The Art of Data

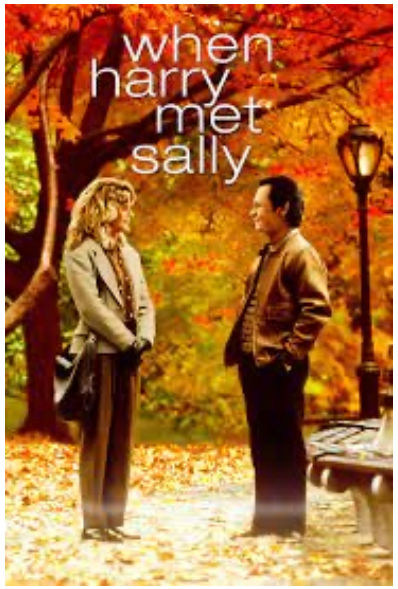
Chulwon Chae | Kiki Martin | Mitali Vipin Dighe | Parth Bansal

Agenda

- Background
- Business Problem
- Business Value and Use Cases
- Data Profile
- Methodology
- Data Modeling
- Data Cleaning
- Exploratory Data Analysis
- Recommendations
- Future Scope

The Metropolitan Museum of Art (The Met) in New York City is a treasure trove of art, history, and culture

- The Metropolitan Museum of Art is **one of the world's largest museums** devoted to modern and contemporary art.
- Its preeminent collection and distinguished scholarship make it one of the **most influential and important institutions** of the art world.
- With its extensive collection spanning **5,000 years of global art**, the MoMA provides a unique opportunity for the exploration and appreciation of diverse cultural treasures.

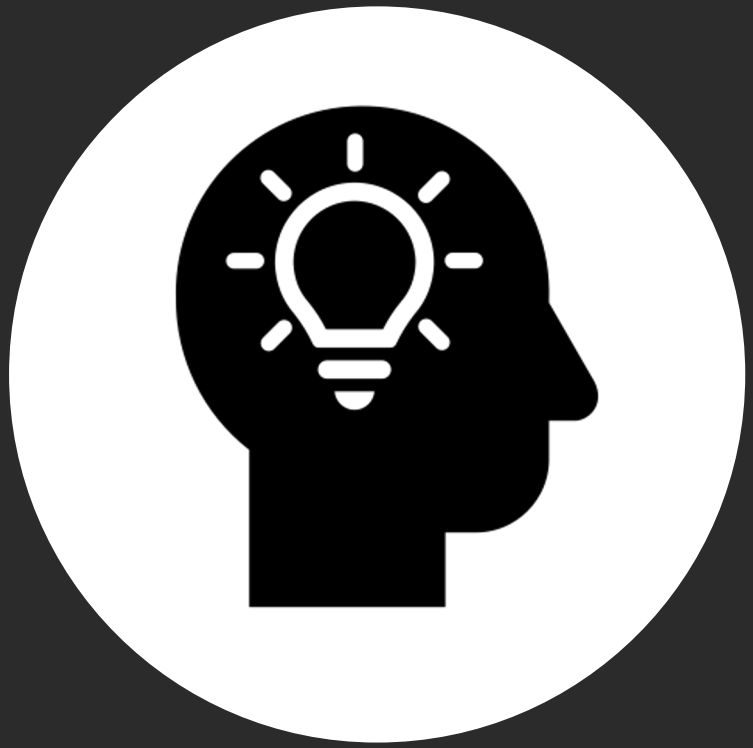

when
harry
met
sally


© Moviestore/REX/Shutterstock


SANDRA BULLOCK
CATE BLANCHETT
ANNE HATHAWAY
MINDY KALING
SARAH PAULSON
AWKWAFINA
RIHANNA
HELENA BONHAM CARTER
EVERY CON HAS ITS PROS
OCEAN'S 8
SUMMER 2018

Business Problem

The Met's artwork database, covering a vast range of art from different cultures and periods, suffers from disorganization and poor data categorization. This hampers its utility as a resource for art enthusiasts, researchers, and educators.



Proposed Solution

Leveraging Data Engineering tools harness the potential of the Met's artwork data to provide valuable insights, create engaging experiences and support various applications in the fields of art, culture, education and business.

Value and Impact

Addressing this problem could significantly enhance the value of The Met's collection as an academic and cultural resource. Improved database organization and accessibility would not only attract more art enthusiasts and researchers but also potentially open new avenues for collaboration, education, and digital innovation in the field of art and culture.

Business Cases: Data-Driven Strategies in Art Market Analysis, Content Creation, Merchandising, and Cultural Heritage Preservation

- **Art Market Analysis:** Galleries, art dealers and investors can analyze trends, artist contributions and art movements to make informed decisions about acquisitions of new art works.
- **Content Creation and Curation:** Multimedia production companies can use the data to source content and inspiration for art-related projects, exhibitions and documentaries from specific artists, departments or time periods.
- **Merchandising and Licensing:** Companies can license artwork images for various merchandise and marketing campaigns, supporting the creative and retail industries.
- **Cultural Heritage Preservation:** Organizations dedicated to cultural heritage preservation can document and research artworks with historical and cultural significance to support conservation efforts.

Data Profile

Data Source	The MET Database
Data Timeline	Over 5000 years
Total Artworks	480,000+
Features available	54
Tools used	Utilizing MySQL for ETL Python for Pre-processing and Analysis Tableau for Interactive Dashboards

Data Cleaning

Data Filtering

- **Dropped Columns:** 19 Columns
  - These columns had a high proportion of null values/NaNs/blanks

- **Dropped Rows:** 484,956 Rows
  - These rows had nulls values for key columns
    - Like Object Title, Accession Year, Dimensions

Setting Data Types and Encoding

- **Set Appropriate Data Types:** 'Accession Year', 'Artist Begin Date', 'Artist End Date'
  - These contain information about the year when the object/artwork had been occupied by the MoMA, artist birth dates and artist death dates
  - Converted the data type to 'Year'

- **Encoding:** 'Is Highlight', 'Is Timeline Work', 'Is Public Domain'
  - These columns are flag variables (True/False)
  - Converted the data type to 'Binary' 1/0

Data Standardization

- Fill NAs for data consistency
  - Numeric values: Filled NAs with 0
  - Date values: Filled NAs with 0
  - Text values: Filled NAs with 'Unknown’
- Normalized all text columns to lower case to increase readability and removed special characters
- Standardized all column names to snake case to increase readability
- Removed duplicate records

Data Transformations

- Transformed Column: 'Artist Gender' to flag column 'Is Female'
  - After cleaning and removing special character gender column only has the single value 'female' indicating whether an artist is female
  - Converted this to a 'Binary' variable with 1/0

