

MoMA Dataset Analysis



Nolan Wu

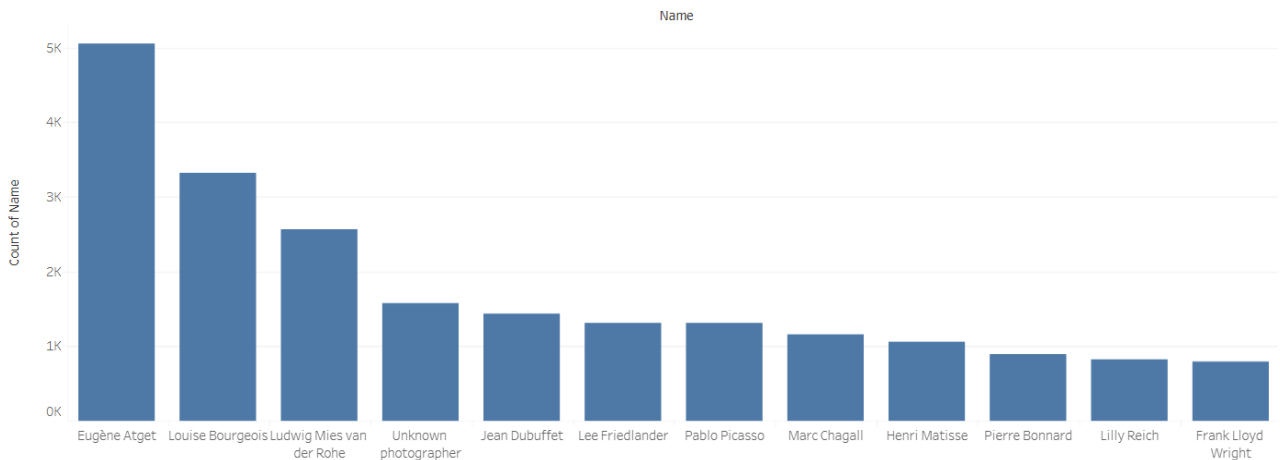
Topic Introduction

MoMA, or the Museum of Modern Art is one of the largest modern art museums in the world located in New York City. All of the artwork is “modern,” which means it's from the late 19th century and later. The MoMA dataset is a publicly available dataset published by MoMA on Kaggle.com. The dataset has 130262 artworks logged, with information about the artist, creation date, medium, dimensions, and more. I chose to analyze the MoMA dataset because I have a personal interest in art and art history, and I wanted to incorporate data science into my interests to see how data science can be used to uncover trends and insights in museums.

Data Analysis and Visualization

[Link to my Tableau Public.](#)

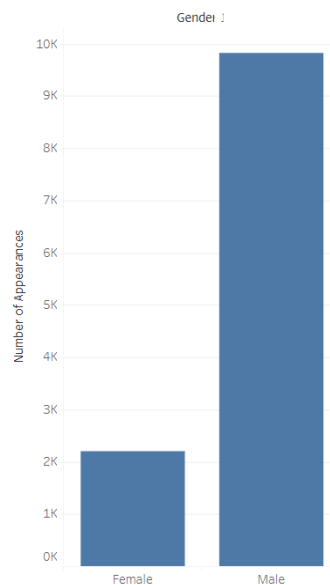
Number of pices per artist



Graph 1.

This graph displays the artists with the most pieces in the MoMA. Eugene Atget has the most pieces in the MoMA which is approximately 5000 pieces. Eugene Atget was a Parisian street photographer in the late 19th and early 20th century.

