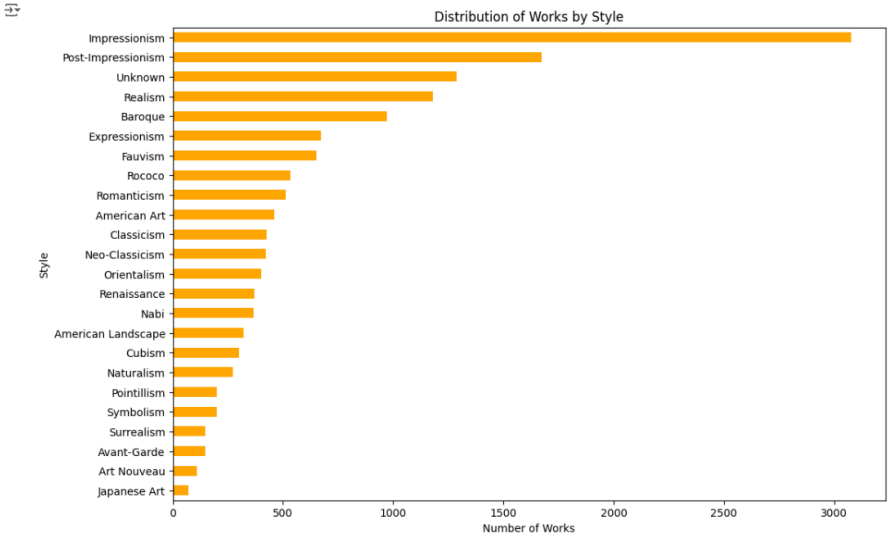
The subject column contains only valid categorical data with no numeric entries or anomalies.

This ensures that the subject data is reliable and can be confidently used for further analysis

Let's try to answer this question :

1. **Which subject is the most popular ?**

```
style_counts = work['style'].value_counts()
plt.figure(figsize=(12, 8))
style_counts.plot(kind='barh', color='orange')
plt.title('Distribution of Works by Style')
plt.xlabel('Number of Works')
plt.ylabel('Style')
plt.gca().invert_yaxis()
plt.show()
```



So it's clear from the plot that **Impressionism** is the most popular subject among artists

4)Analyzing Product Size Table

```
price = pd.read_csv('/content/product_size.csv')
price.head( 200)
```

	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
...	...	...	...	...
195	135925	4860	675	1285
196	135925	5468	805	1545
197	135936	2430	305	545
198	135936	2632	325	585
199	135936	2936	375	685

200 rows × 4 columns

The table contains columns such as work_id, size_id, sale_price, and regular_price.
The product_size table is **renamed to price** for clarity and consistency.

```
[66] price = price.drop(['size_id','regular_price'],axis=1)
```

Columns regular_price and size_id are dropped as they are deemed irrelevant for the analysis.

```
price.isnull().sum()
```

	0
work_id	0
sale_price	0

dtype: int64

No null values were found, indicating that the dataset is complete.

```
grouped_works = price.groupby('work_id').size()
grouped_works
```

work_id	0
178	9
179	9
180	7
382	8
383	7
384	9

However each work has multiple selling_price which is not possible.So, we will **replace the selling_price for each work with the average of all selling prices**.

```
price = price.groupby('work_id')['sale_price'].mean().reset_index()
```

```
price.groupby('work_id').size()
```

work_id	0
178	1
179	1
180	1
382	1
383	1
384	1

Now, the data is clean and can be used for analysis.