- More Transformed Columns: 'Artist Nationality' and 'Country'
  - After cleaning and removing special characters, column had multiple values
  - To standardize the data, kept only the first occurrence of country (primary country)

Methodology

Emphasizing Normalization and RDBMS Utilization

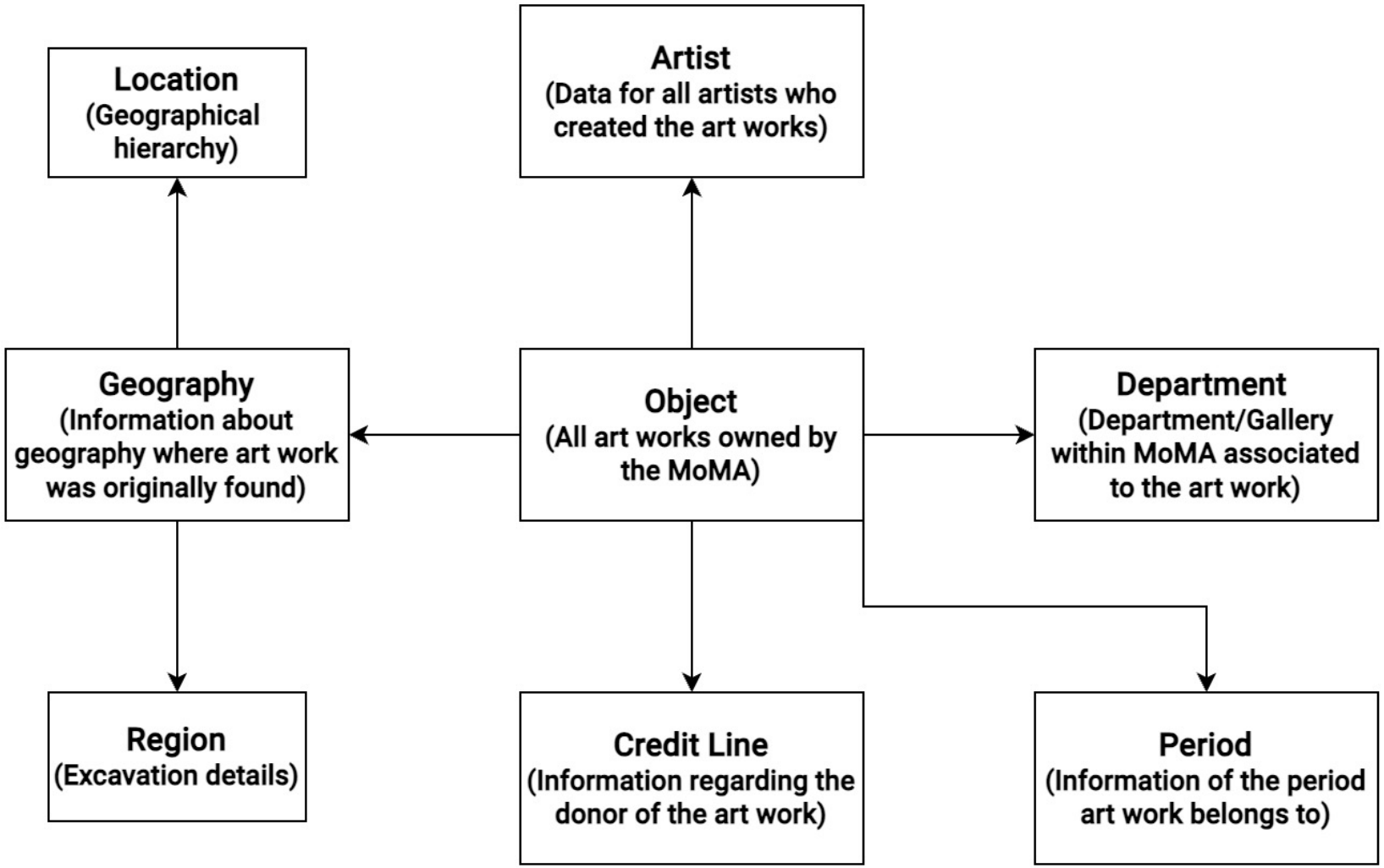
- **Methodology**
  - **Database:** Utilized a relational database management system (RDBMS) for normalization
  - **Database Name:** 'met' - Structured tables for different entities like objects, artists, donors (credit line), geography, etc.

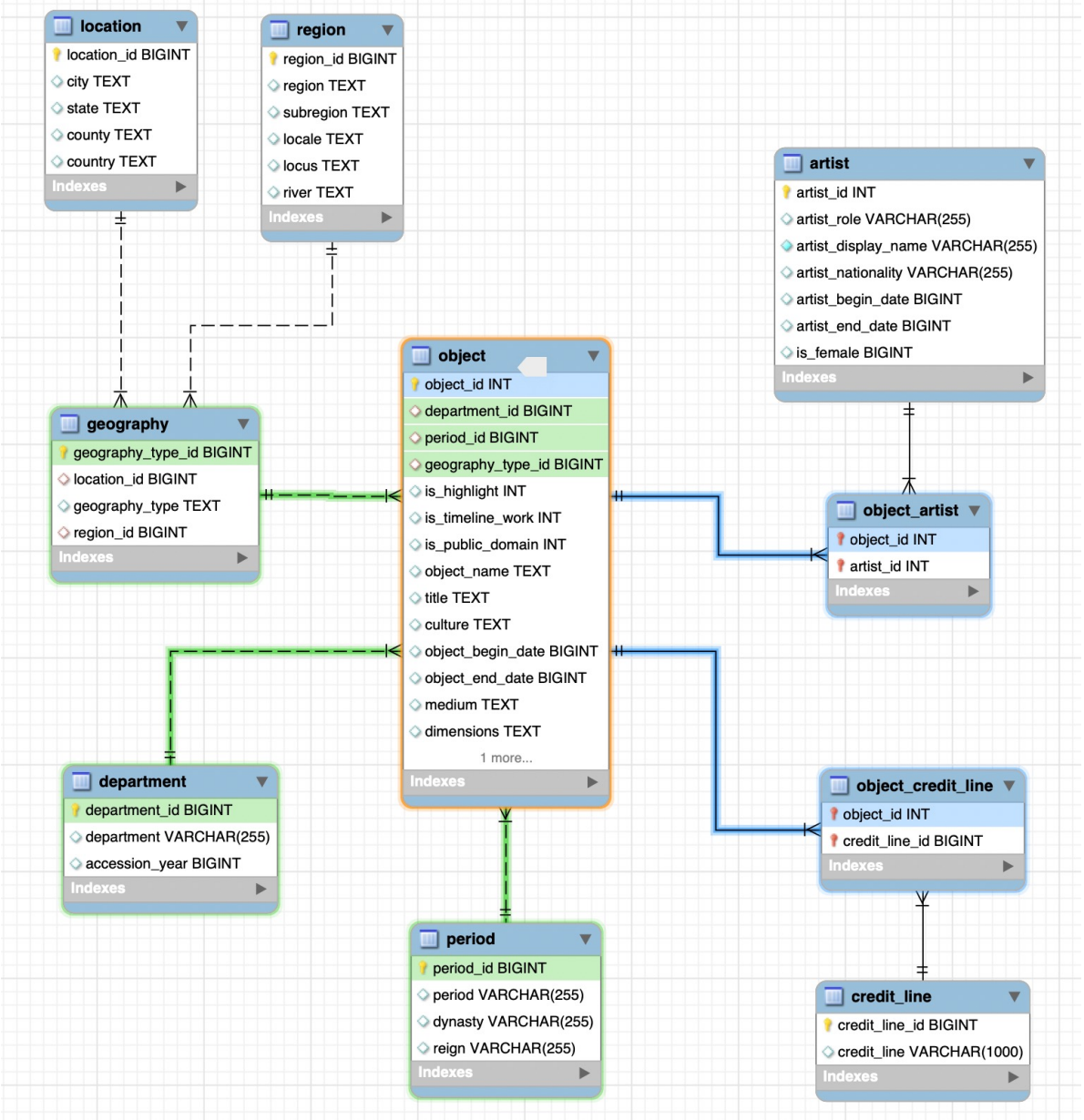
- **Rationale**
  - **Normalization:** Due to the extensive data with numerous columns, normalization was crucial for efficient data management and retrieval
  - **Why RDBMS?** RDBMS has ensured data integrity and relationships between different entities