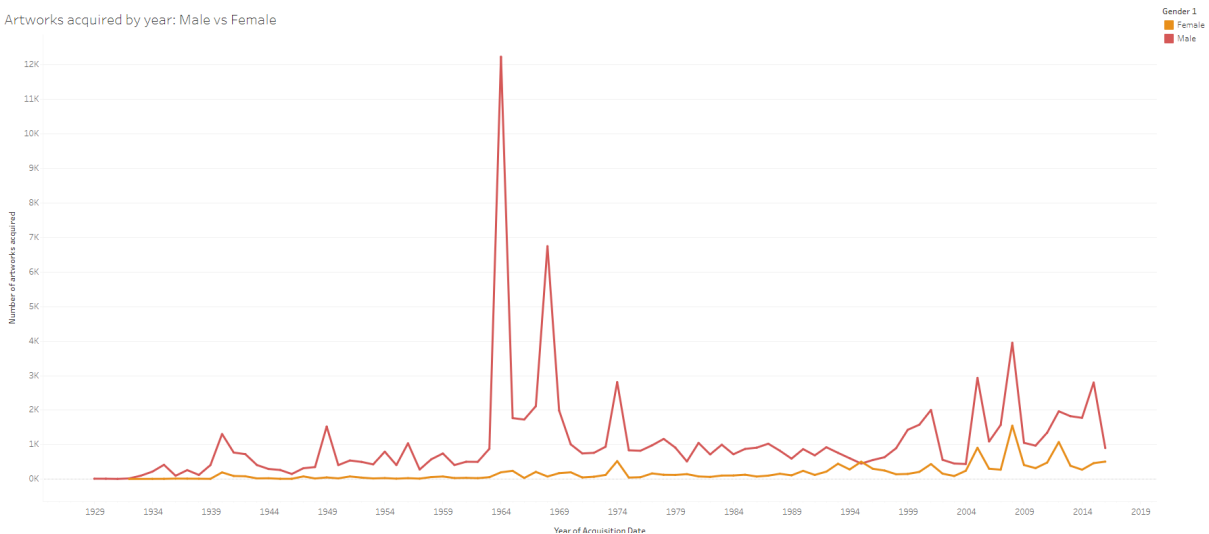
Number of Appearances in each gender



Graph 2

In Graph 2, I wanted to compare the number of artworks created by male artists vs female. Until recently, art was a male dominated field, and even though the MoMA only contains modern art, this stereotype is still prevalent within the MoMA's collection. Male made artworks outnumber female made artworks 9.8k to 2.2k.

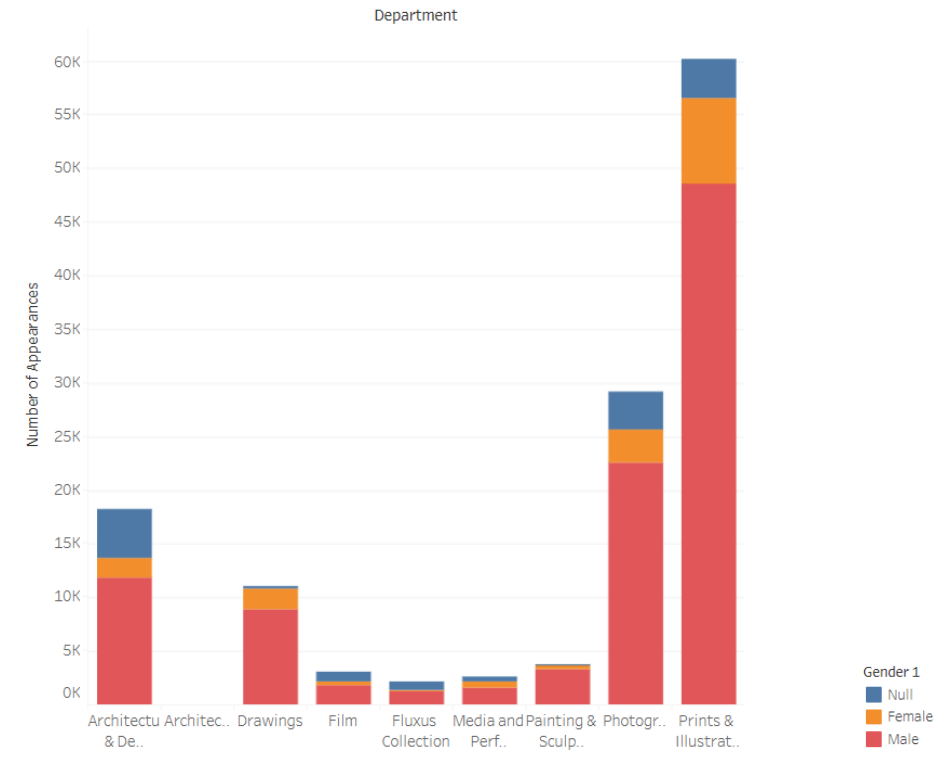
Artworks acquired by year: Male vs Female