5)Analyzing Museum Table

```
[70] museum = pd.read_csv('/content/museum.csv')
     museum.columns
```

```
Index(['museum_id', 'name', 'address', 'city', 'state', 'postal', 'country',
       'phone', 'url'],
      dtype='object')
```

```
museum = museum.drop(['address','city','state','postal','phone', 'url'],axis=1)
museum
```

	museum_id	name	country
0	30	The Museum of Modern Art	USA
1	31	Pushkin State Museum of Fine Arts	Russia
2	32	National Gallery of Victoria	Australia
3	33	São Paulo Museum of Art	Brazil
4	34	The State Hermitage Museum	Russia
5	35	The Metropolitan Museum of Art	USA
6	36	Museum Folkwang	Germany
7	37	Museum of Grenoble	France
8	38	Musée des Beaux-Arts de Quimper	France
9	39	Nelson-Atkins Museum of Art	USA

Columns address, city, state, postal, phone, and url are dropped as they are not needed for the current analysis.

```
[72] museum.isnull().sum()
```

	0
museum_id	0
name	0
country	0

dtype: int64

This confirms that there are no null values in the museum table.

```
museum['country'].unique()
museum['name'].unique()
```

```
array(['The Museum of Modern Art', 'Pushkin State Museum of Fine Arts',
       'National Gallery of Victoria', 'São Paulo Museum of Art',
       'The State Hermitage Museum', 'The Metropolitan Museum of Art',
       'Museum Folkwang', 'Museum of Grenoble',
       'Musée des Beaux-Arts de Quimper', 'Nelson-Atkins Museum of Art',
       'Musée du Louvre', 'National Maritime Museum',
       'Museum of Fine Arts Boston', 'Rijksmuseum', 'Israel Museum',
       'Kunsthaus Zürich', 'National Gallery of Art', 'National Gallery',
       'Mauritshuis Museum', "Musée d'Orsay", 'The Prado Museum',
       'The Barnes Foundation', 'Hungarian National Gallery',
       'Cleveland Museum Of Art', 'Museum of Fine Arts, Houston',
       'The J. Paul Getty Museum', 'Thussen-Bornemisza Museum',
```

Checked for inconsistent values. The values are found to be consistent

Now that the artist table has been cleaned :

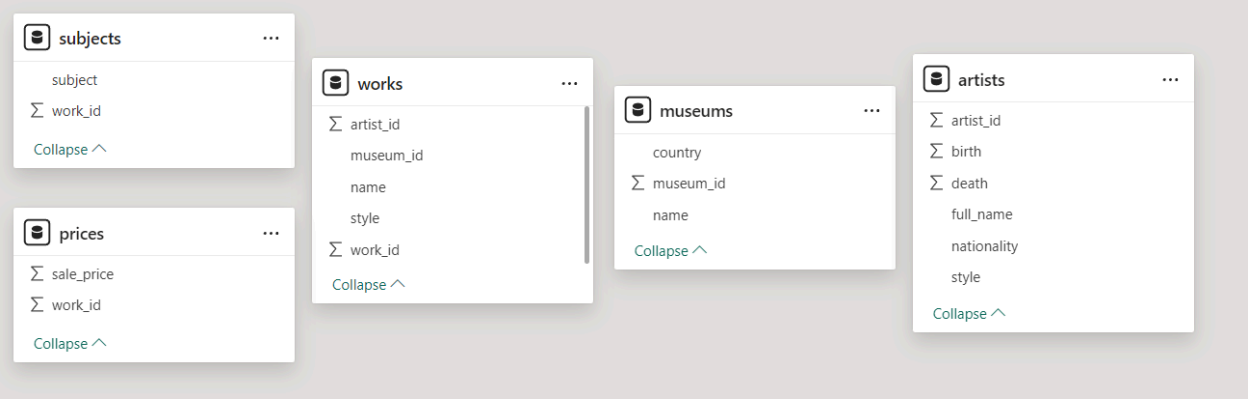
1. Analyze Distribution of Museums Among Countries

```
country_counts = museum['country'].value_counts()
country_counts
```

country	count
USA	25
France	7
UK	5
Netherlands	4
Spain	2
Russia	2
Switzerland	2
Norway	1
Japan	1
Italy	1
Hungary	1
Israel	1
United Kingdom	1
Germany	1
Brazil	1
Australia	1
Czechia	1

So it confirms that **USA has the most museums(25)** among other countries

Dataset after Cleaning



After the data cleaning process, we consolidated our dataset into five major tables: artists, works, museums, subjects, and prices. We dropped unnecessary columns and irrelevant tables to streamline our analysis. With the data now consistent and well-structured, we proceeded to import these cleaned tables into a MySQL database to perform complex queries and derive meaningful insights.