Leveraging Python and SQL for Automated ETL and Visualization

- **Process Automation**
  - **Python Scripting:** Used Python for data cleaning and transformation.
  - **SQL Database Integration:** Automated the import of cleaned data into the 'met' SQL database.
  - **End-to-End Automation:** Streamlined the process to enhance efficiency and reduce manual intervention
    - Used python to insert cleaned data into SQL database
    - Established a live connection of the 'met' database to tableau dashboard for visualizations

- **Benefits**
  - **Consistency:** Automation ensures consistent data processing and reduces the risk of errors
  - **Time Efficiency:** Accelerated the ETL (Extract, Transform, Load) process for a seamless pipeline

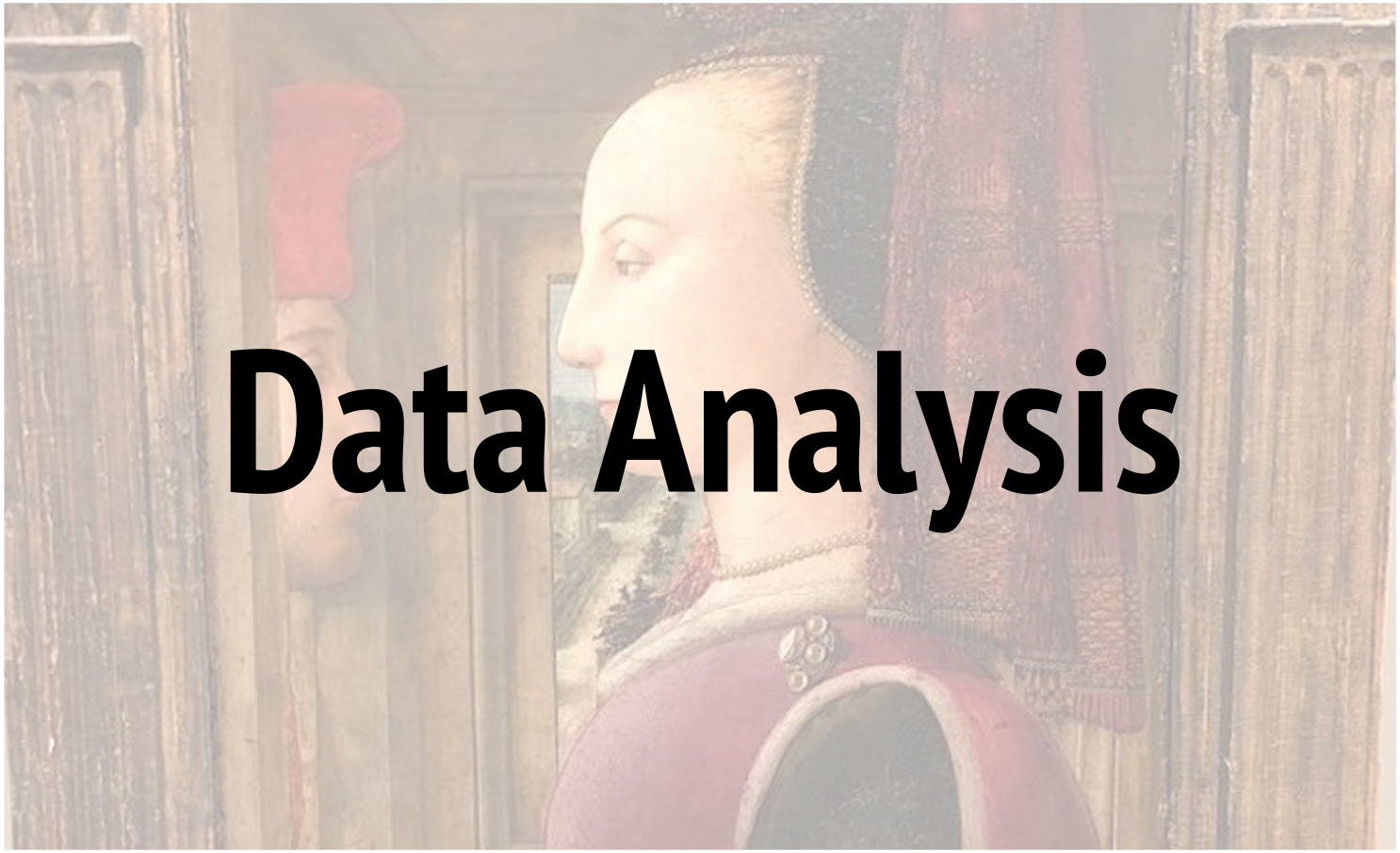

Location
(Geographical hierarchy)
Artist
(Data for all artists who created the art works)
Geography
(Information about geography where art work was originally found)
Object
(All art works owned by the MoMA)
Department
(Department/Gallery within MoMA associated to the art work)
Region
(Excavation details)
Credit Line
(Information regarding the donor of the art work)
Period
(Information of the period art work belongs to)