Graph 3

I wanted to see if the acquisition date of male artworks vs female artworks changed over time. While female inclusion has definitely increased over the years, the art world still seems to be dominated by men, as the number of artworks acquired by the MoMA is still predominantly male each year.

Number of Works of art per gender by category



Graph 4

In this graph, I wanted to see if the lesser representation of women in MoMA was limited to a few categories of art, but I found that in every category of art, men outnumbered women by a large margin. This graph also shows that prints and illustrations as well as photography make up the largest proportion of the artworks in MoMA.

Note: Null comprises both unidentified artworks as well as gender non conforming artists.

A world map showing the number of countries in each region. The regions are color-coded: North America (dark blue), South America (light blue), Europe (light green), Africa (light green), Asia (light green), and Australia (light green). The numbers are as follows:

| Region | Number of Countries |
|---------------|---------------------|
| North America | 23 |
| South America | 12 |
| Europe | 50 |
| Africa | 54 |
| Asia | 48 |
| Australia | 1 |

This graph shows where the artworks in the MoMA mostly come from. The vast majority of the artworks in the MoMA come from the US, with the second largest coming from France.



I wanted to see how the MoMA incorporated more nationalities into their collection of artworks over time. I compared the number of unique nations acquired each year to the number of total artworks acquired each year. I found that the number of unique nations increased almost linearly over time, while the number of artworks acquired was mostly random.

Machine Learning

For my machine learning aspect, I used a regression model to predict the year of an artwork. I chose the columns of Name, Classification, Department, Credit, and Catalog due to their small number of NA values and their correlation to the date. Using linear regression, I ended up with errors of R-squared: 0.069 and RMSE: 32.68.

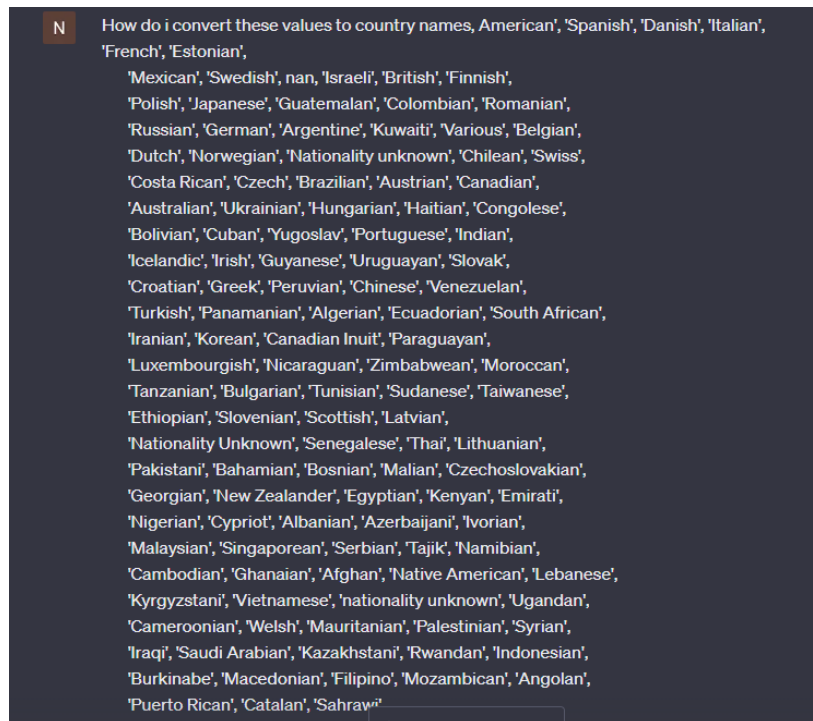
Data Analysis and Chat GPT Integration

[Link to my Colab](#)