Data Analysis using SQL

SQL queries are a cornerstone of data analysis, enabling us to extract, manipulate, and analyze data efficiently. In this section, we will showcase a range of SQL queries, progressing from beginner to advanced levels, to uncover insights from our cleaned dataset. These queries will help us answer critical questions, identify trends demonstrating the power and versatility of SQL in data analytics.

1.Museums with the Most Diverse Collection of Styles

This query identifies museums with the broadest range of artistic styles.This diversity can attract a varied audience and enhance the museum's reputation.

Query:

```
-- Museums with the Most Diverse Collection of Styles
SELECT m.name AS museum_name, COUNT(DISTINCT w.style) AS style_count
FROM museums m
JOIN works w ON m.museum_id = w.museum_id
GROUP BY m.name
ORDER BY style_count DESC;
```

Output:

museum_name	style_count
The Metropolitan Museum of Art	22
National Gallery of Art	20
Cleveland Museum Of Art	17
Philadelphia Museum of Art	16
Los Angeles County Museum of Art	14
National Gallery	14
The State Hermitage Museum	14
The Art Institute of Chicago	13
Thussen-Bornemisza Museum	13
MusÃ©e d'Orsay	12

Conclusion:

The **Metropolitan Museum of Art** is recognized for its extensive collection of artistic styles, featuring 22 distinct styles, making it a leading cultural institution in terms of diversity.

2. Show number of Paintings over Decades

This query provides insight into the production of paintings across different decades. Understanding how the number of paintings varies over time can reveal trends in artistic output and historical influences on art production.

Query:

```
-- Number of paintings over decades
SELECT FLOOR((birth + death) / 2 / 10) * 10 AS decade, COUNT(*) AS painting_count
FROM works w
JOIN artists a ON w.artist_id = a.artist_id
GROUP BY decade
ORDER BY decade;
```

Output:

decade	painting_count
1410	7
1430	5
1460	13
1470	7
1480	21
1490	19
1500	13
1510	15
1520	9
1530	13
1550	8
1560	3
1570	15

Conclusion:

We can find the number of paintings made in each decades

3. Top 5 Museums with the Oldest Collections

This query identifies museums housing the oldest collections, based on the average age of the artists' birthdays. Museums with older collections may offer a richer historical perspective and attract visitors interested in historical art.

Query:

```
-- Top 5 Museums with the Oldest Collections
SELECT m.name AS museum_name, AVG(YEAR(CURDATE()) - a.birth) AS average_age
FROM works w
JOIN artists a ON w.artist_id = a.artist_id
JOIN museums m ON w.museum_id = m.museum_id
GROUP BY m.name
ORDER BY average_age DESC
LIMIT 5;
```

Output:

museum_name	average_age
Uffizi Gallery Italy	579.0000
Mauritshuis Museum	402.0000
National Gallery	374.3503
The Prado Museum	373.1765
Rijksmuseum	368.8580

Conclusion:

The **Uffizi Gallery Italy** boasts the oldest average collection age at 579 years, signifying its extensive historical art holdings.

4.Total Number of Paintings by Each Artist

This query calculates the total number of paintings created by each artist, highlighting prolific artists and their contributions to the art world. Understanding an artist's output can provide context for their influence and significance.

Query:

```
-- Total Number of Paintings by Each Artist
SELECT a.full_name, COUNT(w.work_id) AS total_paintings
FROM artists a
JOIN works w ON a.artist_id = w.artist_id
GROUP BY a.full_name
ORDER BY total_paintings DESC;
```

Output:

full_name	total_paintings
Peter Paul Rubens	56
Claude Monet	45
Jean Baptiste Vanmour	30
Francesco Guardi	28
Jan Steen	28
Anton Raphael Mengs	27
Francisco De Goya	27
John Singleton Copley	27
Ivan Aivazovskiy	25
Paul CÃ©zanne	25
Rembrandt Van Rijn	24
Gerard Ter Borch	23
Nicolas Poussin	23
Gerard Van Honthorst	22

Conclusion:

Peter Paul Rubens is the most prolific artist in the dataset, with 56 paintings, demonstrating his extensive contribution to art.