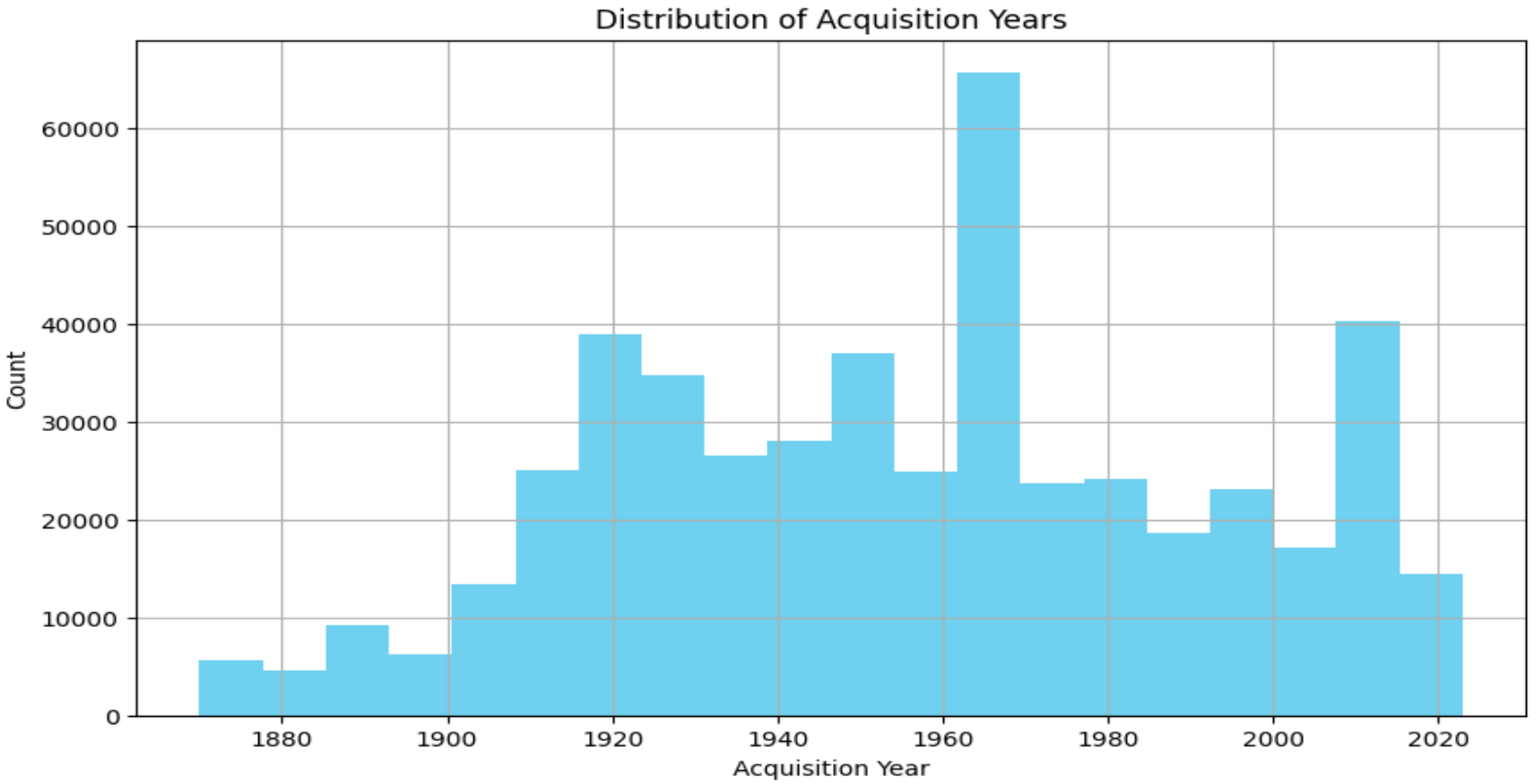
Data Modeling (EER)

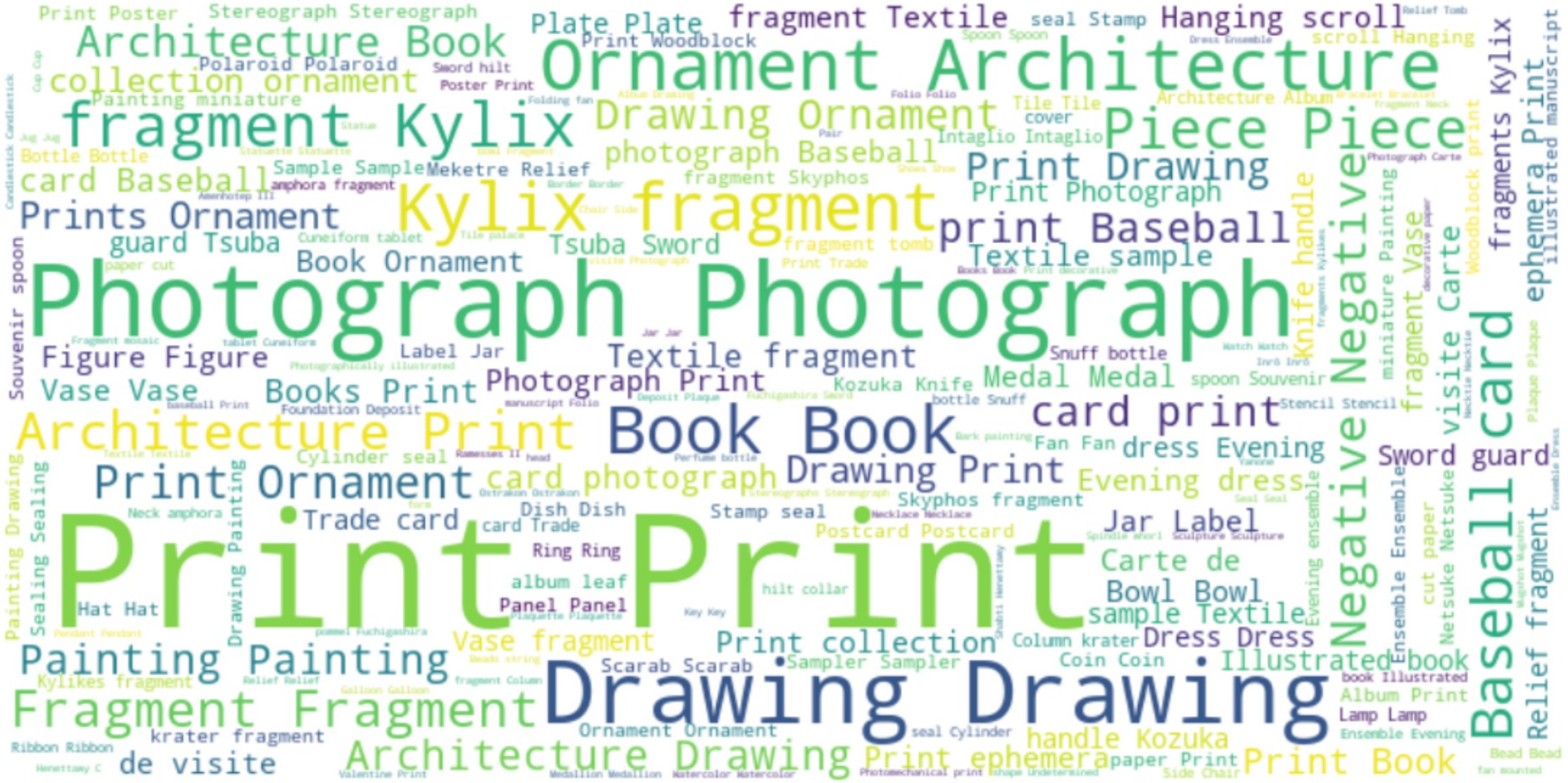
We developed a normalized relational data model with distinct entities and defined relationships between the entities. The 'met' database has following tables:

- Object
- Object_artist (mapping table)
- Artist
- Object_credit_line (mapping table)
- Credit_line
- Department
- Period
- Geography
- Region
- Location


Data Analysis

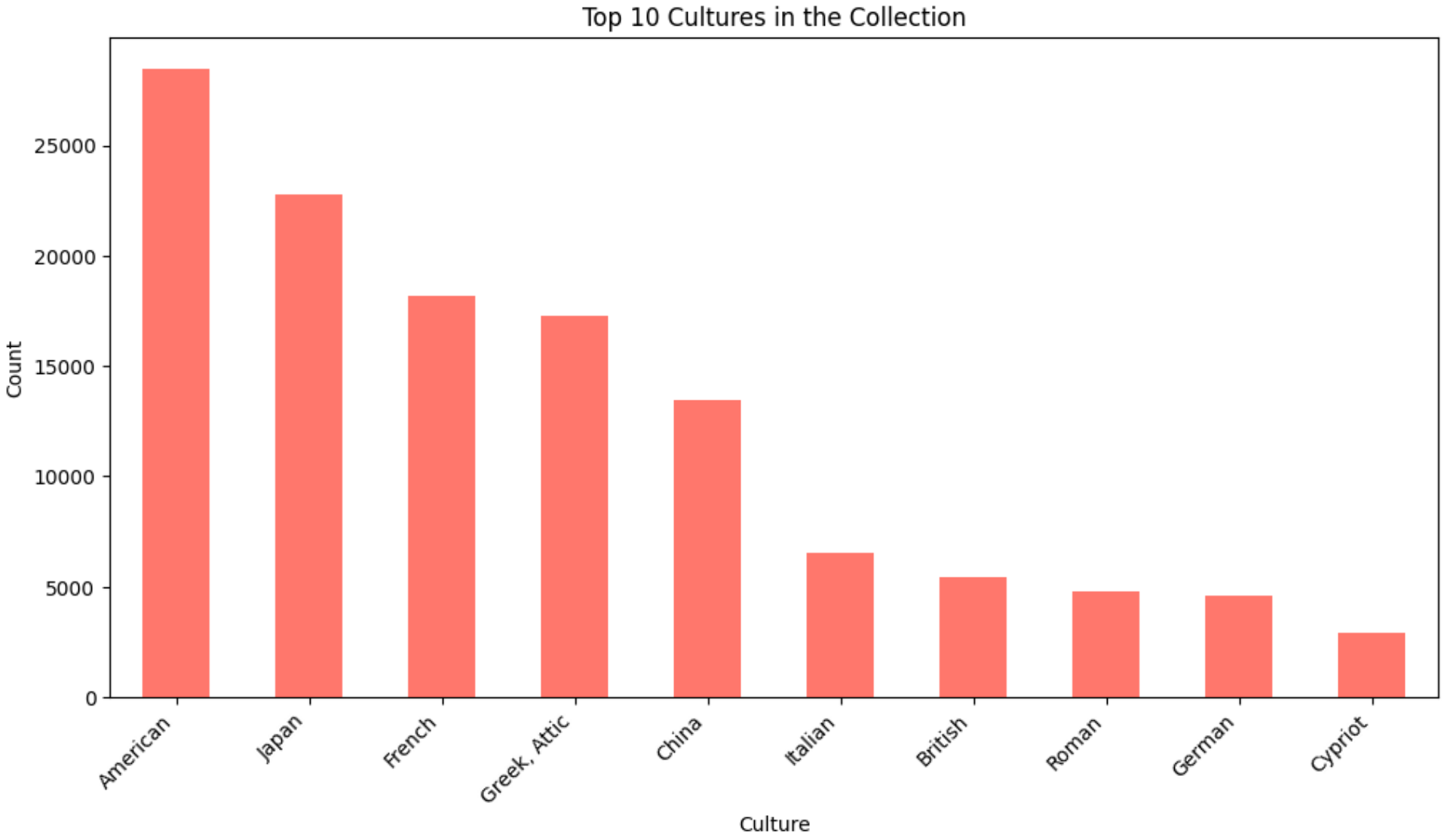
The **distribution of acquisition years** reveals that the highest number of art acquisitions occurred between **1960 and 1970**, followed by a substantial increase between 2004 and 2008