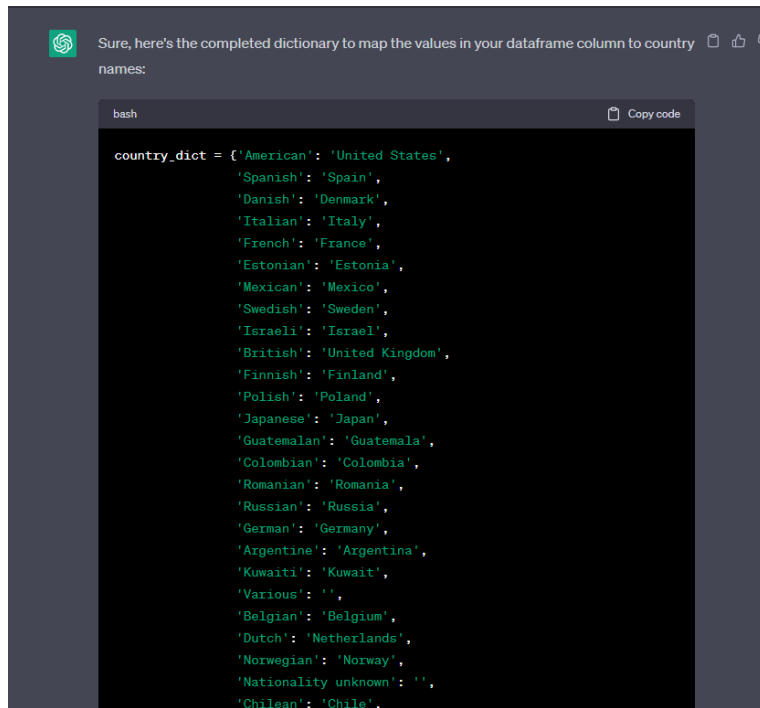
I found several interesting things I found in the dataset through data analysis. First, MoMA had 84194 total artworks in their collection. Out of the 84194 artworks, 49809 were on display, which is roughly 60% of the total collection. The artist with the most pieces in the MoMA is Eugene Atget with 5050 pieces. The largest donor of the MoMA is The Louis E. Stern Collection with 10927 pieces donated. By far the most popular department is Prints and Illustrated books with 60128 pieces. Although there were 13497 unique mediums listed in the dataset, the most popular was Gelatin silver print which had 14103 artworks.

First I used chat gpt to generate some ideas to get me started brainstorming. I asked several ways I could incorporate the idea of a museum dataset into my project, and it helped to give me an outline.

One way I used chat gpt was to do busy work for me, I had it create a long dictionary to map out country names to their adjectives.