5.Top 10 Most Expensive Paintings (Based on Sale Price)

This query highlights the most valuable paintings in terms of sale price. Identifying these top paintings can provide insights into high-value art, market trends, and significant works that might attract collectors and researchers.

Query:

```
-- Top 10 Most Expensive Paintings (Based on Sale Price)
SELECT w.name AS painting_name, a.full_name AS artist_name, p.sale_price
FROM works w
JOIN artists a ON w.artist_id = a.artist_id
JOIN prices p ON w.work_id = p.work_id
ORDER BY p.sale_price DESC
LIMIT 10;
```

Output:

painting_name	artist_name	sale_price
Birth of Venus (Naissance de Venus)	William Adolphe Bouguereau	762.5
Fortuna	Peter Paul Rubens	747
Diana Cazadora	Peter Paul Rubens	697.5
The Coronation of the Virgin	Guido Reni	692.5
The Adoration of the Shepherds	Guido Reni	682.5
The Crowning of Roxana	Peter Paul Rubens	681
The Holy Family with Saints Francis and Anne a...	Peter Paul Rubens	681
Modello for the Israelites Gathering Manna in th...	Peter Paul Rubens	681
Odalisque	Francois Boucher	671

Conclusion:

"**Birth of Venus(Naissance de Venus)**" by **William Adolph Bouguereau** is the most expensive painting, valued at $762.5 million, reflecting its unparalleled market significance.

6. Top 5 Most Popular Museums

This query identifies the most popular museums based on the number of paintings they house. Popular museums often attract more visitors and are significant in preserving and displaying art collections.

Query:

```
-- Top 5 Most Popular Museums
WITH MuseumPaintings AS (
    SELECT museum_id, COUNT(work_id) AS no_of_paintings_in_museum
    FROM works
    GROUP BY museum_id
)
SELECT m.name AS museum_name, mp.no_of_paintings_in_museum
FROM museums m
JOIN MuseumPaintings mp ON m.museum_id = mp.museum_id
ORDER BY mp.no_of_paintings_in_museum DESC
LIMIT 5;
```

Output:

museum_name	no_of_paintings_in_museum
The Metropolitan Museum of Art	536
Rijksmuseum	352
National Gallery	314
National Gallery of Art	175
National Maritime Museum	131

Conclusion:

The **Metropolitan Museum of Art** is the top museum with **536 paintings**, showcasing its extensive art collection.

7.Identify the Artist and the Museum Where the Most Expensive and Least Expensive Painting is Placed

This query finds both the most and least expensive paintings and their respective artist and museum. This information can highlight significant contrasts in artwork value and the prestige of various institutions.

Query:

```
-- Identify the Artist and the Museum Where the Most Expensive and Least Expensive Painting is Placed
WITH PriceRank AS (
    SELECT w.artist_id, w.museum_id, p.sale_price,
           RANK() OVER (ORDER BY p.sale_price DESC) AS rnk
    FROM works w
    JOIN prices p ON w.work_id = p.work_id
)
SELECT DISTINCT a.full_name AS artist_name, m.name AS museum_name, pr.sale_price
FROM PriceRank pr
JOIN artists a ON pr.artist_id = a.artist_id
JOIN museums m ON pr.museum_id = m.museum_id
WHERE pr.rnk = 1 OR pr.rnk = (SELECT MAX(rnk) FROM PriceRank);
```

Output:

artist_name	museum_name	sale_price
Abraham Mignon	Rijksmuseum	85
Jacob Van Ruisdael	Rijksmuseum	85
William Adolphe Bouguereau	MusÃ©e d'Orsay	762.5

Conclusion:

The **most expensive painting is by William Adolph Bouguereau at MusÃ©e d'Orsay,** while the **least expensive top painting is by Abraham Mignon at The Rijksmuseum.**

8. Which Museum Has the Most Number of the Most Popular Painting Style?

This query identifies which museum holds the most paintings of the most popular style. This can help understand which institutions focus on trending art styles and attract audiences interested in specific genres.