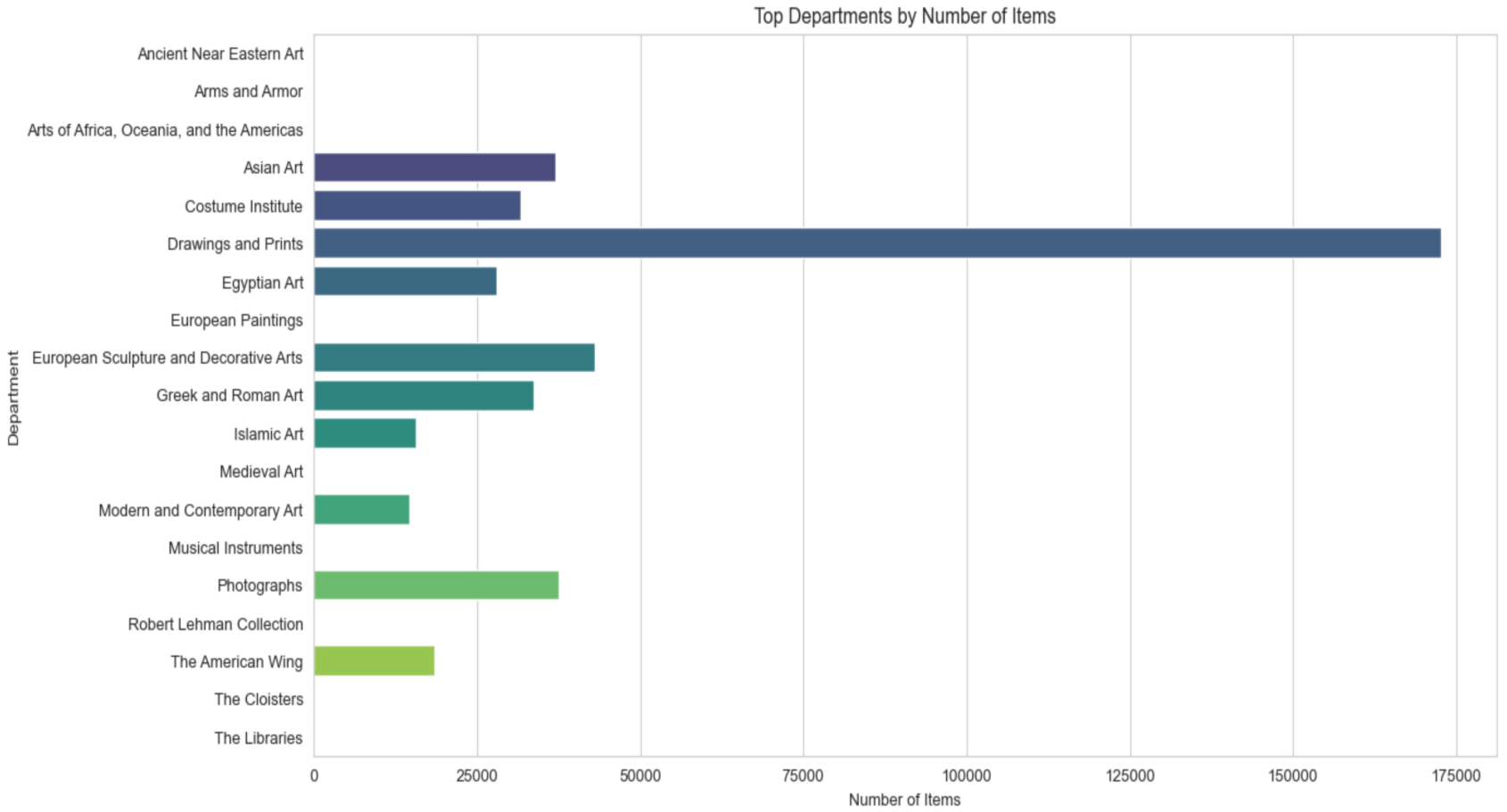




Analyzing the **object names in the collection** indicates that **Prints, Photographs and Drawings** are the most prevalent, with Prints being the most abundant, followed by Photographs and Drawings

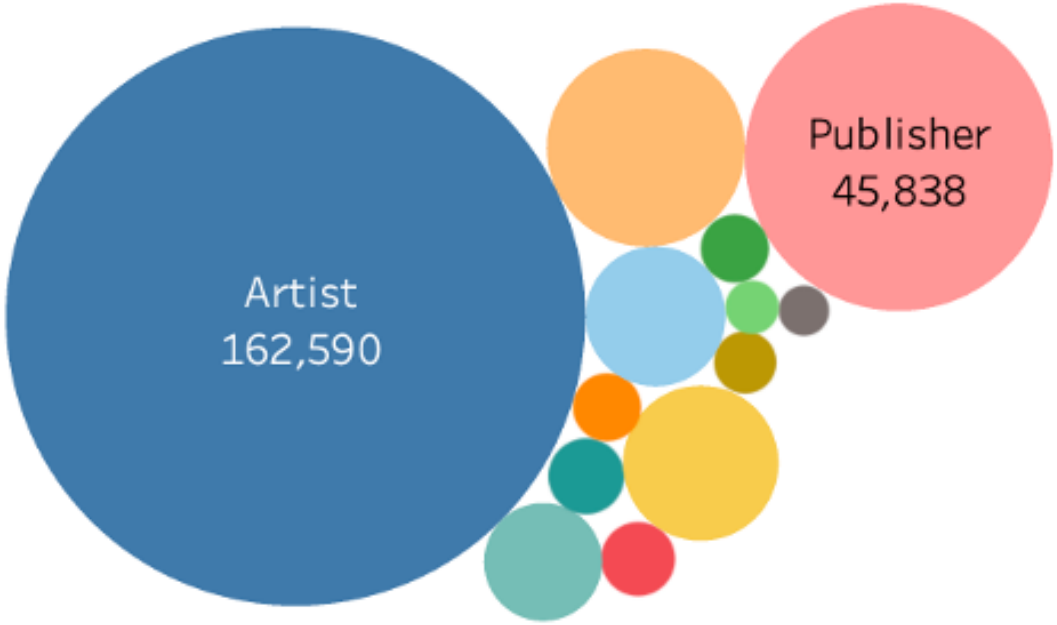
In examining **the cultures of the collected artworks**, American culture is the most prominently represented, followed by Japanese and French artworks





When examining the departments by number of items, it is evident that **Drawings and Prints** have the most artworks followed by European sculptures and Photographs.