N How do i convert these values to country names, American', 'Spanish', 'Danish', 'Italian', 'French', 'Estonian', 'Mexican', 'Swedish', 'nan', 'Israeli', 'British', 'Finnish', 'Polish', 'Japanese', 'Guatemalan', 'Colombian', 'Romanian', 'Russian', 'German', 'Argentine', 'Kuwaiti', 'Various', 'Belgian', 'Dutch', 'Norwegian', 'Nationality unknown', 'Chilean', 'Swiss', 'Costa Rican', 'Czech', 'Brazilian', 'Austrian', 'Canadian', 'Australian', 'Ukrainian', 'Hungarian', 'Haitian', 'Congolese', 'Bolivian', 'Cuban', 'Yugoslav', 'Portuguese', 'Indian', 'Icelandic', 'Irish', 'Guyanese', 'Uruguayan', 'Slovak', 'Croatian', 'Greek', 'Peruvian', 'Chinese', 'Venezuelan', 'Turkish', 'Panamanian', 'Algerian', 'Ecuadorian', 'South African', 'Iranian', 'Korean', 'Canadian Inuit', 'Paraguayan', 'Luxembourgish', 'Nicaraguan', 'Zimbabwean', 'Moroccan', 'Tanzanian', 'Bulgarian', 'Tunisian', 'Sudanese', 'Taiwanese', 'Ethiopian', 'Slovenian', 'Scottish', 'Latvian', 'Nationality Unknown', 'Senegalese', 'Thai', 'Lithuanian', 'Pakistani', 'Bahamian', 'Bosnian', 'Malian', 'Czechoslovakian', 'Georgian', 'New Zealander', 'Egyptian', 'Kenyan', 'Emirati', 'Nigerian', 'Cypriot', 'Albanian', 'Azerbaijani', 'Ivorian', 'Malaysian', 'Singaporean', 'Serbian', 'Tajik', 'Namibian', 'Cambodian', 'Ghanaian', 'Afghan', 'Native American', 'Lebanese', 'Kyrgyzstani', 'Vietnamese', 'nationality unknown', 'Ugandan', 'Cameroonian', 'Welsh', 'Mauritanian', 'Palestinian', 'Syrian', 'Iraqi', 'Saudi Arabian', 'Kazakhstani', 'Rwandan', 'Indonesian', 'Burkinabe', 'Macedonian', 'Filipino', 'Mozambican', 'Angolan', 'Puerto Rican', 'Catalan', 'Sahrawi'



Sure, here's the completed dictionary to map the values in your dataframe column to country names:

```
country_dict = {'American': 'United States',
                'Spanish': 'Spain',
                'Danish': 'Denmark',
                'Italian': 'Italy',
                'French': 'France',
                'Estonian': 'Estonia',
                'Mexican': 'Mexico',
                'Swedish': 'Sweden',
                'Israeli': 'Israel',
                'British': 'United Kingdom',
                'Finnish': 'Finland',
                'Polish': 'Poland',
                'Japanese': 'Japan',
                'Guatemalan': 'Guatemala',
                'Colombian': 'Colombia',
                'Romanian': 'Romania',
                'Russian': 'Russia',
                'German': 'Germany',
                'Argentine': 'Argentina',
                'Kuwaiti': 'Kuwait',
                'Various': '',
                'Belgian': 'Belgium',
                'Dutch': 'Netherlands',
                'Norwegian': 'Norway',
                'Nationality unknown': '',
                'Chilean': 'Chile',
```

In addition, I also used Chat GPT many times throughout the process to help me fix errors in my code, give me inspiration for visualizations, and to help me implement my machine learning



Here are some visualization ideas for Tableau using the columns in your dataset:

1. Artwork Count by Department: Create a bar chart showing the count of artwork by department. This will allow you to see which department has the most artwork in the museum's collection.
2. Artist Nationality Distribution: Create a map chart showing the distribution of artists by nationality. This will allow you to see which countries are represented most in the museum's collection.
3. Medium by Department: Create a stacked bar chart showing the count of artwork by department, broken down by medium. This will allow you to see which mediums are most commonly used in each department.
4. Artist Gender Distribution: Create a pie chart showing the distribution of artists by gender. This will allow you to see the gender breakdown of the museum's collection.
5. Artwork Acquisition Over Time: Create a line chart showing the count of artwork acquired by the museum over time. This will allow you to see how the museum's collection has grown over the years.
6. Artist Lifespan Distribution: Create a histogram showing the distribution of artist lifespans. This will allow you to see the distribution of artist lifespans in the museum's collection.

Regenerate response

N

Create a regression model to predict the date an artwork was made in my dataset and test it as well



To create a regression model to predict the date an artwork was made, you can use the following steps:

1. First, you need to prepare your dataset by selecting the relevant columns and handling missing values appropriately. You can use the following code to create a new dataframe with only the necessary columns:

```
bash Copy code

import pandas as pd

# Read in the dataset
df = pd.read_csv('moma_artworks.csv')

# Create a new dataframe with the necessary columns
df = df[['Date', 'Object Number', 'Medium', 'Classification', 'Department']]
```

Regenerate response