Query:

```
-- Which Museum Has the Most Number of the Most Popular Painting Style?
WITH query1 AS(SELECT style,
COUNT(work_id) OVER(PARTITION BY style) as no_of_work
FROM works
ORDER BY no_of_work desc
LIMIT 1),

query2 AS (SELECT work_id,museums.museum_id,museums.name
FROM works JOIN museums ON works.museum_id = museums.museum_id
INNER JOIN query1 ON query1.style = works.style)

SELECT name , COUNT(work_id) as no_of_painting
FROM query2
GROUP BY name
ORDER BY no_of_painting desc
LIMIT 1;
```

Output:

	name	no_of_painting
▶	Rijksmuseum	153

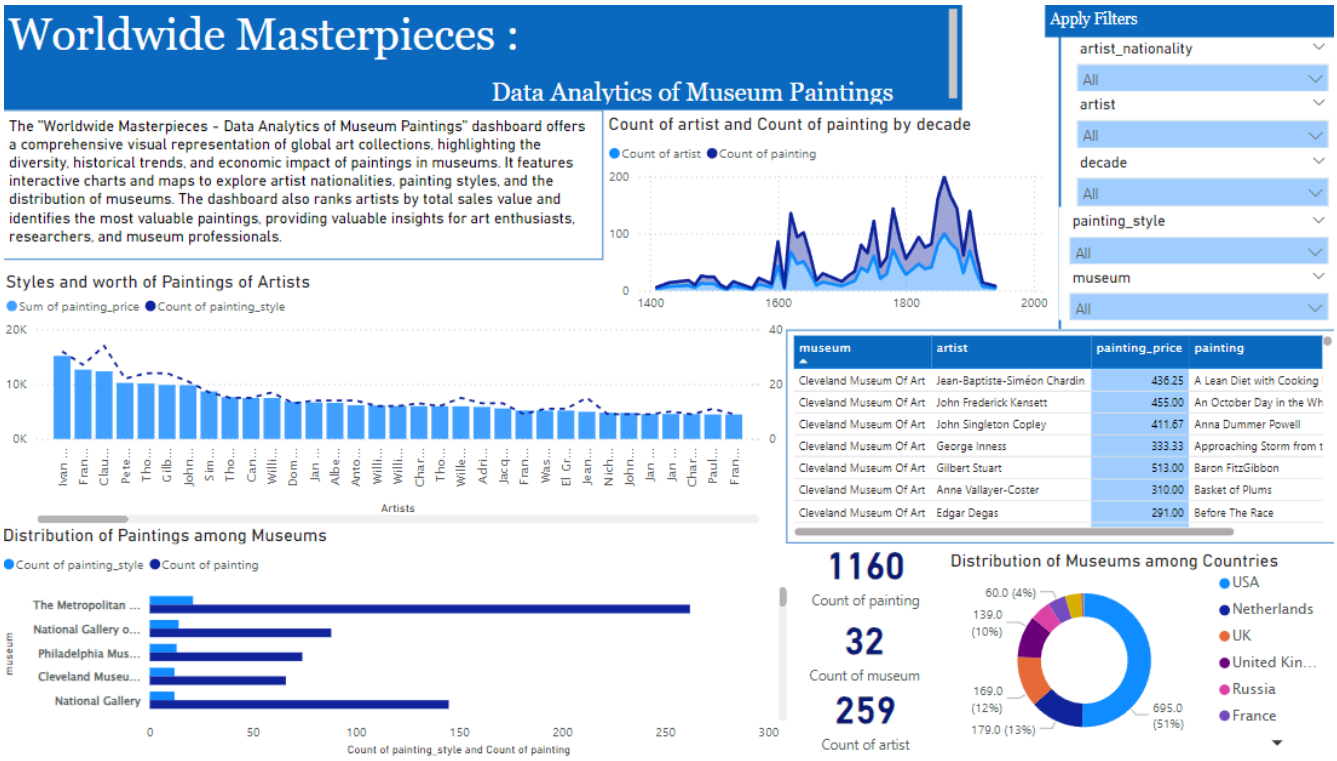
Conclusion:

Rijksmuseum houses the **largest number of paintings from the most popular style**, indicating its focus on contemporary and popular art trends.