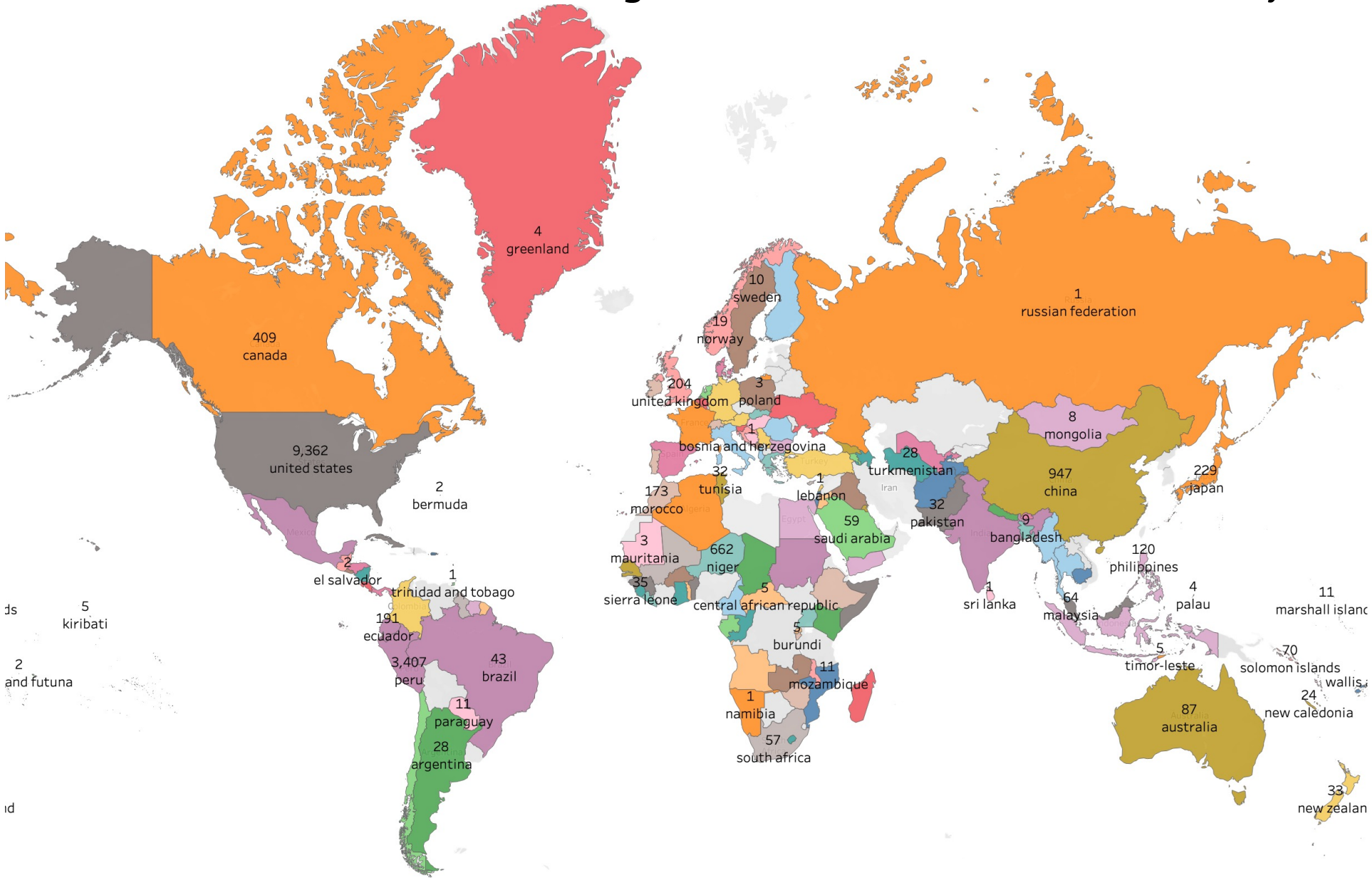
Through the analysis of cleaned data, we were able to determine the order of roles among artists, which is **Artist, Publisher, Designer, and Maker**. Additionally, we identified several artworks created by artists with the occupation of a **Sitter.**



Artist Role

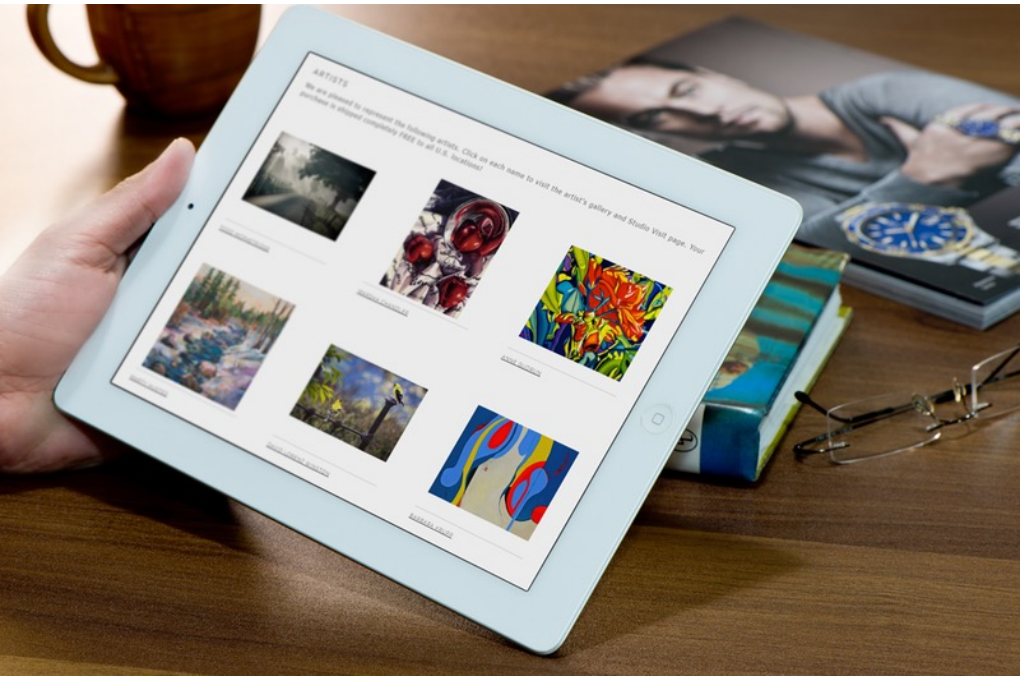
Artist Role	
Artist	162,590
Publisher	45,838
Designer	19,040
Maker	11,705
Author	9,517
Manufacturer	6,775
Manufactory	2,809
Printer	2,669
Engraver	2,266
Design House	2,114
Lithographer	1,898
Factory	1,400
Sitter	1,233

Analyzing the objects by origin, the **highest number of artworks are observed to be originated in USA**. Stakeholders can get information about artworks by their origin.



Conclusion

Strategic Recommendations for Enhancing Museum Operations: Data Management and Business Development



Data Perspective

- **Data Enrichment and Cleaning**
  - To improve overall quality of the museum's database, contributing to better research and visitor experiences
- **Consistent Digital Presentation**
  - To standardize the format of data
  - Will contribute more on the online platforms

Business Perspective

- **Data Monetization Opportunities**
  - Create premium access for researchers, scholars, or businesses interested in levering the collected data
- **Digital Engagement Platform**
  - Enhance the museum's digital presence by building an interactive platforms, virtual exhibits to attract a global audience.