Data Visualisation using Power BI

The Power BI dashboard is designed to provide a holistic view of the data, enabling users to explore and gain insights through various visualizations. It offers a user-friendly interface to interact with the data, apply filters, and drill down into specific details, making it a powerful tool for understanding the intricacies of the dataset and drawing meaningful conclusions.

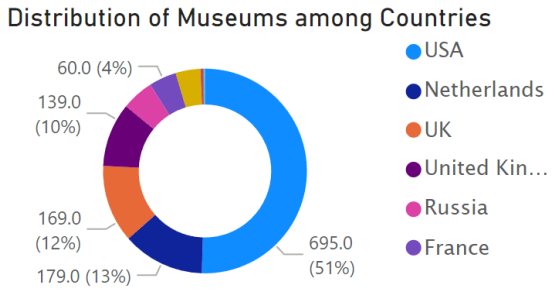
Understanding the Dashboard



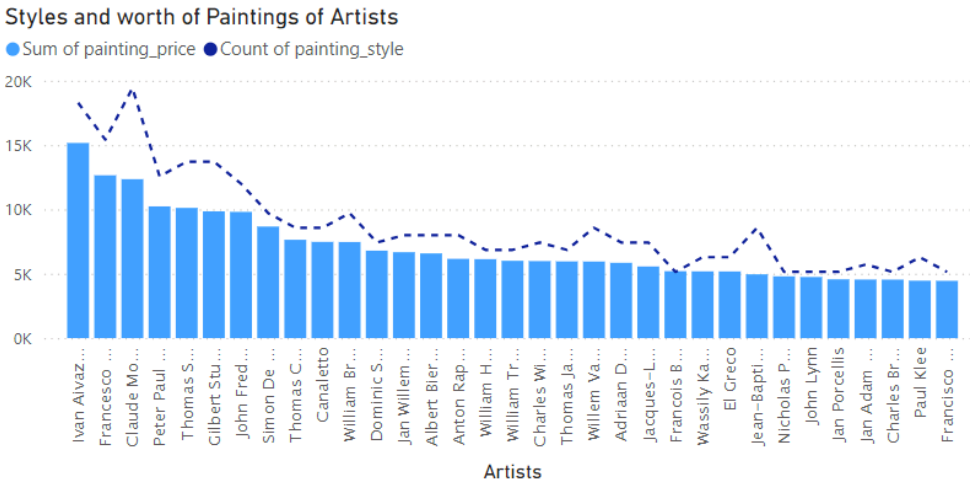
Features:

The dashboard features several interactive visualizations, each designed to provide deep insights into the dataset:

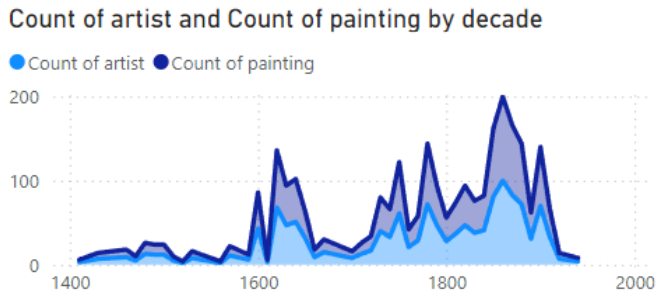
1. **Distribution of Paintings Among Museums**: This donut chart visualizes the distribution of museums across different countries. It highlights the global reach of the museum network and helps identify regions with a higher concentration of cultural institutions.