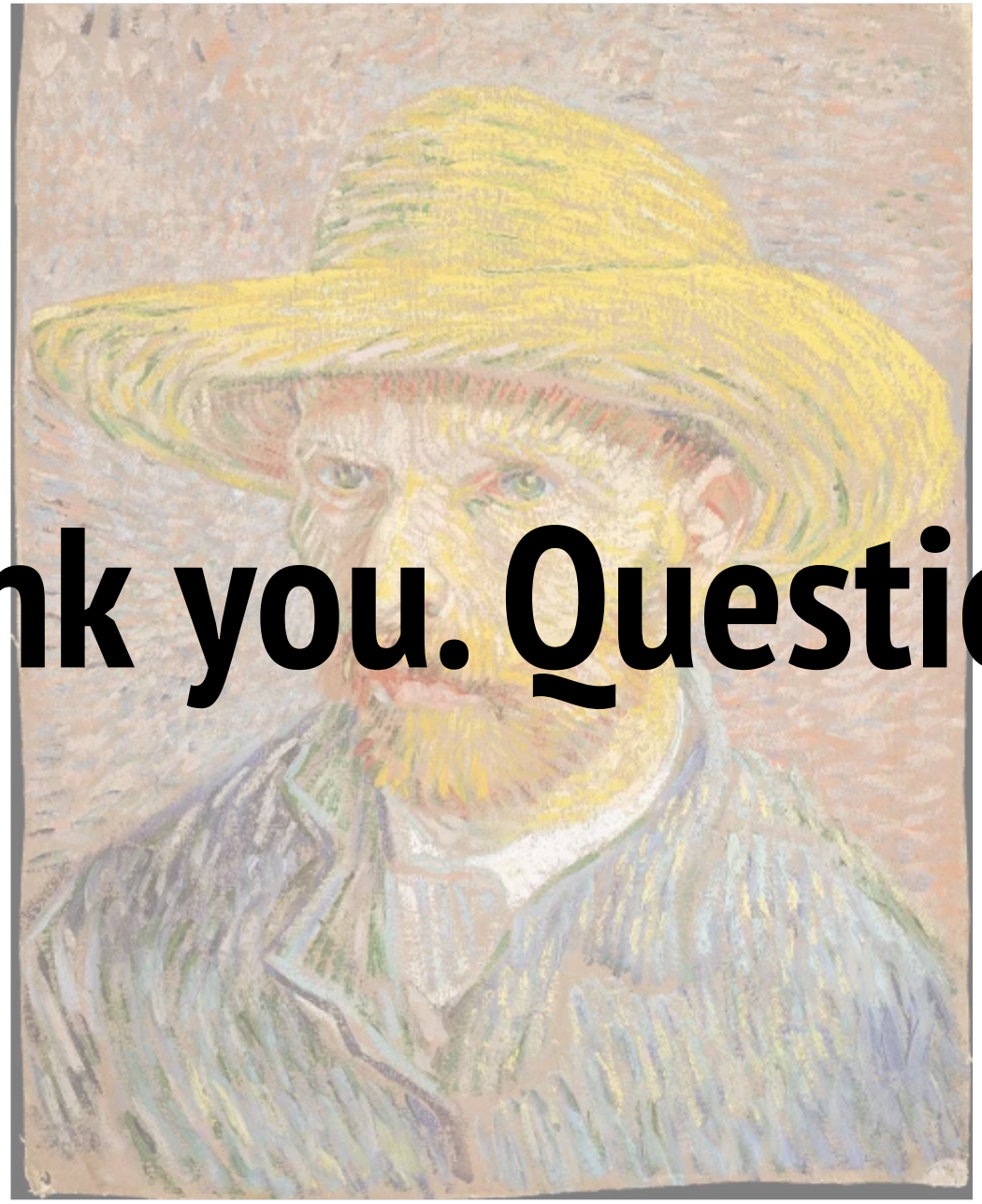
Future Scope

Additional Data

- **Visitor Demographics:** to understand the audience and tailor exhibitions and programs accordingly.
- **Social Media Engagement:** Monitor social media activity related to artworks for insights into public interest and trends.
- **Acquisition Costs:** Assess the financial performance of the collection.

Information to Provide to Stakeholders

- **Data Quality Report:** Summarize the completeness and accuracy of the current data and highlight the need of improvement.
- **Insight Report:** Present key findings and insights from the analysis.



Thank you. Questions?

Appendix

Dataset Dimensions after Cleaning

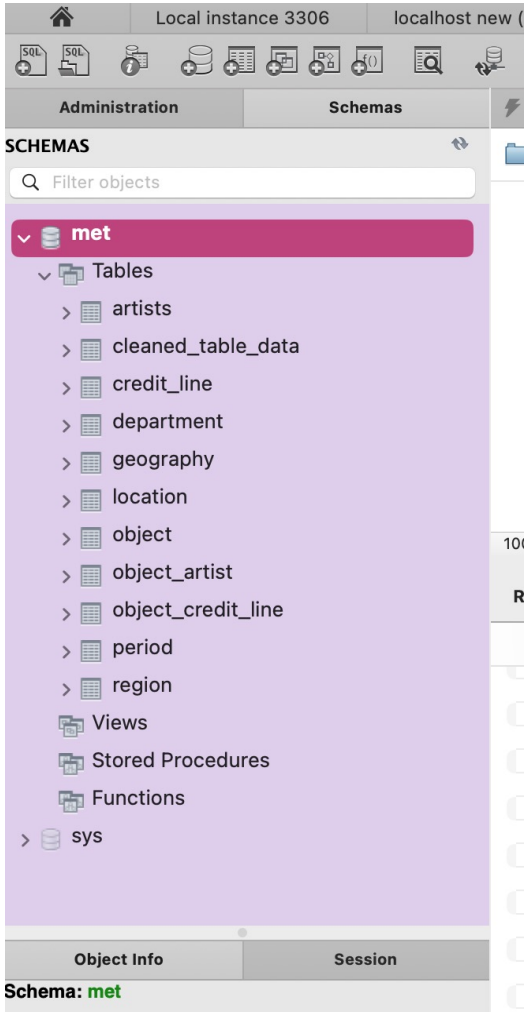
- Data Size
  - 36 Columns
  - 374,445 Rows

- Data type:
  - int64(11)
  - object(25)

Data Import – SQL Database Integration

```
import pandas as pd
from sqlalchemy import create_engine

# MySQL connection string
# Replace 'your_username', 'your_password', 'your_host', 'your_database' with your MySQL credentials
connection_string = "mysql+pymysql://root:rootroot@127.0.0.1:3306/met"

# CSV file path
csv_file_path = '/Users/mitalidighe/Desktop/UChicago Quarter 1/DEP/Final project/MoMA/Final Presentation/cleaned_dat

# Table name in MySQL
table_name = 'cleaned_met_data_updated'

# Read CSV file into a pandas DataFrame
df = pd.read_csv(csv_file_path)

# Create a MySQL connection and upload the DataFrame to the table
engine = create_engine(connection_string)
df.to_sql(table_name, con=engine, if_exists='replace', index=False)

# Close the MySQL connection
engine.dispose()
```

Relational DBMS – The 'met' database



End-to-End Automation – Tableau Live Connection