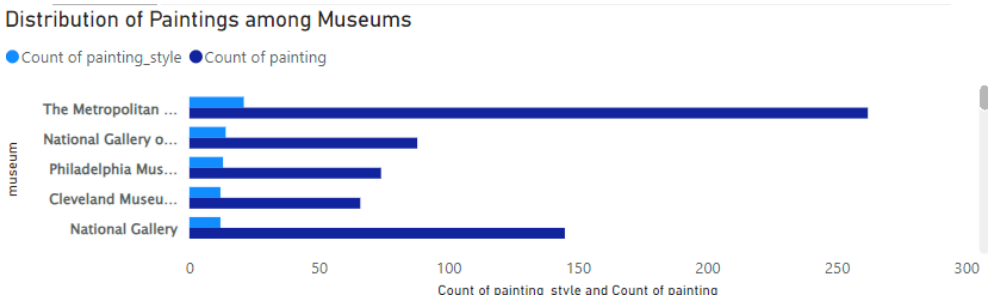
2. **Artist Performance Overview**: The line and clustered column chart displays the sum of painting prices by artist on the y-axis, with a line representing the count of painting styles. This combined view enables users to see both the financial impact and stylistic diversity of each artist's portfolio.



3. **Decade Analysis**: The stacked area chart shows the count of artists and paintings by decade. This visualization provides a temporal perspective on artistic output and trends, allowing users to track changes in the art scene over time.



4. **Museum Popularity by Painting Style**: The clustered bar chart ranks museums by the count of painting styles and paintings. This helps identify which museums have the most diverse collections and highlights their prominence in the art world.

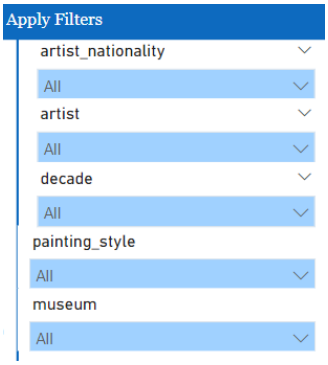


5. **Top Paintings Table**: This table lists artists, museums, paintings, and their prices in decreasing order. It provides a detailed view of high-value artworks and their associations, making it easy to identify top-performing artists and their most valuable pieces.

museum	artist	painting_price	painting
Cleveland Museum Of Art	Jean-Baptiste-Siméon Chardin	436.25	A Lean Diet with Cooking
Cleveland Museum Of Art	John Frederick Kensett	455.00	An October Day in the Wh
Cleveland Museum Of Art	John Singleton Copley	411.67	Anna Dummer Powell
Cleveland Museum Of Art	George Inness	333.33	Approaching Storm from t
Cleveland Museum Of Art	Gilbert Stuart	513.00	Baron FitzGibbon
Cleveland Museum Of Art	Anne Vallayer-Coster	310.00	Basket of Plums
Cleveland Museum Of Art	Edgar Degas	291.00	Before The Race

6. **Key Metrics Cards**: Three cards showcase key metrics including the total number of paintings, museums, and artists. These summary figures offer a quick overview of the dataset's scope and scale.

The dashboard's interactive slicers, allowing filters by nationality, artist, decade, painting style, and museum, further enhance its functionality, enabling users to tailor their analysis and explore specific aspects of the data in detail.



Conclusion

This project has successfully utilized data analytics to provide a comprehensive view of the art world. Through meticulous data cleaning and preprocessing using Python's Pandas, we ensured that the dataset was accurate and ready for analysis. Our SQL queries explored various facets of the data, offering insights into museum collections, artist performance, and painting trends.

The final dashboard in Power BI serves as a powerful tool for visualizing these insights. With features such as distribution charts, decade analyses, and detailed tables, it enables users to explore the data interactively. By providing a clear view of painting distribution, artist performance, and museum popularity, the dashboard helps in understanding the global art landscape and the intricate dynamics of museum collections.

Overall, this project underscores the relevance of combining data cleaning, sophisticated querying, and effective visualization to deliver comprehensive and insightful analyses. It showcases the power of these tools in transforming raw data into valuable knowledge, which is essential for professionals in the field of data analytics and visualization.

Thank You !